Automatic Construction of Networks of Concepts Characterizing Document Databases

Hsinchun Chen and Kevin J. Lynch

Abstract—The results of a study that involved the creation of knowledge bases of concepts from large, operational textual databases are reported. Two East-bloc computing knowledge bases, both based on a semantic network structure, were created automatically using two statistical algorithms. With the help of four East-bloc computing experts, we evaluated the two knowledge bases in detail in a concept-association experiment based on recall and recognition tests. In the experiment, one of the knowledge bases that exhibited the asymmetric link property out-performed all four experts in recalling relevant concepts in East-bloc computing. The knowledge base, which contained about 20 000 concepts (nodes) and 280 000 weighted relationships (links), was incorporated as a thesaurus-like component into an intelligent retrieval system. The system allowed users to perform semantics-based information management and information retrieval via interactive, conceptual relevance feedback.

I. INTRODUCTION

THE AVAILABILITY of cheap and effective storage devices and information systems has prompted the rapid growth of relational, textual, and graphical databases in the past two decades. Information collection and storage efforts have become easier, but the efforts required to retrieve relevant information have become significantly more difficult, especially in the large-scale databases. Information stored in these databases often has become fragmented and unstructured after years of intensive use. Only users with extensive subject area knowledge, system knowledge, and classification scheme knowledge are able to maneuver and explore in these databases. Much crucial information and the knowledge underlying it is buried in these databases and therefore has become inaccessible.

An important fact about large-scale, real-world databases is that significant amounts of organization-related knowledge is often embedded in voluminous data. For example, in an airline company's frequent flyer database, patterns of traveling arrangements and scheduling can be hidden in the records. Banks can analyze customers' credit data to reveal patterns of bad loans and to derive rules for credit assignment. Historical cases of tax fraud can disclose patterns of taxpayers' behaviors and provide indicators for potential fraud. A business or research intelligence database may be able to reveal what persons, organizations, projects, and topics are relevant to

a particular event of interest. Whenever a large amount of information is collected and captured in a database, important domain knowledge also resides there.

Creating systems with knowledge or "intelligence" has long been the goal of researchers in expert systems and artificial intelligence. Many interesting knowledge-based systems have been created in the past decade for such different applications as medical diagnosis, engineering troubleshooting, and business applications. A recent approach to knowledge elicitation is referred to as "knowledge mining" or "knowledge discovery in databases." This approach, which has been gaining attention from researchers in different disciplines, is based on the rationale that growth in the number of large databases greatly exceeds the development of knowledge bases, creating both a need and an opportunity for extracting knowledge from databases. Many AI and database researchers believe that, "knowledge discovery in databases" will become one of the most important research areas in this decade.

The research reported here adopted various algorithms for automatic text processing by extracting semantic knowledge from textual information. Specifically, we generated two knowledge bases automatically from a large (200 Megabytes), operational, textual research database in the subject area of East-bloc computing. We conducted a concept-association experiment based on recall and recognition tests, comparing the output of two knowledge bases with concepts generated by East-bloc computing researchers. The experiment confirmed that knowledge discovered algorithmically is robust and useful for researchers. We then selected one of these knowledge bases and have successfully incorporated it into our information retrieval system. Analysts use it to perform semantic query articulation and concepts refinement for both information retrieval and information management.

This article is organized as follows. In Section II, we present an overview of research in knowledge discovery and automatic knowledge base generation. In Section III, we discuss our algorithms for knowledge base generation and describe the characteristics of the databases from which the knowledge was acquired. In Section IV, we describe our knowledge base evaluation. Detailed results about the concept recall and concept precision performance of the subjects and of our knowledge bases are presented in Section V. Section VI describes the semantics-based information management and information retrieval modules we implemented in a retrieval system after incorporating the knowledge base. We present the conclusions in Section VII.

Manuscript received July 1, 1991; revised January 6, 1992.

The authors are with the MIS Department, University of Arizona, Tucson, AZ 85721.

IEEE Log Number 9201423.

II. KNOWLEDGE DISCOVERY FROM DATABASES

Computerized information systems in the forms of relational database management systems, bibliographic databases, online catalogs, business intelligence systems, research collaboration systems, and decision support systems have been helping users organize and manage relevant information and retrieve and analyze it as needed. Information such as articles, product descriptions, customer information, business letters, book chapters, stock information, and financial news are the basic information units in such systems. Even though these systems may function differently, the fundamental problem of information management and retrieval remains the same when the amount of information stored becomes massive. As a general rule, the amount of effort required to retrieve relevant information is proportional to the amount of information stored. Among the major reasons information retrieval is difficult are the lack of explicit semantic clustering of or linkages between relevant information (we refer to this as the "information fragmentation" problem) and the limits of conventional keyword-driven search techniques (either full text or index-based). Semantics-based query articulation and data analysis are often ranked among the top-priority database research issues.

Expert systems or knowledge-based systems [20], [40], on the other hand, aim to capture human expertise or knowledge by means of computational models. After over a decade of active research in this area, many practitioners and researchers have suggested that the bottleneck in the design of knowledge-based systems is the (manual) knowledge acquisition process [13]. Knowledge acquisition, which was defined by Buchanan [16] as "the transfer and transformation of potential problem-solving expertise from some knowledge source to a program," demands extensive effort on the part of knowledge engineers, who need to interact with subject experts and model their knowledge and expertise in detail and completeness. For a large, realistic domain, knowledge acquisition may take several person-years. For some areas where there are few cooperative human experts, the manual knowledge acquisition approach is not practical.

A. Knowledge Discovery in Real Databases

The "knowledge discovery" approach is believed by many AI and database researchers to be useful for resolving the information overload and knowledge acquisition bottleneck problems. As Piatetsky-Shapiro [33] remarked in the IJCAI-89 (International Joint Conference of Artificial Intelligence, 1989) workshop on "knowledge discovery in real databases":

The growth in the amount of available databases far outstrips the growth of corresponding knowledge. This creates both a need and an opportunity for extracting knowledge from databases. Many recent results have been reported on extracting different kinds of knowledge from databases, including diagnostic rules, drug side effects, classes of stars, rules for expert systems, and rules for semantic query optimization... The importance of this topic is now recognized by leading researchers. Michie predicts that "The next area that is going to explode is the use of machine learning tools as a component of large scale data analysis" (*AI Week* March 15, 1990). At a recent NSF Invitational Workshop on the future of database

research (Lagunita, CA, February 1990), "knowledge mining" was among the top five research topics.

Sample areas where knowledge discovery is applicable include: identifying patterns in frequent flyer databases and behaviors of credit card holders, developing diagnostic expert systems from car trouble symptoms and problems found, and examining corporate intelligence reports and patterns of airline crashes or tax fraud [32], [33], [38]. Many major companies and government agencies are "learning" as much as possible from their databases. The knowledge discovered has often resulted in significant competitive advantage. Many designers adopt production rules as their main knowledge representation scheme. However, other representation schemes such as semantic networks, frames, decision trees, logic, and behavior scores have also been used [29], [30], [32], [36]–[38].

B. Automatic Thesaurus Generation

In this research, our aim was to apply an algorithmic approach to the generation of a robust knowledge base based on statistical correlation analysis of the semantics (knowledge) embedded in the documents of real, textual databases. The research output consisted of a thesaurus-like, semantic network knowledge base, which can aid in semantics-based information management and retrieval.

Use of thesaurus or knowledge base for "intelligent" information retrieval has drawn attention from researchers in information science and computer science in recent years. There have been many attempts to capture expert's domain knowledge for information retrieval. CoalSORT [31], a knowledge-based interface, facilitates the use of bibliographic databases in coal technology. A semantic network, representing an expert's domain knowledge, embodies the system's intelligence. GRANT [10], developed by Cohen and Kjeldsen, is an expert system for finding sources of funding for given research proposals. Its search method - constrained spreading activation in a semantic network — makes inferences about the goals of the user and thus finds information that the user did not explicitly request but that is likely to be useful. Shoval [47] developed an expert system for suggesting search terms. This system is composed of two components: the knowledge base, represented as a semantic network in which the nodes are words, concepts, or phrases. Links express the semantic relationships between the nodes. The second component is made up of rules, or procedures, which operate upon the knowledge base and are analogous to the decision rules or work patterns of the information specialist. Fox's CODER system [14] consists of a thesaurus that was generated from the *Handbook of Artificial Intelligence* and *Collin's Dictionary*. In CANSEARCH [34], a thesaurus is presented as a menu. Users browse and select terms from the menu for their queries. The "Intelligent Intermediary for Information Retrieval" (*I³R*), developed by Croft [11], consists of a group of "experts" that communicate via a common data structure, called the blackboard. The system consists of a user model builder, a query model builder, a thesaurus expert, a search expert (for suggesting statistics-based search strategies), a browser expert, and an explainer. Chen and Dhar [8] incorporated a portion of the *Library of Congress Subject Headings* into the design of

an intelligent retrieval system. The system adopted a branch-and-bound spreading activation algorithm to assist users in articulating their queries. All the systems described above used some type of knowledge base to assist information retrieval.

The National Library of Medicine's thesaurus projects are probably the largest-scale effort that use the knowledge in existing thesauri. In one of the projects, Rada and Martin [27], [39] conducted experiments for the automatic addition of concepts to MeSH (medical Subject Headings) by including the CMIT (current Medical Information And Terminology) and SNOMED (Systematized Nomenclature of Medicine) thesauri. Access to various sets of documents can be facilitated by using thesauri and the connections that are made among thesauri. The Unified Medical Language System (UMLS) project is a long-term effort to build an intelligent automated system that understands biomedical terms and their interrelationships and uses this understanding to help users retrieve and organize information from machine-readable sources [18], [22], [28]. The UMLS includes a Metathesaurus, a semantic network, and an Information Sources Map. The Metathesaurus contains information about biomedical concepts and their representation in several 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.

Most of the knowledge bases used in these intelligent systems were either generated manually from the domain experts via the knowledge acquisition process or derived from some existing thesauri (which were also created manually in the first place by some indexing/subject experts). A complementary approach to manual knowledge base creation is the *automatic thesaurus generation* approach.

Virtually all techniques for automatic thesaurus generation are based on the statistical co-occurrence of word types in text [12], [45]. Similarity coefficients are often obtained between pairs of distinct terms based on coincidences in term assignments to the documents of the collection. For example, a cosine computation can be used to generate normalized term similarities between 0 and 1. We discuss two similarity computations in detail in the next subsection. When pairwise similarities are obtained between all term pairs, an automatic term-classification process such as single-link or complete-link classification can group into common classes all terms with sufficiently large pairwise similarities [43], [45]. The terms in the thesaurus classes can replace the initial search terms and be used to increase retrieval recall.

Most automatic thesaurus generation research has used sample document collections to generate term relationships and thesaurus terms have replaced search terms automatically without searchers' relevance feedback. Recall improvements of the order of 10 to 20 percent have been demonstrated when the thesaurus is used in an environment similar to that in which the original thesaurus was constructed [12], [42], [46].

After examining past research and its pitfalls closely, we believe that creating a robust and useful thesaurus automatically requires both a complete document collection to extract knowl-

edge and searcher relevance feedback. A thesaurus should represent the complete knowledge in the document collection and keep up-to-date with its underlying database. It should also be used as an intelligent aid for interactive information retrieval. The *automatic thesaurus generation* approach, if applied properly, can be extremely powerful in capturing the domain knowledge in textual databases and creating an environment for semantics-based information management and retrieval. We describe our experience in using such an approach in detail in the following section.

III. AN ALGORITHMIC APPROACH TO KNOWLEDGE BASE GENERATION

We adopted two statistics-based algorithms to extract knowledge from textual data. Both algorithms were based on the frequency of concepts co-occurring in the documents. The resulting knowledge is captured in a semantic network representation where nodes represent different types of concepts and weighted links indicate their strengths of relevance. We describe the characteristics of the database we studied first. We then discuss how we extracted knowledge from this database.

A. The Mosaic Environment: Manual Indexing

The textual database we studied was created by the Mosaic international computing research group at the University of Arizona. It is designed to support research on computing of the (former) East-bloc countries [9], [25]. The Mosaic researchers have maintained and used the database for the past nine years in areas such as East-bloc computing evaluation, industry policy analysis, technology assessment, and U.S. export control recommendation. Information in the Mosaic database was collected, entered, and indexed manually by the researchers themselves. This database has become an integral part of the Mosaic research. Currently, the database contains information about 131 countries, 10 000 organizations, 1500 journals, 40 000 keywords, 40 000 documents, and 10 000 folders (for specific topics of interest to the group)—resulting in about 200 MB's of textual information stored in INGRES¹. However, the majority of the information is about the 21 East-bloc countries. In Table I, we show the growth rates of the number of attributes relating to the subject area in the Mosaic database over the previous eight years.

Information was entered by the Mosaic analysts through a template-driven process. This information ranged from articles, book chapters, and technical reports to business cards, foreign trip reports, and electronic mail messages. Most fields in the template were based on a controlled list of acceptable entries, for example, organization names and journal titles. However, the choice of keywords and folders was uncontrolled. (The semi-structured nature of document entry is similar to Malone's "Information Lens" [26].)

An example of an article template is shown in Fig. 1. Reference identification is the unique identifier for the document, created by combining the author name with the year of publication (REFID in Line 2 of Fig. 1). Journal identification

¹A relational database management system vended by ASK/Ingres Corporation, Inc.

TABLE I
GROWTH OF THE MOSAIC DATABASE OVER TIME

Year	no. of documents	no. of folders	no. of keywords	no. of countries	no. of organizations	no. of journals
1982	1725	1160	3263	25	345	129
1983	4589	3029	8423	50	932	302
1984	10239	4935	16310	85	2784	523
1985	15316	6276	21881	95	4408	826
1986	20049	7425	26941	100	5822	988
1987	25767	8146	30845	108	7174	1130
1988	31100	8938	35620	126	8808	1281
1989	35451	9397	40236	131	10074	1383
Growth rate	4999	1167	5280	14.2	1446	186

```

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.

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 to be sent to (each folder contains documents 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 document.

An information management process using AAIS typically involves several stages. (Readers are referred to [9] for detailed description of this process.) Analysts first study the source documents and decide which of them is to be entered to the database. They then invoke the system's data entry module. Analysts complete the information fields associated with each document (e.g., author, editor, publisher, etc.) and then translate (when the document appeared in foreign language), abstract, and enter the relevant portion of the document. After entering a document, an analyst indexes it by supplying appropriate descriptors and assigning it to appropriate folders. The document abstracting and indexing stages require extensive subject area and classification scheme training.

For each document stored in the database, descriptors that include keywords, persons, organizations, countries, and fold-

ers indicate the context and content of the document. Collectively, the documents and their descriptors represent the knowledge in the East-bloc computing domain. Documents, in particular, provide explicit linkages between relevant descriptors in the domain. Documents of similar content, collected over time from different sources, often contain similar descriptors. The co-occurrence of descriptors in the documents stored in this large-scale database can reveal the relationships between the important topics (projects, computers, policy, etc.), crucial persons, relevant organizations, and countries in East-bloc computing. By examining all the documents in the Mosaic database, we were able to perform extensive data analysis and to extract a significant amount of East-bloc computing knowledge.

B. Term Co-occurrence Algorithms: Cosine and Cluster

We used two statistics-based algorithms to create our knowledge bases automatically — one based on normalized cosine computation (Cosine algorithm, an algorithm frequently used in automatic text processing [45]) and one based on the probability of term co-occurrence in document sets (an algorithm developed by the authors, referred to as the Cluster algorithm). These two algorithms have similar underlying justification — relevant concepts often co-occurred in the same document. However, the knowledge bases created by the two algorithms exhibit different structures. Links in the Cosine knowledge base (we refer to it as Cosine *KB* below) are symmetric and equally-weighted, i.e., for any relevant pair of concepts *A* and *B*, there is a link from *A* to *B* and another one from *B* to *A* and these links have exactly the same weight associated with them. But the links in the Cluster *KB* are asymmetric (i.e., there may be a link from *A* to *B*, but no link from *B* to *A*) and their weights are often different. Because of this fundamental difference in network structure, we generated both knowledge bases and compared their performance. After the evaluation, we selected and integrated one of these knowledge bases with the Mosaic database.

We sketch the procedure and the two algorithms used for creating our knowledge bases below:

- 1) *Determine unique descriptors*: We first generated all unique descriptors for all the documents in the Mosaic database, including: keywords, organizations, persons, countries, and folders. To be included, each descriptor had to appear at least three times in the database. The threshold of three appearances in the database ensured that identifiers included are common terms in the database. Some obscure keywords, person names, and organization names were eliminated during this process. We assigned a unique identification number for each descriptor and associated with each descriptor its number of occurrences in the entire database. The number of occurrences was later used in our inferencing algorithms (discussed later) to determine the “goodness” of a descriptor based on the “inverse document frequency” principle (less frequently-used terms are better and vice versa, see [45]). These unique descriptors were stored as relations in INGRES.

- 2) *Weight computation*: For each unique descriptor, we computed its term co-occurrence probabilities with all other descriptors. The term co-occurrence probability, which is a real number between 0 and 1, indicates the *similarity* measure [45] between any two descriptors. We adopted the following two algorithms:

- a) *Cosine Algorithm*:

$$\text{Weight}(T_j, T_k) = \frac{\sum_{i=1}^n d_{ij} \times d_{ik}}{\sqrt{\sum_{i=1}^n d_{ij}^2 \times \sum_{i=1}^n d_{ik}^2}}.$$

This indicates a symmetric similarity weight between T_j and T_k . Where T_j represents descriptor *j*; T_k represents descriptor *k*; *n* indicates the total number of documents in the database; d_{ij} indicates descriptor T_j in document *i* (value: 0 or 1); and d_{ik} indicates descriptor T_k in document *i* (value: 0 or 1).

- b) *Cluster Algorithm*:

$$\text{Weight}(T_j, T_k) = \frac{\sum_{i=1}^n d_{ij} \times d_{ik}}{\sum_{i=1}^n d_{ij}}$$

$$\text{Weight}(T_k, T_j) = \frac{\sum_{i=1}^n d_{ij} \times d_{ik}}{\sum_{i=1}^n d_{ik}}.$$

These indicate similarity weights from T_j to T_k (the first equation) and from T_k to T_j (the second equation). Where d_{ij} indicates descriptor T_j in document *i* (value: 0 or 1) and d_{ik} indicates descriptor T_k in document *i* (value: 0 or 1).

Notice that the amount of computation required was quite large. We computed the co-occurrence of the 60 000-plus descriptors in our complete database. These two algorithms acted as a “batch” learning process, making it possible to examine the patterns in all the documents that were collected by the Mosaic researchers in the past decade.

- 3) *Create knowledge bases*: The output from the above computation was stored in INGRES as tables. However, in order to obtain a reasonable size of co-occurred descriptors for each descriptor, we adopted a weight threshold of 5%. This threshold also ensured that only the most relevant descriptors were represented in our knowledge bases. Each INGRES table recorded a source descriptor and all of its relevant descriptors with weights.

Five types of semantic objects in the Mosaic database were extracted by our algorithms: keywords (describing topics, machines, projects, and so on, e.g., “technology transfer,” “MVS 810,” etc.), folders (virtual files that store a collection of documents of interest to the Mosaic researchers, e.g., “softlaw.dat” for Russian software protection law folder), persons (persons related to a document, e.g., “Y. Andropov”), organizations (institutions related to a document, e.g., “Academy of Science in Kiev”), and countries (e.g., “USSR,” “Poland,” etc.). For each pair of objects there is a probability (between 0 and

```

KEYWORD FRAME --

(object: KEYWORD
  name:      (required: [literals])
  #_of_occurrences: (required: [integer])
  keyword:    (optional: [(list of keyword-weight pairs)])
  folder:     (optional: [(list of folder-weight pairs)])
  person:     (optional: [(list of person-weight pairs)])
  organization: (optional: [(list of organization-weight pairs)])
  country:    (optional: [(list of country-weight pairs)])
)

AN EXAMPLE --

(object: KEYWORD
  name:      [technology transfer]
  #_of_occurrences: [1742]
  keyword:    [(export controls, 0.439) (trade, 0.301) (covert 0.222)
               (export 0.173) (import, 0.157) (micro-electronics 0.132)
               (software 0.113) (microcomputer 0.081) (microprocessor 0.063)]
  folder:     [(ewttknow.dat, 0.534)] ;East-west technology transfer folder.
  person:     nil
  organization: [(IBM, 0.070)]
  country:    [(US, 0.265) (Hungry 0.156) (USSR 0.155) (UK 0.135)
               (Japan 0.119) (Poland 0.092) (Czechoslovakia 0.084)]
)

```

Fig. 2. A frame-based representation for keywords.

1) that indicates their strength of relevance. Collectively, we therefore could consider our knowledge base as a frame-based semantic network, where nodes represented objects and links represented relationships between objects. Fig. 2 shows the frame-based representation for keywords. The frame contains attributes (slots) that indicate the properties of the keyword (e.g., name and number of occurrences in the database) and its weighted relationships with other objects (keywords, folders, persons, organizations, and countries). Similar frame-based representations were created for the folder, person, organization, and country objects.

In Table II we present an overview of the two knowledge bases we created based on the Cosine and the Cluster algorithms, respectively. Because folders created by individual researchers often have obscure names (e.g., "ewttknow.dat" for east-west technology transfer), for our knowledge base evaluation (which involved human subjects) we did not include folders. After the evaluation, we selected one of the two algorithms and regenerated a complete knowledge base that included folders.

Both the Cosine and the Cluster KB's adopted the frame-based representation as shown in Fig. 2. A fundamental difference between the two knowledge bases, however, was that the Cosine KB had symmetric links, but the links in the Cluster KB were asymmetric. According to human memory research, associations between learned concepts are often asymmetric. (For a good overview of the concepts and principles of human memory, readers are referred to [1] and [23].) For example, it may be easier to associate "net" with "volleyball" than to associate "volleyball" with "net." Human beings tend to acquire these asymmetric, weighted associations through experience. The Cluster KB also exhibited a similar asymmetric property in its knowledge structure.

A detailed comparison revealed more precise structural differences and similarities between these two knowledge

bases. As shown in Table II, on average, each source descriptor in the Cosine KB had 8.9 related descriptors while the number of related descriptors was slightly higher (9.6) in the Cluster KB. Overall, the Cluster KB captured more source descriptors (10291) than the Cosine KB (8926). The Cluster KB had more source descriptors that had between 2 and 50 neighbors (related descriptors) than the Cosine KB. The Cosine knowledge base, on the other hand, had more source descriptors having a large number of neighbors (more than 50 related descriptors). These descriptors were often concepts that are more general and have very high occurrences in the database, e.g., software, USSR, etc. But they may be too general to be useful in assisting online information management or retrieval (the "inverse document frequency" principle, described earlier). From structural evaluation, the Cluster KB appeared to demonstrate better properties — it contained more specific, related terms. However, a more detailed performance comparison was needed. We discuss this performance comparison in Sections IV and V.

It took a few weeks of CPU time on a small VAX/VMS mainframe to generate these two knowledge bases. However, considering the scarcity of expertise in the East-bloc computing area, the knowledge extracted, if robust, could be extremely useful to researchers. With a proper interface, researchers could use the knowledge base to help them articulate their queries, identify relevant topics and projects, pinpoint crucial contacts, persons, and organizations, and eventually be able to retrieve from the database relevant documents for their research. Integrating the knowledge base with the underlying database would enable researchers to perform semantic query articulation and semantics-based information management and retrieval. The knowledge base could also serve as a platform for training, by allowing new researchers to explore in the network of knowledge and data. More discussion is presented in Section VI.

TABLE II
STRUCTURAL COMPARISON: COSINE KB VERSUS CLUSTER KB

Characteristics no. of neighbors	Cosine KB Count	Cluster KB Count
2-5	5257	5199
6-10	1701	2064
11-20	1183	1672
21-30	388	797
31-40	180	349
41-50	81	136
51-100	105	72
101-200	27	2
201-300	3	0
301-400	1	0
Total no. of descriptors	8926	10 291
Total no. of neighbors	80 183	99 230
Average no. of neighbors	8.9	9.6

C. The Public Database: A Test of Automatic Indexing

The Mosaic database was produced in a manual indexing environment, within which international computing researchers carefully generated appropriate indexes for documents. In an attempt to generalize the applicability of our approach in an environment where texts can be processed and indexed automatically, we had applied a similar *automatic thesaurus generation* approach to another textual database. This database contained mostly abstracts of articles extracted from the DIALOG database in the areas of database management systems and information retrieval. The database (called the Public database) consisted of about 3000 articles. Some indexes (keywords) had already been assigned to these documents by the DIALOG database when they were extracted.

In addition to these DIALOG-assigned indexes, we also extracted other keywords from documents by using automatic indexing techniques [45]. The complete automatic indexing procedure is outlined below:

- 1) We first collected all indexes that had been assigned to the extracted documents. We then deleted the bad keywords (broad indexes) and solicited other relevant indexes from an expert database user and generated a controlled list of desired index terms (words and phrases).
- 2) For each document in the Public database, we identified the individual words in the document.
- 3) We used a stop word list to delete common function words (e.g., but, while, the or, etc.) in the documents.
- 4) We performed phrase formation by using combinations of adjacent word sequences (1-word, 2-word, and 3-word combinations). For example, we generated three phrases: "information retrieval," "retrieval system," and "information retrieval system" from the extracted word sequence of "information" "retrieval" "system" in the document.
- 5) The phrases and words generated from the above process were then checked against the controlled list of vocabularies created in Step 1. The resulting matched terms were added to the document as extra indexes.
- 6) After the indexes had been assigned automatically, we

followed the same thesaurus generation procedure that was used for the Mosaic database. The Cluster algorithm was used to generate term co-occurrence.

By running our Cluster algorithm on this automatically indexed database, we were able to generate a knowledge base in the areas of database management and information retrieval. The Public thesaurus contained 1488 terms and 29 883 weighted links (20 links per term). Because of the automatic indexing procedure, each Public term had significantly more associated links than a Mosaic term, which had about 9 links per term. Again, the knowledge discovery algorithm was applied on the complete database, instead of sample collection. We believe this informal testing shed light on the applicability of our approach for automatic indexing applications.

IV. KNOWLEDGE BASE EVALUATION: AN EXPERIMENT

Before integrating the knowledge base with the database in use, the performance of the knowledge base had to be evaluated. With the assistance of four East-bloc computing researchers, we performed an experiment based on recall and recognition tests (tests frequently used in experimental psychology). This experiment helped us determine the performance levels of our two knowledge bases. We discuss the experimental design and an overview of the results generated in this section. We present detailed hypothesis testing results comparing human subjects with our knowledge bases using the *concept recall* and *concept precision* measures in Section V.

A. Experimental Design

Even though we cannot claim (nor did we intend to claim) that the computational model of our knowledge bases is a fully accurate model of human memory, the semantic network structure of subject descriptors is similar to the way human long-term memory captures and associates concepts (structural resemblance) [1], [35]. In order to test the robustness of our knowledge bases, we resorted to experimental designs adopted in human memory experiments, in particular, the recall and recognition tests.

According to human information processing theory, a human being encodes and stores information in long-term memory [6] and relevant information is strongly associated. Human long-term memory can be perceived as a network of nodes (information) and links (associations). Information is retrieved when proper stimuli reach long-term memory. This process is referred to as "spreading activation," during which relevant information and associations are activated as in a chain reaction. Consistent with the hypothesis that there exists in human long-term memory information that cannot be recalled (due to poor stimuli or wrong clues) is that humans can recognize things they cannot recall [1]. The recognition process often involves more clues (stimuli) than the recall process. This phenomenon suggests that a recognition test (testing human memory with clues presented) can help identify more long-term information stored in human memory than a recall test (testing memory with no clues).

We adopted a two-phase experiment for our knowledge base evaluation. Phase 1 involved a recall test of 50 randomly-

TABLE III
SAMPLE TEST DESCRIPTORS

Descriptors	Test Cases with Explanation
Persons	GOVORUN (Nikolai Nikolayevich Govorun, USSR) ODHNER (V. T. Odhner, USSR) ANDROPOV (Yuri Andropov, USSR)
Organizations	MEKHATRONIKA (Mechatronics Combine, Bulgaria) MSHET (Ministry of Science, Higher Education, and Technology, Poland) IVTSSOANSSSR (Irkutsk Computer Center, Siberian Academy of Science, USSR) TASHSELMASH (Tashkent Factory of Agricultural Machinery, USSR)
Keywords	IBM 5550 MAXIMUM TRANSFER RATE MVS 810 DIAGNOSTIC SYSTEM UCSD P-SYSTEM

retrieved descriptors from the knowledge bases. Subjects were asked to recall related concepts without clues (e.g., keywords, organizations, persons, etc.) for each of these 50 descriptors in East-bloc computing. At Phase 2, we collected all related concepts recalled by the subjects in Phase 1, the related descriptors suggested by the two knowledge bases, and some unrelated (noise) descriptors randomly generated from the knowledge bases as clues for a recognition test. Subjects were asked to pick the descriptors they deemed relevant in the context of Mosaic research. Results from Phase 2 served to represent the subjects' long-term perceptions about the domains. They were then used as a benchmark to compare the performance of the two knowledge bases. We summarize our experiment procedure below:

- 1) *Tasks and Subjects Selection:* We randomly selected 50 descriptors from the knowledge bases. Table III shows examples of these descriptors. These test descriptors included three person names, 15 organizations, and 32 keywords (topics, machines names, etc.). We recruited four East-bloc computing researchers from the Mosaic group as subjects.² Two subjects were considered senior researchers, while the other two were junior researchers. The senior researchers had worked in East-bloc computing for 10 and 6 years, respectively. Each of them had published numerous academic papers and reports in this area. The junior researchers each had less than two years of research experience in East-bloc computing with the Mosaic Group. However, they had considerable computer and foreign language training before they were recruited as researchers in the Mosaic Group.
- 2) *Recall Test:* At week N , we performed the recall test by asking subjects to recall as many concepts that related to the test descriptors as they possibly could. The 50 descriptors were presented to them in order and their responses were recorded.
- 3) *Recognition Test:* We performed the recognition test at week $N + 2$. The two weeks of lag time served to decay

²Due to the scarcity of East-bloc computing experts in the West, we were only able to solicit the Mosaic researchers as subjects, instead of other non-Mosaic researchers. The subjects were among the 40 plus researchers who had helped create the Mosaic database during the past decade.

the effects of the recall test. For each of the 50 test descriptors, we collected the responses generated by the subjects in the recall test, the related terms suggested by the Cosine and Cluster KB's, and a set of randomly-generated noise terms (the size of the noise set was 25% of the total number of terms generated by all subjects in the recall test and by the two knowledge bases). The 50 descriptors with the set of clues were presented to the subjects in order. They were asked to select terms that they deemed relevant in the context of Mosaic research.

B. An Overview of Experimental Results

The average amount of time spent by the subjects in the recall and recognition tests was about two hours. Some interesting results regarding terms consistency were derived from the experiment.

- *The knowledge bases generated more terms than the subjects in the recall test. However, the subjects more than doubled their term selections in the recognition test:* Table IV shows the number of terms that were generated by the subjects (senior researchers: $S1$ and $S2$; junior researchers: $S3$ and $S4$) and the two knowledge bases in the recall and recognition tests. The subjects generated between 2.3 and 5.3 terms for each test descriptor in the recall test. Senior researchers were able to recall more terms than the junior researchers. The two knowledge bases, on the other hand, generated 9.52 and 10.44 related terms, respectively. Evidently, in the recall test the subjects were only able to activate part of their long-term memory. The two knowledge bases, which can be perceived as the group's "organizational memory," activated more associations (links) than the subjects. In the recognition test, the subjects were able to activate more long-term memory with the help of the clues suggested to them. The average number of related concepts they selected ranged between 7.42 and 18.74. Again, the senior subjects recognized more terms than the junior subjects. The improvement in terms generated from the recall test to the recognition test was between 196% and 353%. A significant portion of the terms selected were from the Cosine and Cluster KB's (discussed below).
- *Term consistencies among subjects and between subjects and KB's were low:* As shown in Table V, The consistency levels of terms selected by different subjects were between 8% and 10%. This is not surprising, however, considering the general characteristic of humans to associate different concepts with the same semantic object based on their own knowledge and experience. As prior research in human-computer interaction has indicated [15], the probability of two persons using the same term in describing the same thing is less than 20%. This fundamental property of language also has been observed in information science research. Prior indexer consistency research [17], [19], [21] has shown that the probability of two indexers' assigning the same descriptors to the same document is between 10% and 20%.
- *The Cluster KB had higher term consistency with the*

TABLE IV
TERMS GENERATED BY SUBJECTS AND KBS

Terms generated	Subjects					KB's		
	S1	S2	S3	S4	Total	Cosine	Cluster	Combined
Recall Test:								
No. of unique terms	265	237	241	115	719	476	522	660
Average no. of terms	5.30	4.74	4.82	2.30	14.38	9.52	10.44	13.20
Recognition Test:								
No. of unique terms	937	572	474	371	2051	—	—	—
Average no. of terms	18.74	11.44	9.48	7.42	41.02	—	—	—
Recognition/Recall	353%	241%	196%	322%	285%	—	—	—

TABLE V
TERM CONSISTENCY

Subject versus Subject (%)	S1 versus S2	S3 versus S4	Senior versus Junior
Consistency	8.2	10.3	9.2
Subject versus KB (%)	Senior	Junior	Any
Cluster	7.9	6.4	8.6
Cosine	4.7	3.2	4.9
Combined	7.3	5.6	8.1

TABLE VI
% OF KB TERMS SELECTED BY SUBJECTS

% of KB terms selected	Senior	Junior	All
Cluster	67	40	71%
Cosine	69	42	73%
Combined	67	39	71%

subjects than the Cosine KB: In comparing the two knowledge bases (see Table V), we observed that the Cluster KB appeared to have better term consistency with subjects (7.9% and 6.4%) than the Cosine KB (4.7% and 3.2%). Term consistency for both knowledge bases also more closely matched that of senior subjects (7.9% and 4.7%) than that of junior subjects (6.4% and 3.2%).

- *Knowledge base terms were selected overwhelmingly by the subjects in the recognition test:* As Table VI indicates, nearly 70% of the terms generated by the two knowledge bases were selected by the senior subjects as relevant in the recognition test and 40% of the KB terms were chosen by the junior researchers.
- *The knowledge bases contributed about 30% of the terms selected by the subjects in the recognition test:* We traced the sources of terms selected by the subjects in the recognition test. As shown in Table VII, the average percentages of terms contributed by the subjects themselves (from the recall test), by other subjects, from the Cluster algorithm, from the Cosine algorithm, and from the noise set were 29.5%, 52.5%, 31.5%, 30.5%, and 2.5% respectively. The subjects' expertise was confirmed somewhat by the fact that very few noise terms were actually selected (2.5%). The subjects' own terms (generated in the recall test) were only about 30% of the terms selected in the recognition test. Over 50% of the terms selected in

TABLE VII
SOURCES OF RECOGNIZED TERMS IN THE RECOGNITION TEST

Source of Terms (%)	S1	S2	S3	S4	Average
Self	23	35	41	19	29.5%
Other Subjects	42	68	46	54	52.5%
Cluster	34	28	26	38	31.5%
Cosine	32	28	24	38	30.5%
Noise	5	1	2	2	2.5%

the recognition test were generated by the other three subjects. This result indicated the value of collective knowledge in performing research. The Cluster and Cosine KB's each contributed about 30% of the terms. They were able to suggest a large number of important East-bloc computing related concepts, organizations, and persons.

V. HYPOTHESIS TESTING: CONCEPT RECALL AND CONCEPT PRECISION

While the previous section presents an overview of the experimental results, we describe in this section the detailed results from our statistical hypothesis testing, comparing the concept recall and concept precision ratios of the four subjects and our two knowledge bases.

A. Research Variables

We first present the research variables:

- *Performance measures – Concept Recall and Concept Precision:* Two conventional performance measures for information retrieval systems were adopted and modified in our evaluation: Recall and Precision. In the information retrieval studies, Recall and Precision are typically determined based on the number of relevant citations retrieved by the searchers, the number of citations retrieved by the searchers, and the number of relevant citations in the database [5], [44].

We modified the operational definitions of these two measures based on the unique characteristics of our research setting. Instead of examining the Recall and Precision of citations retrieved, we compared the Recall and Precision of relevant, associated *concepts* generated by the subjects and our knowledge bases for a source concept. To avoid confusion, we referred to our performance measures as

Concept Recall and Concept Precision. Concept Recall is thus defined as the portion of relevant concepts (i.e., concepts relevant to a source concept) identified by a subject or a knowledge base and Concept Precision is defined as the portion of identified concepts that are found to be relevant to the source concept. They are computed as follows:

$$\text{Concept Recall} = \frac{\text{Number of Retrieved Relevant Concepts}}{\text{Number of Total Relevant Concepts}}$$

$$\text{Concept Precision} = \frac{\text{Number of Retrieved Relevant Concepts}}{\text{Number of Total Retrieved Concepts}}$$

Our two-stage concept association experiment allowed us to generate all three numbers for the Concept Recall and Concept Precision computations. The concepts identified in the recognition test by the two senior experts (S_1 and S_2), who had 10 and six years East-bloc computing research experience, respectively, were used as the set of *Total Relevant Concepts*—they were requested to “recall as many concepts that related to the test descriptors as they possibly could.” *Total Relevant Concepts* indicated all the relevant concepts that were generated by these experts from their long-term memory in the information retrieval context. Clearly, we were unable to guarantee the completeness of this set of *Total Relevant Concepts* (for example, some other researchers may come up with other relevant, associated concepts). However, we believe under the constraint of our experimental setting, this measure of *Total Relevant Concepts* was a good approximation because of the subjects’ extensive domain knowledge and the nature of the recognition test. The *Total Retrieved Concepts* for the subjects and the knowledge bases can be determined by the set of concepts they generated in the recall test, i.e., concepts generated without any external clues or assistance. The set of *Retrieved Relevant Concepts* for the subjects and the knowledge bases can then be determined by the intersection of the *Total Relevant Concepts* and the *Total Retrieved Concepts*.

- **Subjects and knowledge bases:** We were interested in comparing the performances of our two knowledge bases, the Cluster KB and the Cosine KB, with those of the four human subjects. We also hoped to be able to contrast the performance of the two knowledge bases and to identify the effects of combining terms generated from the two knowledge bases, by ANDing (intersection, e.g., CLU-N-COS) and ORing (union, e.g., CLU-O-COS) their terms. Hypotheses 1 and 2 (H_1 and H_2), described later, are for this purpose. And lastly, we aimed at determining the performance improvement of the subjects when the assistance of the knowledge bases was made available to them (H_3H_{10}).

The specific hypotheses we tested are listed below. In these formally stated hypotheses, S_1 , S_2 , S_3 , and S_4 indicate the four subjects, respectively; CLUSTER and COSINE represent

the two KB’s; CLU-N-COS represents the intersection (ANDing) of the Cluster and the Cosine terms; and CLU-O-COS represents the union (ORing) of the Cluster and the Cosine terms. One-way analysis of variance for multiple groups and two-sample t -test were used to compare the Concept Recall and Concept Precision differences.

Hypotheses 1 and 2 (H_1 and H_2) compare the Concept Recall and Concept Precision of the subjects and the knowledge bases, respectively. The subhypotheses ($H_{1.1}$ – $H_{1.13}$ and $H_{2.1}$ – $H_{2.13}$) consist of pairwise two-sample mean comparisons of individual subject and knowledge base. Hypotheses 3 and 4 compare the Concept Recall and Concept Precision of S_1 and S_1 with the assistance of the various knowledge bases. Our goal was to determine whether the various knowledge sources can improve the Concept Recall and Concept Precision of Subject 1. Four subhypotheses ($H_{3.1}$ – $H_{3.4}$) are listed for these comparisons. Similarly, Hypotheses 5 and 6 are for S_2 ; Hypotheses 7 and 8 are for S_3 ; and Hypotheses 9 and 10 are for S_4 .

- 1) H_1 : $\text{Recall}_{S_1} = \text{Recall}_{S_2} = \text{Recall}_{S_3} = \text{Recall}_{S_4} = \text{Recall}_{\text{CLUSTER}} = \text{Recall}_{\text{COSINE}} = \text{Recall}_{\text{CLU-N-COS}} = \text{Recall}_{\text{CLU-O-COS}}$
 - a) $H_{1.1}$: $\text{Recall}_{\text{CLUSTER}} = \text{Recall}_{S_1}$
 - b) $H_{1.2}$: $\text{Recall}_{\text{CLUSTER}} = \text{Recall}_{S_2}$
 - c) $H_{1.3}$: $\text{Recall}_{\text{CLUSTER}} = \text{Recall}_{S_3}$
 - d) $H_{1.4}$: $\text{Recall}_{\text{CLUSTER}} = \text{Recall}_{S_4}$
 - e) $H_{1.5}$: $\text{Recall}_{\text{CLUSTER}} = \text{Recall}_{\text{COSINE}}$
 - f) $H_{1.6}$: $\text{Recall}_{\text{CLUSTER}} = \text{Recall}_{\text{CLU-N-COS}}$
 - g) $H_{1.7}$: $\text{Recall}_{\text{CLUSTER}} = \text{Recall}_{\text{CLU-O-COS}}$
 - h) $H_{1.8}$: $\text{Recall}_{\text{COSINE}} = \text{Recall}_{S_1}$
 - i) $H_{1.9}$: $\text{Recall}_{\text{COSINE}} = \text{Recall}_{S_2}$
 - j) $H_{1.10}$: $\text{Recall}_{\text{COSINE}} = \text{Recall}_{S_3}$
 - k) $H_{1.11}$: $\text{Recall}_{\text{COSINE}} = \text{Recall}_{S_4}$
 - l) $H_{1.12}$: $\text{Recall}_{\text{COSINE}} = \text{Recall}_{\text{CLU-N-COS}}$
 - m) $H_{1.13}$: $\text{Recall}_{\text{COSINE}} = \text{Recall}_{\text{CLU-O-COS}}$
- 2) H_2 : $\text{Precision}_{S_1} = \text{Precision}_{S_2} = \text{Precision}_{S_3} = \text{Precision}_{S_4} = \text{Precision}_{\text{CLUSTER}} = \text{Precision}_{\text{COSINE}} = \text{Precision}_{\text{CLU-N-COS}} = \text{Precision}_{\text{CLU-O-COS}}$
 - a) $H_{2.1}$: $\text{Precision}_{\text{CLUSTER}} = \text{Precision}_{S_1}$
 - b) $H_{2.2}$: $\text{Precision}_{\text{CLUSTER}} = \text{Precision}_{S_2}$
 - c) $H_{2.3}$: $\text{Precision}_{\text{CLUSTER}} = \text{Precision}_{S_3}$
 - d) $H_{2.4}$: $\text{Precision}_{\text{CLUSTER}} = \text{Precision}_{S_4}$
 - e) $H_{2.5}$: $\text{Precision}_{\text{CLUSTER}} = \text{Precision}_{\text{COSINE}}$
 - f) $H_{2.6}$: $\text{Precision}_{\text{CLUSTER}} = \text{Precision}_{\text{CLU-N-COS}}$
 - g) $H_{2.7}$: $\text{Precision}_{\text{CLUSTER}} = \text{Precision}_{\text{CLU-O-COS}}$
 - h) $H_{2.8}$: $\text{Precision}_{\text{COSINE}} = \text{Precision}_{S_1}$
 - i) $H_{2.9}$: $\text{Precision}_{\text{COSINE}} = \text{Precision}_{S_2}$
 - j) $H_{2.10}$: $\text{Precision}_{\text{COSINE}} = \text{Precision}_{S_3}$
 - k) $H_{2.11}$: $\text{Precision}_{\text{COSINE}} = \text{Precision}_{S_4}$

- l) $H2.12: Precision_{COSINE} = Precision_{CLU-N-COS}$
- m) $H2.13: Precision_{COSINE} = Precision_{CLU-O-COS}$
- 3) $H3: Recall_{S1} = Recall_{S1/CLUSTER} = Recall_{S1/COSINE} = Recall_{S1/CLU-N-COS} = Recall_{S1/CLU-O-COS}$
 - a) $H3.1: Recall_{S1} = Recall_{S1/CLUSTER}$
 - b) $H3.2: Recall_{S1} = Recall_{S1/COSINE}$
 - c) $H3.3: Recall_{S1} = Recall_{S1/CLU-N-COS}$
 - d) $H3.4: Recall_{S1} = Recall_{S1/CLU-O-COS}$
- 4) $H4: Precision_{S1} = Precision_{S1/CLUSTER} = Precision_{S1/COSINE} = Precision_{S1/CLU-N-COS} = Precision_{S1/CLU-O-COS}$
 - a) $H4.1: Precision_{S1} = Precision_{S1/CLUSTER}$
 - b) $H4.2: Precision_{S1} = Precision_{S1/COSINE}$
 - c) $H4.3: Precision_{S1} = Precision_{S1/CLU-N-COS}$
 - d) $H4.4: Precision_{S1} = Precision_{S1/CLU-O-COS}$
- 5) $H5: Recall_{S2} = Recall_{S2/CLUSTER} = Recall_{S2/COSINE} = Recall_{S2/CLU-N-COS} = Recall_{S2/CLU-O-COS}$
 - a) $H5.1: Recall_{S2} = Recall_{S2/CLUSTER}$
 - b) $H5.2: Recall_{S2} = Recall_{S2/COSINE}$
 - c) $H5.3: Recall_{S2} = Recall_{S2/CLU-N-COS}$
 - d) $H5.4: Recall_{S2} = Recall_{S2/CLU-O-COS}$
- 6) $H6: Precision_{S2} = Precision_{S2/CLUSTER} = Precision_{S2/COSINE} = Precision_{S2/CLU-N-COS} = Precision_{S2/CLU-O-COS}$
 - a) $H6.1: Precision_{S2} = Precision_{S2/CLUSTER}$
 - b) $H6.2: Precision_{S2} = Precision_{S2/COSINE}$
 - c) $H6.3: Precision_{S2} = Precision_{S2/CLU-N-COS}$
 - d) $H6.4: Precision_{S2} = Precision_{S2/CLU-O-COS}$
- 7) $H7: Recall_{S3} = Recall_{S3/CLUSTER} = Recall_{S3/COSINE} = Recall_{S3/CLU-N-COS} = Recall_{S3/CLU-O-COS}$
 - a) $H7.1: Recall_{S3} = Recall_{S3/CLUSTER}$
 - b) $H7.2: Recall_{S3} = Recall_{S3/COSINE}$
 - c) $H7.3: Recall_{S3} = Recall_{S3/CLU-N-COS}$
 - d) $H7.4: Recall_{S3} = Recall_{S3/CLU-O-COS}$
- 8) $H8: Precision_{S3} = Precision_{S3/CLUSTER} = Precision_{S3/COSINE} = Precision_{S3/CLU-N-COS} = Precision_{S3/CLU-O-COS}$
 - a) $H8.1: Precision_{S3} = Precision_{S3/CLUSTER}$
 - b) $H8.2: Precision_{S3} = Precision_{S3/COSINE}$
 - c) $H8.3: Precision_{S3} = Precision_{S3/CLU-N-COS}$
 - d) $H8.4: Precision_{S3} = Precision_{S3/CLU-O-COS}$

- 9) $H9: Recall_{S4} = Recall_{S4/CLUSTER} = Recall_{S4/COSINE} = Recall_{S4/CLU-N-COS} = Recall_{S4/CLU-O-COS}$
 - a) $H9.1: Recall_{S4} = Recall_{S4/CLUSTER}$
 - b) $H9.2: Recall_{S4} = Recall_{S4/COSINE}$
 - c) $H9.3: Recall_{S4} = Recall_{S4/CLU-N-COS}$
 - d) $H9.4: Recall_{S4} = Recall_{S4/CLU-O-COS}$
- 10) $H10: Precision_{S4} = Precision_{S4/CLUSTER} = Precision_{S4/COSINE} = Precision_{S4/CLU-N-COS} = Precision_{S4/CLU-O-COS}$
 - a) $H10.1: Precision_{S4} = Precision_{S4/CLUSTER}$
 - b) $H10.2: Precision_{S4} = Precision_{S4/COSINE}$
 - c) $H10.3: Precision_{S4} = Precision_{S4/CLU-N-COS}$
 - d) $H10.4: Precision_{S4} = Precision_{S4/CLU-O-COS}$

C. Results from Hypothesis Testing

We performed one-way analysis of variance (ANOVA³) for the 10 major hypotheses (H1-H10) and two-sample *t*-test (TTEST⁴) for each subhypothesis using the MINITAB statistical analysis package [41]. The statistical results listed under each hypothesis are outputs from the MINITAB analysis. The results are summarized below.

- *Hypothesis 1 (Concept Recall)*: The null hypothesis of all means being equal was rejected (with significance level $p = 0.000$), as shown in the analysis in Table VIII (STDEV: Standard Deviation, CI: Confidence Interval, P: Significance Level). The Cluster KB (Concept Recall: 0.3553) had significantly better Concept Recall than any of the subjects (0.1260–0.2726) and the Cosine KB (0.2225). Even senior Experts 1 and 2 were out-performed (in terms of Concept Recall: 0.2530 and 0.2726, respectively) by our asymmetric Cluster KB. The asymmetric property of the Cluster KB appeared to contribute to the performance, especially in contrast to the symmetric Cosine KB (both based on similar term co-occurrence computation). The intersection (CLU-N-COS) and union (CLU-O-COS) of the two KB's did not improve performance over the Cluster KB. The Cosine KB, on the other hand, only performed better than the most junior subject (S4). Its Concept Recall performance was as good as those of the other three subjects, however. See Table VIII.
- *Hypothesis 2 (Concept Precision)*: The null hypothesis of all means being equal was rejected (with $p = 0.000$). Both the Cluster and Cosine KB's had similar Concept Precision levels (0.6653 and 0.6216), which were worse than those of any of the subjects (0.8731–0.9452). Subjects' associations appeared to have less noise than the KBs'. See Table IX.
- *Hypotheses 3 and 4 (SI)*: The null hypotheses of all Concept Recall and Concept Precision means being equal

³One-way analysis of variance provides a test to compare data from different populations and to see if these populations have different means.

⁴Two-sample *t*-test compares means of two populations.

TABLE VIII

LEVEL	N	MEAN	STDEV	INDIVIDUAL 95 PCT CI'S FOR MEAN BASED ON POOLED STDEV			
				-----	-----	-----	-----
S1	47	0.2530	0.1725	(-----)			
S2	47	0.2726	0.1611	(-----)			
S3	47	0.2193	0.1911	(-----)			
S4	47	0.1260	0.1223	(-----)			
CLUSTER	47	0.3553	0.2234	(-----)			
COSINE	47	0.2455	0.2134	(-----)			
CLU-N-COS	47	0.2225	0.1993	(-----)			
CLU-O-COS	47	0.3783	0.2310	(-----)			
POOLED STDEV = 0.1923				0.10	0.20	0.30	0.40

SIGNIFICANT TWO SAMPLE T-TEST RESULTS FOR RECALL:

- * CLUSTER > S1 ($p = 0.015$)
- * CLUSTER > S2 ($p = 0.043$)
- * CLUSTER > S3 ($p = 0.0021$)
- * CLUSTER > S4 ($p = 0.0000$)
- * CLUSTER > COSINE ($p = 0.017$)
- * CLUSTER > CLU-N-COS ($p = 0.0031$)
- * COSINE > S4 ($p = 0.0014$)
- * COSINE < CLU-O-COS ($p = 0.0047$)

TABLE IX

LEVEL	N	MEAN	STDEV	INDIVIDUAL 95 PCT CI'S FOR MEAN BASED ON POOLED STDEV			
				-----	-----	-----	-----
S1	44	0.8891	0.2397	(-----)			
S2	43	0.9452	0.0984	(-----)			
S3	39	0.8875	0.2006	(-----)			
S4	39	0.8731	0.2172	(-----)			
CLUSTER	50	0.6653	0.3047	(-----)			
COSINE	50	0.6162	0.3651	(-----)			
CLU-N-COS	50	0.6216	0.3699	(-----)			
CLU-O-COS	50	0.6609	0.3019	(-----)			
POOLED STDEV = 0.2823				0.60	0.75	0.90	1.05

SIGNIFICANT TWO SAMPLE T-TEST RESULTS FOR PRECISION:

- * CLUSTER < S1 ($p = 0.0001$)
- * CLUSTER < S2 ($p = 0.0000$)
- * CLUSTER < S3 ($p = 0.0001$)
- * CLUSTER < S4 ($p = 0.0003$)
- * COSINE < S1 ($p = 0.0000$)
- * COSINE < S2 ($p = 0.0000$)
- * COSINE < S3 ($p = 0.0000$)
- * COSINE < S4 ($p = 0.0001$)

for Subject 1 were both rejected (with $p = 0.000$). By providing Subject 1 with either of the KB's as a decision aid, the Concept Recall performance increased significantly as shown below (from 0.2530 to about 0.500, a 118% improvement). This was clearly due to the KB having helped the subject activate long-term memory. Even for a senior expert like S1, external aid for concepts articulation was very useful. However, the Concept Precision levels went from 88% down to about 70% after incorporating KB's. See Table X.

- *Hypotheses 5 and 6 (S2)*: Hypotheses 5 and 6 for Subject 2 were both rejected (with $p = 0.000$). Concept Recall improved significantly when KB's were available to S2 (119% improvement), while Concept Precision worsened after incorporating the KB's. See Table XI.
- *Hypotheses 7 and 8 (S3)*: Hypotheses 7 and 8 for Subject 3 were both rejected (with $p = 0.000$). For one junior researcher (S3), we again observed similar effects: Concept Recall improved significantly (138% improvement) and Concept Precision became worse when KB's were present. See Table XII.
- *Hypotheses 9 and 10 (S4)*: Hypotheses 9 and 10 for Subject 4 were both rejected (with $p = 0.000$). For the most junior subject (S4), again, Concept Recall improved and Concept Precision deteriorated after considering KB's. The Concept Recall improvement of 261% was significantly higher than the improvements obtained by other subjects. See Table XIII.

We summarize the significant results from our statistical analysis as follows:

- The Cluster KB appeared to be the best knowledge source among all the human subjects and KB's in terms of Concept Recall. There was no need to combine the output from the Cluster and Cosine KB's — the Concept Recall measure did not improve significantly enough to warrant the effort. On the other hand, human subjects consistently had better Concept Precision than either KB.
- Using the KB's as concepts articulation aids, subjects' Concept Recall performance improved significantly. Their Concept Precision performance went down, however. This finding confirmed the general belief that using a thesaurus improves recall [45].
- Lastly, the better Concept Recall performance of the Cluster KB compared with that of the Cosine KB revealed a very important fact about the appropriateness of the various term co-occurrence algorithms (similarity computations). Most of the similarity measures used in automatic text processing [45] such as the Cosine coefficient, Dice coefficient, inner product, and Jaccard coefficient produce symmetric property on the links. The Cluster coefficient, however, is asymmetric. From the results of our experiment, it was evident that the asymmetric property of term association not only matched with human cognition better in theory, but also fostered better recall performance in practice. We believe this result has important implications for designers of automatic thesauri.

TABLE X

ANALYSIS OF VARIANCE FOR RECALL

INDIVIDUAL 95 PCT CI'S FOR MEAN BASED ON POOLED STDEV			
LEVEL	N	MEAN	STDEV
S1	47	0.2530	0.1725
S1/CLUSTER	47	0.5536	0.2385
S1/COSINE	47	0.4777	0.2484
S1/CLU-N-COS	47	0.4555	0.2525
S1/CLU-O-COS	47	0.5758	0.2314
POOLED STDEV = 0.2305			

SIGNIFICANT TWO SAMPLE T-TEST RESULTS FOR RECALL:

* S1 < S1/CLUSTER, S1/COSINE, S1/CLU-N-COS, S1/CLU-O-COS (p = 0.000)

INDIVIDUAL 95 PCT CI'S FOR MEAN BASED ON POOLED STDEV			
LEVEL	N	MEAN	STDEV
S1	44	0.8891	0.2397
S1/CLUSTER	50	0.6973	0.2916
S1/COSINE	50	0.6931	0.3102
S1/CLU-N-COS	50	0.6962	0.3118
S1/CLU-O-COS	50	0.6952	0.2913
POOLED STDEV = 0.2913			

SIGNIFICANT TWO SAMPLE T-TEST RESULTS FOR PRECISION:

* S1 > S1/CLUSTER, S1/COSINE, S1/CLU-N-COS, S1/CLU-O-COS
(p = 0.0007, 0.0009, 0.0006, 0.0011)

After examining our experiment carefully, we found that our KBs' Concept Recall levels could be increased if we were to present their associated concepts differently. In our evaluation, we presented only the neighboring descriptors from the knowledge bases, i.e., descriptors that were one link away from the source concept. Many related concepts, however, may be two links or even three or four links away from a source concept — in human memory, concept association often involves multiple-link activation. Many of the descriptors identified by the subjects in the recognition test may have existed in the knowledge bases, but a few links away from the source concept. We believe that the activation of multiple levels of descriptors in the knowledge bases could make the knowledge bases perform even better. We discuss our proposed methods for performing multiple-link, semantic activation in the next section.

Another factor that was strongly related to the Concept Recall and Concept Precision was the 5% similarity threshold we adopted when we generated our knowledge bases. Very likely, by choosing a higher-valued cutoff, we could have improved the Concept Precision of our knowledge bases, but possibly deteriorated the Concept Recall. Similarly, we could have improved the Concept Recall but deteriorated

TABLE XI

INDIVIDUAL 95 PCT CI'S FOR MEAN BASED ON POOLED STDEV			
LEVEL	N	MEAN	STDEV
S2	47	0.2726	0.1611
S2/CLUSTER	47	0.5969	0.2136
S2/COSINE	47	0.4963	0.2071
S2/CLU-N-COS	47	0.4739	0.2128
S2/CLU-O-COS	47	0.6193	0.2032
POOLED STDEV = 0.2005			

SIGNIFICANT TWO SAMPLE T-TEST RESULTS FOR RECALL:

* S2 < S2/CLUSTER, S2/COSINE, S2/CLU-N-COS, S2/CLU-O-COS (p = 0.000)

INDIVIDUAL 95 PCT CI'S FOR MEAN BASED ON POOLED STDEV			
LEVEL	N	MEAN	STDEV
S2	43	0.9452	0.0984
S2/CLUSTER	50	0.7342	0.2636
S2/COSINE	50	0.7415	0.2736
S2/CLU-N-COS	50	0.7451	0.2755
S2/CLU-O-COS	50	0.7314	0.2629
POOLED STDEV = 0.2475			

SIGNIFICANT TWO SAMPLE T-TEST RESULTS FOR PRECISION:

* S2 < S2/CLUSTER, S2/COSINE, S2/CLU-N-COS, S2/CLU-O-COS
(p = 0.0000, 0.0000, 0.0000, 0.0000)

the Concept Precision by choosing a lower-valued cutoff. In our research we chose the threshold based on the explosion factor for the number of associated terms generated during knowledge base creation. In our knowledge bases most terms appeared to have neighbors that expanded gradually until the 5% threshold was reached. System designers need to be aware of the effect of threshold selection on the Concept Recall and Concept Precision of a knowledge base.

Comparing the various measures, it was evident that the Cluster KB was superior to the Cosine KB. The Cluster KB had about 35% Concept Recall and 65% Concept Precision based on our evaluation and it out-performed human experts in recalling East-bloc computing related concepts. Incorporating it into retrieval systems can significantly improve the Recall of searchers in associating query concepts. We re-ran the Cluster algorithm on the complete Mosaic database after the experiment, including all folder descriptors. The resulting knowledge base has grown in size — 20 000 descriptors and 280 000 relationships.

VI. SEMANTICS-BASED INFORMATION MANAGEMENT AND RETRIEVAL

We have successfully integrated our knowledge base with the Mosaic database as a thesaurus-like component that can

TABLE XII

INDIVIDUAL 95 PCT CI'S FOR MEAN BASED ON POOLED STDEV				
LEVEL	N	MEAN	STDEV	
S3	47	0.2193	0.1911	(---*---)
S3/CLUSTER	47	0.5216	0.1927	(---*---)
S3/COSINE	47	0.4422	0.2128	(---*---)
S3/CLU-N-COS	47	0.4192	0.2096	(---*---)
S3/CLU-O-COS	47	0.5446	0.1935	(---*---)

POOLED STDEV =	0.2002		0.24	0.36 0.48

SIGNIFICANT TWO SAMPLE T-TEST RESULTS FOR RECALL:

* S3 < S3/CLUSTER, S3/COSINE, S3/CLU-N-COS, S3/CLU-O-COS (p = 0.000)

INDIVIDUAL 95 PCT CI'S FOR MEAN BASED ON POOLED STDEV				
LEVEL	N	MEAN	STDEV	
S3	39	0.8875	0.2006	(-----*-----)
S3/CLUSTER	50	0.6929	0.2890	(-----*-----)
S3/COSINE	50	0.6861	0.3059	(-----*-----)
S3/CLU-N-COS	50	0.6905	0.3066	(-----*-----)
S3/CLU-O-COS	50	0.6901	0.2892	(-----*-----)

POOLED STDEV =	0.2843		0.72	0.84 0.96

SIGNIFICANT TWO SAMPLE T-TEST RESULTS FOR PRECISION:

* S3 < S3/CLUSTER, S3/COSINE, S3/CLU-N-COS, S3/CLU-O-COS
(p = 0.0003, 0.0003, 0.0003, 0.0005)

TABLE XIII

INDIVIDUAL 95 PCT CI'S FOR MEAN BASED ON POOLED STDEV				
LEVEL	N	MEAN	STDEV	
S4	47	0.1260	0.1223	(---*---)
S4/CLUSTER	47	0.4544	0.2024	(---*---)
S4/COSINE	47	0.3544	0.1915	(---*---)
S4/CLU-N-COS	47	0.3317	0.1873	(---*---)
S4/CLU-O-COS	47	0.4771	0.2016	(---*---)

POOLED STDEV =	0.1835		0.15	0.30 0.45 0.60

SIGNIFICANT TWO SAMPLE T-TEST RESULTS FOR RECALL:

* S4 < S4/CLUSTER, S4/COSINE, S4/CLU-N-COS, S4/CLU-O-COS (p = 0.000)

INDIVIDUAL 95 PCT CI'S FOR MEAN BASED ON POOLED STDEV				
LEVEL	N	MEAN	STDEV	
S4	39	0.8731	0.2172	(-----*-----)
S4/CLUSTER	50	0.6801	0.2801	(-----*-----)
S4/COSINE	50	0.6642	0.3008	(-----*-----)
S4/CLU-N-COS	50	0.6655	0.3009	(-----*-----)
S4/CLU-O-COS	50	0.6788	0.2808	(-----*-----)

POOLED STDEV =	0.2802		0.60	0.72 0.84 0.96

SIGNIFICANT TWO SAMPLE T-TEST RESULTS FOR PRECISION:

* S4 < S4/CLUSTER, S4/COSINE, S4/CLU-N-COS, S4/CLU-O-COS
(p = 0.0004, 0.0003, 0.0004, 0.0003)

be activated by users during the online information management or information retrieval process, helping users identify semantically-relevant concepts. We believe this component can bring out the complementary strengths of the human (high precision) and the computer model (better thoroughness) in associating concepts. In this section, we first describe the information retrieval system, the Arizona Analyst Information System (AAIS), that is built upon the Mosaic database. We then discuss the semantics-based information management and information retrieval interfaces that we implemented in the AAIS after incorporating the knowledge base.

A. The Arizona Analyst Information System

The existing system that interfaces the Mosaic database is the Arizona Analyst Information System (AAIS) [25]. The AAIS provides support for various research activities, from data entry, information management and information sharing to information retrieval, communication and document preparation. Research activities are supported by both the AAIS and the VAX/VMS and INGRES system environment in which it resides. The system features are outlined in the following.

- 1) *Data Entry and Information Management:* A template-driven data entry interface enables the user to enter

information from any source. Researchers often use the data entry facility in a note pad fashion as they are analyzing source information. Templates may also be constructed off-line and submitted in a batch mode to be added to the database. Bibliographic information is entered on a field-by-field basis, as are administrative and index information. During the data entry process, the AAIS assists in the generation and selection of indexes by allowing researchers to search and browse archived documents and their associated indexes.

- 2) *Information Retrieval:* An integrated document retrieval utility using a custom-built query processor [24] enhances the retrieval flexibility and friendliness of the system. The AAIS retrieval interface allows researchers to combine indexes to retrieve documents and browse the folder hierarchy. For instance, a researcher could query the system based on the keyword MICROCOMPUTER, the country HUNGARY, and the organization IBM. This query would retrieve all of the documents that have at least one selected index associated with them. The document retrieval utility supports both "AND" and "OR" Boolean operators.
- 3) *Information Sharing and Communication:* Information sharing is central to research collaboration, and is sup-

ported by the AAIS folder assignment process. The AAIS provides special folders for each analyst and each ongoing research topic, called "hot" folders. Analysts route relevant documents to these folders. Analysts who are responsible for these "hot" folders, often the senior analysts, review the documents sent to these special folders and assign them to the most appropriate folders in the folder hierarchy. Since recent information may have significant implications in different contexts and especially for ongoing research, this enables analysts to share information, make good folder assignments, and maintain close monitoring of events.

Information retrieval, sharing, and communication also occur outside the AAIS. The underlying relational database management system, INGRES, provides an environment where relational query languages, QUEL⁵ and Query-By-Forms⁶ can be used. The AAIS allows researchers to formulate complex queries using these query languages. The electronic mail system provided by the VAX/VMS operating system also is heavily used by all group members for communication.

- 4) *Document Preparation*: Document preparation is facilitated by the same retrieval interface used to retrieve relevant documents on-line. Document editing capability is supported by the VAX/VMS editor. Bibliographies can be produced automatically. Finished documents are archived back into the AAIS database, providing an analysis trail for future researchers.

The AAIS system, its underlying INGRES database management system, and the VAX/VMS operating system, provide an environment that supports collaborative research activities.

B. Semantics-Based Information Management

The knowledge base we created was incorporated into the AAIS environment as an external knowledge source, called upon by the Mosaic researchers when necessary. During the information management process, the knowledge base component acted as a "semantic gate keeper," parsing documents that were entered by the analysts and checking the semantic completeness of indexes assigned. The information management process in the current AAIS goes as follows:

- 1) *Document Entry and Index Assignment*: Analysts first invoke the system's data entry module and complete the bibliographical information fields associated with a document such as title, author, publisher, and publication date. They then enter the abstract or the important content from the source document. After entering a document, analysts index it by supplying appropriate keywords, organizations, persons, countries, and folders. An example of a complete document entry appears in Fig. 1.

⁵QUEL Language is a data manipulation language like SQL. QUEL is based on tuple relational calculus. It is nonprocedural and allows the user to process data without concern for physical data structures.

⁶QBF allows the user to issue queries against an INGRES database without requiring any knowledge of QUEL.

- 2) *Syntactic Parsing*: The system automatically performs syntactic checking before accepting document entry into the Mosaic database. The system maintains a dictionary of common vocabularies and other previously entered words that are unique in the East-bloc computing environment, e.g., BESM-6 (Russian computer), Andropov (person), ITMVT (organization), etc. Unknown words in the document are flagged. The analysts have the option of correcting or accepting the flagged words.
- 3) *Semantic Parsing*: As the final stage of the information management and indexing process, the system uses the indexes in the document to activate the system's knowledge base. A thesaurus look-up process is then performed for all the starting indexes. Relevant terms in the knowledge base are activated and ranked automatically (simple summation of all weights associated with the activated terms). The associated terms, which include keywords, persons, organizations, countries, and folders, are presented to the analysts in decreasing order of relevance in a scrollable menu (similar to the menu in Fig. 3). The analysts can then select among these suggested terms and the system appends these selections to the document's indexes automatically.

The semantic parsing stage, in particular, helps ensure the completeness in indexing. As our evaluation and other cognitive psychology-based memory experiments suggested [1], people are better at recognizing clues than recalling learned concepts. With the system providing strongly associated indexes online to the analysts, we believe we can create a more effective information management environment.

C. Semantics-Based Information Retrieval

Our semantics-based information retrieval interface allows extensive query articulation and concept exploration during search. The information retrieval process is presented in the following.

- 1) *Query formation*: The AAIS provides two modes for information retrieval. Analysts can retrieve documents by using the AAIS's user-friendly, menu-based retrieval module or by using the AAIS's underlying QUEL relational query language. In both modes, Boolean operators can be used to combine query terms.
- 2) *Terms matching and query refinement*: Our system then proceeds to assist *terms matching* and *query refinement* [7]. As prior research has indicated, query terms used by searchers are often different from the index terms (*terms matching* problem). Bates [3] argues that for a successful match, the searcher must to some extent generate as much "variety" (in the cybernetic sense, as defined by Ashby [2]) in the search as is produced by the indexers in their indexing. In terms of *query refinement*, users often do not have queries," but what Belkin calls an "anomalous state of knowledge" [4]. Users often expect to refine this anomalous state of knowledge into a query through an interactive process. Our knowledge base provides the means to support both *terms matching* and *query refinement*.

Fig. 3. Thesaurus terms related to EL'BRUS.

Fig. 4. A retrieved document.

Searchers' initial query statements are taken as the "triggers" to identify other semantically-relevant indexes. The system uses the query terms to consult the thesaurus, activate relevant indexes, and rank them in order. The thesaurus provides explicit semantic linkages between query terms and index terms. Since our thesaurus captures almost all terms and indexes used by the Mosaic researchers in the past decade, the thesaurus consultation process will be able to assist *terms matching*.

Terms suggested by the system also serve as clues to help searchers articulate their needs. Searchers can use the thesaurus component iteratively — select relevant terms, activate thesaurus terms by using the selected terms, make more selections, activate more thesaurus terms, and so on. During this human-system interaction cycle, the thesaurus becomes a concept exploration or concept convergence aid to the users, alleviating the cognitive demand on the users in refining their "anomalous state of knowledge."

Figure 3 shows the top ranked thesaurus terms suggested by the system with an initial query request

for "EL'BRUS" (a Soviet high performance computer). These terms were ranked in decreasing order of relevance and their object types were also indicated on the display, i.e., (F) for folder, (K) for keywords, (O) for organization, (N) for person name, and (C) for country. For example, "BESM-6" was the predecessor of the EL'BRUS system; "ITMVT" was the organization in which the machine was developed; and "Albert Nikolayevich Naumov" was the key scientist credited with developing the machine. The searchers could then select any terms they deemed relevant to their queries and perform more refinement using the selected terms. (We imposed a maximal selection of eight thesaurus terms during each iteration for computational reasons).

- 3) *Document retrieval, ranking, and selection:* When searchers feel comfortable with their articulated queries, they can activate the system's document retrieval module, which uses the final selected terms to search in the complete database. Each document retrieved is assigned a score by computing the number of matched indexes in the document. A summary table is then presented to the searchers, showing the number of documents matched with the different numbers of indexes. For example, for a request of five query terms, the system may determine that there are 34 documents that have all five terms as indexes, 234 documents with four matched terms, 550 documents with three matched terms, and so on.

Searchers have the option of browsing these retrieved documents in their decreasing order of relevance or jumping from one document to another. A sample retrieved document is shown in Fig. 4. The selections at the bottom of the menu in Fig. 4 indicate the options searchers have in retrieving document-related information (e.g., index, country, etc.) or performing document-specific operations (e.g., update, output, etc.). The searchers' eventual document selection is then directed to a file. Any terms determined relevant during the document retrieval process can be collected and used in the next round of concept exploration and semantic information retrieval. Our system is designed to foster a tight coupling between users and the knowledge base by providing rich semantic associations and mappings within a friendly and iterative search environment.

Semantics-based information management and information retrieval modules have been incorporated into the Mosaic research environment. In initial reports, some researchers indicated that the thesaurus had been an excellent tool for assisting query articulation. It helped reveal the explicit semantic relationships between objects of interest to them. In particular, it provided semantic interpretation (e.g., related keywords, persons, organizations, etc.) for previously obscure folders, which were created by different researchers at different times over the past decade. Some researchers even suggested using the system's knowledge about the folders to filter incoming documents and to make automatic, semantics-based folder assignment. Some junior researchers also used our knowledge base as a training tool, exploring and traversing the network

of knowledge and its underlying database in order to become familiar with topics of interest.

VII. CONCLUSION

"Knowledge discovery in databases" recently has attracted significant interest of researchers in databases and artificial intelligence. From the database perspective, this approach presents a unique opportunity to impose knowledge and structure on the unstructured, voluminous information that exists in large-scale databases. From the AI point of view, knowledge discovery combines AI machine learning techniques with statistical algorithms to address the knowledge acquisition problem. By using robust algorithms and taking advantage of unused computing cycles, knowledge discovery can help determine the patterns of behaviors, semantic relationships between objects of interest, and the underlying "knowledge" embedded in a database.

This paper reports findings from our ongoing research. We developed two knowledge bases automatically in a unique East-bloc computing domain from a large textual database. We performed a memory-association experiment, comparing the recall and precision of our knowledge bases and four East-bloc computing experts in associating concepts. One of the knowledge bases that exhibited an asymmetric link property out-performed all four human subjects in recalling relevant concepts in East-bloc computing. Using the knowledge bases as concept articulation aids, human subjects' concept recall performance improved significantly. Their concept precision performance became worse, however. The asymmetric knowledge base, which contained about 20 000 concepts and 28 000 weighted relationships, was incorporated as a thesaurus-like component into an intelligent retrieval system.

We posit that there are two important reasons for the robustness of our knowledge bases. First, our term co-occurrence computation was based on ALL documents in the database, instead of a sample collection of documents from the database (used in most automatic thesaurus generation research — due to computational concern in the most part). Our knowledge bases in essence "learned" from all the documents in the database, instead of a portion of it. Considering the variety existing in the subject area and indexing vocabularies, sampling techniques are clearly infeasible. The computational concern of using a complete database can be resolved by utilizing the abundant, unused CPU cycles available in many environments today. The initial automatic thesaurus generation occurs only once and the incremental thesaurus update (by incorporating knowledge from new documents) is not so CPU-intensive.

In addition, we used our thesaurus as a tool to perform user relevance feedback. As prior research indicated [45], use of relevance feedback to add new query terms extracted from previously retrieved relevant documents produces substantial advantages in query formation and retrieval effectiveness. Our thesaurus consolidated and ranked concepts and solicited searchers' choice of relevant concepts interactively. This process can be performed iteratively, assisting online terms matching and query refinement.

We had successfully created an interface for semantics-based information management and information retrieval. The system performs semantic checking during information management to ensure the completeness of document indexing. Users also collaborate closely with our system in performing semantics-based query refinement and relevance feedback. We are in the process of designing an incremental algorithm for learning, creating a meta knowledge structure in East-bloc computing, and developing inferencing algorithms for the knowledge base.

REFERENCES

- [1] J. R. Anderson, *Cognitive Psychology and Its Implications*, second ed. New York: Freeman, 1985.
- [2] W. R. Ashby, *An Introduction to Cybernetics*. London: Methuen, 1973.
- [3] M. J. Bates, "Subject access in online catalog: a design model," *J. Amer. Soc. Inform. Sci.*, vol. 37, no. 6, pp. 357–376, Nov. 1986.
- [4] N. J. Belkin, R. N. Oddy, and H. M. Brooks, "Ask for information retrieval: Part I. background and theory," *J. Documentation*, vol. 38, no. 2, pp. 61–71, June 1982.
- [5] D. C. Blair and M. E. Maron, "An evaluation of retrieval effectiveness for a full-text document-retrieval system," *Commun. ACM*, vol. 28, no. 3, pp. 289–299, 1985.
- [6] S. K. Card, T. P. Moran, and A. Newell, *The Psychology of Human Computer Interaction*. Hillsdale, NJ: Lawrence Erlbaum, 1983.
- [7] H. Chen and V. Dhar, "User misconceptions of online information retrieval systems," *Int. J. Man-Machine Studies*, vol. 32, no. 6, pp. 673–692, June 1990.
- [8] ———, "Cognitive process as a basis for intelligent retrieval systems design," *Information Processing and Management*, vol. 27, no. 5, pp. 405–432, 1991.
- [9] H. Chen, K. J. Lynch, A. K. Himler, and S. E. Goodman, "Information management in research collaboration," *Int. J. Man-Machine Studies*, vol. 36, no. 3, pp. 419–445, 1992.
- [10] P. R. Cohen and R. Kjeldsen, "Information retrieval by constrained spreading activation in semantic networks," *Inform. Processing and Management*, vol. 23, no. 4, pp. 255–268, 1987.
- [11] W. B. Croft and R. H. Thompson, "I³R: A new approach to the design of document retrieval systems," *J. Amer. Soc. Inform. Sci.*, vol. 38, no. 6, pp. 389–404, 1987.
- [12] C. J. Crouch, "An approach to the automatic construction of global thesauri," *Inform. Processing and Management*, vol. 26, no. 5, pp. 629–640, 1990.
- [13] E. A. Feigenbaum, "The art of artificial intelligence: Themes and case studies knowledge engineering," in *Int. Joint Conf. Artificial Intell.*, pp. 1014–1029, 1977.
- [14] E. A. Fox, "Development of the CODER system: A testbed for artificial intelligence methods in information retrieval," *Inform. Processing and Management*, vol. 23, no. 4, pp. 341–366, 1987.
- [15] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, "The vocabulary problem in human-system communication," *Commun. ACM*, vol. 30, no. 11, pp. 964–971, Nov. 1987.
- [16] F. Hayes-Roth, D. A. Waterman, and D. Lenat, *Building Expert Systems*. Reading, MA: Addison-Wesley, 1983.
- [17] R. S. Hooper, *Indexer Consistency Test: Origin, Measurement, Results and Utilization*. Bethesda, MD: IBM Corp., 1965.
- [18] B. L. Humphreys and D. A. Lindberg, "Building the unified medical language system," in *Proc. Thirteenth Annu. Symp. Comput. Appl. Med. Care*, Washington, DC, Nov. 5–8, 1989.
- [19] F. I. Hurwitz, "A study of indexer consistency," *Amer. Documentation*, vol. 20, pp. 92–94, Jan. 1969.
- [20] P. Jackson, *Introduction to Expert Systems*. Reading, MA: Addison-Wesley, 1990.
- [21] J. Jacoby and V. Slamecka, *Indexer Consistency under Minimal Conditions*. Bethesda, MD: Documentation, 1962.
- [22] D. A. Lindberg and B. L. Humphreys, "The UMLS knowledge sources: Tools for building better user interface," in *Proc. Fourteenth Annu. Symp. Computer Appl. Med. Care*, Washington, DC, Nov. 4–7 1990.
- [23] P. H. Lindsay and D. A. Norman, *Human Information Processing: An Introduction to Psychology*. San Diego, CA: Harcourt Brace Jovanovich, 1977.
- [24] K. J. Lynch and L. M. Hoopes, "An interface for rapid prototyping and evolutionary support of database-intensive applications," in *Proc. IEEE IPCCC*, pp. 344–348, Phoenix, AZ, Mar. 1989.

- [25] K. J. Lynch, J. M. Snyder, W. K. McHenry, and D. R. Vogel, "The Arizona Analyst Information System: Supporting collaborative research on international technological trends," in *Proc. IFIP WG8.4 Conf. Multi-user Interfaces and Applications*, Heraklion, Crete, Greece, Sept. 1990, pp. 159–174.
- [26] T. W. Malone, K. R. Grant, K. Lai, R. Rao, and D. Rosenblitt, "Semistructured messages are surprisingly useful for computer-supported coordination," *ACM Trans. Office Inform. Syst.*, 5(2):115–131, 1987.
- [27] B. K. Martin and R. Rada, "Building a relational data base for a physician document index," *Med. Inf.*, 12, no. 3, July–Sept. 1987.
- [28] A. T. McCray and W. T. Hole, "The scope and structure of the first version of the UMLS semantic network," in *Proc. Fourteenth Annu. Symp. Comput. Appl. Med. Care*, Washington, DC, Nov. 4–7 1990.
- [29] R. S. Michalski, "A theory and methodology of inductive learning," in *Machine Learning, An Artificial Intelligence Approach*, R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, Eds. Palo Alto, CA: Tioga, 1983, pp. 83–134.
- [30] R. S. Michalski and R. E. Stepp, "Learning from observation: Conceptual clustering," in *Machine Learning, An Artificial Intelligence Approach*, R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, Eds. Palo Alto, CA: Tioga, 1983, pp. 331–363.
- [31] I. Monarch and J. G. Carbonell, "CoalSORT: A knowledge-based interface," *IEEE Expert*, pp. 39–53, Spring 1987.
- [32] K. Parsaye, M. Chignell, S. Khoshafian, and H. Wong, *Intelligent Databases*. New York: Wiley, 1989.
- [33] G. Pietetsky-Shapiro, "Workshop on knowledge discovery in real databases," in *Int. Joint Conf. Artificial Intelligence*, 1989.
- [34] S. Pollitt, "Cansearch: An expert systems approach to document retrieval," *Inform. Processing and Management*, vol. 23, no. 2, pp. 119–138, 1987.
- [35] M. R. Quillian, "Semantic memory," in *Semantic Information Processing*, M. Minsky, Ed. Cambridge, MA: MIT Press, 1968.
- [36] J. R. Quinlan, "Learning efficient classification procedures and their application to chess end games," in *Machine Learning, An Artificial Intelligence Approach*, R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, Eds. Palo Alto, CA: Tioga, 1983, pp. 463–482.
- [37] ———, "Decision trees and decisionmaking," *IEEE Trans. Syst., Man, Cybern.*, vol. 20, pp. 339–346, Mar./Apr. 1990.
- [38] M. Quint, "Banks looking more closely at their credit card holders," *New York Times*, May 27, 1991, p. 1.
- [39] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," *IEEE Trans. Syst., Man, Cybern.*, vol. 19, pp. 17–30, Jan./Feb. 1989.
- [40] E. Rich and K. Knight, *Artificial Intelligence*, second ed. New York: McGraw-Hill, 1991.
- [41] B. F. Ryan, B. L. Joiner, and T. A. Ryan, *MINITAB Handbook*, second ed. Boston, MA: PWS-KENT Publishing, 1985.
- [42] G. Salton, "Automatic thesaurus construction for information retrieval," *Inform. Processing*, vol. 71. Amsterdam: North Holland, 1972, pp. 115–123.
- [43] ———, "Generation and search of clustered files," *ACM Trans. Database Syst.*, vol. 3, no. 4, pp. 321–346, Dec. 1978.
- [44] ———, "Another look at automatic text-retrieval systems," *Commun. ACM*, vol. 29, no. 7, pp. 648–656, 1986.
- [45] ———, *Automatic Text Processing*. Reading, MA: Addison-Wesley, 1989.
- [46] G. Salton and M. E. Lesk, "Information analysis and dictionary construction," in *The Smart Retrieval System—Experiments in Automatic Document Processing*, G. Salton, Ed. Englewood Cliffs, NJ: Prentice-Hall, 1971, pp. 115–142.
- [47] P. Shoval, "Principles, procedures and rules in an expert system for information retrieval," *Inform. Processing and Management*, vol. 21, no. 6, pp. 475–487, 1985.

Hsinchun Chen, for a photograph and biography, please see page 884 of this issue of the TRANSACTIONS.

Kevin J. Lynch received the Ph.D. degree in management information systems in 1989 from Karl Eller Graduate School of Business, University of Arizona, Tucson.

He is currently a Database Analyst for the University of Arizona. He has worked at GTE and Hughes Aircraft, and has database consulting experience in government and industry. He has published many articles concerning the design and implementation of collaborative research systems. His research centers around knowledge-based information management and retrieval in collaborative systems.