EXPLAINING AND ALLEVIATING INFORMATION MANAGEMENT INDETERMINISM: A KNOWLEDGE-BASED FRAMEWORK

H. CHEN, A.K. DANOWITZ, K.J. LYNCH, and S.E. GOODMAN
Management Information Systems Department, College of Business and Public
Administration, University of Arizona, Tucson, Arizona 85721, U.S.A.

and

W.K. MCHENRY
School of Business Administration, Georgetown University, Washington DC 20057, U.S.A.

(Received 22 June 1993; accepted in final form 13 October 1993)

Abstract—Our research attempted to identify the nature and causes of information management indeterminism in an online research environment and to propose solutions for alleviating this indeterminism. We conducted two empirical studies of information management activities. The first study identified the types and nature of information management indeterminism by evaluating archived texts. The second study focused on four sources of indeterminism: subject area knowledge, classification knowledge, system knowledge, and collaboration knowledge. A knowledge-based design for alleviating indeterminism, which contains a system-generated thesaurus and an inferencing engine, is also proposed in this article.

INTRODUCTION

Information management activities have for decades been the focus of research interests in various disciplines including Behavioral Science, Management Information Systems, and Information Science. In Behavioral Science, Simon's theory of "intelligence," "design," and "choice" has been taken as the model of rational decision making (Simon, 1960). Researchers in these areas have postulated that effective information collection and hypothesis generation and testing are the basis for making rational decisions (Simon, 1947, 1960).

In Management Information Systems, Huber has stressed the role of environmental scanning, information filtering and collection, and system-supported decision aids for achieving effective organizational decision making (Huber, 1974, 1979). Design criteria for Decision Support Systems have been suggested. This type of system creates a repository of important management-related information, and provides statistical and knowledge-based models for assisting high-level decision making. A variant of Decision Support Systems, called Computer-Assisted Research Systems (CARS), is intended to facilitate information collection, management, and retrieval; to support hypothesis generation and analysis; and to assist document preparation during the research process.

Although researchers in Behavioral Science and Management Information Systems acknowledge the importance of information and computer systems, it is information scientists who examine the activities of information management and retrieval in much closer detail. Prior research in the areas of information indexing and retrieval has provided important insights into understanding the processes of indexing and retrieving unknown and unstructured information.

Building on prior research in information indexing indeterminism, we attempted to investigate in detail the problems of information management indeterminism in an online research setting. We aimed at identifying the nature and types of information management indeterminism and the systematic causes that contribute to this indeterminism. Some findings are unique in the CARS environments. There are, however, some results that

pertain to online information storage and retrieval environments in general. Our final research goal is to propose a design model to alleviate information management indeterminism and to facilitate the process of "intelligent" information management and retrieval.

This article is organized as follows. In the second section we present a summary of relevant research. In Section 3, we discuss the unique characteristics of CARS and present a CARS we studied in detail in our research. In Section 4 we describe the research designs of two studies we conducted. Study 1 involved an expert researcher's extensive analysis and evaluation of the archiving of previously entered documents. We identified different types of information management indeterminism from this analysis. Results are presented in Section 5. Study 2 was a task-oriented experiment, investigating the process of system-supported information management. Based on the subjects' think-aloud protocols during problem solving and system interaction logs, we were able to identify some of the sources of information management indeterminism from this study, and our findings are reported in Section 6. In Section 7, we propose a knowledge-based design model for reducing information management indeterminism. This model consists of a system-generated thesaurus and an "intelligent" inferencing module. We report our research conclusions in Section 8.

2. LITERATURE OVERVIEW

Cybernetics researchers attempt to explain the behaviors of systems in terms of the determinism (or indeterminism) involved in the process (George, 1977). Because the initial states of a system may vary and the laws of its behavior may be indeterminate in nature, uncertainty becomes the norm and a variety (a number of distinct elements (Ashby, 1979)) of system outcomes emerge. Based on this perspective, information storage and retrieval tasks that involve interactions of information specialists of different experiences (e.g., indexers), different information storage and retrieval devices, and sometimes ambiguous and inconsistent indexing principles may exhibit a high degree of indeterminism.

The concept of indeterminism (or indeterminacy) has been adopted in information storage and retrieval by Blair (1986) and Bates (1986). Blair (1986) described two types of subject access indeterminism. There is indeterminism in the assignments of subject descriptors to documents and indeterminism in the selection of search terms by an inquirer. The first type of indeterminism is about *indexing inconsistency*. It includes: (a) inter-indexer inconsistency (different indexers are likely to assign different terms to the same document) and (b) intra-indexer inconsistency (an indexer might use different terms for the same document at different times). Studies of inter- and intra-indexer consistency report varying degrees of inconsistency (Hooper, 1965; Hurwitz, 1969; Jacoby & Slamecka, 1962). Much of the variance is due to: (a) the nature of the subject area (e.g., hard science vs. social science), (b) the levels of training and experience of the indexers involved, and (c) the controlled vocabulary nature of the index terms. The second type of indeterminism involves the variety in search terms used by the inquirers for their information needs. Again, the subject area and the experience and training of the inquirer cause the variance in the term selection. Blair proposed a "known-document-first approach" for alleviating subject access indeterminism. Instead of searching by using index terms, the inquirer would first identify a relevant document in the database, and could then use the index terms assigned to this known document to continue the search, similar to relevance feedback (Salton *et al.*, 1985; Salton, 1989).

Bates (1986) described the indeterminism involved in indexing and searching based on the cybernetics concept. She argued that (1) subject indexing is indeterminate and probabilistic (the *uncertainty* principle); (2) variety of searcher query must equal variety of document indexing (the *variety* principle); and (3) the search process is subtle and complex (the *complexity* principle). Based on these three principles, she suggested a design that "wrapped around" existing online catalogs' indexing schemes (Library of Congress Subject Headings). An end-user thesaurus, a search formulation front-end, command language, and natural language interface were suggested to alleviate indexing and retrieval indeterminism.

Both Bates and Blair acknowledge the problems of intra-indexer and inter-indexer indeterminism. Recent online database indexing studies show a third type of indexing

indeterminism, which we refer to as the "inter-system indeterminism." Because of the prevalence of online storage and retrieval systems, many systems (or databases) overlap in their collections of documents. Different systems, however, use their own indexing vocabularies and principles, which create indexing inconsistency among different systems.

Barber, Moffat, and Wood performed 30 case studies on the indexing of pharmacology papers by different online databases (Barber *et al.*, 1988). They found significant inconsistency in the indexing methods applied to drug formulations, pharmacokinetic, and other pharmacological aspects of the papers. Considerable variation was identified in the way different databases (e.g., Biological Abstracts, Chemical Abstracts, MEDLINE, etc.) treated each concept. Information retrieval was difficult because of the variety that existed in the systems. White and Griffith described tests by which the quality of indexing in online bibliographic databases can be compared (White & Griffith, 1987). These tests were based on quantitative measures such as the average number of descriptors applied to a document, the number of descriptors applied to all documents in an area, etc. Results showed that different databases exhibited different characteristics for those measures. This was the result of using different indexing vocabularies, rules, and indexers. As online storage devices become less costly and databases become more distributed (different online sources possess similar or identical documents), we expect to see higher occurrences of inter-system indexing indeterminism.

Based on our literature overview, we propose some research issues concerning online information management. We describe them below.

1. Indexing is only part of the information management process. Most of the prior research focused solely on selection of index terms, probably because the subjects in prior studies were mainly indexers in a library setting, where this was the main activity. However, in the broader information management context (especially in the online, system-assisted research environments), documents may need to be translated, synthesized, and entered before index terms can be assigned. Significant variety and uncertainty may exist during this process. This observation will be articulated further in the next section.

2. Causes of information management indeterminism are not clear. As Bates commented (Bates, 1986), most prior research did not explain or look into the *causes* of these indeterminisms. Human cognition, knowledge representation, and problem-solving heuristics that occur during the process of information management have not been examined systematically and empirically. This is clearly a crucial research question that needs to be addressed.

3. COMPUTER-ASSISTED RESEARCH SYSTEMS

A unique type of information storage and retrieval system, designed to assist information-intensive research activities of researchers, is referred to as Computer-Assisted Research Systems (CARS). These systems provide facilities to assist various research activities: storing and organizing research-related information, retrieving and analyzing archived information, and preparing documents as research output. The characteristics of a computer-assisted research environment are discussed below.

- Different types of information may be relevant to research topics of interest to researchers. Information may contain numeric, textual, or graphical data. It may arrive in the form of newspaper clippings, journal article extracts, book chapters, business or product brochures, midnight telephone conversations, overseas electronic mail exchanges, etc. Multiple sources and formats make information management difficult.
 - Researchers may be involved in the whole process of information collection, management, retrieval, analysis, and document preparation.
 - Research is often longitudinal and collaborative. A group of researchers who possess different levels of experience and expertise may be involved in research activities over an extensive period of time. The database they create serves as "organizational memory," capturing the knowledge of the group members.
-

Some systems provide partial support for research activities. For example, the Collaborative Editing System supports a group of researchers working asynchronously on a shared document (Greif, 1986). Quilt is another computer-based tool for collaborative document preparation. It provides annotation, messaging, computer conferencing, and notification facilities to support communication and information sharing among the collaborators on a document (Leland *et al.*, 1988). Many other systems provide for communication and exchanging information and ideas, but do not support information management and document preparation activities (Malone *et al.*, 1986; Goodman & Abel, 1986; Lai & Malone, 1988; Trigg, 1988; Grudin, 1988).

Virtual Notebook supports more research activities (Gorry *et al.*, 1988). It was designed to help the members of a biomedical group coordinate efforts, share information, and improve group functioning through the automated importing and storage of relevant information from external sources. Document preparation tools are also included. "The Analyst" is another system in this category (Xerox, 1989). Although not originally designed as a collaboration tool, it includes flexible information management capabilities, a desktop publishing module, and an extensive, graphic-based hypermedia interface—each of which can be extended to support group activities.

The system studied in this research is the Arizona Analyst Information System (AAIS). It is used by the Mosaic Group at the University of Arizona, which conducts research on international computing trends and policies (Lynch *et al.*, 1990). AAIS supports collaboration among researchers by facilitating entry, indexing, retrieval, and analysis of large amounts of textual information. The system furnishes information modeling and management capabilities, assists document preparation, and supports inter-researcher communication in a collaborative environment (McHenry *et al.*, 1988). It has been in operation since 1983.

The development of the AAIS database has been an ongoing collective effort that has involved more than 70 researchers. Information collection, entry, and management were guided by contract work, Master's and Ph.D. thesis work, and other individual and group research. Initially, the group conducted research on hardware being produced in the Eastern bloc countries, later adding the People's Republic of China. Then topics were expanded to consider software, management information systems, networks, education, and other applications areas. Rounding out the group's efforts has been the study of policy issues related to economic reform, organizations, and industries.

Over the past eight years, the AAIS has become an integral part of all aspects of Mosaic research. We were interested in studying how effectively the system contributes to information management activities. We describe these activities in detail below.

The Mosaic analysts add information to a database of international technology developments through a template-driven process. Most fields in the template were selected from a controlled list of acceptable entries, such as organization names and journal titles. However, the choice of keywords and folders is uncontrolled. (The semi-structured nature of text entry is similar to Malone's "Information Lens" (Malone *et al.*, 1987).)

An example of an article template is shown in Fig. 1. The reference identification is the identifier for the text, created by combining the author name with the year of publication (REFID in Line 2 of Fig. 1). Journal identification indicates the name of the journal where this article appears (JRNID in Line 3). Both English title and foreign language title (ENGTIT and NONENGTIT) need to be entered. (This article was from a Russian journal called "EKO.") Lines 6 to 10 record other miscellaneous information such as number of pages, date of publication, and so on. The last four fields (lines) indicate the country name, the ACM Computing Review Classification Scheme number, the initials of the entry person, and the relevant folders this entry will be sent to (each folder contains texts of similar topics). The content of the document is recorded in the text entry slot (in the middle of Fig. 1), where the analyst has translated and summarized the article. Within the text entry, important semantics-bearing terms have been marked and enclosed by special symbols. The convention used was as follows: "|" for author name (e.g., |A B Gel'b|, "#" for organization name (e.g., #IKANESSR#, and #Goskomizobreteniy#), "^" for keywords (e.g., ^law^, ^invention^, and ^registration^), and "[an:...]" for the analyst's comments about the text.

```

ARTICLE TEMPLATE *****
&REFID: Gelb86
&JRNID: EKO
&ENGTIT: Software Begs Protection
&NONENGTIT: Programnoye obespecheniye prosit okhrany
&NUMBER: 3
&DATE_PUBLISHED: 1986
&ART_PAGES: 159-165
&AUTHORS: A B Gel'b
&PAGE_OF_REF: 159
&TEXT:
|A B Gel'b| is with the #IKANESSR#. [an:: This article is from one of the
EKO publications. From the hard copy, we are not able to tell which city
it is from. However, since we have many more citations from EKO
(Novosibirsk) than from the Kiev and Minsk versions, and because the
IKANESSR has close ties to the Computer Center in Novosibirsk, EKO
(Novosibirsk) seems a good guess.] The questions of legal treatment of
software was first raised in a 1966 publication. Unfortunately, there has
been only presentation of one view or another, without direct discussion
and certainly without widespread discussion by specialists in "law",
programming, etc. At present, the legal status of software is specified in
two documents:: the Interpretation of #Goskomizobreteniy# from Nov. 13,
1975, "On the recognition as "invention"s computer technology objects
characterized by systems software" and "Decree on the state fund of
algorithms and programs ("GosFAP") with the corresponding instructions for
the formulation, "registration" and software expertise."
&COUNTRY_CODES: USSR
&IDNUM: FF/UR/D/SOFTWARE/LEGAL_ISSUES
&ENTRY_PERSON: pw
&SEND_TO_FOLDERS: softlaw.dat; softdstrb.gen

```

Fig. 1. An article template.

An information management process using AAIS typically involves several stages. Analysts first study the source texts and decide which of them are to be entered into the database. They then invoke the system's data entry module. Analysts complete the information fields associated with each text (e.g., author, editor, publisher, etc.) and then translate (if the text appears in a foreign language), abstract, and enter the relevant portion of the text. After entering a text, an analyst indexes it by supplying appropriate descriptors and assigning it to appropriate folders. The system performs syntactic checking of the entry before sending it to the database.

4. RESEARCH METHODOLOGY AND DESIGN

We performed two studies with the cooperation of the Mosaic group. The first involved a Mosaic expert's extensive evaluation of prior document entries, which enabled us to identify the nature and types of information management inconsistency that exist in Mosaic's computer-assisted information management environment. The second study was intended to reveal the sources of information management indeterminism. A methodology based on cognitive psychology was used (although the analysis was task-oriented). We observed and analyzed the process of online information management by researchers, who were asked to think aloud while they were performing three online information management tasks. Their interactions with the information system (revealed in the log files) and their verbal protocols were used to identify the *knowledge elements* and *operator elements* that existed during these processes (two essential components in problem solving (Anderson, 1985)). This research method has been widely used in studying human problem solving (Newell & Simon, 1972) and the cognitive skills, errors, and problems exhibited in human-computer interactions (Card *et al.*, 1983). This study helped us gain significant insights about the causes of the information management indeterminism.

In the first study we randomly retrieved 40 texts from each year's entries into the AAIS database (from 1985 to 1989); 20 texts were entered by senior researchers and the other 20 by junior researchers (or "analysts," as they are referred to in the group). In Mosaic, senior

analysts were researchers who had worked within the group for three years or more and possessed a focused area of expertise within the subject area. Junior analysts were those who had worked with the group for less than two years and exhibited less subject area knowledge. Subject area expertise level has long been recognized as an important factor in human problem solving (Newell & Simon, 1972; Card *et al.*, 1983). It was included as a variable in both of our studies.

In total, 200 texts were retrieved. These texts included: journal article extracts, business cards, book chapters, telephone conversations, conference attendance records, and group member trip reports. Each text contained a complete document entry (abstract or summary), keyword assignment, folder assignment, and so on. The entry person and entry date were removed from the text to avoid potential biases, and the texts were randomized and presented to an expert. The expert was the initial designer of the AAIS system, and has worked with the Mosaic group for over a decade.

Due to the scarcity of expertise in East-bloc computing, we only were able to solicit one expert for our experiment. He was requested to examine each text entry and the assigned descriptors, and then was asked to express any disagreements and critiques about each text and to make corrective suggestions. His comments dealt with the appropriateness of the descriptors, the understandability of the texts, the correctness of folder assignment, translation errors, and so on. Issues related to online information management (not just information indexing) emerged from these comments. Each text took the expert between 5 to 20 minutes to evaluate. Two authors of this article (neither of them are Mosaic analysts) identified the characteristics of the expert's comments and classified them in different categories to indicate the types of indeterminism that occur during the information management process. The findings are presented in Section 5.

The second study attempted to identify the *causes* of information management indeterminism by means of an experiment. We recruited six Mosaic researchers as subjects. Three junior and three senior researchers participated in our empirical study.

The subjects were asked to perform three actual tasks (of their own choosing) in information management using the AAIS. The texts they selected ranged from an article in a foreign language journal or a chapter in a book to a business card, a brochure, or a group member's trip report. The indexing tasks involved were classified by the analysts themselves as easy, medium, or hard.

One experimenter conducted all 18 sessions, including post-session interviews. After a practice session, subjects were asked to think aloud during their interaction with the system. These protocols were tape-recorded and the interactions between the analysts and the system were logged. Most interactions lasted between 20 minutes and two hours.

To analyze the collected data we used *protocol analysis*, which involves detailed examination of the think-aloud protocols and logs. Protocol analysis requires categorizing protocols and logs according to the processes or knowledge components that could have generated them (Ericsson & Simon, 1984). Ericsson and Simon have argued that verbal protocols provide a means of understanding the skills and processes of problem solvers. This approach is particularly useful in subject areas where the task is relatively complex and open-ended. The protocols and the interaction logs allowed us to study in greater detail the knowledge and problems involved in such processes.

In Fig. 2, we show an example (with verbal protocols and associated operations) in which an expert analyst entered a text about "software protection issues in the Soviet Union" from a Russian journal article. On the left-hand side of the figure, the operators used during data entry are represented as arrows. These operators took a certain analyst's knowledge elements as input and produced results as output. The knowledge elements (or knowledge states) are shown as the shaded, round-cornered rectangles in Fig. 2. The result of each operation is represented in an oval adjacent to the knowledge element. The analyst's verbal protocol appears on the right-hand side of the figure.

In summary, our research contained two empirical studies that aimed to identify types of information management indeterminism and the causes of such indeterminism. We believed by conducting these two studies we would be able to address the complexity involved in the online information management process and to generate useful insights for the design of more useful information systems.

Fig. 2. A sample protocol.

5. TEXT ANALYSIS AND EVALUATION

In Study 1, an expert familiar with the system, the subject area, and the classification scheme examined the indexing and content of the texts for completeness and correctness. We offer the results of this evaluation as evidence of the profuse variations in information management.

In examining the texts, two aspects of indeterminism were observed: subject-area related and classification-scheme related. Subject-area indeterminism is rated by determining the completeness and accuracy of the (uncontrolled) keywords assigned to a text and the quality of abstracted/translated text. It reflects mainly the researchers' knowledge of the subject area. Classification-scheme indeterminism relates to the indexer's choice of folders for each individual text and proper application of controlled vocabulary within the indexing template. It reflects the indexer's knowledge of the folder hierarchy within the database and appropriate handling of controlled index terms. We describe these two types of indeterminism in detail, including some examples of the expert's comments.

5.1 Subject area and classification scheme indeterminism

5.1.1 Subject area indeterminism. Four types of subject area related indeterminism were observed in this study.

1. **Keywording difference:** There was evident disagreement between the keywords used by the original analysts and what the expert would have used. Since keyword selection is uncontrolled in the AAIS, knowledge about the subject area and understanding of the document being indexed appear to be the determinants of a keywording decision. Due to the changing nature of the Russian computing area and the lack of uniform, controlled terminologies (especially after translation), Mosaic's underlying indexing philosophy is "analysts know best." Multiple indexing approaches are acceptable. The expert of Study 1 clearly preferred more specific or detailed keywords and made such comments as:

Text 30: "'Universal computers' should be keyworded, but not small, medium, or large computers . . ."

Text 54: "I would keyword 'production line'. Get rid of 'automated' since 'automation' is already there."

Text 55: "Why keyword 'design' when 'CAD' (computer-aided design) is keyworded? Why keyword 'ES' (Edinaya Sistema in Russian or Unified Systems in English) when specific ES devices are keyworded? Instead of 'language', use 'programming language.'"

2. **Keyword omission:** The second type of subject area-related indeterminism was omission of important concepts in the texts. Keyword omission appears to be a more serious problem than keywording difference. Not assigning proper keywords to texts often results in the loss of entry points to texts. A few sample comments from the expert concerning keyword omission are listed below:

Text 10: ". . . no keyword reflecting development, no keyword reflecting concept of space, or concept of scientific application . . ."

Text 11: "No keyword reflecting norms, none reflecting use of punch card . . ."

Text 55: "Software engineering is what this is about, although such a keyword does not appear . . ."

3. **Translation error:** Many Mosaic source texts are written in foreign languages. Researchers need to study, translate, and enter part of the texts into the database. But even senior researchers may occasionally have translation difficulties. This is only natural considering the technical complexity and the rapidly changing nature of the computing subject areas. Our expert, with more than 10 years of experience, critiqued some texts as follows:

Text 18: ". . . translation is obviously awkward. For instance, the (translated) word 'apparatus' should be 'hardware;' as it is now, it is difficult to understand . . ."

Text 163: "Translation problem . . . no such thing as 'package disk,' probably it is 'removable disk.' . . ."

4. **Synthesis ambiguity:** A crucial type of indeterminism existed in the presentation of the texts. Often the expert found it hard to understand a text abstracted and synthesized by another analyst. Browsing a foreign text and summarizing the content are demanding jobs. Very often, the abstracted information was not well articulated or presented. This may be due in part to an analyst's foreign language training, but it is also related to an analyst's ability to summarize a long document in a concise and understandable way (i.e., writing skill). Sample comments from the expert concerning this type of information management indeterminism were:

Text 19: ". . . What does the phrase 'version 2 made through expansion' mean? I don't understand."

Text 6: ". . . First sentence does not make sense: how can an organization host a session of a ministry? . . ."

5.1.2 Classification scheme indeterminism. Variety and inconsistency also appeared in the classification scheme area. Folder assignment inconsistency and controlled vocabulary selection disagreement were the most crucial areas.

1. **Folder assignment difference:** Texts, after having been assigned keywords, needed to be assigned to some existing folders, which are virtual files, not physical locations.

Therefore, each text can be assigned to more than one folder. The expert found that folder assignments may have been done incorrectly (according to the expert's opinion). This problem was often due to an analyst's lack of knowledge about the overall folder directory. Examples of folder assignment disagreement identified by the expert were:

Text 134: "This should not be in 'ewttknow.dat' (folder about east-west technology transfer). It should have been sent to a folder about GDR (East Germany) educational computing, 'educat.hot.' . . ."

Text 137: ". . . why is it in [folder] 'nets.hot' and 'nets.gen'? Should be in USSR computing policy folder . . ."

2. Folder omission: Frequently, the expert recognized that a text should have been sent to a certain folder. Sometimes an analyst appeared to have decided not to send a text to a certain folder if its concepts already had been adequately keyworded. This created information retrieval problems when another analyst assumed that all the relevant information had been stored in a certain folder (and thus had not bothered to search by using keywords). Examples of folder omission comments were:

Text 43: ". . . None of the subsystems of MIS are keyworded as they are in other texts, nor is the text sent to folders for these subsystems."

Text 136: ". . . Should be in intra-CMEA trade folder, folder about policy in GDR computing industry, or GDR information age folder."

3. Controlled vocabulary error: The system employs a controlled vocabulary for a few of the template attributes including country codes, organization identifications, and journal identifications. Analysts needed to use the correct entry for these attributes. The expert identified some controlled vocabulary-related errors in the texts. Examples were:

Text 18: "CMEA is an organization, not a keyword."

Text 200: "CCITT is an organization, not a keyword."

5.2 Analysis of indeterminism

In Table 1, we present a statistical analysis of the expert's disagreements and corrections pertaining to the subject area and the classification scheme.

Subject area indeterminism (count = 385) was more evident than classification scheme indeterminism (count = 167) for both junior and senior analysts. Considering the complexity and the changing nature of international computing research and the relatively static nature of the classification scheme structure, more variety (and thus indeterminism) is expected in the subject area than in the classification scheme.

In the translation and synthesis category, senior analysts exhibited more errors than junior analysts (junior/senior = 4/13, $p = 0.032$). This was because senior analysts' infor-

Table 1. Text analysis

Types	Junior			Senior			Sig. (<i>P</i>)	Total		
	Count	Mean	St. Dev.	Count	Mean	St. Dev.		Count	Mean	St. Dev.
Subject area indeterminism										
Keywording difference	42	0.42	0.768	48	0.48	0.959	0.630	90	0.45	1.374
Keyword omission	141	1.41	1.580	137	1.37	1.590	0.860	278	1.39	2.121
Translation/Synthesis	4	0.04	0.197	13	0.13	0.367	0.032*	17	0.085	0.403
Total (<i>N</i> : 200)	187	1.87	1.606	198	1.98	1.752	0.640	385	1.925	2.250
Classification scheme indeterminism										
Folder assign. diff.	28	0.28	0.683	21	0.21	0.460	0.400	49	0.245	0.719
Folder omission	44	0.44	0.729	57	0.57	0.902	0.260	101	0.505	1.251
Control. vocab. diff.	6	0.06	0.239	11	0.11	0.342	0.240	17	0.085	0.428
Total (<i>N</i> : 200)	78	0.78	1.031	89	0.89	1.136	0.470	167	0.835	1.570
All Totals	265	2.65	1.893	287	2.87	1.930	0.420	552	2.76	2.544

mation management tasks often required more synthesis and translation, whereas junior analysts often entered texts verbatim and/or from translated sources. This practice had been adopted by the Mosaic group in an effort to ensure the quality of the information entered.

As stated earlier, task difficulty levels were perceived differently by different analysts, and varied significantly over time for the same analyst. As an analyst's knowledge and experience matured, tasks assigned to him or her also became more demanding. For senior analysts, because of the diversity in their specialities (e.g., supercomputing, networking, etc.) and target regions (e.g., Poland vs. Russia), an information management task perceived difficult by one senior analyst might have been considered easy by another senior member.

There was no statistically significant difference between the two groups of texts in any other indeterminism categories. Overall, the output-oriented evaluation of the first study did not reveal significant differences in information management problems between texts entered by junior analysts and those entered by the senior analysts.

In subject area indeterminism, keyword omission was most problematic. The expert often identified important concepts not detected by the original entry person. Over 70% (278 of 385 cases) of subject area indeterminism was classified under this category. In a non free-text search system like AAIS, missing descriptors can cause the semantics of the texts to be lost, and thus reduce retrieval effectiveness. In classification scheme indeterminism, folder omission was most problematic (over 60% of the cases, 101 of 167). Folder omission also can cause retrieval difficulty.

6. SOURCES OF INDETERMINISM: A TAXONOMY

Whereas the first study evaluated the *output* of information management (i.e., the archived texts), our second study examined the *process* of information management and helped pinpoint the sources of information management indeterminism. Six analysts (three junior analysts and three senior analysts) participated in the process study. Each analyst performed three actual tasks of text entry and management (18 different tasks in total). Their think-aloud protocols were recorded and the system interactions logged.

A qualitative analysis of the log files and verbal protocols revealed a range of knowledge-related deficiencies and problems. We classified these knowledge-attributed causes in four major categories. The first concerned the researchers' lack of knowledge in the subject area of a text. The second category involved the researchers' incomplete knowledge of the classification scheme. The third type of problem was related to knowledge about the functioning of the system. These three categories of problems have also been reported in research by Chen and Dhar (1990), which investigated the information retrieval process in an online catalog system. The fourth category was lack of knowledge about the collaborative aspects of research. We describe each category in detail and provide examples (actual verbal protocols or operations).

6.1 Subject area problems

The first category of problems was subject-area related. Two major problem areas appeared in our study relating to keyword assignment and text interpretation.

6.1.1 *Keyword assignment difficulties.* Two types of problems were related to keyword assignment.

1. Choosing uncontrolled terms as keywords: The database is indexed with an uncontrolled vocabulary for keywords. Thus, analysts can assign terms of coarse or fine granularity, which produces one type of variety in keywording. In order to choose semantically correct terms, thorough subject area training is essential. Analysts sometimes lacked a proper grasp of the essential ideas of a text due to poor subject area knowledge or the ambiguous nature of the text itself. One analyst remarked, "I've keyworded 'speech transmission' and 'delay', but I'm not sure I would ever search on that second one . . ."

2. Uncertainty about keywording principles: Even though the keyword index is uncontrolled, certain principles are offered as guidelines in the training manual to ensure some uniformity in selection. A specific example is the practice of keywording concepts in their

abbreviated forms (e.g., “IC” for integrated circuits). Other principles are implicit, such as the whole document indexing principle (descriptors taken together should reflect the complete content of a document) and the specific entry principle (descriptors for a document should be as specific as possible). But these rules are sometimes in conflict and offer only hints as to how to develop consistent personal keywording skills over time. Many analysts have developed their own principles or heuristics, which create inter-indexer inconsistency in the keywords chosen.

3. Difficulties in interpreting text: Information in texts can present many challenges to the analyst, related to both content and the expertise of the analyst. We observed three possible causes for text misinterpretation.

- Translation difficulty: Most Mosaic analysts have proficiency in at least one foreign language, many in two or three. However, even the most experienced translator will face difficulties with foreign terms in computer technology. For example, one analyst commented, “The question I have is in the translation of the title: Should I translate it literally or just give my interpretation of it?”
- Difficulty in synthesizing a text for abstracts: Some types of information, such as books and long articles, need to be synthesized to summary length before entry into the database. This process requires extensive subject area knowledge to identify key concepts correctly. Unstructured, ambiguous texts present a challenge to both junior and senior analysts. One junior analyst remarked, “If a senior analyst were entering this, he might not include all the text, but I don’t have a lot of confidence in abstracting the concepts.”
- Ambiguous terms in text: At times the contextual meaning of terms was not clear, due to either poor translation of a source text or an analyst’s lack of subject area knowledge. One analyst complained, “Moscow institute . . . you never know if the translator just threw in the ‘Moscow’ as an adjective, or whether it’s part of the name.” Analysts often need to consult senior members in the group to resolve this type of ambiguity.

6.2 Classification scheme problems

Lack of knowledge about the classification scheme can be identified in problems relating to the folder assignment process and uncertainty about using the controlled vocabularies for organizations and journals.

6.2.1 *Uncertainty about folder assignment.* Problems with folder assignment of texts were identified in two areas:

1. Uncertainty of folder contents: Due to the complexity of the Mosaic folder directory tree, which encompasses many levels of subdirectories and folders, analysts were often faced with the task of making folder assignments to an area of the directory of which they had little knowledge. Folders are named by the analysts who create them, and often the names do not clearly indicate the folders’ subject area; only some folders had online descriptions. Even senior analysts had a problem with making proper folder assignment. As one analyst remarked, “The problem is that the directory tree has to be known, inside and out” (before proper folder assignment can be achieved). Another remarked, “Some folders have weird, fuzzy names like USSR.CORRUPT and USSR.DEBT.”

2. Uncertainty about the folder assignment principle: Texts should be assigned to the folders that are most relevant to the content of that text, and each text can be assigned to multiple folders. Some analysts assigned their texts to folders that were too general (e.g., USSR.SOFTWARE), and sometimes they did not assign a text to all relevant folders.

6.2.2 *Uncertainty about controlled vocabulary assignments.* The system employs a controlled vocabulary for certain fields in a text, such as organization identification and journal identification. Often an analyst was unable to ascertain the full name of an organization or journal from the information provided in the text. In the areas of international computing research, acronyms are used widely for various organizations and publications. This creates a special problem for researchers. For example, one analyst commented: “I

don't know what the name of the journal is. 'EKO' is given . . . I'm not sure it's the name of this (Russian) journal . . . I suspect it's an acronym."

6.3 *System problems*

In the computer-assisted research environment, the system presents another source of indeterminism in information management. We identified two types of system-related problems.

6.3.1 *Use of inappropriate search options.* Many text entry sessions involved unproductive or inefficient processes. Junior analysts, in particular, had difficulty with the system's advanced search options. For example, the senior analysts frequently used the system's underlying query language to search for relevant texts or to identify appropriate index terms. They were able to formulate complex queries using the facilities provided by QUEL, such as wildcards, Boolean operators, relational algebra, and various qualifiers. Junior analysts, however, rarely used QUEL. They relied mainly on the AAIS easy-to-use, menu-driven, information retrieval interface, which supports keyword-based search and Boolean operators. The advanced search options helped the senior analysts identify more relevant and complete indexes for the texts. As junior analysts became more experienced with the system, the tools they selected became more complex and more efficient.

6.3.2 *Incidental system errors.* This category includes errors of omission and carelessness, from incorrect typing to erroneous selection of a system function. An incorrect keystroke would put the data entry process into an unwanted state. Both junior and senior analysts occasionally made these types of errors, with about the same frequency. An analyst complained, "I've mistyped it. I'm going to have to exit and call it (the text) up again."

6.4 *Collaboration problems*

Knowledge of other group members' research activities and expertise is essential in the collaborative research environment. Mosaic researchers need to work with their colleagues in order to monitor their research areas collectively and to gather information relevant to the whole group (i.e., Russian and East European computing). Researchers often must seek assistance in managing information beyond their areas of expertise. We observed two collaboration problems in our study.

6.4.1 *Uncertainty about ongoing research topics.* AAIS permits analysts to send texts to special 'hot' folders that contain important or timely information on topics that are the subject of current research, by an individual or a group of analysts. Often junior analysts expressed uncertainty about making 'hot' folder assignments. One junior analyst remarked, "I hate to send things to hot folders . . . I'm just not sure the texts are appropriate . . ." For some time-critical international computing research topics (e.g., Russian computing policy assessment or U.S. export control recommendations), omission of a hot folder assignment may cause research findings to be incomplete.

6.4.2 *Insufficient communications among researchers.* In addition to the system's support for routing time-critical information to relevant researchers, constant interaction and discussion among group members are essential, especially to ensure effective junior/senior mentor relationships. Junior analysts often needed suggestions or assistance in identifying key concepts in text, in translation of technical terms, or in effective system use.

6.5 *Analysis of problems*

We tallied the number of errors, misconceptions, and problems that appeared in the analysts' log files and verbal protocols. We present our statistical analysis of these problems in Table 2. Statistically significant problem differences between junior and senior subjects (based on *t*-test, 10% significance level) are marked by * in Table 2, which also includes an analysis of the average amount of time spent by the junior and senior analysts for each task.

For problems exhibited overall, there was a ratio of about 3 errors for junior analysts to 1 error for senior analysts (junior/senior = 71/22). The difference was statistically significant ($p = 0.001$, as shown in Table 2).

Table 2. Analysis of problems

Problems	Junior (N: 9)			Senior (N: 9)			Sig. (P)	Total (N: 18)		
	Count	Mean	St. Dev.	Count	Mean	St. Dev.		Count	Mean	St. Dev.
Subject area	26	2.88	0.601	8	0.88	0.601	0.000*	34	1.89	0.547
Classification scheme	13	1.44	1.130	5	0.55	0.527	0.056*	18	1.00	0.707
System	25	2.78	2.440	4	0.44	0.527	0.023*	29	1.61	1.269
Collaboration	7	0.77	0.833	5	0.55	0.726	0.560	12	0.67	0.662
Total (N: 18)	71	7.89	3.100	22	2.44	0.882	0.001*	93	5.17	1.695
Time per task (mins)	—	57.3	31.5	—	43.4	20.5	0.290	—	—	—

The discrepancy between apparent lack of difference in performance between junior and senior analysts in the first experiment and the large differences shown in the second experiment can be attributed to the granularity of information collected in the two experiments and the nature of the two studies. In the first experiment, the evaluation was based on the expert's comments on the information management *results* (indexed texts). There was no information about how these results were obtained. In Study 2, the detailed nature of the log files (operations and commands) and the subjects' verbal protocols during actual problem solving revealed the underlying knowledge involved in deriving those information management results. It showed that junior analysts experienced more problems than senior analysts, despite the fact that junior analysts often handled easier tasks. Even though the performances of the two groups were comparable, the junior analysts' system operations were often less efficient, their subject area knowledge was incomplete, and they had not developed their indexing heuristics.

The highest number of problems for both junior and senior analysts were related to subject area knowledge. There the junior/senior ratio of error was 26 to 8 (significantly different). This was due to the difficulty and specificity of the Mosaic group research—foreign-area studies and assessment of information technologies in the Communist countries. It often required a few years' extensive research in these areas to become a subject area expert.

Problems relating to system knowledge came next, with junior analysts committing the most errors, by a ratio of 25 to 4. Senior analysts rarely committed system-related errors. Junior analysts, however, were more apt to use ineffective or inefficient system options.

With classification scheme problems, the error ratio was 13 to 5, juniors to seniors (significantly different). The indeterminism involved in selecting index terms, which we discussed earlier, helped explain this problem.

Compared to other types of information management problems, collaboration problems had the lowest occurrence. Senior analysts had fewer problems than junior analysts (with a ratio of 7 to 5), but the difference was insignificant ($p = 0.056$).

As to the average time spent per task (junior/senior = 57.3/43.4 minutes), there was no significant difference between the two groups ($p = 0.29$).

7. ALLEVIATING INDETERMINISM: A KNOWLEDGE-BASED DESIGN

An important objective of our research was to develop a framework and eventually a computational model that can alleviate the kinds of problems analysts experience with information management (and information retrieval). Many system and collaboration problems can be removed by extensive training in using the system, context-sensitive online help, and frequent group discussion and communication. Nevertheless, the classification scheme and subject area appeared to be the most problematic areas, and thus deserve special attention.

In the following subsections, we will first present two design principles for alleviating information management indeterminism. We then describe the general architecture of

our knowledge-based framework in the context of prior research. Special focus is placed upon our proposed knowledge representation schemes, a sketch of the inferencing algorithms, and our current progress.

7.1 *Design principles: variety and intelligence*

Two design principles underlie our knowledge-based framework: the *variety principle* and the *intelligence principle*.

7.1.1 *Variety principle: Only variety can destroy variety.* The *variety principle* was proposed initially by Ashby (1979) in cybernetics and elaborated by Bates (1986) in information retrieval. It is the main principle underlying our knowledge-based design.

A majority of information management indeterminism originates from the variety that exists in the subject area and classification schemes. For intra-indexer indeterminism, indexers' knowledge in these two areas varies, affecting the consistency in their indexing behaviors. (This can be perceived as a learning and knowledge acquisition process by the indexers.) For inter-indexer indeterminism, the differences in the indexers' subject-area and classification-scheme knowledge cause variety in their activities. As to inter-system indeterminism, systems' classification schemes differ and thus affect the output of information management.

In order to alleviate these three types of indeterminism, a meta level of variety needs to be maintained. That is, a composite knowledge structure should encompass the vocabularies used by different indexers, different classification schemes, and different systems, and should provide linkages between these vocabularies to permit communication among different agents (systems or people). This meta knowledge structure will "destroy variety by providing variety."

The National Library of Medicine's Unified Medical Language System (UMLS) (Humphreys & Lindberg, 1989; McCray & Hole, 1990) is the best example of such a meta knowledge structure. The project is a long-term effort to build an intelligent automated system of biomedical terms and their interrelationships. The UMLS includes a Metathesaurus, a Semantic Network, and an Information Sources Map. The Metathesaurus contains information about biomedical concepts and their representation in more than 10 different vocabularies and thesauri. The Semantic Network contains information about the types of terms (e.g., "disease," "virus," etc.) in the Metathesaurus and the permissible relationships among these types. The Information Sources Map contains information about the scope, location, vocabulary, and access conditions of biomedical databases of all kinds. Upon completion, UMLS will be able to help users retrieve and organize biomedical information from machine-readable sources.

7.1.2 *Intelligence principle: Intelligent systems generate "choices" for decision makers.* Based on the *intelligence principle*, we propose to create an inferencing component that can aid researchers in making "intelligent" information management and retrieval decisions. Derived from the analysis of the knowledge structure and the researchers' information management and retrieval heuristics, this component will be able to identify "semantically" relevant concepts for information management and retrieval. When deemed appropriate, researchers could invoke the system's inferencing component and solicit its suggestions for keywords, folders, organizations, etc. Incorporating inferencing capability into information systems would lead to more active human-computer interactions and more fruitful information retrieval and management sessions. Various "intelligent" information retrieval systems have been developed in recent years. We will discuss those most relevant to our research in the next subsection.

Due to the diversity and the changing nature of most areas of intellectual inquiries, human researchers are often overwhelmed by the complexity and sophistication of online system, databases, classification schemes, and subject area. A system that "knows" about the subject area and the classification scheme and that can make "intelligent" suggestions might provide partial solutions for our online information management and retrieval problems. In the following subsection we describe our knowledge-based design in detail in the context of prior research.

7.2 A knowledge-based design

In the past few decades, various approaches to improving information management and retrieval performance have been proposed. A major approach is primarily statistics based, and includes: the vector space model (Salton & Yang, 1973; Salton *et al.*, 1975), the probabilistic model (Maron & Kuhns, 1960), and the fuzzy set theory (Tahani, 1976; Bookstein, 1980). For a comprehensive overview, see Salton (1989). The statistics-based approach is appealing because of its sound theoretical basis and independence of subject area.

Recently researchers in information science have been investigating the application of the knowledge-based approach to information indexing and retrieval. Humphrey's MedIndEx System (Humphrey, 1989), discussed earlier, uses inferencing rules and inheritance relations encoded in the knowledge-base frames to perform interactive knowledge-based indexing of the medical literature. The "Intelligent Intermediary for Information Retrieval" (*I³R*), developed by Croft (Croft & Thompson, 1987), consists of a group of "experts": user model builder, query model builder, thesaurus expert, search expert, browser expert, and explainer. These experts communicate by means of a common data structure, called a *blackboard*. The blackboard architecture and the knowledge-based approach have also been adopted in other "intelligent" information system projects by Belkin (Belkin *et al.*, 1983, 1984) and Fox (1987). The fundamental problem with the application of the knowledge-based approach is the development of the "knowledge base." The essential knowledge elements for successful use of the system need to be identified, as do realistic methods for incorporating these knowledge elements into the design of online systems.

In Fig. 3 we show our knowledge-based design, which aims at alleviating information management indeterminism, and provides "intelligent" information management and retrieval assistance. Two important components are depicted in this design:

1. The *AAIS module* represents the components of the current AAIS system. The *interface manager* presents a user-friendly interface for researchers. Flexibility and extensibility are essential so that the system can evolve in response to the changing information requirements of the group. Icon-based interfaces and a streamlined approach to entering information will be used to increase the ease of use of the system. The existing document database and classification scheme (directory, folders, keywords, and controlled vocabularies) will be retained in our proposed design (see the left-hand part of Fig. 3).

Fig. 3. A knowledge-based design.

2. Our proposed design will, however, expand the usability and "intelligence" of the current system by incorporating a *knowledge-based front-end*. This module is the heart of our proposed system. It consists of two parts: a *knowledge base* and an *inferencing engine*. The knowledge base serves as the meta-thesaurus for the group (based on the variety principle), while the inferencing engine provides active and "intelligent" online help (based on the intelligence principle). We describe our frame-based knowledge base and inferencing engine below.

7.2.1 *A frame-based knowledge base.* Use of a knowledge base for information management and retrieval has proven useful in prior research. Shoval (1985) developed an expert system for search term suggestion, using a thesaurus-based semantic network as its knowledge base. A spreading activation algorithm was used to apply knowledge in the network. The GRANT system, developed by Cohen and Kjeldsen (1987) simulated the performance of a funding adviser to find sources of funding for a given research proposal. GRANT's knowledge base was a highly cross-referenced index of research topics, represented in a frame-based semantic network. Humphrey's medical indexing expert system, MedIndEx (Humphrey, 1989), used knowledge-base frames (with numerous specific relations between frames and procedural attachment) to guide indexers in completing indexing frames. It produced as output not only a set of indexing frames for each document, but also conventional MeSH index terms, some of which were generated by inferencing rules encoded in the knowledge-base frames. In a recent study, Chen and Dhar (1991) used the Library of Congress Subject Headings as the basis for creating a knowledge base. A branch and bound algorithm was used to perform heuristic-based spreading activation in the knowledge base. Various knowledge base components have also been incorporated into the design of other knowledge-based information retrieval systems (Croft & Thompson, 1987; Fox, 1987; Chiaramella & Defude, 1987; Bruandet, 1989; Kimoto & Iwadera, 1990). Most of the knowledge bases were either based on existing thesauri or employed a manual knowledge acquisition approach (from human experts). In our research we propose to create the knowledge base from the system's database by adopting some statistics-based algorithms.

The AAIS database maintains a list of keywords, folders, organization names, country codes, people names, and journal titles. These descriptors are used to represent the context and semantic contents of each text. Due to the uncontrolled nature of descriptor assignment, the system has no explicit knowledge about the relationships among these descriptors. However, the database is in fact Mosaic's "organizational memory." Collectively, the texts and their associated descriptors portray the knowledge in the subject area and classification scheme in a comprehensive way. Texts, in particular, create implicit linkages between relevant descriptors. A text may involve several organizations, may be written by a number of authors, may be derived from certain journals, may be indexed under several keywords, and may be assigned to several folders. Similar texts may also have similar descriptors.

Automatic methods for thesaurus construction were suggested in early research (Sparck Jones, 1971; Salton, 1978). Similarity coefficients can be obtained between pairs of distinct descriptors based on the term assignments of the texts of the collection. Virtually all techniques for automatic thesaurus generation are based on the statistical co-occurrence of word types in text (Crouch, 1990; Kimoto & Iwadera, 1990; Salton, 1989).

We adopted the automatic thesaurus generation approach in creating a thesaurus for the complete AAIS database (200 megabytes, 40,000 texts) (Chen & Lynch, 1992). Two algorithms, one based on cosine computation (see Salton, 1989, for details) and one based on the (document) set-subset probability (an algorithm developed by the authors), were adopted. The performance of these two algorithms had been tested in a controlled experiment and reported in detail in Chen & Lynch (1992). We were able to generate relevance (similarity) weights (in terms of probability, between 0 and 1) between any two terms.

Six types of semantic objects were created for a knowledge base we call the Mosaic thesaurus. These include: keywords (describing semantic concepts, e.g., artificial intelligence), folders (virtual files that store related texts, e.g., Russian software protection law folder), sources of texts (journal names, e.g., EKO), people (people mentioned in the text, e.g., A. B. Gel'b), organizations (institutions related to the texts, e.g., Academy of Science in Kiev), and countries (e.g., USSR).

An object and its relationships with other objects can be perceived as a frame, with the slots indicating the relationships. We present an example of this frame-based knowledge representation for a keyword in Fig. 4. Each slot represents a type of relationship, e.g., keyword-to-keyword, keyword-to-folder, keyword-to-organization, etc. Slot values contain a list of object-weight pairs in decreasing weight order (e.g., (export controls, 0.439) (trade, 0.301) (import 0.157)). Other objects also have similar representation. Collectively, the object frames form a network structure. Except for the name slot, which is required, all other slot values are optional.

Currently the Mosaic thesaurus contains more than 20,000 objects and, on average, each object has 14 “relevant” neighbors. A threshold applied during the automatic thesaurus generation process to retain the more relevant terms was determined empirically with the help of Russian computing analysts. Due to the size of this thesaurus, all objects and their relationships are stored as relational tables in INGRES (i.e., 20,000 tables and 280,000 tuples), the underlying database management system of AAIS. We are currently exploring design issues for incremental knowledge base update (for new incoming documents), editing/authoring, and knowledge base security issues. At the implementation level, the frame-based knowledge base has been represented much more concisely (e.g., nil entries are not represented). For storage reasons, each term, object type, and relationship is represented as a unique integer identifier. Each table then contains information regarding relevant identifiers.

7.2.2 Inferencing engine. The knowledge base we have described is aimed at creating a meta and consistent level of semantic variety for both indexers and inquirers (the variety principle). The rich collection of concepts and cross-references provided in the knowledge base enable users of such systems to “dock” onto a network of terms easily (the hit-the-side-of-barn principle described by Bates (1986)) and allow them to explore and navigate in this network. However, in order to help users focus their attention and avoid “getting lost” in this huge network, an inferencing engine is needed for assisting more active and “intelligent” information management and retrieval (the intelligence principle). We have developed two approaches to the creation of an “intelligent” knowledge-based front-end. We describe our approaches below.

The first approach is heuristics-based spreading activation algorithms. Our knowledge base has a network structure. A proven method of inferencing with this knowledge representation is the use of heuristics-based “spreading activation” algorithms that “activate” the links and nodes associated with some source nodes in the network.

Shoval (1985) and Cohen and Kjeldsen (1987) adopted heuristics-guided (e.g., specific-link-first heuristic) activation algorithms for information retrieval. The networks were created manually and were relatively small—on the order of a hundred/thousand. The size

```
{object: KEYWORD
  Name: (literal)
  key_key: [list of keyword-weight pairs]
  key_folder: [list of folder-weight pairs]]
  key_journal: [list of journal-weight pairs]]
  key_people: [list of people-weight pairs]]
  key_org: [list of organization-weight pairs]]
  key_country: [list of country-weight pairs]]
}

{object: KEYWORD
  Name: (technology transfer)
  key_key: [(export controls, 0.439) (trade 0.301) (import 0.157) ...]
  key_folder: [(ewttknow.dat, 0.534)]
  key_journal: [nil]
  key_people: [nil]
  key_org: [(IBM, 0.070)]
  key_country: [(US, 0.265) (USSR 0.155) (CZ 0.084) ...]
```

Fig. 4. A keyword frame.

of our network, 20,000 nodes and 280,000 links, required a more efficient "spreading activation" method.

In prior research, we adopted a heuristic spreading activation process on a Library of Congress Subject Headings-based network for an "intelligent" online catalog (Chen & Dhar, 1991). The heuristics we used considered the specificity of the nodes (based on the number of neighbors), the specificity of the links (e.g., narrower term link, synonymous link, and broader term link), and the number of links between nodes. We represented the information retrieval problem as a *search* task, where the goal was to identify the relevant terms associated with the searcher's search terms. We assigned *costs* to each expanded path on the semantic network based on the nodes visited, the types of links traversed, and the number of links in the path. Cost was a metric we used for indicating the "semantic distance" (relevance) of terms. It was similar to the "Distance" metric used by Rada *et al.* (1989), which represented the conceptual distance between concepts. However, Rada *et al.* used only the path length to determine this conceptual distance.

We used a branch-and-bound algorithm to guide the search. This algorithm was appropriate for efficiently identifying *optimal* paths in large networks. It computed costs for each partial path based on the link types, number of links, and the specificity of each node, and continuously expanded the least-cost path. This allowed our system to locate the optimal (least-cost) path (i.e., identified the most relevant terms in the online thesaurus, which may be multiple links away from the source terms).

We have developed a similar heuristics-based, branch-and-bound, spreading activation algorithm recently for the Mosaic thesaurus (Chen *et al.*, 1992). The algorithm accepts a user's query terms and traverses the Mosaic thesaurus automatically until it identifies other semantically relevant terms (keywords, people, organizations, etc). This module could be used in assisting descriptor assignment (information management) and query articulation (information retrieval).

The second approach to use of our knowledge base follows the neural network paradigm. Our weighted network structure can be perceived as interconnections of neurons and synapses, where neurons represent nodes (keywords, people, organizations, etc.) and synapses represent weighted links. The development of the initial network can be considered a *training* phase, in which each neuron "learns" about its relationships with other neurons. Our system uses large numbers of training instances (i.e., documents and their indexes) for these neurons, which makes the initial knowledge of the neural network rather mature. Documents entered later can be used as new training instances. Through the use of an incremental version of the automatic thesaurus generation algorithms, the system can provide new knowledge for the network to "learn."

Based on this trained neural network, a user's initial query terms (which need to be articulated and refined for information management or retrieval) can be perceived as "noisy" input to a network of terms. The noisy input pattern needs to be disambiguated based on the network's knowledge about the subject area. Applying neural network algorithms iteratively (activating neurons and synapses, computing neuron activation levels during each iteration, performing input/output transformation function, and applying activation threshold) will cause output (suggested terms) to converge, so that suggestions can be made concerning the terms most appropriate to the user's query.

Belew's AIR system is a good example of a neural-network based system (Belew, 1989). The system uses relevance feedback from its users to change its representation of authors, index terms, and documents so that, over time, AIR improves at its task. Other neural network models have also been proposed by various researchers (Kwok, 1989; Wilkinson & Hingston, 1991; Lin *et al.*, 1991). These models are often used to perform "learning" or for conducting adaptive, associative information retrieval.

We have recently developed a Hopfield-net based algorithm for automatic thesaurus consultation (Chen *et al.*, 1993, 1992). (Interested readers are referred to Schwartz, 1989, and Lippmann, 1987, for a complete description of Hopfield net and other neural network algorithms.) The Hopfield network algorithms, similar to the heuristics-based spreading activation algorithms, applies the associative relationships between terms in our knowledge base to identify and suggest potentially relevant and interesting terms to researchers. The

Hopfield algorithm, however, is different from the heuristics-based, branch-and-bound algorithm (a serial search algorithm) in its parallel activation (search) and convergence properties. Our experiment revealed that both algorithms were able to suggest interesting and relevant terms to searchers, but the Hopfield algorithm was significantly slower than the branch-and-bound algorithm (when implemented on a DECstation 5000/120, running ULTRIX) (Chen *et al.*, 1992).

We have incorporated a Hopfield algorithm into the Mosaic system (running on VAX VMS and INGRES) (Chen *et al.*, 1993). With an empirically tuned activation threshold, and often within 4–5 iterations, a Hopfield algorithm can help searchers traverse a large network of terms and converge on semantically relevant terms (relevant to the initial terms). Our prototype performs a complete thesaurus-activation process in less than 20 seconds (INGRES table access was the bottleneck for the activation process). For 4–5 term queries, the access time increased proportionally to about 1–2 minutes. We are in the process of fine-tuning this algorithm.

In this section, we have outlined our knowledge-based framework for alleviating the information management indeterminism (also for the information retrieval problem). We also have described our system development work. We have learned a great deal from our empirical studies and our system development process, and we believe we are getting closer to a practical solution for the problems of information management and retrieval in the computer-assisted research environment.

8. CONCLUSIONS

Information management activities play an important role in creating the “organizational memory” for both business and academia. The process of online information management in the computer-assisted research environment is complex and indeterminate, however. Different users, with different training, expertise, experience, and skills, may extract different information, provide different descriptors for the information, and operate the online systems in different ways. Our research, which examined online, archived texts and observed researchers performing complex, real information management tasks using an operational and successful system, confirmed this postulation. Specifically, we identified information management indeterminism in four areas: subject area problems, classification scheme problems, system problems, and collaboration problems. Subject area and classification scheme problems appeared to be the most difficult and at the same time the most common in the research setting.

We have proposed a knowledge-based framework grounded on two principles: the variety principle and the intelligence principle. Knowledge in an archived database was extracted and used to develop a thesaurus automatically. An effort to incorporate inferencing algorithms into our proposed system is in process. We believe that by incorporating “variety” and “intelligence” into the computer-assisted research environment we will be able to alleviate information management indeterminism and to create a more active and knowledgeable partner for online information management and retrieval.

Acknowledgements—We wish to thank Dr. Edward Fox and three anonymous reviewers for their helpful suggestions. This project was supported in part by an NSF grant, IRI-9211418.

REFERENCES

- Anderson, J.R. (1985). *Cognitive psychology and its implications*, 2nd edition. New York: W.H. Freeman.
- Ashby, W.R. (1979). *An introduction to cybernetics*. London: Methuen.
- Barber, J., Moffat, S., & Wood, F. (1988). Cases studies of the indexing and retrieval of pharmacology papers. *Information Processing & Management*, 24(2), 141–150.
- Bates, M.J. (1986). Subject access in online catalogs: A design model. *Journal of the American Society for Information Science*, 37(6), 357–376.
- Belew, R.K. (1989). Adaptive information retrieval. In *Proceedings of the Twelfth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pp. 11–20, Cambridge, MA.
- Belkin, N.J., Hennings, R.D., & Seeger, T. (1984). Simulation of a distributed expert-based information provision mechanism. *Information Technology*, 3(3), 122–141.

- Belkin, N.J., Seeger, T., & Wersig, G. (1983). Distributed expert problem treatment as a model for information system analysis and design. *Journal of Information Science*, 5, 153-167.
- Blair, D.C. (1986). Indeterminacy in the subject access to documents. *Information Processing & Management*, 22(2), 229-241.
- Bookstein, A. (1980). Fuzzy requests: An approach to weighted boolean searches. *Journal of the American Society for Information Science*, 30(4), 240-247.
- Bruandet, M. (1989). Outline of a knowledge-base model for an intelligent information retrieval system. *Information Processing & Management*, 25(1), 89-115.
- Card, S.K., Moran, T.P., & Newell, A. (1983). *The psychology of human computer interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Chen, H., Basu, K., & Ng, T. (1992). Concept exploration in large knowledge networks: Symbolic branch-and-bound vs. connectionist Hopfield net activation. In *Center for Management of Information, College of Business and Public Administration, University of Arizona*, Working Paper, CMI-WPS 92-11.
- Chen, H., & Dhar, V. (1990). User misconceptions of online information retrieval systems. *International Journal of Man-Machine Studies*, 32(6), 673-692.
- Chen, H., & Dhar, V. (1991). Cognitive process as a basis for intelligent retrieval systems design. *Information Processing & Management*, 27(5), 405-432.
- Chen, H., & Lynch, K.J. (1992). Automatic construction of networks of concepts characterizing document databases. *IEEE Transactions on Systems, Man and Cybernetics*, 22(5), 885-902.
- Chen, H., Lynch, K.J., Basu, K., & Neg, T. (1993). Generating, integrating, and activating thesauri for concept-based document retrieval. *IEEE EXPERT, Special Series on Artificial Intelligence in Text-Based Information Systems*, 8(2), 25-34.
- Chiaromella, Y., & Defude, B. (1987). A prototype of an intelligent system for information retrieval: IOTA. *Information Processing & Management*, 23(4), 285-303.
- Cohen, P.R., & Kjeldsen, R. (1987). Information retrieval by constrained spreading activation in semantic networks. *Information Processing & Management*, 23(4), 255-268.
- Croft, W.B., & Thompson, R.H. (1987). *I³R*: A new approach to the design of document retrieval systems. *Journal of the American Society for Information Science*, 38(6), 389-404.
- Crouch, C.J. (1990). An approach to the automatic construction of global thesauri. *Information Processing & Management*, 26(5), 629-640.
- Ericsson, K.A., & Simon, H.A. (1984). *Protocol analysis: verbal report as data*. Cambridge, MA: The MIT Press.
- Fox, E.A. (1987). Development of the CODER system: A testbed for artificial intelligence methods in information retrieval. *Information Processing & Management*, 23(4), 341-366.
- George, F.H. (1977). *Cybernetics and the environment*. London: Elek Books Ltd.
- Goodman, G.O., & Abel, M.J. (1986). Collaboration research in SCL. In *Proceedings of the Conference on Computer-Supported Cooperative Work (CSCW'86)*, pp. 246-251, Austin, TX.
- Gorry, G.A., Burger, A.M., Chaney, R.J., Long, K.B., & Tausk, C.M. (1988). Computer support for biomedical work groups. In *Proceedings of the Conference on Computer-Supported Cooperative Work (CSCW'88)*, pp. 39-51, Portland, OR.
- Greif, I. (1986). Data sharing in group work. In *Proceedings of the Conference on Computer-Supported Cooperative Work (CSCW'86)*, pp. 175-183, Austin, TX.
- Grudin, J. (1988). Why CSCW applications fail: Problems in the design and evaluation of organizational interfaces. In *Proceedings of the Conference on Computer-Supported Cooperative Work (CSCW'88)*, pp. 85-93, Portland, OR.
- Hooper, R.S. (1965). *Indexer consistency test: Origin, measurement, results and utilization*. Bethesda, MD: IBM Corp.
- Huber, G.P. (1974). Multi-attribute utility models: A review of field and fieldlike studies. *Management Science*, 20(10), 1393-1402.
- Huber, G.P. (1979). The nature of organizational decision making and the design of decision support systems. *The Management Information Systems Quarterly*, 6(2), 1-10.
- Humphrey, S.M. (1989). MedIndEx system: Medical indexing expert system. *Information Processing & Management*, 25(1), 73-88.
- Humphreys, B.L., & Lindberg, D.A. (1989). Building the unified medical language system. In *Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care*. Washington, DC: IEEE Computer Society Press.
- Hurwitz, F.I. (1969). A study of indexer consistency. *American Documentation*, 20, 92-94.
- Jacoby, J., & Slamecka, V. (1962). *Indexer consistency under minimal conditions*. Bethesda, MD: Documentation, Inc.
- Kimoto, H., & Iwadera, T. (1990). Construction of a dynamic thesaurus and its use for associated information retrieval. In *Proceedings of the 13th Conference on Research and Development in Information Retrieval*.
- Kwok, K.L. (1989). A neural network network for probabilistic information retrieval. In *Proceedings of the Twelfth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pp. 21-30, Cambridge, MA.
- Lai, K., & Malone, T. (1988). Object lens: A spreadsheet for cooperative work. In *Proceedings of the Conference on Computer Supported Cooperative Work*, pp. 115-124.
- Leland, M.D., Fish, R.S., & Kraut, R.E. (1988). Collaborative document production using Quilt. In *Proceedings of the Conference on Computer-Supported Cooperative Work (CSCW'88)*, pp. 206-215, Portland, OR.
- Lin, X., Soergel, D., & Marchionini, G. (1991). A self-organizing semantic map for information retrieval. In *Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pp. 262-269, Chicago, IL.
- Lippmann, R.P. (1987). An introduction to computing with neural networks. *IEEE Acoustics Speech and Signal Processing Magazine*, 4(2), 4-22.
- Lynch, K.J., Snyder, J.M., McHenry, W.K., & Vogel, D.R. (1990). The Arizona Analyst Information System: Supporting collaborative research on international technological trends. In *Proceedings of the IFIP WG8.4 Conference on Multi-user Interfaces and Applications*, pp. 159-174, Heraklion, Crete, Greece.

- Malone, T.W., Grant, K.R., Lai, K., Rao, R., & Rosenblitt, D. (1986). Semi-structured messages are surprisingly useful for computer-supported coordination. In *Proceedings of the Conference on Computer Supported Cooperative Work (CSCW'86)*, pp. 102-114, Portland, OR.
- Malone, T.W., Grant, K.R., Lai, K., Rao, R., & Rosenblitt, D. (1987). Semistructured messages are surprisingly useful for computer-supported coordination. *ACM Transactions on Office Information Systems*, 5(2), 115-131.
- Maron, M.E., & Kuhns, J.L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM*, 7(3), 216-243.
- McCray, A.T., & Hole, W.T. (1990). The scope and structure of the first version of the UMLS semantic network. In *Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care*, Los Alamitos, CA: Institute of Electrical and Electronics Engineers.
- McHenry, W.K., Lynch, K.J., & Goodman, S.E. (1988). Handling textual information in a GDSS database: Experience with the Arizona Analyst Information System. In *Proceedings of the 21st Annual HICSS Conference, Decision Support and Knowledge Based Systems Track*, pp. 232-239, Kailua-Kona, HI.
- Newell, A., & Simon, H.A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
- Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1), 17-30.
- Salton, G. (1978). Generation and search of clustered files. *ACM Transactions on Database Systems*, 3(4), 321-346.
- Salton, G. (1989). *Automatic text processing*. Reading, MA: Addison-Wesley Publishing.
- Salton, G., Fox, E.A., & Voohees, E. (1985). Advanced feedback methods in information retrieval. *Journal of the American Society for Information Science*, 36(3), 200-210.
- Salton, G., Wong, A., & Yang, C.S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 913-920.
- Salton, G., & Yang, C.S. (1973). On the specification of term values in automatic indexing. *Journal of Documentation*, 29(4), 351-372.
- Schwartz, J.T. (1989). Neural networks and artificial intelligence. In S.R. Graubard (Ed.), *An artificial intelligence debate: False starts, real foundations*. Cambridge, MA: MIT Press.
- Shoval, P. (1985). Principles, procedures and rules in an expert system for information retrieval. *Information Processing & Management*, 21(6), 475-487.
- Simon, H.A. (1947). *Administrative behavior*. New York: Macmillan.
- Simon, H.A. (1960). *The new science of management decision*. New York: Harper & Brothers.
- Sparck Jones, K. (1971). *Automatic keyword classification for information retrieval*. London: Butterworths.
- Tahani, V. (1976). A fuzzy model of document retrieval systems. *Information Processing & Management*, 12(3), 177-188.
- Trigg, R.H. (1988). Guided tours and tabletops: Tools for communicating in a hypertext environment. In *Proceedings of the Conference on Computer Supported Cooperative Work*, pp. 216-226.
- White, H.D., & Griffith, B.C. (1987). Quality of indexing in online data bases. *Information Processing & Management*, 23(3), 211-224.
- Wilkinson, R., & Hingston, P. (1991). Using the Cosine measure in neural network for document retrieval. In *Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pp. 202-210, Chicago, IL.
- Xerox (1989). *Analyst Production Description*. Pasadena, CA: Xerox Special Information Systems.