Evaluation of user interface designs for information retrieval systems: a computer-based experiment

Paul Jen-Hwa Hu a,*, Pai-Chun Ma b, Patrick Y.K. Chau c

Abstract

In this study, we conducted a computer-based experiment to evaluate and compare the effectiveness of six different interface designs, graphical or list-based, in supporting communication of an object’s “relevance” from an information retrieval (IR) system to its users. We adopted the Model Human Processor to provide a necessary framework to incorporate relevant cognitive psychology theories and user-centered design principles in the development of different interfaces. The study had a well-researched theoretical foundation, complied with relevant design principles, and included a large-scale empirical evaluation. Our results suggest that interface design may have a significant effect on system–user concept communication, regardless of users’ familiarity with the search task, and that a graphical user interface may be more effective in supporting such communication than a list-based design. Furthermore, we also examined the cognitive load and user satisfaction resulting from each investigated interface design. Findings of the study have important implications for the design of IR systems (including online library systems and Internet-based search systems) as well as for the information representation and visualization of knowledge management systems, which ordinarily depend on text-based display methods to support system–user concept communication. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: User interface design and evaluation; Human–computer interaction; Information retrieval; System–user concept communication

1. Introduction

The contemporary information era has been characterized by rapid information growth and expansion that can be viewed as arising from, as well as contributing to, proliferation of information retrieval (IR) systems, including online library systems and Internet-based search systems [2,44]. Digitalization of the existing vast corpus of information has increasingly accelerated and, at the same time, new information has been generated and integrated at a nearly exponential rate. Retrieval and representation of relevant objects from the sheer volume of information repository have become growing concerns for both organizations and individuals [6,17]. Desire for effective information access and sharing across organizational or geographical boundaries have made manual-based retrievals increasingly ineffective or obsolete, making IT-enabled IR become progres-
sively a norm rather than an exception. As a result, the need for effective online IR support is enormous [34,43].

User interface design is a critical aspect and an essential component of IR systems [34], knowledge management systems, and other emerging information systems (IS). For example, a user interface is necessary for knowledge/information representation and visualization in knowledge management systems and interactive multimedia IS. Most previous IR research has focused on development of search algorithms, data structure, and indexing techniques and has offered limited discussion of user interface design [33]. Consistent with the advent of IT developed for direct use by the general public, a trend towards designing IR for end-users rather than the once-targeted professional intermediaries (e.g., librarians and domain experts) has become increasingly prevalent [42]. In this connection, an IR system should include an effective user interface through which users can interact with the system to complete their search tasks successfully.

A user interface involves the interaction between a system and its users, which often takes place in the form of concept or message communication [31,34]. When using an IR system to retrieve relevant information, known or unknown, a user communicates to the system a concept defined by the specified search word (anchor term) or search criteria. Concept communication also proceeds in the other direction, from a system to its users. A common example is an IR system’s communicating to a user the significance or relevance of a system-suggested item with respect to the current search task. Representation of a concept initiated at either end of the interaction is important in system–user communication and has been identified as a challenge for online IR systems [20,34]. Many IR systems have used text as the predominant concept representation method. Lancaster and Warner [34] suggested that text may not be effective in supporting system–user concept communication. Many commercial online retrieval systems have failed to serve users effectively, in part because they incorporate interfaces that have not been well-accepted by users [22]. Together, the need for alternative concept representation methods and the inappropriate real-world development practices for IR systems demand additional investigative efforts to evaluate user interface designs, preferably involving different concept representation methods for supporting system–user communication.

This study evaluated the effectiveness of six different interface designs in supporting system–user concept communication. Each interface included a set of system-suggested items relevant to a specified search word (anchor term), represented using a graphical or text-based method. Our research approach followed the Human Information Processing (HIP) paradigm [36]. Specifically, we adopted the Model Human Processor [12] as our research framework to incorporate relevant cognitive psychology theories [1,25,32,40] and user-centered design principles [22,23] for different interface designs, graphical and text-based. We evaluated each design’s effectiveness in supporting system–user concept communication. Specifically, we focused on an IR system’s communication to users the relevance [23] of a system-suggested item. A common abstraction in system–user communication, the concept of relevance is a continuum and denotes the extent to which a system-suggested item matches the user-provided search word or specifications. A large-scale computer-based experiment was conducted to evaluate and compare the effectiveness of the investigated interface designs in supporting communication of a system-suggested item’s degree of relevance generated by an IR system to its users.

The organization of the remainder of the paper is as follows. Section 2 discusses the significance of the study, reviews relevant representative literature, and highlights our research motivation. Section 3 describes our research questions and details our research method, design, and process, and is followed by discussion of data analysis results and their implications in Section 4. The paper concludes, in Section 5, with a summary, discussion of its contributions and limitations, and suggestions for some future research directions.

2. Research significance, literature review, and motivation

In this section, we discuss the significance of the study, review representative previous research related to the study, and highlight our research motivation.
2.1. Research significance

An IR system is a system that is capable of storage, retrieval, and maintenance of information [33]. Typically, an IR system has a module-based architecture, consisting of a set of components or subsystems. Common components of an IR system include document selection, indexing, vocabulary, searching, matching, and user interface [34]. This study concerned the user interface subsystem, consisting of features that support the interaction between an IR system and its users [35]. While understanding the importance of other user interface features that include search method and online instruction, we specifically concentrated on evaluating concept representation visualization in system–user communication.

Although it is a widely recognized important system design issue, user interface has received surprisingly limited attention from previous IR research, in part because of many factors that include multidiscipline requirements and problem complexity [10,47]. Closely related to user interface design, human–computer interaction (HCI) has progressively become a discipline of its own, representing a major conjunction of several reference science and technology areas that include IS, computer science, and cognitive psychology [37,45]. In his review of HCI development over the past 40 years, Shackel [45] identified two fundamental approaches: human-oriented and computer-oriented. Increasing market competition (partially resulting from experiences with systems of poor usability), together with a general trend towards the general public’s becoming mainstream technology users, has propelled the popularity and importance of the human-oriented approach. Thus, bridging the need for additional research in interface design for IR systems with existing relevant literature, including cognitive psychology and HCI, is highly desirable; addressing a practical IR system design issue on one hand and extending the validation of findings from the referenced literature on the other.

Based on users’ knowledge about the search target(s), information search can be classified into two broad categories: known-item and unknown-item search [48]. In a known-item search, users know precisely the object(s) for which they are looking. On the other hand, users embarking on an information search may not, and often do not, know exactly the objects needed. Unknown-item search is particularly prevalent and can be problematic in situations where users have to traverse multiple search spaces (topic areas) and/or have limited knowledge about the search domain. In a typical unknown-item search, an IR system often suggests a set of items presumably relevant with respect to the user-provided search word and/or specifications. Representation of these system-suggested items in response to a user’s unknown-item search often may pose a greater and more widespread challenge to interface design for IR systems than displaying results of a known-item search. As suggested by Yee [48], evaluation of the usefulness of alternative arrangements and representations of multiple items for different search tasks is an important direction for future research in user interface design for online IS.

2.2. Literature review

Information presentation/visualization has received considerable research attention from several disciplines. For instance, IS researchers have investigated this issue extensively, particularly in the early 1980s, when fairly easy-to-use and economically affordable computer graphics became available and use of computer-generated graphical reports began to prevail in the business world. DeSanctis [18] examined the relationships between use of computer-generated graphics and decision making and singled out several important research directions. Ives [27] provided general discussion of use of graphical interfaces in IS for business applications. Empirical evaluations of the effects of computer-generated graphics on decision making have also been conducted. For instance, Benbasat and Dexter [8] performed laboratory experiments to evaluate the effects of color and alternative information presentation methods on user perception and decision making under varying time.
constraints. In a related study, Benbasat and Dexter [7] assessed the decision quality resulting from use of different information presentation methods, graphical charts and tabular tables, and compared the respective time requirements.

User behaviors have also been investigated. Benbasat et al. [9] examined variations in information usage associated with different information presentation methods. Influences of information presentation on users’ comprehension have also been assessed. For example, Lucas [38] conducted experiments to examine and compare differences in information comprehension resulting from use of graphs and traditional tables. Dickson et al. [19] compared the effects of computer-generated graphical forms and tabular reports on information interpretation accuracy. Information readability and the resulting learning effects have been investigated as well. Dickson et al. [19] compared the readability of bar charts and conventional tables. Lucas and Nielsen [39] evaluated the effects of information presentation on individual learning and performance. While many studies suggested computer-generated graphics to be an effective information presentation method, support for this conclusion lacked consistency. Jarvenpaa and Dickson [28] provided a fairly comprehensive review of previous information presentation research in IS, identifying and discussing characteristics of situations that have been shown to benefit from use of computer-generated graphics.

Concept representation has been an important research issue in HCI, which can be analyzed from the HIP perspective. In a HIP paradigm, humans are considered as information processors, consisting of such essential components as memory, processors, and intercomponent connections. According to the HIP-based Model Human Processor [12], a human processor, at a system level, has several subsystems that include the perceptual, the cognitive, and the motor. Closely related to concept representation is the conceptual subsystem that carries sensations (e.g., visual stimuli) detected by the sensory system in the external world to the perceptual processor, which in turn transforms the detected sensations into symbolically coded internal representations stored in the perceptual memory.

Representation of perception-based knowledge/information can be classified into two broad schemata: image-based and linear-ordering-based [1]. An image-based schema uses a graphical representation that encompasses one or several prominent visual properties to communicate an intended concept, whereas a linear-ordering-based scheme relies on display sequence for such communication. In this connection, multiple system-suggested items in response to a user’s unknown-item search can be represented by and communicated to the user using a graphical or list-based interface.

Several predominant properties essential to viewers’ visual arousal and the resulting semantics of their internal information/knowledge coding have been examined. For instance, Anderson [1] investigated the relationships between the size of an object and the resulting visual arousal and semantics and suggested that viewers’ perceived intensity and significance of the visual arousal produced by an object may increase with its physical size; i.e., the size of an object, measured by area or volume, may have a positive correlation with its significance as perceived by viewers. Similarly, the relationships between distance and visual arousal have also been evaluated. Humphreys and Bruce [25] and Klein [32] suggested that the intensity of visual arousal produced by an object usually diminished with the distance between the object and a viewer’s attention locus (anchor). In an information search context, the significance of a system-suggested item perceived by a user may decrease with distance between the item and the user-provided search anchor. Furthermore, Murch [40] has discussed the effects of color on visual arousal and viewers’ perceptions, suggesting that red and blue may have contrasting connotations. Traditionally, red is a warm color that can be used to attract immediate attention from viewers or prompt quick action, as manifested by common and frequent use of red in traffic and warning signs in different cultures. While the precise meaning of a color may subtly vary from one culture to another, red is an appropriate candidate color to represent important or significant objects, including system-suggested items highly relevant to a user’s search. On the other hand, blue is often considered a cool color that invites receding attention from viewers. Accordingly, blue may be appropriate for representing system-suggested items not highly relevant to a user’s search. Based on the theories involved, an IR system can communicate to
users the system-generated degree of relevance of a system-suggested item using such visual properties as size, distance and color. Specifically, an IR system can use nodes of large size (area), short distance from the search anchor, attention-attracting display color (e.g., red), or combinations of these to communicate to the user system-suggested items highly relevant to his or her search task.

Alternatively, perception-based information/knowledge can be represented using a text-based linear-ordering schema that encodes a set of significant objects and communicates the respective significance to users using their relative positions within the list [1]. A linear list can be front- or back-end-anchored and is often organized alphabetically or following an ascending or descending rank order. List-based representations are common in IR systems and Internet-based search systems. Observation of many existing systems suggested a predominant use of the front-anchored approach, wherein items of higher degrees of significance or relevance appeared toward the top of a list.

Empirical examinations of the discussed representation methods and their applications in information/knowledge representation, display and visualization have also been fairly extensive. Take the HCI community for instance. Rao and Card [41] developed Table Lens, a new information visualization technique that fused symbolic and graphical representations into a single cohesive view of information. Hirata et al. [24] presented a media-based navigation method that allowed users to incorporate such visual cues as physical shape and color when navigating through a hypermedia system. Jerding and Stasko [30] developed Mural, a two-dimensional information representation method that included such essential visual attributes as color and pixel. Arend et al. [3] examined icon selection behaviors of individual users and compared their behaviors associated with icons that incorporated different visual features. Hutchins et al. [26] investigated the cognitive accounts of the advantages and disadvantages of direct manipulation interface, suggesting the importance of the difference between user intention and system facilities. Caird and Hancock [11] evaluated the effects of static and dynamic whole-hand gestures on user perception and concluded that user-perceived size of an object was important. Barfield and Robless [4] evaluated the effects of two- or three-dimensional graphics on the problem-solving performance of expert and novice decision makers.

Cognitive psychology has also contributed to our understanding of and knowledge about information representation and visualization. For example, Girelli and Luck [21] investigated the mechanisms, including color and form, that attracted the attention of viewers when selecting particular visual targets. Sharps et al. [46] evaluated the primacy effects resulting from different stimuli, including visual and auditory. Chambers and Reisberg [13] examined the relationship between the meaning of an image and viewers' responses to the image. Cohen and Shoup [15] assessed the response-selection processes of viewers and suggested that such processes may not be autonomous from earlier perceptual analysis of underlying visual stimulus or stimuli.

2.3. Research motivation

A review of relevant IS literature suggested that most previous information presentation research has taken a dichotomous approach to color (i.e., color-enhanced vs. color-free), has focused on comparing graphical charts or tabular reports for information summarization purposes, and has shown a predominant interest in decision making support. In addition, information representation and visualization research from HCI and cognitive psychology have limited the discussion of evaluating and comparing the effects of different visual properties on users' perceptions of an object's significance or relevance. In response to this research void as well as to the need for increased research efforts in user interface design for IR systems and a clamorous call for desirable linkages between basic theories and HCI artifacts [5,12], we undertook the work reported here. We used the Model Human Processor as a framework to encompass relevant cognitive psychology theories and user-centered design principles to build interfaces and to evaluate their representational effectiveness in support of system–user concept communication, particularly the concept of relevance. Our study had a well-researched theoretical foundation, complied with essential user interface design principles, and included a large-scale empirical evaluation.
3. Research questions and methodology

The specific research questions addressed were as follows.

1. Does user interface design have a significant effect on communication of the relevance of a system-suggested item from an IR system to its users?

2. Are graphical interface designs more effective than list-based user interfaces in supporting system–user concept communication?

3. Are graphical interface designs incorporating size, distance and color more effective than those exclusively depending on size, distance or color in supporting system–user concept communication?

Our evaluation focus was on the support of an interface design to an IR system’s communicating to users the degree of relevance of a system-suggested item. A logical and intuitive way to evaluate the effectiveness of an interface design is examining the difference deviation between the system-generated relevance and the user’s perceived relevance. We postulated affirmative answers to our research questions and thus used their negations as null hypotheses in the subsequent data analysis. Understandably, the need for, as well as the magnitude of, interface-induced concept communication improvement may decrease when users are knowledgeable about the search topic domain. Thus, we examined the effect of user (search) task familiarity on system–user concept communication resulting from different interface designs. Specifically, we postulated that interface-induced communication improvement would become increasingly prominent as users’ familiarity with or knowledge about the search task decreased.

The effectiveness of an interface design in supporting the target system–user communication may also depend on the amount of information that a viewer needs to process. The significance and magnitude of interface-induced system–user communication improvement may increase with the total number of system-suggested items presented or returned to the user. To investigate the speculated size effect, we performed data analysis using different numbers of system-suggested items. Specifically, we included a total of nine system-suggested items in the study and evaluated each interface at two-item intervals.

Alternative measures for evaluating an interface design’s support of system–user concept communication were also examined, including cognitive load and user satisfaction. We defined cognitive load as the amount of information processing effort a user must expend to take notice of the visual stimuli (stimuli) contained in an interface and comprehend its (their) significance or intended meaning. A self-reporting method was used to obtain individual users’ self-assessments of the cognitive load associated with a particular interface. Conceivably, users would prefer an interface design that requires a relatively low cognitive load and at the same time, can result in high user satisfaction. Therefore, we evaluated each investigated design using its cognitive load requirements and the resulting user satisfaction with that design.

3.1. Research method and design

Our primary research method was a computer-based experimentation. We adopted a one-repeated measure experimental design in which subjects were provided with all the treatments (interface designs) in a random fashion. A subject was presented with six search tasks, each of which used a particular interface design, graphical or list-based. The experiment instrument included a total of 36 (6 × 6) interface–task combinations and used the system clock of an individual host machine to select six specific interface–task combinations to be included in the search tasks for a subject. Choice of the experimental design was made based on its relative strength in detecting the hypothesized treatment effect i.e., effects of interface design on system–user communication while taking into account within-subject variations. The following details our research design, including back-end IR system, subjects, treatments, and experiment instrument development.

3.1.1. Back-end IR system

We used the auto-indexing and co-occurrence modules of a concept-space-based IR system as the back-end system. Developed at the University of Arizona, this system takes a collection of documents as input and performs automatic indexing by creating indexes for related terms (items) contained in the documents. Each index has a weight value representing the degree of relevance between the related...
terms, ranging from 0 to 1.00. The indexes are directional and therefore, the weight values (degree of relevance) of two related terms are not necessarily symmetric. In this study, we examined the effectiveness of different interface designs in communicating to users specific system-generated weight (relevance) values for individual pairs of search anchor term and system-suggested items.

Documents collected from each target topic domain, including article abstracts and news summaries, were preprocessed to remove ‘’stop words’’, approximately 1000 common words that carry no significant meaning. Examples of stop words are articles, prepositions, and pure verbs. Adjacent words were grouped as phrases and each participating word was indexed.

A term frequency, \( t_{ij} \), was calculated for each indexed term to represent the number of occurrences of term \( j \) in the document \( i \). Similarly, a document frequency, \( d_{ij} \), was also calculated to represent the number of documents that contained term \( j \). An asymmetric co-occurrence (or similarity weight) was then determined using the following formula:

\[
W_{s,t} = \frac{\sum d_{ijk}}{\sum d_{ij}}, \quad \text{where}
\]

\[
d_{ij} = t_{ij} \log d_{ij}, \quad \text{and}
\]

\[
d_{ijk} = t_{ijk} \log d_{jk}.
\]

In the calculation of asymmetric occurrence, \( t_{ijk} \) denotes the frequency of both terms \( j \) and \( k \) appearing in document \( i \) and \( df_{jk} \) represents the number of documents containing both terms \( j \) and \( k \).

The co-occurrence weight between terms \( j \) and \( k \) (\( W_{jk} \)) can be interpreted as the conditional probability of finding \( k \) in the same document, given the existence of \( j \). Therefore, equality between \( W_{jk} \) and \( W_{kj} \) is not guaranteed and in effect has been seldom observed. In the study, we developed interface designs, graphical or list-based, to simulate situations where a given occurrence of an anchor term (user-specified search word) and a system-suggested item was represented by different visual properties (size, distance and color), singly or in combination, and examined individual users’ evaluation of each interface.

3.1.2. Subjects

The study targeted members of the Fall 1997 business-major freshman class at a major university in Hong Kong. Choice of subjects was based on accessibility and frequent need to use IR systems to search for information of interest. As part of their first-year requirements, all subjects were enrolled in the Introduction to MIS course. Arrangements were made with the instructors teaching the course to include subjects’ participation in the study as a course requirement. The number of targeted subjects was 715.

3.1.3. Treatments

Each interface design constituted a treatment in the study. We included a total of six treatments which were classified into two broad categories: list-based or graphical. Two designs were list-based and the others were graphical. Our experimental design ensured that each subject would be exposed to all the treatments randomly. No particular treatment pattern was observed, reducing the threat of potential carry-over effects. Table 1 summarizes the treatments included in the study.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Interface category</th>
<th>Specific visual property used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alphabetical list</td>
<td>List-based</td>
<td>Linear list — alphabetical order</td>
</tr>
<tr>
<td>Rank-based list</td>
<td>List-based</td>
<td>Linear list — descending rank order</td>
</tr>
<tr>
<td>Size-based image</td>
<td>Graphical</td>
<td>Size</td>
</tr>
<tr>
<td>Distance-based image</td>
<td>Graphical</td>
<td>Distance</td>
</tr>
<tr>
<td>Color-based image</td>
<td>Graphical</td>
<td>Color</td>
</tr>
<tr>
<td>Multi-property-based image</td>
<td>Graphical</td>
<td>Combination of size, distance, and color</td>
</tr>
</tbody>
</table>
With a list-based interface design, the anchor term was placed on top of a box and all the system-suggested items were placed inside the box according to alphabetical order or a rank sequence determined by the respective degrees of relevance generated by the system, as shown in Fig. 1. Commonly used by many IR/search systems, list-based interfaces usually contain fewer visual cues than graphical designs and thus may not be effective in supporting system—user concept communication. However, the cognitive load associated with a simplistic list-based interface design may be lower than that of a graphical design that often has rich visual cues.

A graphical interface design may incorporate such visual properties as size, distance and color, singly or in combination. In all graphical designs, the anchor term was framed and approximately placed in the center of the interface. For color-based graphical interfaces, we concentrated on examining and comparing use of red and blue in supporting system—user concept communication, using a continuum with pure red (FF0000) at one end to signify extremely high relevance (100% relevance) and pure blue (0000FF) at the other end to indicate extremely low relevance (0% relevance). Our evaluation focus required the inclusion of a neutral color to provide a smooth and continuous transition from one end of the continuum to the other. White is usually considered as a neutral color (i.e., color-free) and therefore was chosen to provide the necessary transition from pure red to pure blue, or vice versa. Specifically, we used white to represent system-suggested items that had a 0.5 system-generated relevance. Thus, items with a system-generated relevance value greater than 0.5 were represented by red and those with relevance values less than 0.5 were displayed in blue. Results from repeated palette experiments suggested that green was an effective mediating color. When properly mixed with red and blue, green was able to create white (our chosen neutral color) without introducing additional colors. Accordingly, we developed a formula that would take the system-generated relevance value of a system-suggested item as input and determine the intensity of each of the colors involved; i.e., red, blue and green jointly determined the precise color of a node representing a system-suggested item, based on its degree of relevance as suggested by the underlying IR system. The color of a node would

![Fig. 1. Sample interface design — list-based by alphabetical order.](image-url)
gradually and smoothly change from red, to white, and then to blue as the system-generated relevance value continuously decreased. Nodes representing system-suggested items were placed around a unit circle from the anchor term, which was displayed in white. We chose white (our neutral color) as the display color for the anchor term to avoid undesirable color matching or association by subjects. Conceivably, a subject might perceive system-suggested items displayed in a color nearly identical or similar to that of the anchor term as being highly relevant, regardless of the display color used. Such color matching or association would then prevent our evaluating and comparing the utility of red and blue for system–user concept communication support.

A distance-based graphical interface design used the distance between the anchor term and a system-suggested item to signify its relevance. Nodes, identical in size (area) and color (color-free), were used to represent system-suggested items and were placed some distance from the node representing the anchor term, depending on their respective relevance (as shown in Fig. 2). Specifically, the node-to-anchor distance was adversely related to the relevance (significance) of a system-suggested item; i.e., the shorter the distance, the more relevant the item. The actual node-to-anchor distance was determined by a pre-defined linear function bounded by specified minimal and maximal distances.

Size-based graphical interfaces used the area of a node to communicate to users the exact degree of relevance of a system-suggested item. Specifically, node size was positively proportional to the relevance of a system-suggested item and was determined by a pre-defined linear function bounded by specified minimal and maximal node sizes. Nodes representing system-suggested items were then placed on an approximate unit circle around the anchor term. In addition, we also examined interfaces that incorporated multiple visual properties; i.e., nodes representing system-suggested items were displayed using a combination of different colors, various distances (from the anchor term), and differential sizes, each of which was determined according to our previous discussions.

3.1.4. Topic domains and information sources

The study included three general search topic areas: Sino–British negotiation, general business, and linear programming. Together, these topic areas pro-

![Sample interface design — graphical, distance-based.](image-url)
vided a desirable diversity, enabling us to assess the potential effect of user task familiarity on the effectiveness of an interface design in communicating to users a system-suggested item’s relevance. Based on results from pre-study interviews with a subset of the target subjects, the freshman class as a group appeared to be most knowledgeable about general business and least familiar with linear programming. We identified and included input documents for the discussed topic areas from such sources as Lexus/Nexus, Dow Jones New Retrieval, and ABI/INFORM.

3.1.5. Experiment tasks

Each subject was asked to review six information search tasks, two from each included topic area. Each search task involved a particular interface design determined by the system clock of the respective machines. In each search task, the interface included a total of nine system-suggested items, varying in their degrees of relevance with respect to the specified (search) anchor term.

In each search task, a subject was presented with an interface containing the anchor term together with nine related items suggested by the system. The subject was asked to review the interface design and use the sliding rulers provided to indicate his or her perceived degree of relevance of each system-suggested item with respect to the anchor term, based on the interface used. Both the interface and the question items were on the same screen and subjects had no time constraint in their interface evaluation. Subjects were constantly reminded to avoid or suppress reliance of the semantics of the anchor and system-suggested items when making judgments about their perceived relevance. By design, the experiment flow was irreversible. Once submitted, responses to a search task could not be re-accessed by the subject.

Upon completing evaluation of all six interfaces, a subject was asked to explicitly indicate the cognitive load associated with each interface and his or her satisfaction with the interface. Reviews of the interface used in each search task were allowed. A subject could unrestrictedly review a particular interface by clicking on the corresponding search task number for which the interface was used. Each review request activated the interface of choice for approximately 3 s. Responses to cognitive load and user satisfaction evaluation could not be re-accessed after submission.

3.1.6. Experiment instrument development

Written in Borland C++, an experiment instrument was developed to present the investigated interface designs to subjects, together with the associated questions. The instrument included a set of sliding rulers in each search task, one for each question item. A subject was asked to move the sliding ruler provided to indicate the appropriate place on a continuum, ranging from 0 to 100, to indicate his or her response to a question. The instrument had a back-end database component developed using Paradox Engine to record subjects’ responses collected during the experiment.

The program was pre-compiled and installed in the personal computers used in the study. When activated, the program would use the system clock of an individual machine to determine the particular combination of interface designs and search tasks and their presentation sequence. To prevent undesirable early exit of subjects, the instrument had built-in mechanisms restricting a subject from proceeding to the next search task before submitting his or her responses to all the questions for the current task. Similarly, the instrument prohibited a subject’s leaving the study before completing his or her evaluations of all the interfaces and providing responses to cognitive load and user satisfaction assessments.

3.2. Research process

The study was conducted in three distinct phases as follows.

3.2.1. Pre-study preparation

Major tasks of this preparatory phase included back-end IR system selection, topic area selection, input document identification and processing, interface design and development, and experiment instrument development. Several candidate search topic areas were identified and those selected for the study were chosen on the basis of pre-study interview results, with a subset of subjects randomly drawn
from the target subject pool. In addition, we developed a script for the experiment process as well as selected and trained experimenters who assisted with experiment administration.

3.2.2. Pre-test

A pre-test that involved 15 first-year business-major undergraduates was conducted to assess the experiment instrument and fine-tune the experiment flow. The pre-test also served as a valuable learning vehicle that allowed the experimenters to become familiar with the experiment flow and data collection procedures. Drawn from the target subject pool, participants who took part in the pre-test voluntarily were excluded from the subsequent formal study. The pre-test results were satisfactory, with feedback from the participants leading to several changes in both instrument design and experiment flow. Furthermore, the pre-test enabled us to develop realistic estimates of the time requirements for the study.

3.2.3. Formal study

The study was conducted in multiple sessions, each of which, constrained by the laboratory capacity, had a maximum of 50 subjects. The same experimenters administered all the experiment sessions, following a scripted process. At the beginning of each session, subjects were given a few minutes to read a two-page written statement, which explicitly explained the purpose of the study, the management and intended use of the data to be collected, and the overall experiment flow. An overview of the experiment purpose and process was verbally presented to the subjects, who were then provided with a demonstration of the entire experiment process. Subjects were explicitly asked and constantly reminded to make perceived relevance judgments based on the visual cues detected and comprehended rather than the semantic meaning of words. Additional demonstrations were provided until all subjects unequivocally signaled their readiness for the study. All subjects were given unlimited amount of time to complete the study.

Each study session commenced with a Welcome Screen explaining the study’s purpose, process, and intended data management, followed by an Input Screen soliciting subjects’ student identification numbers. In the experiment, all screens were displayed at full size (of the computer screen) to avoid potential confounding effects introduced by back-

![Fig. 3. A sample Task Screen — graphical interface using size, distance and color.](image)
ground noise (e.g., irrelevant objects and icons). A Task Screen then was presented, requiring the subjects to complete all six search tasks individually. When undertaking the tasks, a subject did not have to follow any particular sequence but could exit the study only upon completion of all the tasks and survey questions. Each task was presented by a Task Screen consisting of a particular interface on the left and associated question items on the right, as shown in Fig. 3. To complete a search task, a subject needed to indicate the exact degree of relevance of each system-suggested item in relation to the anchor term by moving the sliding ruler to an appropriate location, based on his or her perceived relevance as suggested by the interface presented. Subjects clicked on the Finish Button at the bottom of screen to submit their responses. All subjects were repeatedly reminded that task submission was irreversible and were asked to re-examine their responses before submission.

Upon completing his or her evaluation of all the interfaces, a subject was sequentially presented with two Survey Screens: one on cognitive load and the other on user satisfaction. Each Survey Screen had an interface display space on the left and evaluation questions on the right, as shown in Fig. 4. When responding to an evaluation question, the subject was asked to explicitly indicate the amount of effort required to figure out a particular interface design or the level of satisfaction with that design. Subjects were asked to review all the interfaces used in previous search tasks. Evaluations of cognitive load and user satisfaction were put off until a subject’s completion of all the search tasks for several reasons. First, evaluations of the cognitive load and user satisfaction of different designs were relative in nature (e.g., list-based vs. graphical or size-based vs. distance-based) and logically should proceed by allowing a subject to view or review them together, preferably using one screen. Second, the subjective nature of cognitive load and user satisfaction evaluation, together with the self-reporting method used in the study, may introduce considerable within-subject inconsistency. To produce reliable evaluations and comparisons, a subject needed a consistent internal evaluation system that was easier to establish and maintain with one-time batch-mode evaluations than several individual evaluations taking place at different times. Furthermore, the irreversible design of our experiment instrument prohibited subjects from re-
viewing a completed search task. By making cognitive load and user satisfaction evaluations part of the search task, we in effect prohibited subjects from reviewing evaluations they had previously submitted. Consequently, this could jeopardize the reliability of the resulting evaluations of cognitive load and user satisfaction. When comparing different designs, a subject could review a previously seen interface by clicking on the search task number that employed that particular interface. An interface would appear in the display space for 3 s upon each review request and a subject could make as many review requests as were needed.

4. Data analysis results and implications

Of the 715 target subjects, 601 took part in the study, showing a 84% participation rate. Among the participants, 29 submitted questionable responses, including consecutive or frequent lack of sliding ruler movements. All questionable responses were excluded from the subsequent data analysis, resulting in an effective subject size of 572. On average, subjects took 15 min and 12 s to complete the experiment, excluding the introduction and demonstration.

For the evaluation of a particular interface by an individual subject, an overall deviation index was calculated by averaging over all the system-suggested items the absolute value of the difference between the system-generated and the subject-perceived relevance for a system-suggested item. For example, the overall deviation index was 0.3 when the subject-perceived degree of relevance, on average, deviated by that value from the system-generated degree of relevance. Using overall deviation indexes, we analyzed the main effect of user interface design on system–user concept communication, compared the effectiveness of each interface category, and examined the potential effect of user domain knowledge on interface design evaluation.

4.1. Effects of interface on system–user concept communication

To investigate the main effect of user interface on system–user concept communication, we performed analysis of variance (ANOVA), using the resulting deviation indexes. To investigate the potential size effect previously discussed, we assessed the effectiveness of different interface designs using four different numbers of system-suggested items; namely all system-suggested items (nine) and most relevant three, five, and seven items.

Results from our ANOVA analysis showed that the effect of interface design on system–user concept communication was significant statistically. At all investigated numbers of system-suggested items, graphical interfaces were significantly better than list-based interfaces. As shown in Table 2, the deviation indexes of graphical designs were considerably and consistently smaller than those of list-based interfaces. Interestingly, results of the comparative analysis between list-based and graphical interfaces suggested that the superiority of a graphical design appeared to increase with the number of system-suggested items. Consistent with Jarvenpaa and Dickson [28], our finding implied that a graphical interface may become increasingly effective in situations where a system needs to communicate a large amount of information to users.

We also evaluated the effectiveness of different designs using an alternative measure, a bubble-sort-

| Overall deviation indexes (in percentage) — graphical vs. list-based |
|-------------------------|-------------------------|-------------------------|-------------------------|
| Mean Difference P-value |
| List-based | Graphical | List-based | Graphical | List-based | Graphical | List-based | Graphical |
| All nine items | 38.1 | 32.6 | 10.6 | 12.3 | 5.5 | < 0.01 |
| Most relevant seven items | 28.9 | 24.9 | 10.6 | 11.7 | 4.0 | < 0.01 |
| Most relevant five items | 20.7 | 18.2 | 9.8 | 10.5 | 2.5 | < 0.01 |
| Most relevant three items | 13.3 | 12.0 | 8.5 | 8.9 | 1.3 | < 0.01 |
Table 3
Analysis of overall deviation indexes (in percentage) — list-based interfaces

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Difference</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Alphabetical</td>
<td>Rank-based</td>
<td>Alphabetical</td>
<td>Rank-based</td>
</tr>
<tr>
<td>All nine items</td>
<td>39.2</td>
<td>37.0</td>
<td>9.3</td>
<td>11.7</td>
</tr>
<tr>
<td>Most relevant seven items</td>
<td>29.5</td>
<td>28.2</td>
<td>9.8</td>
<td>11.3</td>
</tr>
<tr>
<td>Most relevant five items</td>
<td>21.0</td>
<td>20.4</td>
<td>9.4</td>
<td>10.3</td>
</tr>
<tr>
<td>Most relevant three Items</td>
<td>13.5</td>
<td>13.0</td>
<td>8.2</td>
<td>8.8</td>
</tr>
</tbody>
</table>

based method. This evaluation method concentrated on deviations of position rather than an absolute value. With this method, the relative ranks of system-suggested items specified by a subject were sorted so as to match those generated by the underlying IR system. Each swap of position incurred a penalty tick and the total ticks for the entire sorting process required for a subject’s perceived ranks of all system-suggested items were tabulated, when using a particular interface. Similar to findings derived from the absolute deviation analysis, results of the bubble-sort-based evaluation suggested that graphical interfaces were significantly better than the list-based ones and that interfaces incorporating use of size, distance, and color were more effective in supporting system–user concept communication than those using only one of these visual properties.

4.2. Comparative analysis of individual interface designs

Pair-wise comparisons of specific designs within each interface category were performed using t-statistics. Results of our comparative analysis may shed light on the contribution as well as the relative strength of the respective visual properties in supporting system–user communication.

For list-based interfaces, a rank-based design appeared to be more effective than alphabetic arrangement, as shown in Table 3. Subjects demonstrated a tendency to use the front-anchor approach in their interpretation of the relative importance of system-suggested items included on a list, as manifested by the observed superiority of the rank-based design, which grouped all the system-suggested items in a descending fashion, based on their system-generated degrees of relevance. In addition, the observed differential seemed to increase with the number of system-suggested items and was statistically significant when the number of system-suggested items exceeded five. This may partially suggest that the superiority of a rank-based list over one arranged alphabetically may be specific to the amount of information represented to the user.

As for graphical interfaces, our results suggested that designs incorporating use of size, distance, and color consistently appeared more effective than those depending on any single one of them, as shown in Table 4. The superiority became increasingly prominent and statistically significant when the number of system-suggested items grew. Among the visual properties investigated, size appeared to be the most effective in supporting system–user concept commu-

Table 4
Overall deviation indexes (in percentage) — different graphical interfaces

<table>
<thead>
<tr>
<th></th>
<th>Color &gt; distance</th>
<th>Distance &gt; size</th>
<th>Size &gt; combination</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>P-value</td>
<td>Mean</td>
</tr>
<tr>
<td>All nine items</td>
<td>37.2</td>
<td>&lt; 0.01</td>
<td>32.1</td>
</tr>
<tr>
<td>Most relevant seven items</td>
<td>28.8</td>
<td>&lt; 0.01</td>
<td>24.5</td>
</tr>
<tr>
<td>Most relevant five items</td>
<td>21.4</td>
<td>&lt; 0.01</td>
<td>18.0</td>
</tr>
<tr>
<td>Most relevant three Items</td>
<td>14.4</td>
<td>&lt; 0.01</td>
<td>12.2</td>
</tr>
</tbody>
</table>
nication and its utility level was comparable to that using a combination of size, distance and color, particularly when the number of system-suggested items was not large (e.g., less than seven). Distance was more effective than color, whose utility in supporting system–user communication was not prominent. In effect, color was the least effective and its contribution to system–user communication was significantly less than that of size or distance, across all the numbers of the system-suggested items investigated.

4.3. Effects of domain knowledge on interface effectiveness

User search task familiarity was investigated. Subjects’ responses to different search tasks were grouped by topic category and analysis was performed on the resulting composite data sets independently. As shown in Table 5, graphical interface designs were significantly more effective than list-based ones across all topic areas investigated. Further analysis showed that interface-induced concept communication improvement was most prominent in search tasks pertaining to linear programming and least prominent with those concerning general business. The observed differentials prevailed across all the investigated numbers of system-suggested items and their prominence increased with the number of system-suggested items. Overall, the findings suggested that a graphical interface was effective in situations where the user had limited knowledge about the search domain, especially when the number of system-suggested items was relatively large.

Further analysis of specific designs pertaining to each interface category was also performed and showed several interesting patterns. For the list-based designs, the rank order was consistently better than alphabetical arrangement and the observed difference as well as its statistical significance seemed to decrease with user task familiarity; i.e., a rank-based list may be more appropriate than one arranged alphabetically in search tasks about which users have limited knowledge.

Results from comparative analysis of the graphical designs showed a general pattern: interfaces using a combination of size, distance and color were more effective than those that exclusively used one of them. Several observations also merit discussion. First, analysis of the responses to search tasks pertaining to general business suggested that interfaces that used size exclusively were as effective as those incorporating size, distance and color. This suggested that a size-based graphical interface is sufficiently effective for support of search tasks about which users have considerable knowledge. However, the observed superiority appeared to diminish as user task familiarity decreased. Second, the difference

<table>
<thead>
<tr>
<th>Topic Area</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Difference</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>List-based</td>
<td>Graphical</td>
<td>List-based</td>
<td>Graphical</td>
</tr>
<tr>
<td>General business</td>
<td>38.4</td>
<td>34.2</td>
<td>10.3</td>
<td>10.9</td>
</tr>
<tr>
<td>Sino–British negotiation</td>
<td>39.2</td>
<td>33.6</td>
<td>10.4</td>
<td>12.4</td>
</tr>
<tr>
<td>Linear programming</td>
<td>36.7</td>
<td>30.2</td>
<td>11.0</td>
<td>13.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 6</th>
<th>Analysis of cognitive load and user satisfaction — graphical vs. list-based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td></td>
<td>List-based</td>
</tr>
<tr>
<td>Cognitive load</td>
<td>53.9</td>
</tr>
<tr>
<td>User satisfaction</td>
<td>38.3</td>
</tr>
</tbody>
</table>
between distance and size in supporting system–user concept communication was statistically insignificant. The utility of a distance-based interface in supporting system–user concept communication was lower than but comparable to that using size exclusively. Furthermore, our data analysis results showed that color conveyed concept less effectively than size or distance and the difference was statistically significant.

4.4. Cognitive load and user satisfaction

We also examined the levels of cognitive load associated with each interface design. Taking all system-suggested items into consideration, our results suggested that the aggregate cognitive load of list-based interfaces was comparable to that of graphical ones and the difference was not statistically significant, as summarized in Table 6. Further analysis showed the differences in cognitive load between or among specific designs within the list-based as well as the graphical interface category were not considerable or statistically significant (as shown in Table 7); i.e., our findings suggested that subjects considered all the investigated interface designs, list-based or graphical, to have comparable cognitive load requirements.

Users’ satisfaction resulting from individual interface designs was also examined. Results of our data analysis suggested that subjects exhibited a higher level of satisfaction with a graphical interface than with a list-based interface. The interface-induced user satisfaction increment was both prominent and statistically significant, as shown in Table 6. When provided with a list-based interface, subjects appeared to be more satisfied with one arranged alphabetically than with one based on a rank order, with the difference being 5.3 and the P-value less than 0.01. Table 7 summarizes the user satisfaction resulting from each graphical design; several interesting observations are apparent. First, subjects exhibited a significantly higher satisfaction level with designs that incorporated color, distance and size than with those that used only one of them. Second, color surprisingly appeared to be the most effective single visual property in contributing to user satisfaction, a finding not in agreement with results suggesting color to be ineffective in system–user concept communication support. The observed disparity in findings may imply that user satisfaction with an interface design is related to but different from the effectiveness of that design in supporting system–user concept communication and that users may have a general preference for color-enhanced interfaces. In this connection, user interface designers should consider use of color and supplement it by other visual properties to effectively support system–user concept communication. Furthermore, when used exclusively, size may not be perceived favorably by users, even though it appeared to be effective in system–user concept communication support (as suggested by our data analysis results). Overall, our study suggested the superiority of interfaces incorporating use of size, distance and color over designs that exclusively use color, distance, or size to enhance system–user concept communications and user satisfaction.

4.5. Implications

Several implications for interface design for IR systems and knowledge management systems can be drawn from our findings. First, system developers and potential adopters should seriously consider graphical interfaces as a viable design alternative to conventional list-based interfaces when their goal is to provide effective system–user concept communication support. In particular, a graphical interface

<table>
<thead>
<tr>
<th>Combination &gt; color</th>
<th>Color &gt; distance</th>
<th>Distance &gt; size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>P-value</td>
<td>Mean</td>
</tr>
<tr>
<td>Cognitive load</td>
<td>53.6, 54.2</td>
<td>&gt; 0.05</td>
</tr>
<tr>
<td>User satisfaction</td>
<td>65.8, 58.4</td>
<td>&lt; 0.01</td>
</tr>
</tbody>
</table>
may be useful in situations in which users have limited knowledge about the search domain or the number of system-suggested items is large. Second, a graphical interface that incorporates use of size, distance and color may be more effective in supporting system–user concept communication than a design based on only one of these visual properties. Conceivably, inclusion of multiple visual properties would increase the likelihood of a subject’s receiving and correctly interpreting an intended concept by using some if not all of the visual cues displayed. In addition, an interface design incorporating multiple visual properties may also contribute to increased user satisfaction without adding considerable cognitive load. Third, size and distance may have a more prominent facilitating effect on system–user concept communication than color. Our subjects exhibited varying degrees of difficulty in detecting color-based visual cues, correctly interpreting them, or both. However, use of color may lead to high user satisfaction with interfaces. To achieve effective concept communication with a desirable user satisfaction level, developers of IR systems and knowledge management systems may want to consider including color in their interface designs and supplementing it with other visual properties, including distance and size. Last but not least, list-based interfaces that are based on rank ordering may be more effective in system–user concept communication support than those arranged alphabetically, especially in situations where the user needs to deal with a fairly large number of system-suggested items and has limited knowledge of the search topic domain.

5. Conclusion

We investigated user interface design for IR systems by means of a computer-based experiment, with results that suggest the following conclusions. First, information representation may have a significant effect on system–user concept communication. Second, graphical interfaces may be superior to text-based linear lists in communicating a concept of interest from an IR system to its users, regardless of the domain knowledge level of users. Third, interfaces incorporating multiple visual properties may be more effective in supporting concept communication than those exclusively depending on a single visual property. Fourth, use of graphical interfaces may result in a higher level of user satisfaction than text-based linear lists, although both may be comparable in cognitive load requirements. Interface-induced user satisfaction may become increasingly prominent when multiple visual properties are included in the design. Fifth, interface-induced system–user concept communication improvement may increase with the number of system-suggested items but decrease with user task familiarity. Furthermore, size and distance may provide many effective and essential visual cues in system–user concept communication, but color may be less effective for such communication support.

The study has contributed to both IS and IR research. We extended previous IS research on information presentation by evaluating potential effects of several theory-based information representation methods on the user-perceived semantics of individual objects together with their accompanying cognitive loads and resulting satisfaction. In its methodology, our study addressed common problems in experimental IS research by including a well-researched theoretical foundation and incorporating an appropriate experimental design [29]. Contributions to IR research have also been made in several areas. First, we responded to the need for formal investigation of the user interface design issue in IR system development. We also have provided a desirable linkage between relevant cognitive psychology theories and interface design, which is essential to IR and search. Furthermore, we empirically evaluated the effectiveness of different interface designs in supporting system–user concept communication as well as the associated cognitive load and the resulting user satisfaction. Results of our study can be applied to the design and developments of such emerging information technologies as knowledge management systems and interactive multimedia IS.

The study has several limitations. First, the user-perceived semantics resulting from a particular information/concept representation method may be culture-dependent and therefore needs to be examined in a cross-cultural setting. Conceivably, our subjects may have differed subtly from their counterparts in other cultures or countries in such areas as detection of visual arousal (or stimuli) and internal coding of
the semantics resulting from the detected visual perceptions. For instance, users outside Hong Kong might find color to be effective in communicating the exact degree of relevance of a system-suggested item. Second, the study included multiple design dimensions, but its evaluations proceeded in a rather disjointed manner. We provided separate discussions of the effects of user interface on concept communication, cognitive load, and user satisfaction rather than integrating the evaluations concurrently. In addition, we did not examine users' actual information search behavior. It would be interesting to observe and analyze actual search behavior of users who have been provided with a list-based or a graphical interface. These limitations suggest some directions for our future research. Furthermore, development of an adaptive interface that incorporates an intelligent agent capable of analyzing user information search behavior and task familiarity should also be appealing.

References

[4] W. Barfield, R. Robless, The effects of two- or three-dimen-
[6] N.J. Belkin, W.B. Croft, Information filtering and informa-
[7] I. Benbasat, A.S. Dexter, An experimental evaluation of graphical and color-enhanced information presentation, Man-
[8] I. Benbasat, A.S. Dexter, An investigation of the effective-
[9] I. Benbasat, A.S. Dexter, P. Todd, An experimental program investigating color-enhanced and graphical information pre-
[21] M. Girelli, S.J. Luck, Are the same attentional mechanisms used to detect visual search targets defined by color, orienta-
[25] G.W. Humphreys, V. Bruce, Visual Cognition: Computa-
[29] S.L. Jarvenpaa, G.W. Dickson, G. DeSanctis, Methodologi-


Paul Hu is an assistant professor of Information Systems and Decision Sciences in the College of Business Administration at the University of South Florida. He received his PhD in Management Information Systems from University of Arizona in 1998. He has had papers published (including forthcoming) in Journal of Management Information Systems, IEEE Transactions on Engineering Management, IEEE Intelligent Systems, Decision Support System, Telecommunication Systems, Journal of Organizational Computing and Electronic Commerce, Topics in Health Information Management, and Journal of Telemedicine and Telecare. His current research interests include information technology management and applications in health care, telemedicine, HCI, and electronic commerce.

Pai-chun Ma was educated at the Stern School of Business, New York University, and received his PhD in Information Systems. He is currently an associate professor of Computer Information Systems at Baruch College of the City University of New York with 11 years of experience teaching university computing courses in Object-Oriented Programming, Decision Support Systems, Database Management and Internet Applications Development. His consulting activities have been primarily focused on model arrangement, data warehouse, software development and project management. His current research interests are user interface design for information retrieval in digital libraries, activity-resource model for negotiation support, data warehouse and data mining.

Patrick Y.K. Chau is an Associate Professor in the School of Business at the University of Hong Kong. He received his PhD in business administration from the Richard Ivey School of Business (formerly Western Business School), University of Western Ontario, Canada. His research interests include decision support systems, information presentation and model visualization, and issues related to IS/IT adoption and implementation. He has had papers published in major information systems journals including Management Information System Quarterly, Journal of Management Information Systems, Decision Sciences, European Journal of Information Systems, Information and Management, and INFOR.