

Lost in hyperspace: metrics and mental models

M. Otter^{a,*}, H. Johnson^{1,b}

^aIcon Medialab, Classic House, 180 Old Street, London EC1 9BP, UK

^bDepartment of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK

Received 2 January 1999; revised 31 January 2000; accepted 5 March 2000

Abstract

Being disorientated or lost is one of the fundamental difficulties which users experience when trying to navigate within hypertext systems. In this paper, two new metrics of lostness are described and applied. The new metrics focus on the effects of link-type and the accuracy of user's mental models of system structure. In a series of studies, the new metrics were compared with the only other published metric of lostness, the optimal path deviation measure formulated by Smith [P.A. Smith, Towards a practical measure of hypertext usability, *Interacting with Computers* 4 (1996) 365–381], and with other measures including self-report estimates and task times. The results tentatively suggest that some types of hyperlink have a greater impact on lostness than others. The accuracy of the subjects' mental models did not correlate significantly with other measures of lostness, however this may have been due to task demands. Based on these findings, suggestions are made for the design of more effective hypertext systems that minimise lostness, and a new approach to designing such systems, based on the mental models of users, is put forward. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Hypertext; Hyperlinks; Metrics; Usability; Lostness; Mental models

1. Executive summary

The main strength of hypertext-based systems is that they have a flexible structure and give users freedom to browse and interact with the information contained within them. The use of hypertext is important since it offers an extremely powerful way of accessing and organising information. However, many hypertext systems do not achieve their full potential due to poor design. One of the most frustrating usability problems faced by users is

* Corresponding author. Tel.: +44-207-549-0253; fax: +44-207-549-0204.

E-mail addresses: malcolm@iconmedialab.co.uk (M. Otter), h.johnson@bath.ac.uk (H. Johnson).

¹ Tel.: +44-1225-323215; fax: +44-1225-826492.

becoming lost or disoriented. “Lostness” has therefore become the initial focus of our research on evaluating hypertext systems.

A primary aim of the research is to develop and evaluate new ways of measuring how lost people become when navigating hypertext systems. Two new metrics for measuring the lostness are proposed here. One of these metrics is concerned with the effects of different types of hyperlinks, and the second measure is concerned with the accuracy of user’s mental models and memory, in situations of lostness. A secondary aim is to gather empirical data on lostness with the objective of correlating different measures. A final aim is to make suggestions for the design of more effective hypertext systems that minimise lostness, and ultimately to produce a new methodology and supporting tool for designing hypertext systems that exhibit greater usability.

The paper discusses first the notion of lostness, and previous attempts at measuring lostness. The differences between types of link within hypertext systems and their potential effects on usability are also considered. Two new metrics are described and a series of studies are reported which assess the utility of the new metrics. The paper concludes by discussing the implications of the results of the studies for designing future hypertext systems.

Although the problem of lostness is one of the most pressing issues surrounding hypertext, very few attempts have been made to try and quantify it. Smith (1996) in the only published study to produce a quantifiable measure of lostness, argues that lostness should be viewed in terms of degradation of performance rather than the subjective feelings of users. She proposes a set of measures for assessing lostness in terms of the efficiency with which users find information in a hypertext system. The research reported here consists of further investigations related to Smith’s lostness measures. It provides an opportunity to validate her measures, and also assesses the utility of the new metrics, and correlations between the metrics. The first metric weights the hyper-links based on their likelihood of inducing lostness. It is expected that this new measure will be more accurate than Smith’s measure which assumes the likelihood of inducing lostness is the same for all link-types. The second new measure relates to the accuracy of users’ mental models, and assumes that the more accurate the model, the less the likelihood of lostness.

A number of different data collection methods were employed to gather data on lostness. A questionnaire was developed to enable users to provide subjective data on their feelings of lostness. The main causes of lostness reported by the subjects were poorly or ambiguously labelled links, confusing numbers of links or options available, the lack of ability to reverse and large site size.

In addition, two empirical studies were undertaken to assess the new lostness measures. The first study involved a number of information retrieval and mental model tasks. It was found that there was a wide range of scores on this measure indicating different degrees of lostness across subjects. The results however suggest that the new link-weighted measure of lostness does provide a more accurate measure of lostness than Smith’s (1996) measure as evidenced by correlations between scores. Although neither measure correlated significantly, the new link-weighted measure of lostness correlated slightly better with the self-report measure of lostness and with total task times, but neither was significant. The differences however, are small and this is partly due to the nature of the domain, the

hypertext system used for investigation, and also the differences in expertise with hypertext systems.

In relation to results concerning the mental model metric, there was a significant difference between the mental models of subjects when tested on three separate occasions. There was a significant difference between the mental models of subjects after three tasks and after twenty-four hours. No other differences were found for the mental model data. In particular, there were no correlations between an individual subject's scores for the mental model drawings and any of the other measures that have been used in an attempt to measure lostness. One important finding was that there were significant correlations between the accuracy of subject's mental models and their level of familiarity with hypertext systems.

A second empirical study was undertaken to provide validation of the mental models task. A further group of subjects, without system exposure, were asked to draw pathways through the system. Their results were compared with the results of the subjects who participated in the first study and it was found that performance of subjects with system exposure was superior to that of those without. This suggests that subjects with system exposure were in fact using their mental model to guide their performance on the task.

The mental model task could be used to provide the basis for a new methodology for designing and evaluating hypertext systems. This could be achieved by identifying if there are mismatches in the designer's mental model of how the hypertext document should be structured, and how a sample of users model the domain. This approach could be used as part of an iterative design process in conjunction with other usability measures. In effect it would be another tool in the HCI researchers toolbox when designing and evaluating hypertext systems. The main advantage of the mental model drawing task method is that it is quick and relatively easy to apply. This new method will support the hypertext designer in modelling the categorical structure that potential users of the system already have of the domain.

Future research aims are to apply the mental-model approach in a predictive manner to analytically evaluate the design of hypertext systems.

2. Introduction

The main strength of hypertext-based systems is that they have a flexible structure and give users a great deal of freedom to browse and interact with the information contained within them. The use of hypertext is important since it offers an extremely powerful way of accessing and organising information. However, many hypertext systems do not achieve their full potential due to poor design. One of the most frustrating usability problems faced by users is becoming lost or disoriented. "Lostness" has therefore become the initial focus of our research on evaluating hypertext systems. Being 'lost in hyperspace' is a complex problem, which limits the efficacy and promise of hypertext providing a rapid and flexible method of accessing large bodies of information. This paper addresses ways in which to measure lostness. The rationale behind this is that if you can measure lostness, then you can identify systems that induce it and users who are lost. Comparative studies between different hypertext systems can also be undertaken. Ultimately, identifying the causes of

lostness will enable people to design hypertext systems that reduce the likelihood of lostness occurring. The research reported here goes some way towards addressing this goal.

The bases for two new metrics will be on the effects of different types of hyperlinks on lostness, and the effect of the accuracy of user's mental models and memory, in situations of lostness. A secondary aim is to gather empirical data on lostness with the objective of correlating different measures. The underlying assumption is that if one measure correlates well with other lostness measures, then it gives that particular measure some degree of validity, providing the measures actually measure what they are intended to measure. A final aim is to make suggestions for the design of more effective hypertext systems that minimise lostness, and ultimately to produce a new methodology and supporting tool for designing hypertext systems that exhibit greater usability.

This paper first discusses the notion of lostness, and previous attempts at measuring lostness. The differences between types of link within hypertext systems and their potential effects on usability are also considered. Two new metrics are proposed and a series of studies are described that assess the utility of the new metrics. The paper concludes by discussing the implications of the study results for designing future hypertext systems.

3. Lostness

Being disorientated or lost is one of the fundamental difficulties which users experience when trying to navigate within hypertext systems. So great is the problem that it has been estimated that about 60% of the research into hypertext systems has been devoted solely to this issue (Dvorak and Sommerville, 1996).

Elm and Woods (1985) describe the difficulty as:

Getting lost in a display of networks means that the user does not have a clear conception of relationships within the system, does not know his/her present location...relative to the display structure, and finds it difficult to decide where to look next in the system.

The problem does not appear to be directly related to the information content contained within the hyperspace. Elm and Woods gave users of a hypertext system a set of information retrieval tasks and found that the degree of lostness experienced by subjects was independent of their level of expertise in the information domain (a manual of emergency procedures for a nuclear power plant).

As a result of their study, three different forms of being lost are outlined when speaking in terms of hypertext, as opposed to navigation per se:

1. Not knowing where to go next.
2. Knowing where to go, but not knowing how to get there.
3. Not knowing where they are in the overall structure of the document.

The problem therefore appears to be one of unfamiliarity with the structure of the document, as evidenced by problems of navigation and location, rather than the user

simply being unfamiliar with the information content or too entropic an information space.

A document's structure is known to be important. For instance, research on text comprehension in children has shown that a proper understanding of a text depends on sensitivity to the relative importance of the ideas within it, as well as an understanding of how the ideas are related (Baker and Stein, 1981). It is even thought that the ability to understand how ideas are interconnected in a text probably develops even before children learn to read (Oakhill and Garnham, 1988).

When users have problems discerning or understanding document structure, they use their existing knowledge to guide them. For instance, Jaynes (1989) discusses how readers of hypertext will use retrieval mechanisms that are already familiar to them from paper-based hard copies. When reading a paper-based medium such as a book, a reader is given many cues as to their location within the document's structure and how to find a specific piece of information. It is, for example, possible to know how far we are from the beginning or the end, which (of a series of numerically ordered) chapter we are in, what the subject matter is, and so forth.

With a book, we have a certain set of expectations or schema of what a book is, and of how it is comprised (Van Dijk and Kintsch, 1983). We know that we can look in the index to find out the location of specific information, we can look on the cover sheet to see when it was written, and we can consult the contents page to find out about chapter headings and contents. There are certain well-established conventions with books that we have learnt, and which are to a great extent, intuitive.

The same may be said for other paper-based mediums such as a phone book, a newspaper or a musical score. Each of these provides certain cues or landmarks to indicate how to find the required information within them. To this extent an analogy exists between navigation through a paper-based information space and navigation in a real physical environment. As familiarity with the text grows, the reader becomes more familiar with the various landmarks in the text and the relationships between them. The reader is in effect building up a mental representation or model of the text, based on orientation cues.

These orientation cues, which are so important for constructing a mental model of the text, are absent in many hypertext systems and as a result users may lose track of their location within the system as a whole. A lost or disorientated user of hypertext systems may also suffer from a number of other problems. Foss (1989) identified the following such problems:

1. Arriving at a particular point in a document then forgetting what was to be done there.
2. Neglecting to return from a digression.
3. Neglecting to pursue digressions that were planned earlier.
4. Not knowing if there are any other relevant frames in the document.
5. Forgetting which sections have been visited or altered.

Foss (1989) categorised orientation problems into three groups, according to their likely origins. The first group of problems is the result of what Conklin (1987) refers to as cognitive overheads. Cognitive overheads are essentially the cognitive demands placed on the user of a hypertext document. The user must decide which path to take through the

system but may find interesting side-tracks, which distract from the main task. Tripp and Roby (1990) argue that disorientation itself will lead to an increased cognitive load thus reducing the mental resources for learning:

If mental resources are engaged by navigational tasks, and if those same resources are needed for learning, it would be logical that achievement should suffer to the extent that navigation is demanding.

(Tripp and Roby, 1990).

The second group of problems arises from unfamiliarity with the structure or conceptual organisation of the network. An example of this is what Shneiderman (1987) refers to as “closure”, where the user does not know the extent of the network or what proportion of it remains to be seen.

The third group of problems is thought by Foss (1989) to be due to a general inexperience with the practice of learning by browsing. This is supposed to result in problems with remembering the information contained within the system and problems in consolidating and understanding the semantic content of the information nodes.

In attempting to ease the problems associated with user navigation, much recent research has been concerned with the use of navigational aids. However, whilst these aids help to locate information, they do not directly help users in navigating a conceptual space. Mayes et al. (1990) contend that it is important to draw a distinction between disorientation in a conceptual space and disorientation in a spatial network containing nodes and links.

Simpson and McKnight (1989) were interested in studying the relationship between the ability to navigate through a hierarchically based hypertext system and the ability to produce a map of the structure. They tested the subjects’ ability to map the structure by giving them a card sorting exercise using screen dumps from nodes of the hypertext. No correlations were found between navigational efficiency and the accuracy of the map construction. However, subjects who produced the most accurate maps were found to use fewer unnecessary cards. This is perhaps an indication that they knew something about the content and could recognise red herrings, but that the actual structure and organisation of the knowledge was not fully known or understood.

Edwards and Hardman (1989) also conducted studies related to mental models constructed by hypertext users. They undertook studies that looked at the effects of different hypertext structures on the user’s perception of the document. They concluded that users appear to be creating a mental model or cognitive representation of the hypertext structures in the form of a survey-type map or schema. This clearly has implications for the way in which hypertext systems are structured and designed.

4. Measuring lostness

The usability of hypertext has been considered from a variety of perspectives, including looking at the efficacy of navigational aids (Edwards and Hardman, 1989; Webb and Kramer, 1990; McDonald and Stevenson, 1998), categorisations of navigational and searching strategies (Canter et al., 1986), comparisons with linear texts (Leventhal et al.,

1993), and also by looking at the relationships between the structure of information and task types (Rada and Murphy, 1992). In many cases, the variables which these studies used as benchmarks of the usability of hypertext, were primarily timings of task performance efficiency, errors and self-report measures of ease of learning, ease of use, etc. None of these studies attempted to produce an accurate, quantifiable measure of lostness.

Although the problem of lostness is one of the most pressing issues surrounding hypertext, very few attempts have been made to try and quantify it. Evaluation measures typically used in HCI, such as task completion times and number of errors are inadequate in themselves as neither really deal with the fundamental issues underlying lostness. Smith (1996) provides a thoughtful rationale of why “traditional” approaches to assessing usability in HCI are inappropriate for assessing usability of hypertext systems.

Smith (1996), in the only published study to produce a quantifiable measure of lostness, argues that lostness should be viewed in terms of degradation of performance rather than the subjective feelings of users. She proposes a set of measures for assessing lostness in terms of the efficiency with which users find information in a hypertext system. She considers that focusing on deviations from the optimal pathways through the system is a more flexible approach than simply counting the number of errors.

The path measures, which Smith used to produce her measure of lostness, were as follows:

T the total number of nodes accessed;

D the number of different nodes accessed;

R the number of nodes which need to be visited to complete a task;

S the total number of nodes visited while searching;

N the number of different nodes visited while searching;

V the number of nodes visited while verifying.

T and *D* were intended to provide information on activity and certainty. It was considered that a lost user might be expected to access significantly more nodes than necessary (*R*). The final three path measures are designed to distinguish between searching and verification, based on the notion that one is not lost when verifying.

The assumption is that for a perfect search, a user will visit exactly the number of nodes required to complete an information retrieval task.

Therefore:

$$T = D = R; \quad S = T$$

$$N/S = 1; \quad R/N = 1$$

Smith’s (1996) measure of lostness (*L*) is calculated as:

$$L = (N/S - 1)^2 + (R/N - 1)^2$$

where *L* increases as lostness increases. For the perfect search, *L* = 0.

Smith’s intention was that designers should be able to incorporate the capture of information required to compute this measure, into the systems that they design and build. Results obtained from data-log measurements of actual usage could then be used to

identify problem areas in the hypertext where more links or clearer cues might usefully be added. This is quite a challenge since designers need to know the users' goals or intentions, what tasks the user is performing to achieve their goals and the information they want or need to retrieve. Therefore, data-log measurements need to be supplemented by data from other sources and user studies.

Smith (1996) never tried to statistically correlate her metric with other measures of lostness. However, she did make subjective assessments of which users she thought were lost, based on selected records from the video analysis, e.g. user comments and long pauses, and tried to relate these to values of L . Smith suggested that users with values of L less than 0.4 were not lost, but if L was any higher than 0.5 then they were definitely lost and from a closer observation declared 0.42 to be the critical value. Before concluding, Smith suggested that further work would be necessary to refine and improve her measures.

The research reported here consists of further research related to Smith's lostness measures. It provides an opportunity to validate her measures, and also assesses the utility of the new metrics, and correlations between the two new metrics. The results are also discussed in relation to user's mental models and subsequent design implications. The next section introduces two new lostness metrics.

5. Two new lostness metrics

5.1. Link-weighted lostness metric

5.1.1. Background

One of the key objectives of this research was to develop and evaluate new ways of measuring how lost people become when navigating hypertext systems. The first of two new metrics of lostness, developed as part of this research, is based on Smith's metric described above, but takes into account the fact that different types of links might have different degrees of influence on lostness. When a hypertext system is being designed, it is reasonably clear that if the designer fails to take into account the different structural aspects of the hypertext, then it is likely that the user of such a system will experience disorientation. The same must surely hold when a researcher attempts to design a measure of this disorientation.

Links are the key building blocks of hypertext. They define the relationships between nodes. The assumption made by Smith (1996) in her measure of lostness described above is that all links have the same potential to make users lost. This may well be the case, however links involve much more complicated theoretical and design issues than at first appear.

De Rose (1989) produced a taxonomy of link types, differentiated by the method of linking used. Other researchers have also generated descriptions of different link types, for example Conklin (1987), describes referential and organisational link types.

De Rose believes that links are divided into two main categories, "extensional" and "intensional". In linguistics, the extension of a concept is the set of instances that are members of the concept, whereas the intension consists of a set of attributes that define what it is to be a member of the concept. Extensional links are those which are stored and

are therefore a permanent feature of the structure of the hypertext system. Extensional links are relevant to measures of lostness based on optimal pathways.

Extensional link types are subdivided into two types, “relational links” and “inclusive links”. Relational links connect one node of information to one other node of information and no more. Conklin (1987) calls these “referential links”. There are two types of relational link, “associative links” which attach arbitrary pieces of documents and are seen by De Rose as being “entirely unpredictable”, and “annotational links”, which for example, connect pieces of text to information about that text.

Inclusive links connect one node of information to many other nodes and represent superordinate and subordinate relationships. Conklin (1987) refers to these as “organisational” links. Again there are two types; “sequential links”, which are like paths and have multiple, ordered target locations, and “taxonomic link” types which also lead to multiple target locations but do not impose any path-like ordering on them.

The psychological phenomenon of priming in memory (Anderson, 1980), suggests that a link at one level in a well-designed hypertext system, should maximise the priming effect for information links further down the hierarchy. Therefore, the theoretical rationale for developing a lostness measure that takes into account link type, relates to the possibility that De Rose’s different link types might prime associations in memory to different extents, and in turn have an effect on lostness.

Perhaps more importantly, it is quite possible that the different link types described above may offer the user of a hypertext system different degrees of predictability about the destination of that link. De Rose himself claimed that associative links were “entirely unpredictable”. If the destination of a link is hard to predict then it is assumed that this will lead to greater lostness in the user, as it makes it harder for them to predict the location of target information. It also makes it harder to form an accurate mental model of the information space. Predictability is related to work by Pirolli (1997) on scent-following and by Larson and Czerwinski (1998) on the implications of scent following for web page design. Pirolli developed computational models of scent following in very large browsable text collections. A browser was provided to users, which tersely summarised, and communicated, the common content of very large collections of information. The users of such systems had to rely on these terse representations of content as a kind of information scent whose trail was designed to lead to information of interest. The research aim of the paper was to develop computational models that are able to evaluate and understand how the summaries of information content could provide information scent to users. Pirolli (1997) does not provide a theoretical or analytical view of the nature of a “scent”.

Larson and Czerwinski (1998) argue that scent conveys distal target information via category labelling. They refer to research by Pirolli and Card (1995) on information foraging theory where scent is taken to be the amount of remote indication a user can derive, from an information structure’s design and labelling, about the relative location of a target. The purpose of Larson and Czerwinski’s paper is to investigate the depth/breadth tradeoff while attempting to design for optimal scent throughout different structures of a web site. They argue that different study results reported in the literature intended to address the depth/breadth tradeoff issue in menu design, could be due to some information structures having a stronger scent at the top levels of a hierarchy. For instance, category labels could have been more distinctive at the top level. Snowberry et al. (1983) in a study

Table 1
Weightings for De Rose's (1989) extensional link types

	Relational (1 to 1)		Inclusive (1 to many)	
	Associative	Annotational	Sequential	Taxonomic
Likelihood of inducing lostness	High	Low	Fairly low	Medium
Lostness weighting	1	4	3	2

of menu design demonstrated a strong advantage for structurally grouping like objects. They specifically showed that there were accuracy and speed advantages, not eliminated by practice, for coherent category structures as opposed to random structures. Larson and Czerwinski therefore conducted a study to investigate the effects of depth and breadth and the relationship to the category structure of the information space. Their results showed reaction times of users were faster for broader, shallower categories with distinctive category labels which maximised scent throughout a well-partitioned web structure. Smith's measure of lostness was employed and the results indicate that users were lost in the hierarchies with the most levels. Larson and Czerwinski conclude that matching category soundness and labelling to reflect users' understanding of information and thereby maximising scent strength, should augment our understanding of how best to design for large-scale information spaces.

The relevance to work reported here relates to the implication that the stronger the scent the greater the predictability, and the less lost the user becomes. It is our intention to quantitatively measure lostness taking into account different types of hyperlinks, and this consequently relates to scent strength conveyed by appropriate and accurate labelling of category structures which thence dictate the structure of the underlying information. As indicated earlier in the paper, De Rose considered some types of link, i.e. associative links to be unpredictable, which in the light of the above discussion means that they had little or no scent. However, annotational links might be construed as having a strong scent as they typically connect pieces of text to information about that text.

In order to improve upon Smith's (1996) measure of lostness it was decided to incorporate De Rose's (1989) taxonomy of links. Accordingly the four types of extensional links have been weighted by the authors (and subsequently independently agreed by an expert in HCI) in terms of the different levels of lostness that they will induce. This weighting of the four relevant types is given in Table 1 below.

Annotational links are given the lowest likelihood of inducing lostness because they link one node to one other node and the destination node is made explicit in the link label.

Sequential links are given the next lowest likelihood of inducing lostness. The reason for this is that once a user is in a path or sequence, nodes within that pathway will be chunked in the memory as a "procedural" unit, so that the exact location and content of each individual sequential node need not be remembered, thus reducing cognitive load.

Taxonomic links are given a medium likelihood of inducing lostness because they have multiple target locations and there is not the same path-like ordering imposed on them as exists for sequential links.

Associative links are given the highest likelihood of inducing lostness because they link one node to one other in a highly unpredictable and arbitrary manner.

The reason why a high likelihood of inducing lostness, gives rise to a low lostness weighting is because these weightings are intended to be applied to Smith's (1996) metric described earlier in the paper. Values from Smith's metric will be divided by these weightings and clearly dividing any number by a low number results in a high number, dividing by a high number will result in a low number.

It is the intention that the new metrics proposed will be generally applicable across different hypertext systems. The generality of the metrics relates to the assumption that hypertext systems will be comprised of different types of links with different abilities to support scent following and predictability, and this is directly reflected in one of the proposed metrics. These differences might relate to the nature of the domain, the goodness of individual category or hyperlink labels, the partitioning of the hypertext structure, the number of different routes to a target, and so on. We also believe that the classification of links will be generally applicable insofar as the taxonomy of links proposed by De Rose and Conklin are comprehensive and representative of all link types. The link types cover arbitrary, sequential, categorical and part-whole relationships, which are likely in our opinion to cover a wide variety of link types.

There were four criteria for choosing a hypertext system on which to test the metrics. First, the system had to be an actual system and freely available; second, the system must not have been previously visited by the study participants; third, the system must incorporate all the link types, and finally, it should be fairly stable (i.e. the same from the start to the end of our studies) and not subject to daily updates or amendments. The system that satisfied these criteria was the Wye College web site. It is important to note that this is a highly predictable informational system. This means that the results achieved will be a fairly strict test of the utility of the new metrics, and that the metrics will undoubtedly fare better in less predictable domains and systems. It is not the case that we believe that the metrics are only applicable in restrictive, predictable informational domains but are applicable across all systems. The reason for this is that the way in which link types are categorised is all-inclusive. These link types are fundamental attributes of hypertext systems regardless of the content of the site.

5.1.2. *The first new metric—link weightings*

The link weighted lostness metric (LWLM), is based on Smith's (1996) metric and is calculated in a very similar way:

$$\text{LWLM} = L/(\text{LW}/4)$$

where L is the Smith's (1996) lostness metric and LW is the Link weightings. LWLM increases as lostness increases. For the perfect search, $L = 0$.

The only difference between this new metric and Smith's, is that once Smith's value of L (lostness) has been calculated, it is then divided by the total link weighting for all of the nodes visited by that user, per task. In order to retain the fact that the new values will fall between zero and one, the total link weighting is itself divided by four. Thus, if the value of L is zero (indicating that the subject, according to Smith's metric was not lost at all), then the link weightings automatically do not apply as zero divided by anything is still zero. Yet

should the value of L be greater than zero, then the weightings apply as follows. If, for example, the subject when completing an information retrieval task, uses predominantly associative links (high likelihood of inducing lostness) which have a link weighting of 1, then L would be divided by a number close to a quarter ($1/4$) thus weighting L quite heavily. If the user used predominantly taxonomic links then L would be divided by a number close to a half ($2/4$). If the user used predominantly sequential links then L would be divided by a number close to three quarters ($3/4$). Lastly, if the user used predominantly annotational links (low likelihood of inducing lostness) then L would be divided by a number close to one ($4/4$) thus putting very little or no weighting on L .

5.2. *The new mental models metric*

5.2.1. *Background*

The second of the new metrics of lostness is based on the assumption that if the user has a poor mental model of the hypertext system's structure, then it is likely that they will experience disorientation. The structure of hypertext should help to reduce the user's memory load by allowing the user to chunk meaningful information. If the system is well-designed then these chunks should make sense to the user, allowing him or her to generate inferences and expectations about the structure and content of the hypertext system. Poor chunks or categorisations in hypertext within the levels of a hierarchy could lead to a poor memory for the structure of the system and thus contribute to disorientation.

Not only must the structure of the hypertext system be consistent with how the users categorise domains, it will also help to facilitate recall through priming. The semantics of links are clearly important (Bloomfield, 1994) since hypertext navigation is semantically guided. In situations where the link semantics are insufficient, users may have to combine a semantically guided search with less knowledge-dependent, more general, search strategies (Richardson et al., 1997). As previously indicated, the relevance of this to hypertext theory is that a link at one level in a well-designed hypertext system, should maximise the priming effect or scent for information links further down the hierarchy.

A schema is a mental model of the environment in which we find ourselves. This model is acquired through experience and affords us an orienting frame of reference in order to assist navigation. Downs and Stea (1977) suggest that such frames of reference exist at all levels from the global, i.e. the world, to the specific, e.g. a village or street. Schemas are rapidly acquired and once we have a generic model of what a 'city' is and what it contains, we soon know what to expect when we visit one. It is quite possible that experienced hypertext users have a schema for hypertext structures, and that this might affect how lost they become. The process by which navigational knowledge is acquired is generally acknowledged by theorists, such as Anderson (1980) and Wickens (1984), to proceed through several stages of development. Beginning initially with an identification of landmarks, followed by a development of route knowledge, leading to the acquisition of a survey type cognitive map which allows us to plan journeys along routes not previously travelled. This sequence may not hold for all individuals and all types of tasks, for example, survey knowledge is better than route knowledge for the location of objects on a map.

In an important paper, Edwards and Hardman (1989) studied the effects of different hypertext structures on the user's perception of the document structure. The results of this study led them to conclude that users appear to create mental models or cognitive representations of the hypertext in the form of schemas. The implication of this is that, in a similar way to becoming lost in a new city, if users form a poor cognitive map or mental model of the hypertext system then it is more likely that they will become lost.

It is not simply exposure to an environment that leads to the creation of an accurate mental model. Using the example of the city given above, even when we arrive in a new city we have pre-conceptions of the things that are likely to be found there and how they are likely to be organised. Norman (1983), stated that users of new systems bring to the system their past experience and attempt to build a mental model of the system in terms of this experience. Research in cognitive psychology has led us to conclude that humans impose a structure on the world by treating objects, entities and events as members of conceptual categories. By doing so we arrive at a hypertext system having already mentally divided the world into chunks, and scent following is particularly strong for conceptually categorised systems.

Leach (1964) argued that our perception of the way in which the world is structured, is brought about by us being taught as children, to "impose upon this environment a kind of discriminating grid, which serves to distinguish the world as being composed of a large number of separate things". Rosch et al. (1976) sees it from a different perspective, i.e. that the world already comes with structure and this has a direct effect on how we categorise objects, entities and events. Rosch's argument is that the role of the human categorisation system is to reflect perceived world structure in a set of categories, which provide maximum information with minimum effort. This principle she has called cognitive economy. Another major influence on the way in which we categorise the world is who we are and our cultural background. This is related to how much we know about a domain, i.e. our level of domain expertise. It is therefore important to determine how individual knowledge and expertise and cultural forces interact with the structured nature of objects and events in the world.

What does the fore-going research imply about the development of hypertext systems that lessen the likelihood of users becoming disoriented or lost? If as hypothesised by the earliest proponents (Bush, 1945), hypertext systems follow "the workings of the mind", then they should be designed as such and in doing so must take note of how domains are categorised and who it is that will be categorising it.

The categorisation of domains should be directly related to how hypertext is structured i.e. a neat fit between the mental models of the user's conception of the domain and the way in which it is structured in a hypertext system. Issues which then arise relate to who the users are likely to be and what level of domain knowledge they possess. If a system is designed so that it does not correspond with users' mental models, or support scent following, then the users are more likely to become disorientated or lost. This is due to misconceptions, incorrect inferences and expectations on the part of the user, arising from their mental models, which are not met. Further evidence for this position comes from the work of Hamilton et al. (1998), who have conducted studies on user interface design principles relating to task knowledge. The findings indicate that structuring interfaces in

a manner that does not support or even violates the conceptual structure users impose on their tasks leads to detriments in usability.

5.2.2. *The second new metric—mental models*

The first problem to overcome was to develop a procedure for measuring users' mental models. A pilot study was thought necessary in order to ensure that the measures of the users' mental models were sufficiently sharp to pinpoint any inaccuracies.

The intention underlying the pilot study was to assess the users' mental models by asking them to take part in a card sorting exercise, after completing three and then ten tasks. This procedure follows that adopted by Simpson and McKnight (1989). An experimental subject was asked to try and lay out the cards in a manner corresponding with the structure of the chosen site, i.e. their mental model of the system. The 50 cards from the hypertext system were placed in a random order, and once the task was completed the subject was asked how confident they were with what they had done.

This method of assessing the users' mental models proved problematic. First, it was very time consuming, even though the hypertext system had been reduced to 50 nodes, and this led to the subject becoming bored. Clearly, this could have a detrimental affect on the quality of the data. Secondly, and perhaps more importantly, because the subject could see the cards in front of them, it was likely that they were not only using their mental model of the system but were also receiving retrieval cues from the cards. Additionally, the contextual information of the juxtaposed cards could affect the sorting behaviour. Consequently, it was decided to develop a new way of testing mental models for hypertext systems.

The idea behind this new method was to ask subjects to draw (from memory) pathways through the system beginning at the start page and finishing at a given set point. This changed the task from recognition and sorting to retrieval and prediction. They would be asked to do this after three and ten tasks in order to test the development of their models. The end points of these diagrams would be in areas of the system which they had not visited. This means that they would not be producing their answers simply on the basis of retrieval cues, but would be using their mental models of the structure of the system to make predictions.

Asking subjects to draw their mental models of hypertext systems is in some respects not entirely new. Grey (1990) also attempted to use drawings to study mental model construction during hypertext navigation to investigate lostness. She asked subjects to perform ten information retrieval tasks from a large hypertext system and then asked them to draw their mental model of the system after one, five and ten tasks.

The difference between the new method described below and Grey's, is that Grey asked her subjects to attempt to draw their model of the whole system rather than just one pathway from the start to a specific node. From our perspective, this procedure had a major drawback in that the drawings for each subject were very different. This meant that it was very difficult to make any direct comparison between them, other than at a very general level of whether subjects viewed the structure as linear, hierarchical, grid or web. In the examples of the diagrams given in her paper, none of the nodes are labelled and the diagrams do not relate to any specific piece of information in the system. This is acceptable in that it was not Grey's intention to apply metrics to the diagrams in her study. However,

in the present study it was important to make diagrams of mental models more quantifiable in order that comparisons could be made with other methods of measuring lostness.

The diagrams produced by this new method can be quite cleanly analysed because the diagrams will be of specific pathways that actually exist in the system. It is assumed that the diagrams will consist of a representation of both nodes (e.g. boxes) and links (e.g. lines). It is then possible to count the numbers of nodes that are:

- (a) Actually drawn (AD)
- (b) Required to be drawn (RD)
- (c) Correct (C)
- (d) Correct and correctly placed (CCP), and
- (e) The number of different levels of a hierarchy correctly drawn before they make an error (LBE), i.e. the degree of hierarchical depth achieved before an error occurs.

Each of these numbers is subsequently placed into a formula, giving each diagram a specific score. The formula involves: Dividing *C* by AD to get a measure of correctness; dividing CCP by RD to get a measure of correctness of positioning; and, dividing LBE by RD to get a measure of depth of correctness.

The formula for the accuracy of mental models for hypertext systems is as follows:

$$\text{AMMH} = \frac{1}{3}(C/AD + CCP/RD + LBE/RD)$$

This formula will always give a number between 0 and 1 with a score towards 0 indicating a very poor mental model and a score towards 1 indicating a very good mental model.

This new metric and formula will be applied in the studies described in the next section of the paper.

6. Studies of lostness

In advance of reporting the different data collection and analysis methods we employed in assessing lostness, it is important to provide a rationale for the design of our studies. Over the last ten years HCI has seen a significant number of meta-evaluation studies which purport to extol the virtues of one method of assessing usability over other methods. Gray and Saltzman (1998) provide a critique of the most influential of these studies in terms of their experimental design. The problems to which they allude are concerned with validity and generality of the findings. They found evidence of studies where false conclusions were drawn from the data, where researchers had gone beyond what was investigated to provide advice to practitioners and where statistical problems had arisen due to having undertaken too many comparisons between groups or conditions. In relation to the pragmatics of the studies, they found problems with the statistics chosen, the number and selection of participants, the settings in which the studies were conducted, etc. The recommendations from Gray and Saltzman's paper were followed as rigorously as possible in the studies we report in this paper. We do not believe that we have drawn any false conclusions, gone beyond the data or provided unsatisfactory advice to practitioners. In the conclusion we state that usability is not a unitary concept, and lostness is only one facet

of usability. We suggest the use of a battery of complementary usability methods and measures. In a previous paper, the second author (Dutt et al., 1994) stressed the complementary, rather than opposing uses of different usability approaches, which together might provide a fuller picture of usability.

With respect to the current study, the subjects all came from the same population, the settings in which the studies were conducted were identical and the statistics were appropriately chosen. In relation to instrumentation we had clear methods to be employed in applying the metrics, and independent judges were also used. Moreover, we believe that the basis of many problems faced by meta-evaluation studies can be overcome by being explicit, as Gray and Saltzman suggest, about the exact operations and methods used. We have purposely given enough detail for instance about the process we followed and the test system used, to allow replication by others. In effect the studies reported here are a replication, validation and modification of Smith's metrics.

Research by Blustein et al. (1997) investigated different evaluation approaches to measuring the quality of hypertext links. The paper is of relevance here because Blustein et al. assume that the quality of hyperlinks is related to semantic similarity of the links and their targets. This obviously relates to notions of predictability and scent following. However, the paper is also relevant because it discusses the issues to be considered in undertaking experiments to assess the usability of hypertext. Blustein et al. are concerned with computer and human evaluation of hyperlink quality as two different approaches to measuring the usability of hypertext. They argue that there are circumstances where computer evaluation can be used where it is not feasible or possible for a human to evaluate the hypertext. As a result they provide a framework for graph-based evaluation of hypertext link quality based on four semantic similarity measures. We concur with Blustein et al. that computer based evaluation is given serious consideration.

However, with respect to human evaluation, their approach to experimentation in a number of ways is either not relevant to, or quite different from, ours. Whilst we are concerned with assessing and measuring lostness within a hypertext system, the approach to human evaluation that they exemplify relates to assessing the benefits of hypertext over linear text, and is therefore not directly relevant to our immediate concerns. In addition, the experimental design advocated is within groups, that is the same group of subjects appearing in more than one condition with the result that the subject acts as his or her own control. This means that there is the possibility of range or carry over effects from one condition to another. In order to control for this they advocate the use of similar but different texts in each condition and counterbalancing the order of linear and hypertext across matched subjects. However it is possible that there is asymmetric transfer between linear text to hypertext and vice versa. If within groups design are to be used (since they are more powerful), it is wise to include the subjects as an independent variable when conducting the statistical analysis. It would also be difficult to ensure the same structure, degree of semantic similarity and so on across texts. Gray and Saltzman state that there are problems with using within subjects experimental designs across different conditions and refer to Poulton (1982) for additional problems with within group designs and counterbalancing. In our view a better alternative would have been to match the groups in a similar manner but then exposed those different groups to different treatment conditions thereby keeping the text, etc. the same.

Table 2
Causes of lostness reported without prompting

Cause of lostness	Number of times reported	% of people reporting
Poorly labelled or ambiguous links	10	45
Confusing number of links / options	6	27
Lack of ability to reverse	5	23
Large site	3	14
Lack of ability to return to original link	2	9
Lack of recognisable start page	2	9
Forgetting to put in bookmarks	2	9
Continuing a train of thought along many links	2	9
Location of the links on the page	2	9
Information required is not listed in site index	2	9
Bad page structuring	1	5
Bad site structuring	1	5
Pictures that are non-obvious links	1	5
Getting distracted by attractive looking links	1	5
Too much information when searching	1	5
Don't know the information space structure.	1	5
Unfamiliarity with the system	1	5
More than one link to same destination	1	5
Too much information presented	1	5
Loading time	1	5
Missing links	1	5
Links not following a "logical order"	1	5
Not keeping a tab on how many pages visited	1	5
Browsing	1	5
Pages with links to many unrelated topics	1	5
Frustration	1	5

In the first empirical study to be reported in this paper all the subjects experienced identical treatment conditions and tasks. We did not systematically vary any factor and observe the effects. The variations in the data only came about as a result of the analyst applying different metrics to the study results. In the second empirical study an entirely different group of subjects undertook the mental model tasks in order to allow us to compare the results.

A number of different data collection methods were employed to gather data on lostness. First, a questionnaire was developed to enable users to provide subjective data on their feelings of lostness. Secondly, two empirical studies were undertaken to assess the new lostness measures. A post-study questionnaire was also to be completed by users to provide further data relating to their experience in the first empirical study.

6.1. Questionnaire

The main focus of the questionnaire was concerned with what people thought were the common causes of them becoming lost or disorientated when using hypertext systems. It was hoped that the results would not only be interesting in themselves but would also be invaluable in informing the design of the first empirical study. Beyond this it is hoped that

Table 3
Causes of lostness reported when asked to circle options

Cause of lostness	Number of times reported	%
Size of site	11	50
Unfamiliar domain	10	45
Predictability of link destinations	9	41
Unfamiliar words or language used	9	41
Hierarchical structure	7	32
Non-task centred structure	6	27
Index structure	5	23
Mixed structure	5	23
Type of hyperlink used	4	18
Expertise with hypertext systems	3	14

they would also provide information for the generation of future design principles and also to see if they provide corroborating data or an explanatory account of why lostness occurs.

The subjects were first asked what their level of familiarity with hypertext systems was, in order to get an idea of the subject sample. They were then asked what they thought were the common causes of them becoming lost, followed by questions regarding particular hypertext systems that they remembered induced feelings of lostness and why they thought this was.

On the reverse of the questionnaire was a further question about the causes of lostness, but on this occasion the subjects were given a restricted set of ten possibilities and were asked to circle those with which they agreed. The reason for asking a similar question twice and putting it on the reverse, was so that if subjects had not been able to recall their own possible causes of lostness, the questionnaire would still provide some relevant data.

The last question on the questionnaire asked subjects to “circle the number of information retrieval tasks that they would typically perform when looking at any one particular hypertext-based system”. The aim of this question was to inform the design of Study 1, in particular the number of information retrieval tasks they would be set.

Twenty two people completed the questionnaire. The average age of the subjects was 24 years. The subject group had a wide range of familiarity with hypertext systems.

6.1.1. Questionnaire results

In total, 52 responses were given to the questions concerned with common causes of lostness, and the responses were grouped under 26 separate categories. These results are given in full in Table 2.

The main cause of lostness reported by the subjects was poorly or ambiguously labelled links, which was reported by 45% of the subjects. The second highest cause was having a confusing number of links or options available (27%). Third came the lack of ability to reverse (23%), followed by large site size (14%). Each of the remaining twenty-two categories of responses were reported by two (9%) or fewer subjects.

When the subjects were asked to circle “which of the following causes lostness?” from a

selection of ten possible causes (Table 3), they circled size of site most frequently (50%), closely followed by unfamiliarity with the domain (45%), predictability of the link destinations, and unfamiliar words or language used (both 41%). Expertise with hypertext systems received the lowest percentage of reporting (14%). Interestingly, the type of hyperlink used was not seen as being particularly problematic (18%).

The two sets of results from page 1 and 2 of the questionnaire are interesting with respect to the comments voluntarily made by participants versus their views when asked specific questions. The size of the site was only spontaneously mentioned by 14% of participants, whereas 50% acknowledged that this could be a problem when prompted. Unfamiliarity with the domain was not spontaneously considered to be much of a problem, mentioned by only 5% of participants, but was considered to be a problem when specifically asked as a question, by 45% of participants. Predictability of link destinations was also not volunteered as a problem, but when specifically asked about, 41% of participants thought that this was an important cause of becoming lost. This finding is not independent of the type of hyperlink, which only achieved a mention by 18% of participants. Unfamiliar words or language used, mentioned by 41% of participants is one finding where there is some correspondence between responses to specific questions, and volunteered responses, related to poorly labelled or ambiguous links. The vast majority of subjects (77%) stated that they would typically perform between five and ten information retrieval tasks when using hypertext systems.

The principal findings concerning what subjects regarded as being the main causes of lostness, will be used as a basis for questions in a further questionnaire given to subjects in Study 1, once they have completed a set of information retrieval tasks on the hypertext system. For example, the finding that poorly labelled or ambiguously labelled links were regarded as being one of the main causes of lostness will be used to ask: “Were any of the links particularly badly labelled or ambiguous?”.

6.2. Empirical studies on lostness

6.2.1. Background

The two new metrics described previously, essentially pose two different but closely related questions. If lostness is indeed related to the way in which links are used to structure a hypertext system, is this due to something inherent about the semantics of the link and its destination, or is it more to do with the way in which the link is applied? Does the linking structure go counter to what the user would expect and if so, is lostness more related to the degree to which these expectations are confounded? Users come to a system not as a ‘tabula rasa’ but with a whole set of pre-conceptions of how an information domain should be structured and this is influenced by individual differences in knowledge, expertise and as indicated previously, conceptual structuring.

The new metrics will be compared with other measures, primarily that formulated by Smith (1996), which also attempts to capture some aspect of lostness.

If the link-weighted measure of lostness provides higher positive correlations with other lostness measures than Smith’s measure, then this suggests that lostness in terms of hypertext structure is connected to inherent characteristics of link types. In particular some link types, because of a greater or lesser degree of ambiguity in their supposed

destination, and lack of scent could have a greater impact on lostness than others. Therefore, a primary aim underlying Study 1 is to validate Smith's measures, and to assess if the link-weighted measure provides a more accurate measure of lostness.

A secondary aim relates to the mental model metric. Having a poor mental model of a system's structure is assumed to be due to a confounding or a mismatch between user expectations and perceptions of the system structure, and the structure itself. If lostness is related to users having a poor mental model of the system, then users who perform badly in tasks designed to assess the accuracy of their mental models, are likely to be those who are lost, assuming that those tasks are valid. Therefore, it is expected that there will be a relationship between the accuracy of the mental models of subjects and the degree of lostness or disorientation they experience when using hypertext systems.

The relative influence of perception over expectation of system structure will be drawn out by looking at the accuracy of mental models after different levels of exposure to the system structure. Other measures that are assumed to capture elements of lostness against which the metrics will be compared, are task completion times and a variety of self report measures (e.g. "How lost were you?" and "How easy was the system to use?").

Two empirical studies are reported in this section. The first relates to both new measures of lostness, and involves the study participants undertaking three activities: (i) completing a series of information retrieval tasks; (ii) drawing pathways through the system on three separate occasions; and (iii) completing a post-study questionnaire. The second study to be reported is a validation of the mental models tasks employed in the first study.

6.2.2. *Empirical study 1*

This study was designed to provide data relating to our research aims and to assess the two new ways of measuring lostness.

6.2.2.1. Subjects. Twelve subjects took part in the study, ten male and two females. The average age of the subjects was 23 years of age. The subject sample included both experienced and inexperienced users of hypertext systems.

In this study the independent variables were the three metrics (Smith's metric and the two new metrics), and the dependent variables were task times and responses to the questionnaire items. The experimental design was within subjects. The ordering of the tasks was held constant across subjects.

6.2.2.2. Procedure. As indicated previously, in this empirical study, the subjects were involved in three activities: carrying out a series of information retrieval tasks; drawing pathways through the system after tasks three and ten, and after the post-study questionnaire; and finally, completing the post-study questionnaire.

The hypertext system chosen, as discussed earlier, was Wye College web site. Before each study began, subjects were first given a brief description of the domain. Each subject interacted with a networked Pentium based PC with the Wye College web site home page preloaded on the screen. The subjects were given an instruction sheet outlining the study procedure. They were then asked to use the web site to complete a set of ten information retrieval (IR) tasks concerning Wye college.

6.2.2.3. Information retrieval tasks. There were two types of IR task. First, there were eight tasks set for the subjects by the experimenter. These required the user to traverse different depths in the hypertext structure, some of the tasks required the location of information only two levels deep, whereas others required investigation to five. The complete set of tasks and the task ordering is given in Appendix A. The ordering of the tasks was constant across subjects to prevent the results from becoming confounded by learning effects.

Secondly, there were two user-directed IR tasks. These were information retrieval tasks that the users set themselves before they saw the system but after they had been given a description of the domain. Allowing the users to undertake tasks of their own choosing meant that to some degree subjects could direct the acquisition of their own mental models of the system structure. Additionally, it was considered to mirror a realistic scenario where users might be unfamiliar with a specific system and its inherent structure, but have existing knowledge or preconceptions of the domain, and could then generate further queries to ask of the system. Having read and understood the instruction sheet, the subjects were given the following scenario:

You have arrived at the home page of Wye College, part of the University of London, which specialises in agriculture and the environment.

They were then asked to try and think of two pieces of information about the college, or courses offered by the college, about which they might like to gather information. As they carried out the tasks, the subjects were invited to give a concurrent verbal protocol, explaining throughout, what they were doing and why. A video camera was used to record on-screen navigation, timings and the subjects' concurrent verbal protocols. After each task they were returned to the start point, which was at the top level of the hierarchy. This is one important difference between the procedure used in this study and the procedure used by Smith (1996). Smith admitted that the failure to do this meant different subjects started tasks from different points depending on where they were.

6.2.2.5. Mental model drawing tasks. After three and ten IR tasks, the subjects were given the mental model task sheets. The mental model task involved subjects drawing pathways through the system from the start point, to where they thought they would find information on X (the final specified information node), using boxes to represent individual pages and lines to represent links between them.

6.2.2.6. Post-study questionnaire. Once they had completed all these tasks, the subjects were asked to complete a post-study questionnaire, which was designed to elicit further information about their experiences interacting with the system. The questionnaire included questions specifically designed to find out more information about lostness. Some of the questions were included on the basis of data gathered from the previous questionnaire, while others were adapted from Smith's (1996) original questionnaire.

Having completed the questionnaire the subjects were given one final mental model task to perform. This was given to them in a sealed envelope for them to complete the

Table 4
Examples of link labels, types and weightings

Link labels	Link type	Link weighting
Online catalogue	Associative	1
Departments	Taxonomic	2
Next page	Sequential	3
Table version (of library opening times)	Annotational	4

following day. The purpose of this was to see if the memory traces making up their mental models remained after a significant time delay.

6.2.2.7. Procedure for data analysis. The procedure for getting the data into a form suitable for statistical analysis involved the following steps. First, the information on the videotapes was coded; secondly, each link followed by the users had to be classified and given the appropriate weighting; finally, the mental model diagrams were coded. The first two of these are explained in more detail below, the coding procedure for the mental model diagrams was explained fully in Section 5.

6.2.2.8. Coding of video tapes. In total there were between six and seven hours of videotaped recordings to observe and code, half an hour on average per subject. The video recordings were used to record on-screen navigation, task times and the subjects' concurrent verbal protocols.

In order to calculate both Smith's (1996) and the link weighted measure of lostness it was necessary to identify from the video tapes, exactly which nodes the individual subjects had visited and the links which they used to get there. This was undertaken for each IR task.

During the analysis of the video-tapes, any comments made by subjects while carrying out their IR tasks relating to lostness were noted in order to provide some qualitative data.

6.2.2.9. Calculating link weightings. Calculating link weightings accurately was an important part of the analysis procedure. These weightings would later be applied to the information taken from the analysis of the video-data. The data consisted of which links each subject had used when navigating the hypertext system, and this data would be used to calculate the link weighted measure of lostness.

Overall the process was an objective one and involved checking each link used by subjects in the study against De Rose's (1989) taxonomy. Table 4 below shows examples of links within the system, the type to which they have been assigned, and the associated weighting. Further examples of weightings of links used by subjects in the study are given in Appendix B. The Wye college site used in this study, comprised mainly taxonomic (medium likelihood of inducing lostness, weighting 2/4) and annotational (low likelihood of inducing lostness, weighting 4/4) links. This is likely to be due in part to the nature of the domain. The implications of the link-types in the domain is that the link-weighted lostness measure, as indicated in Section 5.1 will be less effective than it might have been if there had been many associative (high likelihood of lostness) links. Therefore, any

findings which indicate that it provides a higher positive correlation with other lostness measures in this study, are a fairly conservative estimate of its utility.

It is not possible to guess the link type from the link label, since it depends at least in part on the information in the destination node. Occasionally, there is a difference between what the link suggests its type is, and what it actually is. For example, there is a link called ‘back to index page’ which would be classified as taxonomic. In fact, it is inappropriately labelled and takes the user to a map and is therefore annotational. This is problematic for scoring purposes. Should links be scored in relation to what they actually are, in terms of where they take the user, or scored in terms of what they suggest? In order to ensure the validity of the link typing it was necessary to identify whether other HCI experts agree with the way in which De Rose’s (1989) classification has been applied in each case. It was therefore important to get an independent judge to apply the weighting system for the site in question.

An expert in HCI was given the original paper (De Rose, 1989) and asked to pay particular attention to the link taxonomy and how it might apply to link types. They were then given the web site address to enable access to the site used in the main study, a list of all the links which the subjects used in the main study, and asked to classify them in terms of De Rose’s link taxonomy.

When the two sets of weightings for the same set of links were compared there were only three out of a total of 54 links where there was any difference between the two evaluations. The two evaluators discussed these differences once the evaluations were complete, and agreement was reached on each.

The differences were all centred around the nature of the links to a ‘non-clickable’ image map of the university campus. The independent evaluator assumed incorrectly that it was possible to click on the ‘Department of Engineering’ as seen on the map, and this would be a link to that department, which would make actually clicking on the map taxonomic instead of annotational. In the light of this error it was agreed that this link was annotational. As the map was divided into two sections, this accounted for two of the differences between the evaluations. The third difference was also due to the nature of the map. The independent evaluator had seen the link to the map page (as opposed to actually clicking on the map itself) as being annotational. It was agreed through discussion that since the map was in two sections the link had to be taxonomic.

All the data were then in a format appropriate for comparing with the other variables, the questionnaire responses and task times.

In the next section the results for the individual measures in terms of their range and averages will be discussed. The results for the mental model tasks will be discussed as a subsection of this. This will be followed by analysis of results relating to the research aims, and the section will end with a discussion of other interesting findings from the study.

6.2.3. *Study 1 results*

First, the results of the IR tasks will be given, followed by results for the post-study questionnaire, and the mental model tasks.

6.2.3.1. *Information retrieval task results.* The table of means below (Table 5) demonstrates that the average Link Weighted Lostness value was 0.44. There was a

Table 5
Averages, minimums, maximums and standard deviations for key variables

Variable	Mean	Min	Max	SD
Link weighted lostness measure (0 = Lost, 1 = Not lost)	0.44	0.13	0.71	0.18
Smith's (1996) lostness value (0 = Lost, 1 = Not lost)	0.26	0.01	0.42	0.11
Total task times (seconds)	1363	1025	1681	234.33
How lost or disorientated did you feel overall? (0 = "Not lost at all", 10 = "Very lost")	3.83	2	8	2.03
Did you try and keep a mental note of where you were in the system? (1 = yes, 0 = no)	0.58	0	1	0.51
How easy was the system to learn? (0 = "Very hard", 10 = "Very easy")	6.42	1	9	2.19
How easy was finding the information needed? (0 = "Very hard", 10 = "Very easy")	5.92	1	8	1.97
How helpful was the start page? (0 = "Very unhelpful", 10 = "Very helpful")	5.83	1	8	2.04
How familiar are you with using web sites like this? (0 = "Very unfamiliar", 10 = "Very familiar")	6.08	2	9	2.31
How frustrating was the system to use? (0 = "Very frustrating", 10 = "Not very frustrating at all")	6.33	2	9	2.53
How easy did you find it to return to the start page? (0 = "Very hard", 10 = "Very easy")	8.67	4	10	1.82
How easy did you find it to reverse / go back? (0 = "Very hard", 10 = "Very easy")	6.58	2	10	2.46

wide range of scores on this measure indicating different degrees of lostness across subjects. The individual scores on the Link Weighted Lostness Measure are given in Appendix C. The average for Smith's lostness value was 0.26. This value was inevitably lower than the new measure because of the weightings. As expected, given the relationship between Smith's measure and the new lostness measure, the same subjects were lost according to both measures. In terms of task times, there was a ten minute difference between the fastest and the slowest subjects. It is important to note that task times include the time taken to provide concurrent protocols, and are therefore only indicative of realistic task times. People who reported themselves as lost took a longer time to undertake the tasks, which indicates that lostness is the cause of the longer task times whether or not that time is taken up by providing a protocol on how lost users feel, or in their attempts to overcome their lostness. The average total task time was 22 minutes and 43 seconds.

6.2.3.2. Post-study questionnaire results. For the results of the post-study questionnaire, please refer to Table 5 below. All questions were scored out of 10.

When asked "How lost or disorientated did you feel overall?" the mean response was 3.83. However, the mean conceals wide variation in user's subjective reports of being lost. There was a wide range of values for this important question ranging from 2 (not very lost) to 8 (very lost).

On average slightly more subjects stated that they tried to keep a mental note of where

they were in the system, than those who did not. In general, subjects found it fairly easy to learn how to use the system (mean = 6.42), and fairly easy to find the information needed (mean = 5.92). The start page was found to be fairly helpful (mean = 5.83). The subjects found it very easy to return to the start page (mean = 8.67) and fairly easy to go back from whence they had come (mean = 6.58).

There was a wide range of familiarity with hypertext systems within the subject sample.

6.2.3.3. Mental model task results. The mental model diagrams that the subjects produced were, as requested, all box and line diagrams. The new method of measuring mental model accuracy produced a full range of scores from 0 to 1. The mean scores across all subjects were .47 after the completion of three tasks, .68 after the completion of ten tasks and .8 after 24 hours. The purpose of testing the accuracy of subject's mental models after 24 hours was to see if the memory traces, which formed the basis of the mental models of system structure, remained after a significant time delay. The results suggest that this was the case.

A repeated measures Analysis of Variance (ANOVA) was performed to test whether there were any differences between the accuracy of the mental models of the subjects after three tasks (mean = 0.47), ten tasks (mean = 0.68) and twenty four hours (mean = 0.8). This test was chosen not only because it is robust to violations of its assumptions but also because the data appeared to have homogeneity of variance and to be normally distributed. There was a significant difference between the mental models of subjects when tested at the three points ($F = 5.71$, $df = 2$, $p < .01$). The direction of these differences suggests that the subjects' mental models of the system are improving.

Three related one way t-tests were then performed on this data, which showed that there was a significant difference between the mental models of subjects after three tasks (mean = 0.47) and after twenty four hours (mean = 0.8) ($t = 4.88$, with $df = 11$, $p < 0.001$). No other differences were found for the mental model data.

These results lead us to ask questions about the task and the mental models the users are developing. Why, for instance, did the mental models improve over a 24 h delay? Could this be due to increasing familiarity with the particular hypertext system chosen, or with hypertext structures in general where more learning could occur as a result of consuming less mental resources on navigation (Tripp and Roby, 1990)? Does the 24 h delay allow users to re-conceptualise and mentally practice their experiences with the hypertext system? Alternatively, the results could be a consequence of an experimental artefact, that is, that the task of producing the drawn mental models has become easier and this accounts for the improvements. These questions will be returned to at the end of Section 6, and provided the motivation for the second empirical study which was designed as a validation of the mental models task.

6.2.3.4. The drawings. A secondary aim of the study was to investigate whether there would be a relationship between the accuracy of the subjects' mental models, and the degree of lostness or disorientation which users experience when using hypertext systems. The study was designed to test this and the results are given below.

There were no correlations between an individual subject's scores for the mental model drawings and any of the other measures that have been used in an attempt to measure

Table 6

Significant correlations between the link weighted measure of lostness and other variables

	<i>r</i>	<i>p</i>
Smith's (1996) measure of lostness	0.96	0.0001
Did you keep mental note of where you were in the system?	0.56	0.028
How easy did you find it to reverse/go back?	-0.54	0.034
How helpful was the start page?	-0.54	0.034

lostness. A number of possible explanations for this finding are discussed in detail in the General results section.

One important finding, related to our questions about the validity of the mental model task, did arise from this study. The results revealed that there were significant correlations between the accuracy of subject's mental models and their level of familiarity with hypertext systems. This finding was found after three tasks ($r = 0.62$, $p = 0.016$) and also after 24 hours ($r = 0.5$, $p = 0.05$). It is also the case that subjects producing a high (correct) mental model score achieved a low self-report score of lostness, and vice versa.

6.2.4. General results

6.2.4.1. Results relating to the link-weighted measure of lostness. In order to provide an assessment of whether the link-weighted lostness measure is a more accurate measure of lostness, analyses were performed correlating the variables under examination with each other. Spearman's Rho tests were used, as the data was ordinal rather than interval or ratio.

There is some tentative evidence to suggest that the link-weighted measure of lostness provides a more accurate measure of lostness than Smith's (1996) measure as evidenced by correlations between scores. Although neither measure correlated significantly, the link-weighted measure of lostness correlated slightly better ($r = 0.47$, $p = 0.061$) with the self-report measure of lostness (i.e. how lost did you feel overall?) than Smith's ($r = 0.4$, $p = 0.097$). It also correlated slightly better with total task times than Smith's, but again neither were significant. The differences however, are small and might be attributable to the nature of the domain, the hypertext system used for investigation, and also the differences in expertise with hypertext systems.

Table 6 below shows the variables that correlated significantly with the link-weighted measure of lostness. The highest correlation was between the new measure and Smith's measure ($r = 0.96$, $p < 0.0001$), which was expected as other than the link weightings, the two measures are identical. This high correlation also reflects the fairly low likelihood of lostness induced by the predictable nature of the hyperlink types in the informational hypertext system used in our empirical study. If the links had been primarily associative and taxonomic then the correlation could have been much lower.

One key finding of interest was the high correlation between the new measure and the subjects' response to the question in the questionnaire regarding whether they tried to keep a mental note of where they were in the system while using it ($r = 0.56$, $p = 0.028$). This provides some tentative support for a relationship between lostness and mental models.

There were negative correlations between the link weighted measure and the questions ‘How easy did you find it to reverse/go back?’ and ‘How helpful was the start page?’ (for both, $r = -0.54$, $p = 0.034$). This lends some validity to the measure, as both of these are crucial to navigation. It is reasonable to suggest that subjects with navigational problems or subjects who do not find the navigational tools helpful, are more likely to get lost. Both metrics suggest that there was a wide range of values of lostness, with the new metric inevitably giving higher values in all cases.

One of the aims of this research was to collect both qualitative and quantitative data on lostness. The post-study questionnaire provided useful and informative data on the relationships between factors that may influence lostness.

Self-report measures, such as the responses given by subjects to the questionnaire, can be unreliable. For example, confidence judgements in decision-making tasks are known to have no relationship to the accuracy of those decisions (Otter and Vrij, 1995). In many cases such as in the case of attempting to measure lostness, where there is no definitive measure, self-reports provide an important benchmark.

Subjects’ responses to the question ‘How lost or disorientated did you feel overall’, which was one of the key questions, correlated with the following:

	<i>r</i>	<i>p</i>
Task time	0.63	0.014
How easy was the system to learn?	-0.69	0.006
How easy was it to find information?	-0.64	0.013
How helpful was the start page?	-0.65	0.011

There were also correlations between ‘How frustrating was the system to use?’ and:

	<i>r</i>	<i>p</i>
How easy was the system to learn?	0.57	0.026
How easy was it to find information?	0.72	0.004
How helpful was the start page?	0.50	0.047
How easy was it to reverse/ go back?	0.70	0.005

‘How easy was the system to learn’, correlated with:

	<i>r</i>	<i>p</i>
How frustrating was the system to use?	0.57	0.026
How easy was it to find information?	0.71	0.005
How lost did you feel overall?	-0.7	0.006
How helpful was the start page?	0.69	0.006

In addition to the correlations mentioned above, ‘How easy was it to find information?’ correlated with ‘How helpful was the start page’ ($r = 0.63$, $p = 0.014$).

6.2.5. Empirical Study 2: validation of mental model tasks

A low score on the mental model tasks could well be a reflection of a poor mental model of the hypertext system. This is the assumption on which the mental model tasks are based. However, a low score could be due to a number of other variables, such as boredom or not fully understanding the task demands, but most importantly it could be due to the invalidity of the drawing task as a measure of mental models.

In order to check the validity of the mental model tasks given to subjects in the previous study, it was decided to give exactly the same set of tasks to a new set of subjects who had not used nor seen the Wye college web site. In effect, we were asking them to guess the location of the information given the domain. It would then be possible to compare the two sets of results for the mental model tasks to see if there were any differences between them. The findings will also shed some light on how well-designed and structured the Wye College site is, and how predictable the links. In addition, it will be possible to provide data which may shed light on why the mental models improved significantly after three tasks and a 24 hour delay.

If there were no differences between the mental models of subjects with system exposure and those without, then the validity of the earlier results could be questioned. The findings could be explained in terms of the demand characteristics of the tasks. However, if differences are found between the mental models of exposed versus non-exposed subjects then this supports the hypothesis that use of the system does provide the subjects with a mental model of the kind represented in the drawings they produce.

6.2.5.1. Subjects. Twelve subjects took part in this study, nine male and three female. The average age of the subjects was 24 years of age. None of the subjects from Study 1 took part in Study 2.

6.2.5.2. Procedure. Subjects were given the same three mental model task sheets as for study 1. The only difference was that subjects in this study were asked to imagine how the site would be structured even though they had never seen it. Before beginning the subjects were allowed to ask the experimenter any relevant questions that would help them to clarify the nature of the task that they had been set.

6.2.5.3. Results. Table 7 below shows the mean scores on the three mental model tasks for subjects with and without system exposure. For all three tasks the performance of subjects with system exposure was superior to that of those without.

Three independent *t*-tests were performed on the data. This test was applied since we were interested in differences between the two data sets, the data satisfied the criteria of an interval scale of measurement, and different subjects were used in the two conditions.

All three of the *t*-tests were significant. There were significant differences between subjects with and without system exposure in the accuracy of their mental model drawings for all three mental model tasks. For mental model task 1: $t = 8.38$ (df 11, $p < 0.005$). For mental model task 2: $t = 2.24$ (df 11, $p < 0.05$). For mental model task 3: $t = 3.54$ (df 11, $p < 0.01$).

As indicated previously, the motivation underlying this second empirical study was to investigate the validity of the mental models task, and to provide an account of why the

Table 7

Mean scores on the three mental model tasks for subjects with and without system exposure

	With system exposure	Without system exposure
Mental model task 1	0.47	0.39
Mental model task 2	0.67	0.41
Mental model task 3	0.8	0.54

mental models task results improved after a 24 h time delay. Three accounts were considered; increasing experience with hypertext systems, re-conceptualisation of their understanding and experiences of the system, and the task being an experimental artefact. Before considering each of these accounts it is necessary to consider the differences between the group with exposure and those without.

The groups without system exposure are using their existing knowledge and pre-conceptions about the domain to guide their task. Therefore, they have some predictable (given the domain chosen) knowledge about the information content but not of how the information is structured and partitioned within the hypertext system. This was reflected in the mental model task results. Essentially, they are using their existing knowledge and conceptual structuring of the domain to guide them in designing pathways (i.e. how the system is structured). However, they are receiving no confirmation or feedback on whether or not they are correct.

In contrast, the groups with system exposure not only have their existing knowledge of the domain to guide them, but also knowledge, through exposure, of the hypertext system structure and have had the benefit of their existing knowledge being confirmed or rejected. It is possible that they are building up mental models using nodes as landmarks, they are able to test hypotheses about their intuitions and expectations and also receive feedback. The mental resources they consume in navigating through the hypertext structure lessens, thus leaving more resources for learning.

The three accounts generated each provide a partial explanation of the results. It is likely that familiarity with the hypertext system partially accounted for the improvements experienced by the group with system exposure, both through experience with the system and the re-allocation of the mental resources used for navigating to learning. It is also the case that they come to learn the system better and have a more accurate mental model as a result of further exposure as evidenced by the rise in scores from task 1 (= 0.47) to task 2 (= 0.67). The corresponding rise in the group without exposure was 0.39–0.41. As indicated there were significant differences between the two groups for all three mental model tasks. This indicates that the groups with exposure were developing mental models and were using these in the mental model tasks. In this sense the mental model tasks were a valid measure. However, the nature of the mental model tasks also provided a partial explanation for the results, that is, the subjects in both groups were getting better at undertaking the tasks, and/or the tasks were becoming simpler. We can make this conclusion because the scores for both groups increased by the same amount over the 24 h delay period (the increase is 0.13) but greater for the group without exposure in proportion to their initial scores. Therefore, system exposure makes a difference to how well the groups do throughout the tasks but training in the mental model task accounts for much of the

increase in scores over the 24 h time delay. It is obvious that further systematic investigation must be carried out to identify the origin of the higher scores, i.e. did they each become more correct with establishing actual nodes, their position in the hypertext structure, their level of depth, and so on. It is also obvious that training on the tasks is vital, and we do not have any satisfactory explanatory account as to why the task suddenly becomes more simple for both groups with a delay period.

7. Discussion and implications

First, the results of the studies are summarised and then more general issues relating to this research are discussed. Both of the following sections concentrate on the implications that these findings have in terms of an advancement of theory, and of producing methods for the design of hypertext systems, which lessen the likelihood that users will become lost.

With respect to the difference between Smith's (1996) measure of lostness and the link-weighted measure of lostness, it was found that the link-weighted measure of lostness correlated slightly better with other variables also assumed to be associated with lostness. In particular, the new measure correlated better with self-reports of how lost subjects felt overall and with task times. This provides tentative evidence to suggest that different types of hyperlinks do induce different degrees of lostness, especially for subjects relatively unfamiliar with hypertext systems. Even in a hypertext system with mainly annotational and taxonomic links, (in this respect a low and medium likelihood of inducing lostness and correspondingly low and medium scent strength), there was some evidence that the semantics of links do have some impact. We postulate that the nature of the connection between the link semantics and lostness appears to be related to the degree of predictability about the destination of that link, and this relates to how well the user is supported in following a scent.

The main implication of this suggestion is not that hypertext systems should be built and structured in a way that excludes the use of the link types (e.g. the "entirely unpredictable" associative links (De Rose, 1989)) most likely to cause lostness. To suggest this would be both unwise and unnecessary since certain domains are quite likely to be structured in a way that necessitates their use, as are some user tasks. To proscribe the use of certain types of link would be counter-productive, as one of the great strengths of hypertext is the flexibility it offers for structuring information.

The real implication of this suggestion is that the use of associative links in particular, should be avoided where they are not strictly necessary, where their use is not directly tied to task demands, or the inherent structure of information within the domain. Because taxonomic links connect one node of information to many others, they can also be problematic and care should be taken in assigning an appropriate link label to reflect category structure of the information, and thereby maximise predictability and scent following ability. A new way in which designers can be assisted in this particular task is presented in the general discussion and conclusion section.

The results from the pre-study questionnaire showed that people regard poorly or ambiguously labelled links as one of the main causes of lostness. This viewpoint seems

to be supported by literature on scent following Larson and Czerwinski (1998) and the results of the main study. Designers of hypertext systems should consider the semantics of the link labels with great care as poor link labelling can mean that users find it difficult to tell which links are relevant to their task. There is nothing new in this recommendation, indeed it is fairly intuitive. What this study suggests however, is that for some types of link accurate labelling is even more crucial than for others, i.e. predictability and scent following are not equal across link types.

The purpose of trying to measure lostness as accurately as possible is so that this measure can then be used as a benchmark of usability. Early iterations of designs can be compared with later versions to assess the influence of design changes on lostness. Data logs could be used, in conjunction with other empirical methods, to identify where users were lost and ideally, redesigns could be based on improving the problematic areas. We appreciate that identifying where users become lost, what factors led to them becoming lost, and how to overcome the problem is a major undertaking. However, providing ways of measuring lostness, in a systematic and sensitive manner, is a step in the right direction.

Smith (1996) observed from her results that values of L less than 0.4 corresponded to experimental subjects who were not lost and that where L was greater than 0.5 the subjects were definitely lost. This in itself could be used as a benchmark figure for design. In Smith's study, closer observation of subject's behaviour on the video record, showed that with all values of 0.42 or more there was evidence of lostness.

The results of this present research include one instance where $L = 0.41$ and the video evidence shows that at least on some tasks the subject in question, appeared to be lost. This user commented "I don't really know" having returned to the start point three times, before eventually completing the task in 5.23 minutes (average for task is 2.39 min). The exact specification of a level of L or LWLM over which users are lost, does not make much practical sense. A band of values for L , i.e. 0.4–0.5 is perhaps more appropriate. For LWLM a corresponding range of values is harder to offer as it clearly depends on the type of links being used. Yet since the two metrics are scaled against one another and since both are measured between 0 and 1, the same range of between 0.4 and 0.5 could be used as a benchmark over which it is likely that the user is lost.

It was expected that there would be a relationship between the accuracy of the mental models of subjects and their degree of lostness. However, there were no correlations between the subject's scores for the mental model drawings and any of the other measures that were used in an attempt to measure lostness.

There are at least two possible conclusions that can be drawn from this finding. Either the result is due to problems with the mental model tasks themselves, or the subjects were not producing mental models of system structure of the kind that the mental model tasks were designed to measure. It is quite possible that there are problems with the mental model tasks as they currently stand. The task involves a mixture of prediction and to a lesser extent, memory. Subjects are asked to predict and draw pathways to specific information in a hypertext system structure, based on limited knowledge of that structure.

Inaccurate mental model diagrams could be due to a number of factors. One factor relates to the design of the particular hypertext system chosen. For instance, if the system does not constitute predictable links, then the prediction aspects of the task will be particularly difficult. This observation leads us to another possible explanation for the

lack of a relationship between the accuracy of mental models and the degree of lostness. The results could be related to the “goodness” of the design of the hypertext system. It is quite possible that it is easier to conceptualise, make predictions about, and draw pathways through well-designed systems. Conversely, it is likely to be much harder to construct and generate mental models of poorly designed systems that do not concur with user expectations. The results, as expected, highlight problems with the design of the hypertext system chosen, given the predictable nature of the domain.

Another factor that could result in inaccurate mental model diagrams could be due to subjects misunderstanding what exactly was required of them. For example, it may have been the case that there was some confusion as to whether the subjects felt they ought to draw ‘how the system was’, versus ‘how they thought it ought to be’. Additionally, some subjects may have found this task harder than others, and therefore, the mental model score given to subjects may have been more an indication of their ability to understand and carry out task demands than the accuracy of their mental models of the system structure.

One result from the study, which lends support to this, is the finding of a significant correlation between the accuracy of subject’s mental models (as measured by the mental model task diagrams) and their level of familiarity with hypertext systems. Subjects who were familiar with hypertext systems and presumably therefore with the way, in which, they can be structured, produced better diagrams than those did with little familiarity with hypertext systems. These subjects could also benefit from their experience of, and familiarity with, the task of browsing and searching in hypertext systems to learn and retrieve information. The finding that subjects with experience of hypertext systems produce better diagrams supports the notion that subjects unfamiliar with hypertext systems may have found the drawing tasks harder.

An alternative experimental procedure which could be adopted if it was necessary to control for experience with hypertext systems, is to allow subjects time to familiarise themselves with a range of hypertext systems. However, in terms of design practice, it is important to take into account the needs of users who have not previously used hypertext systems.

A secondary aim of this research was to gather as much empirical data on lostness as possible. A large volume of data was collected during the course of the studies described above and some of the most interesting and theoretically relevant findings are outlined below.

The pre-study questionnaire provided invaluable data on what people in general think are the causes of lostness. The main finding from this questionnaire was that the subjects in the study considered link semantics (ambiguous/poorly labelled links) to be more important (45% of respondents) in causing lostness than problems directly related to link structuring (too many links—27%, inability to reverse—23%). Overall the 22 subjects who took part in this study named 26 possible causes of lostness which provides support for the idea that lostness may not be a unitary concept with a unitary measure and a unitary solution.

The post-study questionnaire responses provided much useful data. In particular it was found that there were strong correlations between self-reports of lostness and task times, system learnability, ease of finding information, and helpfulness of the start page. The helpfulness of the start page appeared to be another crucial variable, correlating not only with self-reports of lostness but also with system learnability and how frustrating the system was to use.

This suggests that when designing systems and attempting to produce accurate and

unambiguous link labels, and careful link structuring, extra care should be taken on the start or home page. It is perhaps at this point that the potential to reduce lostness and increase scent is at its greatest. The research of Rosch et al. (1976) on the ‘basic level’ of representations of conceptual hierarchies in terms of categorisations is of particular relevance at this point in the discussion. Rosch’s argument was that the role of the human categorisation system is to reflect perceived world structure in a set of categories, which provide maximum information with minimum effort and that there was one level, the basic level, where this was particularly pertinent. This principle she has called cognitive economy. The basic level is assumed to be the most convenient for cognitive activities such as memory, perception and learning. It is assumed that there is a close relationship between the principle of cognitive economy and the principle of cognitive overheads, described by Conklin (1987) as the cognitive demands placed on users of hypertext systems. If there is such a close relationship then maximising cognitive economy at the start page should help to reduce cognitive overheads and thereby increase scent and reduce lostness.

8. General discussion and conclusion

One of the main findings from this study is that there is no one clear measure of lostness. Although the link weighted measure of lostness produced slightly better correlations than Smith’s (1996) non-link weighted measure against other measures assumed to be associated with lostness, it certainly did not explain all the variance in the data. This could have been partly due to the relatively small sample size, the types of tasks that were selected and the highly predictable nature of the domain and hypertext system chosen. Although it would be preferable to have used a larger sample size, the number used (12) was certainly sufficient to provide a large set of valid data. Care was also taken to use a set of representative tasks. However, the results indicate that much of the variance in the data is more likely to be due to the nature of lostness.

It is unlikely that any single measure will ever capture all that it means to be lost. If there is no single way of measuring lostness then it makes sense to use a battery of measures which correlate well with one another and which have been shown through empirical studies and in the literature to measure lostness to some degree. This battery could then be used to highlight areas in a system where problems lie and consequently, the design could be improved. Thus metrics used independently, and as a battery, can be used as a tool in the design of hypertext systems.

8.1. Predicting lostness

One important issue that needs to be discussed is whether the findings from this research or the new measures that have been produced can be used to predict which systems are more likely than others to make their users lost. The answer to this is that they can.

As has been mentioned throughout this discussion, the tentative results suggest that different types of hyperlink induce lostness to different degrees, and thus the measure could well be used to predict lostness. Hypertext systems that are predominantly structured using associative links are, because of the inherent unpredictability of the link destination, more likely to cause lostness. The same applies to a lesser extent for systems

that predominantly use the other three types of hyperlink. This comment still holds, even though the use of such links may be unavoidable due to the way in which the information or the users' tasks, are structured.

The link-weighted measure allows predictions to be made, but not directly. It can be used to highlight particular areas of existing designs that are causing users to become lost. The problems in these particular areas could well be related to poor design practices already well known in the HCI community, but the use of this metric could uncover other aspects of poor design practice, which are specifically related to lostness. If applied to a wide variety of systems and users, it is likely that problematic aspects of designs that cause lostness can be highlighted. New systems can then be checked against a list of these problematic aspects of design in order to predict which hypertext systems are most likely to cause lostness.

The mental model metric could also be used indirectly to make predictions about lostness, however the mental model tasks could have a much more direct application for design. This is discussed in more detail in the next section.

8.2. The mental model task as a new methodology for designing hypertext systems

The mental model tasks could be used as a new methodology for designing and evaluating hypertext systems. This could be achieved by seeing if there are mismatches in the designer's mental model of how the hypertext document should be structured and how a sample of subjects models the domain.

It could be used as part of an iterative design process and perhaps could also be used when other usability measures have discovered weaknesses in a design. In effect it would be another tool in the HCI researchers toolbox when designing and evaluating hypertext systems. The main advantage of the mental model drawing task method is that it is quick and relatively easy to apply.

The research by Rosch et al. (1976), suggests that we as humans and users of hypertext systems, already have an in-built conception of how information in domains should be structured. Taxonomic links, by their very nature, connect one node of information to many others, and consequently designers should take care in assigning an appropriate link label. A new way in which designers can be assisted in this particular task could be along the lines of the mental model tasks in empirical Study 2, where the subjects were asked to draw pathways through the system having never previously seen it. This method could be applied prior to any design or implementation, as a part of the requirements analysis.

If this new method were to be applied it would allow the hypertext designer to have some idea of the categorical structure that potential users of the system already have of the domain. When asked, "Which bit of the system caused you to be the most lost?" one subject in the main study commented that "the link typology was awful". Perhaps if this method had been employed during the design of the Wye College web site then fewer subjects would have shown signs of being lost.

The studies reported here suggest that training would need to be given in the mental model tasks themselves. One of the main differences between the mental model tasks in the main study and the same task in the validation study, is the cognitive demands of asking subjects to draw pathways through structures which they had never seen. Training

would be necessary in order to reduce the effects of misunderstanding the nature of the tasks. Training experienced hypertext users would not need to consist of more than a couple of practice attempts. Those with low familiarity could be “walked through” a couple of web sites. The mental model tasks are not intrinsically more difficult than for instance, requesting subjects to draw a map from their house to the shops via a particular street.

The mental model tasks could be extended to investigate how potential users mentally represent and categorise domains and the knowledge within those domains. The extension would include a more thorough study of these categorisations and domain knowledge, in order to improve the design of hypertext structures. Users’ domain knowledge could be gathered, analysed and modelled as Domain Knowledge Structures (DKS): The DKSs would be closely related in structure to Task Knowledge Structures (TKS: (Johnson et al., 1988; Johnson and Johnson, 1991)). Specific tasks undertaken within domains would be represented as TKSs within the appropriate DKS(s). The intention is that these DKSs would inform designers of the domain knowledge users possess which guides their expectations, inferencing and problem solving. These in turn will partially dictate their interaction with, and navigation through, the hypertext system. This new DKS approach could be applied not only to hypertext systems but also to the labelling of menus, as an extension to the recent work by Richardson et al. (1997), and Larson and Czerwinski on scent following.

The underlying philosophy of this mental models approach is that there will be consistency and invariance in some of the guesses, expectations and inferences generated by users, due to how they categorise domains and the knowledge within domains. This knowledge, coupled with universal aspects of cognition and information processing, should dictate aspects of user behaviour when interacting with hypertext systems. Consequently, this knowledge can be exploited by constructors of hypertext systems. There will, of course, also be differences in domain knowledge and expertise which will also play a part in their interaction with the system. This approach ensures that the user is still in control, and there is consistency in the hypertext system that can be utilised by users in generating further expectations and hypotheses.

An alternative approach to structuring hypertext has been proposed by Chen (1998). Chen argues that closely related documents or nodes could be placed near each other and only intrinsic connections between them could be shown to users as automatically generated “virtual” links. The “relatedness” or similarity of documents and nodes could be based on hyperlink usage, content similarity and usage patterns. The outcome would be a “self-inorganised” information space which could be transformed based on, for instance, usage patterns. Consequently, the structure of the information space is incrementally tailored to users’ search and browsing styles. A similar approach is proposed by Calvi and De Bra (1998) with respect to hypertext for learning. They consider that student learning could be enhanced by hypertext courseware that automatically modifies its’ link structure during the student’s learning process. A further suggestion is that there could be changes to information content in addition to link availability.

Both Chen’s and Calvi & De Bra’s approaches are quite different to the approach we are advocating. They are suggesting individual tailored systems which may involve the use of AI techniques to adapt the hypertext system to individual users, and that the hypertext

structure is not static. There have been CHI panel debates about the efficacy of the use of intelligence at the interface with both proponents (e.g. Maes) and antagonists (e.g. Shneiderman). It would be interesting to assess and compare the outcomes of the two different approaches.

8.3. Future research

This research has opened up a number of avenues for future research. The measures and approach employed here need to be applied, replicated and refined with different systems and different user populations. If the findings remain constant across these studies than it is probable that these are factors, or combinations of factors, that result in lost users.

There were two possible approaches to discovering whether or not the types of hyperlinks as specified by De Rose (1989) affect lostness. This present research was one of the two. The alternative would have been to investigate a number of different web sites each with a predomination of one of the four main types. The reason why this approach was rejected was due to the multitude of uncontrollable and potentially confounding variables, such as site size and differences in domain. It would be possible to create systems, but creating them so that the only difference between them was the linking structure would create an artificial structure for a hypertext system. However, this approach could be used to test specific hypotheses about lostness which might have arisen from the studies reported here. The artificial nature of a specially created hypertext system would have to be taken into account in any discussion as to the validity and generality of the findings (Gray and Salzman, 1998) and those findings would need to be corroborated by more naturalistic studies, which also have both advantages and disadvantages.

A totally naturalistic study of lostness such as accumulating log data from a 'walk up and use' hypertext system already in use, is one option. Users themselves would be involved in choosing the information tasks, which would allow them to direct their own acquisition of mental models. After using the system, they could be asked what information they were looking for, etc. However, it is important to note that log data does not give adequate indications of how well-designed a system might be unless user tasks and intentions are known.

As discussed above, if the general approach to measuring mental models used in this research, is to be developed for system design purposes, or as a means of assessing usability, then much more research needs to be conducted. In particular the following issues need to be considered: What exactly is being measured, how the measure can be improved, how it can best be applied, what problems do people encounter when performing these tasks, why and at which point was this the case? It might be interesting to look at the effects of link type on the mental model tasks too. Did subjects have particular problems drawing the diagrams and if so at what points? Do these points correlate with different types of hyperlink? These are research issues we are currently investigating.

If the poor design of hypertext systems does have an influence on lostness, it would be interesting to look at lostness in relation to other poor design practice by comparing data on lostness to other usability measures, such as inconsistency. In the hypertext system used in this study there were a number of inconsistencies in the overall design. Certain departments had simply added information on their department to the main body of the Wye college web site,

but done it so that the overall ‘look and feel’ was inconsistent. Perhaps it was in these areas where the subjects were most lost. This is an important area for future investigation.

Acknowledgements

We are grateful to the people who were willing to participate in the studies. We are also grateful to two anonymous referees for constructive and helpful reviews of the paper.

Appendix A. Information retrieval tasks

Set tasks:

1. How far is Wye from London?
2. How do you get from the Biochemistry laboratories to IT services?
3. What is the email address of the chief editor of the environmental newspaper?
4. What time does the library close on a Sunday in term time?
5. You are studying for a BSc in Biology. What is the title of the most recent book on your recommended reading list for your course in environmental physiology and behaviour?

User-directed task A (example of user’s specific task choice)

1. How many courses must the students on the distance learning MSc in Agricultural Economics successfully complete?
2. Which of the contents pages in the Department of Biological Sciences Postgraduate Guide for 1997/8 has a yellow title?

User directed task B (example of user’s second choice of task)

1. What is research in the “Farm Business Unit” concerned with?

Examples of User-directed tasks.

1. How many students are there at Wye College?
2. How do you get to Wye College?
3. How do you get a postgraduate application form?

Appendix B. Examples of link types, weightings and labels

Associative (link weighting = 1)

Course Outline
On-line catalogue

Taxonomic (link weighting = 2)

About Wye College

Academic Services
Department of Agriculture—staff

Sequential (link weighting = 3)

Next Page, sequential, 3

Annotational (link weighting = 4)

About the FBU
Accommodation and catering
Wye and local area

Appendix C. Individual scores on the New Link Weighted Lostness Measure

Subject 1—0.13
Subject 2—0.42
Subject 3—0.25
Subject 4—0.65
Subject 5—0.53
Subject 6—0.71
Subject 7—0.22
Subject 8—0.39
Subject 9—0.49
Subject 10—0.35
Subject 11—0.65
Subject 12—0.43

References

- Anderson, J., 1980. *Cognitive Psychology and its Implications*, W.H. Freeman, San Francisco, CA.
- Baker, T., Stein, N.L., 1981. The development of prose comprehension skills. In: Santa, C., Hayes, B. (Eds.). *Children's Prose Comprehension: Research Practice*, International Reading Association, Newark.
- Bloomfield, H., 1994. *Links in hypertext: An investigation into how they can provide information on inter-node relationships*. PhD thesis, Queen Mary and Westfield College, University of London, UK.
- Blustein, J., Webber, R.E., Tague-Sutcliffe, J., 1997. Methods for evaluating the quality of hypertext links. *Information Processing and Management* 33 (2), 255–271.
- Bush, V., 1945. As we may think. *Atlantic monthly* July, 101–108.
- Calvi, L., De Bra, P., 1998. A flexible hypertext courseware on the Web based on a dynamic link structure. *Interacting with Computers* 10, 143–155.
- Canter, D., Powell, J., Wishart, J., Roderick, C., 1986. User navigation in complex database systems. *Behaviour and Information Technology* 5, 249–257.
- Chen, C., 1998. Generalised similarity analysis and pathfinder network scaling. *Interacting with Computers* 10, 107–129.
- Conklin, J., 1987. Hypertext an introduction and survey. *Computer* September, 17–41.
- De Rose, S., 1989. Expanding the notion of links. In: *Proceedings of the Hypertext'89 conference*. Chapel Hill, North Carolina.
- Downs, R., Stea, D., 1977. *Maps in Minds: Reflections on Cognitive Mapping*, Harper and Row, New York.

- Dutt, A., Johnson, H., Johnson, P., 1994. Evaluating evaluation methods. In: Cockton, G., Draper, S.W., Weir, G.R.S. (Eds.). *People and Computers, IX*, pp. 109–121.
- Dvorak, R., Sommerville, S.T., 1996. Hypertext/hypermedia—a review. QMW Technical Report.
- Edwards, D.M., Hardman, L., 1989. “Lost in hyperspace”: cognitive mapping and navigation in a hypertext environment. In: McAleese, R. (Ed.). *Hypertext: Theory into Practice*, Intellect books, pp. 105–125.
- Elm, W.C., Woods, D.D., 1985. Getting lost: a case study in interface design. *Proceeding of the Human Factors Society ACM Press*, 927–931.
- Foss, C.L., 1989. Tools for reading and browsing hypertext. *Information Processing and Management* 25, 407–418.
- Gray, W., Salzman, M.C., 1998. Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human Computer Interaction* 13, 203–261.
- Grey, S.H., 1990. Using protocol analyses and drawings to study mental model construction during hypertext navigation. *International Journal of Human–Computer Interaction* 2, 359–377.
- Hamilton, F., Johnson, P., Johnson, H., 1998. Task-related principles for user interface design. Paper presented at the Schaerding workshop on Task analysis. *Schaerding* 2–4.
- Jaynes, J.T., 1989. Limited freedom: linear reflections on nonlinear texts. In: Barrett, E. (Ed.). *The Society of Text: Hypertext, Hypermedia and the Social Construction of Information*, MIT Press, Cambridge, MA, pp. 148–161.
- Johnson, P., Johnson, H., Waddington, R., Shouls, A., 1988. Task related knowledge structures: Analysis, modelling and application. In: Jones, D.M., Winder, R. (Eds.). *From Research to Implementation, People and Computers, IV*. Cambridge University Press, Cambridge, UK.
- Johnson, H., Johnson, P., 1991. Task knowledge structures: psychological basis and integration into system design. *Acta Psychologica* 78, 3–26.
- Larson, K., Czerwinski, M., 1998. Web page design: Implications of memory, structure and scent for information retrieval. In: *Proceedings of CHI’98*.
- Leach, E., 1964. Anthropological aspects of language: animal categories and verbal abuse. In: Lenneberg, E.H. (Ed.). *New Directions in the Study of Language*, MIT Press, Cambridge, MA.
- Leventhal, L.M., Teasley, B.M., Instore, K., Rohlman, D.S., Farhat, J., 1993. Sleuthing in HyperHolmes: An evaluation of using hypertext vs. a book to answer questions. *Behaviour and Information Technology* 12, 149–164.
- Mayes, J.T., Kibby, M.R., Anderson, T., 1990. Learning about learning from hypertext. In: Jonaeen, D.H., Mandl, H. (Eds.). *Designing Hypermedia for Learning*, Springer.
- McDonald, S., Stevenson, R., 1998. Navigation in hyperspace: An evaluation of the effects of navigational tools and subject matter expertise on browsing and information retrieval in hypertext. *Interacting with Computers* 10, 129–142.
- Norman, D.A., 1983. Some observations on mental models. In: Genter, D., Stevens, A. (Eds.). *Mental models*, Laurence Erlbaum Associates, Hillsdale, NJ.
- Oakhill, G., Garnham, T.S., 1988. *Becoming a Skilled Reader*, Blackwell, Oxford, UK.
- Otter, M., Vrij, A., 1995. Improving the accuracy-confidence relation in eyewitness identification: The impact of retrospective self-awareness and contradictory reasons. *Expert Evidence* 4 (2), 56–59.
- Pirolli, P., Card, S., 1995. Information foraging in information access environments. In: *Proceedings of the conference on Human Factors in Computing, CHI’95*, ACM Press, Denver, CO, pp. 51–58.
- Pirolli, P., 1997. Computational models of information scent following in a very large browsable text collection. In: *Proceedings of CHI’97*, ACM Press, Atlanta, Georgia, pp 3–10.
- Poulton, E.C., 1982. Influential companions: Effects of one strategy on another in the within-subjects designs of cognitive psychology. *Psychological Bulletin* 91, 673–690.
- Rada, R., Murphy, C., 1992. Searching versus browsing in hypertext. *Hypermedia* 4, 1–31.
- Richardson, J., Howes, A., Payne, S.J., 1997. An empirical investigation of memory for routes through menu structures. In: *Proceedings of Interact’97*, pp. 348–354.
- Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M., Boyes-Graem, P., 1976. Basic objects in natural categories. *Cognitive Psychology* 8, 382–439.
- Shneiderman, B., 1987. *Designing the User Interface Strategies for Effective Human Computer Interaction*, Addison-Wesley, Wokingham, UK.
- Simpson, C. McKnight, 1989. Navigation in hypertext: structural cues and mental maps. In: *Proceeding of the Hypertext II Conference*.

- Smith, P.A., 1996. Towards a practical measure of hypertext usability. *Interacting with Computers* 4, 365–381.
- Snowberry, K., Parkinson, S.R., Sisson, N., 1983. Computer display menus. *Ergonomics* 26 (7), 699–712.
- Tripp, S.D., Roby, W., 1990. Orientation and disorientation in a hypertext lexicon. *Journal of computer based instruction* 17, 120–124.
- Van Dijk, T.A., Kintsch, W., 1983. *Strategies of discourse comprehension*, Academic Press, New York.
- Webb, J.M., Kramer, A.F., 1990. Maps or analogies? A comparison of instructional aids for menu navigation. *Human factors* 32, 251–266.
- Wickens, C., 1984. *Engineering Psychology and Human Performance*, Charles Merrill, Columbus, OH.