Aston University

Some pages of this thesis may have been removed for copyright restrictions.

If you have discovered material in AURA which is unlawful e.g. breaches copyright, (either yours or that of a third party) or any other law, including but not limited to those relating to patent, trademark, confidentiality, data protection, obscenity, defamation, libel, then please read our Takedown Policy and contact the service immediately

SPATIALLY STRUCTURED COGNITIVE MODELS OF SEMANTIC INFORMATION: THE IMPLICATIONS FOR COMPUTERISED DATABASES

JULIE COLLINS

DOCTOR OF PHILOSOPHY

ASTON UNIVERSITY

JANUARY 2006

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the its author and that no quotation from the thesis and no information from it may be published without proper acknowledgement.

Aston University

Spatially Structured Cognitive Models of Semantic Information: The Implications For Computerised Databases

Julie Collins

Doctor of Philosophy

January 2006

Summary

Existing theories of semantic cognition propose models of cognitive processing occurring in a conceptual space, where 'meaning' is derived from the spatial relationships between concepts' mapped locations within the space. Information visualisation is a growing area of research within the field of information retrieval, and methods for presenting database contents visually in the form of spatial data management systems (SDMSs) are being developed. This thesis combined these two areas of research to investigate the benefits associated with employing spatial-semantic mapping (documents represented as objects in two- and three-dimensional virtual environments are proximally mapped dependent on the semantic similarity of their content) as a tool for improving retrieval performance and navigational efficiency when browsing for information within such systems. Positive effects associated with the quality of document mapping were observed; improved retrieval performance and browsing behaviour were witnessed when mapping was optimal. It was also shown using a third dimension for virtual environment (VE) presentation provides sufficient additional information regarding the semantic structure of the environment that performance is increased in comparison to using two-dimensions for mapping. A model that describes the relationship between retrieval performance and browsing behaviour was proposed on the basis of findings. Individual differences were not found to have any observable influence on retrieval performance or browsing behaviour when mapping quality was good. The findings from this work have implications for both cognitive modelling of semantic information, and for designing and testing information visualisation systems. These implications are discussed in the conclusions of this work.

Dedication

With love,

To my parents Kath and Jim Collins thank you for your unselfish support and generosity.

To Stu , Sam, and Ian – Thank you!!

Acknowledgments

I wish to convey my sincerest heartfelt thanks to my two supervisors Dr. Paul Furlong and Dr. Steve Westerman. They have contributed to this work in so many ways and both have saved me from defeat throughout.

I must also thank the wonderful staff here in the Psychology Department at Aston University, valuable advice and support has come from all quarters, academic, administrative, and technical. In particular I would like to thank Niteen Mulji, and Jon Wood for their patience and help.

I would like to thank Dr. Roy Davies for his wisdom during the early stages of this undertaking, and Timothy Cribbin who helped me develop my understanding of the field in the beginning, and helped me master the complex algorithms needed to calculate the results in Chapter Two. To Dr Gareth Barnes who provided invaluable help and advice in a variety of ways throughout, in particular with programming software to aid my analysis, I extend my gratitude.

I also wish to thank everyone who has helped and supported me but who I have not named (you know who you are). Without such a strong network of support, friendship and guidance this thesis would not have been written.

Finally I would like to thank my friends and family who have put up with me for the last four years. I know it hasn't been easy, I'm sorry and I promise to provide recompense. Your love and support has kept me going throughout, and I would not have survived the journey without you.

INDEX

INDEX OF FIGURES AND TABLES

FIGURES

TABLES

1 Spatial-semantic Processing, Information Visualisation, and Spatial Data Management Systems

The aim of this thesis is to gain a better understanding of automatically generated internal/cognitive models of semantic distance and to show how these can be supported in computer-based representations of information spaces. Spatial-semantic mapping allows the semantic properties of documents within an electronic database to be visually conveyed to the user via a graphical interface. In the research reported in this thesis the consistency of mapping of cognitive conceptual spaces and computerised information spaces is investigated. The degrees to which individual differences in cognitive ability are implicated in the way people use such systems are also examined. By examining the interaction between cognitive information spaces and computerised information spaces which both use spatial-semantic mapping (it is argued), the research follows two perspectives – cognitive psychology and human-computer interaction.

A cognitive psychology perspective is adopted to add support for theories implicating conceptual spaces and the spatial processing of meaning in individuals' structuring of semantic information. At the same time a human-computer interaction (HCI) approach is used to facilitate the study of individual differences in users' cognitive maps of a specific type of information retrieval (IR) system (i.e. a spatial data management system (SDMS) – see section 1.4.2). Furthering the understanding of the cognitive structure of semantic information within a specific IR paradigm can help with the design of effective IR systems, while providing valuable insights for psychologists regarding the organisation and acquisition of mental maps of information spaces, thus providing cohesion between the cognitive and HCI themes.

The current chapter firstly presents the research questions – section 1.1, and then proceeds to provide an overview of the principal fields of research that impact on this thesis (i.e. spatial models of semantic cognition – section 1.2, and information retrieval adopting visualisation and spatial data management tools – section 1.4), and finally details the reasons for conducting the current research, and the structure of the thesis – sections 1.5 and 1.6.

1.1 Principal Research Questions

Existing theories of semantic cognition (e.g. Jackendoff, 1983; Gardenfors, 2000) propose models of cognitive processing that occur in a conceptual space, where 'meaning' is derived from the spatial relationships between spatial locations of the mapped concepts. Processes for generating or mapping concepts within this space involve the same or similar cognitive structures and rule based algorithms that are employed by the perceptual system when encoding and processing sensory stimuli (e.g. Shepard, 1957; Jackendoff, 1987). A review of some of these theories is given in section 1.2 'Semantic Cognition'.

A direct implication of semantic models is that, when organisation of people's internal cognitive space reflects the organisation of the external space they are working in (or vice versa) an accurate mapping between internal and external representations is facilitated to optimise performance. It is this assumption that provides the platform for the current research.

The principal research question can be formalised as follows: -

What are the implications of a spatial-semantic theory of cognition for IR when

employing systems that utilise inter-document spatial relationships to visually convey the underlying semantic features of database contents?

Marchionini & Shneiderman (1988) suggest that, within a 'framework of information seeking', IR efficacy is based on outcomes which themselves are measured in terms of production (e.g. the quantity and accuracy of retrieved items) and / or processes (e.g. users' behaviour while performing the task). Based on this distinction, the following questions will be experimentally addressed by the experiments reported in Chapters Three (questions 1a and 1b), Four (questions 2a and 2b), and Five (question 3): -

1a) Is performance, operationally defined as production during IR tasks (see previous paragraph for definition of production), influenced by the degree to which the organisation of documents within an information space facilitate users' acquisition of a good mental model by promoting assimilation of the semantic information contained in that space, with the user's existing cognitive model?

1b) If performance is influenced by the spatial-semantic organisation of material, are individual differences in cognitive ability contributing factors to the 'goodness of fit' between the user's cognitive maps of the information space and their internal conceptual spaces?

2a) Is behaviour, operationally defined as the processes used during IR (see previous paragraph for definition of processes), influenced by the synchronicity between cognitive space and database mapping? For example are patterns of browsing or adopted browsing strategies and decisions of relevance affected by the 'goodness of fit' between the physical mapping of documents in the information space and users' conceptual mapping?

2b) If a relationship between performance and behaviour exists are individual differences in cognitive ability a contributing factor?

3) By differentiating performance as referred to in question 1 and behaviour defined by the interaction between users and the IR system, can a relationship between these two factors be identified that accounts for observed differences in performance? If a relationship is shown to exist can it be attributed to the coherence between internal and external maps of the information space?

These questions were addressed by examining individuals' browsing behaviour and performance during IR tasks using a visual information retrieval interface (VIRI). A spatial data management system (SDMS) form of VIRI was used to present database contents (documents) as objects in a virtual environment (VE) by means of spatial-semantic mapping. Proximity between documents reflected their semantic similarity, such that the degree of relatedness between documents was inversely correlated with the distance between them in Euclidean space. The semantic similarity between documents was assessed using automatic text analysis (ATA) models that use vector space modelling (VSM) to produce high-dimensional models of semantic content based on term by document matrices. A full and detailed explanation of VSM based ATA is presented in Chapter Two in which two models; n-gram based (ATA) (in which terms are represented as letter strings) and latent semantic indexing (LSI) (in which words are represented as terms) are evaluated. Section 1.4 describes more fully VIRIs, SDMSs, and spatial-semantic mapping.

1.2 Semantic Cognition

Theories exist that propose that semantic cognition, i.e. cognitive processing of meaning is structured in a spatial way (e.g. Jackendoff, 1987; Gardenfors, 2000). While there is disagreement between these theorists regarding the specifics of such models (some of these will be reviewed), it is generally accepted that cognition occurs in a mental, cognitive, or conceptual space and that concepts are spatially mapped into these spaces. The spatial organisation of concepts and/or features of the concepts, whether absolute in terms of Euclidean distance and geometric relations (e.g. Tversky, 1977; Gardenfors, 2000), ordinal along multiple higher-order or reduced binominal dimensions (e.g. Osgood, 1969; Brooks, 1998), or metaphorical, dynamic and context driven (e.g. Lakoff, 1987), is a catalyst for the effective and efficient processing of meaning.

Why do we need meaning and what is meaning? In order to operate in a three-dimensional interactive environment, to experience the physical world around us, and to communicate with others, we need to represent external information internally and *vice versa*. Meaning therefore refers to the usable information that can be elicited from these internal and external models and their interactive relationships. This derivation can be achieved via a mental model (e.g. Tolman, 1948) where the function of the model is to convert incoming sensory signals into some 'meaningful' (i.e. useable) mental representation. In order to communicate and share our personal experiences and internal models of the world with others, we represent elements of our model to them using sensory outputs that can be converted into an internal mental model by the receiver using their perceptual and semantic systems (Glenberg, 1997).

Clearly this reciprocal relationship between internal and external models is as necessary for our own personal operation in the world as it is for everyone else. This relationship is not restricted to inter-person reciprocity, motor tasks, for example require person to object interaction. For example drinking a cup of coffee requires projection of information from the object i.e. the cup and the liquid inside the cup, regarding how to pick it up (where the handle is) and how hot the liquid is (temperature) etc. The state of the object is changeable based on changes in the individual's behaviour prompted by their interpretation of the original information projected by the object. The change of state will then be projected to the individual and re-interpreted promoting continuation of the cycle and continued interaction/communication (e.g. Glenberg, 1997).

Not all interaction is time bound. For instance, while writing a PhD thesis the author will not know whether they have successfully conveyed their internal model of the research conducted until the work has been read and discussed at *Viva*. The relationship between the author's internal model, the external model represented in the written work, and the reader's internal model are, however, dependent on feedback. If the thesis were never read or discussed the author would never know whether the message had been clearly conveyed. As a result learning on behalf of the author or the reader in terms of adapting their respective internal models would not occur.

As cognate animals our models also include abstract constituents that enrich and personalise our perception of the world. For instance emotions can be related to specific encounters or experiences. As a species we are unique in having a rule based productive language that enables us to convey much of the personalised elements of

our internalised models (thoughts) and while animals can express basic survival messages such as fear, we can share much deeper experiences (e.g. love, hate, joy, pain (emotional as well as physical), empathy, disappointment etc.). It should be emphasised at this point that language and thought are different concepts in that language is used to express thought but thought is much richer than language. Specific languages are unique and contain elements that do not translate thought however, as far as we are aware, remains transient across different languages (notwithstanding individual personalisation/individual differences). However as Jackendoff (1995) suggests, interpretation of meaning within language, cognition, and perception are governed by similar rule based processes that he terms 'mental grammar'. Thought is processed in a conceptual space governed by a conceptual grammar. Jackendoff goes on to suggest that spatial relationships and spatial cues are key factors within this mental grammar. Support for the theory of spatial processing of language and thought can be seen in existing literature in both the fields of language and cognition (e.g. Landau & Jackendoff, 1993; Glenberg, 1997; Brooks, 1995a; and Siakaluk, Buchanan, & Westbury, 2003).

Our cognitive processes have evolved over time to perfect understanding of the three-dimensional world in which we function and it therefore makes sense that many of these processes are spatially organised internally. Glenberg (1997) suggests that the purpose of memory and conceptualisation is to assist perception and action in a three-dimensional world and as such memory, language, conceptualisation, imagery, perception, and meaning are all embodied. Glenberg makes a strong case for the embodiment of meaning, building on the work of Lakoff (1987) in which he proposes that conceptual or mental structures are embodied. The term embodied is used to refer

to the way information is encoded as a direct result of the type of bodies we have, and the actions necessary to operate these bodies in a three-dimensional world i.e. using spatial reference to the relationship between our bodies, and our conceptualisations in order to instigate action. Jackendoff (1983) also supports the idea of embodiment of conceptual structure suggesting that semantic structure and conceptual structure are not distinct and the organisation of this "natural language semantics" is derived using the same processes as visual perception described in Marr's theory of vision.

Marr (1982) suggested that visual perception is achieved via a three-stage model that firstly requires signals received through the eye to be mapped as a "primal sketch". This organises boundary information acquired from the incoming signals, a "2½D sketch" which organises the geometric properties of visible surface features of the target object, and a "3D sketch" which organises the spatial elements of the target in relation to the external environment. These three separate sketches are combined in conceptual space to provide the final perception, which includes inferred information about non-visible features and functions. Here the argument for spatial organisation within the conceptual space is supported by the fact that in addition to functional and experiential information related to the concept of the object, spatial elements of the concept's/object's relation to the world are necessary for the 3D sketch, and geometric and boundary properties are necessary for the primal and 2½D sketches, such that the formation rules for semantic/conceptual structure can be mapped directly to Marr's formation rules Jackendoff (1987).

Linguistic researchers are generally concerned with a rule-based code (grammar) that allows us to communicate our mental models to others through verbal and auditory

mediums. Spoken and written language involves the allocation of abstract markers (words) to semantic elements of our cognitive model or incoming sensory information. The grammar that governs human language consists of three elements phonetics (the sound units that produce words), syntax (the formal rules by which words can be constructed into sentences), and semantics (the overall message that is conveyed by the phonetics and syntax). In linguistics research, the semantic element of grammar is considered a very specific product of phonetics and syntax rather than in the wider context of meaningfulness that conceptual semantics refers to, in psychological research (McNamara & Miller, 1989).

While this relationship and definition of semantics is important to the study of language production and comprehension, it is too narrow a focus for Psychology where conceptual semantics are considered more complex. For instance the Gestalt view that the meaning of a concept is more than the sum of its parts, implies that conceptual meaning derived from a passage of text is much richer and deeper than the meaning that is derived from each sentence separately – see (Glenberg, 1997).

As such linguistic semantics doesn't adequately explain how accurate interpretation can be derived from sentences with more than one meaning, due to effects of context, individual differences in higher order interpretations, or fuzzy concepts etc. For instance the same sentence "I'm tired of writing this thesis." can have more than one meaning 1) "I'm mentally fed up with doing this PhD", or 2) "I'm physically tired of sitting here at the computer writing". Using truth-value propositional logic which forms the basis of linguistic semantic theories of human meaning, the receiver would not know which of the definitions was correct. However, if the receiver combined this

sentence with other sensory information exuded by me (e.g. yawning) they would be able to infer that I was physically tired and may therefore select the second definition, Alternatively if the receiver of the message had themselves completed a PhD they may select the first definition on the basis of their own experiences and emotions during their writing-up period. Additionally judgements can be made from the information obtained by combining several sentences within a piece of text. This problem with propositional theories of meaning is highlighted by Putnam (1998) who shows that pairs of sentences that have different meanings, share exactly the same truth values and therefore could not be encoded / decoded accurately (if at all) on the basis of truth values and propositional logic alone. Theories that support a conceptual space in which semantic processing occurs address these limitations by enabling context to be considered and represented as spatial relationships (e.g. Osgood, 1969; Gardenfors, 2000).

It is the issue of context, and the definition and measurement of meaning in a conceptual framework that provides the platform for the research in this thesis. Theories from the fields of semantic linguistics, semantic memory, information processing, and cognitive modelling are combined in this work to address this issue. While each of these fields can be and are researched separately they are all necessary to the comprehensive understanding of meaning. It is not the intention to review in detail or defend specific theories of meaning (for this see McNamara & Miller (1989)), however the case for adopting spatially driven theories of meaning as a template for semantic processing within this study is explained.

1.3 Spatial-semantic Cognition

The term spatial-semantic processing derives from the field of computer science as a result of using graphical user interfaces to spatially convey the intrinsic/semantic properties of the information displayed (Herot, 1980). Within psychology this term can be used to encompass spatial theories of cognitively processed meaning.

Spatial-semantic processing offers a reciprocal means of mapping incoming information about external factors to internal concepts existing in an individual's conceptual space in order to derive or attribute meaning. Meaning is extracted from the spatial relationships between concept representations within a psychological space that operates using the physical laws of space e.g. (Gardenfors, 2000). As previously suggested therefore, spatial-semantic processing marries perceptual processing, e.g. object recognition, with conceptual processing i.e. language and thought, suggesting that the same psychological structures are used in both (Jackendoff, 1987).

Theories such as those already identified in section 1.2 about semantic processing have their grounding in models that suggest operations involving similarity, discrimination, and generalisation decisions about perceptual stimuli and concepts occur in a psychological space. These ideas began to emerge during the 1950's and 60's at the height of the Information Processing theoretical movement. In language research, formulae and algorithms applying physical laws of metrics to an individual's psychological space, were proposed and tested as a means of predicting stimulus/response relationships. Such models began to combine behaviourist and cognitive explanations of language and semantics (Shepard, 1957).

In 1957, Shepard proposed a multidimensional-scaling-choice (MDS) model to explain performance in identification and categorisation tasks. He suggested that stimuli are represented as points in a psychological space, and the probability of two separate stimuli eliciting the same or similar responses is a "monotonically decreasing function" of the distance between them (Shepard, 1986). This basic model continued to be supported and expanded upon within psychophysical and perceptual research by numerous theorists (e.g. Nosofsky, 1986; Ennis, 1988; Reid & Staddon, 1997; etc.).

At the same time theorists in cognitive psychology had turned their attentions to memory and how concepts could be represented in semantic memory to facilitate comprehension and language (see Kintsch, 1980). These theories centred on two distinct models and their extensions – feature comparison models (e.g. Smith & Medin, 1981; McCloskey & Glucksberg, 1979) and semantic network and spreading activation models (e.g. Quillian, 1968; Collins & Loftus, 1975).

Feature comparison models proposed that semantic representation occurs through the decomposition of concepts into semantic constituents that are compared with other concept constituents. In this way meaning is derived through the degree of similarity between concepts based on their semantic structure. As (Jackendoff, 1987) highlights, semantic and conceptual structure are the same, therefore spatial-semantic processing is not contradictory to a feature comparison model. It can be argued that within a feature comparison model of semantic memory, concepts conveyed through language are represented in the same way as real world objects are conveyed through sensory motor inputs. Representations are formed on the basis of the concept or object's internal structure using basic formation rules applied to the spatial relationships

between components; these can then be compared to previously stored representations.

Semantic network models (e.g. Quillian, 1968; Collins & Loftus, 1975) suggest that representations of concepts occur as nodes within a network. The network is formed by links between the nodes that represent the shared properties of those concepts and weighted with "criteriality" values (i.e. numbers indicating the importance of that particular property to that particular concept). The links are multi-directional (and can link more than two nodes) with weightings that vary dependent on the direction of the link, a property may be more important to one concept than another. The meaning of a concept is derived from the entire network of connected nodes. Again this type of model utilises the spatial relationship between concepts to process meaning, however, distance is used to judge semantic similarity or relatedness not to derive conceptual structure per se. Semantic distance is the length of the shortest path between two nodes sharing a property and semantic relatedness is the aggregate length of the all the shared links between two nodes in the network. This means that whereas two concepts are close in space i.e. have a short semantic distance, they may not be closely related or have a strong semantic relationship, if for instance they share relatively few links. Here it could be argued that dimensionality is key to semantic distance, if all links were thought of as separate dimensions and multi-dimensional scaling such as that used by (Shepard, 1986) was applied, the resulting coordinates within the individuals' psychological space would transform distance to a direct correlation measure of relatedness or similarity (e.g. Schvaneveldt et al., 1985). Again spatial-semantic processing can be considered complimentary to semantic network models of memory.

At the same time as memory research was beginning to consider spatially structured models of semantic encoding, spatial metaphors were being used within the measurement or interpretation of meaning. Osgood's (1969) 'Semantic differential technique' is a classic theory that uses distance and dimensionality as metaphors for distinguishing and measuring semantics. The theory defines meaning as the combination of a set of similarity judgements based on bi-polar terms (e.g. same, not same), the semantics of a concept can therefore be measured on the basis of its judged 'distance' along many bi-polar dimensions. It was found that when factor analysis was applied to these multiple dimensions there were three primary dimensions, evaluation, potency and activity along which all similarity judgements were made.

A more recent model also utilises the spatial properties between concepts in a psychological space to identify how users of bibliographic databases use descriptor terms to locate relevant documents. This 'semantic distance model' (SDM) of relevance (Brooks, 1998) was designed as a model for explaining judgements of item relevance within bibliographic descriptor hierarchies. The 'descriptor hierarchy' used by Brooks is a generic tree of subject index (e.g. descriptor) terms taken from the Educational Resources Information Center (ERIC) thesaurus. The trees include varying numbers of descriptor terms that consist of the broadest term at the top (i.e. a term under which the other terms can be categorised) with the narrowest term (i.e. a specific term that can't be/isn't further sub-categorised). 'Topical-level' descriptors are assigned to documents in the database by indexing staff, and can fall anywhere between the top of the tree and the bottom. 'Broader descriptors' are defined as those above the topical-level descriptor within the hierarchy, and 'narrower descriptors' are those below the topical-level descriptor. An example of a generic descriptor tree used

in his study is given by Brooks as: -

> "
> :::: Services
> ::: Information Services
> :: Information Processing
> : Documentation
> Bibliometrics
> . Citation Analysis
> .. Bibliographic Coupling
>
> In this example, "Bibliometrics" is the entry; increasingly broad descriptors are indicated by a greater number of colons, and increasingly narrow descriptors are indicated by a greater number of periods. "Services" is six semantic steps away from "Bibliographic Coupling"."
>
> Brooks (1995) page 108

Semantic distance is defined as the number of steps between descriptor terms within the hierarchy; this is exemplified in the above extract in which "Services is defined as six steps away from "Bibliographic Coupling".

In a series of studies using the SDM, Brooks (1995a; 1995b; 1997; and 1998) demonstrated that judgements of relevance to query decline systematically with semantic distance. For example when descriptor terms from a generic tree were rated for degree of relevance to a document that had been indexed either at the top of the tree or at the bottom of the tree, the ratings decreased monotonically (e.g. Shepard's (1986) model) with semantic distance. In addition a 'semantic direction' effect was demonstrated. The distance to judgments of non-relevance was shorter when judging the relevance of terms to a document indexed under a broad term that indexed to the bottom of the tree. In other words when moving up the hierarchy, terms remain relevant for a greater distance than when moving down the hierarchy. Moving down the hierarchy becomes more specific and less general.

The work of Brooks developed through the advent of electronic databases and the need to identify ways of indexing the contents on the basis of how users would judge

documents to be relevant to the index terms assigned to them. Strong evidence that the use of distance within a cognitive space is used to determine meaning is provided in a body of work by Burgess and Lund that developed an automated method of text analysis that models semantic memory and language comprehension (e.g. Lund & Burgess 1996; Burgess, 1998; Burgess & Livesay 1998; Burgess, Livesay, & Lund 1998). The HAL 'hyperspace analogue to language' model of memory demonstrates how computational assessment of term/word co-occurrence between large corpus of text mapped into a high dimensional space, can reflect language comprehension and memory for meaning.

HAL uses vector space modelling similar to that referred to later in this chapter and described in detail in Chapter 2. Inter-term relationships with other terms are weighted on the basis of their co-occurrence and dependent on how many words separate them (i.e. if two words always occur next to each other they would have the maximum weighting). A window ten terms wide moves one term at a time through the documents in the dataset. A vector is created for each word, comprising the weightings of its co-occurrence with all possible word pairs (this includes directional information – in other words for each word pair two values exist dependent on the order the words occur i.e. AB and BA). Scaling procedures using the final vectors allow the terms to be mapped into a high multi-dimensional space. Terms with the highest weightings occur closest together in this space. This mapping enables proximity to be used to assess similarity between words, in order to make judgements about the meaning of words.

In addition to identifying the similarity in meaning between two words, richer

contextual meaning can be extracted by examining the average semantic distance between them. Average semantic distance is the average distance between a word and its 10 closest neighbours, referred to as a semantic neighbourhood. When terms share neighbourhoods they can be judged to share meaning even if they themselves are not mapped close together.

Experiments have shown that HAL can successfully use this model to categorise words into animals, body parts, and graphical locations, and mirror human judgements of similarity based on lexical decision times for accurate word pairings (Lund & Burgess, 1996). HAL can also identify semantic and contextual relationships (e.g. mapping 'beatles' in the same neighbourhood as 'band', 'song', 'album' etc. and 'frightened' in the same neighbourhood as 'scared', 'upset', 'embarrassed', 'worried etc.) (Burgess1998) and can predict word recognition equally as well as word frequency for frequent words and better than word frequency for medium to low frequency words (Burgess & Livesay, 1998).

More recent experiments (Siakaluk, Buchanan, & Westbury, 2003) have shown that semantic distance determined using HAL can predict performance on word recognition tasks. When both experimental and non-experimental words require a response (i.e. the yes/no categorisation task where a response was required for both categorical (yes) and non-categorical (no) words) the effect of semantic distance was much weaker than when participants had to respond only to the experimental target (in this case non-animals) although was still consistent with findings from other studies e.g. (Shneiderman et. al., 1997)

It has been shown that since the advent of the cognitive and information processing movements within psychology, cognitive or psychological space has been central to many theories of semantic cognition and memory. 'Distance' in various ways has been used as a construct to help explain meaning and although these theories don't intentionally use a 'spatial' or 'absolute' distance model of processing they don't contradict such a model.

The aforementioned models of the structural nature of semantics together with Shepard's (1957; and 1986) model of perceptual processing combine within Gardenfors' (2000) theory of conceptual space. Offered as an encompassing explanation of semantic-cognition, that draws together existing theories of perceptual, conceptual and semantic processing from both symbolic and connectionist approaches, Gardenfors proposes that a psychological or conceptual space exists that comprises many 'quality dimensions' e.g. height, width, depth, weight, brightness, pitch, and temperature. Concepts are defined by their distance along these dimensions, determined by the quality of their individual features (based on physical or abstract properties). Meaning for an active concept is derived from geometric calculations based on relationships between internal representations and experiences from long-term memory (both semantic and episodic) and external information from the environment. These calculations provide the co-ordinates for concepts within the conceptual space where distance or proximity defines their (dis)similarity (i.e. the smaller the distance between two concepts the more similar they are considered to be). It can be argued that concepts are re-mapped each time they are activated and this re-mapping gives meaning its dynamic and contextual nature dependent on what other concepts are active within the space at the time of processing.

Given that we are equipped to exist in a three-dimensional physical world, and given the degree to which traditional theories rely on the distance metaphor to explain underlying cognitive processes and the way in which everyday language (i.e. the choice of nouns verbs and verb phrases) is used to express similarity and dissimilarity, spatial theories of meaning intuitively make sense. As Gardenfors (2000) points out, spatial theories for the cognition of meaning rather than conflicting with existing theories complement them. Support for this comes from empirical evidence provided by researchers such as Lund & Burgess (1996), Brooks (1998), Burgess, Livesay, & Lund (1998), and Siakaluk, Buchanan, & Westbury (2003).

As information retrieval becomes a vital part of modern life, research needs to be extended to encompass much broader definitions of conceptual processing, (i.e. simultaneous processing of multiple concepts). More needs to be learned about how judgements of similarity, relevance, and information requirements are influenced by differences in individuals' cognition of multiple concepts (document content comparisons for instance rather than just sentence comparisons). If, as evidence suggests, theories of spatially processed meaning and similarity are correct, they lead to interesting implications in terms of information representation and retrieval needs as well as in the search for understanding of how individuals acquire cognitive maps of information spaces.

1.4 Information Retrieval, Visualisation and Spatial Data Management Systems

Computerised information retrieval (IR) expands to all fields of data search and

extraction involving electronically stored information. The study of IR is generally aimed at designing and evaluating the effectiveness and usability of computerised systems for organising and facilitating retrieval of electronic representations of information. Such research is largely undertaken by information specialists within the fields of computer science and human-computer interaction. As is witnessed by the unprecedented growth and success of the Internet in particular the World Wide Web (WWW) almost all types of information can be and are represented electronically. The common denominator of the various types of computerised information tends to be text and as such most IR research focuses on textual or bibliographic database storage and retrieval systems.

Quantitative evaluation of IR systems began in earnest with the Cranfield experiments conducted in the 1960's by Cyril Cleverdon, (Cleverdon, 1967; Cleverdon, 1972) (see Chapter 3); a comprehensive review is given by Sparck Jones (1981). This early IR research was restricted to the interaction of information specialists with indexes and bibliographic databases, e.g. librarians, indexers, search analysts etc. For instance Tague (1981), when giving guidelines for operationalising 'variables related to people' in experimental IR, identifies 'users' as clients requiring information and 'reference librarians' or 'search analysts' as the intermediaries who interact with the IR system. Tague suggests using experience, training, and degree of interaction with a particular system as useful operational measures.

Due to the advances in technology and the proliferation of computerised interfaces in work places, in public, at school, and in homes, neither the opportunity nor the need to employ specialist intermediaries still exists in the vast majority of situations. 'Users'

now also interact with the system despite often being lay persons or IR novices. As such the scope of IR research has necessarily expanded greatly, and become much more diverse. Issues such as database organisation and presentation, query operations, user need, retrieval protocols and retrieved item presentation, usability, and system evaluation are all aspects examined within the current research field (Baeza-Yates & Ribeiro-Neto 1999).

Initial IR research was concerned with evaluating methods of retrieving documents from a specific bibliographic collection that at the time used batch-processing not real-time interactive retrieval (e.g. the Cranfield experiments, (Cleverdon, 1967; and 1972)). Within the current IR paradigm, however, the term information is not restricted to specific document or text retrieval based on Boolean style query matching, but to the satisfactory solution of an information need by the user. While this information need can be satisfied by the retrieval of documents deemed relevant to a specific query, it is more often the case that the user themselves are not aware of what will satisfy their requirements. They are in a position to make this judgement only when they have identified relevant sources and followed links and pathways to associated documents, hence the need for browseable interfaces (e.g. Chang & Rice, 1993; Borlund, 2000;).

The term 'browsing' itself has caused a large degree of debate within the field in terms of how it should be defined. However, Toms (2000) suggested the term tends to be used dichotomously in referring either to a) the nature of the task, where browsing refers to fulfilling a poorly defined, an undefined, or a broad information need or b) the methods and behaviours employed to carry out the task where browsing refers to

an unstructured or semi-structured search through a database aimed at determining the nature of the contents and discovering information that may fulfil the need or lead to alternative information that will.

One of the fundamental problems associated with searching or browsing for computer stored information is that it effectively disappears into a hyper-existence and remains there until users select a suitable retrieval term to enter into a search engine programme, or they 'stumble' across the information while browsing the database. Given the size of many databases the latter is extremely unlikely, and existing methods for information retrieval based on the former have proven problematic. This is largely due to a reliance on Boolean logic in query formulation, combined with the diversity in the experience and ability of users. This problem is highlighted in a study by Jansen, Spink, & Saracevic (2000) that examined 18,113 users searching the World Wide Web (WWW) and found that they generated an average of only 2.8 queries per search, that queries were short (av. 2.21 terms), and that they rarely used Boolean operators or relevance feedback (i.e., links to 'more like these').

Toms (2000) compared users behaviour when interacting with a database in which search and browsing was facilitated by either 'Menus', a hierarchical representation of the database structure, or 'items to browse' tools, in either a goal-driven search or a 'no goal' browsing task. It was demonstrated that when the task was non-defined, in other words no retrieval goal was given, people tended to look at more documents and used more menu items than people with a specific information requirement. It was also shown that those with no goal followed more system generated 'suggestions' within the 'items to browse' tools and located more interesting articles, whereas those

with a specific goal generally used the traditional search tool (i.e. query term search) within the 'items to browse' tools. There were no observed differences between menu types. The nature of the task impacts upon behaviour as well as performance.

Retrieval problems are not confined to the size of the databases involved or user expertise in query formulation but also to issues of synonymy (multiple words sharing the same meaning) and polysemy (a single word with multiple meanings). When retrieval is dependent on term matching algorithms (as with most search engines), documents are generally retrieved and ranked for relevance on the basis of the frequency of term occurrence. The result is often very long lists of documents which include the target term. Many of the retrieved items however, relate to an entirely different context or meaning; this leads to very poor retrieval precision (the ratio of relevant documents to total documents retrieved – see chapter 3). In addition many relevant documents may not be included in the retrieved list as they contain alternative terms to the specific search terms used; the result is poor recall (the ratio of relevant retrieved documents to possible relevant documents – see chapter 3).

In order to address these problems researchers are constantly exploring alternative methods of information location and retrieval, and this is often reflected in the way systems are designed to present database contents. Advances in technology and computerised processing capabilities both in software and hardware, have led to the introduction of new tools and techniques. Most recently these tools and techniques utilise graphics and virtual environments (VEs) to represent either whole databases or at least retrieved documents. Resulting visualisations enable the user to gain a much broader overview of what is available in a Gestaltian manner.

1.4.1 Information Visualisation

Information visualisation (IV) is the result of this evolution within IR research and design, and systems have developed that present information spaces using a combination of tools, techniques, and visualisations enabling easier, more intuitive search and retrieval – see Hearst (1999).

Even the most novice computer users are familiar with visual aides such as icons, colour coding, hierarchical menus and window displays. In addition to these fundamental visualisations tools there is a variety of more complex instruments that the day to day user may or may not be aware of. These include:

- brushing and linking – the same information is presented in different formats e.g. graphical and text, and linked so that selecting a target in one view automatically selects the corresponding target in the other view(s) and changes made to one of the representations are automatically reflected in the other,
- panning and zooming – the screen presentation can be manipulated in a similar fashion to videoing scenes. The user can zoom in or out, enlarging specific points of interest at the exclusion of other content or gain a broader overview of the entire information space, or they can pan sideways or up and down to bring peripheral contents into the field of view (FOV),
- focus+context – used in conjunction with zooming and panning retains all information on screen but enlarges the focal points of interest while shrinking the remainder.

A comprehensive overview of these and other tools and techniques is given in Hearst (1999).

Information systems that utilise these tools have developed visualisation methods to suit the presentation of particular types of information, which Shneiderman (1998) organises into a 'data type by task taxonomy (TTT)' – see Figure 1-1. The data types and definitions that Shneiderman identifies and uses are: -

- Temporal, which refers to data that is represented using time-lines e.g. the history of the monarchy of England.
- 1D linear, which includes text based information that is sequential in nature and would apply to traditional document databases where contents are presented in list format.
- 2D map data, including geographical data presented in a X, Y coordinates.
- 3D World data, allows representation of three-dimensional real-world objects using X, Y, Z coordinates which is useful for a wide range of situations for instance architectural design and scientific research e.g. brain imaging.
- Multi-dimensional data refers to multi-faceted data for which each element can be represented as a separate dimension in n-dimensional space e.g. statistical data such as census information where many facts about individuals living at a single property are collected. Presentations of this type of data often use two or three-dimensional scattergrams.
- Tree data, which refers to hierarchically organised information, presented using a format where information spreads from the point of origin (trunk) out to peripheral points (branches). For instance management structures within corporations can be easily presented this way.
- Network data, is similar in nature to tree data however the relationships aren't clearly structured in a hierarchical format and cluster in a more abstract / non

linear and complex fashion .

OLIVE – an On-line Library of Information Visualisation Environments developed by Shneiderman and his 1997 students provides an excellent taxonomy of current systems using these data types and techniques (Shneiderman et. al., 1997), while chapter 10 in Shneiderman (1998) provides a much more comprehensive explanation and definition of this TTT of visualisation techniques.

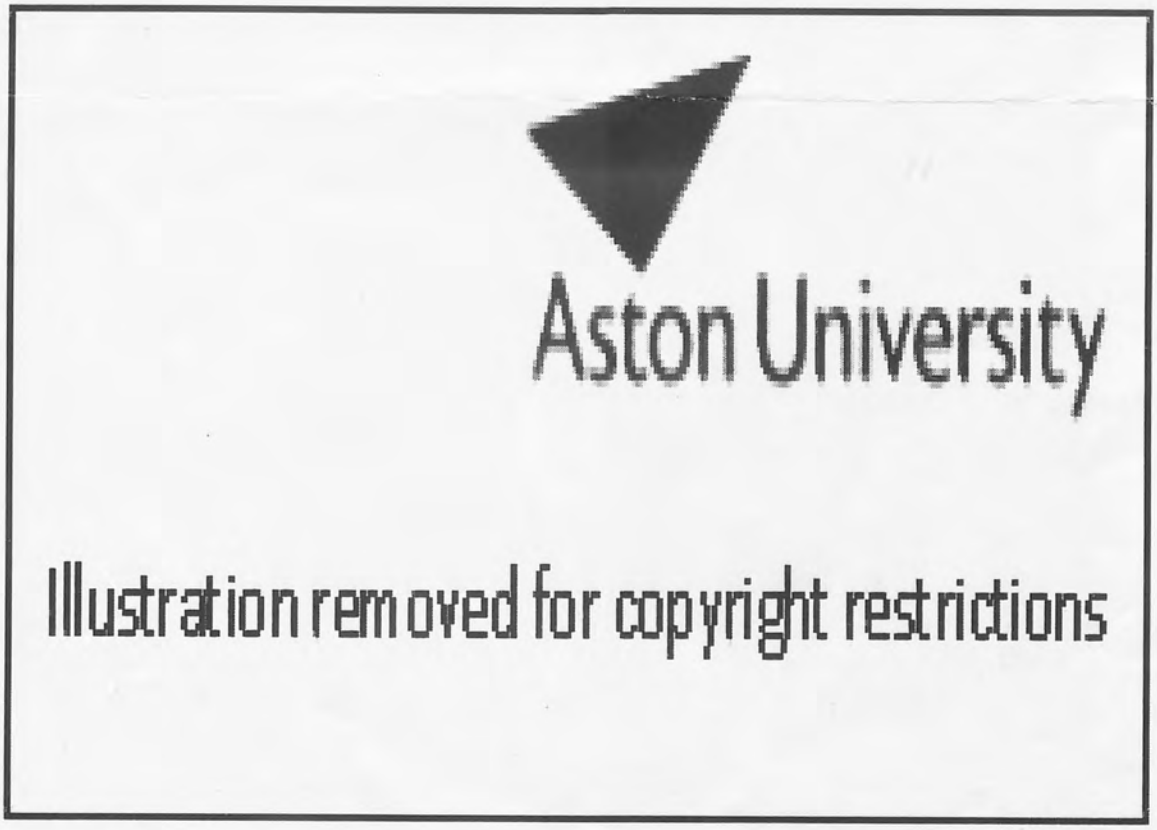
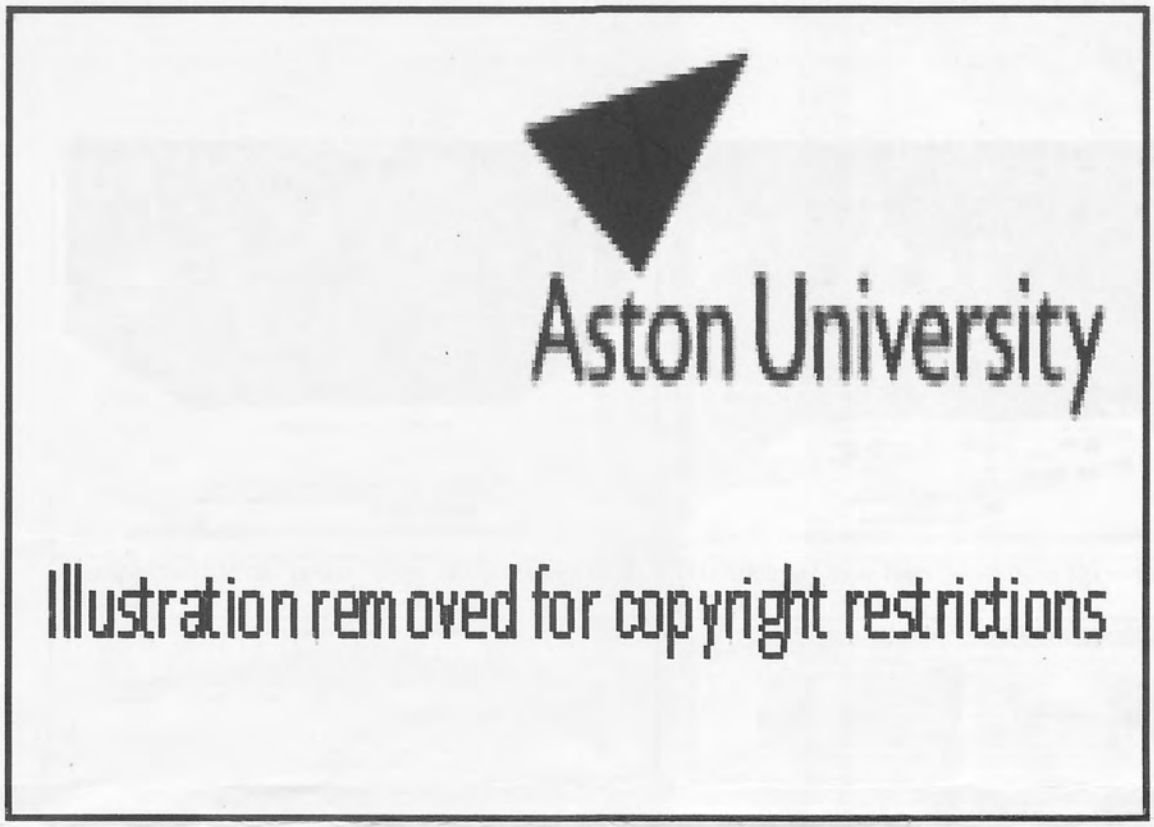
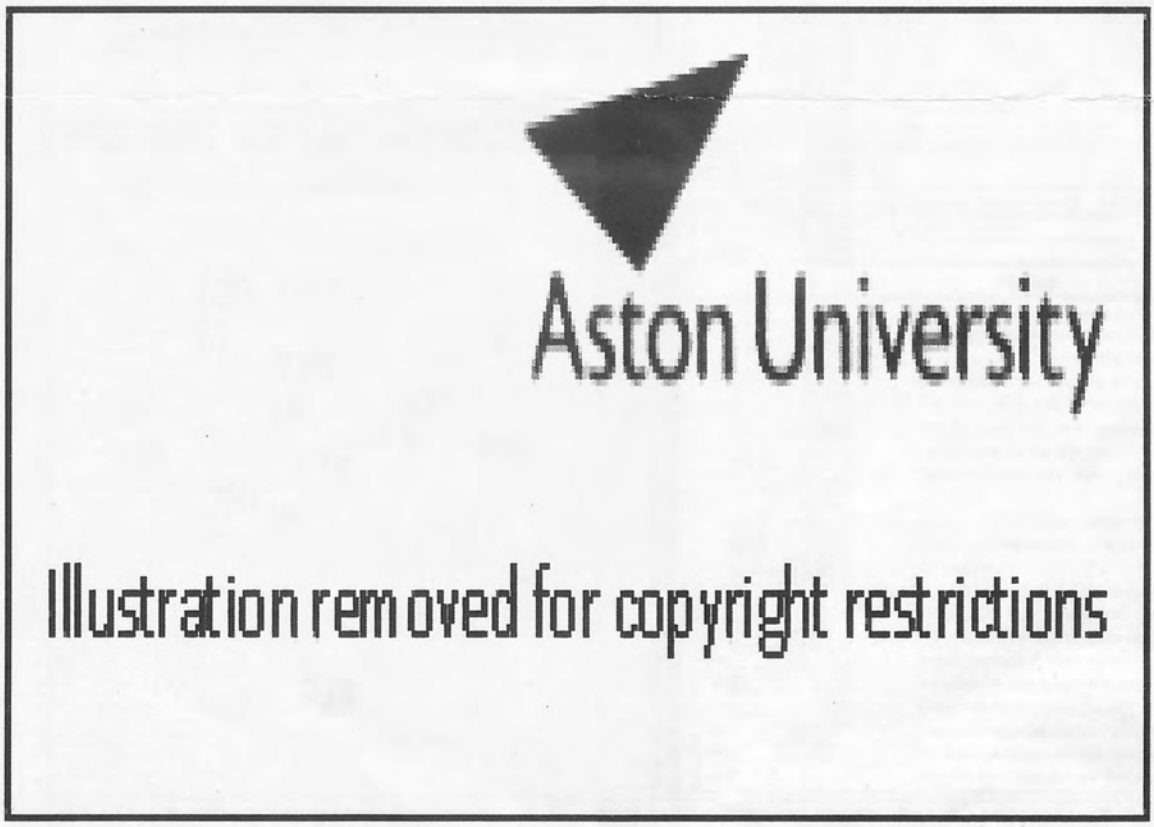


Figure 1-1 Data type by task taxonomy (TTT) taken from Shneiderman (1998), Chapter 14 page 524

Much of the early work into IV systems (for which Shneiderman developed his TTT)

took place at Xerox PARC – Palo Alto Research Institute in the early 1990's (for example Perspective Wall (Robertson, Mackinlay, & Card, 1991a); Cone Trees (Robertson, Mackinlay, & Card, 1991b); Scatter/gather (Cutting, Karger, & Tukey, 1992); Table Lens (Rao & Card, 1994), BEAD, (Chalmers & Chitson, 1995); and TileBars, (Hearst, 1995)) see Figure 1-2 for examples. These systems employed a variety of visual cues such as colour, size, object shape, hierarchical organisation and links, and spatial clustering etc. together with the techniques referred to previously.


Aston University
Illustration removed for copyright restrictions


Aston University
Illustration removed for copyright restrictions

For instance Table Lens which uses a 'network' visualisation method, displays tabular information which reflects 'multi-dimensional data', and uses focus+context to enable the user to manipulate the presented information (Rao & Card, 1994). Perspective Wall alternatively displays 'temporal data' on a wall that can be manipulated to show time moving from left to right. The 2D wall is presented using three-dimensional space to maximise the display area, this is done by 'folding' the wall. The information on the wall in the centre screen (X,Y axis) is accessible so that the user 'unfolds' the wall by moving left to right or right to left. The pieces of information are represented as bricks in the wall where the size, colour, and position (horizontal and vertical) of the bricks provide additional pertinent information (Robertson, Mackinlay, & Card, 1991a).

More recent developments in information visualisation has resulted in systems such as those described above (i.e. visual information retrieval interfaces – VIRIs), focus on ways to present document collections (i.e.1D linear data) using object based 2 and 3D display interfaces and animation to enable users to locate required documents more easily. Documents are presented as icons or objects and inter-document relationships can be relayed using the visual cues already discussed e.g. colour, size, links, clusters etc. Often the principle cue is spatially defined whereby the environmental organisation of documents/objects via clustering, proximity, or linked pathways reflect document similarity and additional cues, e.g., colour, shape and size of objects etc. provide secondary information (e.g. Chen et al., 1999; Leuski & Allan, 2000).

Examples of prototype systems using these types of combinations include the NIST Information Retrieval Visualization Engine (NIRVE) prototype (Cugini, Laskowski,

& Piatko, 1997; Sebrechts et al., 1999), and SENTINEL (Fox et al., 1999).

1.4.2 Spatial Data Management Systems

The term spatial data management system (SDMS) originated from the work of Herot (1980). He devised a technique for organising and retrieving general information that involved converting it to a visual format that could then be presented to the user in spatial manner. The term is now extended to spatial visualisation systems of document databases that use the spatial relationships themselves as an additional source of information regarding the content of the documents or relevance of the documents to a particular query or topic. For example StarWalker (Chen, 1998; Chen, 1997; and Chen et al., 2002) is an SDMS that also uses a virtual environment (VE) for presentation of a database's contents. It displays documents as nodes in a spatially organised network – see Figure 1-3.

Figure 1-3 On-screen display of the Starwalker virtual environment taken from Chen, & Cribbin (2002) page 3.

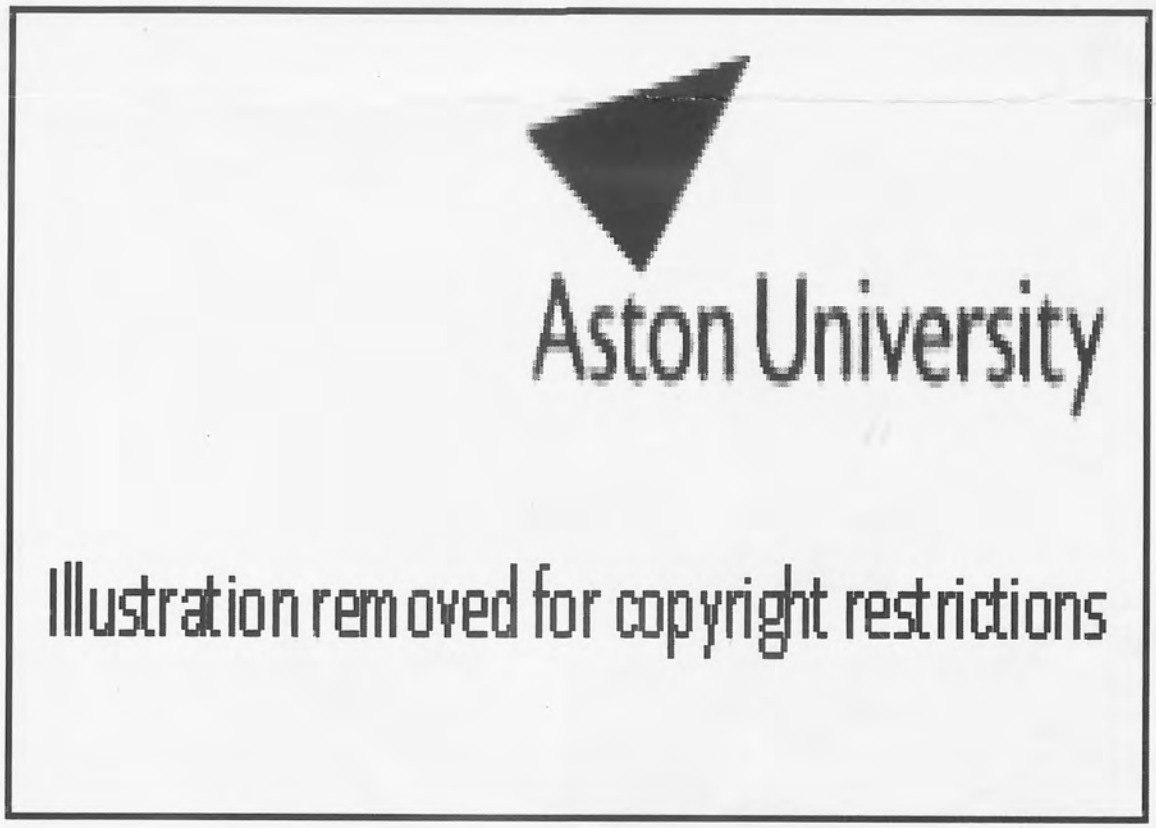
The nodes or documents cluster into semantically related topics and virtual links connect nodes that share common features. These links represent the underlying semantic relationships between the documents. The mapping of this network of documents is achieved by firstly gauging the similarity between documents for all combinations of pairs of documents, and then converting these similarity measures to coordinates in a virtual space. This is done using mathematical algorithms, in this case 'Pathfinder network scaling' was used (Schvaneveldt, Durso, Goldsmith, Breen, Cooke, Tucker, & Demaio, 1985). The similarity measure (Chen, 1998) uses is called generalised similarity analysis (GSA) and uses several indicators of similarity i.e. hypertext linkage patterns (patterns of use of hyperlinks within documents that lead to other documents for example when browsing the World Wide Web (WWW)), content

similarity, and document usage patterns.

There are an increasing number of prototype information visualisation systems incorporating SDMSs that use various tools and techniques for document organisation and presentation, but fundamentally rely on a model of spatial organisation in which proximity reflects semantic relationship. For instance Lighthouse (Allan et al., 2001); and Leuski & Allan, 2000) presents documents retrieved from a query both as a list, based on their ranked retrieval position, and as objects (spheres) in a three-dimensional space. The objects are proximally organised in the 3D space on the basis of their inter-document similarity as judged using a multi-dimensional vector space model (VSM) of term co-occurrence (see chapter two for an explanation of VSMs). A brushing and linking technique is used to connect the individual objects with their associated documents and vice versa - see Figure 1-4.

The purpose of the Lighthouse system is to structure document organisation so that documents of a similar relevance value to an 'original' document will be spatially proximal to that document, facilitating enhanced retrieval performance. The 'original' document is determined by the user as the one they judge most relevant to their query. Selecting the spheres closest to the relevant document identifies the appropriate documents from the ranked retrieval list – see Figure 1-4. Results of user studies showed that retrieval performance in terms of recall and precision was increased when the visualisation was included compared to task performance on the same system but without the visualisation window present. The authors also found that users preferred 2D visualisation over 3D due to the 'cognitive overload' involved in judging distance between spheres in 3D space. This was supported by the fact that the system algorithm

out-performed users using 3D presentation (Leuski & Allan, 2000).



(bottom) pictures are shown.

Figure 1-4 The Lighthouse visual information retrieval system taken from Leuski & Allen, 2000), page 5

Due to the diversity in the design of SDMSs and the variety of additional tools and techniques employed in individual systems it is very difficult to isolate the areas that are key factors in optimum information retrieval efficiency, particularly from a psychological viewpoint. Allan, Leuski, Swan, & Byrd (2001) for instance found that numerous factors such as individual differences in users, and differences in topics

used within the retrieval tasks, had to be subtracted in order to identify system effects. In addition they authors originally thought that indirect comparison of performance on the two experimental systems of interest (one with visualisation, one without), to a third control system would demonstrate the differences between visualisation and non-visualisation. They found however, that while the participants employing the system using visualisation significantly out-performed individuals using the control system, and that those who used the control system significantly out-performed participants using a non-visual, a direct comparison of performance on the two experimental systems showed no significant differences.

The VE SDMS used by Westerman & Cribbin (2000b) was designed to provide inter-document similarity information through spatial cues with no other navigational or search/retrieval tools and techniques available to the user. This allowed the authors to examine exclusively the advantages and disadvantages of providing information regarding the semantic content of documents through extra spatial dimensions (i.e. 2D vs. 3D). In addition, the effects of the quality of spatial mapping in terms of the amount of semantic information accounted for when placing the objects (these were nodes that represented buildings, parts of buildings or furniture) in the environment could be tested. The effects of individual differences in users' cognitive maps of the information spaces on the trade-off between navigational demands and the provision of additional semantic detail were also examined, in order to a) to identify the relationship between cognitive ability and the spatial mapping of cognitive information spaces, and b) to control for these effects in performance differences between environment organisation.

The authors found that performance improved with the quality of the spatial-semantic mapping of the environment, indicating that participants did use spatial cues in judging semantic content. They also found that performance in a poorer quality mapping using two-dimensions produced comparable performance to higher quality mapping using three-dimensions. It was suggested that the additional cognitive demands of navigation in 3D reduced the benefits of the higher quality mapping. This replicated Leuski's & Allan's, (2000) findings although they suggested this difference was due to having to reproduce a 3D cognitive model of the space from a flat screen image (Sebrechts, Cugini, Vasilakis, Miller, & Laskowski, 1999). Westerman & Cribbin (2000b) also found significant effects of individual differences in cognitive ability, both in terms of spatial visualisation and associative memory on retrieval performance. These results further support theories of spatially processed meaning. They suggest that individuals with more finely tuned spatial abilities can form better cognitive maps of the information space. This is likely due to a better cohesion between their internal cognitive space and the experimental environment.

The SDMS employed in the current project was modelled on that of Westerman & Cribbin (2000b). Spatial-semantic mapping was used to convey information regarding the degree of semantic relatedness between documents via proximity. Unlike the StarWalker (Chen, Thomas, Cole, & Chennawasin, 1999) and Lighthouse (Leuski & Allan, 2000) systems, similarity measures derive from vector space modelling of term co-occurrence i.e. semantic content of documents only, in order to focus on the relationship between cognitive semantic mapping and computerised semantic mapping. The original Westerman and Cribbin SDMS presented a database of object nouns rather than documents and therefore used human judgements to assess semantic

similarity. Presentation of the database took the form of a VE in which the documents were represented as objects much the same as in the aforementioned systems, however in this instance the objects weren't linked (Starwalker), any visual clustering occurred simply due to the varying degrees of document similarity in the reduction of high dimensional space to 2 or 3D space, and user relevance judgements did not impact on document organisation (Lighthouse).

The position of the objects was spatially determined with proximity reflecting document similarity and no other clues were provided regarding document organisation – see chapter 3 for a detailed account of how the environment was created.

Spatial-semantic mapping within the VE was used in order that the environment resemble users' conceptual spaces allowing more accurate cognitive models of the database to be generated (based on the aforementioned theories of spatial-semantic cognitive processing). The VE presentation allowed users to 'navigate' between documents in order to make use of the spatial mapping within the environment to aid location and retrieval of target documents.

1.5 A Need for the Current Research

As previously highlighted, the use of electronically presented information has increased and continues to increase rapidly among individuals and organisations from all corners of society. Expert and novice users alike, from the fields of computer science, information technology, academia, industry, and commerce are increasingly finding computerised databases an essential source of information. Children are

introduced to computers during the early years of schooling and will continue to use them throughout their lives for learning, work and recreation. In addition to this the introduction of the 'family PC (personal computer)', the World-Wide Web (WWW) and interactive television, has resulted in a shared reliance on electronic information storage mediums from a variety of general users, and a fierce increase in competition from manufacturers and developers of technology. Information technology (IT) is employed within public and private sectors throughout Government, industry, and commerce making its use an ever increasing necessity for even the most computer-illiterate or techno-phobic. We are becoming an 'informational society' (Castells, 2000).

Information storage, retrieval, and presentation are all key factors within the field of IT, emphasising the need for effective development and management of systems designed for this purpose. Such necessity has led to a considerable amount of research in the development of information retrieval (IR) systems, some of which has been discussed. Paramount to development is an understanding of the diversity of the user population in terms of need, knowledge and experience.

Currently, much of the research reviewed comes from the fields of computer science and information technology with a more limited input from Psychology. One enterprise that exemplifies this is the Text Retrieval Conference (TREC) project (e.g Harmen, 1993), which is a testimony to the extent of work being conducted privately, publicly and academically on a global scale. The first TREC in 1992 involved 92 individuals presenting 25 participating systems, in 2001 however, TREC 10 had increased to 87 groups from academia, commerce, and government representing 21

different countries (Harmen, 1993; and Voorhees & Harman, 2001). An overview of TREC is included in chapter two. Most evaluation tasks (tracks) in TREC pit system against system or test a system's performance in retrieving documents judged relevant to a topic query by a single individual (the topic's author). In 1997 however, an interactive track was introduced to examine user performance, as recognition of the need to consider the user. Nonetheless user evaluation remains less of a focus than system design as demonstrated by the other ten tracks.

Within computer science the emphasis of research still tends towards the technical aspects of system design and providing systems that are task efficient and user friendly. To an extent users are seen as a single entity, with differences in expertise accounted for but much less consideration given to psychological/cognitive differences. For example Sebrechts et al., (1999) evaluated presentation effects of a prototype information visualisation system NIRVE using only five participants in each experimental condition (when comparing the behaviour of novices to experts this figure reduced to two experts and three novices). No other measures of individual differences were examined. A further example of low sample size can be seen in the user studies reported in the evaluation of Lighthouse (Allan, et. al., 2001) described in the previous section. The sample size consisted of eight users per system evaluated while this figure is low it is acceptable for factorial analysis, however when comparing experts to general users this figure drops to four per system. Even studies that claim specifically to examine individual differences e.g. "spatial ability and visual navigation" (Chen, 1997), and "individual differences in a spatial-semantic environment" (Chen, 2000), sample sizes remain small – eleven participants in the first paper which, and ten and twelve respectively for the two studies in the second

paper all of which used correlation and differential designs. In addition for the second study in Chen (2000) a multivariate analysis was conducted giving a degrees of freedom ratio of 6:4. Clearly with such low participant numbers there is cause to be wary regarding the reliability or reproducibility of such studies. Clearly evaluative judgements regarding the usability or benefits associated with certain types of presentation made on the basis of such a small sample size should not be considered representative of the general user population.

The need for an increased focus on individual differences in HCI research generally was initially recognised in the mid to late 1980s and research examining the effects of cognitive differences in task performance was published (e.g. Paepcke et al., 2000; Egan, 1988; and Seagull & Walker, 199). In particular, issues of individual differences between users when browsing and searching for information within computerised databases has attracted much attention (e.g. Borgman, 1989; Kwasnik, 1992; Chang & McDaniel, 1995; Harter, 1996; Borgman, 1999; and Chen & Macredie, 2000). However, while there has certainly been increased attention to differential research in HCI since this time, the field of information visualisation, particularly utilising virtual environment interfaces, is at an early stage and as such differential research in this area is minimal.

Valuable insights are being gained through examination of individual differences in IV and VE research (e.g. Westerman, 1995; Westerman & Cribbin. 1999; Westerman & Cribbin, 2000a; and Collins & Westerman, 2001). And despite the low sample sizes previously referred to, Chen (e.g. 1997; 1998; and 2000), has demonstrated that differences in cognitive ability (in particular spatial ability and memory) impact on

users' performance as have the studies of (Leuski & Allan, 2000) and (Allan et al., 2001).

While the focus of this differential work in HCI tends to remain with examining issues of system design (and rightly so from the viewpoint of IT development), it is the intention of the work in this thesis to determine what can be learned about the relationship between cognitive ability, behaviour, and performance in terms of electronic information seeking from a psychological and theoretical perspective. Studying individuals' interaction with the system will enable this. It is also expected however, that conclusions drawn from this research can be beneficial to system designers and possibly offer an alternative perspective from which to view user issues. It is hoped this work will build on recent studies directed at individual differences in the psychology of the user (e.g. Westerman, 1995; Westerman, 1998; Borgman, 1999; Westerman & Cribbin, 1999; Westerman & Cribbin, 2000b; Collins & Westerman, 2001; and Westerman, Collins, & Cribbin, 2005)

The benefits to HCI of applying knowledge gained from contemporary psychology in differential research were identified by Dillon & Watson (1996). They highlight the fact that since the early work of Hull (1928) cited in Dillon & Watson (1996), continuing through various contemporary studies (e.g. Schmidt & Hunter, 1983; Bobko & Karren, 1982; Ledvinka & Simonet, 1983; Nielsen, 1993 cited in Dillon & Watson, 1996), differences in worker productivity between the best and the worst are in the region of 2:1. Dillon & Watson report that a series of studies by Schmidt & Hunter (1983) demonstrated that the "standard deviation of workers productivity is between 40 and 70% of salary" (Dillon & Watson, 1996, pg. 629).

A further issue that will be addressed in the current work is the criticism of system evaluation for the restricted measures that are employed. In many system evaluation studies, recall (i.e. the ratio of relevant documents retrieved to total relevant documents available) and precision (i.e. the ratio of relevant documents retrieved to total documents retrieved) are considered the 'gold standard' measures of performance (Harter, 1996). Research however, clearly shows that judgements of relevance vary widely between individuals and are influenced by an extensive variety of factors both internal in terms of users' cognitive differences, experience, expertise, subject knowledge etc. and external terms of information need, system design (e.g. Schamber, Eisenberg, & Nilan, 1990; and Schamber, 1994). Measures based on recall and precision, which are determined by relevance judgements of the documents in question, are highly susceptible to influences from these factors, and as such generalisation of results are predicted to be limited.

While measures of recall and precision are retained within the current research other measures such as time on task, distance travelled, browsing behaviour, are also examined – see chapters three, four, and five.

1.6 Structure of Thesis

The thesis is structured to reflect the way in which all the elements discussed previously (i.e. spatially structured semantic cognition, SDMSs, and information visualisation, assessing document similarity, system evaluation, and browsing behaviour), interact to mediate effective information retrieval, and assess the implication of spatial-semantic processing within IR and cognition. In order to address

the topics set out above in a manner that allows the issues to be examined both exclusively and jointly, chapters two and three use data collected from the same experiment to address separate research questions.

Chapter two, the first of four experimental chapters, considers the mechanisms for identifying similarities between documents based on their semantic content. In particular two forms of automatic text analysis are compared using human ratings as a baseline measure. The primary purpose of these experiments was to ensure the subsequent adopted method for measuring document similarity (n-gram based VSM) was sufficiently robust to provide the necessary data on which organisation of the database contents within the IR system could be based.

Chapter three then examines various performance measures on a browse task within the experimental environment, to identify how the quality of mapping and number of dimensions employed affect the success with which people use a VE SDMS. The effects of individual differences in cognitive abilities (i.e. spatial visualisation, associative memory, and spatial working memory) that have already been implicated in IR performance are examined within performance results. In addition, traditional measures of performance (i.e. recall and precision) are combined to provide a measure of accuracy, various measures of speed of task completion are examined and a speed accuracy trade-off used to determine overall effectiveness of performance.

Chapter four, as previously mentioned, uses data from the experiment reported in chapter three. Here however the experimental questions are focused on behaviour. Novel measures of browsing patterns using n-gram analysis are used together with

measures of distance travelled, direction of travel, lostness (Smith, 1996) etc. to determine the effects of spatial mapping on users browsing behaviour. Once again individual differences in cognitive ability are analysed.

Chapter five, the final experimental chapter, examines new data in order to bring together the findings from chapters three and four to identify the relationship between performance and behaviour, and consider the implications of spatial-semantic cognition, and spatial-semantic mapping in information visualisation. The size of the participant sample is greatly increased by focusing only on performance and behaviour in the optimal mapping solution within the 3D environment. In addition examination of individual differences is extended to include differences in preferred wayfinding strategies (Lawton, 1994; and Lawton, 1996) and verbal ability, measures that have not been widely examined in visual information retrieval research.

Finally chapter six will present an overview of the research together with conclusions drawn from the findings.

2 N-Gram Based Automatic Text Analysis

2.1 Introduction

Automatic text analysis (ATA) refers to a non-manual, computerised process for filtering text. The outcome can be used to locate, organise, or retrieve documents on the basis of their content. The design of effective information retrieval (IR) systems, required for dealing with continually expanding masses of electronic information, relies on effective ATA. The majority of this information is text-based and an IR system can only be as effective as the programme it uses for interpreting the content of the available information. Currently one of the fundamental problems associated with electronic databases is that information effectively disappears into a hyper-existence and remains there until a suitable retrieval term is entered into a search programme, or a user stumbles across the information while browsing a database. Given the size of many databases the latter is extremely unlikely, and existing methods for information retrieval have proven problematic, largely due to a reliance on Boolean logic in query formulation (e.g. Jansen, Spink, & Saracevic, 2000; and Park, 2000).

Traditional IR methods tend to use Boolean logic whereby keywords and operators (e.g. 'AND, OR, NOT') are combined to produce a query. A retrieved list of documents containing the requested terms is then presented to the user. Often the documents in these lists are ranked on the basis of some weighting, e.g. the normalised inverse document frequency (IDF) score as used in 'InQuery' a prototype IR system developed by (Allan et al., 2001). The IDF is a complex calculation employing the following values: i) how often the term appears in the document; ii)

how many documents contain the term; iii) how many terms appear in the document; and iv) the average number of terms in a document. From these, term weightings are produced that are inversely related to the 'commonality' of the terms. A term that appears frequently and relatively equally within all documents in a corpus has little discriminatory value in determining which documents are semantically similar. Using the IDF reduces the effect of such non-discriminatory terms. Unfortunately document ranking often fails to reflect document relevance and document similarity due to problems of polysemy (a single word or phrase has multiple meanings) and synonymy (multiple words share the same meaning). For instance, if two documents identical in semantic content comprise different word usage and/or misspellings, ranking will fail to identify document synonymy. In addition to these factors, many users have little expertise or skill in using Boolean operators and tend to rely on the first few retrieved documents from only a couple of queries. This was demonstrated in a study of 18,113 users searching the World Wide Web (WWW) in which it was found that people generated an average of only 2.8 queries per search, that queries were short (MN: 2.21 terms), and that Boolean operators or relevance feedback (i.e., links to 'more like these') were rarely used (Jansen, et al., 2000).

As IR system design becomes more complex, to fulfil current needs for accessing relevant information quickly and easily, and to address the problems of traditional Boolean based search engines, innovative ATA methods are being developed. This is evident from, for example, the increasing numbers of participating groups in TREC the Text REtrieval Conference. TREC is an annual event held in North America co-sponsored by the National Institute of Standards (NIST), and the Defense Advanced Research Projects Agency (DARPA). The first conference was held in November

1992 as part of the TIPSTER text programme an initiative by DARPA to encourage research into effective text processing and information retrieval systems. TIPSTER had three main areas of interest: i) document detection (locating relevant documents); ii) information extraction (locating specific relevant information within a single larger document); and iii) summarization (reducing the size of a document or corpus while retaining the fundamental information themes). Under this programme government, industry, and academia came together to research these areas. TREC was just one aspect of the programme and was designed as a platform for presenting, comparing, and evaluating research from the various participating bodies. The TIPSTER programme terminated in 1998 however TREC continues to grow in terms of interest and participation. Twenty-five groups (12 companies and 13 universities) participated in TREC 1, by TREC 10 in November 2001 this number had increased to 87 groups representing 21 countries. In fact TREC 10 demonstrated a 25% increase in participation over TREC 9 (Harmen, 1993; and Voorhees & Harman, 2001).

A major progression apparent in this research is the replacement of Boolean based search engines by systems that retrieve documents on the basis of their semantic content, not just the occurrence of specified keywords within a query. Information visualisation systems are an example of this (see Chen & Yu (2000) and Chapter One section 1.3.1) for a review. As a result of this, ATA methods have become more complex leading to greater precision and accuracy, e.g. Latent Semantic Analysis (LSA) (Deerwester et al., 1990).

2.1.1 Vector Space Models of Information Retrieval

Analysis techniques used for retrieving and organising documents based on content

similarity often use vector space mathematical algorithms and are known as vector space models (VSMs). Documents within a collection are represented as equal length vectors comprising a list of unique terms contained in the corpus. Terms can be individual words, word strings, or letter strings, and as with ranked list retrievals (see above) terms are often weighted using various methods for judging their relative importance. Each document vector comprises the actual or weighted value of the terms based on the number of occurrences for that document. The similarity between the vectors of any two documents can then be computed based on their relative positions within the vector space. The dot product for instance provides the Euclidean distance between two documents or between a document and a query within the vector space (Letsche & Berry, 1997), while the cosine value provides an index of similarity between documents based on the angle between their term vectors in high dimensional space. Documents and queries can be mapped into a high dimensional space where relatedness is reflected by the proximity between documents within this multi-dimensional information space. As will be seen in Chapters Three, Four and Five of this thesis, proximity can be used as a visual measure of document similarity when documents are mapped as objects in a virtual environment representation of a database.

Cosine is generally the favoured measure of similarity, having been shown to more accurately convey document relatedness both in terms of producing expected grouping patterns of a data set within a vector space (Rorvig, 1999), and in matching human ratings of document similarity (Cribbin & Westerman, 1999). A further benefit of using the angle rather than Euclidean distance measure between vectors is that vector lengths are not an important factor Letsche & Berry (1997).

Latent Semantic Indexing (LSI) is currently one of the most popular VSM-based ATA methods and has proven successful in various IR studies e.g. (Dumais, 1995; Chen, 1998; Soboroff et al., 1997;Letsche & Berry, 1997; and Newby, 2001). LSI identifies inherent semantic relationships between documents, and component terms. To identify 'latent' concepts in a corpus and reduce noise caused by non-discriminate terms, a factor analytic, singular value decomposition technique (SVD) is used to decompose the document-term matrix into smaller matrices containing linearly independent factors. Factors are ranked by their factor weightings with small values being set to zero which are then ignored. By re-combining the matrices (which now contain fewer factors) similarity between documents, terms, or terms and documents can be determined by calculating the angle or distance between their vectors. The benefits of LSI as an approach to document retrieval exist in the way it identifies the latent semantic content of documents rather than simply identifying documents that share the same key words or terms. This allows the semantic structure of database contents to be presented to users using visualisation techniques, and alleviating the problems related to query production and use (e.g. the need for expertise in the use of Boolean operators, and excessively large lists of retrieved documents that may not be relevant at all or in which highly relevant documents and ranked toward the bottom of the list). For a full explanation of the mathematical processes involved see Deerwester, et al. (1990)and Letsche & Berry (1997).

LSI generally uses words as terms, although letter strings known as n-grams have also been used (Soboroff, Nicholas, Kukla, & Ebert, 1997). Prior to compiling the term document matrix pre-analysis of the documents is often conducted to weight

component terms and remove high frequency words. SVD identifies weighted ranked factors; these weightings are often directly attributable to the document corpus the LSI programme was initially programmed with ('trained' on), which can influence the inter-document similarity judgements (Dumais,1995). One of the problems with LSI is that it is computationally very expensive and the results can vary dependent on how the LSI programme was 'trained'.

Analysis using n-grams involves identifying and comparing the co-occurrence of unique sequences of 'n' character strings between documents within a set, by passing a window, 'n' characters long, across the member documents one character at a time. A term-document matrix is composed similarly to LSI, however factor analysis of the matrix is not conducted and the inter-document similarity matrix (comprising cosines between the document vectors) is calculated directly from the raw term-document matrix. The benefit of n-grams over words or terms is that misspellings, capitalisation of letters, and suffix and prefix variations etc. are minimised; two words that vary only slightly are therefore still likely to retain a number of shared n-grams. This is however dependent on word length and adopted n-gram length. Generally n-gram based systems have used short strings such as bi-grams and tri-grams (Cavnar, 1993), quad-grams (Cavnar, 1995), and pentagrams (Damashek, 1995a).

For each document within the set, a vector comprising the frequency to which each unique n-gram occurred is produced (see Table 2-1). The cosine of the angle between each of these vectors as viewed from the absolute origin of the information space, is then calculated which results in a similarity matrix of all the component documents.

	Doc 1	Doc 2	Doc 3	...	Doc *n*
TI: T	1	0	0	...	0
I: TH	1	0	0	...	0
: THE	1	0	0	...	0
THE	11	5	12	...	10
THE N	1	0	0	...	0
HE NE	1	0	0	...	0
E NEU	1	0	0	...	0
NEUR	1	0	0	...	0
NEURA	1	0	0	...	0
EURAL	1	0	0	...	0
URAL	1	0	0	...	0
RAL B	1	0	0	...	0
AL BA	1	0	0	...	0
L BAS	1	0	0	...	0
....				...	
Term k...	...	...	...	...	...

Table 2-1 Term by document matrix showing the number of times terms occur within each document in the Working Memory set

Cavnar (1995) used n-grams as the basis for filtering and retrieving documents from TREC 2 and TREC 3 test collections. However he did not use the full list of n-grams present in the document collections. Inverse document frequency (IDF) measures were calculated for each n-gram (see above) and a cut-off IDF value employed which, had the same effect as word-stemming or removing stopwords in a term based system. Cavnar found that the system performed reasonably well and, based on average precision scores, achieved almost the median score compared to all other participating systems. By using the IDF measure of n-gram weighting Cavnar substantially reduced the average length of document vector which subsequently reduced the amount of processing space required. However it could also be argued that in doing so valuable information regarding n-gram co-occurrence between documents was lost.

A possible alternative which reduces the number of unique n-grams as a bi-product is the use of longer strings. Damashek (1995a), tested an IR system called 'Acquaintance' at TREC 3 in 1994, in which he used 5-grams. The n-gram weights that compiled the vectors were simply the number of occurrences of the specific n-

gram as a proportion of total n-gram occurrences within the document. The cosine value between document and / or query vectors was calculated on the basis of these n-gram co-occurrences.

Acquaintance was tested on two main retrieval tasks at TREC, the 'ad-hoc' task in which short descriptions of relevant documents are used, and the 'routing' task whereby document retrieval is based on an example document. Generally performance was poor when compared to other participating systems with Acquaintance ranking 22nd out of 23 systems for the ad-hoc task, and 19th out of 21 systems for the routing task (Harmen et al., 1995). While it was shown that Acquaintance performed poorly compared to other IR systems at TREC, it should be noted that, in terms of topic visualisation, Acquaintance can successfully cluster documents on the basis of topic (Damashek, 1995b). This suggests that n-gram analysis using un-weighted n-gram co-occurrence may be a useful tool for database visualisation when documents are presented as objects within a virtual environment presentation.

2.1.2 Current Experiment

The experiments reported in this chapter were conducted to assess the suitability of an n-gram ATA as reported by Damashek (1995a), for obtaining document similarity measures on which to structure the virtual environment (VE) database that is used in the remainder of this thesis. In doing so the intention was to forgo weighting methods, which are to a large degree subjective in terms of where to impose cut-off points, and test the effects of n-gram length as a predictor of document similarity. One problem with TREC evaluations is that, while systems are effectively compared with other systems, judgements of document relevance against which systems are evaluated are

made by a single individual. In this set of experiments n-gram based ATA was evaluated on the correspondence between assessments of semantic similarity and multiple human judgements of document similarity. As previously stated, LSI is a proven method of ATA (Deerwester et al., 1990). Therefore comparisons of n-gram based ATA with LSI were carried out.

Four experiments were conducted to test the effectiveness of n-gram based ATA in judging inter-document semantic similarity. The first two experiments involved comparing n-gram based analyses and LSI with human judgements of document similarity. Experiment One used two 'technical' document sets each comprising journal abstracts relating to a specific psychology-based topic. Experiment Two used 'non-technical' document sets each consisting of newspaper articles relating to a current affairs topic. The experiments were designed to compare: i) average levels of agreement between human raters (inter-rater reliability - IRR); and ii) human ratings with n-gram (using various length grams) and LSI measures of inter-document similarity. It was hypothesised that n-gram analysis would reflect human judgements of document similarity when judgements were semantically driven, and there would be no difference between different the n-gram and LSI protocols. The technical documents sets and data for relevant human ratings collected by Cribbin & Westerman (1999) were used with additional ATA analyses being conducted for Experiment One. The non-technical document sets were created specifically for Experiment Two and human ratings data were collected. Full details of how the document sets were created are given in the section 'document set preparation' within the methodology of each experiment. For Experiment One the details of document set collection, and participant data collection reflect the processes undertaken by Cribbin

& Westerman (1999), however the analyses were novel.

In addition to the differences in the nature of the documents used for Experiments One and Two, methods for initial document retrieval when compiling the sets differed. The technical documents were retrieved from the BIDS social sciences database using Boolean style term-match queries (the terms used were 'Working AND Memory' and 'Schizophrenia'). The non-technical documents were taken from the TREC test collection database of documents deemed relevant to specific topics (in this instance 'Journalists Risks' (JR) and 'Piracy' (P)). It was this difference in document selection methodology together with differences in the results from Experiments One and Two in terms of optimal n-gram length for comparing human ratings of document similarity that prompted Experiment Three (see Results for each experiment and Discussion for Experiment Two).

The third experiment repeated the design of the first but involved the removal of keywords (i.e. the terms used in the initial retrieval query) from the technical document sets (keywords could not be removed from the non-technical documents since documents were not initially selected on this basis). The revised documents were re-analysed using n-gram based ATA and correlated with the original human similarity ratings. Keywords were removed in order to identify their role in determining the length of n-gram required to match human ratings. Given that keywords were relatively long, it was expected that the optimal length of n-gram required to match human ratings would be smaller when keywords were removed and would be similar across document sets. As LSI based ATA produces only one solution per document set and is not dependent on an optimal choice, as with n-gram length, it

was not deemed necessary to include LSI analysis.

The final experiment expanded on work of Damashek (1995a), examining the capacity for ATA to identify material as belonging to a particular document. Following Damashek (1995a), document sets comprising either pairs of documents deemed twins or pairs of documents deemed non-twins were created from the four original document sets. The ability of n-gram based analysis to distinguish between documents deriving from the same parent document or the same document set (twin pairs) was evaluated. It was expected that pairs of twin documents would attain higher measures of similarity than non-twin pairings. An in depth description of how 'twin' documents were created is given in 'document set preparation' for Experiment Four.

2.2 Experiment One

2.2.1 Methodology

2.2.1.1 Document Set Preparation

Two document sets comprising eight documents each were compiled using abstracts from published psychology journal papers. The abstracts were retrieved from the Bath Information and Data Services (BIDS) social-sciences online database, using Boolean based keyword match queries. The keyword term 'Working AND Memory' was used for set one (WM), and the keyword 'Schizophrenia' was used for set two (S). 'Chronologically most recent' was given as the retrieval priority command. The final eight documents in each set were selected from the top 32 retrieved documents and matched for word length, document length and readability. Readabilty was determined

using the Flesch reading ease score, as calculated by MS Word, which provides a rating up to 100 based on average sentence length and average syllables per word. Text with a high score is more easily understood than text with a lower score, an average document is expected to have a score of approximately 60 or 70 (Talburt, 1986). Table 2.2 shows document statistics together with t-test results showing that WM and S document sets did not differ significantly. It can be seen that read ease scores are well below average, and reflect scores generally attributed to scientific literature i.e. scores within the range 0 – 30 (Lemos, 1985).

Doc. Set	No. of Raters	Age		Gender		Word Length		Doc. Length (words)		Read Ease	
		Mean	SD	M	F	Mean	SD	Mean	SD	Mean	SD
WM	Expert 6	30.2	8.6	1	5	5.7	0.26	189.4	36.3	21.1	10
	Non-expert 12	20.6	2.8	6	6						
S	Expert 6	31.8	14	2	4	5.8	0.41	199.1	61.6	14.7	9.2
	Non-expert 12	19.4	0.7	4	8	t = -1.82; ns		t = -0.39; ns		t = 1.33; ns	

Table 2-2 Descriptive statistics for Working Memory and Schizophrenia document sets, and participants

2.2.1.2 Participants and Procedure

An independent sample was used to obtain human-ratings of between document similarities for the two document sets. Thirty-six psychology staff and students (13 males, 23 females) were semi-randomly (ensuring equal numbers of staff and students in each condition) allocated to judging either WM or S. Of the 18 participants in each document set condition, six were considered expert in the subject areas of working memory and schizophrenia respectively (staff) and 12 were considered non-experts (students). See Table 2.2 for participant statistics.

Using a purpose-written computer programme participants were presented with all

paired combinations of documents in a random order, (28 pairings in total). Viewed one at a time participants could swap between documents within a pair until satisfied they had understood the content, at which stage they indicated their judgement of semantic similarity between the pair, using a visual analogue scale; 0 = completely dissimilar, 100 = identical. Participants were not told the nature of the document sets or asked to judge on particular criteria.

Computer based ATA employing the n-gram technique was used to obtain measures of document similarity for all pair-wise document comparisons. N-gram lengths of three to 25 characters were used, resulting in 23 similarity matrices (using cosine values) that were subsequently converted to vectors in order to conduct correlation analyses with the vectors of human similarity ratings. A vector of cosine values produced by LSI was also obtained.

2.2.2 Results

The data showed, that for the WM document set, cosines produced by n-gram lengths three to seven, and average human ratings were normally distributed, but lengths eight to 25 were not. For the S document set only, cosines produced by n-gram lengths three and four were normally distributed. As this raised issues of whether parametric or non-parametric analyses were appropriate, both Pearson's r, and Spearman's rho were used in the first instance. Correlation values between ATA and average human ratings, (expert, non-expert, and all raters), were calculated together with average correlation values between ATA and individual raters (expert, non-expert and all individuals). As can be seen from Figures 2.1 & 2.2 both measures produced similar curves across n-gram lengths for WM. However, for S document set the curve patterns produced were

dissimilar. In both instances parametric measures demonstrate gradual changes between cosine values as n-gram lengths increase, non-parametric measures however produce less uniform incremental changes. It is assumed that Pearson's coefficient r better describes the data as Spearman rho is adversely susceptible to the large number of tied ranks due to many zero value cosines particularly as n-gram length increases, therefore the results reported are of analyses using Pearson's coefficient r.

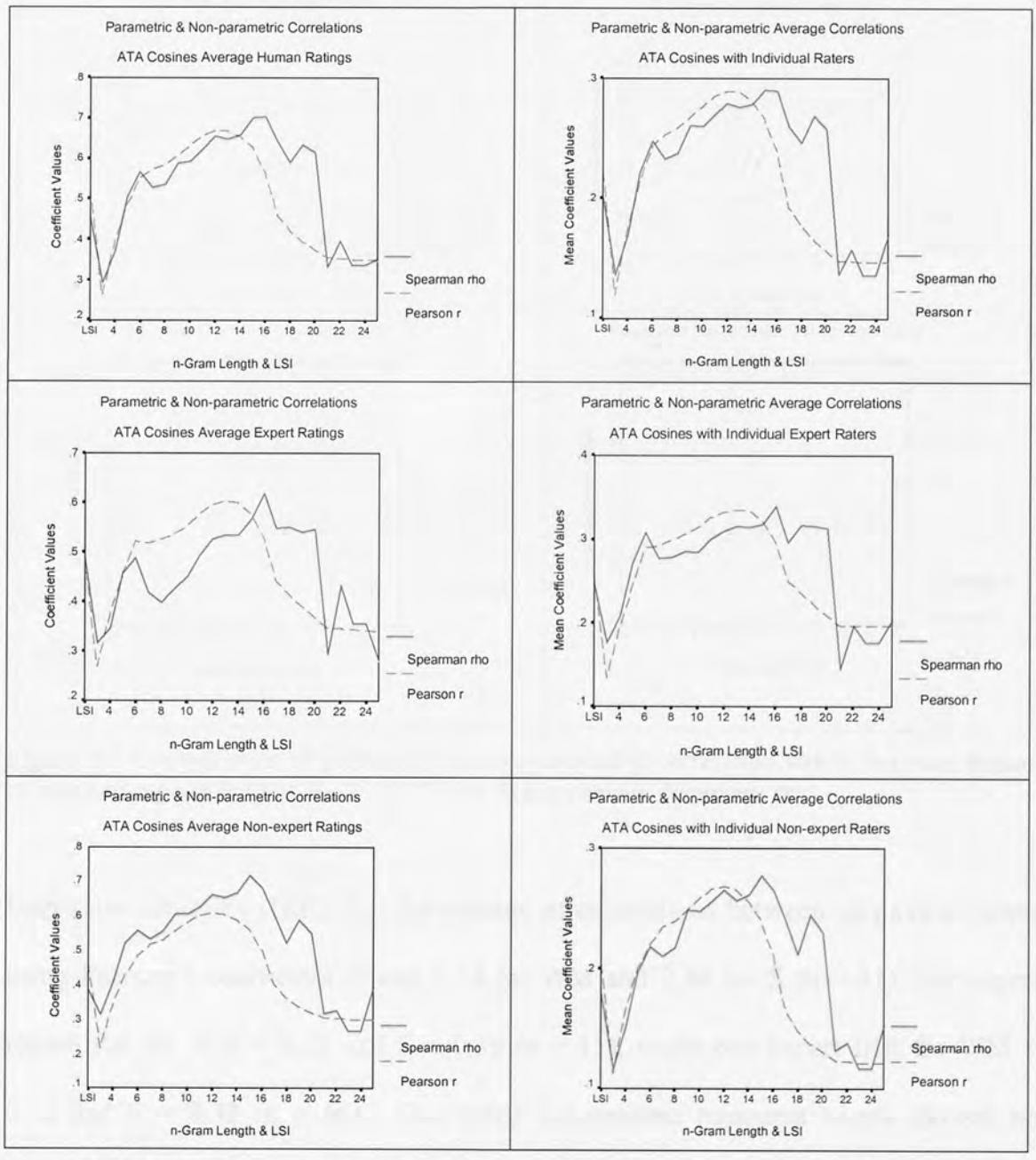


Figure 2-1 Comparisons of parametric & non-parametric correlation values between human ratings and n-gram lengths 3 – 25 & LSI for Working Memory document set

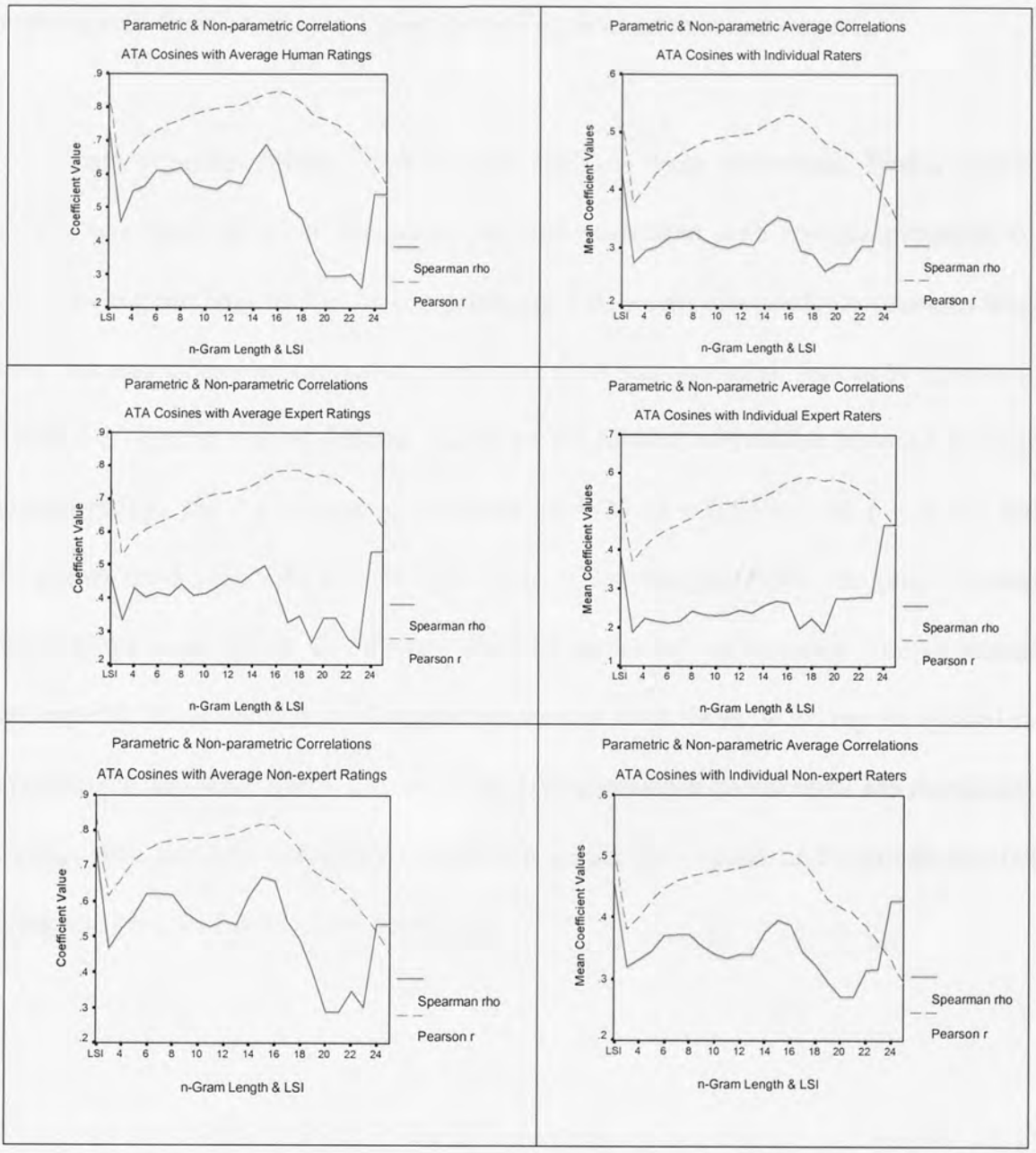


Figure 2-2 Comparisons of parametric & non-parametric correlation values between human ratings and n-gram lengths 3 – 25 & LSI for Schizophrenia document set

Inter-rater reliability (IRR) (i.e. the average of correlations between all pairs of raters using Pearson's coefficient r) was 0.14 for WM and 0.34 for S (n = 81). For expert raters IRR for WM = 0.23 and S = 0.39 (n = 15), while non-expert IRR for WM = 0.12 and S = 0.32 (n = 66). One-tailed independent measures t-tests showed no significant difference for IRR between experts and non-experts for S (t (79) = 1.43;

ns), however significant differences were observed for WM ($t(79) = 1.89$; $p < 0.05$), with experts demonstrating a higher level of agreement than non-experts.

To compare human ratings to ATA two analyses were performed. Firstly human ratings averaged for each document pair and correlated with cosines produced by ATA for n-gram lengths 3 to 25 characters and LSI were calculated. Correlations were then calculated between individual raters and ATA and averaged. As can be seen from Table 2-3 optimal n-gram lengths, based on the highest correlation between average human ratings and ATA occur at 12 grams for WM ($r = 0.67$ $n = 28$; $p < 0.01$), and 15 grams for S ($r = 0.82$ $n = 28$; $p < 0.01$). These lengths differ marginally across experts, and non-experts. Coefficients for LSI correlated with average human ratings for experts, non-experts and all raters are weaker than those occurring at optimal n-gram length for WM. For S document set LSI correlation coefficients are marginally weaker than those produced by optimum n-grams for experts and non-experts, but marginally higher for all raters combined.

	WM non-expert	WM expert	WM all	S non-expert	S expert	S all
3g	.215	.268	.260	.619**	.528**	.615**
4g	.345	.374	.394*	.671**	.583**	.662**
5g	.428*	.455*	.485**	.719**	.606**	.717**
6g	.483**	.522**	.551**	.745**	.626**	.744**
7g	.516**	.518**	.572**	.766**	.642**	.766**
8g	.528**	.526**	.583**	.776**	.659**	.772**
9g	.557**	.535**	.607**	.789**	.686**	.777**
10g	.581**	.552**	.631**	.797**	.704**	.779**
11g	.602**	.577**	.656**	.801**	.714**	.780**
12g	.609**	.596**	.668**	.806**	.719**	.786**
13g	.603**	.603**	.667**	.812**	.725**	.791**
14g	.587**	.599**	.654**	.831**	.741**	.809**
15g	.559**	.577**	.626**	.844**	.761**	.819**
16g	.500**	.528**	.565**	.848**	.781**	.815**
17g	.401*	.441*	.460*	.832**	.787**	.789**
18g	.361	.413*	.421*	.801**	.784**	.746**
19g	.335	.389*	.393*	.755**	.768**	.689**
20g	.319	.368	.373	.738**	.772**	.664**
21g	.304	.346	.354	.720**	.759**	.643**
22g	.302	.345	.352	.687**	.734**	.610**
23g	.300	.343	.349	.645**	.704**	.565**
24g	.299	.342	.349	.598**	.667**	.517**
25g	.299	.340	.348	.538**	.615**	.458*
LSI	.464*	.493**	.525**	.843**	.740**	.854**

* p < 0.05; ** p < 0.01 Highest correlations are highlighted

Table 2-3 Correlations between cosines for n-gram lengths 3 – 25 and LSI and average human ratings of experts, non-experts and all raters for Working Memory and Schizophrenia doc sets.

In the second analysis inter-document cosines from ATA were correlated with individual participants' ratings of document pairs and average correlations calculated for n-gram lengths 3 to 25 characters and LSI between experts, non-experts and all raters combined.

The optimal length n-gram occurred at 12 grams for WM (mean r = 0.289, SD 0.179) and 16 gram for S (mean r = 0.53, SD = 0.116) – see Table 2.4. The optimum n-gram length differed marginally across experts and non-experts for both WM and S. In addition the difference between averaged correlations for experts and non-experts at lengths 12g for WM and at 16g for S were not significant (t (16) = -0.64 and t (16) = -1.26 respectively). For both WM and S there were no significant differences between average correlation coefficients for optimal length n-grams (12g (WM), and 16g (S)) and LSI (t (17) = 1.35 and t (17) = 0.17 respectively). In addition there was no

significant difference between WM and LSI for the worst performing n-gram length 3g t (17) = -1.59. The difference between averaged correlations of individual raters with ATA when comparing LSI with the worst performing n-gram length (25 gram) for S document set was significant t (17) = -4.99; $p < 0.001$.

	WM non-expert		WM expert		WM all		S non-expert		S expert		S all	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
3g	0.112	0.218	0.134	0.212	0.119	0.210	0.380	0.120	0.370	0.125	0.377	0.118
4g	0.167	0.211	0.193	0.215	0.175	0.206	0.401	0.140	0.408	0.101	0.403	0.125
5g	0.200	0.212	0.244	0.200	0.214	0.203	0.428	0.160	0.436	0.076	0.431	0.135
6g	0.220	0.218	0.290	0.194	0.243	0.208	0.447	0.145	0.454	0.085	0.449	0.125
7g	0.234	0.206	0.291	0.182	0.253	0.195	0.461	0.141	0.468	0.086	0.463	0.123
8g	0.239	0.202	0.298	0.168	0.258	0.188	0.467	0.135	0.484	0.093	0.473	0.120
9g	0.249	0.199	0.305	0.157	0.267	0.183	0.472	0.127	0.506	0.108	0.484	0.119
10g	0.258	0.200	0.313	0.152	0.276	0.183	0.476	0.124	0.517	0.114	0.489	0.119
11g	0.266	0.200	0.325	0.150	0.285	0.183	0.477	0.122	0.521	0.120	0.492	0.120
12g	0.268	0.195	0.331	0.149	0.289	0.179	0.482	0.121	0.522	0.119	0.495	0.119
13g	0.265	0.182	0.334	0.144	0.288	0.169	0.486	0.121	0.527	0.122	0.500	0.119
14g	0.255	0.167	0.334	0.019	0.281	0.158	0.498	0.119	0.539	0.118	0.512	0.117
15g	0.239	0.147	0.321	0.144	0.266	0.147	0.505	0.117	0.557	0.113	0.523	0.115
16g	0.210	0.132	0.296	0.166	0.239	0.145	0.506	0.113	0.578	0.116	0.530	0.116
17g	0.164	0.133	0.249	0.184	0.192	0.152	0.491	0.116	0.586	0.125	0.523	0.124
18g	0.148	0.136	0.235	0.184	0.177	0.154	0.464	0.126	0.589	0.135	0.506	0.139
19g	0.137	0.143	0.221	0.183	0.165	0.158	0.430	0.139	0.581	0.141	0.481	0.154
20g	0.130	0.147	0.208	0.187	0.156	0.161	0.418	0.147	0.583	0.129	0.473	0.159
21g	0.123	0.151	0.196	0.188	0.147	0.162	0.407	0.151	0.574	0.117	0.462	0.159
22g	0.122	0.150	0.196	0.188	0.147	0.162	0.387	0.158	0.555	0.105	0.443	0.161
23g	0.121	0.150	0.196	0.188	0.146	0.162	0.361	0.162	0.530	0.088	0.417	0.161
24g	0.121	0.150	0.196	0.187	0.146	0.161	0.332	0.164	0.499	0.079	0.387	0.161
25g	0.121	0.149	0.196	0.185	0.146	0.161	0.296	0.166	0.459	0.077	0.350	0.161
LSI	0.204	0.147	0.258	0.167	0.222	0.151	0.517	0.149	0.545	0.124	0.526	0.138

Highest correlations are highlighted

Table 2-4 Average Pearson r coefficients between cosines for n-gram lengths 3 – 25 & LSI and individual human ratings of experts, non-experts and all raters for Working Memory and Schizophrenia doc sets

2.2.3 Discussion

The results confirm that both n-gram based ATA and LSI provide suitable measures of inter-document similarity, in terms of reflecting human judgements of document similarity. At all n-gram lengths correlations between ATA and average human judgements exceed IRR for both WM and S document sets (see Table 2-3). At optimal length, (i.e. 12 gram for WM and 15 gram for S) the correlations obtained by n-grams are substantially higher than IRR. IRR accounted for 2% shared variance and n-grams

accounted for 44% shared variance for WM, and IRR accounted for 11.5% shared variance and n-grams accounted for 67% shared variance for S. LSI performs on par with optimum n-grams for S, but produces weaker correlations than many n-gram lengths for WM. This is consistent for experts, non-experts and all raters.

All n-gram lengths (except 3-gram for WM) and LSI exceed IRR, when comparing average correlations between ATA and individual human ratings for both document sets (Table 2-4). At optimal n-gram lengths (12-gram for WM and 16-gram for S), n-gram based ATA on average accounts for 8%, and 28% shared variance with individuals respectively. Again when comparing performance between ATA and experts, non-experts, and all raters, LSI performs on par with optimum length n-grams for S but is weaker for 6-gram to 16-gram ATA for WM.

When comparing ATA similarity scores with individual raters the difference between ATA and IRR is greatly reduced, reflecting the high degree of variation between individual judgements of document similarity. Very low levels of IRR demonstrate the poor level of agreement between individuals in terms of document similarity, which can be attributed to differences in individual contextual cues both internal and external during document interpretation (see 2.6 Conclusions).

For both document sets, the optimum n-gram lengths are long by comparison to those used previously e.g. (Chen, Thomas, Cole, & Chennawasin, 1999; Cavnar, 1995; Fox, et al., 1999; and Soboroff et al.,1997). For S the optimum length n-gram (16) is longer than that for WM (12), in addition the correlations with average human ratings and with individual ratings averaged are higher for the S document set (see Tables 2.3 &

2.4). One possible explanation for the difference in required length of n-gram is the readability level of the document sets. Both WM and S rate as 'scientific' on their Flesch read ease scores i.e. a low read ease score. WM however achieved a higher score than S, which may account for a shorter optimum n-gram length (see Table 2-2). This factor is considered further when examining the results of Experiment Two. The readability of the documents can also explain the higher correlation coefficients obtained between raters and ATA for S. Due to the complexity of the documents only a small number of pairings may be obviously similar. If humans and ATA identify these pairings, stronger correlations would be expected. A further consideration is the occurrence of keywords and the length of the key-terms used for document selection; the key-terms used for WM (working memory) are 13 characters long, as is the key-term for S (schizophrenia). This is supported to some degree by the occurrence of optimum n-gram length between 11g and 12g for WM across both correlation analyses, although does not explain the optimum n-gram length of 16 characters for S.

In order to test the possibility that individuals are focusing on the occurrence of keywords to make judgements of similarity, and ATA is influenced by the occurrence of those keywords Experiment Three examines the effect of removing the key-terms from the document sets prior to ATA analysis. IRR suggests that raters are not using keywords as a primary cue when judging similarity, as ratings between individuals would be expected to be more consistent however the presence of a keyword or term may provide additional context free indicators on which to base judgements. If this is the case, when keywords are removed from the text for n-gram analysis not only is it predicted that optimum n-gram lengths will decrease but also that correlations with average human ratings will decrease due to the loss of some shared cues between

humans and n-grams.

2.3 Experiment Two

2.3.1 Methodology

2.3.1.1 Document Set Preparation

The two document sets used for this experiment each comprised eight newspaper articles originally published in the Los Angeles Times in 1989. The documents were selected from the TREC 7 database list of documents catalogued relevant to topic queries 'Journalist Risks' (JR) for set one and 'Piracy' (P) for set two (http://trec.nist.gov). In order to obtain a judgement of relevance, documents retrieved by competing IR systems are submitted to the human assessor, who devised the original query, for a binary relevance decision ('relevant'/'not relevant'). The document list in this instance derived from the top 100 pooled items retrieved in response to the queries for 'Journalist Risks' and 'Piracy', by each of the IR systems tested during TREC 7 (Voorhees & Harman, 1999). Documents for the experiment were randomly selected and matched on the basis of word length, document length, and readability (see Appendix 2.4 for document sets). Table 2.5 shows document set statistics, as can be seen the document sets are matched for readability and reflect read ease scores appropriate to 'fairly difficult' material i.e. scores ranging from 50 – 60 (Lemos, 1985).

Doc. Set	No. of Raters	Age		Gender		Word Length		Doc. Length (words)		Read Ease	
		Mean	SD	M	F	Mean	SD	Mean	SD	Mean	SD
JR	12	20.2	3.9	2	10	5.1	0.3	290.9	112.4	51.5	8.3
P	12	19.8	1.4	2	10	5.1	0.4	297.6	112.8	56.8	10.0
						t = 0.39; ns		t = -0.12; ns		t = -1.16; ns	

Table 2-5 Descriptive statistics for Journalists' Risks and Piracy document sets and participants

2.3.1.2 Participants and Procedure

Independent samples of 24 psychology students (4 males, 20 females) were randomly assigned to judging semantic similarities between the contents of either JR or P document sets. Twelve participants were allocated to each condition – see Table 2.5 for participant statistics.

As with Experiment One participants viewed all 28 possible pairings of documents presented randomly using purpose-written software and indicated their rating of document similarity using an analogue scale 0 (no similarity) to 100 (perfect similarity). Again participants were not told the nature of the document sets or asked to judge on a particular criterion.

Cosine measures of document similarity for all combinations of document pairings were obtained using n-gram lengths three to 25, together with LSI based ATA resulting in 24 similarity matrices that were subsequently converted to vectors and compared with human ratings of judged similarity (see Appendix 2.5 for cosine and rating vectors).

2.3.2 Results

The data were examined and demonstrated normal distributions for n-gram lengths 3

to 6, and LSI only for document set JR. Distributions for document set P were normal for n-gram lengths 3 to 5, and for LSI. Correlation analyses using both parametric (Pearson r) and non-parametric (Spearman rho) analyses were initially conducted. Figures 2.3 and 2.4 show comparisons of the correlation coefficients obtained using both methods between ATA and average human ratings, and between ATA and individual ratings averaged. The effect of n-gram length on the amount of variance accounted for between ATA and human ratings increases and declines steadily when measured using the parametric analysis. It was again decided therefore that Pearson's coefficient r represented the relationship between ATA and human ratings more accurately.

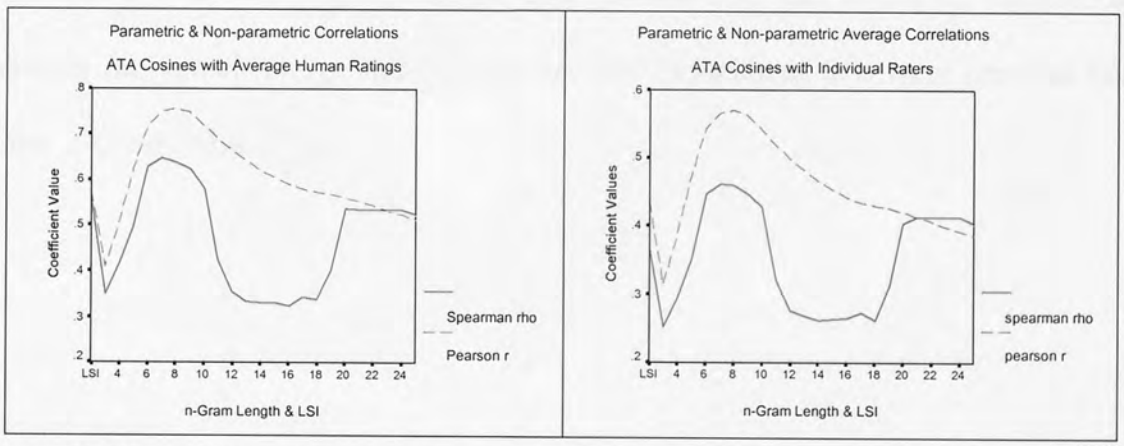


Figure 2-3 Comparisons of parametric and non-parametric correlation values between human ratings and n-gram lengths 3 – 25 for Journalists' Risks document set

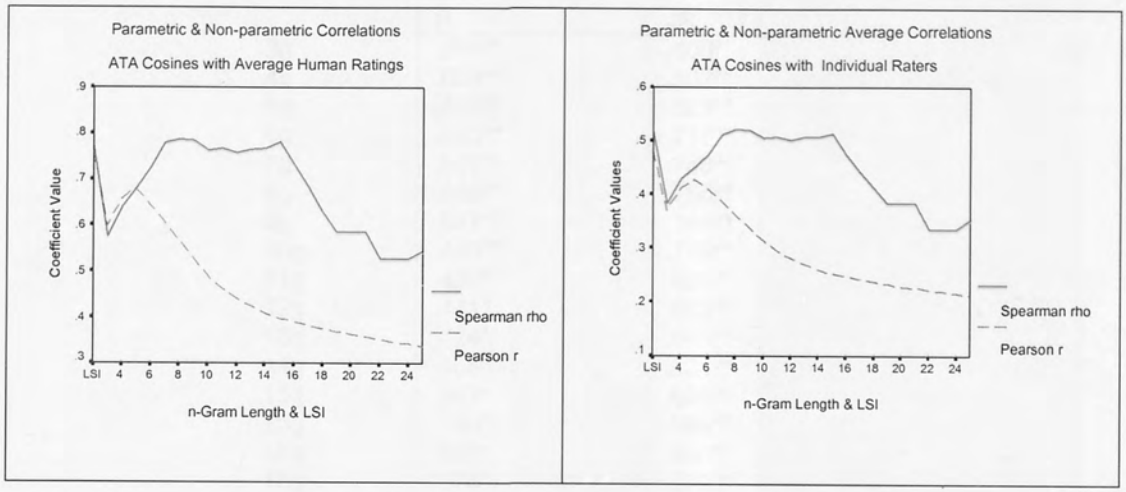


Figure 2-4 Comparisons of parametric and non-parametric correlation values between human ratings and n-gram lengths 3 to 25 for Piracy document set

Inter-rater reliabilities (IRR) were 0.38 (SD = 0.2) for P document set, and 0.52 (SD = 0.13) for JR document set. Analyses correlating average human ratings with cosines produced by ATA for n-gram lengths 3 to 25 and LSI, and averaging correlations between individual ratings and ATA as used in Experiment One were repeated (see Table 2-6 and Table 2-7).

	P	JR
3g	.598**	.409*
4g	.658**	.507**
5g	.680**	.623**
6g	.643**	.717**
7g	.607**	.748**
8g	.566**	.755**
9g	.527**	.746**
10g	.489**	.719**
11g	.462*	.689**
12g	.441*	.663**
13g	.424*	.641**
14g	.408*	.621**
15g	.397*	.604**
16g	.390*	.589**
17g	.382*	.580**
18g	.374*	.573**
19g	.368	.567**
20g	.362	.560**
21g	.357	.551**
22g	.352	.542**
23g	.346	.532**
24g	.342	.523**
25g	.337	.513**
LSI	.759**	.575**

* $p < 0.05$; ** $p < 0.001$

Highest mean correlations for each ATA method within each document set are highlighted

Table 2-6 Correlations between cosines for n-gram lengths 3 to 25 and LSI, and average human ratings of experts, non-experts, and all raters for Piracy (P) and Journalists' Risks (JR) document sets

	P		JR	
	Mean	SD	Mean	SD
3g	0.37	0.229	0.313	0.202
4g	0.411	0.231	0.387	0.179
5g	0.428	0.233	0.473	0.151
6g	0.41	0.224	0.543	0.125
7g	0.388	0.212	0.565	0.109
8g	0.362	0.209	0.57	0.099
9g	0.337	0.209	0.562	0.092
10g	0.313	0.209	0.541	0.088
11g	0.295	0.209	0.518	0.085
12g	0.281	0.21	0.498	0.086
13g	0.27	0.21	0.481	0.086
14g	0.26	0.21	0.466	0.086
15g	0.252	0.21	0.453	0.086
16g	0.247	0.21	0.442	0.086
17g	0.242	0.209	0.434	0.087
18g	0.237	0.208	0.429	0.086
19g	0.233	0.208	0.425	0.086
20g	0.229	0.207	0.42	0.086
21g	0.226	0.207	0.413	0.086
22g	0.222	0.207	0.406	0.086
23g	0.219	0.207	0.399	0.086
24g	0.216	0.206	0.392	0.087
25g	0.212	0.206	0.385	0.087
LSI	0.482	0.223	0.437	0.148

Highest mean correlations for each ATA method within each document set are highlighted

Table 2-7 Average Pearson r coefficients between cosines for n-gram lengths 3 to 25 and LSI and individual human ratings of experts, non-experts and all raters for Piracy (P) and Journalists' Risks (JR) document sets

The optimal n-gram length for comparing with average human ratings was 5-gram for P document set ($r = 0.68$, $n = 28$; $p < 0.01$), and 8 gram for JR document set, ($r = 0.76$, $n = 28$; $p < 0.01$). LSI when correlated with average human ratings accounted for more variation with P document set ($r = 0.76$, $n = 28$; $p < 0.01$) but less variation for JR document set ($r = 0.56$, $n = 28$; $p < 0.01$) than that accounted for by optimum length n-grams (Table 2-6). When correlating n-gram based ATA with individual ratings and then averaging, the highest mean correlation occurred at 5-gram for P (mean $r = 0.43$, $SD = 0.23$), and 8-gram for JR (mean $r = 0.57$, $SD = 0.1$ – both). LSI when compared to the optimal length n-gram again produced marginally stronger coefficients for P (mean $r = 0.48$, $SD = 0.2$) and weaker coefficients for JR (mean $r = 0.44$, $SD = 0.15$). However when comparing the differences for significance using independent measures t-test only differences between JR (8-gram) and LSI were

significant, $t(22) = 2.582$; $p < 0.05$. Differences between the weakest n-gram length for JR (3-gram) and the optimum and weakest n-gram lengths for P (5-gram and 25-gram) when compared with LSI were not significant ($t(22) = -1.705$, $t(22) = -.58$, and $t(22) = -1.206$, respectively). See Table 2-7 for all ATA outputs.

2.3.3 Discussion

For both document sets ATA provided a more robust level of agreement of document similarity than the level of agreement reached between human judges for both comparisons with average human ratings, and averaged comparisons with individual human ratings. This was the case for both n-grams and LSI with the exception of averaged comparisons with individual ratings of JR, where LSI did not meet the same level of agreement as human raters. For n-grams this was also dependent on gram length, especially in the case of P document set where the range of n-gram lengths that outperformed inter-rater judgments was much smaller than for JR – Table 2-6 & Table 2-7.

Optimum n-gram lengths for both P and JR which occurred at 5g and 7/8g respectively were much smaller than for the technical document sets for which optimal n-gram lengths occurred between 12g and 16g (Table 2-4 & Table 2-5) and accounted for an additional 31.5% for P and 31% for JR in document similarity, than IRR.

It was suggested in Experiment One that keywords may have provided additional cues for individual raters resulting in a greater level of agreement in similarity between documents with a large number of keywords present, thus explaining the longer length

of n-gram required to match documents on the basis of keyword occurrence. As keywords were not used in selecting the documents it is unlikely that a substantial number of identifiable key terms, if any, were present, therefore individuals would not have been able to make use of such cues. The fact that in these circumstances optimum n-grams were shorter lends support to the argument made in Experiment One.

An alternative explanation to keywords determining n-gram length was that of the readability level of the document sets. It was suggested that the length of n-gram was inversely related to the readability scores of the documents. This is again supported as readability scores for the non-technical sets were higher than those of the technical sets (see Table 2-5). Despite the scores still classifying the non-technical documents as 'fairly difficult', n-gram lengths were much shorter than for the technical documents suggesting that if a relationship does exist between readability score and n-gram length, it may be influenced by other factors. The higher level of IRR for the non-technical document sets (JR, $r^2 = 0.27$; P, $r^2 = 0.14$), suggests individuals can identify more context free similarities between documents that are easier to read.

It should be noted that correlation coefficients produced by n-grams were generally higher for JR than for P, but that correlations produced by LSI demonstrated the opposite trend. Raters also demonstrated stronger levels of agreement for JR than P. ATA is able to identify a context free element of similarity that people also identify and agree with. It is possible that LSI while successfully measuring document similarity to a degree that can be used by individuals is less context free than n-grams. The performance of an LSI algorithm is to a large extent dependent on the extracted

factors derived from complex transformations involving term weightings resulting from the process of singular value decomposition (SVD). N-gram analysis however examines only the term co-occurrences within the document set being analysed. This could explain why for some document sets LSI can outperform or match optimum n-gram measures but for other document sets performs less well. These findings again support conclusions drawn from Experiment One.

2.4 Experiment Three

2.4.1 Methodology

2.4.1.1 Document Set Preparation

The document sets 'Working Memory' (WM) and 'Schizophrenia' (S) used for Experiment One were adapted for use in this experiment. The keywords used in the original document retrieval searches i.e. 'Working' AND 'Memory' and 'Schizophrenia' were removed from the original documents creating 16 new documents which provided two new document sets 'WM-key' and 'S-key'. The document sets remained matched for average document length, word length, and readability see Table 2-8 for details.

Doc. Set	Word Length		Doc. Length (words)		Read Ease	
	Mean	SD	Mean	SD	Mean	SD
WM-key	5.48	0.3	182.13	34.52	24.56	11.09
S-key	5.7	0.38	196.75	60.36	16.55	11.57
	t = -1.325; ns		t = -0.595; ns		t = 1.414; ns	

Table 2-8 Document set statistics for Working Memory-keyword and Schizophrenia-keyword

2.4.1.2 Participants and Procedure

Removal of keywords from the original documents resulted in new documents that lost coherence in terms of syntax. Therefore the human ratings of document similarity for WM and S document sets obtained in Experiment One were used for this experiment. N-gram analyses of the new document sets WM-key and S-key using three to 25-grams were conducted, providing 23 similarity matrices for comparison with the human similarity judgements of WM and S.

2.4.2 Results

Initial examination of the data showed that only n-gram lengths 3 to 6 for WM-key, and 3 and 4 for S-key were normally distributed, however to maintain continuity with Experiment One Pearson coefficient r analysis was used. As can be seen from Figure 2-5 and Figure 2-6, parametric analysis describes the data more effectively.

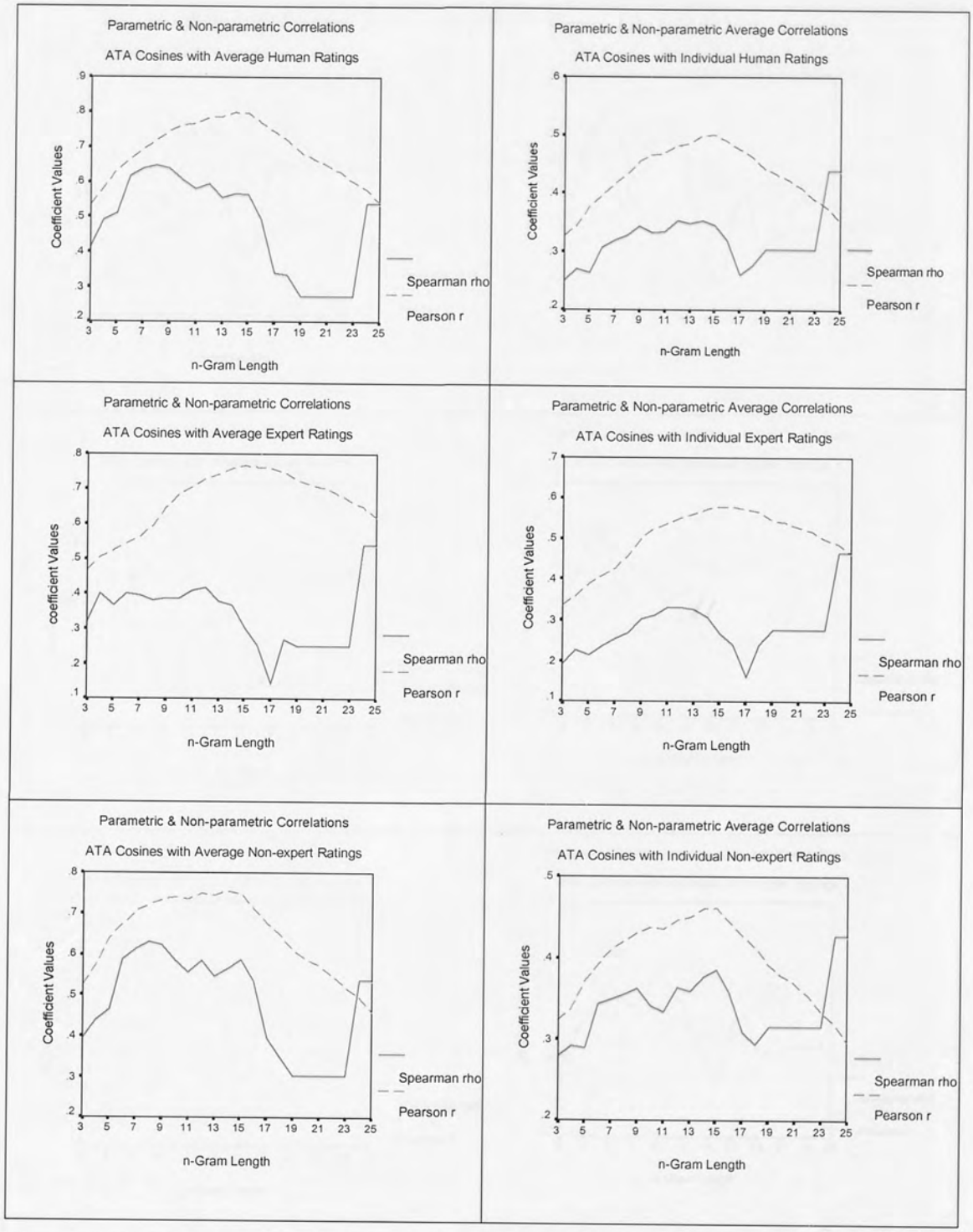


Figure 2-5 Comparisons of parametric and non-parametric correlation values between human ratings and n-gram lengths 3 to 25 for Schizophrenia-keyword document set

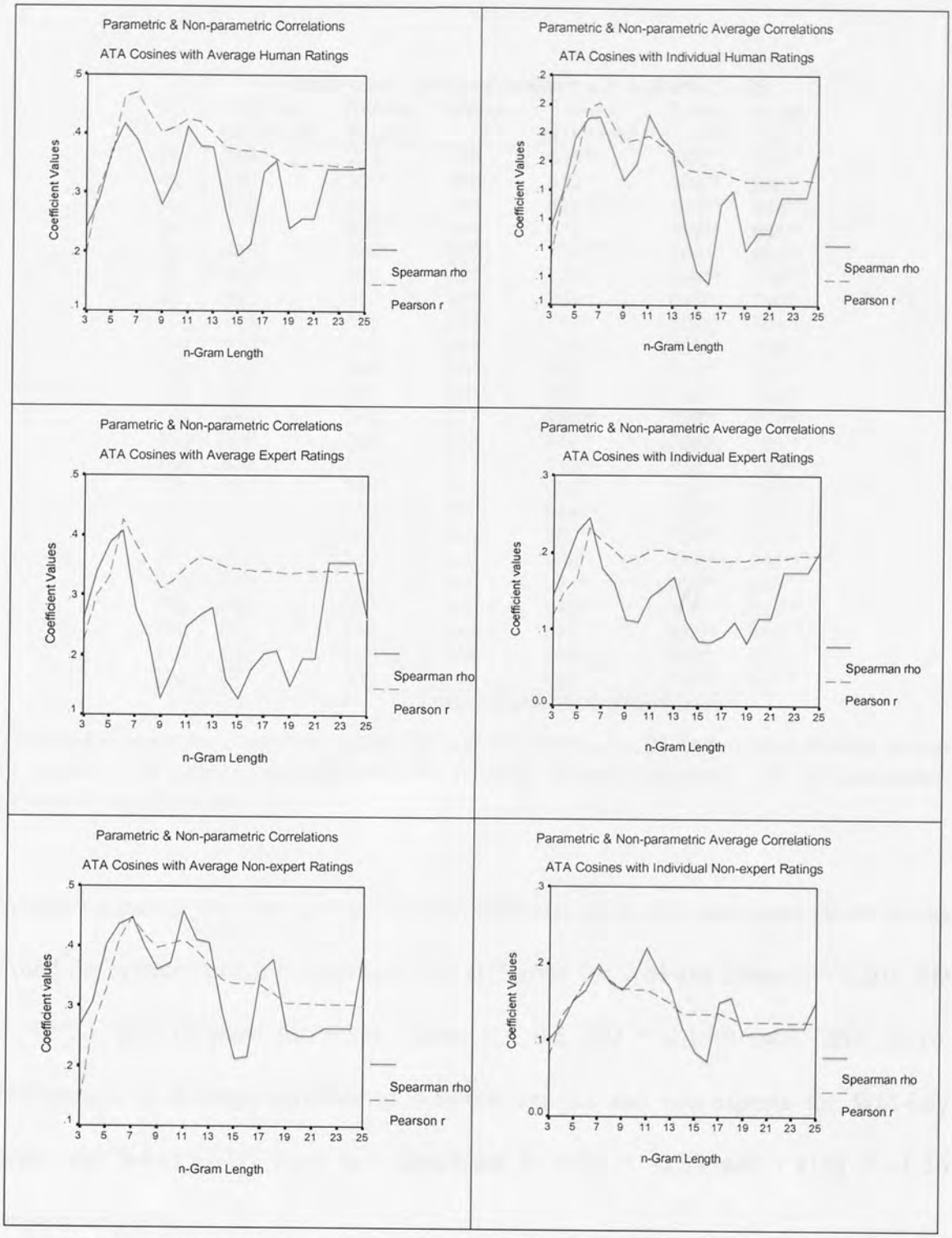


Figure 2-6 Comparisons of parametric and non-parametric correlation values between human ratings and n-gram lengths 3 to 25 for Working Memory-keyword document set

Comparisons with average human ratings demonstrated an optimal n-gram length of 7- gram for WM-key ($r = 0.469$; $n = 28$; $p < 0.05$) and 14-gram for S-key ($r = 0.798$; $n = 28$; $p < 0.01$) (see Table 2-9). Optimal n-gram length was relatively consistent between experts and non-experts.

	Average Human Ratings Correlated with n-Grams 3 - 25					
	WM-key non-expert	WM-key expert	WM-key all	S-key non-expert	S-key expert	S-key all
3g	.143	.224	.196	.525**	.468*	.534**
4g	.262	.300	.306	.572**	.504**	.580**
5g	.326	.332	.364	.638**	.522**	.633**
6g	.411*	.427*	.464*	.672**	.542**	.664**
7g	.448*	.384*	.469*	.704**	.562**	.694**
8g	.418*	.349	.432*	.722**	.594**	.718**
9g	.397*	.312	.401*	.734**	.645**	.744**
10g	.404*	.325	.411*	.740**	.685**	.763**
11g	.410*	.348	.425*	.735**	.705**	.765**
12g	.394*	.364	.420*	.749**	.727**	.783**
13g	.369	.357	.400*	.745**	.740**	.785**
14g	.344	.346	.377*	.755**	.758**	.798**
15g	.338	.344	.372	.749**	.765**	.796**
16g	.339	.341	.371	.712**	.762**	.769**
17g	.337	.340	.369	.676**	.758**	.742**
18g	.322	.339	.357	.648**	.747**	.719**
19g	.305	.337	.345	.609**	.724**	.683**
20g	.304	.339	.345	.589**	.712**	.665**
21g	.304	.340	.346	.571**	.702**	.649**
22g	.303	.341	.346	.548**	.686**	.627**
23g	.303	.340	.345	.519**	.665**	.599**
24g	.303	.339	.344	.496**	.648**	.577**
25g	.302	.338	.343	.460*	.618**	.541**

* $p < 0.05$; ** $p < 0.001$ Highest correlations are highlighted

Table 2-9 Correlations between cosines for n-gram lengths 3 – 25 and average human ratings of experts, non experts, and all raters for Working Memory-keywords and Schizophrenia-keywords document sets

When comparing the averaged correlation between ATA and individual raters it was found the optimal length n-gram occurred at 7-gram for WM-key (mean r = 0.201, SD = 0.125), and 16-gram for S-key (mean r = 0.5, SD = 0.159) (see Table 2-10). Differences in average correlations between experts and non-experts for WM-key (7g), and S-key (14g) were not significant (t (16) = -0.33 and t (16) = -1.36 respectively).

	WM-key non-expert		WM-key expert		WM-key all		S-key non-expert		S-key expert		S-key All	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
3g	0.079	0.205	0.118	0.168	0.092	0.189	0.323	0.138	0.335	0.104	0.327	0.127
4g	0.131	0.178	0.15	0.161	0.137	0.168	0.338	0.161	0.357	0.088	0.344	0.14
5g	0.151	0.147	0.167	0.129	0.156	0.137	0.371	0.188	0.388	0.088	0.377	0.16
6g	0.178	0.125	0.232	0.102	0.196	0.118	0.391	0.178	0.407	0.092	0.396	0.154
7g	0.194	0.137	0.215	0.109	0.201	0.125	0.41	0.178	0.427	0.106	0.416	0.156
8g	0.177	0.149	0.203	0.113	0.186	0.135	0.42	0.171	0.457	0.13	0.432	0.158
9g	0.167	0.159	0.187	0.135	0.173	0.148	0.431	0.162	0.502	0.176	0.454	0.17
10g	0.166	0.155	0.193	0.146	0.175	0.148	0.437	0.163	0.525	0.189	0.466	0.178
11g	0.166	0.151	0.202	0.163	0.178	0.152	0.435	0.161	0.537	0.205	0.469	0.184
12g	0.157	0.143	0.203	0.176	0.172	0.151	0.446	0.156	0.55	0.2	0.481	0.18
13g	0.148	0.137	0.198	0.183	0.165	0.15	0.449	0.151	0.56	0.207	0.486	0.181
14g	0.135	0.141	0.195	0.187	0.155	0.155	0.46	0.147	0.573	0.188	0.498	0.171
15g	0.134	0.143	0.192	0.192	0.153	0.157	0.461	0.138	0.578	0.166	0.5	0.159
16g	0.134	0.144	0.188	0.199	0.152	0.161	0.445	0.134	0.577	0.141	0.489	0.151
17g	0.135	0.145	0.188	0.2	0.153	0.161	0.427	0.137	0.572	0.123	0.475	0.15
18g	0.129	0.147	0.19	0.197	0.149	0.162	0.413	0.144	0.565	0.106	0.463	0.151
19g	0.123	0.151	0.19	0.194	0.146	0.164	0.391	0.152	0.545	0.088	0.442	0.153
20g	0.123	0.151	0.192	0.195	0.146	0.164	0.378	0.156	0.538	0.077	0.432	0.154
21g	0.123	0.151	0.192	0.194	0.146	0.164	0.368	0.157	0.528	0.07	0.422	0.154
22g	0.123	0.151	0.192	0.194	0.146	0.164	0.353	0.158	0.517	0.069	0.408	0.156
23g	0.122	0.152	0.195	0.192	0.146	0.165	0.334	0.162	0.498	0.068	0.389	0.158
24g	0.123	0.151	0.195	0.188	0.147	0.163	0.319	0.162	0.487	0.071	0.375	0.159
25g	0.122	0.151	0.195	0.188	0.146	0.162	0.297	0.166	0.462	0.068	0.352	0.161

* $p < 0.05$; ** $p < 0.001$ Highest correlations are highlighted

Table 2-10 Average Pearson r coefficients between cosines for n-grams lengths 3 to 25 and individual human ratings from experts, non-experts, and all raters for Working Memory-keywords and Schizophrenia-keywords document sets

2.4.3 Discussion

For WM-key the optimum length of n-gram reduced from 12-gram to 7-gram and still produced higher levels of agreement both with average human ratings and individual human ratings, than raters produced amongst themselves (see Table 2-9 and Table 2-10). However the correlations were weaker than those produced in Experiment One and the difference between the amount of variance accounted for by the optimum length n-gram and human ratings was smaller. In this case removal of keywords appears to reduce an element of shared agreement between humans and n-gram based ATA.

For S-key document set however, the n-gram length did not reduce when compared to Experiment One, and the coefficient values for S-key document set approximated

those for S. For instance comparisons of average ratings for all raters for S 15-gram and for S-key 14-gram were similar, as were correlations between ATA and individual raters averaged across all raters for S 16g, and for S-key 15g (see Tables 2-3, 2-4, 2-9, & 2-10).

Based on these result it cannot be assumed that keywords are a primary factor in determining n-gram length and it is unclear whether individuals use keywords to make their judgements of similarity, or to what extent. It is possibly the case that keywords are used, but the degree to which they are relied upon is determined by other factors. For instance if the number of times keywords are present in documents is clearly discriminatory within document pairings people may use them when making similarity judgements, if however keywords appear evenly across a document set they are unlikely to be of benefit when making such judgements.

It is interesting to note that the read ease score for both S and WM document sets increased when keywords were removed (see Table 2-2 and Table 2-8). While the differences between read ease score for the two document sets in each experiment were not statistically significant both WM and WM-key were higher than S, and S-key further supporting the suggestion that n-gram length may be related to the readability of the documents within the set. This is certainly an issue that warrants more detailed investigation.

2.5 Experiment Four

2.5.1 Methodology

2.5.1.1 Document Set Preparation

Six new document sets each containing 16 documents instead of eight was created from the original four document sets used in Experiments One, and Two. The technical document sets WM, and S were expanded to WM-twin, S-twin, and WMS, respectively, while the non-technical sets JR and P were expanded to JR-twin, P-twin, and JRP respectively –see below for explanations of new document sets.

Within each of the original sets WM, S, JR, and P, individual member documents were divided into two separate documents by assigning alternate sentences to each, to create WM-twin, S-twin, JR-twin, and P-twin. Pairings of new documents containing sentences from the same source document were considered twins; all other pairings were deemed non-twins.

Combining the document sets WM and S, and JR and P respectively created WMS and JRP. Pairings of documents originating from the same document set e.g. WM were considered 'twins' while pairings of documents originating from separate sets e.g. WM and S were considered non-twins.

2.5.1.2 Procedure

As the new documents comprised disjointed sentences human raters would have struggled to make any sense of them, human ratings were therefore not obtained. However, based on evidence from Damashek (1995a), it was expected that n-gram analysis would identify related documents by producing higher average cosines for twin documents compared to non-twins. For each of the six document sets n-gram analyses were conducted using n-gram lengths 3 to 25 and LSI.

2.5.2 Results

In order to account for differences in size of cosine obtained across ATA techniques (i.e. n-grams 3 to 25, and LSI) and between document sets when calculating differences between twin and non-twin document pairings, mean cosine values obtained were standardised using z-scores. Table 2-11 shows standardised average cosine values for twin documents, non-twin documents, and difference values between the two as produced by n-grams 3 – 25 & LSI for technical document sets WM-twin, S-twin, & WMS.

	WM-twin Twin	WM-twin Non-twin	WM-twin Difference	S-twin Twin	S-twin Non-twin	S-twin Difference	WMS Twin	WMS Non-twin	WMS Difference
3g	2.005	-0.143	2.149	1.415	-0.101	1.516	0.410	-0.358	0.768
4g	2.401	-0.171	2.572	1.809	-0.129	1.938	0.509	-0.445	0.954
5g	2.706	-0.193	2.899	2.038	-0.146	2.184	0.550	-0.482	1.032
6g	2.878	-0.206	3.083	2.197	-0.157	2.354	0.586	-0.512	1.098
7g	2.855	-0.204	3.059	2.195	-0.157	2.352	0.582	-0.510	1.092
8g	2.830	-0.202	3.032	2.153	-0.154	2.307	0.562	-0.492	1.054
9g	2.824	-0.202	3.025	2.150	-0.154	2.304	0.571	-0.500	1.071
10g	2.795	-0.200	2.995	2.196	-0.157	2.353	0.566	-0.496	1.062
11g	2.754	-0.197	2.951	2.273	-0.162	2.436	0.552	-0.483	1.035
12g	2.765	-0.198	2.963	2.340	-0.167	2.507	0.548	-0.480	1.028
13g	2.789	-0.199	2.988	2.434	-0.174	2.608	0.492	-0.431	0.923
14g	2.842	-0.203	3.045	2.515	-0.180	2.694	0.475	-0.415	0.890
15g	2.906	-0.208	3.113	2.533	-0.181	2.714	0.436	-0.381	0.817
16g	2.961	-0.212	3.173	2.488	-0.178	2.666	0.355	-0.311	0.667
17g	2.946	-0.210	3.157	2.419	-0.173	2.592	0.254	-0.223	0.477
18g	2.883	-0.206	3.089	2.338	-0.167	2.505	0.161	-0.141	0.301
19g	2.806	-0.200	3.007	2.199	-0.157	2.356	0.159	-0.139	0.298
20g	2.714	-0.194	2.908	2.095	-0.150	2.244	0.136	-0.119	0.255
21g	2.591	-0.185	2.776	2.018	-0.144	2.162	0.136	-0.119	0.255
22g	2.442	-0.174	2.616	1.947	-0.139	2.086	0.139	-0.122	0.261
23g	2.259	-0.161	2.420	1.855	-0.133	1.987	0.139	-0.122	0.261
24g	2.048	-0.146	2.194	1.721	-0.123	1.844	0.139	-0.122	0.261
25g	1.969	-0.141	2.109	1.663	-0.119	1.781	0.125	-0.110	0.235
LSI	3.497	-0.250	3.747	3.005	-0.215	3.219	0.721	-0.631	1.352

Largest difference between mean cosines for twin vs. non-twin pairs is highlighted

Table 2-11 Standardised mean cosines between twin (n = 8) and non-twin (n = 112) document pairs for Working Memory-twin, Schizophrenia-twin, and WMS document sets together with difference in mean cosine values between twin and non-twin pairs.

As can be seen from highlighted cells in Table 2-11 the optimal length n-grams in terms of discriminating between twin and non-twin document pairs, were 16-gram (difference = 3.17) for WM-twin, 15-gram (difference = 2.71) for S-twin, and 6-gram (difference = 1.1) for WMS. The discriminatory value for LSI was higher for all

document sets (difference = 3.75, 3.22, and 1.35, respectively).

Mixed model 2 x 2 ANOVAs were performed to examine the effects of ATA and document type. The repeated measures factor 'ATA' used 16-gram and LSI as levels for WM-twin, 15-gram and LSI for S-twin, and 6-gram and LSI for WMS document sets. In all cases the independent measures factor 'type' had two levels 'twin' and 'non-twin'.

For WM-twin a significant main effect of type was observed $F(1,118) = 81.3$; $p < 0.001$ with twins obtaining higher cosine values overall than non-twins (twin marginal mean = 3.23, SE = 0.128 ; non-twin marginal mean = -0.23, SE = 0.034). A main effect of ATA just reached significance $F(1,118) = 4.10$, $p = 0.048$ where LSI produced higher mean cosine overall than 16 gram (LSI marginal mean = 1.62, SE = 0.063; 16-gram marginal mean = 1.38, SE = 0.112). The interaction between type and ATA was also significant $F(1,118) = 5.34$; $p < 0.05$. This appears to be due to there being little differences between mean cosine values produced by the two ATA protocols for non-twin document sets but LSI produces relatively larger cosines for twins (see Figure 2-7 and Table 2-11 for means).

For S-twin, there was a significant main effect of type $F(1,118) = 177.88$; $p < 0.001$, with twins obtaining higher cosine values than non-twins (twin marginal mean = 2.76, SE = 0.219 ; non-twin marginal mean = -0.197, SE = 0.058). The main effect of ATA technique was also significant $F(1,118) = 4.59$, $p < 0.05$, with LSI producing higher mean cosine overall than 15-gram (LSI marginal mean = 1.395, SE = 0.109 ; 15g marginal mean = 1.167, SE = 0.136). A significant interaction between document pair

type and ATA was also observed, $F(1,118) = 6.11$; $p < 0.05$ (see Figure 2-7 and Table 2-11 for means).

The document set WMS showed only a significant main effect for type $F(1,118) = 81.3$; $p < 0.001$ with twin document pairs producing higher mean cosine values than non-twins (0.65, and -0.57 respectively). The interaction between document type and ATA was significant ($F(1,118) = 6.55$; $p < 0.05$), such that LSI produced higher mean cosines for twin document pairs than 6-gram but 6-gram produced higher mean cosines for non-twin pairs (see Figure 2-7 and Table 2-11).

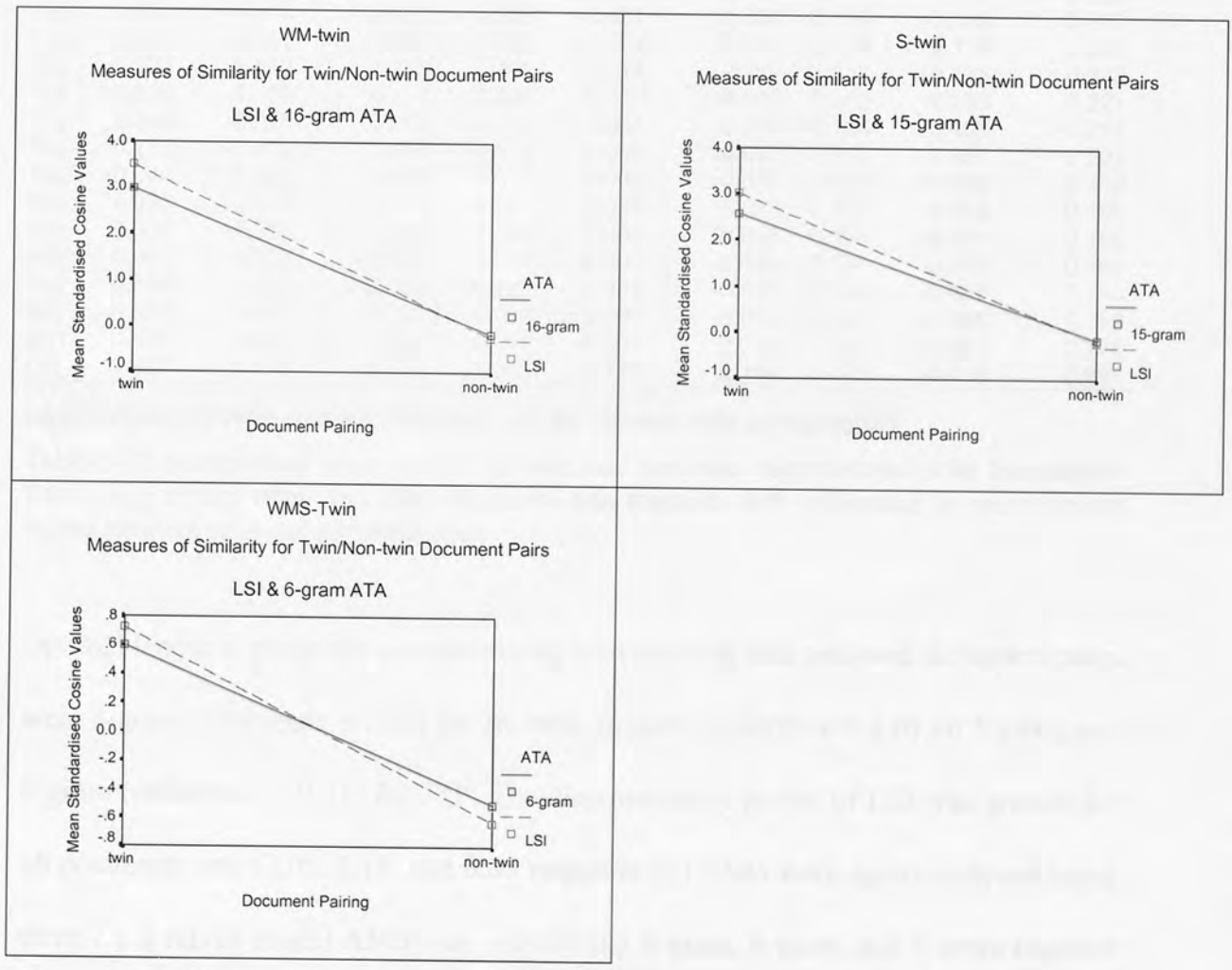


Figure 2-7 Working Memory-twin, Schizophrenia-twin, and WMS document sets, comparisons between mean standardised cosine values for twin & non-twin pairings, and optimal length n-gram and LSI.

Average cosine values for twin and non-twin document pairs together with differences between the two for the non-technical document sets JR-twin, P-twin, and JRP are shown in Table 2-12. Again these values have been normalised using z scores.

	JR-twin Twin	JR-twin Non-twin	JR-twin Difference	P-twin Twin	P-twin Non-twin	P-twin Difference	JRP Twin	JRP Non-twin	JRP Difference
3g	0.974	-0.070	1.044	0.997	-0.071	1.068	0.141	-0.123	0.264
4g	1.169	-0.084	1.253	1.106	-0.079	1.185	0.190	-0.166	0.356
5g	1.376	-0.098	1.474	1.221	-0.087	1.308	0.228	-0.199	0.427
6g	1.544	-0.110	1.655	1.494	-0.107	1.601	0.275	-0.240	0.515
7g	1.555	-0.111	1.666	1.471	-0.105	1.576	0.264	-0.231	0.496
8g	1.578	-0.113	1.690	1.354	-0.097	1.450	0.243	-0.213	0.455
9g	1.549	-0.111	1.660	1.225	-0.088	1.313	0.217	-0.190	0.406
10g	1.458	-0.104	1.562	1.034	-0.074	1.108	0.184	-0.161	0.344
11g	1.355	-0.097	1.452	0.857	-0.061	0.919	0.165	-0.145	0.310
12g	1.221	-0.087	1.308	0.690	-0.049	0.739	0.150	-0.132	0.282
13g	1.036	-0.074	1.110	0.523	-0.037	0.560	0.139	-0.122	0.261
14g	0.874	-0.062	0.937	0.342	-0.024	0.366	0.131	-0.115	0.246
15g	0.723	-0.052	0.775	0.198	-0.014	0.212	0.124	-0.108	0.232
16g	0.520	-0.037	0.558	0.096	-0.007	0.103	0.118	-0.103	0.221
17g	0.344	-0.025	0.368	0.018	-0.001	0.019	0.114	-0.100	0.214
18g	0.173	-0.012	0.185	-0.039	0.003	-0.042	0.111	-0.097	0.207
19g	0.044	-0.003	0.047	-0.072	0.005	-0.077	0.108	-0.094	0.202
20g	0.033	-0.002	0.035	-0.077	0.006	-0.083	0.106	-0.093	0.199
21g	0.018	-0.001	0.019	-0.062	0.004	-0.066	0.104	-0.091	0.195
22g	-0.001	0.000	-0.001	-0.042	0.003	-0.045	0.101	-0.089	0.190
23g	0.004	0.000	0.005	-0.023	0.002	-0.025	0.099	-0.087	0.186
24g	0.006	0.000	0.007	-0.006	0.000	-0.006	0.097	-0.085	0.182
25g	0.006	0.000	0.007	0.015	-0.001	0.016	0.095	-0.083	0.178
LSI	1.917	-0.137	2.054	2.043	-0.146	2.189	0.475	-0.416	0.891

Largest difference between mean cosines for twin vs. non-twin pairs are highlighted

Table 2-12 Standardised mean cosines of twin and non-twin document pairs for Journalists' Risks-twin, Piracy-twin, and JRP document sets together with difference in mean cosine values between twin and non-twin pairs.

Optimal length n-grams for discriminating between twin and non-twin document pairs were 8-gram (difference = 1.69) for JR-twin, 6-gram (difference = 1.6) for P-twin, and 6-gram (difference = 0.51) for JRP. The discriminatory power of LSI was greater for all document sets (2.05, 2.19, and 0.89 respectively). Data were again analysed using three 2 x 2 mixed model ANOVAs, substituting 8-gram, 6-gram, and 6-gram together with LSI as the 2 levels of ATA for the respective document sets (JR-twin, P-twin and JRP). Again in all cases the independent measures factor was document pair 'type'

with two levels 'twin' & 'non-twin'.

For JR-twin a significant main effect was demonstrated for type, $F(1,118) = 35.09$; $p < 0.001$ and ATA, $F(1,118) = 4.81$; $p < 0.05$. Twin document pairings showed higher cosine values overall than non-twins (twin marginal mean = 1.733, SE = .304; non-twin marginal mean = -.124, SE = .081) and LSI produced higher mean cosine than 8-gram (LSI mean = 0.89, SE = 0.15; 8-gram marginal mean = 0.719, SE = 0.167). The interaction between type and ATA was also significant $F(1,118) = 6.41$; $p < 0.05$. LSI produced higher mean cosines for twin document pairs than 8-gram but showed there was little difference in mean cosine values for non-twin pairing (see Figure 2-8 and Table 2-12).

When considering the P-twin document set, there was demonstrated a significant main effect of document pair type $F(1,118) = 41.25$; $p < 0.001$ with twin pairs obtaining higher cosine values overall than non-twin pairs (twin marginal mean = 1.769, SE = 0.285; non-twin marginal mean = -0.126, SE = 0.076). The main effect of ATA approached significance $F(1,118) = 3.86$, $p = 0.052$ with LSI tending to produce higher mean cosine values overall than 6-gram (LSI marginal mean = 0.949, SE = 0.154; 6-gram marginal mean = 0.694, SE = 0.168). A significant interaction between type and ATA was observed, $F(1,118) = 5.14$; $p < 0.05$, with greater discriminatory power for LSI over 6-gram for twin pairings than for non-twin pairings (see Figure 2-7 and Table 2-11).

With respect to the document set JRP, there was only a significant main effect for document type $F(1,118) = 20.01$; $p < 0.001$, with twin document pairs producing

higher mean cosine values than non-twin document pairs (0.38, and –0.33 respectively). The interaction between type and ATA was also significant $F(1,118) = 7.75$; $p < 0.01$, with LSI producing higher mean cosines for twin document pairs than 6-gram but 6-gram producing higher mean cosines for non-twin pairs (see Figure 2-8 and Table 2-12).

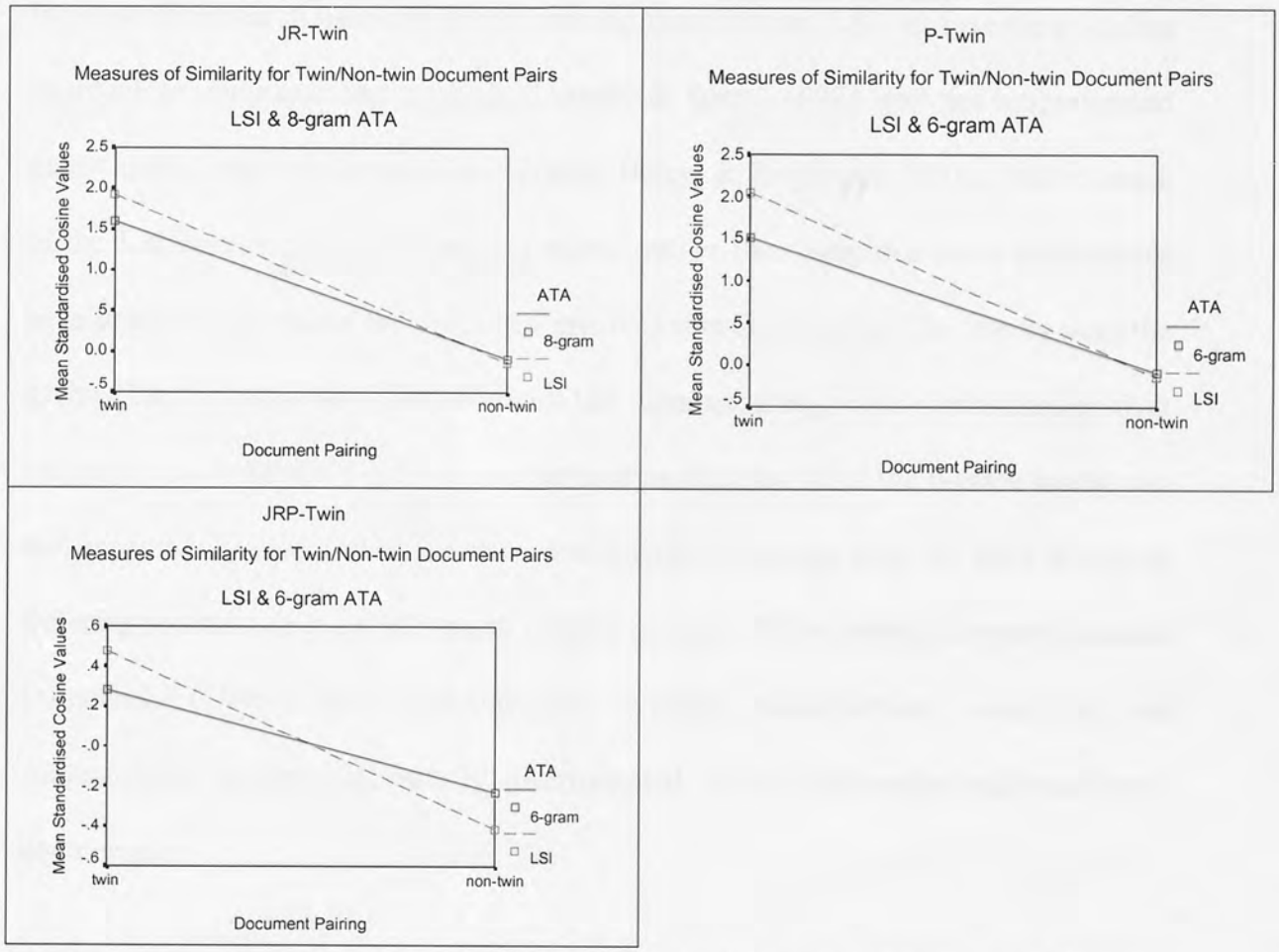


Figure 2-8 Journalists' Risks-twin, Piracy-twin, and JRP document sets, comparisons between mean standardised cosine values for twin and non-twin pairings, and optimal length n-gram and LSI

2.5.3 Discussion

The results of this experiment show that in all cases ATA can discriminate between twin and non-twin documents, whether twins come from the same source document or

the same source document set. The results also show that when the task is specific to documents only, rather than comparing ATA analysis to human judgements, LSI performs better than n-grams. Although with some document sets the main effect between n-gram and LSI (i.e. the difference in mean cosine values) was not significant, the interaction was significant in all cases, demonstrating a superior discriminatory ability of LSI.

In recent IR research particularly information visualisation, LSI has become a leading approach to automatic text analysis (Letsche & Berry, 1997), and has outperformed many lexical matching approaches (Zhang, Berry, & Raghavan, 2001). Within much of the LSI research weighted n-gram strings rather than weighted word occurrences have been used, however the processes involved remain complex. The results from the current Experiments show that, although LSI demonstrates greater discrimination than the basic non-weighted n-gram co-occurrence model used here, the n-gram model can still successfully discriminate between documents originating from the same source or differing sources while using a much simpler process. These findings support those of Damashek (1995a) who demonstrated n-gram co-occurrence, used in his Acquaintance system successfully discriminated 'twin' documents and 'non-twin' documents.

2.6 Conclusions

It can be argued that human ratings of document similarity should be considered the 'gold standard' against which to evaluate ATA, as people are the information seekers. However these judgements are affected by a multitude of factors both internal and

external to the individuals making the judgements (e.g. Schamber, et al., 1990). One of the key papers criticising the use of distance as a measure of similarity emphasises the fact that human perceptual judgements of similarity are contextually driven and vary not only between individuals but also in terms of direction of judgement, e.g. the judged similarity between a and b is not the same as the judged similarity between b and a, (Tversky, 1977). Given the influence of contextual factors on individuals' judgements of document similarity an effective ATA programme must either be adaptable to an individual, or be able to identify the context free elements of documents which individuals agree on when judging semantic similarity. The results in terms of the high degree of shared variance between ATA and average human judgements suggest that both n-grams and LSI identify context free inter-document similarity. In terms of n-grams this is to some degree dependent on selecting the optimum length letter string. However, in the current study even when selecting the poorest performing n-gram length a difference only occurred for the Schizophrenia document set. The results also show that while LSI is superior in identifying document similarity in terms of document content alone, when does not predict human judgements of similarity as well as n-grams. The current experiments have shown that while n-gram based VSM systems have performed below average in terms of dedicated information retrieval at previous TREC conferences, there is a benefit to using un-weighted n-gram based co-occurrence models for the purpose of general database visualisation, using spatial-semantic relationships aimed at a varied user group.

The n-gram based analysis used in these experiments will provide a robust measure of document similarity that can be utilised by the majority of individuals. Given that

people agree to only a limited extent on document similarity, a spatial mapping of an information space needs to utilise the context free aspect that people share in order to facilitate acquisition of their cognitive map of the space. The results of the current set of experiments suggest n-gram based analysis can do this, and is a simpler and possibly more reliable method for this purpose than LSI due to its power to identify context free semantic similarities.

The VSM is a particularly useful measurement of document similarity for IR systems that use visualisation as a means of database presentation to the user. The similarity measure (i.e. the angle) between two vectors allows the documents to be mapped into a multi-dimensional space where inter-document similarity is judged by relative proximity. Having demonstrated n-gram analysis to be a suitable method of analysis subsequent chapters discuss its use in constructing the VE database used for the remaining experimental work in this thesis.

3 Browsing Performance in a VE Database

3.1 Introduction

This chapter describes a study of the benefits of using spatial-semantic mapping to visually present the contents of an electronic database. Information retrieval performance was evaluated while participants browsed a database presented as a navigable virtual environment (VE). In spatial data management systems (SDMSs) of the type used in the current study (explained in Chapter One – Section 1.3.2), the placement of documents in the VE is determined by the semantic properties of the documents; semantically similar documents are mapped closer together than semantically dissimilar documents, i.e. spatial proximity = semantic similarity. The extent to which participants were able to use the spatial relationships between contents to locate semantically relevant information was assessed. The importance of cognitive factors in terms of individual differences in spatial ability, spatial working memory, and associative memory, in utilising the spatial-semantic mapping were also examined. Reasons for employing visualisation techniques and spatial-semantic mapping in information retrieval tasks are presented in section 3.1.1. This is then followed by a discussion of the role of individual differences in cognitive ability as determinants of user performance (see section 3.1.2). Finally the nature and purpose of the current study is presented in section 3.1.3.

3.1.1 Information Visualisation

One of the principle disadvantages with information retrieval systems widely employed in IR (e.g. internet search engines such as 'Yahoo' or 'Google'), is that the user is presented with a list of documents which are ranked on the basis of likely

relevance to an input query. The ranking and identification of such documents is calculated using a variety of algorithms. These vary in complexity but are principally based on the co-occurrence of keywords or terms in the query and the retrieved documents. The documents are then ranked on how frequently the word appears in relation to the length of the document, the number of documents in the corpora, the number of different keywords contained in the document, etc. (see Chapter 2 – Section 2.1, and Allan, et al., (2001), for further details).

As previously identified, major problems with Boolean logic based IR approaches can include unmanageably large quantities of documents being retrieved with relevant documents being placed far from the top rankings and non-relevant documents appearing near the top. Reasons for this are varied but include user expertise in formulating the initial queries, the specificity of the information need (a loosely structured information need where either the information requirement is poorly defined, or very broad makes query formulation difficult – (Marchionini & Shneiderman, 1988)), and natural language problems such as polysemy (a single word or phrase can have multiple meanings) and synonymy (multiple words can share the same meaning). The problem of incorrect document positioning within the ranking is further compounded by the fact that users rarely look beyond the first page of retrieved results (Jansen, et al., 2000) and so fail to identify relevant documents lower down the rankings. Representing the retrieved documents as visualisations in a virtual environment has the advantage of allowing many more documents to be shown in a single window. Using spatial-semantic mapping to determine the position of the documents, results in clusters of documents that share meaning without necessarily sharing keywords. This potentially alleviates many of the problems described above

as; i) users do not need to formulate Boolean based queries; ii) the problems associated with polysemy and synonymy due to reliance on keywords are removed; iii) users do not have to visit multiple pages to view all the retrieved documents.

It has been shown that it is important for the user to develop a good mental model of the system and the information space in order to facilitate improved information retrieval. For example, Borgman (1999) showed that performance increased on complex retrieval tasks when users were trained to develop a conceptually driven mental model rather than a procedural model of a Boolean based IR system. During the initial stages of the study only 32 out of an initial 43 participants reached the initial competency requirement, indicating the degree of general difficulty involved in Boolean based IR. The authors who also classified and compared users on a high/low technical scale on the basis of their degree major, found that significantly more low technical users were excluded on the basis of failure to meet the competency criterion. A possible explanation for this is that the mental models of low-technical users are qualitatively different to those of high technical users and as such low-tech users have greater difficulty acquiring a good mental model of a system that uses logic based operators to locate information. It has been argued in Chapter One – Section 1.2, that to derive semantic context we use a conceptual space in which semantic and linguistic processing is spatially driven by an inherent evolutionary adaptation of operating in a physical three-dimensional environment (e.g. Jackendoff, 1995; Glenberg, 1997; Dillon 2000; Gardenfors, 2000; and Buchanan, Westbury, & Burgess, 2001). An information retrieval system that is spatially organised and therefore utilises this relationship between spatial and semantic processing is likely to mesh well with users' cognitive models and will therefore promote less effortful assimilation and improved

performance.

Willie & Bruza (1995) examined the potential of using graphical methods (Venn diagrams) to express underlying information needs without using explicit Boolean commands. The task involved participants representing queries related to the selection of shapes as Venn diagrams. The authors found that two principle visualisations were employed which they termed 'Set Assembly' where the ellipses overlapped, and 'Set Refinement' where ellipses representing sub-levels of information were placed inside higher level ellipses (nested). Although Set Assembly was seen to reflect Boolean based queries Set Assembly and Set Refinement were used equally. The complexity of the query determined which type of representation was employed resulting in users employing both types equally. Users did demonstrate a preference for a particular approach, which was demonstrated by the consistency in adopted starting approaches, and reflected some influence of individual differences. However as the complexity of the query increased users tended to combine the two approaches. The authors concluded that, users' mental models of the underlying information space impact on their ability to use Boolean logic in query formulation, and that a mismatch between their mental model and the actual information space can result in poorly formulated queries. Using visualisation helped users better match their own cognitive models to the environment and vice versa.

Information visualisation as an approach to information retrieval developed to combat the problems described above and to facilitate improved synthesis between users' mental models and the physical models of the information spaces and IR systems. The SPIRE project (the Spatial Paradigm for Information Retrieval and Exploration) for

instance was commissioned in 1994 by the Department of Energy and the US intelligence agencies. Its aim was to try and find ways that the thousands of electronic documents presented to analysts on a weekly basis could be categorised and the information contained visually organised to enable optimal access to that information (Wise, 1999). The purpose of SPIRE and the visualisations systems produced by the project (e.g. Galaxies, and Themescapes – now known as Themeview) was to produce ecologically valid (i.e. a match between users' internal representations of the external environment) visualisations of the information contained in large quantities of documents: Visualisations that represent a coherent mapping of the semantic information contained in the text with users' cognitive structure and interpretation of that information. This was done using processes based on vector space modelling (VSM) as described in Chapter Two – Section 2.1.1, that resulted in documents being mapped into high-dimensional spaces which could then be reduced to two or three-dimensional spaces. In the case of the SPIRE systems 3D spaces were employed. One of the important features of the SPIRE systems is that, in the final visualisation, documents clustered around topics and themes that themselves had been extracted from the mathematical analysis of the document corpora and as such weren't based on keywords or terms, or pre-established topics. The visualisations (in the case of 'Galaxies' this was a representation that resembled the night sky, and for Themescapes a 3D topographical map similar to that used for geographical representations of the Earth's structure) contained the inherent structure of the semantic information contained in the documents enabling information seekers to intuitively perceive the semantic structure of the document corpora in the way the natural environment is perceived.

Recent studies have supported the work of information visualisation specialists and shown that using graphical and visual representations to present document collections can benefit users in IR tasks (e.g. Cugini, et al., 1997; Chen, et al., 1999; Fox, et al., 1999; and Chen & Czerwinski, 2000). Stanney & Salvendy (1995), showed that reducing the need to visualise mentally the embedded structure of traditional IR systems (e.g. menu structures) by producing 2D and 3D visualisations of the system structure helped individuals, particularly those with low spatial ability, to access and retrieve information more effectively.

Allan et al. (2001) found that, when Boolean based logic was used for initial query formulation, a system (Lighthouse – see Chapter One – Section 1.3.2) that included a spatial visualisation of retrieved documents improved recall and precision compared to a traditional search engine (ZPRISE). The spatial visualisation, in which documents were represented as spheres, used VSM to judge inter-document semantic similarity, and proximity mapping to display document similarity. The list produced by ZPRISE was ranked by relevance to the query based on keyword occurrence. The visualisation included clusters of semantically similar documents that were not necessarily close in the ranked list, which was shown to improve retrieval performance.

The NIST Information Retrieval Visualisation Engine (NIRVE) (Cugini et al., 1997) was also designed as a visualisation system for presenting the retrieved results of a traditional query driven search engine. Subsets of keywords are mapped to a concept space by the user, enabling the user to imprint their own conceptual model on the visualisation. The system then automatically maps retrieved documents to the concept space using the occurrence of initial keywords. A single document can be mapped to

more than one concept dependent on the user's original mapping of keywords to the concept space. Sebrechts, et al. (1999) evaluated NIRVE using 2D and 3D visualisations and a comparative text version where documents were presented as lists under each concept heading. The results were mixed but generally found that participants in the text based condition completed the tasks quicker. However improvement in task performance across six sessions was greatest for the 3D condition, suggesting that familiarity with the nature of the system had a role to play, and that the observed advantages of text based presentation were in fact due to familiarity with this type of presentation. Once users became familiar with using 2 and 3 dimensional visualisations, the text based presentation no longer had an advantage. There were unfortunately confounding issues that may have impacted on the reliability of these results. For instance the computer used for presentation of the 3D version was "substantially slower" and the authors commented upon the frustration displayed by users in this condition. In the text based condition users had to scroll down the list of documents however it is not clear how long this list was. It may be that for a larger document set the time required to scroll and scan the list would exceed any time disadvantages displayed in the 2D and 3D conditions. It was also found that text based search was poorer than 2D or 3D search when documents were linked to neighbouring clusters i.e. not relevant to a single concept (this was visually displayed by a coloured arc adjoining the relevant clusters). These findings are clearly mixed and show the complexity of employing visualisation techniques and evaluating IR systems employing these techniques.

In a simpler and more generic system, Westerman & Cribbin (2000b) looked at the benefits of using spatial-semantic mapping to present an entire document collection, in

which query formulation and keyword occurrence played no part at any stage. They showed that, when using a VSM of document content analysis and a spatial data management system (SDMS) VE to present the entire contents of a database in a single visualisation (with semantically similar documents mapping closer together in space), the quality of spatial-semantic mapping used in the visualisation positively influenced performance. The authors employed a SDMS similar to that used in the current study (see section 3.2.1), but the mapped database consisted of single words representing real-word objects related to particular categories rather than documents (e.g. "house" as an example of a "type of human dwelling"; "chair" as an "article of furniture"). They found that performance on a "locate and retrieve" task improved on a variety of dependent measures (e.g. time per trial, and timed-out trials) when spatial mapping accounted for greater semantic variance between objects. This finding was stable when objects were mapped into either two- or three-dimensional organisations. However, when using a three-dimensional solution, user performance was poorer than when employing a two-dimensional layout. Performance was comparable even when the two-dimensional solution accounted for only 50 to 70% of the semantic variance accounted for by a three-dimensional mapping. This difference was attributed to the additional cognitive demands required for navigating three dimensions. Westerman & Cribbin (2000b) also found that individual differences in cognitive ability were a factor in the performance of IR tasks using visualisation.

3.1.2 Individual Differences in Cognitive Ability

Dillon (2000) highlighted that individual differences between users should not be ignored when designing and evaluating IR systems. Individual differences in computerised tasks have been shown to differentiate performance to a far greater

degree than in non-computerised tasks with differences in retrieval performance ratios of 20:1 and 2:1 respectively, (Egan, 1988). Previous studies examining individual differences in cognitive ability in the field of VE visual information retrieval interfaces (VIRIs) have produced mixed results, some of which are initially counter-intuitive (e.g. (Chen2000)). This study focuses on three particular abilities that have been shown by various studies to impact on the use of SDMSs, although not in a consistent manner (e.g. Westerman & Cribbin, 2000a; Chen & Yu, 2000; and Modjeska & Chignell, 2003). The abilities of interest are associative memory, spatial working memory, and spatial ability.

3.1.2.1 Associative Memory

Associative memory has been implicated in the use of SDMSs using spatial semantic mapping in several studies (e.g. Chen & Macredie; 2000; Chen & Yu, 2000; Westerman & Cribbin, 2000a; and Westerman & Cribbin, 2000b). The results, while generally showing a trend for high associative memory relating to improved performance, have been somewhat mixed. For instance, using Starwalker (an immersive VE SDMS using spatial-semantic mapping), Chen (2000) found a significant positive correlation between associative memory and recall ($r = 0.855$) when users were required to locate and retrieve as many documents relevant to a given search topic as they could. However, in a second study reported in the same paper, associative memory was significantly but negatively correlated with recall on just one topic ($r = -0.619$), and demonstrated no significant relationship with performance on the remaining topics. The authors could offer no explanation of these opposing results but did identify low sample size ($n = 6$) as a possible reliability issue. Westerman & Cribbin, (2000b) found a weak but positive main effect of associative memory on

some measures (time taken on task, number of different objects visited, and the total number of objects visited), suggesting that high associative memory is beneficial in developing a cognitive map of the information space. This advantage was consistent across manipulations of spatial-semantic mapping quality. A similar study examined differences in the protocols for mapping solutions (i.e. ordinal versus interval measures), and used human ratings for similarity judgements. People with low associative memory displayed a performance advantage when an ordinal solution was used, while people with high associative memory did better when using 'true' rating distances (interval) (Westerman & Cribbin, 2000a). The contradictory nature of these results prompted further exploration of associative memory within the current study. This was indexed using similar psychometric measures employed generally within this field and detailed in section 3.2.5.

3.1.2.2 Spatial Ability

Findings regarding the effects of individual differences in spatial ability generally show a positive association with performance on IR tasks (e.g. Chen, 1997).

Chen (1997) asked participants to sketch a map of a spatial-semantic environment after performing a set of search tasks and found that high spatial ability predicted a more accurate representation of the environment. The results of a later study however found that spatial ability was a negative predictor of recall performance (Chen, 2000). Modjeska & Chignell (2003) found positive effects of spatial ability on the general usability of a virtual information world. In their experiment the interface consisted of approximately 1500 pieces of information mapped into a virtual world consisting of "archipelagos, islands, villages, buildings, and floors". Each structure related to

hierarchical subsets of information. Examples of island labels for instance were "entertainment", or "news and media". Visiting these islands would allow the user to locate information relevant to those topics that was further sub-categorised into villages, etc. Factor analysis of dependent measures identified two principal measures i) "doing", which consisted measures such as number of trials completed, distance travelled, and errors per target etc., and ii) "feelings", which consisted measures such as "sense of presence", "ease of use", and "overall enjoyment" etc. Spatial ability was shown to have a significant effect on the "doing" factor with high spatials performing better. There were however no effects of spatial ability on "feelings" measures suggesting that while spatial ability can improve performance in information visualisation retrieval tasks users are not necessarily aware of this in terms of feeling more comfortable. Performance is enhanced but subjective effort is not reduced.

Westerman & Cribbin (2000b) also found positive effects of spatial ability with high spatials performing better on a number of measures including, time taken to perform a task, and number of "timed out" trials. High spatial ability was also associated with improved performance during practice trials.

3.1.2.3 Spatial Working Memory

A relationship between spatial memory and computerised information visualisation has been identified, with some studies showing that spatial memory has a role in IR performance (e.g. Chen & Yu, 2000). The nature of this role however is not clear with the findings reported in the literature being mixed in terms of how significant spatial working memory is in facilitating the use of visual interfaces. Robertson et al. (1998) suggest good spatial memory is an important element in the usability of Data

Mountain, a system designed to facilitate document management by allowing the user to arrange documents spatially in an interactive VE desktop. Cockburn & McKenzie (2003) examined both a VE and a real world model of Data Mountain and also concluded that spatial memory was an "effective aid" to visual IR. However, in both of these studies individual differences in spatial memory were not formally measured.

The study of 'Starwalker', (Chen, 2000) showed that, while associative memory was significantly correlated with performance on an information retrieval task spatial memory was not a significant predictor. However these results were based on a sample size of 10 participants, which raises issues about statistical power and reliability. Performance measures were based on a weighted value of how many documents were retrieved. This value was derived from the position of the document in a ranked list of relevant documents produced by the automatic text analysis program (latent semantic indexing - LSI) used to initially organise the VE interface. The sample size for a correlation analysis is quite small, but of greater concern, in terms of both reliability and validity, is the use of ranked values of relevance for the retrieved documents rather than binary measures (i.e. relevant or not relevant). 'Relevance' has been shown to be a subjective and highly problematic concept especially in terms of IR evaluation as individual judgements of relevance are dynamically influenced by many internal and external factors such as perceived information need and context of task (e.g. Harter, 1992; Harter, 1996; Mizzaro, 1998; and Schamber, et al., 1990). It can be argued that weighting performance on the basis of ATA rankings of 'how relevant' a document is unjustified and leads to poor reliability and validity. In addition while the effectiveness of LSI is not being questioned in terms of judging document similarity (there is strong evidence,

including that presented in Chapter Two – Sections 2.2.2 to 2.5.2, that LSI is an effective means of judging document similarity), the accuracy of the documents' relevance rankings if not moderated or mediated by human judges should be viewed cautiously. This is largely due to the dynamic nature of relevance (e.g. Mizzaro, 1997), whereby judgements of relevance alter dependent on many factors both internal and external to the individual making the judgement. These factors include the nature of the task for which the information is needed, the existing knowledge of the user, and the degree of specificity of information needed. For the tasks used in the study described in the current Chapter, performance was judged using a relevant/not relevant measure.

It was decided therefore, to re-examine the effect of spatial working memory and include it in the cognitive measures using the 'Kit of Factor Referenced Tests'(Ekstrom, et al., 1976) – see section 3.2.6.

3.1.3 The Current Study

The mixed findings presented in the previous sections demonstrate some of the difficulties associated with evaluating the impact of cognitive abilities on the usability of information visualisation systems. In the study by Sebrechts, et al. (1999) (see section 3.1.1), various factors could be responsible for the observed differences, such as the speed of the operating system, the number of documents in the corpora, the relative experience with the nature of the task, and the presentation method. Within the visualisations themselves various cues to document similarity were used and it is difficult to identify which of these cues were beneficial and to what degree. In order to reduce such problems and provide a more coherent evaluation of the important

elements within successful IR, experimental controls can be employed to enable examination of the associated factors individually.

Marchionini & Shneiderman (1988), proposed a framework for information retrieval, in which factors associated with task domain, setting, search system, and user, interact to affect outcomes. They suggested that outcomes consist of both the products of the IR process (e.g. retrieved items,) and processes (e.g. user behaviour, search strategies employed etc.) both of which can be used to assess overall performance, and both of which are important considerations for developing efficient and effective interactive computer-user interfaces. In order to successfully develop and evaluate IR systems it is essential to understand all facets of the IR process, including the role of the user. In the current work a similar distinction is made between measures of retrieval performance (e.g. recall, precision, accuracy, efficiency, time taken on task – current chapter), and measures related to users' browsing behaviour (e.g. distance travelled, rotation, number of nodes/documents visited etc. – see Chapters Four and Five). Within the thesis, retrieval performance and browsing behaviour are first examined separately (Chapter Three examines performance, and Chapter Four studies behaviour). In Chapter Five, study of these elements is combined and extended such that the relationship between user behaviour and performance is examined. It is recognised that while behaviour and performance are generally being studied independently there is a high level of interaction between the two, and where necessary this will be considered. With respect to this framework, the study reported in this chapter examines performance focusing on i) the user (i.e. individual differences in cognitive ability), and ii) the search system (i.e. the use of spatial-semantic mapping to place documents in a VE database presentation). The 'setting'

(which in this instance was experimental), and 'task domain', (which refers to 'a body of knowledge') were not considered experimental variables but were controlled. The task domain used a database of newspaper articles from 1989 to avoid confounding factors such as individual differences in user knowledge. The average age of participants at the time of the experiment (2001) was 19 so most were unlikely to have an in depth knowledge of the specific material.

The current study was designed to expand upon the previously described work of Westerman & Cribbin (2000b), which examined performance on a directed search task using a spatial-semantically mapped SDMS object database, and conducted in collaboration with these authors. In the current work, performance on a browse and retrieve task conducted in the SDMS, which had documents rather than single words mapped as objects, was examined following the methodology of Westerman & Cribbin (2000b) – full details of how the environment was created are given in section 3.2.1. As previously explained, the authors found that spatial-semantic mapping of documents was successfully utilised to improve performance when users were given a specific search task. In addition their results showed that a two-dimensional mapping solution facilitated better performance than a three-dimensional solution even when the 2D solution only accounted for 50 to 70% of the semantic variance in the 3D solution. The intention of the current study was to determine whether the effects observed during a specific search task generalise to tasks that involve browsing for documents relevant to a general query. This is likely to be more reflective of typical IR tasks in which users have limited time to browse for an unknown number of documents to satisfy a general information need. In this study participants were required to locate as many news articles related to 'risks taken by journalists' as

possible. It was hypothesised that the quality of spatial-semantic mapping would positively influence performance, and the number of dimensions used for organising the visualisation of the database would impact on this.

To gain a deeper understanding of the cognitive requirements of processing and retrieving semantic information, individual differences in cognitive ability were considered as factors potentially impacting on the usability of SDMS. It was hypothesised that individual differences in cognitive ability would be reflected in retrieval performance and that high cognitive ability would be associated with increased performance.

3.2 Methodology

3.2.1 Creating the VE Experimental Platform

A database of 100 documents taken from the TREC 7 database (see Voorhees & Harman,1999) was compiled, comprising newspaper articles from the LA Times. Of these 100 documents 20 were considered by TREC to be relevant to 'risks taken by journalists' (JR). The other 80 were randomly related to each other but considered irrelevant to JR.

The documents were analysed using the 'n-gram' method of ATA as detailed and tested in Chapter Two. A 5-gram analysis was used as this was the optimal n-gram length for discriminating the 'Piracy' document set (see Chapter Two) and compares to n-gram lengths generally used in existing research. To date 5-gram appears to be the longest n-gram used in an information retrieval system e.g. (Cavnar, 1993;

Damashek, 1995a; and Soboroff, et al.,1997).

In order to map the document set into a virtual information space, a vector space model (VSM) was employed. First, a list of unique strings five characters long was compiled for all 100 documents, along with the number of times each string occurred in each document. The cosine value of the angle between all possible pairs of documents, based on the number of co-occurrences of unique 5-grams between documents, was then calculated. A similarity matrix based on the inter-document cosines was generated and converted to a single vector, and a multi-dimensional scaling (MDS) solution calculated (Stress = .286; r^2 = .485) to map the documents initially into a three-dimensional VE. The goodness of fit between the original dissimilarities and the MDS derived distances is represented by the stress value (calculated using Kruskal's stress formula 1); ideally a value below 0.1 would be required for a good MDS solution. However, despite the high stress value and relatively low r^2 value this was the best solution available and was therefore taken to be the 100% condition (i.e. 3D100), accounting for 100% of the semantic variance notwithstanding random error. It should be noted that due to the very high dimensional nature of the raw similarity measures previously discussed, it was expected that reduction to a three dimensional space would only account for a small degree of the original variance and as such 48.5% was considered a relatively good solution.

Adding random noise to the co-ordinates produced by MDS for 3D100 generated two further three-dimensional VEs – 3D80 and 3D60. This resulted in the environments varying in the quality of the semantic mapping, when correlated with 3D100; 3D80

resulted in $r^2 = .77$, and 3D60 produced an $r^2 = .56$, thus accounting for approximately 80% and 60% respectively of the semantic variance accounted for by 3D100. The 2D60 environment was created by removing one of the three co-ordinates calculated for the 3D100 solution. This produced a 2D mapping solution that accounted for approximately 60% of the variance accounted for in 3D100 ($r^2 = .56$), and matched the variance accounted for by 3D60. In terms of semantic variance accounted for by each of the environments with respect to the raw similarity data (prior to MDS), 3D100 = 48.5%, 3D80 = 37.35%, 3D60 27.16%, and 2D60 = 27.16%. Given the nature of the original data and the way in which people can identify an underlying generic level of semantic similarity identified by ATA (see Chapter 2 – Section 2.6), it was considered that despite the low variance accounted for, the mapping in the environments reflected sufficient semantic information for people to identify and use.

It was decided that creating a 2D environment from a new MDS solution could qualitatively alter the way in which semantic variance was accounted for within the mapping. To alleviate this possibility and to ensure that the removal of a single coordinate did not also affect the quality of the mapping, three separate 2D60 environments were created in which the x, y, or z coordinate from the 3D100 environment was removed. These three solutions were used in a counterbalanced sequence by allocating participants randomly to conditions. Data from all three 2D environments were then pooled for analysis.

3.2.2 Presenting, Manipulating, and Recording Events from the Environments

In all environments each document was represented as a green spherical node, resulting in a visual screen image of black space with 100 green planets (see Figure

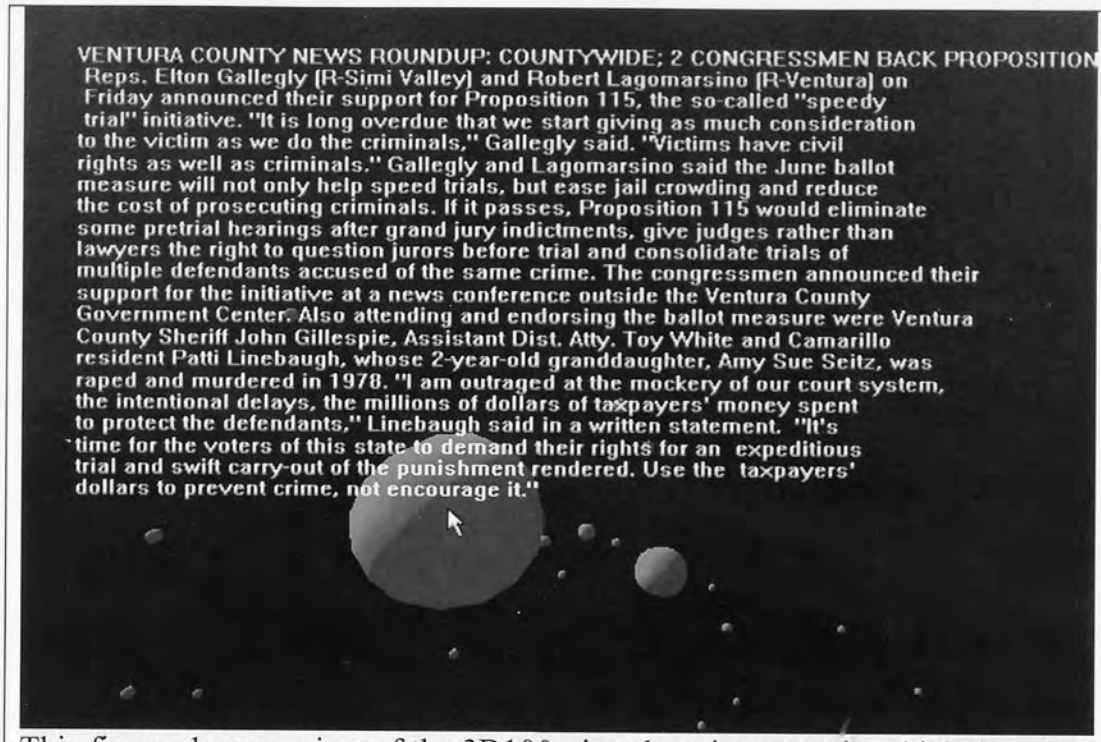
3-1, Figure 3-2, and Figure 3-3, for examples screen shots of the 3D100 environment, and Figures Figure 3-4, and Figure 3-5 for examples of 3D80 and 3D60 environments). Unfortunately the images shown in the figures below do not clearly demonstrate the dynamic nature of the virtual environment or the qualitative differences between the individual 3D environments. The semantic mapping within the environments could not be easily identified visually until the user began travelling through the environment and generating a cognitive map of the spatial-semantic relationship of documents visited based on content and visual patterns of node clusters.

The user navigated through the environment using the mouse. A full range of movements imitating travel through a real environment was possible. These included, moving forwards, backwards, up, down, left and right. The user could also rotate or tilt their position within the environment in order to simulate turning (i.e. panning) left or right, or looking up or down (i.e. tilting their position to give the effect of tilting one's head up or down). The environment could be rotated clockwise or anti-clockwise such that the user's perspective of position did not alter, but nodes within the field of view rotated so that those at the bottom of screen moved to the top and vice versa. In order to achieve these movements, users had to point the cursor in the direction they wished to travel while depressing either or both of the mouse buttons dependent on the type of movement required. For example, to step or glide left, right, up or down, the cursor was pointed in the required direction and the right hand mouse button depressed; to turn or pan left or right, or move forwards (cursor pointed to top of screen) or backwards (cursor pointed to bottom of screen), the left hand button was depressed. To tilt the environment up or down, or rotate it left or right both buttons

had to be depressed simultaneously. Movement was in simulated 'real' space meaning that travel occurred in the exact direction the cursor was pointed; when the cursor was not placed on the exact central X or Y axis, a combined directional movement occurred. Speed of movement was moderated by distance from the centre of screen; speed increased as the distance of the cursor from the centre of the screen increased.

When the cursor was held over a node that was in range, the document it represented appeared on screen (see Figure 3-1). If the node was "out of range" however a message "too far away" was displayed (see Figure 3-2). This was done to control levels of navigation and to ensure all users fully navigated the environment rather than adopting a distant position in which all or most nodes were present on screen and randomly passing the cursor over all nodes in view. While this may arguably be a legitimate strategy for people to use in this particular environment, due to the limited size of the database, it would not prompt users to make use of the spatial-semantic mapping used in the environments. As the focus of the experiment was the how people use spatial-semantic mapping, this had to be controlled.

When a document was considered relevant, users 'selected' it by pressing the space bar or enter key while the document appeared on screen. The node could be 'de-selected' using the same procedure if it had been selected in error, or the participant later considered it irrelevant.

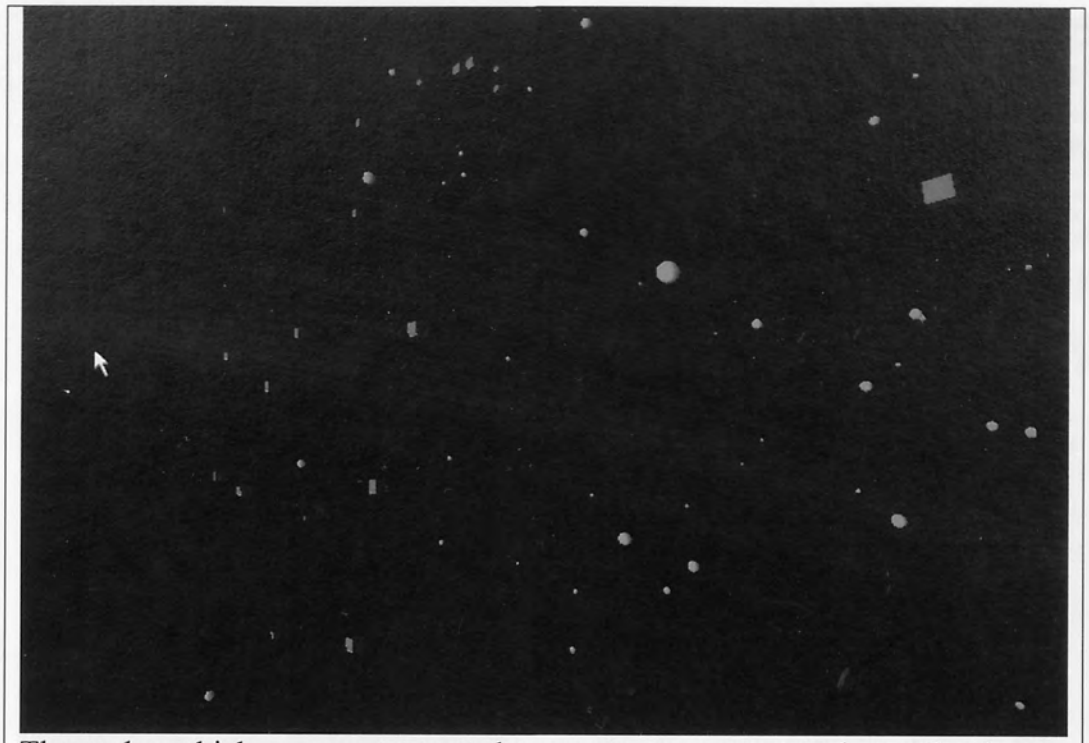


This figure shows a view of the 3D100 virtual environment in which documents are optimally mapped on the basis of semantic content in three dimensions. The green nodes represent documents held in the database when the cursor is held over a node that is within range its representative document appears on screen.

Figure 3-1 3D100 environment

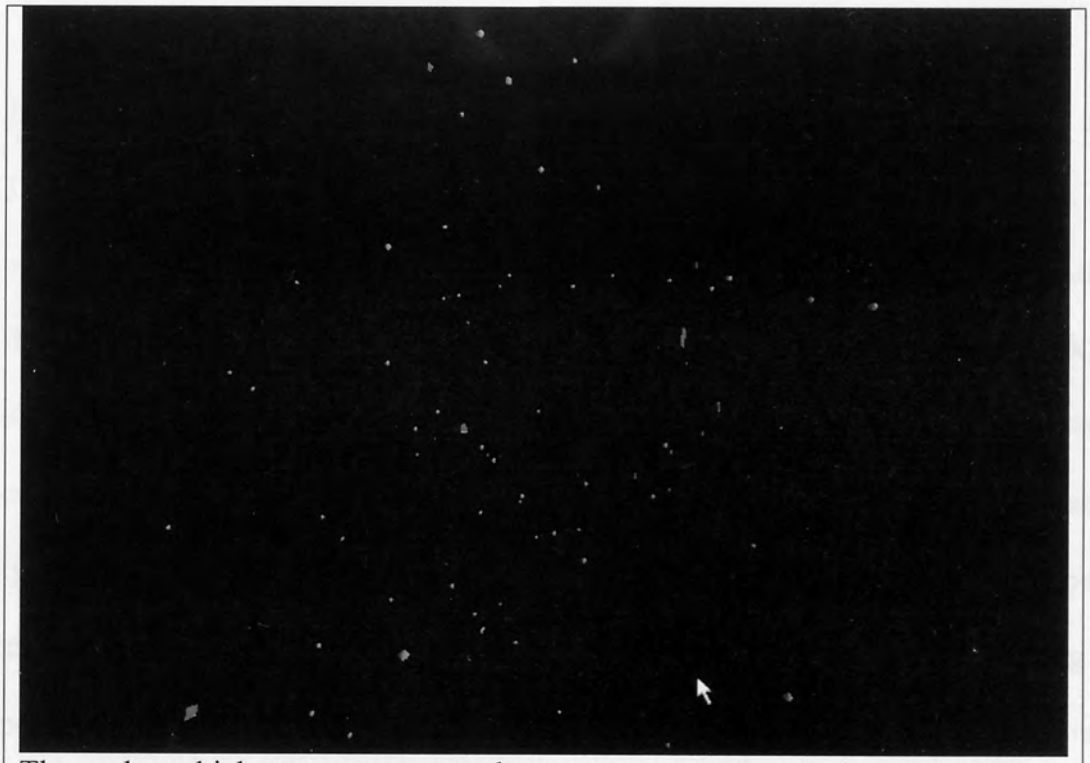
This figure shows a view of the screen message that appears when the cursor is pointed at an out of range node the representative document only appears on screen when the node is within range.

Figure 3-2 3D100 environment showing out of range message



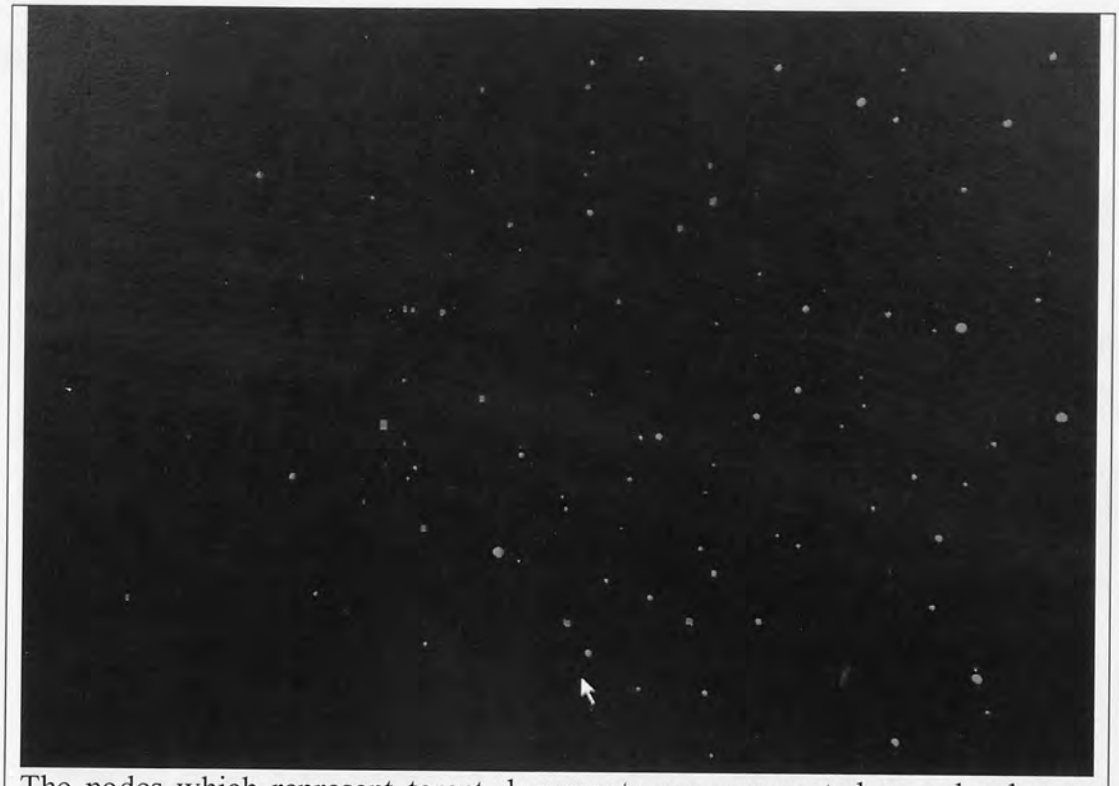
The nodes which represent target documents are represented as red cubes to give the reader an indication of the spatial-semantic mapping of documents; participants were not given this detail.

Figure 3-3 3D100 environment showing target documents



The nodes which represent target documents are represented as red cubes to give the reader an indication of the spatial-semantic mapping of documents; participants were not given this detail.

Figure 3-4 3D80 environment showing target documents

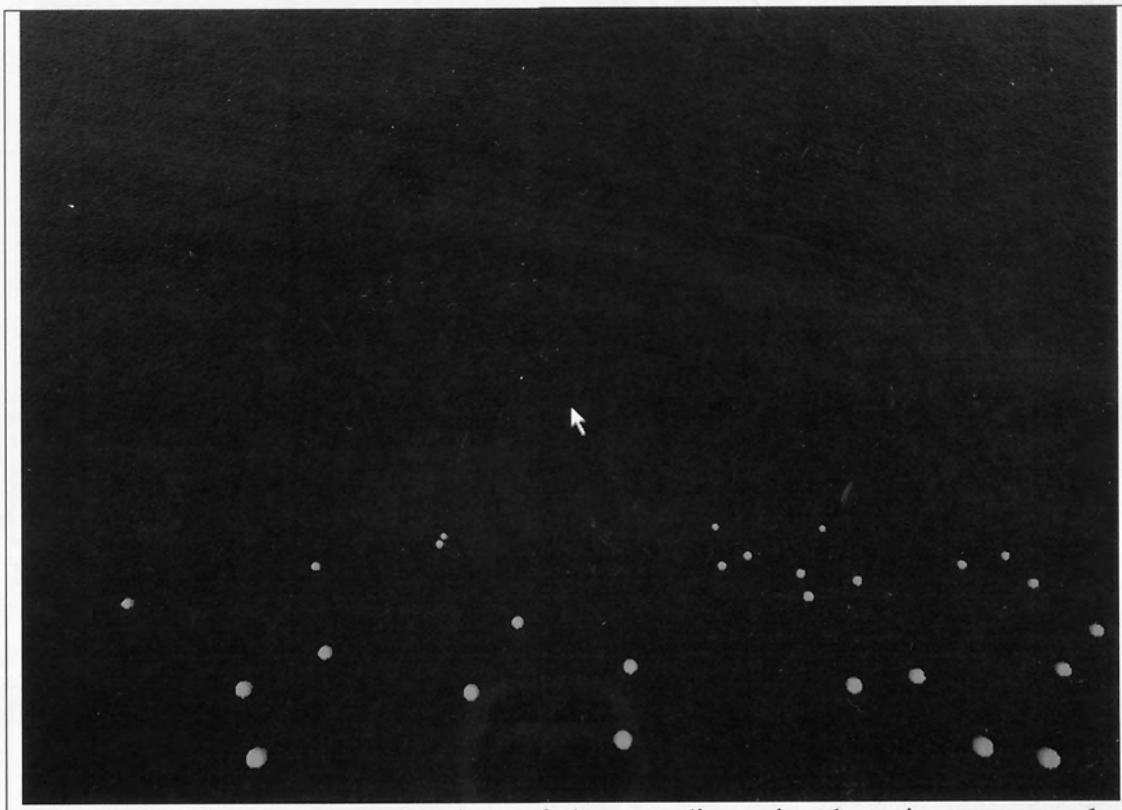


The nodes which represent target documents are represented as red cubes to give the reader an indication of the spatial-semantic mapping of documents; participants were not given this detail.

Figure 3-5 3D60 environment showing target

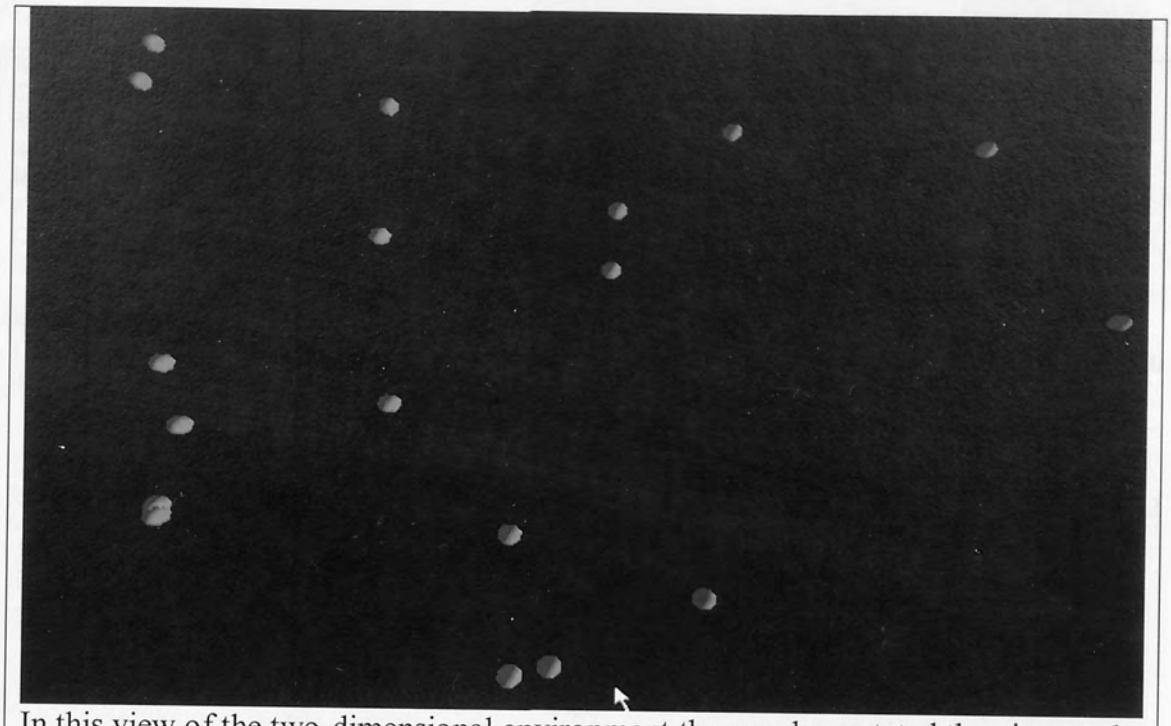
The 2D environment was presented, navigated and controlled in the same way as the 3D environments. However as the mapping was two-dimensional the nodes were presented in a two-dimensional plane (see Figure 3-6, Figure 3-7, and Figure 3-8). As the task began users were presented with the environment on a horizontal (X / Z) plane (see Figure 3-6). The environment was presented using this perspective to enhance the perception of a three-dimensional virtual environment and to replicate the type of presentation used in similar information visualisation systems (e.g. Chen, et al., 1999). Users could rotate the perspective to a vertical (X / Y) plane by tilting their position (using the controls as previously explained) while navigating the environment (see Figure 3-7). Unfortunately, it was not possible to analyse the degree to which users chose to navigate in the vertical perspective due to the pattern of data recorded

during task completion – addressing this limitation should be considered in future studies.



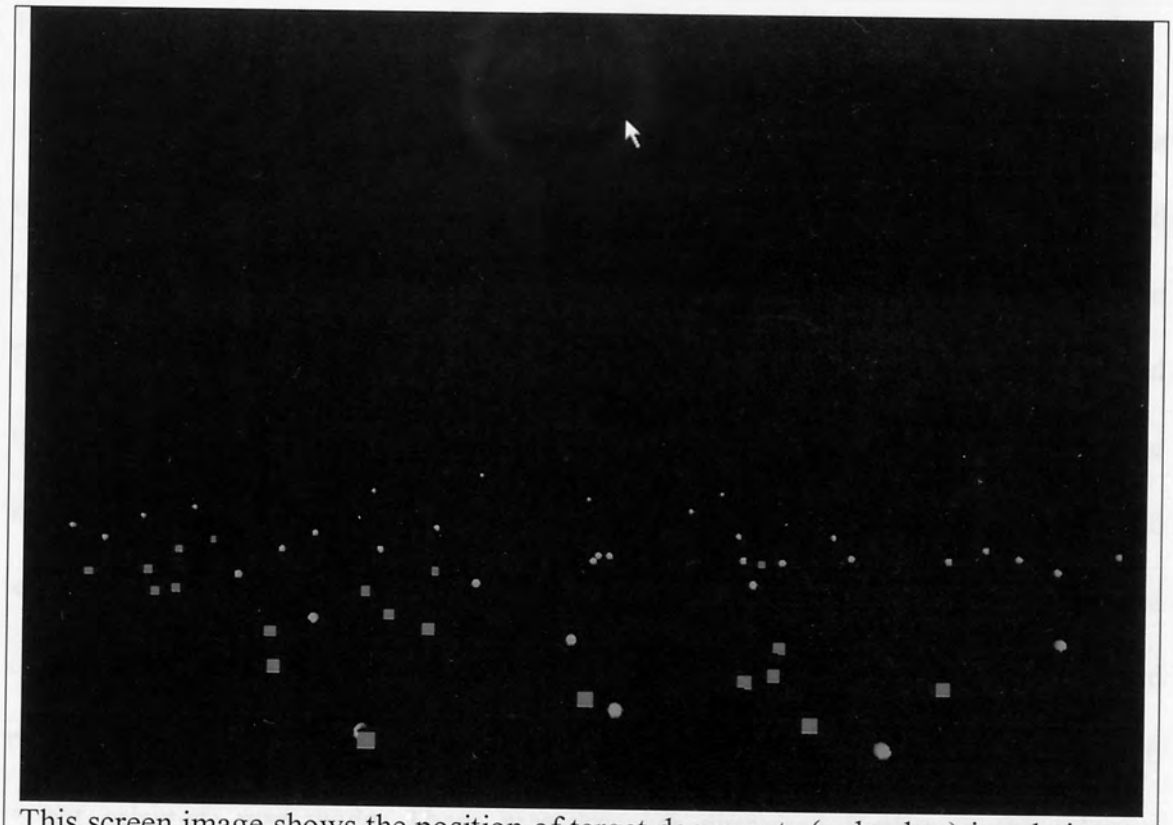
This figure shows the user's view of the two-dimensional environment at the starting position. The participant has a slightly elevated position and the nodes are presented in the X / Z plane.

Figure 3-6 2D60 environment presented in the X / Z plane



In this view of the two-dimensional environment the user has rotated the view to the X / Y plane and is no longer viewing from above.

Figure 3-7 2D60 environment rotated to X / Y plane.



This screen image shows the position of target documents (red cubes) in relation to each other, within the two-dimensional virtual environment. Users did not receive these cues.

Figure 3-8 2D60 environment showing target documents

```
Subject = ? : Sex = 1
Object file name = br 3d60inv
Space = 18
Speed = 5
Reach = 1080
Final time        = 3544.840088
Distance travelled = 151279.890625
Rotation          = 16932.000000

Selected nodes:   [Documents 1 – 100 where 1 = node selected, 0 = node not selected]
1 0 1 1 1 1 1 1 1 1 0 1 0 0 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Times each node was visited:   [Documents 1 – 100 where numerator = number of times node visited]
2 1 1 2 1 2 2 1 1 1 4 2 0 1 1 2 1 4 1 1 2 0 1 1 1 6 1 3 2 1 2 2 0 2 1 2 4 1 2 4 4 2 2 4 4 1 3 1 2 2 1 5 1 4 1 3 1 3 8 2 0 2
6 4 2 2 1 1 1 1 2 2 2 1 6 4 4 1 2 3 7 2 1 0 0 2 2 5 5 2 4 1 3 4 2 0 1 0 1 1

Event   Node   Time    Travel    Rotation   Select
----------------------------------------------------------
 0       6    3.410     0.000     0.000  2 Read
 1       6   28.670    31.799     0.000  1 Select
 2       8   33.950   416.734    18.000  2 Read
 3       8   39.490   416.734    18.000  1 Select
 4      80   46.580  1869.096    34.000  2 Read
 5      11   66.850  2200.752    39.000  2 Read
 6      11   71.410  2200.752    39.000  1 Select
 7       9   75.470  2649.354    57.000  2 Read
 :       :      :        :          :       :
 :       :      :        :          :       :
```

Figure 3-9 Example of record of events produced for each participant

An example of the data recorded during task completion is shown in Figure 3-9. For each event (document viewed, selected, or de-selected) the system was programmed to record an event by event record, as well as general task data. As can be seen in Figure 3-9 the general task data included details about the condition and environmental settings (internal programmed settings regarding the scale of the environment) consisting; i) distance between nodes (space), ii) proximity required to view a document (range), iii) the speed of movement (speed)), iv) the total time taken on task, v) the total distance travelled, and vi) total rotation. For each node a record regarding how many times it was visited, and whether it was selected (1 = selected, 0 = not selected) was made. These measures are discussed further in the relevant 'measures' sections of the three experimental chapters in this thesis, it should be noted however that movement generated by activation of the mouse buttons was required in

order to record a measure of distance or rotation. Moving the cursor between nodes within the field of view (FOV) without altering it did not register a measure of distance or rotation.

The event by event data was recorded cumulatively, showing the amount of time lapsed, the distance travelled, and the amount the environment was rotated, between the start of the experiment and the specific event. No data were recorded regarding direction of travel, which with the benefit of hindsight, limited the degree to which patterns of navigation and individual search strategies could be analysed.

3.2.3 Participants

Initially 84 participants completed the experiment, however data for two people were excluded – see section 3.3.1 Data Screening. Of the 82 participants whose data were retained 62 were female and 20 were male. Participants were aged between 18 and 41 years, (mean 19.6, SD 3.6).

3.2.4 Experimental Design

The analyses employed an independent measures experimental design. One-way ANOVAs were used to examine the effect of spatial-semantic mapping on performance in terms of i) the quality of the semantic mapping (three levels 3D100, 3D80, & 3D60), and ii) the number of dimensions (two levels 3D60 & 2D60), used for desktop presentation of the VE database (these results are presented in section 3.3.2). As described in section 3.2.1 the maximum semantic variance accounted for by a two-dimensional solution when compared to the maximum three-dimensional solution was 60%, therefore the 60% 3D solution (3D60) was used for dimensional

comparisons.

Factorial analyses based on a median split of test scores were used to examine the effects of individual differences in cognitive ability (associative memory (MA), spatial working memory (MV), and spatial visualisation ability (VZ)), together with any interactions between cognitive ability and environmental mapping on performance. Correlation and regression techniques were also used to identify which cognitive abilities had the most predictive power. For these however, the interactive effects of spatial-semantic mapping could not be analysed. Details of the psychometric tests used are given in Section 3.2.6 Materials.

3.2.5 Measures of Performance

Performance was measured across two parameters; how accurately participants achieved the objective of the task set, and how quickly they completed the task (see sections 3.2.5.1 and 3.2.5.2 respectively). In addition these two parameters were combined into an overall measure of efficiency (E) which was calculated using speed / accuracy trade-off i.e. $\frac{accuracy(A)}{timeontask(TT)}$ (Davies & Parasuramen, 1982).

3.2.5.1 Accuracy

In terms of accuracy three measures were initially selected, these were: -

(P) precision (the number of relevant documents selected divided by the total number of documents selected)

(R) recall (the number of retrieved relevant documents divided by the total number of relevant documents within the database),

(A) accuracy (a combined measure of precision and recall i.e. the F stat (F) that

represents the harmonic mean between recall and precision – $F = \frac{2}{\frac{1}{R} + \frac{1}{P}}$ where R = recall and P = precision).

During analysis (see section 3.3) precision demonstrated a strong ceiling effect with a mean value of .94, SD .13, and kurtosis 7.84 - the maximum possible value for precision was 1. It was decided not to retain precision as a dependent variable since analysis was not likely to produce useful or reliable results. However, the measure was included for the calculation of the F stat.

There has been much debate and controversy within the IR community regarding the appropriate measures to use for performance evaluation, particularly in terms of recall and precision. This debate began during the late 1960s at the time of the Cranfield experiments (Cleverdon, 1972) and continues today (e.g.van Rijsbergen, 1979; Dunlop, Johnson, & Reid, 1998; and Voorhees, 2000).

There are two major criticisms of precision and recall as measures of retrieval effectiveness and information seeking, particularly in hypertext systems and interactive electronic IR interfaces. Firstly there is the issue of 'relevance'; precision and recall are both ratio measures of relevant to non-relevant retrieved documents or to total relevant documents available but not retrieved. It is argued that relevance judgements are not only subjective, i.e. based on individual differences in understanding, experience, and motivation, but are also dynamically influenced by the current information need, the nature of the task, and the IR system used to present or retrieve the information (Schamber, et al., 1990). Not only has it been shown that

people differ with others in judgements of document relevance to a particular query, but it is also the case that individuals differ on their own judgements dependent on a variety of factors (e.g. Harter, 1996; Mizzaro,1997; and Mizzaro, 1998). Brooks (1998) stated that in order to calculate precision and recall 'relevance' should be a binary decision where a document is either relevant or not relevant. However Salampasis, Tait, & Bloor (1998) suggest that in many instances this is too simplistic because in reality, particularly when individuals are browsing large quantities of documents in electronic information environments, they are most likely to rank located documents in order of how relevant they are to their current need. In this situation 'relevance' is an ordinal measure not a nominal one, and also dependent on individuals' perception of information need.

The second problem related to precision and recall as measures of performance, arises from the implicit nature of the measurements. Maximum recall can be achieved by simply retrieving all documents in the set; this would mean however that precision would be at a minimum[1]. If however only one document was retrieved, and it was relevant, precision would yield a maximum value but recall would be at a minimum level. This inverse relationship between the two measures, e.g. if recall increases precision is likely to decrease, was first identified by Cleverdon (1972), and it has since been likened to the issues faced in attention and perceptual experiments (Dumais, 2003). In IR experiments such as those run by TREC, the performance of the systems, rather than the users, is generally the focal issue. These types of

[1] The minimum level for precision is greater than zero and dependent on the ratio of relevant documents to non-relevant to begin with. For instance, in the current data set there are 100 documents of which 20 are relevant, if the user retrieved all 100 documents they would have a recall level of 1 and a precision value of 0.05.

experiments involve large databases and an unknown number of relevant documents. In these circumstances precision-recall curves, based on the patterns of document recall across competing systems, are used to demonstrate performance in much the same way as the ROC curves are used in perception, attention, and memory research (Dumais, 2003; and Craig, et al., 1987)

The problematic nature of these issues can be addressed if the nature of the research is weighted in favour of one or other measure e.g. if the precision of the system or user is the primary focus of the research. In the current study however, it can be argued that for individuals looking for answers to a particular query, a balance between the two measures is ideal in terms of performance. A variety of averaging techniques have been developed to find this balance between precision and recall (van Rijsbergen, 1979) and in the current research the F stat, which measures the harmonic mean between precision and recall has been used as an overall measure of accuracy (see Baeza-Yates & Ribeiro-Neto, 1999). It should be noted that while using the F stat has addressed the problem of the inverse relationship between recall and precision, issues related to the variant nature of relevance judgements remain un-addressed within this work. However the task is fairly specific in terms of the documents that would qualify as relevant and consideration is given to differences in relevance judgements when discussing the results.

3.2.5.2 Time on Task

Two measures of time were examined, time on task (TT) and mean time on task (MnTT), which represents the average time taken per relevant document selected (i.e. TT/number of relevant documents selected).

3.2.6 Materials

Four cognitive ability tests were used; three tests were taken from the "Kit of Factor-Referenced Cognitive Tests" (Ekstrom, et al., 1976). These were: -

VZ-2 – Spatial visualisation ability (VZ).

There are two parts to this test each comprising 10 items and each part is subject to a time limit of three minutes.

MA-2 – Associative memory (MA).

There are two parts to this test each comprising 15 items (object name paired with a number). For each part participants are given three minutes to memorise the paired lists. They are then shown a re-ordered list of the object names for two minutes during which time they have to recall the associated numbers.

MV-2 – Spatial working memory referenced within this thesis as MV.

There are two parts to this test each comprising 12 items. For each part participants are given four minutes to memorise a map and four minutes to answer the 12 questions.

All of the above measures were negatively scored, using a standardised procedure recommended by the authors, to reduce effects of guessing. Split-half reliability measures comparing scores from the two parts of each test were calculated.

The fourth cognitive ability test administered was: -

SCOLP – Speed and Capacity of Language-Processing Test (Baddeley, Emslie, & Smith, 1992). This test comprises 100 items to be completed within a two minute time-limit. This measure was used to ensure all participants were matched for

comprehension ability in terms of speed, not for use as an experimental variable.

WorldToolKit – Release 7 (Sense8 Corporation, 1997) was used to generate the virtual environment database, which was presented using desktop PCs.

3.2.7 Procedure

On entering the experiment room participants were welcomed and the purpose of the experiment was explained, along with ethical guidelines and data protection regulations. When participants were satisfied with the explanations and all queries had been addressed, consent forms were completed. The participants were then asked to complete the four timed, paper based, cognitive ability tests (see section 3.2.6 Materials).

Participants were asked to complete a practice task order to familiarise themselves with navigation and document selection controls. In this task they had to locate and select a target (a single red coloured node presented in a three-dimensional environment consisting of green nodes), as quickly as possible. Instructions on manoeuvring the environmental platform were given verbally as the experimenter demonstrated and allowed the participant to practice. The task comprised 10 blocks of 10 target presentations. If participants successfully reached the pre-programmed required response time for two consecutive blocks they proceeded to the actual experiment. All participants successfully completed the practice trials.

For the main task participants were semi-randomly allocated to one of the experimental conditions (3D100, 3D80, 3D60, 2D60), matching conditions for

cognitive ability. Participants were required to locate and select as many documents (newspaper articles) related to journalists facing personal risk, as they could find and deemed relevant. Participants were told to stop when they had found all the documents they thought they were going to find, however they were required to spend a minimum of 20 minutes browsing and to not exceed 45 minutes on task (participants were told when they reached these time limits). Once individuals were certain they understood the task they began. Participants were told they were free to ask for help during the experiment if they needed it.

At the end of the experiment participants were thanked for their cooperation and reminded they could request further information regarding their data or the experiment in general at any time.

3.3 Results

3.3.1 Data Screening

Initially 84 participants completed the experimental procedure; the final sample size however was 82 as data from two participants were removed prior to analyses. One participant (due to error on the part of the experimenter) exceeded the designated time limit on the task. A second participant was removed from all analyses, due to registering extreme values for 'distance travelled' and 'rotation'. (While these are not variables analysed in the current chapter, they are examined in Chapter Four.)

On closer inspection of the individual's task record it was found they had become lost in the environment such that all nodes had disappeared from screen. Although some other participants had the same experience, this particular participant did not request

help and remained lost for 55.5% of the total time on task (12.84 minutes). During this period of disorientation, 76.9% of the total distance was travelled and rotation increased by 95.4%. It was felt this was not a true representation of performance or behaviour on the task, as the participants were told to let the experimenter know if they had a problem.

During initial analysis it was discovered that the start position of the 3D60 environment was immediately adjacent to a relevant node. In order to account for this, time on task was taken as time from first node selected to last node selected. Time between first and last documents selected was considered the most reliable measure of 'time on task' as it ensured any observed differences reflected effects of browse performance and environmental design, rather than an individual's motivation to continue searching until all documents had been viewed, or until the end of the experiment.

The results in this section are divided into two principal sections a) the effects of environmental mapping on performance (section 3.3.2), and b) the effects of cognitive ability on performance together with interactions between the two (section 3.3.3).

3.3.2 Environmental Mapping as a Determinant of Performance

Two aspects of the way the spatial-semantic mapping and VE presentation affect performance, are examined. Section 3.3.2.1 presents the results comparing differences in performance arising from differences in the quality of spatial-semantic mapping in three-dimensional VE representations, and Section 3.3.2.2 gives the results of analyses comparing the effects of presenting data in two versus three-dimensions on

performance.

3.3.2.1 Quality of Mapping

Conditions 3D100, 3D80 & 3D60 were compared, to examine the effect of quality of mapping, i.e. the amount of semantic variance accounted for within the spatial arrangement of documents. Table 3-1 shows means and standard deviations of all measures of performance detailed in section 3.2.5. Figure 3-10 shows bar graphs of means for R, TT, MnTT, A, and E.

Performance Measure	Condition	N	Mean	SD
Recall (R) - (proportion of retrieved relevant documents to total relevant documents)*	3D100	21	.6619	.2018
	3D80	20	.5225	.1943
	3D60	20	.5425	.1633
Time on Task (TT) - in seconds*	3D100	21	1055.35	256.27
	3D80	20	1074.48	414.86
	3D60	20	1310.29	352.91
Average Time on Task (MnTT) – average time per relevant document selected in seconds	3D100	21	93.68	69.18
	3D80	20	108.99	49.92
	3D60	20	132.65	53.25
Accuracy (A) – F stat	3D100	21	.7546	.1783
	3D80	20	.6504	.1582
	3D60	20	.6670	.1479
Efficiency (E) – (A) divided by (TT)*	3D100	21	.00057	.00015
	3D80	20	.00048	.00012
	3D60	20	.00046	.00013

* ANOVA shows significant main effect of Condition $p < 0.05$

Table 3-1 Descriptive Statistics for Performance in 3D100, 3D80, 3D60 Environments

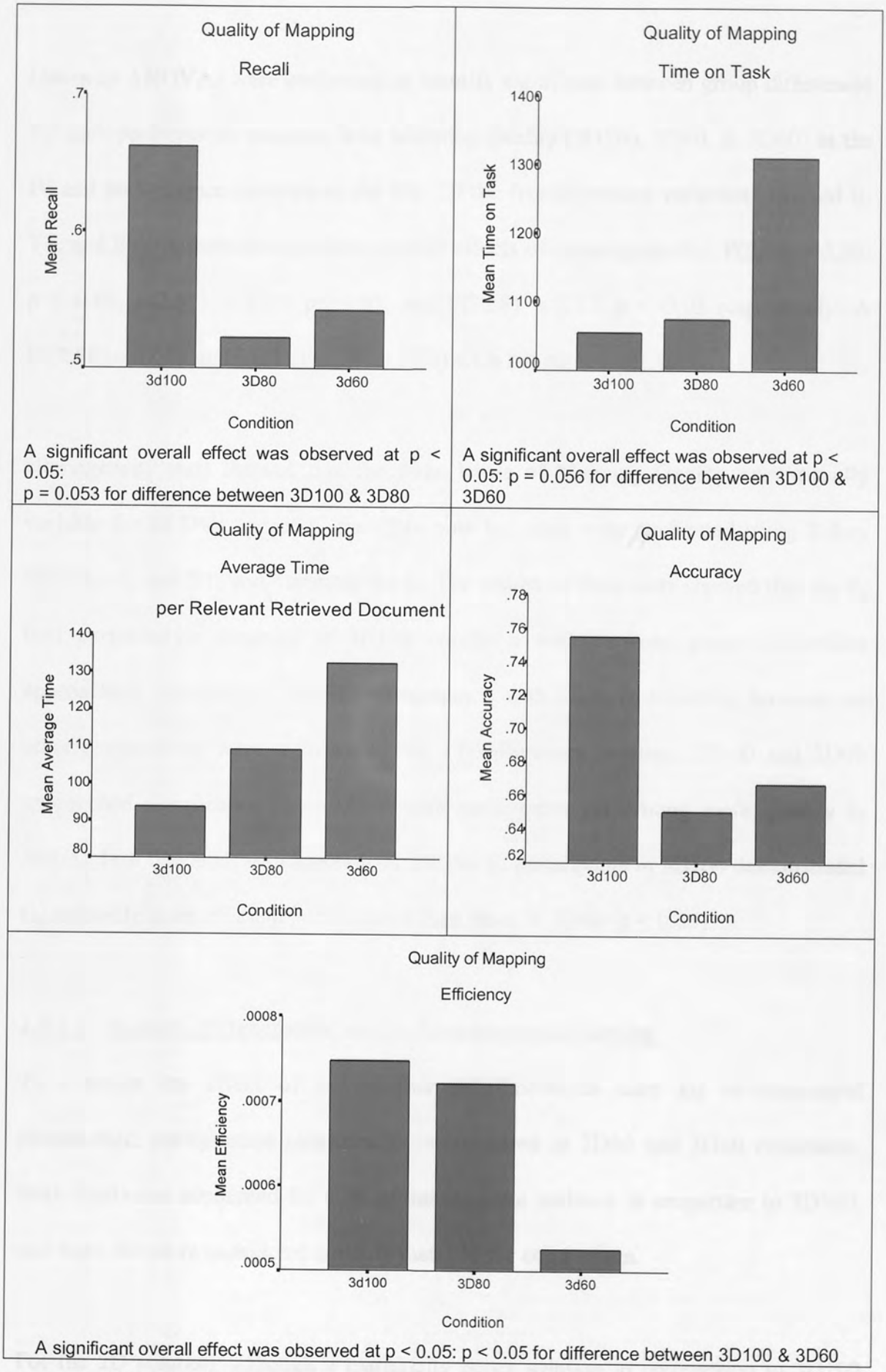


Figure 3-10 Graphs of Means for the Effect of Quality of Mapping on Performance

One-way ANOVAs were performed to identify significant between group differences for each performance measure, with Mapping Quality (3D100, 3D80, & 3D60) as the IV and performance measure as the DV. Of the five dependent variables analysed R, TT, and E demonstrated significant overall effects of mapping quality, $F(2,58) = 3.34$; $p < 0.05$, $F(2,58) = 3.34$; $p < 0.05$, and $F(2,58) = 3.97$; $p < 0.05$ respectively. A ($F(2,58) = 2.46$), and MnT ($F(2,58) = 2.32$) were not significant.

Homogeneity tests showed that the three levels of Mapping Quality were equally variable for all DVs except E, therefore post hoc tests were performed using Tukey HSD for R, and TT, and Tamhane for E. The results of these tests showed that for R, best performance occurred in 3D100 condition with between group differences approaching significance for 3D100 compared with 3D80 ($p = 0.053$), however no other comparisons were significant. For TT differences between 3D100 and 3D60 approached significance ($p = 0.056$) with participants performing more quickly in 3D100. Post hoc analyses also showed that for E, participants in 3D100 demonstrated significantly more efficient performance than those in 3D60 ($p < 0.05$).

3.3.2.2 Number of Dimensions used in Environmental Mapping

To examine the effect of the number of dimensions used for environmental presentation, performance measures were compared in 2D60 and 3D60 conditions. Both conditions accounted for 60% of the semantic variance in proportion to 3D100, and were therefore considered suitably matched for comparison.

For the 2D solution, although a marginally better solution in comparison to 3D100

could be obtained (approximately 70%), a solution accounting for the same amount of variance as 3D100 was not possible. Therefore comparisons between 2D60 and 3D100 were conducted to determine the benefits of using 3D to obtain a better mapping solution. Westerman & Cribbin (2000b) found that the additional semantic information available in their 3D100 solution did not compensate sufficiently for the extra cognitive effort required to navigate 3 rather than two dimensions, and that performance in 2D60 was not significantly poorer than performance in 3D100. Table 3-2 shows means and standard deviations for each performance measure. Figure 3-11 shows means for R, TT, MnTT, A, and E.

Performance Measure	Condition	N	Mean	SD
Recall (R) - (proportion of retrieved relevant documents to total relevant documents)	3D100	21	.6619	.2018
	3D60	20	.5425	.1633
	2D60	21	.5571	.2117
Time on Task (TT) - in seconds	3D100	21	1055.35	256.27
	3D60	20	1310.29	352.91
	2D60	21	1196.18	350.91
Average Time on Task (MnTT) – average time Per relevant document selected in seconds	3D100	21	93.68	69.18
	3D60	20	132.65	53.25
	2D60	21	116.46	36.32
Accuracy (A) – F stat	3D100	21	.7546	.1783
	3D60	20	.6670	.1479
	2D60	21	.6796	.1866
Efficiency (E) – (A) divided by (TT)*	3D100	21	.00075	.00024
	3D60	21	.00046	.00013
	2D60	20	.00045	.00013

* t –test shows significant difference between 3D100 and 2D60 at $p < 0.05$

Table 3-2 Descriptive Statistics of Performance in 2D60, 3D60, & 3D100 Environments

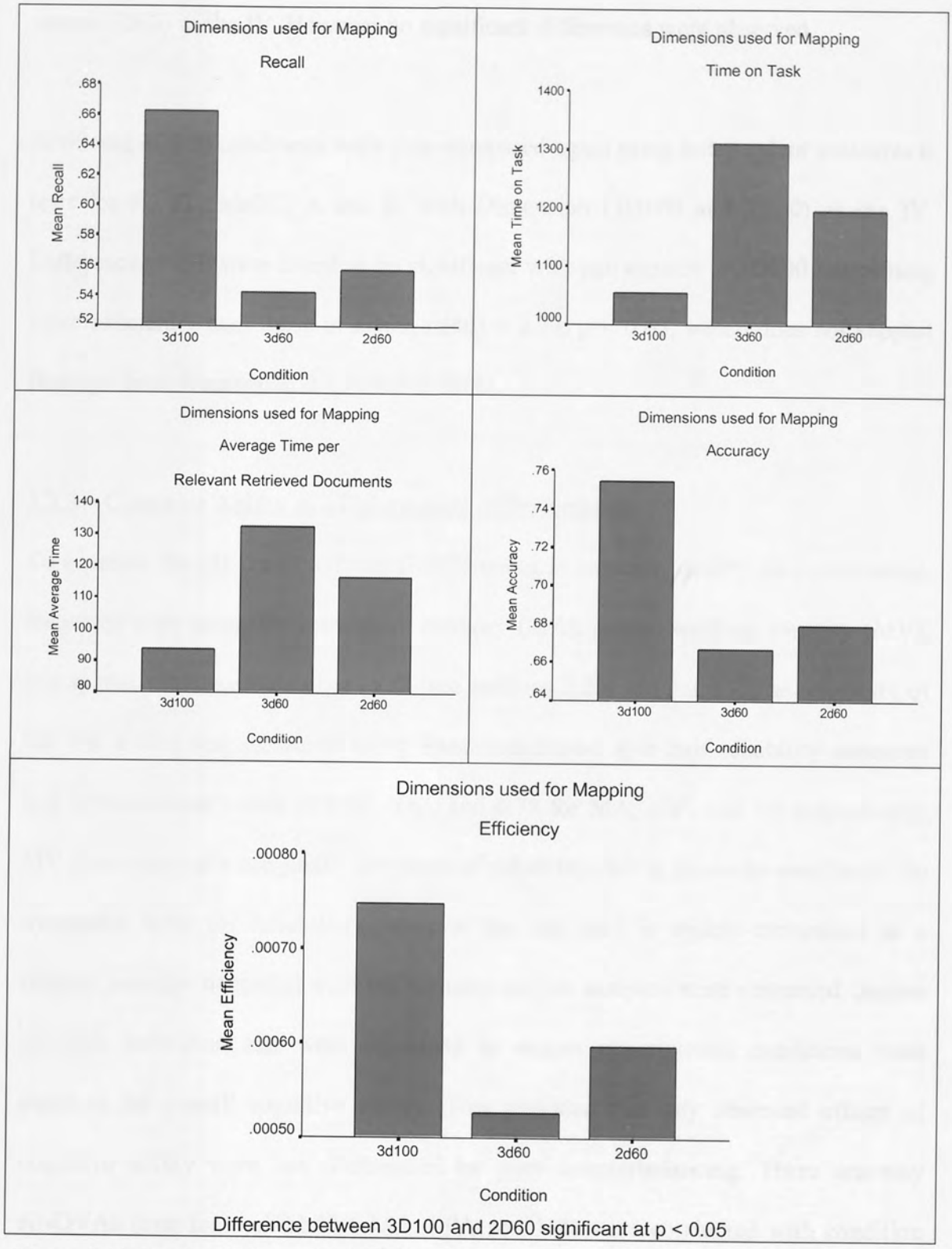


Figure 3-11 Graphs of Means for the Effect of Dimensions used for Environmental Mapping on Performance

Independent measures t-tests were performed to determine significant differences between conditions for each of the performance measures, with Dimension (3D60,

versus 2D60) as the IV. However no significant differences were observed.

2D60 and 3D100 conditions were then compared again using independent measures t-tests for R, TT, MnTT, A and E, with Dimension (3D100 and 2D60) as the IV. Differences for E were found to be significant with participants in 3D100 performing more efficiently than those in 2D60, $t(40) = 2.23$; $p < 0.05$, which does not support findings from Westerman & Cribbin (2000b).

3.3.3 Cognitive Ability as a Determinant of Performance

To examine the effects of individual differences in cognitive ability on performance, measures were taken for associative memory (MA), spatial working memory (MV), and spatial visualisation ability (VZ) (see sections 3.2.4 and 3.2.6.). The reliability of the test scores was measured using Spearman-Brown split-half reliability measures and demonstrated values of 0.82, 0.63, and 0.78 for MA, MV, and VZ respectively. MV demonstrated a marginally low level of reliability (0.7 is generally considered the acceptable level for reliability), however the test used is widely recognised as a reliable measure of spatial working memory and so analyses were continued. Scores on each individual test were examined to ensure experimental conditions were matched for overall cognitive ability. This provided that any observed effects of cognitive ability were not confounded by poor counterbalancing. Three one-way ANOVAs (one for each ability, MA, MV, and VZ) were conducted with condition (3D100, 3D80, 3D60, and 2D60) as the IV, and total score on ability as the DV. Differences were confirmed not significant; $F(3,78) = .495$ (MA), $F(3,78) = .832$ (MV), and $F(3,78) = .737$ (VZ). Table 3-3 shows descriptive statistics for all overall, and group levels of task condition, and high or low cognitive ability.

Cognitive Ability	Condition	Group	N	Mean	SD
Associative Memory (MA)	3D100	Hi	10	24.10	3.54
		Lo	11	13.36	3.44
		All	21	18.48	6.46
	3D80	Hi	11	22.00	3.38
		Lo	9	14.33	1.23
		All	20	18.55	4.69
	3D60	Hi	8	24.00	4.34
		Lo	12	11.25	4.65
		All	20	16.35	7.78
	2D60	Hi	11	23.05	2.78
		Lo	10	11.50	4.50
		All	21	17.55	6.92
	All	Hi	40	23.21	3.45
		Lo	42	12.52	3.88
		All	82	17.74	6.50
Spatial Working Memory (MV)	3D100	Hi	13	21.32	1.70
		Lo	8	15.06	4.99
		All	21	18.94	4.49
	3D80	Hi	9	21.39	1.13
		Lo	11	13.69	3.71
		All	20	17.13	4.82
	3D60	Hi	8	22.06	1.74
		Lo	12	13.81	3.71
		All	20	17.11	5.12
	2D60	Hi	13	20.87	1.95
		Lo	8	14.56	3.63
		All	21	18.46	4.09
	All	Hi	43	21.34	1.69
		Lo	39	14.17	3.86
		All	82	17.93	4.62
Spatial Visualisation Ability (VZ)	3D100	Hi	12	14.35	1.95
		Lo	9	7.11	3.04
		All	21	11.25	4.39
	3D80	Hi	10	13.05	1.37
		Lo	10	6.85	2.68
		All	20	9.95	3.79
	3D60	Hi	9	13.83	1.69
		Lo	11	7.66	2.34
		All	20	10.44	3.74
	2D60	Hi	9	13.67	1.66
		Lo	12	7.54	2.97
		All	21	10.17	3.95
	All	Hi	40	13.76	1.70
		Lo	42	7.32	2.68
		All	82	10.46	3.94

For each ability measure, overall, marginal, and cell, Means and SDs are shown by environmental condition and hi/lo ability group (based on median split - 3.3.3.1). For each ability measure, marginal means between conditions were not significantly different. Differences between hi/lo groups in each condition were significant ($p < 0.01$).

Table 3-3 Descriptive statistics of, Associative Memory, Spatial Working Memory, and Spatial Visualisation test scores

3.3.3.1 Individual Differences in Cognitive Ability and Quality of Mapping

For the factorial analyses a median split was calculated for each of the three cognitive

abilities and participants were allocated to high or low ability groups on this basis. In the first instance 2 (Ability – Hi & Lo) x 3 (Mapping Quality – 3D100, 3D80, & 3D60) independent measures ANOVAs were performed to compare the main effects of cognitive ability on each of the performance measures detailed previously, and any interactions between cognitive ability and mapping quality. Means and standard deviations of R, TT, MnTT, A, and E for and are shown for MA, MV, and VZ in Table 3-4, Table 3-5, and Table 3-6 respectively. Only significant main effects of ability and / or significant interactions are reported in this section.

Associative memory (MA), showed no significant effects (see Table 3-4).

Performance Measure		3D100		3D80		3D60		Total	
		Lo MA	Hi MA	Lo MA	Hi MA	Lo MA	Hi MA	Lo MA	Hi MA
Recall (R) - relevant/total retrieved docs	Mean	.6864	.6350	.4389	.5909	.5625	.5125	.5703	.5845
	SD	.2481	.1435	.1193	.2212	.1611	.1727	.2063	.1842
	N	11	10	9	11	12	8	32	29
Time on task (TT) - seconds	Mean	1056.02	1054.61	957.64	1170.08	1321.97	1292.78	1128.08	1164.11
	SD	299.29	215.45	427.46	398.10	263.01	478.29	354.15	371.13
	N	11	10	9	11	12	8	32	29
Time per relevant retrieved document (MnTT) - seconds	Mean	101.50	85.08	113.69	105.15	128.92	138.25	115.21	107.36
	SD	95.58	18.08	60.41	42.18	52.51	57.48	70.76	44.93
	N	11	10	9	11	12	8	32	29
Accuracy (A) – F stat	Mean	.7516	.7580	.5970	.6941	.6814	.6455	.6818	.7027
	SD	.2300	.1089	.1275	.1729	.1303	.1784	.1763	.1565
	N	11	10	9	11	12	8	32	29
Efficiency (E) – A / TT	Mean	.00075	.00074	.00076	.00069	.00052	.00052	.00067	.00066
	SD	.00031	.00016	.00038	.00037	.00013	.00016	.0003	.00028
	N	11	10	9	11	12	8	32	29

No significant differences observed

Table 3-4 Descriptive statistics of performance for high and low Associative Memory (MA) participants in 3D100, 3D80 & 3D60 environments.

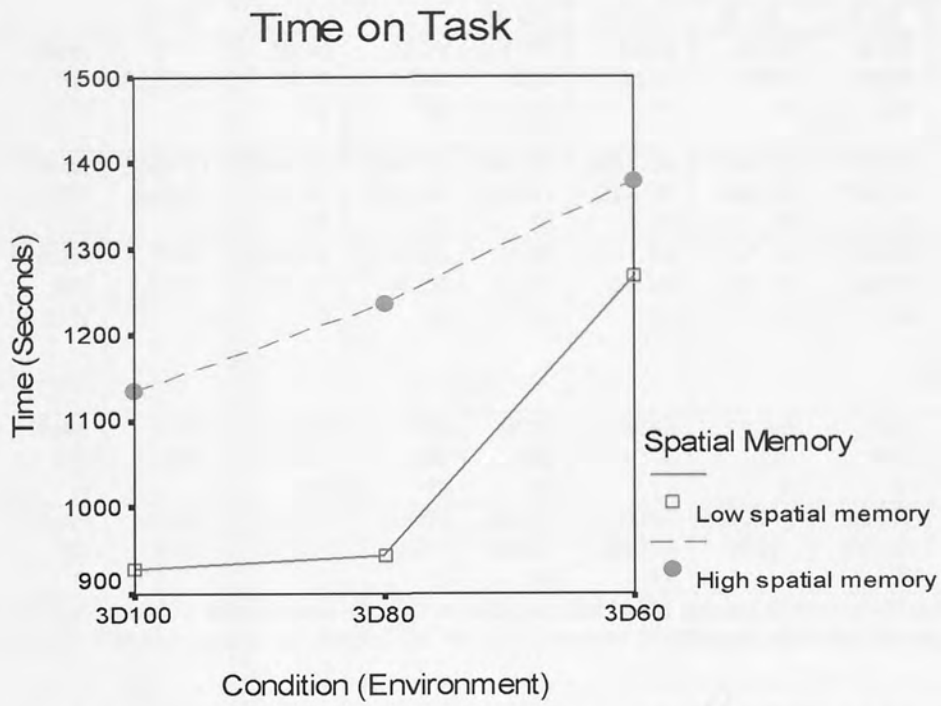
When comparing data for high and low spatial working memory (MV) groups, a significant main effect of cognitive ability was observed for TT ($F(1,55) = 5.48$; $p <$

0.05), with high MV participants taking longer on task than low MV. However post hoc independent t-tests showed that differences between groups within individual conditions (3D100, 3D80, and 3D60), were not significant (t (19) = -1.94, t (18) = -1.65, and t (18) = -0.69 respectively). See Table 3-5 for descriptive statistics and Figure 3-12 for a graph of mean values of TT.

Performance Measure		3D100		3D80		3D60		Total	
		Lo WM	Hi WM	Lo WM	Hi WM	Lo MW	Hi WM	Lo WM	Hi WM
Recall (R) - relevant/total retrieved docs	Mean	.6182	.7100	.5000	.5500	.5708	.5000	.5468	.6083
	SD	.2316	.1612	.1628	.2345	.1196	.2155	.1727	.2134
	N	11	10	11	9	12	8	31	30
Time on task (TT) – seconds *	Mean	925.89	1135.02	941.61	1236.87	1265.23	1377.89	1062.8	1230.3
	SD	302.3	195.2	406.50	385.05	286.63	447.49	365.1	339.26
	N	8	13	11	9	12	8	31	30
Time per relevant retrieved document (MnTT) - seconds	Mean	104.24	82.08	93.96	127.36	115.03	159.08	107.1	116.02
	SD	95.35	15.70	35.32	60.57	32.29	68.75	61.66116	57.95
	N	11	10	11	9	12	8	31	30
Accuracy (A) – F stat	Mean	.7310	.7806	.6362	.6677	.7020	.6146	.6756	.7085
	SD	.2215	.1212	.1349	.1899	.0952	.1998	.1560	.1771
	N	11	10	11	9	12	8	31	30
Efficiency (E) – A / TT	Mean	.0008	.00075	.00083	.00072	.00057	.00053	.00073	.00061
	SD	.00037	.00024	.00044	.00038	.00013	.00014	.00034	.00019
	N	11	10	11	9	12	8	31	30

* ANOVA shows an overall main effect ($p < 0.05$)

Table 3-5 Descriptive Statistics of Performance for High and Low Spatial Working Memory (MV) Participants in 3D100, 3D80 & 3D60 Environments.



ANOVA shows overall differences between high and low spatial working memory groups are significant ($p < 0.05$). However differences between high and low ability groups within each condition are not significant.

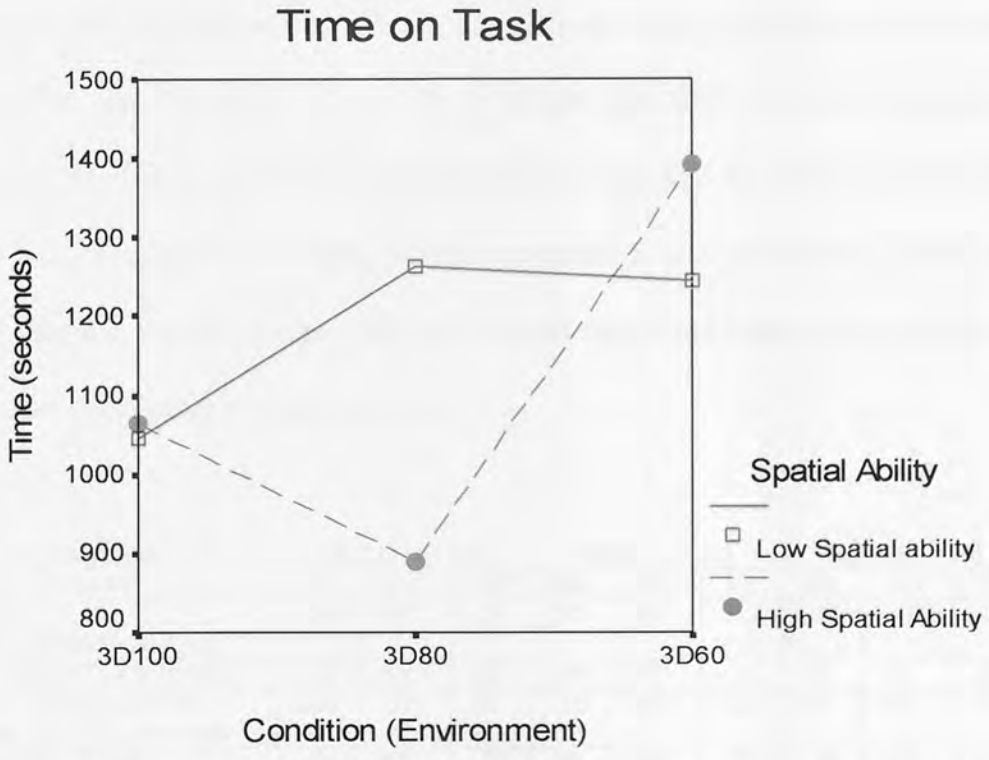
Figure 3-12 Plot of mean values of TT for Spatial Working Memory (MV) by Quality of Mapping

Effects of ability in spatial visualisation (VZ), demonstrated a borderline significant main effect for R, ($F(1,55) = 3.46$; $p = 0.068$). A significant interaction between ability and quality of spatial-semantic mapping occurred for TT, ($F(2,55) = 3.31$; $p < 0.05$). On closer inspection (see Table 3-6 and Figure 3-13) it can be seen that the differences between high and low ability groups are larger for 3D80 than for 3D100 and for 3D60. Independent measures t-tests confirmed this, showing that group differences for 3D80 were significant ($t(18) = 2.2$; $p < 0.05$), whereas differences between group differences for 3D100 and 3D60 were not significant.

Performance Measure		3D100		3D80		3D60		Total	
		Lo VZ	Hi VZ	Lo VZ	Hi VZ	Lo VZ	Hi VZ	Lo VZ	Hi VZ
Recall (R) - relevant/total retrieved docs	Mean	.7111	.6250	.5700	.4750	.5818	.4944	.6167	.5387
	SD	.1728	.2211	.2044	.1814	.1554	.1685	.1830	.2007
	N	9	12	10	10	11	9	30	31
Time on task (TT) – seconds *	Mean	1042.94	1064.65	1260.40	888.56	1242.68	1392.94	1188.67	1103.16
	SD	346.63	178.35	396.49	359.61	254.08	448.48	337.14	381.09
	N	9	12	10	10	11	9	30	31
Time per relevant retrieved document (MnTT) - seconds	Mean	73.82	108.58	120.60	97.38	117.54	151.12	105.45	117.32
	SD	22.05	88.22	56.00	42.73	54.53	48.17	50.77	67.27
	N	9	12	10	10	11	9	30	31
Accuracy (A) – F stat	Mean	.7797	.7359	.6883	.6125	.6892	.6399	.7161	.6682
	SD	.1347	.2091	.1732	.1402	.1318	.1696	.1486	.1809
	N	9	12	10	10	11	9	30	31
Efficiency (E) – A / TT	Mean	.00083	.00069	.00065	.00079	.00057	.00047	.00067	.00066
	SD	.0003	.00018	.00043	.00033	.00015	.00011	.00032	.00025
	N	9	12	10	10	11	9	30	31

* ANOVA shows significant interaction between high and low spatial ability groups across environmental conditions ($p < 0.05$). Between group differences are not significant for 3D100 and 3D60 but are significant for 3D80.

Table 3-6 Descriptive Statistics of Performance for High and Low Spatial Visualisation (VZ) Participants in 3D100, 3D80 & 3D60 Environments.



ANOVA shows significant interaction between high and low spatial ability groups across environmental conditions ($p < 0.05$). Between group differences are not significant for 3D100 and 3D60 but are significant for 3D80.

Figure 3-13 Plot of mean values of TT for VZ by Quality of Mapping

The results do not demonstrate a clear effect of cognitive ability on performance, but equally do not totally reject the experimental hypothesis. Therefore correlation and regression analyses were performed to endeavour to clarify the relationship between cognitive ability and performance, and identify issues of shared variance between the three different abilities measured. Restrictions associated with these techniques meant that each environment condition was examined independently.

The only significant relationship between abilities that was observed using a Pearson product moment correlation was for spatial working memory (MV) and associative memory (MA) in 3D60 ($r = 0.465$, $n = 20$; $p < 0.05$). All other pairwise comparisons of ability within individual conditions were not significant.

For measures of performance significant correlations were found between; i) R and MV in 3D100 and 3D60, ii) TT and MV in 3D100 and 3D80, iii) MnT and MV in 3D80 and 3D60, iv) A and MV in 3D100, and v) E and MV in 3D80 (see Table 3-7 for all r and associated p values). Simple regression analyses using MV as the predictor variable and performance as the DV confirmed these relationships. Table 3-8 shows model summaries of these analyses.

Performance Measure	3D100 (n = 21)			3D80 (n = 20)			3D60 (n = 20)		
	MA	MV	VZ	MA	MV	VZ	MA	MV	VZ
Recall (R) - relevant/total retrieved docs	-.24	.49*	-.24	.38	.16	-.12	-.17	-.45*	-.24
Time on task (TT) – seconds	-.12	.61**	.05	.26	.49*	-.15	.08	.07	.16
Time per relevant retrieved document (MnTT) - seconds	-.20	-.27	.25	-.06	.47*	-.02	.25	.54*	.31
Accuracy (A) – F stat	.10	.47*	-.23	.30	.16	-.07	-.14	-.40	-.10
Efficiency (E) – A / TT	.07	-.36	-.35	-.11	-.52*	-.09	-.14	-.42	.22

* Significant at $p < 0.05$
** Significant at $p < 0.01$

Table 3-7 Pearson r correlation coefficients for cognitive ability with performance in 3D conditions

Performance Measure	Condition (Environment)	R	R^2	Adjusted R^2	ANOVA
Recall (R) - relevant/total retrieved docs	3D100	.49	.24	.20	$F(1,19) = 6.09$; $p < 0.05$
	3D60	.45	.20	.16	$F(1,18) = 4.47$; $p < 0.05$
Time on task (TT) – seconds	3D100	.61	.37	.34	$F(1,19) = 11.37$; $p < 0.01$
	3D80	.49	.24	.19	$F(1,18) = 5.54$; $p < 0.05$
Time per relevant retrieved document (MnTT) - seconds	3D80	.47	.22	.18	$F(1,18) = 5.07$; $p < 0.05$
	3D60	.54	.29	.25	$F(1,18) = 7.38$; $p < 0.05$
Accuracy (A) – F stat	3D100	.47	.22	.18	$F(1,19) = 5.28$; $p < 0.05$
Efficiency (E) – A / TT	3D80	.52	.27	.23	$F(1,18) = 5.28$; $p < 0.05$

This table shows the model statistics for the simple regressions that were conducted between variables that demonstrated significant bivariate correlations. In all cases MV was the predictor variable.

Table 3-8 Simple regression model statistics for significant correlations between spatial working memory (MV) and performance

On the basis of these correlations and regressions, spatial working memory appears to be the only cognitive ability that impacts on individuals' performance. Results also suggest that the impact of working memory differs across environment dependent on the performance measure being examined. In 3D100 spatial working memory has a positive relationship with R, TT, and A, while in 3D80 there is a positive relationship on TT, and MnT, but a negative relationship with E. Furthermore, MV has a positive relationship with MnT, but a negative relationship with R, when the least efficient spatial-semantic mapping solution is used (3D60). Standard multiple regression analyses were conducted to ensure the observed effects of MV were not due to any co-variance between spatial ability and associative memory. All three cognitive ability measures were entered simultaneously as predictor variables of performance (response variable). A separate analysis was conducted for each performance measure (R, TT, MnT, A, and E) within each environmental condition; fifteen analyses in total (5 x 3).

When all three IVs were entered, the regression analyses only produced two significant models. These were for R and TT in 3D100 (Adj $r^2 = .28$, $F(3,17) = 3.64$; $p < 0.05$ and Adj $r^2 = .34$, $F(3,17) = 4.48$; $p < 0.05$ respectively). Inspection of the Beta

values and part correlation coefficients showed that for both DVs, MV was the only significant predictor within the models, and did not rely on any shared variance with VZ or MA. Table 3-9 (R) and Table 3-10 (TT) show unstandardised and standardised coefficients, values of t, and part correlation coefficients for MV, VZ, and MA.

Predictor Variables	B	Beta	t	Part correlation (unique)
(Constant)	.416		2.24*	
Spatial working memory (MV)	.028	.630	3.06**	.579
Spatial ability (VZ)	-.015	-.327	-1.69	-.321
Associative memory (MA)	-.007	-.211	-1.03	-.196

* significant at $p < 0.05$ Adjusted $r^2 = .28$, $F(3,17) = 3.64$; $p < 0.05$
** significant at $p < 0.01$

Table 3-9 Multiple Regression Coefficients for Cognitive Ability as a Predictor of Recall (R) in 3D100

Predictor Variables	B	Beta	t	Part correlation (unique)
(Constant)	507.850		2.25*	
Spatial working memory (MV)	41.111	.720	3.65**	.662
Spatial ability (VZ)	-3.004	-.051	-0.28	-.050
Associative memory (MA)	-10.683	-.269	-1.38	-.250

* significant at $p < 0.05$ Adj $r^2 = .34$, $F(3,17) = 4.48$; $p < 0.05$
** significant at $p < 0.01$

Table 3-10 Multiple Regression Coefficients for Cognitive Ability as a Predictor of Time on Task (TT) in 3D100

It should be noted there is some doubt about the suitability of the data for multiple regression analysis as the sample size is very low (n = 21 (3D100), n = 20 (3D80), and = 20 (3D60)). However, results were as generally expected based on the findings from the investigation of differences between high and low ability groups confirming that spatial working memory is the only ability that impacted on performance during this task. The results suggest that spatial working memory can predict some element of all the performance measures examined but that this is dependent on the accuracy of the spatial-semantic mapping used in database presentation. This is discussed in Section 3.4.3.

3.3.3.2 Individual Differences in Cognitive Ability and Number of Dimensions used in Environmental Mapping

To assess the effect of individual differences in cognitive ability on performance when the number of dimensions used for environmental presentation is manipulated, 2 (Ability – Hi & Lo) x 2 (Dimension – 2D60 & 3D60) ANOVAs, were conducted Further analysis using correlation and regression techniques were also undertaken. Again to prevent repetition, only significant main effects of ability and significant interactions between ability and dimension are presented in this section. For the purpose of individual differences 2D was not compared to 3D100.

For MA, no significant effects were observed – see Table 3-11 for descriptive statistics.

Performance Measure		3D60		2D60		Total	
		Lo MA	Hi MA	Lo MA	Hi MA	Lo MA	Hi MA
R	Mean	.5625	.5125	.4700	.6364	.5205	.5842
	SD	.1611	.1727	.1719	.2203	.1688	.2062
	N	12	8	10	11	22	19
TT	Mean	1321.97	1292.78	1113.10	1271.71	1227.03	1280.59
	SD	263.01	478.29	406.38	290.64	344.01	368.79
	N	12	8	10	11	22	19
MnT	Mean	128.92	138.25	126.76	107.10	127.94	120.22
	SD	52.51	57.48	42.28	28.76	47.02	44.66
	N	12	8	10	11	22	19
A	Mean	.6814	.6455	.5996	.7523	.6442	.7073
	SD	.1303	.1784	.1741	.1736	.1537	.1790
	N	12	8	10	11	22	19
E	Mean	.00053	.00052	.00059	.0006	.00056	.00057
	SD	.00013	.00017	.00026	.00012	.0002	.00014
	N	12	8	10	11	22	19

Table 3-11 Descriptive Statistics of Performance for High and Low Associative Memory (MA) Participants in 3D60 & 2D60 Environments.

When examining the effect of MV and dimension on performance no significant main effects of cognitive ability were observed, however significant interactions between

MV and dimension were observed for R, $F(1,37) = 4.36$; $p < 0.05$; MnTT, $F(1,37) = 7.11$; $p < 0.05$; and A, $F(1,37) = 5.91$; $p < 0.05$. In all cases low MV groups demonstrated better performance than high MV groups in the 3D60 environment but showed poorer performance in the 2D60 environment – see Table 3-12 for means and SDs, and Figure 3-14 for plots of interactions.

Performance Measure		3D60		2D60		Total	
		Lo WM	Hi WM	Lo WM	Hi WM	Lo WM	Hi WM
R	Mean	.5708	.5000	45	.6231	.5225	.5762
	SD	.1196	.2155	16.48	.2157	.1482	.2189
	N	12	8	8	13	20	21
TT	Mean	1265.23	1377.89	1114.15	1246.67	1204.8	1296.66
	SD	286.63	447.49	331.96	365.66	306.48	393.16
	N	12	8	8	13	20	21
MnTT	Mean	115.03	159.08	134.16	105.57	122.68	125.96
	SD	32.29	68.75	44.11	26.95	37.59	52.91
	N	12	8	8	13	20	21
A	Mean	.7020	.6329	.5795	.7412	.653	.6929
	SD	.0952	.1847	.1685	.1752	.1396	.1907
	N	12	8	8	13	20	21
E	Mean	.00058	.00045	.00057	.00061	.00057	.00055
	SD	.00013	.00012	.00028	.00011	.0002	.00014
	N	12	8	8	13	20	21

Table 3-12 Descriptive Statistics of Performance for High and Low Working Memory (MV) Participants in 3D60 & 2D60 Environments

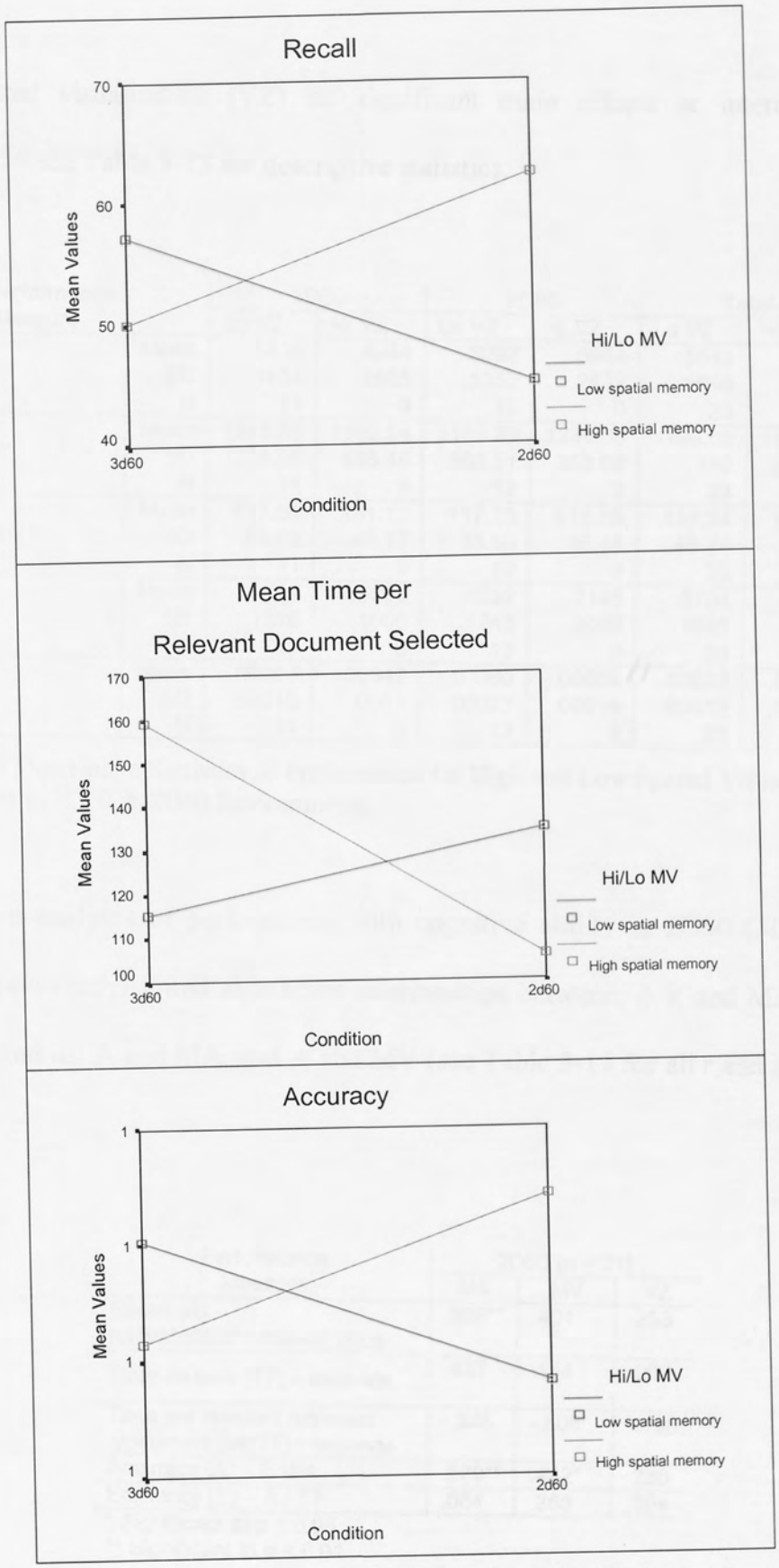


Figure 3-14 Plots of mean values of R, MnTT, & A for Working Memory (MV) by Dimension

For spatial visualisation (VZ) no significant main effects or interactions were observed – see Table 3-13 for descriptive statistics.

Performance Measure		3D60		2D60		Total	
		Lo VZ	Hi VZ	Lo VZ	Hi VZ	Lo VZ	Hi VZ
R	Mean	.5818	.4944	.5292	.5944	.5543	.5444
	SD	.1554	.1685	.1852	.2493	.1698	.2127
	N	11	9	12	9	23	18
TT	Mean	1242.68	1392.94	1160.26	1244.08	1199.68	1318.51
	SD	254.08	448.48	360.51	353.02	310	398.95
	N	11	9	12	9	23	18
MnTT	Mean	117.54	151.12	117.15	115.55	117.34	133.33
	SD	54.53	48.17	38.50	35.48	45.74	44.93
	N	11	9	12	9	23	18
A	Mean	.6892	.6399	.6532	.7148	.6704	.6774
	SD	.1318	.1696	.1743	.2069	.1531	.1875
	N	11	9	12	9	23	18
E	Mean	.00057	.00047	.00060	.00059	.00059	.00053
	SD	.00015	.0001	.00023	.00014	.00019	.00013
	N	11	9	12	9	23	18

Table 3-13 Descriptive Statistics of Performance for High and Low Spatial Visualisation (VZ) Participants in 3D60 & 2D60 Environments.

Correlation analyses of performance with cognitive ability in 2D60 (3D60 already reported previously) found significant relationships between; i) R and MA ii) MnTT and MV, and iii) A and MA, and A and MV (see Table 3-14 for all r and associated p values).

Performance Measure	2D60 (n = 21)		
	MA	MV	VZ
Recall (R) - relevant/total retrieved docs	.568**	.401	.253
Time on task (TT) – seconds	.432	.114	.082
Time per relevant retrieved document (MnTT) - seconds	-.345	-.508*	-.188
Accuracy (A) – F stat	.579**	.442*	.285
Efficiency (E) – A / TT	.064	.268	.104

* Significant at $p < 0.05$
** Significant at $p < 0.01$

Table 3-14 Pearson r correlation coefficients for cognitive ability with performance in 2D60

Performance Measure	Cognitive Ability	R	R^2	Adjusted R^2	ANOVA
Recall (R) - relevant/total retrieved docs	MA	.57	.32	.29	$F(1,19) = 9.05$; $p < 0.01$
Time on task (TT) – seconds	MV	.51	.26	.22	$F(1,19) = 6.63$; $p < 0.05$
Accuracy (A) – F stat	MA	.58	.34	.30	$F(1,19) = 9.6$; $p < 0.01$
	MV	.44	.19	.15	$F(1,19) = 4.6$; $p < 0.05$

This table shows the model statistics for the simple regressions that were conducted between variables that demonstrated significant bivariate correlations. For R associative memory (MA) was the predictor variable, for MnTT spatial working memory (MV) was the predictor, and for A both MA and MV were significant predictor variables.

Table 3-15 Simple regression model statistics for significant correlations between cognitive ability and performance

When multiple regressions were performed entering all IVs the only significant model produced was for A (Adj $r^2 = .28$, $F(3,17) = 3.58$; $p < 0.05$). However as can be seen from Table 3-16, no individual ability is a significant predictor within the model.

Predictor Variables	B	Beta	t	Part correlation (unique)
(Constant)	.193		.946	
Associative memory (MA)	.010	.385	1.475	.280
Spatial working memory (MV)	.013	.279	1.188	.226
Spatial ability (VZ)	.007	.145	.623	.118

Adj $r^2 = .28$, $F(3,17) = 3.58$; $p < 0.05$

Table 3-16 Multiple Regression Coefficients for Cognitive Ability as a Predictor of Accuracy (A) in 2D60

A further multiple regression analysis was conducted for A entering only MA and MV as predictor variables, as both these IVs were significantly correlated with A. As can be seen from Table 3-17, MA was the only significant predictor of A within this model. The Beta weights and part correlation coefficients show that the majority of the predictive power of MA is unique and not due to MV, on the other hand the observed zero order correlation between MV and A is most likely due to an interaction between MA and MV.

Predictor Variables	B	Beta	t	Part correlation (unique)
(Constant)	.269		1.666	
Associative memory (MA)	.013	.477	2.262*	.422
Spatial working memory (MV)	.010	.220	1.041	.194
* significant at p< 0.05	Adj r^2 = .28, F(3,18) = 5.36; p < 0.05			

Table 3-17 Multiple Regression Coefficients for MA and MV as a Predictor of Accuracy (A) in 2D60

3.4 Discussion

The study set out to answer three principal questions: - 1) Can information seekers successfully use spatial-semantic mapping to improve retrieval performance when browsing for loosely structured information?; 2) Is there a difference in performance dependent on how many dimensions (two or three) are used to map database contents in to the spatial-semantic environment?; 3) Do differences in cognitive ability between users impact on their performance when using spatial-semantic IR systems?

3.4.1 Effects of Mapping Quality

In terms of question one, the results show that for measures of recall (R), time on task (TT), and retrieval efficiency (E), users can and do employ the spatial-semantic mapping of documents in the database interface. A clear benefit to using spatial-semantic mapping was demonstrated, with best performance achieved in optimally mapped conditions i.e. 3D100. For both TT and E examination of the means showed a systematic decline in performance as the quality of mapping reduced (see Figure 3-1), although the size of differences also reduced as quality reduced and these differences were not found to be statistically significant when using post hoc comparisons. Differences between 3D80 and 3D60 were generally smaller and not significant compared to differences between these conditions and 3D100 (for E differences between 3D100 and 3D80 were small, but differences between 3D60 and 3D100 were

large). Recall also shows high performance in 3D100 with minimal differences in performance between 3D80 and 3D60.

The results compare with those found by Westerman & Cribbin (2000b) in a directed search task. Participants' utility of spatial-semantic mapping to locate and retrieve specified objects represented by single words is also applicable to browsing a document database to locate and retrieve documents relevant to a general query. Performance is aided when the information space employed for database presentation fits well with the information seekers internal model (cognitive map). Adding noise to the quality of spatial-semantic mapping within the IR environment reduces the 'goodness of fit' between the user's own map and the information space, causing interference in retrieval performance. The pattern of results would suggest that the goodness of fit between a user's mental model, and the IR environment is affected in a linear manner which in turn affects performance, but that once any degree of interference is introduced people struggle to perform well.

This need for a 'goodness of fit' between the users' mental model and the external information space is supported by the findings of Chen (2000), where participants who could sketch the most accurate maps of a spatial-semantic IR environment after task performance achieved better performance in terms of both recall and precision. Where mapping between the internal and external models is matched users can acquire a better model of the information space, which leads to improved retrieval performance.

Although not statistically significant, the pattern of improved performance with

optimal spatial-semantic mapping was also visible across the remaining two measures of performance accuracy (A), and average time per relevant document selected (MnTT), with means showing a systematic decline in performance as mapping quality reduced.

One possible reason for A not demonstrating statistical significance is a likely confounding effect of precision. Accuracy, which was measured using F-stat, represents the harmonic mean between precision and recall. Precision was excluded as a dependent measure due to a ceiling effect across all conditions. As the F-stat involves summing recall and precision the ceiling effect for precision could have masked the difference between groups. The most likely reason for such a high level of precision is that users may adopt strict criteria for identifying a document as relevant. Relevance has been shown to be multi-layered and dynamic dependent on many criteria such as information need, and motivation etc. (e.g. Schamber, et al., 1990; Mizzaro, 1998; and Harter, 1996). The retrieval task given to participants was to find as many documents as they could that were relevant to journalists facing personal risk. As participants were aware that the study was experimental they may have wanted to ensure they selected only documents that met the query exactly. In a real world situation however they may have been more lenient. This is however only conjecture as no measures of participants' criteria for judging a document relevant were taken. In future experiments of this nature such a post hoc measure would prove useful not only in moderating or identifying reasons for effects of precision, but also to identify any effects of spatial-semantic mapping on individuals' judgements of relevance.

3.4.2 Effects of Dimension

The second question addressed in this study was to determine the effects of using two or three dimensions to organise the database contents. Westerman & Cribbin (2000b) had found that there was no advantage to using 3D mapping solutions, as the advantages resulting from the extra semantic variance accounted for in the 3D mapping were outweighed by the "cognitive overhead" involved in navigating an extra dimension. The authors found that performance in 3D conditions was poorer overall than for 2D and performance between 2D and 3D solutions where 2D accounted for less semantic variance (50% and 70% of the semantic variance accounted for by the 3D solution), was comparable. Barshi & Healy (2002) found that while the number of dimensions along which people had to navigate (2D or 3D) did not affect performance, the number of dimensions required for mental representation of the information space did. The cognitive demands of maintaining a 3D mental model led to reduced performance.

In the current study comparisons of 3D60 and 2D60 solutions (each solution accounted for the same degree of semantic variance) showed no significant differences in performance, while comparisons of 3D100 and 2D60 demonstrated a significant increase in efficiency (E) for 3D100. These results contradict the findings of Westerman & Cribbin (2000b). It would appear that when browsing for information (as in the current study), rather than trying to locate a specific target people can perform equally well whether documents are mapped to two or three dimensions, and as such can benefit from the additional quality of spatial-semantic mapping that can be achieved by using the extra dimension. An explanation for this can be found in the nature of the mental representations required for task completion. In a directed search

task users would need to retain a very specific model of the information space in order to find the correct target (a similar target would not suffice). Retaining a 3D model has been shown to be more effortful than retaining a 2D model Barshi & Healy (2002). However for a browse task in which retrieval of relevant targets does not have to be completed in a specific sequence, and all relevant targets are placed in the same general vicinity when spatial-semantic mapping is good, a less specific model of the actual information space is required. Provided the information space approximates users' mental models, or facilitates users acquiring a conceptual map of the space, (which has been demonstrated when spatial-semantic mapping is optimal), performance will be enhanced. Using a three-dimensional mapping solution in preference to a two-dimensional solution provides additional conceptual information to allow this.

A further point for consideration within the current task should be highlighted; the document database consisted of only 100 documents, in terms of modern day information retrieval systems this is a very small quantity. If the database visualisation were to consist of many more nodes/documents the advantages of using three dimensions would most likely increase. This is a factor that could be considered in future studies. The current study used a small document database, principally due to the restrictions imposed by using MDS as a means for mapping the documents in to the environment on the basis of the ATA system used. Current advances in information visualisation however have produced valid alternatives to MDS that can handle much larger quantities of documents (e.g. Chen, Cribbin, Kuljis, & Macredie, 2002).

3.4.3 Individual Differences

The results presented in this study do not reflect the general findings in the literature regarding effects of cognitive ability on IR in spatial-semantic SDMSs. At the group level (high versus low ability), neither associative memory nor spatial ability were found to have an effect on performance and no interactive effects were found with either the quality of the spatial-semantic mapping in the search environments, or the number of dimensions used for database organisation. Correlation and regression analyses did however show a positive relationship between associative memory and recall, and associative memory and accuracy in the two-dimensional condition. For recall this relationship wasn't observed however when all three abilities were entered into the model. This was most likely due to complex interactive effects between the abilities which may have been due to the unsuitability of the data for regression analysis as the sample size was too small.

Spatial memory demonstrated a significant group level effect for time on task only, but in this was not in an intuitive direction, as low ability participants were quicker on task without relative loss of retrieval performance. In addition while the effect was demonstrated overall no group differences were observed within individual 3D conditions. This relationship between spatial working memory and time on task was further supported when correlation and regression analyses were conducted which showed that MV is a significant predictor of time on task in 3D100 only. However, the relationship is such that individuals with high spatial working memory spend more time on task. A further significant relationship between MV and performance in terms of recall was observed when using correlation and regression techniques. Again this significant relationship only occurred when spatial-semantic mapping quality was

optimal, i.e. 3D100. For recall, high spatial working memory indicated increased performance, which appears to be in opposition to the findings for time on task. Initially these relationships do not make sense, and again this may be an artefact of low reliability and validity due to using regression techniques with such a small sample. However, a possible explanation is that in the optimal condition where mapping in the information space best matches mapping in the user's cognitive model, recall is enhanced due relevant documents being easier to locate, but time on task is longer due to differences in adopted browse strategy. It is reasonable to assume that high spatial working memory is an advantage to helping people identify their position in the information space in relation to relevant documents if the space matches their cognitive model and no other cues are present. As such users are more likely to feel comfortable in an environment where they do not feel lost, and may therefore be prepared to spend longer on task ensuring they have found all possible target documents. However when spatial working memory is poorer and users become more lost not only does recall reduce people may not be prepared to spend as much time in the environment. The effects of this feeling of lostness would also explain why in environments that do not match user's cognitive models i.e. 3D80, and 3D60, performance in relation to spatial working memory becomes less predictable. Spatial working memory is of little or no benefit so users rely on alternative strategies.

The explanation that differences in effects of spatial working memory is due to differential effects of 'goodness of fit' between environment and cognitive map on adopted strategy is further supported by the results of correlation and regression analyses in 2D60. In this condition associative memory is a significant predictor of recall and accuracy, while spatial working memory has no individual predictive power

for any performance measure. It can be argued that due to a difference in visual organisation of the environment together with a reduced need to travel equally in all three dimensions users adopt a qualitatively different browse strategy that involves more rigorous exploration of all documents in the environment rather than using the spatial organisation (this is explored further in Chapter Four). As the spatial-semantic mapping of documents is a poorer match with the users' cognitive map, it can be argued that associative memory for how individual documents are semantically related to their neighbours provides better support than spatial memory of documents' positions in space.

Unfortunately much of this explanation is based on assumption as no measures or investigations of individuals' feelings while conducting the task were assessed. In future work, a more qualitative approach could be combined with the quantitative analysis to try to address this limitation, and possibly support these assumptions.

Differences related to all three cognitive abilities had been predicted on the basis of previous research (e.g. Westerman & Cribbin 1999; Chen, 2000; Chen & Macredie, 2000; and Westerman & Cribbin, 2000b) although the nature of those differences was not clear due to the discrepancies between findings previously described. For instance, when users performed a directed search and retrieve task, locating animals in an SDMS that was organised on the bases of human judgements of semantic similarity, Westerman & Cribbin (1999) found performance advantages for individuals of high spatial ability. The authors tested two environmental organisations one used an ordinal arrangement in which objects formed a cubed and the second used interval spacing which produced a more abstract and natural organisation. While high associative

memory predicted better performance when interval mapping was used participants with low associative memory demonstrated an advantage in the ordinally mapped environment. In a later study described earlier in this chapter in which a more 'natural' interval based mapping was employed, Westerman & Cribbin (2000b) again found high associative memory related to improved performance, although in terms of 'timed out' trials (i.e. participants did not locate the object in the time allotted) this only applied to 3D conditions, for 2D arrangements performance between groups was comparable. The positive effects of spatial ability on performance were also supported.

This contradictory effect of associative memory was also witnessed by Chen (2000), who as previously detailed in section 3.1.2.1 found a positive relationship with associative memory and recall performance in study one, but found a negative relationship between performance and associative memory in study two. However in factorial analyses no effects were observed for either spatial ability or associative memory. Both experiments used the same interface, and the task which required participants to locate a set number of documents (journal abstracts) related to a particular topic, was similar to that used in the current study.

The reason for theses contradictions in findings regarding associative memory and the lack of observed effects of associative memory and spatial ability in the current study is not obvious. The only explanation Chen (2000) offered was in relation to an experimental design flaw identifying that the sample size was small. The differences observed between the two experiments in Westerman & Cribbin (1999), and in Westerman & Cribbin (2000b) however was explained in terms of 'goodness of fit'

between users mental models and the information environment. In the first experiment the spatial arrangement of similarity judgements was calculated using an ordinal solution whereas in the second experiment and the subsequent study the arrangement was calculated using interval measures producing a more 'natural' solution. The authors argue that users with high associative memory use "richer semantic models" than those low in ability groups and therefore individuals high in ability would experience more interference from the ordinal arrangement of objects.

In the current study interval mapping was used suggesting an advantage for high ability groups would be expected. The general lack of observed differences (apart from a relationship between associative memory and recall and accuracy in 2D, and spatial working memory in 3D) may be due to the nature of the task as suggested earlier in section 3.4.2. Users may not rely on associative memory as much as spatial working memory when they do not need to utilise such rich semantic models for browsing. In such a task their position within the space is likely to be of more importance. When the nature of the space alters however and position in the space is less important than recalling the relationship between documents associative memory becomes a more important predictor of performance than spatial memory but the relationship isn't strong enough to identify differences between groups (i.e. in the 2D environment). It can be argued that processing capacity is conserved by maintaining approximate models of the environment, when the task doesn't require such detailed models, resulting in no observable differences between performance measures in different conditions.

The effects of spatial working memory on time taken reflect to some degree the

findings for associative memory in Westerman & Cribbin (1999) experiment one, in that time on task was longer for high ability groups. If individuals with high levels of working memory rely more heavily on the ability they may spend time trying to assimilate the organisation of the environment, whereas individuals with a low working memory capacity may automatically rely on alternative strategies. This could result in high working memory individuals taking longer. Despite the disparate nature of the findings discussed, the users' mental model is clearly implicated in all explanations of IR performance with relation to cognitive ability. The general lack of findings related to the cognitive abilities examined in the current study may be due to the dynamic and interactive nature of the users' model with system and task domain within information retrieval framework (Lin, Soergel, & Marchionini, 1991; and Marchionini & Shneiderman, 1988). The way in which spatial ability and memory are used to aid information retrieval is arguably dependent upon the nature of the users' mental model of the system and the task. When conducting a browsing task which is loosely structured a combination of strategies may be relied upon to construct a good mental model which do not rely specifically upon any single ability.

3.5 Conclusions

The reported experiment was designed using a generic interface with no document location or navigational cues other than spatial-semantic mapping for information seekers to use. The reason for this was to isolate issues related to the use of spatial-semantic mapping as a tool for improving information retrieval. Overall the results support the experimental hypotheses and confirm that in the absence of all other cues there is a benefit to users in applying good quality spatial-semantic mapping for the

visual organisation of document databases, when browsing for information in a spatial data management system. It is realistic to assume this can be an added advantage in more complex visual information retrieval systems. The results also demonstrate that there are efficiency advantages to using an extra dimension (3D) to improve the quality of spatial-semantic when performing relatively complex or unstructured tasks such as browsing. It is argued that the nature of the task qualitatively alters the nature of information seekers mental models. This is supported by the differences in the graphical representations of queries observed by Willie & Bruza (1995) (presented in section 3.1.1) in which complex queries were represented using alternative or combined techniques compared to representations of simple queries.

The results have demonstrated a minimal effect of cognitive ability on performance, which reflects to some extent the contradictory influence of cognitive ability on IR tasks using SDMSs employing spatial-semantic mapping. Explanations offered for these findings are that the nature of the browsing task compared to directed search tasks results in qualitatively different mental models that do not depend upon specific cognitive abilities in a systematic way. However when users with strong existing models try to assimilate the external model the additional processing time caused by a mismatch can be reflected in slower performance times.

4 Browsing Behaviour in a VE Database

4.1 Introduction

Chapter Three evaluated performance in terms of the products of information retrieval during a browse task. This chapter presents the results and conclusions of the study of user behaviour during the task, i.e. the navigation methods, browsing strategies, and browsing patterns employed. As detailed in Chapter Three the database contents (documents) were presented in a navigable virtual environment (VE) as objects (spheres/nodes) which were spatially mapped to the VE dependent on the degree of shared semantic information. Semantically similar documents were mapped closer together than semantically dissimilar documents. In this chapter attention is focused on how browsing behaviour, both in terms of navigating the environment (i.e. manipulating their position within the environment to move from object to object) and browsing pattern (i.e. the order in which objects/documents were visited), was affected by the configuration of the VE.

In previous studies of user interaction with SDMSs employing spatial-semantic mapping, behaviour and performance have tended to be examined in parallel (e.g. Westerman & Cribbin, 2000). However as referred to previously (see chapters One and Three), Marchionini & Shneiderman (1988) in their framework for information seeking, identify 'products' and 'processes' as two separate aspects of 'outcomes' (outcomes being one of the five principal factors that they suggest should be evaluated in order to assess effective and efficient information retrieval). Given this distinction it is worthwhile to consider products and processes separately when assessing the

benefits of spatial-semantic mapping as a tool for successful IR. In this thesis the term 'products' is replaced with the term 'performance' (full discussion and details of performance is provided in Chapter Three), and 'processes' is replaced with 'behaviour' and its associated concepts. Behaviour for the purpose of this study has been divided in to two categories 'navigation' and 'browsing pattern', the distinction between these two categories is discussed in sections 4.1.1 and 4.1.2 respectively, and the measures used to assess these behaviours are detailed in section 4.2.4.

The purpose of this study was to examine whether behaviour is influenced by factors associated with the spatial-semantic mapping of database contents in the absence of any cues to environmental layout or other navigational cues e.g. landmarks, identified routes/pathways, or labelled clusters/topics etc. The role of specific cognitive abilities was also assessed and reasons for this are discussed in section 4.1.3. Introductory material focuses on issues attributable to behaviour within this setting, and how navigating within virtual environments, and browsing patterns / strategies used for the exploration of the information environment are influenced by spatial-semantic mapping (see sections 4.1.1, and 4.1.2).

The nature of the IR system used within the current study limits to some degree the way navigation and browse strategies can be identified since the environment contains few measurable elements (only direction of travel, e.g. distance, rotation, and order of nodes/documents visited). To make a distinction between issues of navigation and the browsing strategies employed by users, navigation is taken to represent manipulation of the environment e.g. the distance travelled or the degree to which users rotate or change travel direction during browsing, while browsing strategy refers to the order in

which nodes/documents are accessed. It is not suggested these two factors are unrelated as an interaction clearly exists between the two. For instance the number of repeat visits to the same document contributes both to measures of efficient navigation (e.g. the ratio between unique and repeat visits to nodes is the basis for the "lostness" measure of navigation – see section 4.2.4.1) and to the order in which nodes are visited, from which identification of browsing strategy is derived. However, within the literature distinctions are made between foraging behaviour (i.e. the pattern of information access), and navigational issues with regard to wayfinding and travel, and useful insights drawn. Some of these are reviewed in the next two sections.

As described in Chapter Three, the environment represented in this study included no content or organisational cues other than distance between objects. There were no menus to view, no hierarchical structure of document organisation, no landmarks or directional guides (e.g. compass points or coordinates), and no labelled / identifiable topic clusters. However it was hypothesised that the spatial arrangement of the documents alone would provide sufficient cues to navigation provided the spatial-semantic mapping in the environment accurately reflects users' personal models of semantic relations, or contains sufficient generic / context free agreement to facilitate the acquisition of a good mental model.

4.1.1 Navigation

Research examining navigational learning in VE representations of real world models (e.g. building layouts – see for example (Ruddle, Payne, & Jones, 1998; and Richardson, Montello, & Hegarty, 1999)) shows that, users find the tasks very

difficult to begin with. However, experience with the VE improves navigational performance and can facilitate improved learning of routes for transfer to the real world model. For instance Witmer et al. (1996), demonstrated that practice in a VE model of a complex building led to improved route finding within the building when compared to verbal instructions and the use of photographs. Ruddle, Payne, & Jones (1997), found that similar levels of spatial knowledge of the layout of a physical environment (in their study a building) were acquired through practice in a VE, as when using a real world model. The navigational efficiency gained from practice in a VE can exceed that gained from practice in the real world environment, provided a sufficient level of training occurs (Waller, Hunt, & Knapp, 1998). The disadvantages associated with using a VE model are generally associated with disorientation, and the lack of way-finding cues that are present in real-world situations. Ruddle, Payne, & Jones (1998), found that participants' spatial knowledge, in terms of identifying their direction of travel, was good when only one or two changes in direction (90^0 turns) had been made, however when more changes in direction were required participants' performance reduced suggesting they experienced greater disorientation. Darken & Sibert (1996a) demonstrated that both overall navigational performance and the amount of directional information obtained in a VE, improved when map and grid cues regarding environmental layout and organisation were provided. With experience however, topographical information is sufficient to improve navigation. This was demonstrated by Ruddle, Payne, & Jones (1999), who showed that while generally navigational performance in a "very-large-scale" VE was most effective when users were provided with both local and global maps, with practice, the same level of performance could be achieved using global maps alone.

These findings suggest that with experience users' can assimilate the VE to their cognitive model and rely upon topographical cues for navigation. This is to some extent supported by the fact that route and survey knowledge can develop simultaneously during navigational experience, and is aided by the use of familiar landmarks (unfamiliar landmarks did not aid performance) (Ruddle, Payne, & Jones, 1997). Specific directional cues alone, e.g. the use of a compass, are not sufficient to prevent disorientation (Ruddle, Payne, & Jones, 1998). This also supports the need for a global representation of the entire environment to be assimilated with the users' mental model.

The underlying theory is that, the degree to which users' cognitive models of semantic information are represented by the spatial mapping within a VE information space, impacts on the effectiveness of information retrieval from that VE. In this chapter this assumption is applied to the behavioural element of IR. Dillon (2000) points out that while visual navigation aids are important in the use of virtual information environments, emphasis needs to be given to examining top-down processing. He suggests the users' mental models of semantic processing impact on their ability to derive shape from the information environment; this in turn impacts on their behaviour within the environment. While the research discussed would suggest experience is likely to be a factor in navigating VE information spaces, together with navigational aids such as landmarks, other important factors will be the degree of disorientation an individual experiences. Disorientation however, is inversely related to the accuracy of individuals' mental models of the environment Dillon (2000). If a high degree of topographical / global information is represented in the visualisation, i.e. if the visualisation melds well with the user's cognitive map, disorientation will be

reduced and way-finding behaviour improved. Landmark information was deliberately excluded from the VE used in the study reported here to determine whether a good quality spatial-semantic representation can provide sufficient topographical information to reflect the users' semantic map. This would in turn reduce disorientation and improve navigation. It is hypothesised that when the 'goodness of fit' between the VE mapping and users' cognitive models of the semantic space is maximised, navigation behaviour as determined by environment manipulation will be most efficient.

4.1.2 Browsing Patterns

As previously stated browsing patterns for the purpose of this thesis are being examined separately from navigational behaviour despite a strong inter-dependence between the two. Knowledge about individual differences in the psychological processes involved in computerised information retrieval can be gained by examining peoples' information seeking behaviour, in terms of the strategies and patterns of browsing that they employ, during such tasks (see Chang & Rice, 1993; and Marchionini, 1995). The identification and quantification of browsing strategies or patterns of browsing is problematic however. This is largely due to the multidimensional nature of browsing, and the numerous factors, both internal and external to the user, that determine browsing behaviour (Chang & Rice, 1993). Many studies have used verbal protocols as a basis for analysis (e.g. Chang & McDaniel, 1995; and Kwasnik, 1992). Unfortunately, these generally rely on post hoc allocation of individual responses to categories by the experimenter, and only measure behaviours of which the individual is consciously aware. Other studies have analysed usage patterns of menus and sub menus (Toms, 2000), query formulation and number

of retrieved pages visited (e.g. Jansen, et al., 2000), the use of system generated cues (e.g. 'more like these' lists) to similar documents, and 'hyperties' (hypertext based on the Interactive Encyclopaedia System, in which hyperlinks to additional information are embedded in the current document in the form of highlighted words or text) (e.g. Jansen, et al., 2000; and Marchionini & Shneiderman, 1988). Such studies quantify browsing behaviour in terms of the queries generated, the use of Boolean operators, and pages visited etc., to provide valuable information regarding the usability of a traditional IR system but do not identify specific browsing patterns within spatial VE databases. If a means of measuring the sequence of document visitation i.e. a pattern of browsing, can be identified this will lead to greater understanding of the cognitive processes involved in browsing databases, and provide further insight into users' cognitive maps of the information environment.

Chen, et al. (2002) examined behaviour in terms of the sequential patterns of browsing people employ when travelling through a visual-spatial VE employing spatial-semantic mapping. The perspective they adopted was 'social navigation', i.e. people with similar ideas demonstrating similar browsing patterns, and individuals benefiting and learning from one another in terms of following the 'trails' people use when browsing for information. The authors used Hidden Markov Models (HMM) as a means of identifying users' trails (see Chen, et al. (2002) for an explanation of the process). They were able to produce visualisations of users' foraging behaviour and the trails they followed, and produce computations of optimal foraging patterns by modelling the behaviour of the most proficient information seekers. This research offers promising insights into the use of spatial-semantic information visualisations and the nature of users browsing patterns. It is suggested that within multi-user

environments, in which information seekers can be represented as Avatars, the occurrence of social navigation (i.e. learning from each other by following similar browsing trails) could be examined to identify ways of optimising user behaviour. As this was an early study, principally testing the methodology, the sample size used was small (N = 5), and therefore did not enable between user comparisons on the basis of cognitive ability. Additionally the authors were not examining the effects of spatial-semantic mapping quality on browsing behaviour (although three different methods for visualisation were tested).

The current study set out to test a means of quantifying browsing patterns to identify effects of mapping quality, and users' cognitive ability, on the sequences of documents visited (i.e. browsing pattern). It was predicted that in conditions for which spatial-semantic mapping was optimal, users' browsing patterns would be more similar, on the basis that users could employ the spatial-semantic mapping to drive their information seeking behaviour. Early analyses of the data from this study were presented at conference (Collins & Westerman, 2001), and support this hypothesis. Since this time the way in which the data were prepared for analysis has been reviewed and findings have altered somewhat. This is discussed fully later in the chapter (see section 4.2.4.2). The method of browsing pattern quantification that was tested was based on the n-gram approach to ATA and is described in section 4.2.4.2.

4.1.3 Individual Differences

This study also examined individuals' behaviour during the IR experiment reported in Chapter Three. The arguments for examining spatial ability, associative memory, and spatial working memory were presented in detail there. However the main focus of

these arguments was on 'performance' rather than behaviour. Within the studies referenced a general distinction between these two factors was not made, and many of the findings related to behavioural measures. For instance Westerman & Cribbin (2000b) found differences related to associative memory in measures of number of objects visited, and number of different objects visited. In this thesis these variables are used as measures of navigation behaviour. The previously mentioned preliminary study of the data collected during this experiment presented in Collins & Westerman (2001) also demonstrated effects on browsing patterns of individual differences in all three abilities. Westerman & Cribbin (1999), found an overall effect of associative memory for navigational efficiency, with high ability groups demonstrating a disadvantage. They also found that individuals with low levels of spatial ability tended to perform more poorly in terms of lostness, although these findings only approached significance at the 5% probability level.

The findings regarding individual differences in behavioural aspects of IR from VE systems present similar discrepancies as those observed for measures of performance. Modjeska & Chignell (2003) found navigating a virtual information world was subject to effects of spatial ability in terms of "doing", which was measured using variables such as distance travelled, whereby individuals with high levels of spatial ability performed better. (Westerman1998) however, found no observable effects of spatial ability, when examining the effects of using two or three dimensions for environmental presentation on target retrieval. In their study, 64 cubes were arranged in either an 8 x 8 (2D), or 4 x 4 x 4 (3D) layout with no account taken of spatial-semantic mapping. Cubes had a single letter displayed on all sides and participants were given a target letter to locate; measures of time on task, distance travelled, and

navigational efficiency were recorded. The author suggests the absence of effects of spatial ability is due to the lack of semantic information within the environmental mapping. It is suggested that the effects of spatial ability observed in other studies is more likely due to the representation of semantic distance or structure rather than the shape of the spatial structure of the object arrangements specifically.

In the current study effects of individual differences in spatial ability, spatial memory, and associative memory on behavioural measures were examined, and two-tailed hypotheses (that differences in behaviour would be observed between groups of individuals with high or low levels of cognitive ability) were tested.

4.2 Methodology

The data for the current analyses were collected during the experiment for which performance analyses reported in Chapter Three. Therefore several sections below refer the reader to the methodology in Chapter Three.

4.2.1 Experimental Platform

Full details of the experimental platform, including the design of environment, and how the quality of semantic mapping and the number of dimensions used for mapping documents were manipulated are given in section 3.2.1. The term 'quality of semantic variance' is used to refer to the amount of inter-document semantic similarity / dissimilarity accounted for by the spatial proximity of documents within the environment, and was examined using three differentially mapped environments: i) 3D100 which represented the maximum semantic variance accounted for; ii) 3D80

which accounted for 77% of the semantic variance accounted for by 3D100; and iii) 3d60 which accounted for 56% of the best solution. A comparison was also made between using two and three dimensions for semantic. The 2D solution was compared to 3D60 as removal of a dimension from 3D100 to create 2D reduced the amount of semantic variance accounted for within the spatial-mapping to 56% of the variance accounted for by 3D100. By comparing 2D60 and 3D60 the quality of spatial-semantic mapping was maintained across conditions.

4.2.2 Participants and Procedure

Chapter Three, section 3.2.2 gives details of participants and procedure.

4.2.3 Experimental Design

One-way ANOVAs were used to assess the effects on navigation, of i) the quality of spatial-semantic mapping (three levels 3D100, 3D80, and 3D60) and ii) the number of dimensions (two levels 3D60 & 2D60) used for mapping. To analyse the effects of individual differences in cognitive ability on navigation a median split of participants' results for each cognitive measure was performed resulting in two levels (Hi & Lo). These IVs were then combined using two-way ANOVAs with the previous IVs for quality of spatial-semantic mapping (2 x 3), and number of dimensions used for mapping (2 x 2). It was decided not to perform additional correlation and regression analyses (as conducted in Chapter Three) with respect to cognitive ability and navigation. This was because the factorial analyses (which were conducted first) showed no significant effects of cognitive ability (4.3.1.2), and as explained in Chapter Three the data were not best suited to give reliable multiple regression solutions as the sample size was too small (n = 21 in 3D100, 20 in 3D80, 20 in 3D60,

and 20 in 2D60).

Factorial designs were used to assess the effects of spatial-semantic mapping on browsing patterns. Browsing patterns were assessed based on the n-gram approach used in Chapter Two, and detailed in Section 4.2.4 of the current chapter which produced cosine values for all pairs of participants. A further IV, n-gram length, was added to the IVs for mapping previously detailed and ANOVAs were used to assess main effects and interactions of these IVs on mean cosine value.

A correlation design was used for examining possible relationships between individual differences in cognitive ability and adopted browsing patterns. The inter-participant cosine measures of browsing pattern similarity were correlated with all pair-wise difference scores between individuals for each cognitive ability measure (see section 4.3.2 for details).

4.2.4 Measures of Behaviour

Browsing behaviour was measured according to the two interpretations of behaviour previously described, which were both influenced by users' browsing strategy. The first interpretation is referred to as 'navigation' and was based on a collection of measures that identified spatial movement within the environment (see Section 4.2.4.1). The second interpretation of browsing behaviour is users' 'browsing pattern', which was identified based on the sequence in which documents were visited (see Section 4.2.4.2). For the reasons stated in section 3.2.1 regarding the variable 'Time on Task', all measurements were taken between the points at which the first node / document was selected and the final node / document selection.

4.2.4.1 Measures of Navigation

Navigation was measured using seven dependent variables in all. These were: i) the total distance travelled (distance); ii) average step distance between documents (step size); iii) the total degree to which the environment was rotated (rotation); iv) average step angle between documents (step angle) v) a measure of how lost participants became (lostness) (Smith1996); vi) the total number of nodes / documents visited (total nodes); vii) the number of unique documents visited (unique nodes).

Distance travelled and rotation values represent cumulative totals recorded during task performance. The unit of measurement was a virtual unit programmed into the VE design. Step distance represents the Euclidean distance between two nodes visited consecutively, divided by the total number of nodes visited. The step angle is the dot product of the inter-node angle between consecutively visited nodes, divided by the total nodes visited. It was discovered during initial analyses that the process of adding noise to the 3D80 and 3D60 environments resulted in an overall increase in environmental area, which increased inter-document distance. To account for this total distance travelled and mean step distance were averaged by environment size.

Lostness was a rating measure based on the combination of the ratio of unique nodes visited to total nodes visited and the ratio of required nodes to fulfil the task (i.e. relevant nodes) to unique nodes visited - $\sqrt{\left((D/T-1)^2+(R/D-1)^2\right)}$ where D = unique nodes visited, T = total nodes visited, R = total relevant nodes in database. This measure, which was introduced by Smith (1996) to quantify issues associated with the 'lostness' experienced by users searching hyper environments such as the

internet, has since been used by Westerman (1998), Westerman & Cribbin (2000b), and Westerman et. al. (2005) as a measure of user behaviour within a VE SDMS.

4.2.4.2 Measures of Browsing Pattern

Browsing pattern was measured and analysed using the same n-gram principle employed and tested for automatic text analysis (ATA) of document similarity in Chapter Two. The objects in the environment were allocated unique numerical identifiers and the sequence in which objects were visited was recorded during task performance. For each participant, windows 2, 3, 4, and 5-grams long were passed across this record of documents visited. Cosines were then calculated for every pair-wise comparison of participants in each of the semantic variance conditions (i.e. 3d100, 3D80, 3D60, & 2D60), based on the co-occurrence of n-gram sequences using the VSM detailed in Chapter Two. Two methods for identifying matching sequences were compared in the first instance. The first was 'exact-match', in which only an exact match of the combination of numbers within the n-grams constituted co-occurrence. As such, a 4-gram exact-match a pattern of node visits of 3,4,5,6 would only co-occur with 3,4,5,6; a visiting order of 4,3,6,5 would not be considered a match. However it was expected that the probability of individuals sharing exactly the same browsing pattern given the infinite possible combinations available during the task completion, would be too limiting in terms of the amount of information regarding browsing pattern obtainable. The second method for identifying browse sequences was referred to as an 'any-match' n-gram analysis. In this instance the requirement for n-grams to be considered co-occurring was for the same object identifiers to appear within the n-gram windows but the order in which they appeared did not have to match. In other words the 4-grams '3,4,5,6' and '4,3,6,5' would be

considered a match, as participants have visited the same documents within the n-gram window. In section 4.3.2 the comparison of these two methods is presented and reasons for proceeding with analyses using the 'any-match' method are discussed.

It should also be noted that during recording of task performance pointing the cursor at a node records a visit to that node, additional selecting or de-selecting the node records a visit. Therefore when a document is deemed relevant and the user selects it, a sequence of two visits to the node are recorded, if immediately the user changes their mind and de-selects the document a sequence of three visits would be recorded. While reading a document, the cursor moved slightly the document would disappear requiring the user to re-position the cursor. This would then record a further visit. It was found that for many users' relatively long sequences of the same node number (e.g. '....2 5 4 3 3 3 3 3 10 10 1 76 .....') were recorded, so for the purpose of identifying browsing sequences all duplicate visits occurring consecutively were removed (e.g. the above sequence became '....2 5 4 3 10 1 76 .....').

4.2.5 Materials

Chapter Three section 3.1.1 provides details of the experimental platform used in this experiment. Details of the cognitive ability tests employed to measure spatial visualisation (VZ), associative memory (MA), and spatial working memory (MV) are given in Chapter 3 section 3.1.5, together with details of the software, and equipment used.

4.3 Results

The results reported here are presented in two sections based on the interpretations of browsing behaviour detailed earlier. Section 4.3.1 presents results attributed to 'navigation' and section 4.3.2 presents the results of the analyses of users' 'browsing pattern'. Each section is further subdivided to examine the effects of environmental mapping, and individual differences in cognitive ability on user behaviour.

4.3.1 Navigation

Navigation was examined for effects due to 'environmental mapping' and 'cognitive ability'. Section 4.3.1.1., presents the results of effects due to environmental mapping, and the results of effects due to individual differences in cognitive ability are given in section 4.3.1.2.

4.3.1.1 Environmental Mapping and Navigation

Differences in navigation due to the way in which the environment was mapped were examined from two perspectives; i) the quality of mapping, i.e. how much semantic variance is accounted for by the placement of documents, and ii) the number of dimensions used for document layout. All seven variables detailed in section 4.2.4 were used to examine the quality of environmental mapping. When comparing dimensions the inherent differences in environment volume resulting from the removal of a dimension make it difficult to isolate differences occurring as a result of differential behaviour for distance travelled, rotation, step distance, and step angle. Therefore, only three dependant variables were analysed to determine the effects of the number of dimensions used for environmental mapping on navigation. These were lostness, total nodes visited, and unique nodes visited.

4.3.1.1.1 Quality of Mapping

Initial analysis of the data showed that step size (skewness = .885, kurtosis = 1.4), rotation (skewness = 1.29, kurtosis = 1.54), and step angle (skewness = 1.7, kurtosis = 2.97) were not normally distributed. When examining each condition separately, step size was normally distributed for 3D100 (skewness = .783, kurtosis = .9) and 3D60 (skewness = .725, kurtosis = -.01), and rotation was normally distributed for 3D60 (skewness = .56, kurtosis = -.54). Various transformations of the data were attempted but distributions were not improved. It was decided that as the normality of distribution was not severely violated i.e. < 2 (except kurtosis for step angle), a parametric analysis of the data was more appropriate as this permitted a more in depth factorial analysis and also provided more conservative results.

Table 4-1 shows descriptive statistics for distance, step size, rotation, step angle, lostness, total nodes, and unique nodes, and Figure 4-1 contains graphs of means for 3D conditions.

Performance Measure	Condition	N	Mean	SD
Distance – total VE units travelled**	3D100	21	25409	14931
	3D80	20	37204	25525
	3D60	20	65145	25796
Step size – average VE units travelled between nodes**	3D100	21	.3360	.0482
	3D80	20	.4051	.0803
	3D60	20	.4261	.0563
Rotation – total degrees of environmental rotation**	3D100	21	2552	2584
	3D80	20	4542	4028
	3D60	20	5972	3729
Step angle – average angle between consecutive nodes visited	3D100	21	.6602	.5606
	3D80	20	.7608	.7025
	3D60	20	.5588	.4388
Lostness – total lostness ratio based on Smith (1996)	3D100	21	.7391	.126
	3D80	20	.683	.1946
	3D60	20	.7964	.1218
Total nodes – total documents visited	3D100	21	100.04	46.45
	3D80	20	92.1	49.37
	3D60	20	118.4	47
Unique Nodes* - total documents visited once only	3D100	21	50.48	18.64
	3D80	20	43.7	15.92
	3D60	20	57.65	15.67

All measures are taken between the first and last documents selected
* ANOVA shows significant main effect of Condition $p < 0.05$
** ANOVA shows significant main effect of Condition $p < 0.001$

Table 4-1 Descriptive Statistics for Navigation in 3D100, 3D80, and 3D60 Environments

One-way analyses of variance with environment as the IV with three levels - 3D100, 3D80, and 3D60 - showed significant main effects for distance, $F(2,58) = 16.67$; $p < 0.001$, step size $F(2,58) = 11.61$; $p < 0.001$, rotation, $F(2,58) = 4.98$; $p < 0.001$, and unique nodes $F(2,58) = 3.44$; $p < 0.05$. Post hoc comparisons were conducted using Tukey HSD and demonstrated that: for distance people travelled significantly further in 3D60 than in 3D100 and 3D80; step size was significantly smaller in 3D100 than 3D80 and 3D60; rotation was significantly less in 3D100 than 3D60; and significantly more unique nodes were visited in 3D60 than 3D80.

Lostness approached significance $F(2,58) = 2.83$; $p = 0.067$, however post hoc comparisons using Tamhane (homogeneity of variance was violated and equal variances were not assumed) showed no significant differences between conditions.

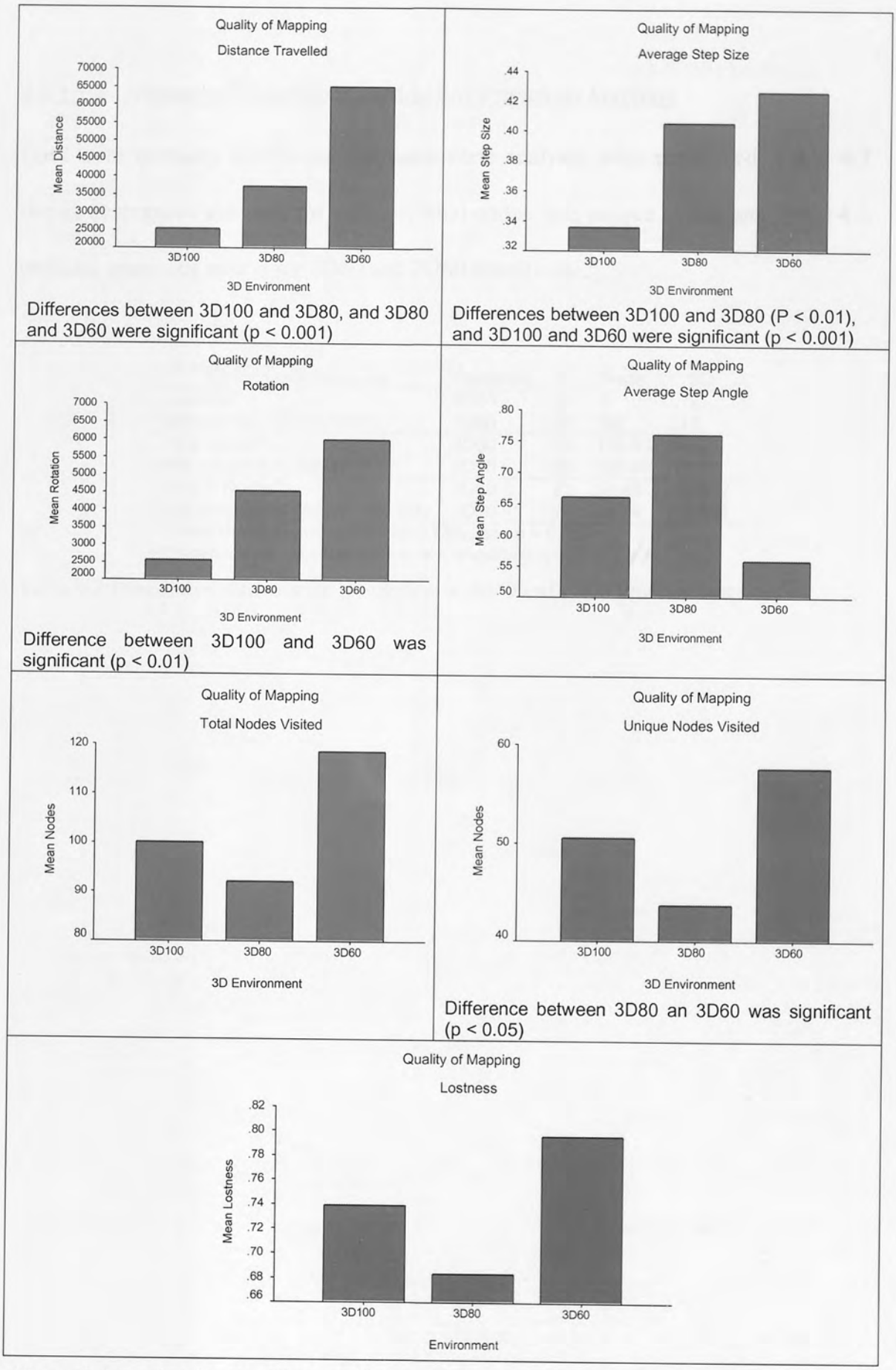


Figure 4-1 Graphs of Means for the Effect of Quality of Mapping on Navigation

4.3.1.1.2 Number of Dimensions used in Environmental Mapping

Data were normally distributed and parametric analyses were performed. Table 4-2 shows descriptive statistics for lostness, total nodes, and unique nodes, and Table 4-2 contains graphs of means for 3D60 and 2D60 conditions.

Performance Measure	Condition	N	Mean	SD
Lostness* - lostness ratio Smith (1996)	3D60	21	.8	.12
	2D60	20	.89	.11
Total nodes* - total documents visited	3D60	21	118.40	47.01
	2D60	20	169.90	78.76
Unique Nodes* - total documents visited once only	3D60	21	57.65	15.672
	2D60	20	74.24	18.558

* t-test shows significant effect of Condition $p < 0.05$
** t-test shows significant effect of Condition $p < 0.01$

Table 4-2 Descriptive Statistics for Navigation in 3D60 and 2D60 Environments

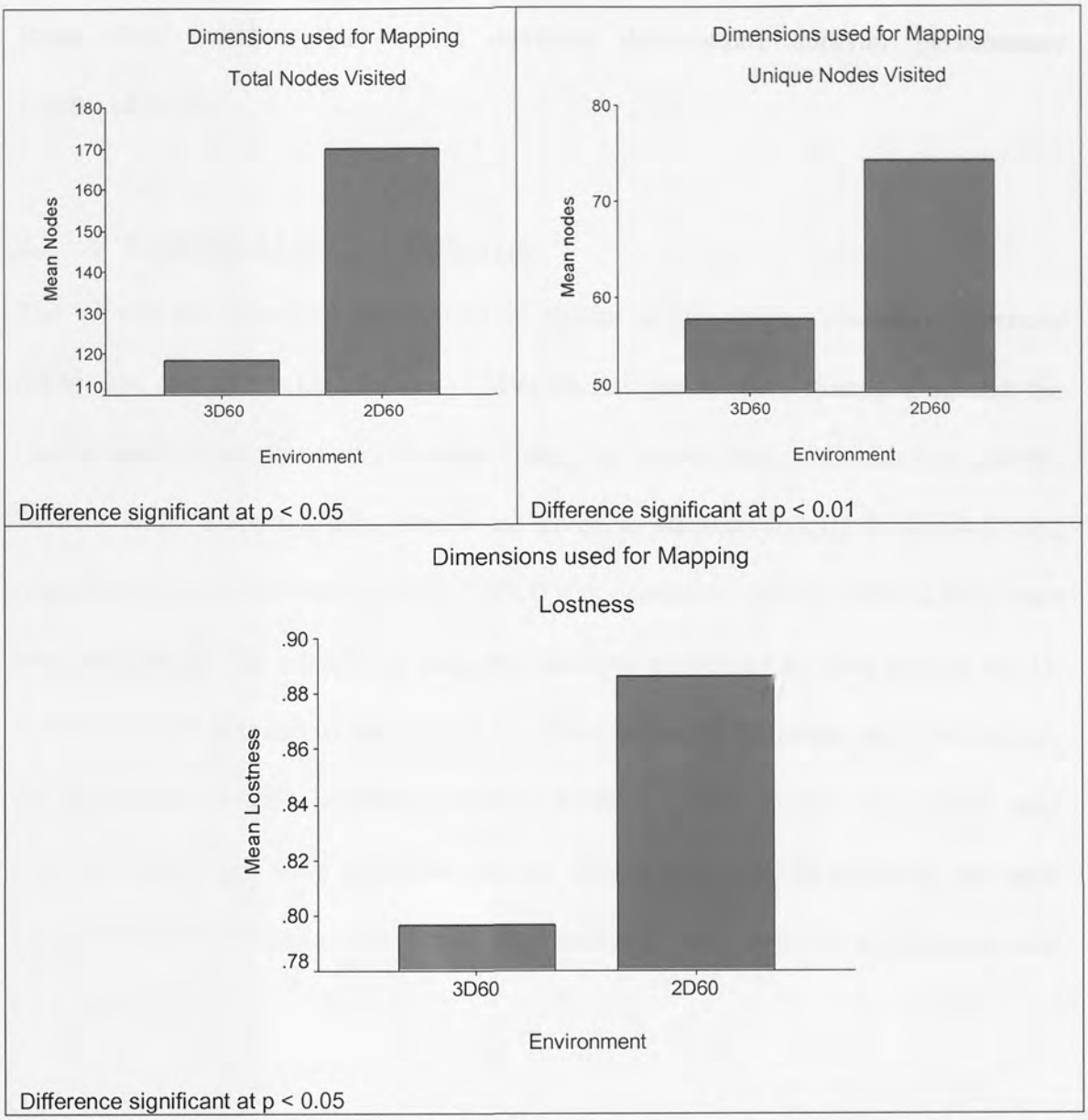


Figure 4-2 Graphs of Means for the Effect of Dimensions used for Environmental Mapping on Navigation

Independent t-tests were performed and showed that the mean number of total nodes ($t(39) = -2.53$; $p < 0.05$), unique nodes visited ($t(39) = -3.09$; $p < 0.01$), and lostness ($t(39) = -2.46$; $p < 0.05$), differed significantly between 3D60 and 2D60. As comparisons between 2D60 and 3D60 demonstrated an advantage for 3D when less than optimum semantic variance was accounted for comparisons of 2D60 and 3D100 were not conducted as this would produce redundant material. Although differences were not significant 3D100 had produced greater performance scores than 3D60 on

these three measures and would therefore demonstrate superior performance compared to 2D.

4.3.1.2 Cognitive Ability and Navigation

The effects of individual differences in spatial ability (VZ), associative memory (MA), and spatial working memory (MV) on navigation were examined against the two parameters quality of environmental mapping, and number of dimensions used for environmental mapping. The results are given in sections 4.3.1.2.1, and 4.3.1.2.2 respectively. As explained previously 2D60 was compared only to 3D60 as they were best matched for the amount of semantic variance accounted for (see section 4.2.1). For the reasons detailed in section 4.3.1.1 'Environmental Mapping and Navigation', the dependent variables distance, rotation, lostness, unique nodes, total nodes, step size, and step angle were examined for the effects of quality of mapping, but only lostness, total nodes, and unique nodes were analysed with respect to dimensions used for mapping.

4.3.1.2.1 Individual Differences in Cognitive Ability and Quality of Environmental Mapping

A median split was calculated for each of the three cognitive abilities and participants were allocated to high or low ability groups on this basis. Two-way independent measures ANOVAs were performed using a 2 (Ability – Hi & Lo) x 3 (Mapping Quality – 3D100, 3D80, & 3D60) model for each ability (MA, MV, and VZ) by each DV detailed in the previous section. Means and standard deviations of distance, rotation, lostness, unique documents visited, total documents visited, step distance, and step angle are shown for MA, MV, and VZ in Table 4-3, Table 4-4, and Table 4-5

respectively. Results of the analyses for all ANOVAs showed no significant main effects of cognitive ability on behaviour and no significant interactions.

Performance Measure		3D100		3D80		3D60		Total	
		Lo MA	Hi MA	Lo MA	Hi MA	Lo MA	Hi MA	Lo MA	Hi MA
Distance – Total VE units travelled	Mean	24997	25863	29515	43494	64943	65447	41247	43470
	SD	14927	15729	20772	28213	26377	26696	27948	28119
	N	11	10	9	11	12	8	32	29
Rotation – total degrees of environmental rotation	Mean	299	2092.70	4205.56	4818.00	6383.25	5355.88	4597.06	4026.62
	SD	3041	2030	3198	4740	3431	4303	3467	4008
	N	11	10	9	11	12	8	32	29
Lostness – total lostness ratio Smith (1996)	Mean	.7190	.7612	.6488	.7107	.8253	.7531	.7392	.7398
	SD	.1143	.1405	.1556	.2251	.0795	.1636	.1345	.178
	N	11	10	9	11	12	8	32	29
Unique Nodes – total documents visited once only	Mean	49.18	51.90	37.78	48.55	60.08	54.00	50.06	51.21
	SD	15.17	22.62	10.57	18.29	13.51	18.82	15.84	19.44
	N	11	10	9	11	12	8	32	29
Total Nodes – total documents visited	Mean	91.18	109.80	78.22	103.45	123.83	110.25	99.78	107.52
	SD	37.65	54.93	37.82	56.32	37.2	60.8	41.26	55.11
	N	11	10	9	11	12	8	32	29
Step Size – average VE units travelled between nodes	Mean	.3346	.3374	.3869	.4200	.4234	.4301	.3826	.3943
	SD	.0604	.0332	.0651	.0912	.043	.0753	.0663	.0808
	N	11	10	9	11	12	8	32	29
Step Angle – average angle between consecutive nodes	Mean	.6578	.663	.8951	.6509	.4072	.7862	.6306	.6924
	SD	.5743	.5762	.5967	.7895	.2327	.5822	.5073	.6462
	N	11	10	9	11	12	8	32	29

Table 4-3 Descriptive Statistics for Navigation of High and Low Associative Memory (MA) Participants in 3D100, 3D80, and 3D60 Environments

Performance Measure		3D100		3D80		3D60		Total	
		Lo MV	Hi MV	Lo MV	Hi MV	Lo MV	Hi MV	Lo MV	Hi MV
Distance – Total VE units travelled	Mean	17467	3029	35428	39375	64801	65660	42163	42450
	SD	9540	15828	24784	27747	27889	24153	29924	25974
	N	8	13	11	9	12	8	31	30
Rotation – total degrees of environmental rotation	Mean	2269	2726	5208	3729	6035	5879	4769	3868
	SD	2580	2676	5052	2300	3699	4028	4178	3169
	N	8	13	11	9	12	8	31	30
Lostness – total lostness ratio Smith (1996)	Mean	.6735	.7794	.6363	.7399	.8005	.7903	.7095	.7705
	SD	.1411	.1011	.2136	.1620	.111	.1444	.1732	.1301
	N	8	13	11	9	12	8	31	30
Unique Nodes – total documents visited once only	Mean	43.87	54.54	43.09	44.44	56.50	59.37	48.48	52.80
	SD	19.48	17.62	18.21	13.64	15.81	16.37	18.24	16.72
	N	8	13	11	9	12	8	31	30
Total Nodes – total documents visited	Mean	85.13	109.23	82.73	103.56	117.42	119.88	96.77	110.37
	SD	55	39.88	51.79	46.57	42.23	56.5	50.33	45.46
	N	8	13	11	9	12	8	31	30
Step Size – average VE units travelled between nodes	Mean	.3224	.3443	.4173	.3903	.4297	.4205	.3976	.3784
	SD	.0523	.0456	.093	.0637	.0504	.0675	.0807	.0643
	N	8	13	11	9	12	8	31	30
Step Angle – average angle between consecutive nodes	Mean	.885	.5219	.9188	.5677	.5091	.6333	.7515	.5654
	SD	.8005	.3093	.8238	.4976	.4154	.4907	.691	.41
	N	8	13	11	9	12	8	31	30

Table 4-4 Descriptive Statistics for Navigation of High and Low Spatial Working Memory (MV) Participants in 3D100, 3D80, and 3D60 Environments

Performance Measure		3D100		3D80		3D60		Total	
		Lo VZ	Hi VZ	Lo VZ	Hi VZ	Lo VZ	Hi VZ	Lo VZ	Hi VZ
Distance – Total VE units travelled	Mean	28657	22973	46710	27698	65055	65255	48021	36773
	SD	18676	11686	22914	25487	28954	23075	27844	27091
	N	9	12	10	10	11	9	30	31
Rotation – total degrees of environmental rotation	Mean	2934	2265	5700	3385	4805	7399	4542	4117
	SD	3219	2097	4831	2820	3825	3252	4054	3406
	N	9	12	10	10	11	9	30	31
Lostness – total lostness ratio Smith (1996)	Mean	.7191	.7540	.7551	.6108	.7962	.7966	.7594	.7202
	SD	.1518	.1074	.2195	.1420	.1150	.1368	.1640	.1465
	N	9	12	10	10	11	9	30	31
Unique Nodes – total documents visited once only	Mean	49.44	51.25	49.40	38.00	55.64	60.11	51.70	49.55
	SD	18.08	19.81	17.69	12.24	18.21	12.51	17.63	17.60
	N	9	12	10	10	11	9	30	31
Total Nodes – total documents visited	Mean	95.67	103.33	115.60	68.60	116.91	120.22	110.10	97.03
	SD	45.56	48.83	52.61	33.94	44.24	52.87	46.90	49.11
	N	9	12	10	10	11	9	30	31
Step Size – average VE units travelled between nodes	Mean	.3479	.3269	.4039	.4064	.4248	.4276	.3948	.3818
	SD	.0661	.0290	.0591	.1006	.0557	.0604	.0664	.0797
	N	9	12	10	10	11	9	30	31
Step Angle – average angle between consecutive nodes	Mean	.7282	.6092	.6504	.8713	.5487	.5711	.6365	.6827
	SD	.8033	.3091	.7758	.6427	.4185	.4878	.6563	.4901
	N	9	12	10	10	11	9	30	31

Table 4-5 Descriptive Statistics for Navigation of High and Low Spatial Visualisation (VZ) Participants in 3D100, 3D80, and 3D60 Environments

4.3.1.2.2 Individual Differences in Cognitive Ability and Number of Dimensions used in Environmental Mapping

In order to examine main effects of individual differences in cognitive ability on behaviour when comparing the number of dimensions used for environmental mapping 2 (Ability – Hi & Lo) x 2 (Dimension – 2D60 & 3D60) ANOVAs were conducted for each of the three DVs, unique documents visited, total documents visited, and lostness. Results showed that high and low group cognitive ability did not differ significantly in behaviour, and no interactions between cognitive ability and dimension occurred. However the interaction between high and low MA and dimension approached significance at the 95% probability level for Lostness, $F(1,37) = 3.61$; $p = 0.065$ (see Figure 4-3 for the interaction plot). Table 4-6, Table 4-7, and Table 4-8 show means and SDs.

Performance Measure		3D60		2D60		Total	
		Lo MA	Hi MA	Lo MA	Hi MA	Lo MA	Hi MA
Lostness – total lostness ratio Smith (1996)	Mean	.8253	.7531	.8520	.9177	.8374	.8484
	SD	.0795	.1636	.0921	.1244	.0845	.1612
	N	12	8	10	11	22	19
Unique Nodes – total documents visited once only	Mean	60.08	54.00	67.90	80.00	63.64	69.05
	SD	13.51	18.82	15.29	20.04	14.55	23.13
	N	12	8	10	11	22	19
Total Nodes – total documents visited	Mean	123.83	110.25	139.60	197.45	131.00	160.74
	SD	37.20	60.80	45.37	93.80	40.89	91.01
	N	12	8	10	11	22	19

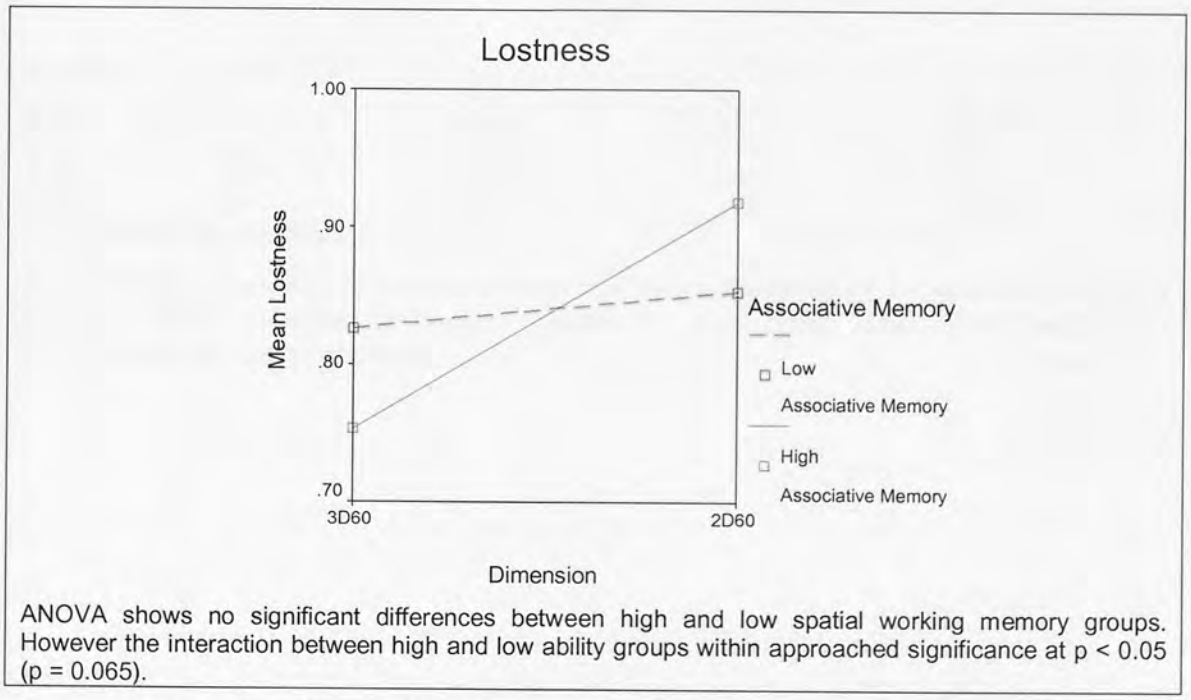
Table 4-6 Descriptive Statistics for Navigation of High and Low Associative Memory (MA) Participants in 2D60 and 3D60 Environments

Performance Measure		3D60		2D60		Total	
		Lo MV	Hi MV	Lo MV	Hi MV	Lo MV	Hi MV
Lostness – total lostness ratio Smith (1996)	Mean	.8005	.7903	.8684	.8974	.8276	.8566
	SD	.1110	.1444	.0831	.1295	.1041	.1421
	N	12	8	8	13	20	21
Unique Nodes – total documents visited once only	Mean	56.50	59.37	74.25	74.23	63.60	68.57
	SD	15.81	16.37	12.96	21.82	16.92	20.83
	N	12	8	8	13	20	21
Total Nodes – total documents visited	Mean	117.42	119.88	147.75	183.54	129.55	159.29
	SD	42.23	56.50	42.93	93.45	44.09	85.80
	N	12	8	8	13	20	21

Table 4-7 Descriptive Statistics for Navigation of High and Low Spatial Working Memory (MV) Participants in 2D60 and 3D60 Environments

Performance Measure		3D60		2D60		Total	
		Lo VZ	Hi VZ	Lo VZ	Hi VZ	Lo VZ	Hi VZ
Lostness – total lostness ratio Smith (1996)	Mean	.7962	.7966	.8828	.8911	.8414	.8439
	SD	.1150	.1368	.1170	.1133	.1217	.1312
	N	11	9	12	9	23	18
Unique Nodes – total documents visited once only	Mean	55.64	60.11	70.25	79.56	63.26	69.83
	SD	18.21	12.51	20.75	14.60	20.54	16.55
	N	11	9	12	9	23	18
Total Nodes – total documents visited	Mean	116.91	120.22	165.67	175.56	142.35	147.89
	SD	44.24	52.87	72.52	90.62	64.34	77.40
	N	11	9	12	9	23	18

Table 4-8 Descriptive Statistics for Navigation of High and Low Spatial Visualisation (VZ) Participants in 2D60 and 3D60 Environments



ANOVA shows no significant differences between high and low spatial working memory groups. However the interaction between high and low ability groups within approached significance at $p < 0.05$ ($p = 0.065$).

Figure 4-3 Interaction Plot of Mean Lostness for High and Low Associative Memory (MA) Participants in 2D60 and 3D60 Environments

4.3.2 Browsing Pattern Using N-Gram Analysis

As explained in Section 4.2.4.2 two methods employing n-gram based analysis of browsing patterns were examined; 'exact-match' and 'any match'. In the exact-match method an n-gram was considered co-occurring across participants if the same object identifiers appeared in the same sequence, however for the 'any match' method an n-gram was considered co-occurring if the same identifiers appeared regardless of sequence. These two methods were compared by firstly correlating the cosines produced by each of the n-gram lengths for participants in each of the environments, and then by calculating means and SDs of cosines for each method and employing Wilcoxon analysis to test the significance of observed differences. Table 4-10 and Table 4-11 show the results of these analyses. In both analyses non-parametric models were used as cosine values were not normally distributed.

Condition	n-gram length			
	2g	3g	4g	5g
3D100	.932**	.71**	.402**	.26**
3D80	.926**	.662**	.33**	.233**
3D60	.935**	.654**	.386**	238**
2D60	.95**	.870**	.878**	.572**

All correlations are significant at ** $p < 0.01$

Table 4-9 Spearman Rho Correlations between Cosines Produced by Exact Match N-Grams and Any Match N-Grams for N-gram Lengths 2 – 5 in 2D60, 3D60, 3D80, and 3D100 Environment Mapping Conditions

Condition	Exact Match		Any Match		Wilcoxon
	Mean	SD	Mean	SD	Z value
2g 3D100	.1383	.0674	.2173	.0978	-12.56**
2g 3D80	.1105	.0626	.1760	.0932	-11.88**
2g 3D60	.1261	.0607	.1942	.0836	-11.94**
2g 2D60	.0928	.1133	.1316	.1571	-12.15**
3g 3D100	.0223	.0225	.0793	.0502	-12.2**
3g 3D80	.0134	.0160	.0483	.0393	-11.1**
3g 3D60	.0171	.0160	.0599	.0357	-11.77**
3g 2D60	.0229	.0391	.0539	.0836	-10.06**
4g 3D100	.0029	.0068	.0285	.0287	-10.95**
4g 3D80	.0010	.0041	.0129	.0184	-8.44**
4g 3D60	.0020	.0056	.0159	.0172	-9.88**
4g 2D60	.0061	.0131	.0219	.0392	-7.57**
5g 3D100	.0004	.0019	.0089	.0131	-8.5**
5g 3D80	.0002	.0019	.0039	.0097	-5.36**
5g 3D60	.0002	.0020	.0033	.0071	-6.14**
5g 2D60	.0015	.0056	.0083	.0179	-6.71**

* $p < 0.05$; ** $p < 0.01$

Table 4-10 Descriptive and Inferential Statistics Comparing Exact Match and Any Match N-gram Analyses for N-gram Lengths 2 – 5 in 2D60, 3D60, 3D80, and 3D100 Environment Mapping Conditions

It can be seen in Table 4-9 that as n-gram length increased the strength of correlations between cosines based on exact-match node sequences and any-match node sequences decreased suggesting that while the same documents were leading to each other, increased variation occurred in the direction or order they were visited. Descriptive statistics and Wilcoxon analysis (see Table 4-10) confirmed that the mean cosines produced by any match n-gram analysis were significantly higher than those produced by the exact match method. On this basis the any match method is identifying greater level of shared variance in browsing pattern than exact match.

The results of these analyses support the prediction that the exact-match method of identifying similar browsing patterns is too restrictive to provide useful information regarding the similarity between users' browsing patterns. The excessively large number of possible routes through the environment would make it unreasonable to expect exactly the same routes to be followed, even when using the spatial organisation of the environment, to aid browsing.

However it would be reasonable to expect that if participants were using the spatial organisation of documents to lead them to other relevant documents and their browsing pattern was influenced by this, moving from document 3 to document 4 could be considered the same as moving from document 4 to document 3. Therefore a better representation of shared browsing patterns was considered more likely if the focus of co-occurrence was based on a pattern of visiting the same objects within each n-gram window, rather than the specific order in which the objects were visited during that window. On the basis of this, the analyses conducted on browsing patterns reported in the following sections are based on the "any match n-gram method", which is predicted to give more meaningful results.

4.3.2.1 Environmental Mapping and Browsing Pattern

As with the previous analyses regarding navigation, differences in browsing pattern resulting from environmental mapping were considered from two perspectives i) the quality of mapping, i.e. how much semantic variance is accounted for by the placement of documents, and ii) the number of dimensions used for document layout. These independent variables were analysed using factorial ANOVAs; in Section 4.3.2.1.1 the results regarding quality of mapping are presented and results of analysis of factors relating to dimension are presented in 4.3.1.1.2.

4.3.2.1.1 Quality of Mapping in Terms of Semantic Variance Accounted For

To test the effects of the quality of mapping on individuals adopted browsing patterns a 3 (mapping quality – 3D100, 3D80, & 3D60) x 4 (n-gram length – 2g, 3g, 4g, & 5g) ANOVA was performed in which cosines were the dependent measure. Main effects

of semantic variance, and n-gram length were observed with cosines reducing with both increased semantic variance, $F(2,2348) = 4.16$; $p < 0.001$, and n-gram length, $F(3,2348) = 1636.85$; $p < 0.001$. There was also a significant interaction, $F(6,2348) = 4.78$; $p < 0.001$, as the differences in mean cosine values occurring between mapping quality conditions reduced with n-gram length (see Table 4-10 for means and SDs and Figure 4-4 for means plot).

Post hoc comparisons were performed using Tamhane (Levene's test for equality of variance was not satisfied – $p < 0.05$) for both quality of mapping and n-gram length. It was shown that mean cosines for 3D100 differed significantly from both 3D80 ($p < 0.001$), and 3D60 ($p < 0.01$) conditions. Differences across all levels of n-gram length were significant at $p < 0.001$.

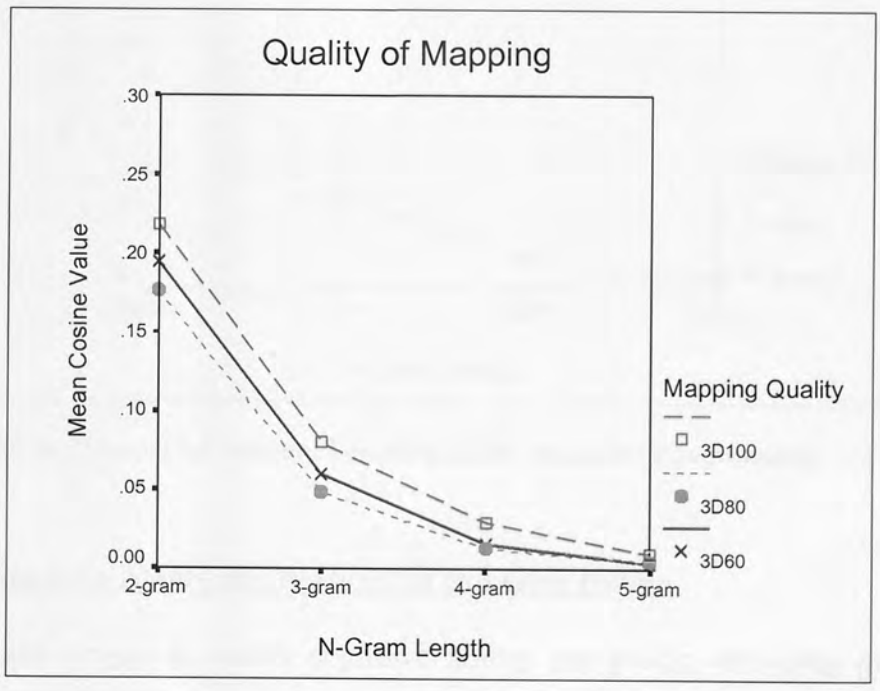


Figure 4-4 Mean Cosines for N-Gram Length in 3D100, 3D80, and 3D60 Environments

4.3.2.1.2 Number of Dimensions used in Environmental Mapping

The effects of the number of dimensions used for environmental mapping and n-gram length on measured browsing pattern were examined using a 2 (dimension – 3D60, 2D60) by 4 (n-gram length – 2g, 3g, 4g, & 5g) ANOVA. Main effects of dimension, $F(1,1592) = 15.2$; $p < 0.001$, and n-gram length occurred, $F(3,1592) = 373.06$; $p < 0.001$. There was also a significant interaction, $F(3,1592) = 19.5$; $p < 0.001$, with mean cosines being greater for 3D60 compared to 2D60 at 2g but no differences occurring as n-gram length increased. Once again mean cosine values reduced with n-gram length (see Figure 4-4 for means plot).

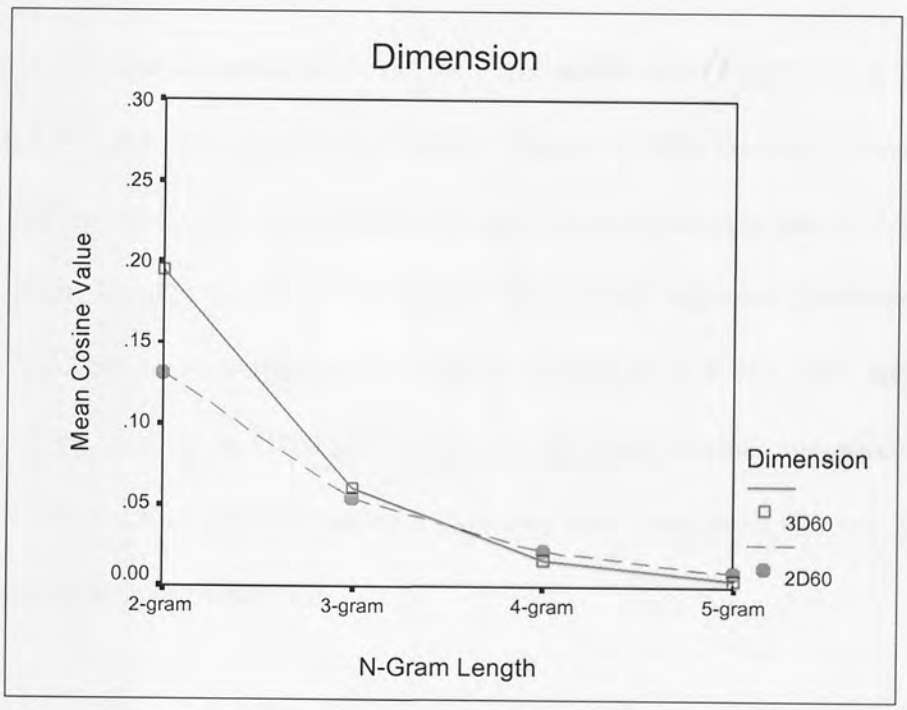


Figure 4-5 Mean Cosines for N-Gram Length in 2D60 and 3D60 Environments

4.3.2.2 Cognitive Ability as a Predictor of Browsing Pattern

To assess the degree to which cognitive ability can predict browsing pattern the difference between individuals' cognitive ability score was correlated with the observed difference in browsing pattern as measured using cosines. It was predicted that individuals with a similar cognitive ability score would also demonstrate similar

browsing strategies (i.e. as the difference score in ability decreases (similar ability) the cosine value for browsing pattern would increase (Cos 1 = exact match; Cos 0 = no similarity)). For each of the cognitive ability measures (MA, MV, and VZ) difference values in test scores were calculated for every unique pair of individuals in each condition. This resulted in 210 pair-wise comparisons for 3D100 and 2D60, and 190 pair-wise comparisons for 3D80 and 3D60 conditions. Within each condition the difference scores for each of the three cognitive abilities were correlated with cosines calculated using 2g, 3g, 4g, and 5g.

4.3.2.2.1 Correlation Analyses

Spearman's Rho (non-parametric) bivariate correlations were conducted as data were not normally distributed. Table 4-11 shows values of Rho between differences in cognitive ability scores and cosine value for each pair-wise comparison of participants for all n-gram lengths in all environments. Significant negative correlations were observed VZ and cosines measured at 2g in 3D100 ($p < 0.01$), MV and cosines measured at 4g, and 5g in 3D80 ($p < 0.01$), and MV and cosines measured at 5g in 3D60 ($p < 0.05$). These correlations were however very weak with rho = - .299 being the strongest coefficient observed.

Environment Mapping	Cognitive Ability	N-Gram Length			
		2g	3g	4g	5g
3D100	MA	.046	.095	.130	.077
	MV	-.007	.002	.049	.064
	VZ	**-.192	-.106	-.119	-.134
3D80	MA	.083	.094	.012	.018
	MV	-.113	-.074	**-.229	**-.149
	VZ	.134	.038	.023	-.048
3D60	MA	.045	.077	.099	.042
	MV	-.055	-.104	-.074	*-.153
	VZ	-.026	-.015	.012	-.057
2D60	MA	.078	-.034	-.016	.017
	MV	.030	-.015	-.036	.046
	VZ	.059	.078	.008	.049

* p<0.05; ** p<0.01: MA = Associative Memory; MV = Spatial Working Memory; VZ = Spatial Visualisation

Table 4-11 Correlations between Cognitive Ability Difference Scores and Cosine Values for N-Gram Lengths 2 - 5 in 3D100, 3D80, 3D60, and 2D60

4.4 Discussion

4.4.1 Navigation

When examining effects of spatial-semantic mapping and individual differences in cognitive ability on behavioural measures of navigation, the general conclusion is that improved spatial-semantic mapping leads to improved navigation, there is an advantage to mapping documents in an additional dimension to provide extra spatial-semantic cues, and that individual differences in cognitive ability do not have an observable effect on behaviour.

Upon closer examination of the results it can be seen that when using three dimensions for document placement, mapping quality has no effect on measures of lostness, the amount of deviation from travelling in a straight line from document to document (step angle), or the total number of documents visited. However as the amount of semantic variance accounted for in the environmental mapping decreases the distance people travel increases relatively systematically, as does step size, and

rotation (see Figure 4-1, and Table 4-1). As differences in the overall size of the environments were accounted for prior to analyses (see section 4.2.4.1) this is clearly an effect of the way people used the spatial-semantic mapping. Where mapping quality is good people travel less distance overall or between individual documents, and they change direction (rotate) in the environment less. These findings partially support those of Westerman & Cribbin (2000b), who found that during a directed search task in an environment similar to that used in the current study, people performed better when increased amounts of semantic variance were accounted for in the spatial mapping. However there is a large discrepancy in findings from the two studies. Firstly Westerman & Cribbin (2000b) demonstrated significant effects of lostness, total number of objects visited, and number of different objects visited. In the current study while the trend for these three measures is comparable i.e. people became more lost and visited more nodes as mapping quality decreases, only differences between mapping condition for the number of unique nodes visited were significant. Secondly, and more importantly one of the major findings of Westerman & Cribbin (2000b), was that users performed less well in three-dimensional environments and an observable advantage for navigation in 2D existed. The authors found that in order for 3D to facilitate comparable navigational performance an additional 30 – 50% variance would need to be accounted for within the mapping. In the current study the opposite effect was observed whereby 3D demonstrated significant advantages compared to 2D when the same amount of variance was accounted for in each environment.

A possible explanation for lack of significant findings in the current study for lostness and nodes visited when examining effects of mapping quality is the difference in the

nature of the task. The current task involved 'browsing' for information so users were free to explore the environments and develop their own cognitive maps. If as hypothesised users develop a cognitive map more quickly when 'goodness of fit' between the spatial-semantic mapping in the environment and the users' cognitive model is high, they will be able to identify the semantic structure within the environment which will facilitate acquisition of a global model. As Dillon (2000) suggests users' cognitive models of spatial semantics impacts on their ability to derive shape from the environment, where spatial-semantic mapping is poor this ability is impeded. In the task conducted in Westerman & Cribbin (2000b) an advantage for good quality spatial-semantic mapping existed supporting this, however the nature of the task (i.e. a timed trial in which participants had to locate a specified target – between each presentation the starting position was reset) would not allow users sufficient access to develop a good cognitive map of the overall environment structure. In the current study once a user has found a relevant document provided they have developed a good model of the semantic structure of the environment they are likely to remain close to that point and visit all documents present on screen (i.e. in their current field of view – FOV), distance, rotation, and inter-document step size would therefore be lower. When the model is poor however individuals would need to forage more across the entire information space. This pattern of behaviour explains why users in 3D60 visited significantly more unique nodes, but measures of total nodes visited and lostness did not reach significance. If it is assumed that individuals explored the environment more trying to find relevant documents due to a lack of spatial-semantic cues to document position, they may have had a general idea of their own position and therefore not made significantly more repeat visits to nodes, but could not develop a suitable semantic map of the environment to aid relevant

document location. In the optimal mapping condition users may have had more repeat visits to nodes which were in positions where they expected relevant documents to be, in order to satisfy themselves they had not missed any. It appears people may adopt different browse strategies dependent on how 'comfortable' they are in the environment. When coherence between internal and external models is high users adopt a more rigorous strategy in terms repeat visits to documents they expect to be relevant ensure nothing has been missed. When coherence is low however, users concentrate on checking all documents but in a less rigorous way.

Two explanations are offered for the observed differences in effects of dimension between the two studies. Again these explanations relate to the nature of the task. For dimension, lostness, total nodes visited, and unique nodes visited all demonstrated significant effects, however these were not in the direction expected – Westerman & Cribbin (2000b) found an advantage for 2D, the current study found an advantage for 3D. The first explanation is related to cognitive load. Westerman & Cribbin (2000b) suggested the poorer performance in 3D was a result of the additional cognitive load experienced through having to navigate an extra dimension, however browsing tends to be a task which does not place a high cognitive load on users (Marchionini & Shneiderman, 1988). The combination of a timed trial, and lack of facility to explore the environment in the aforementioned study, would place a high cognitive load on memory, the addition of an extra dimension would undoubtedly increase this load creating a disadvantage for 3D. However, in the current study where users have time to browse, the additional dimension could be used to extract additional semantic information as intended without increasing cognitive load sufficiently to impede navigation – in other words a positive trade-off between additional usable semantic

information and cognitive load was achieved. An alternative explanation is that in 2D users adopt a different strategy such that with minimal time pressure, the freedom to browse, and only two-dimensions to travel they implement a more thorough and exhaustive search visiting as many documents as possible. This would account for observed differences in total nodes visited, and unique nodes visited. The effect of increased lostness in 2D is not fully accounted for however, as repeat visits to nodes would not generally be expected. It could be argued that in using this type of strategy no attempt is made to model the environmental layout which could result in disorientation leading to repeat visits. Examination of users browsing patterns examined in section (4.3.2) may help to distinguish between these explanations.

The lack of effects due to individual differences in cognitive ability is more difficult to explain. The explanations offered in Chapter Three section 3.4.3 apply equally well to behaviour. When browsing information spaces users may not need to rely upon specific cognitive abilities as the nature of the task is not cognitively demanding, it is arguable that observable benefits of high cognitive ability only occur when tasks are cognitively demanding such as the timed directed search task employed by Westerman & Cribbin (2000b).

4.4.2 Browsing Pattern

This element of the study set out to examine user behaviour during database browsing with respect to 'browsing patterns' i.e. the path of travel through the environment based on the order in which documents were accessed. It has been shown in the literature that individuals patterns of document access in VE SDMSs can be identified and visualised in order to derive informative conclusions about the way people use

spatial-semantic mapping (e.g. Witmer, Bailey, Knerr, & Parsons, 1996). In this study a method of quantifying the degree of overlap between individuals' browsing patterns (n-gram analysis) was tested along with hypotheses regarding the way spatial-semantic mapping, the number of dimensions used for VE presentation, and individual differences in cognitive ability relate to users' identified patterns. The results are generally encouraging and demonstrate that shared patterns of document access displayed by users in a VE SDMS can be successfully measured using short n-grams. It was also shown that when quality of semantic mapping is high within 3D information spaces, users tend to adopt more similar browsing patterns, suggesting their choices of document access are driven by the spatial placement of documents and the amount of identifiable semantic structure available through this mapping. When documents were mapped in two dimensions a lower level of overlap was observed but where patterns were shared they remained consistent over longer sequences of document visits. Reasons for this are discussed presently. With respect to individual differences in cognitive ability it was found that in three-dimensional environments a small degree of shared browsing pattern can be predicted by spatial ability, spatial memory, and associative memory but findings were erratic, and it is suggested that more investigation is needed before inferences could be reliably drawn.

In terms of the how to implement n-gram analysis of browsing patterns it was considered necessary to examine two possible methods – exact match and any match n-grams (see section 4.2.4.2 for an explanation of the differences). It was expected that due to the large number of possible routes through the environment an exact match n-gram analysis would not identify sufficient salient elements of shared patterns to allow useful conclusions regarding the way in which spatial-semantic

mapping drives individuals' pattern of browsing. For example if someone visits document A first but then chooses to travel to document C due to its content and spatial positioning in relation to document A, it could be argued that the spatial-semantic mapping has driven their strategy to the same degree as someone who started at document C and moved to document A.

Initial comparisons of the two methods supported this assumption with the exact-match method identifying significantly less shared browsing pattern than any-match n-grams (see Table 4-10). However with shorter n-grams both methods were highly correlated demonstrating a high level of agreement in identifying browsing patterns. Although any-match n-grams could identify greater levels of shared browsing pattern (higher mean cosines) both measures were identifying the same proportional degree of variation in browsing patterns e.g. at 2-gram in the 3D100 environment exact-match n-grams could identify 86% of the browsing patterns identified by any match n-grams (see Table 4-9). For 3D conditions this level of agreement declined rapidly and systematically with increased n-gram length as would be expected e.g. at 5-gram in the 3D100 condition only approximately 7% variance was accounted for between the two methods.

What is of particular interest in the comparison of the two methods is that when only two-dimensions were used for mapping the environment correlations between the exact-match and any-match did not decrease as rapidly with n-gram length, at 2-gram 90% shared variance was accounted for and at 5-gram 27% shared variance remained. At 3-gram, and 4-gram 76% and 77% shared variance were accounted for respectively. This would suggest that browsing patterns within the 2D environment

remained fairly stable, and that although any-match n-grams did identify a greater degree of overlap in individuals browsing patterns, to a large extent people tended to maintain the same pattern of document access across a larger number of documents. This would support the explanation given in section 4.4.1, that in 2D users follow a more exhaustive pattern of document access, moving consecutively from one document to the next.

Having determined any-match n-grams were better at identifying an overlap between users browsing patterns further analyses used this method. The results showed a significant effect of mapping quality in 3D conditions, whereby a higher level of shared browsing pattern was exhibited when spatial-semantic mapping was optimal. As the degree of semantic structure accounted for in the spatial mapping decreased, the degree to which users adopted similar patterns of document access decreased. This finding supports the contention that users browsing behaviour is driven by the spatial-semantic structure of the VE when sufficient information is available to assimilate the semantic structure of the environment to their own cognitive models. In the 2D condition users did not generally adopt similar browsing patterns to the same degree as was evidenced in 3D conditions – 3D60 demonstrated higher cosines i.e. higher levels of shared browsing pattern than 2D60. However the interaction between dimension and n-gram length was significant showing that in 2D where users did adopt similar browsing patterns they maintained the pattern over a longer sequence of document visits. Once again it can be argued that while browsing patterns between users tended to be more randomly generated overall in 2D (i.e. decisions of which documents to visit or where to begin browsing were not based on semantic structure), the nature of users' patterns was more sequential, supporting an exhaustive browsing

strategy based on environmental organisation rather than semantic structure.

With regard to predicting browsing patterns on the basis individual differences in cognitive ability, some significant correlations were observed but these were erratic. It is suggested therefore that the reliability of findings should be viewed with caution. It should also be noted corrections for multiple comparisons were not applied as the measures were inter-related and correlations were significant at $p < 0.0005$, however this does raise the possibility of Type 1 error. Having prepared the reader for caution the results are discussed and tentative explanations offered.

Spatial visualisation was shown to be predictive of browsing pattern when maximum semantic structure was portrayed (3D100 environment) such that individuals with similar levels of ability tended to adopt similar browsing patterns. This was only evident however when n-grams were measured using 2-grams. Similarities in spatial memory capacity were also shown to predict browsing pattern, but only when spatial-semantic mapping was not optimal and when patterns were measured with longer n-grams i.e. 4-gram and 5-gram in 3D80, and 5-gram in 3D60. It could be argued that the effect being limited to longer n-grams is due to the specificity of browsing pattern identified as these lengths. In other words only a very small degree of shared variance in browsing pattern is accounted for at these lengths, but it could be argued that where there is an overlap all be it very small it is due to similarities in spatial memory. In environments where spatial-semantic mapping is poor (3D80, and 3D60) it could be argued people must rely more heavily on their spatial memory and as such their browsing patterns may be influenced by this reliance. This relationship however may only be observable at longer n-gram lengths due to the effects of multiple factors

accounting for more of the shared variance at shorter n-grams. It is recognised this explanation is very speculative and conclusions should not be drawn until further investigation is carried out. Associative memory was not shown to have any influence on browsing behaviour.

It should be noted that in a conference paper based on preliminary results from this work (Collins & Westerman, 2001), the relationship between spatial ability and browsing pattern in 3D100, while decreasing in strength persisted across 2, 3, and 4-grams. Spatial memory was also shown to have an effect at 3-gram in 3D80. In all cases cosines values were slightly higher. Since publication of that data the way in which the record of document access was submitted to n-gram analysis has been reviewed. As explained in section 4.2.4.2, the software used during task performance recorded a visit to a document when the cursor was pointed at the document, when the document was selected, and when the document was de-selected, this would lead to a string of consecutive visits being recorded for a single node, in n-gram analysis this would form a pattern of browsing. For the analyses reported in Collins & Westerman (2001), duplicate consecutive visits resulting from the cursor slipping from the node during reading were removed, but duplicates resulting from selecting or de-selecting documents were retained. However with further consideration it was decided for this study to remove all consecutive duplicates. It was felt that the influence these additional sequences had on cosines produced by n-gram analysis reflected users' choice of document relevance rather than their patterns of browsing. In addition for the current study the patterns of browsing recorded between selection of first and last relevant documents were used, for the earlier publication patterns of browsing recorded between start and end of the session were used, explanations of reasons for

this have been presented in 4.2.4.

4.4.3 Conclusions

The results of the study presented here have confirmed the experimental hypotheses in terms of environmental organisation. The degree of spatial structure accounted for by spatial mapping within a VE SDMS (i.e. quality of spatial-semantic) has a significant influence on browsing behaviour both in terms of navigational behaviours (e.g. distance travelled, degree direction of travel changes, the number of unique documents visited etc.), and browsing patterns (sequence documents are visited). Navigational performance has been shown to improve when additional information regarding the semantic structure of the information space is available through mapping. In addition, and contrary to previous research this advantage extends to using three-dimensions rather than two for environment organisation. This is apparently due to the current task, i.e. browsing, being less cognitively demanding than the specific search tasks used in other studies. The benefits associated with the additional semantic information that can be presented by using 3D do not appear to be outweighed by the additional cognitive demands of navigation. However comparisons of 2D and 3D environments when the same degree of semantic variance is accounted for (i.e. 60%) also demonstrate an advantage for 3D which suggests differences in navigational performance are more likely due to differences in strategy with users in 2D employing a more exploratory strategy visiting all documents sequentially with no regard for spatial-semantic mapping. This latter explanation is supported by analysis of browsing patterns which shows users in 2D generally adopt less similar browsing patterns but when similarities do exist they exist across a longer sequence of document visits. N-gram analysis of the sequence in which documents are accessed

has proven a successful measure of browsing pattern similarity which has in turn demonstrated the quality of spatial-semantic mapping can influence the patterns of browsing people use.

Previous research has demonstrated VE simulations of real-world environments can prove difficult to navigate, but that with experience route, and survey type knowledge develops (e.g. Ruddle, et al., 1997) that facilitate VE navigation comparable real world navigation. It has been suggested that the development of survey-type knowledge is sufficient to promote this Ruddle, et al. (1999). In order to determine whether route and survey knowledge of the environment is as important to navigating VE information spaces, the study reported in Chapter Five examines the effects of individual differences in preferences for route and survey based way-finding strategies. The principle focus of the work in Chapter Five is to examine the relationship between behaviour and performance given that spatial-semantic mapping has been demonstrated to be positively influential to both aspects of a successful outcome when browsing for information.

The purpose of the studies in this thesis is to exclusively examine spatial-semantic mapping as a tool for IR in visual IR systems. The combined results of Chapters Three and Four support spatial-semantic mapping as a useful tool for enhancing IR in SDMS VEs. Future work should therefore be directed at examining the how additional navigational tools such as landmarks may enhance navigational efficiency.

5 The Relationship between Behaviour and Performance: And More Individual Differences

5.1 Introduction

The experiment reported in this Chapter expanded upon the work presented so far. This has demonstrated that spatial-semantic mapping can improve information retrieval from virtual information visualisation systems (VIRIs). The benefits exhibited are not dependent on other visualisation tools (e.g. colour coding and physical links), and can be measured in terms of retrieval performance (e.g. relevant documents retrieved – see Chapter Three), and behaviour (e.g. navigation and browsing patterns – see Chapter Four). While variables that can be allocated to these two factors (performance and behaviour) are well researched within the information visualisation literature they are rarely explicitly examined in separate contexts (see Chapter Four). Moreover research into the relationship between the behavioural aspects of IR and measures of retrieval performance is extremely limited. The study reported here aims to investigate the existence of such a relationship, while at the same time further exploring the role of individual differences in cognitive ability (see sections 5.1.2 and 5.1.1). As previous studies had demonstrated a positive effect for good quality of spatial-semantic mapping this aspect was not re-investigated. The optimally mapped environment 3D100 was used in the current experiment.

5.1.1 Individual Differences in Cognition Re-visited

The effects associated with three cognitive abilities were examined in the previous chapters; these were i) associative memory, ii) spatial working memory, and iii)

spatial ability, and reasons for identifying these as pertinent to browsing in a spatial-semantic SDMSs were discussed. It was shown that while evidence supporting the influence of associative memory in IR performance exists within the literature, when 'browsing' for information in the SDMS employed here, no such effect occurs in three-dimensional environments for either retrieval performance or behaviour. It was decided on this basis not to pursue investigation of associative memory. The results of analyses of the effects attributable to spatial memory and spatial ability were not so well defined and partial influences were observed. For performance, differences between users with high or low levels of spatial memory for time on task were exhibited together with an effect of spatial visualisation on recall that approached significance. The interactive effect on recall between quality of semantic-mapping and spatial ability, was also significant such that when spatial mapping optimised presentation of the semantic structure of the environment differences in levels of spatial ability did not influence performance, but when the semantic structure of the environment was not well visualised spatial ability had differential effects on levels of recall. In terms of behaviour none of the cognitive abilities examined displayed significant effects on measures of navigation. For browsing pattern however results demonstrated that similarity in browsing patterns could be predicted on the basis of spatial working memory when less than optimal levels of semantic variance were accounted for by object placement (i.e. 80% and 60% semantic variance accounted for by spatial mapping) and when measured using n-grams four and five characters long.

As previously discussed these results do not generally support those reported in the literature and are relatively difficult to interpret. In some respects this is not unexpected as previous research demonstrates conflicting results in terms of the

cognitive abilities examined (see Chapters Three and Four). It was decided to retain both spatial working memory, and spatial visualisation as measures of cognitive ability to determine the reliability of the results in terms of replicability, and where possible expand upon the conclusions reached in Chapters Three and Four. It was recognised that while the sample size in the previous experiment was not particularly small (n = 20 or 21 per condition), a larger sample size is preferable when examining individual differences in cognitive ability due to the large degree of variance within the population. In the current study the sample size was more than doubled and therefore expected to produce more reliable findings.

It was also decided at this stage to examine the effects of verbal ability in addition to spatial abilities (e.g. visualisation and spatial memory). As intuitively expected, studies examining information visualisation and user performance tend to focus on spatial abilities and memory due to the visual nature of the interfaces (see Chen & Rada, 1996; and Chen & Macredie, 2000). However, given that the visual representation of the database conveys semantic knowledge it is arguably remiss to overlook the possibility of verbal ability at the very least interacting with spatial ability. For example Barshi & Healy (2002), found the mental representations people generate from verbal instructions for navigating a computerised space, were dependent on an interaction between verbal and spatial memory. Another reason for including verbal ability as an experimental variable was the inclusion of 'reading time' and 'travel time' as strategy measures. Although participants were matched for speed of comprehension, it was considered worthwhile investigating the effect of differences in verbal ability on the quality of the mental models users derived from the text based semantic information and the differences in reading/travel strategies they

adopted. As will be explained in section 5.1.2, comparisons based on the amount of time users spent reading documents proportional to the amount of time spent travelling between documents were employed as measures of browsing strategy. Given this distinction it was hypothesised differences in verbal ability would be observed within this measure.

5.1.1.1 Preferred Wayfinding Strategy

As discussed in Chapter Four research regarding navigating VEs has shown that the 'landmark, route, survey knowledge' model applied to real world navigation (e.g. Thorndyke & Stasz, 1980) also applies to VEs (e.g. Darken & Sibert, 1996b; and Ruddle, et al., 1997). Dillon (2000) suggests that this model should be considered applicable to navigation in VE information spaces as users build cognitive maps of an information space based on a spatially driven conceptual model of semantics. Dillon (2000) asserts however, the traditional model of developing navigational knowledge based on a hierarchical or sequential acquisition of landmark, route, and survey knowledge is over simplistic and various factors including the nature of the environment, the task involved, and individual differences between users impact on how and when these types of knowledge are extracted and used. Survey knowledge has been shown to relate to visual-spatial ability (e.g. Cutmore et al., 2000) and that the use of either route or survey based wayfinding strategies is a user-based preference determined by their cognitive processes. Gender differences have been shown to exist in which females demonstrate preference for using a route-knowledge based strategy, and males a survey-based strategy (Lawton, 1994). It is argued this preference is related to males' superior spatial ability particularly in terms of mental rotation and spatial perception. Lawton (1996) demonstrated that the same strategies used for

outdoor wayfinding, were applied to indoor wayfinding and previously identified gender differences existed. Cutmore, et al. (2000) confirmed females demonstrated a disadvantage in navigating a VE and that when survey-knowledge had to be relied upon high spatial visualisation ability aided navigation.

In the current study it was decided to examine differences between IR retrieval performance and behaviour based on users' preferred wayfinding strategies. A measure of route strategy and orientation (survey) strategy was taken using the wayfinding scale developed by Lawton (1996). The scale is measured by self-report responses on a Likert scale to a series of questions relating to way finding behaviours that individuals would use. Participants are asked to recall an experience of having to find their way for the first time. Statements of wayfinding activities are presented and participants are asked to provide responses on a Likert scale regarding how true each statement is true of their behaviour during the recalled incident. The questionnaire provides scores on the two scales, route and orientation strategy, but does not distinguish between having a preference for one or the other, individuals may score high or low on both scales. In this study the two strategies were examined independently of each other and individuals were deemed as either "high survey" or "low survey" preference, and "high route" or "low route" preference, post hoc on the basis of a median split. It was hypothesised differences between high and low groups would be significant in terms of both retrieval performance and browsing behaviour.

5.1.2 The Relationship between Behaviour and Performance

The current experiment aimed to explore any identifiable relationship between performance and behaviour, using the variables applied to measures of retrieval

performance (Chapter Three) and browsing behaviour (Chapter Four), in the previous experiment. As previously stated the literature tends to examine both these aspects of user interaction with IR systems in parallel. When behaviour is examined separately it is often aimed more specifically at issues related to how people use the environment and methods of overcoming problems associated with navigation, rather than identifying links between how people use the environment and how successful they are at retrieving information. For instance Chen, et al., (2002) explored the navigational paths and browsing strategies people employ when seeking information in a VE SDMS, however their focus of interest was 'social navigation' and whether users could learn to navigate the environment by following well used or optimal pathways (see Chapter Four for a more detailed review). Witmer, Bailey, Knerr, & Parsons (1996) however examined users' information seeking behaviour in terms of node access based on environmental cues and compared the effectiveness of identified strategies in terms of both navigational efficiency and retrieval success. The environment they used employed spatial-semantic mapping with the placement of objects calculated using minimum spanning trees (MSTs) (see Chen, et al., 2002) and Chapter Four for an explanation of MSTs). MSTs provide pathways between document nodes that share the most salient semantic features. Three types of nodes are identified, i) extremities which have a single pathway leading to/from them, ii) thread nodes which link to two documents, and iii) branch nodes which have pathways leading to/from three or more nodes. The authors tested hypotheses regarding optimal navigation strategies and retrieval performance, but found that users did not generally adopt the browsing patterns that were expected. For instance 'chaining activity' (moving to a node directly linked to the current node) was expected to occur more frequently when the current node represented a relevant document, this was not the

case however and the degree of chaining from non-relevant nodes was more predictive of overall performance than chaining from relevant nodes – unfortunately in this paper (Cribbin & Chen, 2001) it is not made clear what dependent measures of retrieval were used. The results from the study demonstrated the difficulties associated with identifying effective navigation behaviours with respect to retrieval performance. In this instance using links from non-relevant nodes was more predictive of retrieval performance however this is counter intuitive as these links should in theory lead to other non-relevant nodes. It was suggested that the results may have been due to poor or unexpected semantic mapping, or a by-product of people chaining from non-relevant nodes proportionally more frequently than from relevant nodes. Despite specific hypotheses being rejected the study does confirm a link between navigation behaviour and retrieval outcomes.

The aim of the current experiment was to not to identify optimal browsing behaviours, but to identify which behavioural measures can be used to predict retrieval performance, and whether individual differences in cognitive ability and preferred wayfinding strategies impact on this.

5.1.2.1 Reading Time versus Travel Time

Two additional measures of behaviour were introduced in this current experiment, these were amount of time spent reading documents, and amount of time spent travelling between documents. A third variable based on the ratio between these two measures was also included. For the experiment reported in Chapter Four these variables had not been considered, however there is evidence to suggest that the amount of time spent reading when compared to the amount of time spent deciding

which documents to view, reflects differences in individuals' browsing strategies (e.g. Toms, 2000). It was therefore felt that a valuable measure of user behaviour had been overlooked. Toms (2000) demonstrated that users with no specific retrieval goals made quick decisions about which articles to read, spent proportionally more time reading document contents, and tended to prefer unstructured environments, when browsing a text based IR system that allowed them to select documents using either 'navigation' menus, or 'items to browse' lists. This latter finding may be due to the degree to which these types of IR systems impose a structure inconsistent with the users' cognitive model. In such cases interference from the imposed structure would likely produce greater cognitive load, than no structure. It has been suggested previously in this thesis (Section 3.4), that when spatial-semantic mapping is employed for information visualisation, and mapping accounts for maximum semantic variance, a goodness of fit between the system and the user's cognitive model is achieved, thus reducing cognitive load. It has also been argued that cognitive load is influential in IR activity and the way in which individual differences in cognitive ability facilitate this (see Chapter Four and Westerman & Cribbin (2000b)). Given that the database presentation in the current study used the optimal spatial-semantic mapping solution it was hypothesised that individuals who benefited from the mapping would be able to make judgements about which documents to view more quickly by using their spatial position and spend more time reading. On the basis of results from Toms (2000) it could be argued that this strategy represents users' preferred browsing behaviour, and being able to adopt a preferred behaviour would facilitate improved retrieval performance.

5.2 Methodology

5.2.1 Participants

Initially 48 participants completed the experiment however data for one person were removed. Post experimental discussion revealed the participant had misunderstood the nature of relevant documents and believed that all documents involving danger should be recovered – see section 5.2.4 for clarification. Of the 47 participants whose data were retained 40 were female and 7 were male. Participants were aged between 18 and 46 years, with a mean of 19.55 and SD 4.3.

5.2.2 Experimental Design

One-way ANOVAs were used to examine the effects of individual differences on users' performance and behaviour in terms of navigation. The independent variables were cognitive ability or preferred wayfinding strategy (see section 5.2.2.1 for details of independent measures) with 2 levels (high or low) based on a median split of scores from the relevant measures. Correlation was used to examine i) the relationship between cognitive measures and behaviour in terms of browsing pattern, and ii) to identify any relationship between navigation behaviour and performance.

5.2.2.1 Independent Variables

Results for the previous studies (Chapters Three and Four) demonstrated very little effect of cognitive ability in terms of either performance or behaviour. However as detailed in section 5.1.1 spatial ability (VZ) and spatial working memory (MV) demonstrated limited effects. It was decided therefore to re-examine these abilities with a larger sample size, to test the reliability of the results.

Three new independent measures were introduced that hadn't been examined previously, these were i) verbal ability (V5), ii) route based preferred wayfinding strategy (route), and iii) survey based preferred wayfinding strategy (survey) – see sections 5.1.1.1 and 5.2.3 for details of the measures used.

5.2.2.2 Dependent Measures

Fifteen dependent variables were initially examined, six of these (the same DVs employed in Chapter Three) measured aspects of users' performance, and nine of the variables were measures of behaviour.

The variables attributed to performance were i) recall (R) – the number of retrieved relevant documents divided by the total number of relevant documents within the database), ii) precision (P) – the number of relevant documents selected divided by the total number of documents selected, iii) accuracy (A) – the harmonic mean between recall and precision (F – stat), iv) time on task (TT), vi) mean time on task (MnTT) – the average time taken per relevant document selected (i.e. TT/number of relevant documents selected), and vii) efficiency (E) – based on a speed / accuracy trade-off i.e. $\frac{accuracy(A)}{timeontask(TT)}$. Full details of these variables are given in Chapter Three section 3.2.4. Once again it was found that the distribution of scores for precision was very severely skewed with a skewness value of -2.78, SE .347, and kurtosis value 7.39, SE .681. The mean for P was .95 with SD .12, as the maximum possible value for precision was 1 it was excluded as an individual dependent variable, but was used in the calculation of the F stat.

Behaviour was measured across the two parameters described in Chapter Four, 'navigation' and 'browsing pattern', and a third parameter which is referred to as 'strategy'. Strategy refers to differences in how people completed the task with reference to the degree of time they spent reading the documents, and the amount of time spent travelling between documents. Given that participants were matched for reading speed it was hypothesised that individuals who were able to utilise the spatial-semantic mapping of the documents in the VE would differ in the amount of time spent travelling and reading, and the ratio between the two, and this would impact on performance.

The dependent measures used to examine navigation were comprised of the seven variables detailed in Chapter Four i.e. i) the total distance travelled (distance), ii) average step distance between documents (step size), iii) the total degree to which the environment was rotated (rotation), iv) average step angle between documents (step angle), v) a measure of disorientation (lostness), vi) the total number of nodes / documents visited (total nodes), and vii) the number of unique documents visited (unique nodes). Browsing pattern was measured using cosines based on 2, 3, 4, and 5-gram any match analysis described fully in Chapter Four section 4.2.4.2. Strategy was examined using total time spent reading documents during task completion (reading time), total time spent travelling during the task (travel time), and the ratio $\frac{readtime}{traveltime}$ (read/travel ratio).

All dependent measures were recorded between participants selecting their first document, and their last document.

5.2.3 Materials

Three cognitive ability tests were used; three tests were taken from the "Kit of Factor-Referenced Cognitive Tests" (Ekstrom, et al., 1976). These included VZ-2 (spatial visualisation) and MV-2 (spatial working memory), as described in Chapter Three, and an additional measure for verbal ability: -

V-5 – Verbal ability referenced within this thesis as V-5.

There are two parts to this test each comprising 18 items. Participants are allowed four minutes for each part.

All of the above measures were scored using standard protocols including negative scored to reduce effects of guessing. Split-half reliability measures comparing scores from the two parts of each test were also calculated.

The Wayfinding Strategies Scale: International Version (Lawton, 1994), measured preferred way-finding strategies based on route knowledge and survey knowledge. It consisted two sets of statements to be rated using a five point Likert scale (A – E, scored 0 – 4) no time limit was given. The first set comprised nine statements deemed to be associated with outdoor wayfinding behaviour; the second set comprised eight statements associated with indoor wayfinding. Of the 17 statements, scores from statements 1, 4, 11, 14, 15, and 16 were totalled to give a score of preference for route strategy (maximum score = 24), and statements 2, 3, 5, 6, 7, 8, 9, 10, 12, 13, and 17 (maximum score = 44) were totalled to give a preferred survey strategy score.

WorldToolKit – Release 7 (Sense8 Corporation1997) was used to generate the virtual environment database, which was presented using desktop PCs. In this study a single

mapping solution was used which was 3D100 the environment which was presented using three dimensions, and accounted for the maximum semantic variance (see Chapter Three section 3.2.1).

5.2.4 Procedure

On entering the experiment room participants were welcomed and the purpose of the experiment was explained, along with ethical guidelines and data protection regulations. When participants were satisfied with the explanations and all queries had been addressed, consent forms were completed. The participants were then asked to complete the four timed paper based cognitive ability tests and the measure of way-finding strategy which was not timed (see section 5.2.3).

Participants were then asked to complete a practice task which involved locating and selecting a target (red coloured node) presented in a three-dimensional environment comprising a cubed arrangement of 64 nodes, the 63 non-target nodes were green. Instructions on manoeuvring the environmental platform were given verbally as the experimenter demonstrated and allowed the participant to practice. The practice task comprised 10 blocks of 10 target presentations and provided participants successfully reached the pre-programmed required response time for two consecutive blocks they could continue to the actual experiment. All participants successfully completed the practice trials.

The task for the main experiment mirrored that used for the studies reported in Chapters Three and Four. Participants were required to locate and select as many relevant documents they could find (newspaper articles related to journalists facing

personal risk). The participants were told to stop when they had found all the documents they thought they were going to find provided they had completed a minimum of 20 minutes on task. If participants had not completed the task within 45 minutes they were asked to stop. Once individuals were certain they understood the task they began. Participants were told they were free to ask for help during the experiment if they needed it.

5.3 Results

5.3.1 Individual Differences in Browsing Performance

To examine the effects of individual differences in cognitive ability and preferred way-finding strategy on performance, participants were allocated to one of two levels high or low (based on a median split of total score), each of the five independent measures, verbal ability (V5), spatial working memory (MV), and spatial visualisation ability (VZ), route-based strategy (route), and survey-based strategy (survey). Differences between high and low groups for all IV measures were significant at $p < 0.001$ (see Table 5-1 for descriptive statistics and t-values). The reliability of the test scores for cognitive ability were measured using spearman-brown split-half reliability measures and demonstrated values of 0.54, 0.87, and 0.69 for V5, MV, and VZ respectively. V5 demonstrated a relatively low level of reliability (0.7 is generally considered the acceptable level for reliability). The test used is widely recognised as a reliable measure of verbal ability so analyses were continued, however the low level of split-half reliability is taken into consideration in the discussion of the results.

Cognitive Ability	Group	N	Mean	SD	t-value
Verbal Ability (V5)	Hi	24	12.66	2.60	t(45) = -8.51
	Lo	23	6.31	2.50	
	All	47	9.56	4.09	
Spatial Working Memory (MV)	Hi	24	21.79	1.58	t(24.9) = -7.04
	Lo	23	12.65	6.03	
	All	47	17.32	6.32	
Spatial Visualisation Ability (VZ)	Hi	24	12.57	2.43	t(45) = -9.16
	Lo	23	5.21	3.06	
	All	47	8.97	4.61	
Survey-based Strategy (survey)	Hi	21	20.00	4.32	t(45) = -8.06
	Lo	26	11.19	3.16	
	All	47	15.13	5.76	
Route-based Strategy (route)	Hi	21	21.90	1.38	t(40.79) = -8.79
	Lo	26	16.96	2.43	
	All	47	19.17	3.19	

For each ability measure, Means and SDs are shown for hi/lo ability groups and overall together with t-test statistics. Differences between hi/lo groups for each measure were significant ($p < 0.01$).

Table 5-1 Descriptive statistics of Verbal Ability, Spatial Working Memory, Spatial Visualisation, and Preferred Route and Survey Wayfinding Strategy Scores

Means and standard deviations of performance measures R, TT, MnTT, A, and E are shown for V5, MV, VZ, in Table 5-2, and for route and survey in Table 5-3.

Performance Measure		Lo V5	Hi V5	Lo MV	Hi MV	Lo VZ	Hi VZ
Recall (R) - (proportion of retrieved relevant documents to total relevant documents)	Mean	.5696	.5687	.5696	.5688	.5261	.6104
	SD	.1887	.2058	.2131	.1817	.1751	.2085
	N	23	24	23	24	23	24
Time on Task (TT) - in seconds	Mean	1182.77	1159.35	1213.59	1129.81	1119.68	1219.81
	SD	371.79	335.03	338.17	363.01	381.57	316.80
	N	23	24	23	24	23	24
Average Time on Task (MnTT) – average time Per relevant document selected in seconds	Mean	100.45	110.32	110.44	100.74	101.96	108.87
	SD	34.86	52.82	56.34	30.28	32.02	54.74
	N	23	24	23	24	23	24
Accuracy (A) – F stat	Mean	.6836	.6963	.6782	.7015	.6521	.7264
	SD	.1691	.1713	.1884	.1502	.1496	.1804
	N	23	24	23	24	23	24
Efficiency (E) – (A) divided by (TT)	Mean	.0006	.0006	.0006	.0007	.0006	.0006
	SD	.0002	.0002	.0002	.0002	.0002	.0002
	N	23	24	23	24	23	24

Table 5-2 Descriptive Statistics of Performance for High and Low Verbal Ability (V5), Spatial Working Memory (MV), and Spatial Ability (VZ)

Performance Measure		Lo Route	Hi Route	Lo Survey	Hi Survey
Recall (R) - (proportion of retrieved relevant documents to total relevant documents)	Mean	.5481	.5952	.5596	.5810
	SD	.1797	.2150	.2083	.1827
	N	26	21	26	21
Time on Task (TT) - in seconds*	Mean	1163.85	1179.42	1064.10*	1302.93*
	SD	370.92	330.66	338.69	323.94
	N	26	21	26	21
Average Time on Task (MnTT) – average time Per relevant document selected in seconds	Mean	100.98	111.07	95.38	118.00
	SD	30.87	57.89	32.03	54.98
	N	26	21	26	21
Accuracy (A) – F stat	Mean	.6713	.7133	.6780	.7050
	SD	.1672	.1712	.1779	.1591
	N	26	21	26	21
Efficiency (E) – (A) divided by (TT)	Mean	.0006	.0006	.0007	.0006
	SD	.0002	.0002	.0002	.0002
	N	26	21	26	21

* Differences between high and low survey strategy groups significant at $p < 0.05$

Table 5-3 Descriptive Statistics of Performance for High and Low Wayfinding Strategies (route and survey)

Data for MnTT were not normally distributed overall. Within levels of the IVs: - MnTT was not normally distributed for high V5, high VZ, high route, and high survey; A was not normally distributed for low V5; R was not normally distributed for low VZ, and low survey. Mann-Whitney U tests were performed, for MnTT with all IVs, for A with V5, and for R with VZ and survey. For all other comparisons (i.e. R with V5, MV, and route; A with MV, VZ, route, and survey; E with V5, MV, VZ, route, and survey; TT with V5, MV, VZ, route, and survey) one-way ANOVAs were conducted. All IVs had 2 levels (high and low).

No significant effects of cognitive ability or route strategy were found. However, for TT, differences between survey strategy groups were significant with the high survey group taking longer on task than the low survey group $F(1,45) = 6.004$; $p < 0.05$. Differences for MnTT, and E when comparing high and low survey groups approached significance, $U = 190$, $Z = -1.78$, $N = 47$; $p = 0.076$, and $F(1,45) = 3.376$; $p = 0.073$ respectively; in both measures participants with a low preference for survey strategy outperformed those with a high preference.

5.3.2 Individual Differences in Browsing Behaviour

The three elements of behaviour described in section 5.2.2.2 were examined and are reported in this section: Section 5.3.2.1 reports on navigation, section 5.3.2.2 reports on strategy, and section 5.3.2.3 reports on browsing pattern.

5.3.2.1 Navigation

Differences in navigation within the VE database arising due to differences in cognitive ability and way-finding strategy were examined using V5, MV, VZ, route, and survey as the IVs with 2 levels (hi and lo), and distance, step size, rotation, step angle, lostness, total nodes, and unique nodes as dependent measures (see section 5.2.2.2). Table 5-4 and Table 5-5 show descriptive statistics.

Data for distance, rotation, and lostness were not normally distributed overall and within levels of the IVs; distance was not normally distributed for low V5, low MV, high, and low survey; rotation was not normally distributed for high and low V5, low MV, high and low VZ, high and low route, high and low survey; lostness was not normally distributed for low route. In addition assumptions of normality were violated for angle size in high V5. For between group comparisons of these measures Mann-Whitney U tests were performed, however where data were normally distributed one-way ANOVAs were conducted.

Navigation Measure		Lo V5	Hi V5	Lo MV	Hi MV	Lo VZ	Hi VZ
Distance	Mean	61599.95	54916.93	60578.76	55895.58	61990.59	54542.57
	SD	44695.64	35708.78	41266.84	39606.28	43936.96	36522.84
	N	23	24	23	24	23	24
Step Size	Mean	.8017	.7750	.7906	.7857	.7975	.7790
	SD	.0809	.1222	.0915	.1163	.0995	.1092
	N	23	24	23	24	23	24
Rotation	Mean	2545.48	1954.79	2428.35	2067.04	2718.65	1788.83
	SD	2641.04	1806.34	2805.74	1586.29	2645.36	1725.61
	N	23	24	23	24	23	24
Step Angle	Mean	53.80	52.58	52.61	53.72	51.88	54.42
	SD	6.96	6.46	4.96	8.04	5.93	7.21
	N	23	24	23	24	23	24
Lostness	Mean	.7100	.7110	.7294	.6924	.6908	.7294
	SD	.1444	.1536	.1264	.1660	.1699	.1232
	N	23	24	23	24	23	24
Total Nodes	Mean	87.22	91.42	92.57	86.29	88.39	90.29
	SD	40.20	35.85	31.39	43.34	39.74	36.45
	N	23	24	23	24	23	24
Unique Nodes	Mean	47.35	47.71	51.26	43.96	47.09	47.96
	SD	16.24	12.55	12.26	15.45	15.48	13.42
	N	23	24	23	24	23	24

Table 5-4 Descriptive Statistics of Navigation for High and Low Verbal Ability (V5), Spatial Working Memory (MV), and Spatial Ability (VZ)

Navigation Measure		Lo Route	Hi Route	Lo Survey	Hi Survey
Distance	Mean	68902.98	44920.37	66444.33	47964.41
	SD	47741.99	22513.88	45412.14	30230.61
	N	26	21	26	21
Step Size	Mean	.8089	.7623	.8233	.7444
	SD	.1044	.0995	.1053	.0856
	N	26	21	26	21
Rotation	Mean	2783.19	1576.10	2520.54	1901.29
	SD	2631.15	1459.37	2496.40	1901.84
	N	26	21	26	21
Step Angle	Mean	53.7764	52.4351	53.33	52.98
	SD	7.3808	5.7500	7.36	5.86
	N	26	21	26	21
Lostness	Mean	.7197	.6991	.6980	.7259
	SD	.1485	.1492	.1461	.1515
	N	26	21	26	21
Total Nodes	Mean	90.23	88.29	83.96	96.05
	SD	38.44	37.65	36.24	39.24
	N	26	21	26	21
Unique Nodes	Mean	47.23	47.90	45.19	50.43
	SD	16.15	12.04	15.00	13.19
	N	26	21	26	21

Table 5-5 Descriptive Statistics of Navigation for High and Low Wayfinding Strategies (route and survey)

Results from inferential statistics showed that no significant differences in navigation occurred due to cognitive ability. However differences between high and low route groups were significant for distance, $F(1,45) = 4.48$; $p < 0.04$, whereby users with a

low preference for route based way-finding strategies travelled further. Differences in step size were also significant for survey such that users with a high preference for a survey based strategy took shorter steps between documents than those with a low preference, $F(1,45) = 7.68$; $p < 0.01$ (see Figure 5-1 for graphs of means).

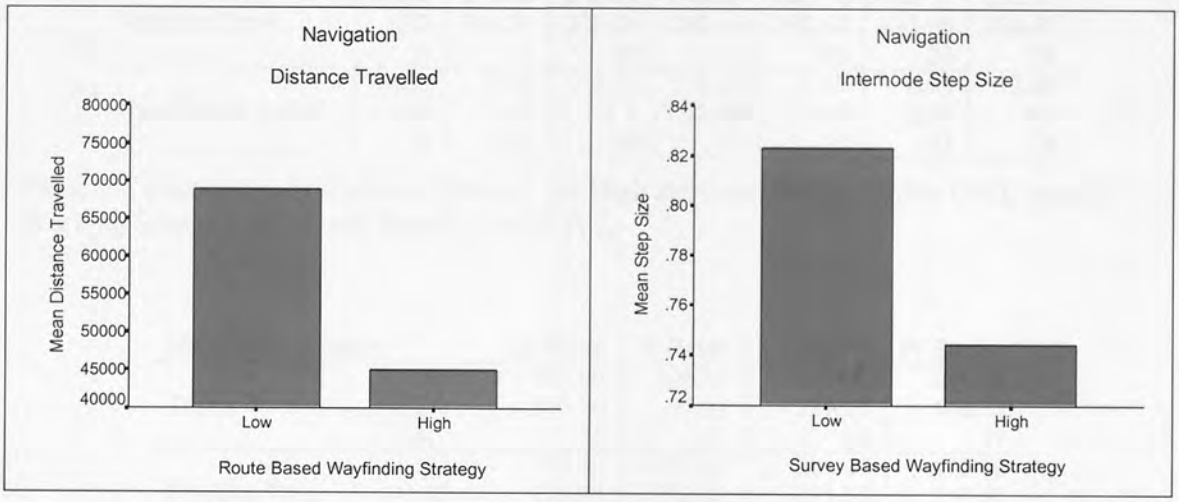


Figure 5-1 Graphs of Means of Significant Effects for Individual Differences on Navigation

5.3.2.2 Strategy

Individual differences in browsing strategy were measured using time spent reading time and time spent travelling and a ratio of reading time to time spent travelling. The independent variables were V5, MV, VZ, route, and survey (see Tables Table 5-6 and Table 5-7). Data for travel time and read/travel ratio across all IVs were not normally distributed. Non-parametric Mann-Whitney U analyses were used for these comparisons and one-way ANOVAs were used to examine effects of V5, MV, VZ, and route based wayfinding strategy on read time.

The only significant effects observed were those of survey based wayfinding strategy on reading time where users with a high preference for using survey knowledge to navigate spent longer reading $U = 134$, $Z = -2.97$, $N = 47$; $p < 0.01$, and read/travel

ratio which was borderline significant U = 183, Z = -1.93, N = 47; p = 0.054 (see Figure 5-2) .

Navigation Measure		Lo V5	Hi V5	Lo MV	Hi MV	Lo VZ	Hi VZ
Travel Time	Mean	390.25	352.63	388.14	354.65	377.65	364.70
	SD	239.41	137.36	224.16	160.55	230.38	153.50
	N	23	24	23	24	23	24
Reading Time	Mean	810.99	819.08	840.58	790.73	762.31	865.73
	SD	314.21	333.04	339.44	306.44	334.66	304.62
	N	23	24	23	24	23	24
Read/travel Ratio	Mean	3.08	2.74	3.13	2.67	2.86	2.94
	SD	2.47	1.9	2.69	1.59	2.49	1.9
	N	23	24	23	24	23	24

Table 5-6 Descriptive Statistics of Strategy for High and Low Verbal Ability (V5), Spatial Working Memory (MV), and Spatial Ability (VZ)

Navigation Measure		Lo Route	Hi Route	Lo Survey	Hi Survey
Travel Time	Mean	407.80	325.53	392.03	345.05
	SD	240.30	97.27	227.25	140.51
	N	26	21	26	21
Reading Time	Mean	769.59	871.50	686.97	973.79
	SD	320.91	318.49	271.52	310.26
	N	26	21	26	21
Read/travel Ratio	Mean	2.86	2.96	2.47	3.44
	SD	2.5	1.76	2.12	2.19
	N	26	21	26	21

Table 5-7 Descriptive Statistics of Strategy for High and Low Wayfinding Strategies (route and survey)

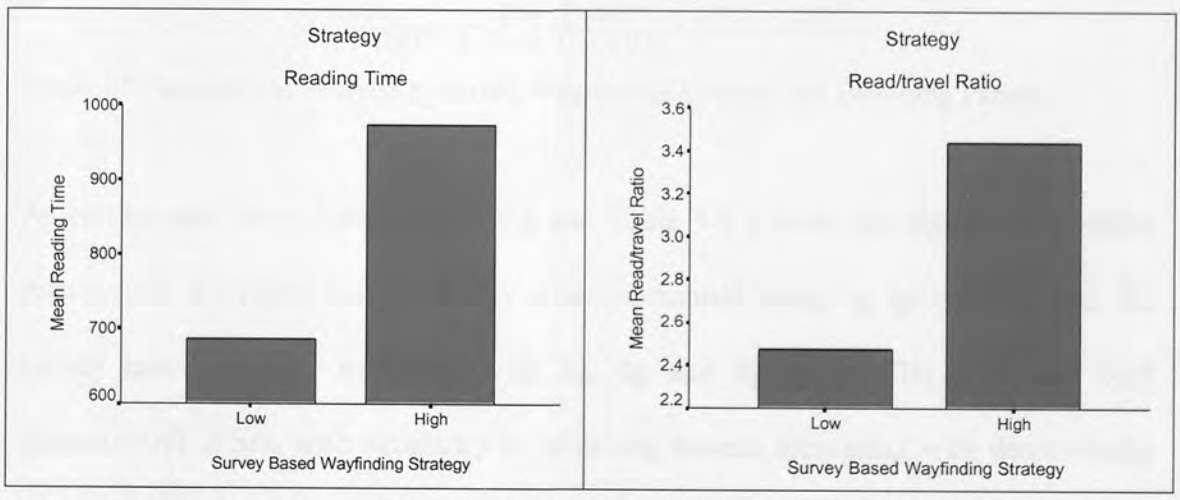


Figure 5-2 Graphs of Means for Significant Effects of Individual Differences on Strategy Measures

5.3.2.3 Browsing Pattern

Browsing pattern was measured using the any-match n-gram method (see Chapter 4 section 4.3.2). Cosines were the dependent measure and were calculated using 2g, 3g, 4g, and 5g. The existence of a relationship between cognitive ability and preferred wayfinding strategy and the degree of similarity between individuals browsing pattern was tested using non-parametric correlation. Table 5-8 shows Spearman Rho correlations between cognitive ability and cosines measured using the four n-gram lengths. Table 5-9 shows Spearman Rho correlations between preferred wayfinding strategy and cosines.

Cognitive Ability	N-Gram Length			
	2g	3g	4g	5g
Verbal Ability V5	.017	.053	.053	.053
Spatial Working Memory (MV)	-.010	.048	.007	.010
Spatial Ability (VZ)	.030	.046	.010	.011

N = 1081

Table 5-8 Correlations between Cognitive Ability and Browsing Pattern

Wayfinding Strategy	N-Gram Length			
	2g	3g	4g	5g
Route	.031	.050	.022	.066*
Survey	.054	.092**	.073*	.063*

N = 1081; * $p < 0.05$; ** $p < 0.01$

Table 5-9 Correlations between Preferred Wayfinding Strategy and Browsing Pattern

As can be seen from Tables Table 5-8 and Table 5-9 a weak but significant positive relationship for route based strategy when measured using 5g ($p < 0.05$), and for survey based strategy measured with 3g, 4g and 5g ($p < 0.01$, 0.05, and 0.05 respectively) exists, with similarity in browsing pattern increasing with dissimilarity in scores of preferred wayfinding strategy.

5.3.3 The Relationship between Behaviour and Performance

The relationship between behaviour and performance was examined by correlating measures of performance with measures of behaviour in terms of navigation and strategy. This involved an examination of the variance shared between variables the original measures of behaviour and performance were therefore reduced as in many instances they were measures resulting from combinations of variables. Of the five original variables measuring performance, R was excluded as this formed part of A; E was excluded as it combined A and TT; MnTT was excluded as it was calculated using number of relevant retrieved documents (directly related to A) divided by TT. This resulted in two variables A and TT being retained, as these were seen to best account for all aspects of performance. These were then correlated with each other to identify any remaining shared variance, a significant positive correlation was identified, rho = .338, N = 47, $p < 0.05$. Simple regression confirmed TT could predict approximately 10% of A, adj $r^2 = .097$, $F(1,45) = 5.92$; $p < 0.05$.

Predictor Variable	B	Beta	t
(Constant)	.498		1.666
Time on Task (TT) - in seconds	.000	.341	2.43*
* significant at p< 0.05	Adj r^2 =.097, F(1,45) = 5.92; p < 0.05		

Table 5-10 Simple Regression Coefficients for Time on Task (TT) as a Predictor of Accuracy (A)

Of the measures of behaviour total nodes visited and unique nodes visited were not included as these were used to calculate lostness. The remaining measures of behaviour (distance, rotation, step size, angle size, lostness, reading time, travel time, and the ratio of reading to travel time) were initially correlated with each other to identify variables that shared large amounts of variance measuring behaviour. The significant correlations are shown in Table 5-11. As many of these variables were not

normally distributed (see previous sections) non-parametric analyses using Spearman Rho were performed.

Distance, rotation, step size, lostness, travel time, and read/travel ratio were all inter-correlated in an intuitive manner. The longer a participant spent travelling the further they were likely to travel and the more they were likely to rotate the environment. The further they travelled and the more they rotated the environment the greater the inter-node step sizes were likely to be and the more lost they were likely to become. Read/travel ratio was negatively correlated as expected since it represented read time divided by travel time and travel time positively correlated with these measures. Read time was only inter-correlated with three of the other seven variables – negatively with step size suggesting people who spend longer reading do not travel as far between subsequent targets, and positively with step angle suggesting that people who spend longer reading are more likely to deviate from a linear travel pattern. It should be clarified that rotation is the amount the individual actually rotates within the environment and step angle is the degree to which the user deviates from a straight line when visiting nodes. If a user visits all nodes within their field of view (FOV) (i.e. all nodes visible on screen at one time) rotation will likely be zero but step angle will be positive and possible large dependent on how many nodes are visible and the order the user visits them. Step angle was only correlated with reading time.

	Distance	Rotation	Step Size	Angle Size	Lostness	Read Time	Travel Time
Rotation	.567**						
Step Size	.545**	.378**					
Angle Size							
Lostness	.537**	.439**					
Read Time			-.319*	.304*			
Travel Time	.788**	.536**			.725**		
Read/Travel Ratio	-.721*	-.479**	-.462**		-.337**	.674**	-.741**

N = 47; * p < 0.05; ** p < 0.01

Table 5-11 Significant Non-parametric Correlations between Behaviour Measures

It was decided to retain all eight variables as measures of behaviour that may predict performance despite the inter-correlations, although it would appear initially the key measures are travel time and read time.

	A	TT
Distance	-.153	.158
Rotation	-.056	.218
Step Size	-.116	-.108
Angle Size	.004	.330*
Lostness	.090	.483**
Reading Time	.422**	.851**
Travel Time	-.091	.320*
Read/travel Ratio	.325*	.268

* p < 0.05; ** p < 0.01 N = 47

Table 5-12 Non-parametric Correlations between Behaviour and Performance Measures of Accuracy (A) and Time on Task (TT)

As can been seen from Table 5-12 time spent reading produces significant moderately strong positive correlations with both A and TT suggesting individuals who spend more time reading are likely to be more accurate but will spend longer on task. Step angle and lostness also correlate positively with TT whereby less linear movements between nodes are likely to increase total time on task and the more lost an individual becomes the more time on task they are likely to take. The read/travel ratio is significantly positively correlated with A suggesting the more time spent reading in proportion to travelling the more accurate a participant is likely to be in document retrieval. The read/travel ratio is not however significantly correlated with TT which

suggests that the proportion of time spent reading to time spent travelling is not important in terms of predicting TT.

As step angle inter-correlated with read time only and no other behavioural measures correlated with TT, it was decided to perform a multiple regression to determine whether the shared variance between step angle and reading time accounted for the relationship between step angle and TT (see Table 5-13).

Predictor Variable	B	Beta	t	Part
(Constant)	149.83		.64	
Reading time	.88	.808	9.57**	.791
Angle size	5.69	.108	1.28	.106
** significant at p< 0.001	Adj r^2 =.685, F(2,44) = 51.02; p < 0.01			

Table 5-13 Multiple Regression Model for Reading Time and Angle Size as Predictors of Time on Task (TT)

The regression model was significant and accounted for approximately 69% of the variance in TT - adj $R^2 = .685$, $F(2,44) = 51.02$; $p < 0.01$. Beta values however showed that only reading time was a significant predictor within the model Beta = .808; $p < 0.01$ compared to step angle Beta = .108; ns. The part correlations confirmed that the predictive power of the model was principally due to reading time which demonstrated large unique predictive power (.79).

As the initial regression model between TT as a predictor of A showed that TT significantly predicted approximately 10% of A, and reading time was correlated with both TT and A, a simple regression with reading time as a predictor of A, and a standard multiple regression with reading time and TT as predictors of A were performed. Reading time was shown to significantly predict approximately 14% of accuracy adj $R^2 = .138$, $F(1,45) = 8.36$; $p < 0.01$. Standard multiple regression showed

that while the combined model of reading time and TT predicted approximately 12% of A - adj R^2 = .119, F(2,44) = 4.107; $p < 0.05$, neither variables were significant predictors independently of one another, Beta for reading time = .363; p = 0.15, and Beta for TT = 0.04; p = .802. The part correlations did however show that the unique contribution of reading time was greater than TT .203 and .023 respectively – see Tables Table 5-14 and Table 5-15

Predictor Variable	B	Beta	t
(Constant)	.520		8.27**
Reading time	.000	.396	2.89**
** significant at p< 0.01	Adj r^2 =.138, F(1,45) = 8.36; p < 0.01		

Table 5-14 Simple Regression Coefficients for Reading Time as a Predictor of Accuracy (A)

Predictor Variable	B	Beta	t	Part
(Constant)	.512			
Reading time	.000	.363	1.47	.203
Time on Task (TT)	.000	.040	.163	.023
Values of t non significant	Adj r^2 =.119, F(2,44) = 4.11; p < 0.05			

Table 5-15 Multiple Regression Model for Reading Time and Time on Task (TT) as Predictors of Accuracy (A)

Finally as lostness and travel time were strongly correlated, and both correlated with TT, a standard multiple regression analysis was performed to identify the relationship between these variables and TT (see Table 5-16). The model was a significant predictor of TT accounting for approximately 28% of variance in TT - adj R^2 = .275, F(2,44) = 9.73; $p < 0.01$. Beta values however showed that only lostness had any significant power within the model, Beta = .463; $p < 0.01$ compared to Beta for travel time = .167; ns. Part correlations showed that a large amount of the predictive power of lostness within the model was unique (.345).

Predictor Variable	B	Beta	t	Part
(Constant)	324.048		1.45	
Lostness	1033.599	.436	2.75**	.345
Travel time	.303	.167	1.05	.132

** significant at $p < 0.01$ Adj $r^2 = .275$, $F(2,44) = 9.73$; $p < 0.01$

Table 5-16 Multiple Regression Model for Lostness and Travel Time as Predictors of Time on Task (TT)

5.4 Discussion

Individual differences in cognitive ability and preferred wayfinding strategies were examined to determine their role when using an SDMS of the nature investigated previously in this thesis. No effects of spatial visualisation, spatial working memory, or verbal ability were found for retrieval performance, navigation, browsing strategy, or browsing pattern. In terms of spatial working memory and visualisation, as detailed at the beginning of this chapter, significant effects had previously been demonstrated. Differences in spatial working memory existed for performance in terms of total time spent on task, a significant interaction occurred between spatial ability and quality of mapping for time on task, and a significant association between users' levels of spatial memory and adopted browsing patterns was observed. For the latter this was only apparent when spatial-semantic mapping was not maximised. The interactive effect of visualisation was also due to differences occurring in conditions where optimal mapping solutions were not applied. Closer inspection of the main effect for spatial working memory for time on task reported in Chapter Three suggests that this main effect is also largely due to differences occurring in conditions that reflect less semantic structure. Given that only the environment employing optimal spatial-semantic mapping (3D100) was used for this experiment the lack of significant results within the current experiment support the previous findings. This contradicts conclusions in the literature where both spatial ability and spatial memory have

demonstrated influences over user performance in SDMS VEs (e.g. Chen, 2000), but supports the argument made by Dillon (2000), and explored with reference to the previous experiment (see Chapters Three and Four), that the nature of the task is an important factor in determining whether individual differences in cognitive ability will impact on IR. When using spatial-semantic mapping as the only tool to aid visualisation of the semantic structure of electronic databases, spatial ability and spatial memory can predict performance and behaviour when either conducting a specific search task, or when browsing in environments in which the mapping is not maximised. However, when browsing for information in an optimally mapped environment no demonstrable influence of these abilities is observed. Examination of browsing patterns also failed to identify any significant relationships between cognitive ability and adopted browsing patterns, however this was not unexpected as previously identified relationships (with the exception of spatial visualisation measured with 2-gram) occurred in environments with less than optimal semantic mapping. This provides further indirect support for the argument that the effects of cognitive ability are only evidenced within IR behaviour and performance when cognitive demand is increased (Westerman & Cribbin, 2000b). This increase in cognitive load may come from the nature of the task or from the additional demands of navigating environments where disorientation is increased due to a lack of 'goodness of fit' between the information space and users' cognitive models (see Chapters Three and Four for more discussions).

The findings associated with preferred wayfinding strategy initially suggest that users with a high preference for survey strategy were disadvantaged in terms of time on task, and average time taken to select a relevant document. However further

exploration of differences in behaviour between high and low survey strategies suggest this may be due to a more complex interaction. Individuals with a high preference for survey strategy travelled shorter distances between documents, and spent significantly more time reading. This is supported by the negative correlation between inter-document step size and reading time, such that individuals who spent more time reading travelled shorter distances between documents. Such a relationship suggests that users were better able to identify the semantic structure of the database so, were more likely to visit documents adjacent to potentially relevant material, made quicker judgements about which documents to study in depth, and were therefore able to spend more time reading. Differences between high and low preference groups in the ratio between travel and reading time approached significance with high preference groups spending proportionately more time reading than travelling between documents. The only effect of level of preference for route knowledge again initially suggested a disadvantage for those with high preference in terms of the time taken to complete the task however the trends displayed in the descriptive statistics while not significant suggest the same pattern of behaviour as witnessed for survey knowledge. Time on task was also shown to positively correlate with accuracy displaying a speed accuracy trade-off. Although specific differences in recall and accuracy were not identified on the basis of wayfinding strategies a more subtle relationship may exist due to the effects of adopted wayfinding strategies on behavioural strategies and time on task (i.e. high preference for adopting specific wayfinding strategies predicts reading time which in turn predicts time on task and accuracy of IR). This is considered in more detail next when the relationship between behaviour and performance is examined.

One of the principal aims of this experiment was to explore any identifiable relationships between behaviour and performance. The results confirm the existence of such a relationship and show it is highly complex and interactive. From closer examination of the way in which the dependent measures interact with both each other and the independent measures a principal behaviour measure i.e. reading time has been identified as pertinent in contributing to this relationship. Time spent reading was moderately well correlated with accuracy, and highly correlated with time on task, suggesting that users who adopt a strategy of taking longer to read documents will spend longer on task but will have improved accuracy. Time on task although considered a dependent measure of performance for the purpose of this thesis was correlated with accuracy suggesting a speed-accuracy trade-off. Multiple regression analysis however demonstrated that time on task did not independently predict accuracy, and much of the variance shared between time on task and accuracy was due to variance shared between time on task and time spent reading. It should be highlighted that the adoption of time on task as a performance measure rather than a behaviour measure was not arbitrary although there is justification for it to be considered indicative of behaviour. Much consideration was given to this issue but it was decided that within the context of this thesis time on task is a measure of 'outcomes' in that efficient IR not only requires accurate information retrieval, in many cases it also requires speedy retrieval due to the time restrictions most people operate under. However, as can be seen from this discussion, the distinction between time on task as a measure of retrieval performance or a measure of behaviour is not easily made. The ratio between reading time and travelling time predicted accuracy, such that accuracy improved when users spent proportionally more time reading than travelling. This ratio however shared no variance with time on task. It would appear

therefore that users who move quickly from one document to another but are selective in their judgements of which documents to spend time reading (people with high preferences for survey strategy tended to visit more total nodes and more unique nodes although this was not significant) will achieve greater accuracy of document retrieval. Ratio did not share any variance with time on task however suggesting that while both reading time and travel time predicted time on task (as would be expected since time on task comprises principally these two measures) the way in which time was proportionally allocated to these behaviours was not relevant. Lostness was shown to be predictive of time on task which suggests the more disorientated individuals were the more time they took overall. Step angle size was also predictive of time on task but multiple regression analysis showed this was due to the amount of variance angle size shared with reading time. This positive relationship between reading time and step angle may be indirectly associated with the negative relationship between step size and reading time identified earlier. If as suggested individuals who utilise the semantic structure of the environment tend to visit nodes adjacent to current potentially relevant nodes step size would be small, but angle sizes would increase as users visited all documents within their current field of view (FOV) (see Chapter Four).

A model for predicting performance from behaviour within VE SDMSs using spatial-semantic mapping can be proposed that suggests individuals who spend more time reading documents, and less time travelling will retrieve more relevant documents, however this will be at the expense of time on task. The results presented here suggest that users who can identify the semantic structure within the mapping of the environment will be better placed to locate documents likely to be relevant with less

time spent travelling, and will therefore be able to allocate proportionally more time to reading than travelling, leading to more effective IR

This pattern of results is intuitive based on the findings of Toms (2000) who found that people when free to browse in an unstructured task made quick decisions about what to read and then spent almost twice as long reading material as looking for it.

Unfortunately, with the benefit of hindsight, it is clear comparisons should have included quality of mapping similar to the previous studies. As stated a model of the way behaviour can predict performance can be proposed, but conclusions regarding the way people use the semantic information present have been inferred from findings associated with survey strategy due to lack of comparison of between qualitatively different spatial-semantic mapping. The significant effect of high levels of preferred survey strategy being associated with shorter inter-document distances, longer time on task, and longer time spent reading suggests spatial-semantic mapping is being utilised. It has been demonstrated in VE navigation tasks people with high survey knowledge preference acquire a cognitive model more quickly and can use landmarks more effectively (Cutmore, et al., 2000). In the current environment the documents themselves represent the only possible landmarks as no other cues exist. It can be argued that where a goodness of fit exists between the environment and users' cognitive models those individuals with a high preference for survey knowledge can identify document content as landmarks/navigational cues and use them effectively. While the model regarding the relationship between reading time and accuracy of information retrieval is supported by the results, the lack of observation of an overt relationship between wayfinding strategy and accuracy does not confirm the proposed

use of identified semantic structure, and further research is needed to explore this.

6 Conclusions

6.1 Research Questions

The aim of the research reported in this work was to investigate whether spatial-semantic theories of cognition have a role to play in improving information retrieval from computerised databases (particularly when browsing for information). Can the use of systems that use the inter-document spatial organisation to visually convey the underlying semantic structure of database contents (e.g. spatial data management systems – SDMSs), simplify the task of browsing for solutions to general information needs be? More specifically can spatial-semantic mapping (i.e. the proximal location of documents conveys inter-document semantic relationships) of database contents within information visualisation systems facilitate improved user-system interaction leading to improved information retrieval?

Directed search tasks using a similar interface to that used currently had already been examined (Westerman & Cribbin, 2000b) and this thesis was designed to expand upon these authors' findings. Searching for target pieces of information/facts in answer to a specific query, and browsing for information related to a general query are qualitatively different (Marchionini & Shneiderman, 1988). This work examined browsing performance and behaviour to see if i) the benefits of using spatial-semantic mapping identified by Westerman & Cribbin (2000b) generalised to a browsing task, and ii) whether these findings could be expanded to address more specific issues regarding retrieval performance and user browsing behaviour.

Three principal questions detailed in Chapter One were identified as relevant. They can be summarised as: 1) Is information retrieval performance (e.g. items retrieved, time taken to retrieve them etc.) mediated by the degree to which the spatial-semantic organisation of the database is compatible with the users' cognitive model? 2) Is user behaviour in terms of navigation and browsing strategies (e.g. distance travelled, number of objects visited, degree of lostness, etc.) influenced by the 'goodness of fit' between the semantic structure of the database presented in the environment visualisation, and the users' internal cognitive model? 3) Does a relationship exist between behaviour and performance that is facilitated by the cohesion between the IR system and the users' conceptual space, such that behaviour in terms of navigating the environment and browsing strategies can predict performance? The role of individual differences in specific cognitive abilities within each of these questions was also examined.

The current chapter discusses the findings of the work conducted in this thesis, which principally demonstrates spatial-semantic mapping is a tool that information seekers successfully employ when browsing information visualisation systems. More detailed examination of the findings suggest; i) good quality spatial-semantic mapping successfully conveys sufficient detail regarding the semantic structure of database contents to benefit information seekers; ii) when browsing for information users can benefit from the additional cues to semantic structure provided by mapping contents into three dimensions; iii) individual differences in cognitive ability do not generally appear to be an important factor in determining users' retrieval performance or behaviour when browsing for information (there are exceptions which will be discussed); iv) certain browsing behaviours facilitate predictions about retrieval

performance when maximum semantic structure is conveyed through spatial mapping. The structure of the current chapter reflects these findings, with conclusions about what has been learned from the studies including shortcomings and proposals for developing this work presented at the end of this chapter.

6.2 Spatial-semantic Mapping

6.2.1 Gauging Semantic Similarity

The experimental platform used in the current work was similar to that employed by Westerman & Cribbin (2000b) but comprised whole documents rather than nouns. In order to position documents within the environment based on spatial-semantic distance, a reliable means of identifying inter-document semantic similarity was needed. Westerman & Cribbin (2000b) used human ratings of semantic similarity for their study, however there are problems associated with this. The predominant problem is the number of judgements required to make all possible comparisons. It took a week for participants in the study mentioned to make all 4465 comparisons, while this is a formidable task it remained manageable as judgements were made between individual objects represented by nouns. The database used in the current work consisted of 100 documents which required 4950 pairwise comparisons to obtain inter-document similarity measures. Clearly it would not be practical to ask participants to perform this task for real world databases (e.g. the internet and the World Wide Web) where numbers of documents equal thousands, millions, and billions. Developments in methods of automatic text analysis (ATA) over recent years have provided an alternative to human judgements of documents' relevance or similarity. As demonstrated in Chapter Two these methods have successfully applied

vector space modelling to represent the semantic structure of documents in a high dimensional space, similarity between documents can be identified within this space and then re-scaled to two or three dimensions for information visualisation purposes. In the current work, the suitability of using the n-gram method of ATA (Damashek, 1995a) to identify semantic similarity between the database documents for the purpose of spatial-semantic mapping was tested. The results supported the use of this method demonstrating the superiority of the appropriate length n-gram over human judges in identifying average levels of agreement between human raters. While it is acknowledged that human judgements should be considered the 'gold standard' by which to test ATA methods of identifying document similarity, the level at which humans agree amongst themselves on document similarity is low. This is demonstrated by the low levels of inter-rater reliability (IRR) achieved during comparisons of inter-rater judgements of document similarity conducted in Experiments One and Two of Chapter Two (see sections 2.2.2 and 2.3.2). N-gram based ATA achieved higher levels of agreement with humans on average than they did with each other. For technical document sets this pattern was observed at all n-gram lengths, and for non-technical documents persisted at all but the poorest performing n-gram lengths. Low inter-rater reliability is a problem that can be associated with human judgements of semantic similarity and/or document relevance due to the dynamic individual and situational factors that interact when making such judgements (e.g. Schamber, et al., 1990). As identified in Chapters One and Two human judgements of similarity and relevance alter dynamically due to the interaction of numerous factors both internal and external to the person making the judgement. Not only do raters demonstrate high disparity with other raters, differential ratings of the same pair of items made by the same individual can and do occur. It is impractical

to assume that these factors can be identified and controlled for each individual in every information retrieval situation. For example, judgements have been shown to alter during a particular information search due to the effects of changes in content knowledge and re-evaluation of information need (Schamber et. al., 1990). However such problems do impact on the effectiveness of information retrieval systems. One way to address this problem is to produce IR systems that can most accurately convey the 'general' semantic structures of a set of documents that the majority users agree upon. Users are then free to judge which documents most suit their current needs. This can be achieved if the system can broadly match the users' cognitive models of semantic structure.

To provide an SDMS that allows multiple users to optimise the spatial-semantic mapping a high degree of 'context-free' semantic structure must be conveyed. It has been shown that n-gram analysis can successfully do this. However, the benefits of this method of ATA are maximised when the optimal length n-gram is used. Further investigation is needed to identify practical and effective ways of determining optimal n-gram lengths for different types of document sets. It may even be possible to identify n-gram lengths that are optimally suited to a specific individual which would facilitate personalisation of information visualisations based on spatial-semantic mapping.

6.2.2 Spatial-semantic Mapping Quality and use of Dimensions

The work presented clearly demonstrates a positive benefit to using good quality spatial-semantic mapping when browsing for information, adding support to the research of Westerman & Cribbin (2000b) which showed users successfully employed

spatial-semantic mapping to improve performance when carrying out directed search tasks. The current work shows that users can benefit from optimal spatial-semantic mapping when browsing for information. The benefit applies to both retrieval performance, and navigational efficiency. However the degree of benefit gained is not proportional to the amount of semantic variance accounted for in the mapping. When there is optimal 'goodness of fit' between users' cognitive models of semantic structure and the semantic structure conveyed through environmental mapping user-system interaction is enhanced, information seekers retrieve information more effectively and navigate more efficiently. When spatial-semantic mapping is not optimal however performance is generally poorer, although performance does not systematically decline with quality of mapping. This may be due to the multi-dimensional nature of the browsing process (Chang & Rice, 1993) such that spatial-semantic mapping accounts for a relatively small amount of variance between users' observed performance. Differences are possibly only apparent when the effects are distinct.

An alternative explanation is that the optimal mapping solution (3D100) used in these experiments accounted for a relatively small amount of the semantic variance identified by ATA (48.5%). In less than optimal solutions it is possible too little remaining semantic structure was available for users to identify clearly. It is possible that within 3D environments users demonstrated random performance particularly evident in terms of browsing behaviour as they struggled to identify a useful structure. In 2D60 however users may have identified an alternative structure not based on the semantic properties of the environment. Both retrieval performance and navigational efficiency were poorer in 2D60 suggesting users were not assimilating the semantic

structure, and the pattern of results in terms of navigation (see Chapter Four) suggests they adopted a more thorough, sequential approach to browsing as evidenced by visiting many more documents. This would suggest users in 2D60 did not rely on the semantic structure of the environment to facilitate browsing. This is further demonstrated by the sequence in which documents were visited. In Chapter Four similarities between users' browsing patterns based on the order documents were visited were analysed using the n-gram method employed for ATA. This analysis showed that although overall users in 2D60 tended not to adopt similar browsing patterns, where similarities did occur they remained consistent over a longer sequence of documents. The amount of shared variance in users' patterns decreased significantly with n-gram length however in 2D60 this decrease was not as steep. It can be argued that users will employ the most salient structures in order to acquire a model of the information space. Given that users' only have to navigate in two directions in 2D60 the geometric properties of the environment are easier to identify. When a poor fit between the semantic structure of the environment and the user's cognitive model exists, and an alternative structure can be identified as in 2D60 users will adapt their browsing behaviour appropriately, however when the 'goodness of fit' is poor but no other structure is clearly visible users continue to employ the limited semantic structure available, as evidenced by improved retrieval performance in 3D60 compared to 2D60.

The results pertaining to differences between 2D60 and 3D60 do not support those reported by Westerman & Cribbin (2000b), who identified a distinct disadvantage when using three-dimensions for mapping. There was a negative trade-off between the additional demands of navigating three dimensions and the additional information

available regarding environment structure. It was suggested that the best 3D solution (3D100) would need to account for an additional 30 – 50% semantic variance to compensate for the additional cognitive load imposed by navigation. In the current study however mapping to three-dimensions provided an advantage even when amounts of semantic variance accounted for were equal. It would appear this opposing effect is due to differences in the nature of the task, and that when browsing, a task considered not to be cognitively demanding, sufficient cognitive processing capacity remains available for assimilating the environmental model with the users' cognitive model. It would appear that in the present task users' do acquire good semantic models of the environment which is apparent not only in the performance benefits drawn from spatial-semantic mapping, but also in the fact that people with a high preference for survey based wayfinding strategies identify the global structure of the environment more effectively (see Chapter Five) than people with a low preference for this strategy.

Findings regarding individual differences in general suggest that cognitive ability does not predict retrieval performance or browsing behaviour when the quality of spatial-semantic mapping is optimal. Navigation was not influenced at all by individual differences in ability, but retrieval performance and similarity of browsing pattern did show some effects when spatial-semantic mapping was not optimal. As discussed in Chapters Three and Four these findings are difficult to interpret, however given the conflicting nature of findings reported in the literature associated with individual differences and IR performance in SDMSs (e.g. Chen, 2000; and Westerman & Cribbin, 2000b) and the findings in the current work, the observation by Dillon & Watson (1996) that a complex interaction exists between cognitive ability and IR is

supported. It would appear from the results here that individual differences have a greater effect on user performance and behaviour when cognitive load is higher, for instance when the nature of the task is more demanding as in Westerman & Cribbin (2000b), or the acquisition of a good mental model is made difficult by poor 'goodness of fit' between the IR environment and the individual's semantic model.

6.2.3 The Relationship between Browsing Behaviour and Retrieval Performance

The relationship between browsing behaviour and retrieval performance was explored in Chapter Five. As discussed, evidence from the studies reported in Chapters Three and Four, showed that both retrieval performance and browsing behaviour were positively influenced by the quality of spatial-semantic mapping of the environment, and benefited from the use of three-dimensions for mapping. In the experiment reported in Chapter Five a relationship between behaviour and retrieval performance was identified. Most variables inter-correlated, therefore as explained in detail in the results section, measures of retrieval performance were reduced to, accuracy and time on task. On the basis of the correlation and regression analyses performed between measures of behaviour as predictor variables and the aforementioned performance measures as response variables, a model can be proposed that accounts for the way browsing behaviour and retrieval performance interact. The model states that retrieval performance in terms of accuracy and time on task can be predicted by the amount of time an individual spends reading overall and proportionate to time spent travelling between documents. The model shows that when users' spend more time reading they will retrieve more relevant documents with greater accuracy but this will be at the expense of total time on task. However since the ratio of reading time to travel time

was positively correlated with both accuracy and time on task, it can be argued more efficient information retrieval can be achieved by reducing travel time. This is upheld somewhat by inferences drawn from results for differences in preference for survey based wayfinding strategies. It was shown that users with a high preference for survey based strategies took longer on task, spent longer reading, and took shorter inter-document steps. This would suggest that where assimilation of the environment structure to users' cognitive models was facilitated by the 'goodness of fit' of spatial-semantic mapping, users with a high preference for survey knowledge acquired a better global model of the environment. Given the lack of landmark or route-based cues these individuals would have an advantage. Despite a lack of overt findings with respect to retrieval performance in terms of accuracy, it can be argued, based on the significant findings detailed above, that participants with a high preference for survey based strategies have the potential to benefit from behaviour and improve performance. It was suggested (Chapter Five) that these people gain a qualitatively better model of the environment and therefore when a potentially relevant document is located continue browsing in the vicinity of that document which results in shorter step sizes. This allows the individual to allocate more time to reading, which leads to longer on task but based on the model could also potentially lead to improved accuracy – although not statistically significant this trend is apparent in the results with high survey groups scoring higher on average than low survey groups.

Direct conclusions regarding the role of the quality of spatial-semantic mapping cannot be drawn due to a lack of comparison between differentially mapped environments. However given that recall and time on task were factors that demonstrated significant effects of the quality of spatial mapping it could be inferred

that the model would apply on the basis of quality of spatial-semantic mapping this however would need to be tested.

6.3 Implications for Existing Research Fields

6.3.1 HCI and Information Processing Theories of Human Cognition

At the start of the information processing approach to cognitive psychology and HCI, Newell & Simon (1956) designed and presented a computerised information processing system that was modelled on human problem solving techniques rather than systematic algorithms. Since that time a principal aim within HCI research has been to model human cognition in order to identify and predict how people carryout computer based tasks. Being able to model such behaviour allows for highly effective systems to be designed more efficiently and more economically.

As a result of the information processing movement there has been a reciprocal relationship between the application of research within both cognitive psychology and HCI fields. The computer metaphor has been used to explain many aspects of cognitive functioning including memory (e.g. Atkinson & Shiffrin, 1968), knowledge representation (e.g. Kintsch, 1986), and problem solving (Newell & Simon, 1972) etc. As research has advanced it has been recognised that the 'computer metaphor' is too simplistic to explain human cognition and a 'brain metaphor' is possibly more appropriate (e.g. McClelland, McNaughton, & O'Reilly, 1995). Regardless of the specifics of these approaches e.g. 'computer metaphor' or 'brain metaphor', the important factor to recognise is the existence of cognitive structures that handle human information processing and attempt to provide the best possible match between

these 'cognitive architectures' and the processes used in computerised systems, to provide optimum effectiveness and efficiency in human-computer interaction. The EPIC (Executive Process-Interactive Control) architecture for example is a framework that allows modelling of human information processing (Kieras & Meyer, 1997) that has achieved successful computerised simulation and prediction of human performance for a variety of computer based task. For instance, (Hornof & Kieras 1997) used EPIC to run a variety of models examining how people use drop down menu items on a computer interface. By using cognitive models that varied across a variety of parameters e.g. serial vs. parallel processing, the authors were able to demonstrate that people use random and serial strategies, and process multiple menu items simultaneously. These types of findings have direct implications on the design of interfaces.

The work presented in this thesis has implications for the use of architectures such as EPIC and vice versa. By providing evidence about the way in which individuals cognitively process information support is being given to spatial-semantic theories of cognition which can be used to design EPIC based models of IR. Conversely by facilitating these models frameworks such as EPIC can be used to test whether predictions about the strategies individuals use are correct. In order to accurately replicate and predict human performance these systems rely heavily on empirical information about human cognitive models, and how people perform tasks. The research conducted here has given support to existing theories that suggest people process semantic information in a cognitive space that adheres to the similar spatial rules of processing as perception (e.g. Jackendoff, 1987). This information can be used to programme an EPIC architecture to model information retrieval in SDMSs to

design and test optimum data base organisations in the presented visualisations used by an IR system. One of the principal problems in the design of effective IR systems is the impracticality of testing new systems. The rapid growth of numbers of people and systems participating in TREC demonstrates how technology advances more rapidly than can be managed in terms of human evaluation of these emerging technologies resulting in. The result is evaluations of system effectiveness being made on the judgements of single individuals (see Section 1.5). While this is far from ideal, the alternative options of making general evaluations on the basis of large samples of human assessors are impractical. It has been demonstrated in Chapter Two that a large level of disagreement about semantic similarity exists between people and that on average, automated methods of judging inter-document semantic similarity have a greater level of agreement with humans than people do. It is therefore important when designing IR systems that these idiosyncrasies are accounted for. By enabling accurate cognitive modelling for use with frameworks such as EPIC, systems can be designed that can be adapted to differences in individual's cognitive models. The reduction in evaluation costs and the ability to programme many different cognitive models to test systems will result in more flexible, personalised retrieval interfaces.

The research here has shown that this is possible by firstly identifying that when spatial mapping is optimal people can and do use spatial mapping to form part of their retrieval strategy, and secondly by demonstrating a means of quantifying search patterns and similarities between search patterns through the use of measurements such as distance and direction of travel, and n-gram analysis of the order in which documents are visited. The possibilities identified by the work in this thesis are very exciting but are clearly at a very early stage. Problems associated with identifying the

alternative strategies that people adopt when spatial-semantic mapping is not optimal and the nature of the interference that is caused by poor mapping were not considered prior to conducting the experiments. These omissions limit the precision with which the IR task presented could be modelled, if at all at this stage. However the results do suggest that with further work these problems could be overcome.

6.3.2 Information Retrieval and Information Visualisation

The information visualisation environment used in this thesis was specifically designed to enable investigation of the use of spatial-semantic mapping. In order to ensure that the only cues to document location that participants were presented with were spatial, the environment was bereft of any visualisation tools. As shown in the figures of the environment presented in Chapter 3, all documents were perceptually represented as objects identical in shape, colour, and size. These factors did actually vary but the variations were dynamic and proportional in order reflect the spatial position of the object with reference to the user's position in the environment. In other words when the user changed position within the environment the size of objects would alter relative to their new perspective and position in relation to the user.

This approach differs from much of the published work on information visualisation which tends to incorporate stand alone prototype IR systems that are being tested. For example Lighthouse (Leuski & Allen, 2000), is a system designed to incorporate spatial cues to document relevance by structuring document organisation to reflect relevance to an 'original' document. The user decides which document is most relevant to their query, and the remaining documents are spatially organised so that proximity reflects relevance. However this system is dependent on all documents

within the original database being searched using more traditional retrieval methods. A retrieved list of documents is produced that may be relevant to the original enquiry to a greater or lesser degree. The user selects the document they consider most relevant and the remaining retrieved documents are spatially organised in the centre of the screen (see Figure 1-4; Chapter 1). It has been shown by the authors that the additional information provided by the spatial visualisation increases performance on both measures of recall and precision. They however found that users preferred a 2D visualisation to a 3D layout. It was presumed this was due to cognitive overload due to additional navigational demands (Allen et. al., 2001).

However the results of the current research suggest that when browsing for information the additional semantic cues provided by using a third dimension for mapping are beneficial to both performance and behaviour. The reasons for the previous findings may be due to the system not consisting entirely of spatially mapped objects. The amount of information present on screen (text based lists of retrieved documents and spatially arranged nodes) may have caused interference in users' perception of the organisation and caused a perceptually driven cognitive overload. Using both text based menu lists and VE object-document visualisations in the Lighthouse system, means that there is less space to present the object based VE. This in turn means that object/document density is high creating problems of occlusion. It may be problems associated with the 'crowding' of objects that explains differences in performance between 3D and 2D. An alternative explanation that is supported by the findings from both this thesis and that of Westerman & Cribbin (2000a), is that the task Allen et. al. (2001) were asking their participants to perform represented a specific search task rather than a brows task. The browsing element would have been

addressed with the initial key term search that resulted in a list of retrieved documents. Once this list was produced participants would have some idea of what documents they were looking for as a result they may have been searching for specific items rather than just browsing the entire space. This would place additional load on memory as they would need to remember more specific details regarding the target documents. The additional navigational workload present in 3D environment may then give a disadvantage over 2D (Westerman & Cribbin, 2000b).

Other IV systems such as Scatter Gather (Cutting et. al., 1992), Bead (Chalmers & Chitson, 1995), and Perspective Wall (Robertson et. al., 1991a), use multiple visualisation tools including the spatial organisation of documents and have been shown to successfully increase IR performance. In these systems however the spatial arrangement does reflect a direct spatial-semantic mapping of the document contents. It is unlikely therefore that there is an optimal 'goodness of fit' between the spatial arrangement presented by the system, and the user's cognitive map. Results from the studies in this thesis suggest that performance reduces, and strategies become more exhaustive and less efficient as 'goodness of fit' reduces. It may be beneficial therefore to use spatial-semantic mapping to produce the spatial element of the visualisation within these systems. This would optimise the benefits of the spatial organisation and the additional visualisation tools.

6.4 Conclusions

It is felt that one of the major flaws in this study was not maintaining the four differentially organised SDMS environments to examine the effects of spatial-

semantic mapping quality, and number of dimensions used for environmental organisation, in the study reported in Chapter Five. While it was considered unnecessary at the time, as conclusions regarding these factors had been drawn from the previous studies, inferences regarding the effect of these features on the model of the relationship between behaviour and performance have been made but not tested. Behaviour in terms of the amount of time allocated to reading and travelling may well have altered across mapping conditions, together with effects of individual differences in preferred wayfinding strategies. If further studies pursue this work this shortfall in experimental design and data collection should be addressed. It remains important however, to ensure sufficiently large sample sizes are used to achieve reliable results that can generalise to the population as a whole, especially when considering individual differences. This is currently one of the main flaws in the literature pertaining to individual differences and information visualisation.

The current work has shown that the additional semantic information that can be conveyed by using a third dimension for visual representation is an advantage to the user, and in a browsing task this outweighs the disadvantages of the additional cognitive load generated through navigating a more complex space. However the adverse effects of mapping density (i.e. distance between nodes/documents) and problems of occlusion that this may cause were not examined. The current levels of density and speed of travel were set relatively arbitrarily choosing the average settings available within the limitations of the programme used for the VE. Placing documents too far apart would increase the volume of the environment and reduce the number of nodes that appeared in the users' FOV, this in turn would make it more difficult for users to acquire a cognitive map of the information space Westerman & Cribbin,

(2000b). Alternatively however, if mapping density was increased and documents were placed too close together, the number of occluded nodes would increase, this too would have a detrimental effect on users' acquisition of a cognitive map, and reduce the advantages afforded by the use of a third dimension (Cockburn & McKenzie, 2003). It was hoped that the level of density chosen would offer a balance between these two scenarios, however this balance was achieved by simply selecting the mid-level available within the programme. For the current study it was felt this was sufficient to control for the effects of occlusion which was not expected to present too serious a problem due to the dynamic nature of navigation. As soon as a participant moved their position their view would alter and any occluded nodes would become immediately visible. Although other nodes then became occluded, this was considered representative of navigation in a real world situation. The results support this judgement as an advantage for 3D over 2D was observed despite there being potential for a higher level of occlusion in 3D. However, it would be worthwhile to investigate the effects of mapping density on both behaviour and performance, to identify at what levels occlusion impacts on browse strategies and retrieval effectiveness. If an optimum mapping density could be identified effects of occlusion could be controlled and subsequently factored out enabling a more precise evaluation of the advantages afforded by the additional semantic information portrayed in the extra dimension.

In the current work no attempt has been made to identify optimal browsing strategies, or how strategies differ once a relevant document has been identified, this would be a worthwhile avenue of research for the future, and one the author is currently considering.

Despite these shortcomings it is felt the work presented here has been worthwhile and beneficial. With respect to the principal research aim this work suggests that spatial theories of semantic cognition can explain the observed improvements in information retrieval when browsing computerised databases that employ spatial-semantic mapping to convey the semantic structure of the database. In terms of the three questions identified it has been shown that; 1) Information retrieval performance is mediated by the degree to which the spatial-semantic organisation of the database is compatible with the users' cognitive model, and increased 'goodness of fit' between the two facilitates improved performance; 2) Aspects of users' browsing behaviour are also influenced by the 'goodness of fit' between the semantic structure of the database presented in the environment visualisation, and the users' internal cognitive model; and 3) A relationship exists between behaviour and performance for which a model is proposed that can predict retrieval performance from behaviour strategies, and that this model is likely to be more effective when a high degree of agreement exists between the IR system and the users' conceptual space.

This research supports the current progression towards replacing traditional search engines that use Boolean type keyword searches to produce a list of possible relevant documents, with information visualisation systems that portray the semantic structure of the database and allow users to identify possible relevant documents for themselves. As they progress, these types of IR systems will allow the user more freedom to adapt the interface to better match their own cognitive models of semantic structure. Such systems could be personalised for particular users or tailored for specific IR tasks. This could be achieved by synthesising findings from this and similar research with HCI modelling architectures to facilitate the evaluation of

various interface organisations based on multiple cognitive models. However at present this level of precision is a distant goal and there are many barriers that need to be overcome due to the complexity and individuality inherent in human cognition.

References

Allan, J., Leuski, A., Swan, R., & Byrd, D. (2001). Evaluating combinations of ranked lists and visualizations of inter-document similarity. *Information Processing & Management, 37 [3],* 435-458.

Atkinson, R.C. & Shiffrin, R.M. (1968). Human memory: A proposed system and its control processes. In K.W. Spence & J.T. Spence (Eds.), *The psychology of learning and motivation (Vol.2),* London: Academic Press.

Baddeley, A. D., Emslie, H., & Smith, M. (1992). *The Speed and Capacity of Language-Processing Test (SCOLP).* Thames Valley Test Company: Bury St. Edmunds.

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval.* ACM Press: New York.

Barshi, I. & Healy, A. F. (2002). The effects of mental representation on performance in a navigation task. *Memory & Cognition 30[8],* 1189-1203.

Borgman, C. L. (1989). All Users of Information-Retrieval Systems Are Not Created Equal - An Exploration into Individual-Differences. *Information Processing & Management, 25[3]* 237-251.

Borgman, C. L. (1999). The user's mental model of an information retrieval system: an experiment on a prototype online catalogue. *International Journal of Human-Computer Studies, 51[2]* 435-452.

Borlund, P. (2000). Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation, 56[1],* 71-90.

Brooks, T. A. (1995a). People, Words, and Perceptions: A Phenomenological

Investigation of Textuality. *Journal of the American Society for Information Science, 46[2],* 103-115.

Brooks, T. A. (1995b). Topical Subject Expertise and the Semantic Distance Model of Relevance Assessment. *Journal of Documentation, 51[4],* 370-387.

Brooks, T. A. (1997). The Relevance Aura of Bibliographic Records. *Information Processing & Management, 33[1],* 69-80.

Brooks, T. A. (1998). The Semantic Distance Model of Relevance Assessment. *Proceedings of the American Society for Information Science Annual Meeting, 35,* 33-44.

Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterising semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review, 8,* 531-544.

Burgess, C. (1998). From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. *Behavior Research Methods Instruments & Computers, 30[2],* 188-198.

Burgess, C. & Livesay, K. (1998). The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kucera and Francis. *Behavior Research Methods Instruments & Computers, 30[2],* 272-277.

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse Processes, 25[2],* 211-257.

Castells, M. (2000). *The Rise of the Network Society, 2nd Edition.* Blackwell Publishers Ltd: Oxford.

Cavnar, W. B. (1993). N-Gram-Based Text Filtering For TREC-2, In D. Harmen

(Ed.), *NIST Special Publication 500-226,* (pp. 171-179). Gaithersburg, MD: National Institute of Standards and Technology.

Cavnar, W. B. (1995). Using an N-Gram based Document Representation with a Vector Processing Retrieval Model. In D. Harman, (Ed.), *NIST Special Publication 500-226* (pp. 269-278). Gaithersburg, MD: National Institute of Standards and Technology,

Chalmers, M. & Chitson, P. (1995). BEAD - An Information Visualization System. In: *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 32,* 330-337.

Chang, C. K. & McDaniel, E. D. (1995). Information Search Strategies in Loosely Structured Settings. *Journal of Educational Computing Research, 12[1],* 95-107.

Chang, S. J. & Rice, R. E. (1993). Browsing - A Multidimensional Framework. *Annual Review of Information Science and Technology, 28,* 231-276.

Chen, C. (1997). Spatial Ability and Visual Navigation: An Empirical Study. *New Review of Hypermedia and Multimedia, 3,* 67-89.

Chen, C. (1998). Generalised Similarity Analysis and Pathfinder Network Scaling. *Interacting with Computers, 10,* 107-128.

Chen, C., Thomas, L., Cole, J., & Chennawasin, C. (1999). Representing the Semantics of Virtual Spaces. *IEEE Multimedia, 6[2],* 54-63.

Chen, C. & Yu, Y. (2000). Empirical Studies of Information Visualization: A Meta-analysis. *International Journal of Human-Computer Studies, 53,* 851-866.

Chen, C. M. (2000). Individual differences in a spatial-semantic virtual environment. *Journal of the American Society for Information Science, 51[6]* 529-542.

Chen, C. M., Cribbin, T., Kuljis, J., & Macredie, R. (2002). Footprints of information foragers: behaviour semantics of visual exploration. *International Journal of Human-Computer Studies, 57[2],* 139-163.

Chen, C. M. & Czerwinski, M. P. (2000). Empirical evaluation of information visualizations: an introduction. *International Journal of Human-Computer Studies, 53[5],* 631-635.

Chen, C. M. & Macredie, R. (2000). Individual differences in virtual environments - Introduction and overview. *Journal of the American Society for Information Science, 51[6],* 499-507.

Chen, C. M. & Rada, R. (1996). Interacting with hypertext: A meta-analysis of experimental studies *Human-Computer Interaction, 11[2],* 125-156.

Cleverdon, C. W. (1967). The Cranfield Tests on Indexing Language Devices *ASLIB Proceedings, 19[6],* 173-193.

Cleverdon, C. W. (1972). On the Inverse Relationship of Recall and Precision. *Journal of Documentation, 28[3],* 195-201.

Cockburn, A. & McKenzie, B. (2003). Evaluating the effectiveness of spatial memory in 2D and 3D physical and virtual environments. In: *Proceedings of the SIGCHI conference on Human factors in computing systems,* 203-210.

Collins, A. M. & Loftus, E. F. (1975). A Spreading-Activation Theory of Semantic Processing. *Psychological Review, 82[6],* 407-428.

Collins, J. & Westerman, S. J. (2001). Browsing Patterns in a Virtual Information Space Representation of a Document Database. In: *Proceedings of 9th International Conference on Human-Computer Interaction, 963-967,* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Craig, A., Davies, D. R., & Matthews, G. (1987). Diurnal-variation, Task Characteristics, and Vigilance Performance. *Human Factors, 29[6],* 675-684.

Cribbin, T. & Westerman, S. J. (1999). Spatial Data Management Systems: Mapping Semantic Distance. In: *Proceedings of INTERACT '99. Seventh IFIP Conference on Human-Computer Interaction,* 171-178. Amsterdam: IOS Press.

Cugini, J., Laskowski, S., & Piatko, C. (1997). Document Clustering in Concept Space: The NIST Information Retrieval Visualization Engine (NIRVE). *CODATA Euro-American Workshop on Visualization of Information and Data, Paris France, June 1997.*

Cutmore, T. R. H., Hine, T. J., Maberly, K. J., Langford, N. M., & Hawgood, G. (2000). Cognitive and gender factors influencing navigation in a virtual environment. *International Journal of Human-Computer Studies, 53[2],* 223-249.

Cutting, D. R., Karger, D. R. P. J. O., & Tukey, J. W. (1992). Scatter/Gather: A cluster-based approach to browsing large document collections. In: *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 318-329.

Damashek, M. (1995a). Gauging Similarity with N-Grams - Language-independent Categorization of Text. *Science, 256[5199],* 843-848.

Damashek, M. (1995b.) Performance of Text Retrieval Systems. *Science 268,* 1419-1420.

Darken, R. P. & Sibert, J. L. (1996a). Navigating large virtual spaces. *International Journal of Human-Computer Interaction, 8[1],* 49-71.

Darken, R. P. & Sibert, J. L. (1996b). Wayfinding Strategies and Behaviours in Large Virtual Worlds. In: *Proceedings ACM SIGHCI, 96,* 142-149.

Davies, D. R. & Parasuramen, R. (1982). Decision Theory and the Analysis of Vigilance Performance. In D. R. Davies & R. Parasurman (Eds.), *The Psychology of Vigilance*, London: Academic Press.

Deerwester, S., Dumais, S. T., Furnas, G. W., & Landauer, T. K. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science, 41[6],* 391-407.

Dillon, A. (2000). Spatial-semantics: How users derive shape from information space. *Journal of the American Society for Information Science, 51[6],* 521-528.

Dillon, A. & Watson, C. (1996). User analysis in HCI - The historical lessons from individual differences research. *International Journal of Human-Computer Studies, 45[6],* 619-637.

Dumais, S. (1995). Latent semantic Indexing (LSI): TREC-3 Report. In D. Harmen (Ed.) *NIST Special Publication 500-226,* Gaithersburg, MD: National Institute of Standards and Technology.

Dumais, S. (2003). Data-driven approaches to information access. *Cognitive Science, 27[3],* 491-524.

Dunlop, M. D., Johnson, C. W., & Reid, J. (1998). Exploring the layers of information retrieval evaluation. *Interacting with Computers, 10[3],* 225-236.

Egan, D. E. (1988). Individual Differences in Human-Computer Interaction. In M. Helander, (Ed.), *Handbook of Human-Computer Interaction*, Amsterdam: Elsevier Science, pp. 543-568.

Ekstrom, R. B., French, J. W., & Harmen, D. (1976). *Kit of Factor-referenced Cognitive Tests,* Princeton, NJ: Educational Testing Services.

Ennis, D. M. (1988). Confusable and Discriminable Stimuli – Comment. *Journal of Experimental Psychology-General, 117[4],* 408-411.

Fox, K. L., Frieder, O., Knepper, M. M., & Snowberg, E. J. (1999). SENTINEL: A multiple engine information retrieval and visualization system. *Journal of the American Society for Information Science, 50[7],* 616-625.

Gardenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought,* Massachusetts Cambridge: The MIT Press.

Glenberg, A. M. (1997). What memory is for? *Behavioral and Brain Sciences, 20[1],* 1-7.

Harmen, D. (1993). Overview of the First Text REtrieval Conference (TREC 1). In: *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* pp. 36-47.

Harmen, D., Buckley, C., Callan, J., Dumais, S., Lewis, D., Robertson, S., Smeaton, A., Sparck Jones, K., & Tong, R. (1995). Performance Of Text retrieval Systems. *Science, 268,* 1417-1418.

Harter, S. P. (1992). Psychological Relevance and Information-Science. *Journal of the American Society for Information Science, 43[9],* 602-615.

Harter, S. P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science, 47[1],* 37-49.

Hearst, M. (1999). User interfaces and Visualisation. In R. Baeza-Yates & B. Ribeiro-Neto, (Eds.) *Modern Information Retrieval*, Harlow, England: Addison-Wesley, 257-322.

Hearst, M. A. (1995). Tilebars: Visualization of Term Distribution Information in Full Text Information Access. In *Proceedings of CHI '95 Mosaic of Creativity* New York: ACM Press, 59-66.

Herot, C. F. (1980). Spatial management of data. *ACM Transactions of Database Systems, 5[4]* 493-513.

Hornof, A. J. & Kieras, D. E. (1997). Cognitive Modeling Reveals Menu Search is Both Random and Systematic. In: *Proceedings of CHI '97: Atlanta GA,* New York: ACM, 107-114.

Jackendoff, R. (1983). *Semantics and Cognition.* Cambridge, Massachusetts: The MIT Press.

Jackendoff, R. (1987). On Beyond Zebra - The Relations of Linguistic and Visual Information. *Cognition, 26[2],* 89-114.

Jackendoff, R. (1995). *Patterns in the mind: Language and human nature,* New York: Basic Books.

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing & Management, 36[2],* 207-227.

Kieras, D. & Meyer, D.E. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human- Computer Interaction., 12,* 391-438.

Kintsch, W. (1980). Semantic Memory: A Tutorial. In: R. S. Nickerson, (Ed.), *Attention and Performance VIII,* Cambridge, Mass: Bolt, Beranek & Newman, 595-620.

Kintsch, W. (1986). Learning From Text. *Cognition and Instruction 3(2)*, 87-108.

Kwasnik, B. H. (1992). A Descriptive Study of the Functional Components of Browsing. In: *Proceedings of the IFIP TC2/WG2.7 Working Conference on Engineering for Human-Computer Interaction,* 191-203.

Lakoff, G. (1987). *Women, fire, and dangerous things; What categories reveal about the mind,* Chicago: University of Chicago Press.

Landau, B. & Jackendoff, R. (1993). What and Where in Spatial Language and Spatial Cognition. *Behavioral and Brain Sciences, 16[2],* 217-238.

Lawton, C. A. (1994). Gender Differences in Way-finding Strategies - Relationship to Spatial Ability and Spatial Anxiety. *Sex Roles, 30[11-12],* 765-779.

Lawton, C. A. (1996). Strategies for indoor wayfinding: The role of orientation. *Journal of Environmental Psychology, 16[2],* 137-145.

Lemos, R. S. (1985). Rating the major computing periodicals on readability. *Communications of the ACM, 28[2],* 152-157.

Letsche, T. A. & Berry, M. W. (1997). Large-scale information retrieval with latent semantic indexing. *Information Sciences, 100[1-4],* 105-137.

Leuski, A. & Allan, J. (2000). Lighthouse: Showing the Way to Relevant Information. In Steven F.Roth & Daniel A.Keim, (Eds.), *Proceedings of IEEE Symposium on Information Visualization (InfoVis'00)* Salt Lake City, Utah: IEEE Computer Society, 125-130.

Lin, X., Soergel, D., & Marchionini, G. (1991). A Self-organizing Semantic Map for Information Retrieval. In: *Proceedings of the fourteenth annual international ACM/SIGIR conference on Research and development in information retrieval,* 262-

269.

Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co- occurrence. *Behavior Research Methods Instruments & Computers, 28[2],* 203-208.

Marchionini, G. & Shneiderman, B. (1988). Finding Facts Vs Browsing Knowledge in Hypertext Systems *Computer, 21[1],* 70-80.

Marr, D. (1982). *Vision,* San Francisco: Freeman.

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102[3],* 419--457.

McCloskey, M. & Glucksberg, S. (1979). Decision processes in verifying category membership statements: Implications for models of semantic memory. *Cognitive Psychology, 11,* 1-37.

McNamara, T. P. & Miller, D. L. (1989). Attributes of Theories of Meaning. *Psychological Bulletin, 106[3],* 355-376.

Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science, 48[9],* 810-832.

Mizzaro, S. (1998). How many relevances in information retrieval? *Interacting with Computers, 10[3],* 303-320.

Modjeska, D. & Chignell, M. (2003). Individual differences in exploration using desktop VR. *Journal of the American Society for Information Science and Technology, 54[3],* 216-228.

Newby, G. B. (2001). Empirical study of a 3D visualization for information retrieval tasks *Journal of Intelligent Information Systems, 18[1],* 31-53.

Newell, A., & Simon, H. A. (1956). The logic theory machine: A complex information processing system. *IRE Transactions on Information Theory 2[3],* 61-79.

Newell, A., & Simon H. A. (1972). *Human Problem Solving*. Englewood Cliffs, N.J.: Prentice-Hall

Nosofsky, R. M. (1986). Attention, Similarity, and the Identification-Categorisation Relationship. *Journal of Experimental Psychology-General, 115[1],* 39-57.

Osgood, C. E. (1969). The Nature and Measurement of Meaning. In: G. S. James & E. O. Charles, (Eds.), *Semantic Differential Technique*, Chicago: Aldine Publishing Company, 3-41.

Paepcke, A., Garcia-Molina, H., Rodriguez-Mula, G., & Cho, J. (2000). Beyond document similarity: Understanding value-based search and browsing technologies. *SIGMOD Record, 29,* 80-92.

Park, S. (2000). Usability, user preferences, effectiveness, and user behaviors when searching individual and integrated full-text databases: Implications for digital libraries. *Journal of the American Society for Information Science, 51[5],* 456-468.

Putnam, H. (1998). *Reason, truth and history,* Cambridge, MA: Cambridge University Press.

Quillian, M. R. (1968). Semantic Memory. In: M. Minsky, (Ed.), *Semantic Information Processing*, Cambridge, MA: MIT Press.

Rao, R. & Card, S. K. (1994). The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information.

In: *Proceedings of the SIGCHI conference on Human factors in computing systems*, , Boston, MA: ACM Press, 318-322.

Reid, A. K. & Staddon, J. E. R. (1997). A reader for the cognitive map. *Information Sciences, 100[1-4],* 217-228.

Richardson, A. E., Montello, D. R., & Hegarty, M. (1999). Spatial knowledge acquisition from maps and from navigation in real and virtual environments. *Memory & Cognition, 27[4],* 741-750.

Robertson, G., Czerwinski, M. P., Larson, K., Robbins, D., Thiel, D., & van Dantzich, M. (1998). Data mountain: Using spatial memory for document management. In: *Proceedings of the 11th annual symposium on user interface software and technology (UIST '98),* 153-162.

Robertson, G. G., Mackinlay, J. D., & Card, S. K. (1991). Cone Trees: Animated 3D Visualizations of Hierarchical Information. In: *Proceedings of CHI '91* 189-194.

Rorvig, M. (1999). Images of similarity: A visual exploration of optimal similarity metrics and scaling properties of TREC topic- document sets *Journal of the American Society for Information Science, 50[8],* 639-651.

Ruddle, R. A., Payne, S. J., & Jones, D. M. (1997). Navigating buildings in "desk-top" virtual environments: Experimental investigations using extended navigational experience. *Journal of Experimental Psychology-Applied, 3[2],* 143-159.

Ruddle, R. A., Payne, S. J., & Jones, D. M. (1998). Navigating large-scale "desk-top" virtual buildings: Effects of orientation aids and familiarity. *Presence-Teleoperators and Virtual Environments, 7[2],* 179-192.

Ruddle, R. A., Payne, S. J., & Jones, D. M. (1999). The Effects ofMaps on Navigation and Search Strategies in Very-Large-Scale Virtual Environments *Journal of*

Experimental Psychology-Applied, 5[1], 54-75.

Salampasis, M., Tait, J., & Bloor, C. (1998). Evaluation of information-seeking performance in hypermedia digital libraries. *Interacting with Computers, 10[3],* 269-284.

Schamber, L. (1994). Relevance and Information Behavior. *Annual Review of Information Science and Technology, 29,* 3-48.

Schamber, L., Eisenberg, M. B., & Nilan, M. S. (1990). A Re-examination of Relevance - Toward a Dynamic, Situational Definition. *Information Processing & Management, 26[6],* 755-776.

Schvaneveldt, R. W., Durso, F. T., Goldsmith, T. E., Breen, T. J., Cooke, N. M., Tucker, R. G., & Demaio, J. C. (1985). Measuring the Structure of Expertise. *International Journal of Man-Machine Studies, 23[6],* 699-728.

Seagull, F. J. & Walker, N. (1992). The Effects of Hierarchical Structure and Visualization Ability on Computerized Information Retrieval. *International Journal of Human-Computer Interaction, 4[4],* 369-385.

Sebrechts, M. M., Cugini, J. V., Vasilakis, J., Miller, M. S., & Laskowski, S. (1999). Visualization of Search Results: A Comparative Evaluation of Text, 2D, and 3D Interfaces. In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 3-10.

Sense8 Corporation. WorldToolKit. [Release 7]. (1997). Mill Valley, USA.

Shepard, R. N. (1957). Stimulus and Respose Generalization: A Stochastic Model Relating Generalization to distance in Psychological Space. *Psychometrika, 22[4],* 325-345.

Shepard, R. N. (1986). Discrimination and Generalization in Identification and Classification – Comment. *Journal of Experimental Psychology-General, 115[1],* 58-61.

Shneiderman, B. (1998). *Designing the user interface: Strategies for effective human-computer interaction, 3rd edn*, Reading, Massachusetts: Addison Wesley.

Shneiderman, B. & students of cmsc 828/838 (1997). OLIVE On-line library of information visualization environments. Michael Reed and Dan Heller, (Compilers), http://www.otal.umd.edu/Olive/. Accessed 26th March 2004: 14:55.

Siakaluk, P. D., Buchanan, L., & Westbury, C. (2003). The effect of semantic distance in yes/no and go/no-go semantic categorization tasks. *Memory & Cognition, 31[1],* 100-113.

Smith, E. E. & Medin, D. L. (1981). *Categories and Concepts,* Cambridge MA: Havard University Press.

Smith, P. A. (1996). Towards a practical measure of hypertext usability. *Interacting with Computers, 8[4],* 365-381.

Soboroff, I. M., Nicholas, C. K., Kukla, J. M., & Ebert, D. S. (1997). Visualizing document authorship using n-grams and latent semantic indexing. In: *Proceedings of the 1997 workshop on New paradigms in information visualization and manipulation* New York: ACM Press, 43-48.

Sparck Jones, K. (1981). The Cranfield Tests. In K. Sparck Jones, (Ed.), *Information Retrieval Experiment*, London: Butterworths, 256-284.

Stanney, K. M. & Salvendy, G. (1995). Information Visualization - Assisting Low Spatial Individuals with Information Access Tasks Through the use of Visual Mediators. *Ergonomics, 38[6],* 1184-1198.

Tague, J. M. (1981). The pragmatics of information retrieval experimentation. In K. Sparck Jones, (Ed.), *Information Retrieval Experiment*, London: Butterworths, 59-102.

Talburt, J. (1986). The Flesch index: An easily programmable readability analysis algorithm. In: *Proceedings of the fourth annual international conference on systems documentation. Annual ACM Conference on Systems Documentation,* New York: ACM, 114-122.

Thorndyke, P. W. & Stasz, C. (1980). Individual Differences in Procedures for Knowledge Acquisition from Maps. *Cognitive Psychology, 12,* 137-175.

Tolman, E. C. (1948). Cognitive Maps in Rats and Men. *The Psychological Review, 55[4],* 189-208.

Toms, E. G. (2000). Understanding and facilitating the browsing of electronic text. *International Journal of Human-Computer Studies, 52[3],* 423-452.

Tversky, A. (1977). Features of Similarity. *Psychological Review, 84[4],* 327-352.

Van Rijsbergen, C. J. (1979). *Information Retrieval, 2nd edn,* London: Butterworth.

Voorhees, E. M. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management, 36[5],* 697-716.

Voorhees, E. M. & Harman, D. (1999). Overview of the Seventh Text REtrieval Conference (TREC-7). In E. M. Voorhees & D.K.Harmen, (eds.), *NIST Special Publication 500-250, The Seventh Text REtrieval Conference (TREC 1999)* Gaithersburg, Maryland: Department of Commerce, National Institute of Standards and Technology, 1-24.

Voorhees, E. M. & Harman, D. (2001). Overview of the Tenth Text REtrieval

Conference (TREC-10). In E. M. Voorhees & D.K.Harmen, (eds.), *NIST Special Publication 500-250: The Tenth Text REtrieval Conference (TREC 2001)* Gaithersburg, Maryland: Department of Commerce, National Institute of Standards and Technology, 1-16.

Waller, D., Hunt, E., & Knapp, D. (1998). The transfer of spatial knowledge in virtual environment training. *Presence-Teleoperators and Virtual Environments, 7[2],* 129-143.

Westerman, S. J. (1995). Computerized information retrieval: Individual differences in the use of spatial vs nonspatial navigational information *Perceptual and Motor Skills, 81[3],* 771-786.

Westerman, S. J. (1998). A comparison of the cognitive demands of navigating two- versus three-dimensional spatial database layouts *Ergonomics, 41[2],* 207-212.

Westerman, S. J., Collins, J., & Cribbin, T. (2005). Browsing a document collection represented in two- and three-dimensional virtual information spaces. *International Journal of Human-Computer Studies, 62,* 713-736.

Westerman, S. J. & Cribbin, T. (1999). Navigating virtual information spaces: Individual differences in cognitive maps. *UK Virtual Reality Special Interest Group Conference* 95-105.

Westerman, S. J. & Cribbin, T. (2000a) Cognitive Ability and Information Retrieval: When Less is More. *Virtual Reality, 5,* 1-7.

Westerman, S. J. & Cribbin, T. (2000b). Mapping semantic information in virtual space: dimensions, variance and individual differences. *International Journal of Human-Computer Studies, 53[5],* 765-787.

Willie, S. & Bruza, P. (1995). Users' Model of the Information Space: the case for two

search models. In: *Proceedings of the International ACM SIGIR Conference* 205-210.

Wise, J. A. (1999). The Ecological Approach to Text Visualization. *Journal of the American Society for Information Science, 50[13],.* 1224-1233.

Witmer, B. G., Bailey, J. H., Knerr, B. W., & Parsons, K. C. (1996). Virtual spaces and real world places: Transfer of route knowledge. *International Journal of Human-Computer Studies, 45[4],* 413-428.

Zhang, X. Y., Berry, M. W., & Raghavan, P. (2001). Level search schemes for information filtering and retrieval. *Information Processing & Management, 37[2],* 313-334.

APPENDIX I

Collins, J. & Westerman, S. J. (2001). Browsing Patterns in a Virtual Information Space Representation of a Document Database. In: *Proceedings of 9th International Conference on Human-Computer Interaction, 963-967,* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Aston University

Content has been removed for copyright reasons

Aston University

Content has been removed for copyright reasons

Aston University

Content has been removed for copyright reasons

APPENDIX II

Westerman, S. J., Collins, J., & Cribbin, T. (2005). Browsing a document collection represented in two- and three-dimensional virtual information spaces. *International Journal of Human-Computer Studies*, *62,* 713-736.

Aston University

Content has been removed for copyright reasons

Aston University

Content has been removed for copyright reasons

Aston University

Content has been removed for copyright reasons

APPENDIX III

Westerman, S. J., Cribbin, T. & Collins, J. (submitted). An Empirical Evaluation of Automatic Text Analysis Techniques.

Aston University

Content has been removed for copyright reasons

Aston University

Content has been removed for copyright reasons

Aston University

Content has been removed for copyright reasons