Semantic Retrieval for the NCSA Mosaic

Hsinchun Chen, Assistant Professor, University of Arizona, MIS Department
Bruce R. Schatz, Research Scientist, University of Illinois, NCSA

Abstract

In this paper we report an automatic and scalable concept space approach to enhancing the deep searching capability of the NCSA Mosaic. The research, which is based on the findings from a previous NSF National Collaboratory project and which will be expanded in a new Illinois NSF/ARPA/NASA Digital Library project, centers around semantic retrieval and user customization. Semantic retrieval supports a higher level of abstraction in user search, which can overcome the vocabulary problem for information retrieval. Rather than searching for words within the object space, the search is for terms within a concept space (graph of terms occurring within objects linked to each other by the frequency with which they occur together). Co-occurrence graphs seem to provide good suggestive power in specialized domains, such as biology. By providing a more understandable, system-generated, semantics-rich concept space as an abstraction of the enormously complex object space plus algorithms and interface to assist in object/concept spaces traversal, we believe we can greatly alleviate both information overload and the vocabulary problem of internet services. These techniques will also be used to provide a form of customized retrieval and automatic information routing. Results from past research, the specific algorithms and techniques, and the research plan for enhancing the NCSA Mosaic's search capability in the NSF/ARPA/NASA Digital Library project will be discussed.

Introduction

Despite the usefulness of database technologies, users of online information retrieval systems are often overwhelmed by the amount of current information, the subject and system knowledge required to access this information, and the constant influx of new information. The result is termed ``information overload.'' A second difficulty associated with information retrieval and information sharing is the classical ``vocabulary problem,'' which is a consequence of diversity of expertise and backgrounds of system users. Previous research in information science and in human-computer interactions has shown that people tend to use different terms (vocabularies) to describe a similar concept - the chance of two people using the same term to describe an object or concept is less than 20%. The ``fluidity'' of concepts and vocabularies, especially in the scientific and engineering domains, further complicates the retrieval issue. A scientific or engineering concept may be perceived differently by different researchers and it may also convey different meanings at different times. To address the ``information overload'' and the ``vocabulary problem'' in a large information space that is used by searchers of varying backgrounds, a more ``intelligent'' and proactive search aid is needed.

The problems of information overload and vocabulary difference have become more pressing with the emergence of the increasingly more popular internet resource discovery services. Retrieval difficulties, we believe, will worsen as the amount of online information increases in an accelerating pace under the National Information Infrastructure. Although network protocols and software such as Mosaic and WAIS support significantly easier importation of online information sources, their use is accompanied by the adverse problem of users not being able to explore and find what they want in an enormous information space.

The main information retrieval mechanisms provided by the prevailing resource discovery software and other information retrieval systems are either based on ``keyword search'' (inverted index or full text) or ``user browsing.'' Keyword search often causes low precision and poor recall due to the limitations of

controlled language based interfaces (the vocabulary problem) and the inability of searchers themselves to fully articulate their needs. Furthermore, browsing only allows users to explore a very small portion of a large and unfamiliar information space, which was constructed based in the first place on the system designer's view of the world. A large information space organized based on hypertext-like browsing can also potentially confuse and disorient its user, the ``embedded digression problem;'' and it can cause the user to spend a great deal of time while learning nothing specific, the ``art museum phenomenon.'' This research aims to provide a semantic, concept-based retrieval option that could supplement existing information retrieval options.

Our proposed approach is based on textual analysis of a large corpus of domain-specific documents in order to generate a large set of subject vocabularies. By adopting the cluster analysis techniques to analyze the co-occurrence probabilities of the subject vocabularies, a similarity (relevance) matrix of vocabularies can be built to represent the important concepts and their weighted ``relevance'' relationships in the subject domain. To create a network of concepts, which we refer to as the ``concept space'' for the subject domain (to distinguish it from its underlying ``information space''), we propose to develop general AI-based graph traversal algorithms (e.g., serial, optimal branch-and-bound search algorithms or parallel, Hopfield net like algorithms) and graph matching algorithms (for intersecting concept spaces in related domains) to automatically translate a searcher's preferred vocabularies into a set of the most semantically relevant terms in the database's underlying subject domain. By providing a more understandable, system-generated, semantics-rich concept space as an abstraction of the enormously complex information space plus algorithms to assist in concept/information spaces traversal, we believe we can greatly alleviate both information overload and the vocabulary problem.

In this paper, we first review our concept space approach and the associated algorithms in Section 2. In Section 3, we present our experience in using such an approach. In Section 4, we review our research plan for building a semantics-rich

Interspace for a multi-million dollar digital library project recently awarded by NSF/ARPA/NASA (1994-1998) to the University of Illinois. In particular, we will discuss the planned semantic retrieval and user customization capabilities for the next-generation NCSA Mosaic.

The Concept Space Approach and the Algorithms

To alleviate information overload and the vocabulary problem in information retrieval, researchers in human-computer interactions and information science have suggested expanding the vocabularies for objects and linking vocabularies of similar meanings. For example, Furnas et al. (1987) [8] proposed ``unlimited aliasing," which creates multiple identities for the same underlying object. In information science, Bates (1986) [1] proposed using a domain-specific dictionary to expand user vocabularies in order to allow users to ``dock" onto the system more easily. The general idea of creating rich vocabularies and linking similar ones together is sound and its usefulness has been verified in previous research and in many real-life information retrieval environments (e.g., reference librarians often consult a domain-specific thesaurus to help users in online subject search). However, the bottleneck for such techniques is often the manual process of creating vocabularies (aliases) and linking similar or synonymous ones (for example, the effort involved in creating an up-to-date, complete, and subject-specific thesaurus is often overwhelming and the resulting thesaurus may quickly become obsolete for lack of consistent maintenance).

Based on our experiences in dealing with several business, intelligence, and scientific textual database applications, we have developed an algorithmic and automatic approach to creating a vocabulary-rich dictionary/thesaurus, which we call the concept space. In our design, we generate such a concept space by first

extracting concepts (terms) automatically from the texts in the domain-specific databases. Similar concepts are then linked together by using several elaborate versions of co-occurrence analysis of concepts in texts. Finally, through generating concept spaces of different (but somewhat related) domains, intersecting common concepts, and providing graph traversal algorithms to lead concepts from a searcher's domain (queries expressed in his/her own vocabulary) to the target database domain, the concept space approach allows a searcher to explore in a large information space effortlessly and ``intelligently." We present the blueprint of this approach below.

A. Concept Identification:

The first task for concept space generation is to identify the vocabularies used in the textual documents. AI-based Natural Language Processing (NLP) techniques have been used for generating detailed, unambiguious internal representation of English statements. However such techniques are either too computationally intensive or are domain-dependent and are inappropriate for identifying content descriptors (terms, vocabularies) from texts in diverse domains. An alternative method for concept identification that is simple and domain-independent is the automatic indexing method, often used in information science for indexing literature. Automatic indexing typically includes dictionary look-up, stop-wording, word stemming, and term-phrase formation. Another technique (often called ``object filtering") which could supplement the automatic indexing technique involves using existing domain-specific keyword lists (e.g., a list of company names, gene names, researchers' names, etc.) to help identify specific vocabularies used in texts.

B. Concept Space Generation:

While automatic indexing and object filtering identify vocabularies used in texts, the relative importance of each term for representing concepts in a document may vary. That is, some of the vocabularies used may be more important than others in conveying meanings. The vector space model in information retrieval associates with each term a weight to represent its descriptive power (a measure

of importance). Based on cluster analysis techniques, the vector space model could be extended for concept space generation, where the main objective is to convert raw data (i.e., terms and weights) into a matrix of ``similarity'' measures between any pair of terms. The similarity measure computation is mainly based on the probabilities of terms co-occurring in the texts. The probabilistic weights between two terms indicate their strengths of relevance or association. We have developed several versions of similarity functions which considered the unique characteristics of the individual terms such as: the position of a term (e.g., a term in title vs. in abstract), the number of words of a term, the appearance date of the term (i.e., the publication year of the document which produced the term), and the identifiable type of the term (e.g., a person's name, a subject descriptor, a gene name, a company's name, etc.).

The proposed concept space generation techniques aim to defeat the vocabulary (difference) problem by identifying ``vocabulary similarity'' automatically. The output of the cluster analysis process can be perceived as an inter-connected, weighted network of terms (vocabularies), with each link representing the degree of similarity between two terms.

C. Intersecting and Traversing Multiple Concept Spaces:

A fundamental problem in information retrieval is to link the vocabularies used by a searcher (those he/she feels most comfortable and natural to use to express his/her own information need) with the vocabularies used by a system (i.e., indexes of the underlying database). By creating a target concept space using the texts of the underlying database (e.g., a C. elegans worm database) and another concept space from texts representative of the searcher's reference discipline (e.g., human genome) and intersecting and traversing the two concept spaces algorithmically, we believe we will be able to create a knowledgeable online search aide which is capable of bridging the vocabulary difference between a searcher and a target database, thereby helping alleviate the information overload problem in a large information space. We have tested a serial branch-and-bound search algorithm and a parallel Hopfield-like neural

network algorithm for multiple-thesauri consultation in previous research (Chen, Lynch, Basu, and Ng, IEEE Expert, 1993) [5]. The initial results were promising. In conclusion, by acquiring vocabularies from texts directly (instead of from human experts), either in incremental mode or by periodic batch processing, and by creating concept spaces for the target databases and other related subject disciplines (i.e., pre-processing selected source textual documents), a system will be able to help searchers articulate their needs and to retrieve semantically (conceptually) relevant information in a more effortless and effective way.

Our Experience in Using the Approach

We have tested the proposed techniques in several domains. In (Chen and Lynch, IEEE SMC, 1992) [4], we generated a Russian computing concept space based on an asymmetric similarity function we had developed. Using the indexes extracted from about 40,000 documents (200 MBs) and several weeks of CPU time on a small VAX VMS, we were able to generate a robust and domain-specific Russian computing thesaurus that contained about 20,000 concepts (countries, institutions, researchers' names, and subject areas) and 280,000 weighted relationships. In a concept-association experiment, the system-generated thesaurus out-performed four human experts in recalling relevant Russian computing concepts.

In (Chen, Hsu, Orwig, Hoopes, and Nunamaker, CACM, 1994) [3] and (Chen, IEEE Computer, 1994) [2], we tested selected algorithms in an electronic meeting environment where electronic brainstorming (EBS) comments caused the information overload and idea convergence problems. By extracting concepts in individual EBS comments, linking similar vocabularies together, and clustering related concepts, we were able to help meeting participants generate a consensus list of important topics from a large number of diverse EBS comments. In an experiment involving four human meeting facilitators, we found

that our system performed at the same level as two facilitators in both concept recall and concept precision (two measures similar to the conventional document recall and precision). Our system, which ran on either a DECstation or a 486, accomplished the concept categorization task in significantly less time and was able to trace the comments which supported the concluded topics.

In a recent NSF-funded project, we built a (C. elegans) worm concept space using the literature stored in the Worm Community System (WCS) (Chen, Schatz, Yim, and Fye, JASIS, 1994) [7]. Our algorithms were implemented in ANSI C and ran on both SUN SPARC stations and DECstations. It took about 4 hours of CPU time to analyze 5,000 worm abstracts and the resulting worm thesaurus contained 798 gene names, 2,709 researchers' names, and 4,302 subject descriptors. We tested the worm thesaurus in an experiment with six worm biologists of varying degrees of expertise and background. The experiment showed that the worm thesaurus was an excellent ``memory-jogging'' device and that it supported learning and serendipity browsing. The thesaurus was useful in suggesting relevant concepts for the searchers' queries and it helped improve search recall. The worm thesaurus was later incorporated as a concept exploration and search aid for the WCS.

As an extension of the worm thesaurus project and in an attempt to examine the vocabulary problem across different biology domains, we generated a fly thesaurus recently using 6,000 abstracts extracted from Medline and Biosis and literature from FlyBase, a database currently in use by molecular biologists in the Drosophila melanogaster-related research community. The resulting fly thesaurus included about 18,000 terms (researchers' names, gene names, function names, and subject descriptors) and their weighted relationships. In a similar fly thesaurus evaluation experiment involving six fly researchers at the University of Arizona, we confirmed the findings of the worm experiment. The fly thesaurus was found to be a useful tool to suggest relevant concepts and to help articulate searchers' queries.

Our initial comparison of the fly and worm thesauri revealed a significant overlap of common vocabularies across the two domains. However, each thesaurus maintains its unique organism-specific functions, structures, proteins, and so on. A manual tracing of fly-specific concepts and relevant links often lead to relevant, worm-specific concepts, and vice versa. We believe that by intersecting concepts derived from the two domain-specific concept spaces and by providing AI search methods we will be able to bridge the vocabulary differences between a searcher's (e.g., a fly biologist's) domain and the target database's (e.g., the worm database's) subject area. We are in the process of testing and fine-tuning several search algorithms (Chen and Ng, JASIS, 1994) [6] and we also plan to expand our subject coverage to other model organisms including e. coli, yeast, rat, and human in the near future. (Readers are encouraged to access the URL listed at the end of the paper for more information.)

Research Plan for the NCSA Mosaic Digital Library Project

In this section, we review the recently awarded Illinois digital library project and our research plan relating to semantic retrieval and user customization.

A. The Illinois Digital Library Project: An Overview

In the world of the near future, the Internet of today will evolve into the Interspace of tomorrow. The international network will evolve from distributed nodes supporting file transfer to distributed information sources supporting object interaction. Users will browse the Net by searching digital libraries and navigating relationship links, as well as share new information within the Net by composing new objects and links.

The Illinois digital library project (PI: B. Schatz) includes two concurrent and complementary activities that will accelerate progress towards building the

Interspace. These together construct a model large-scale digital library and investigate how it can scale up to the National Information Infrastructure.

- Construction of a digital library **testbed** for a major university engineering community, in which a large digital collection of interlinked documents and databases will be maintained, software to browse and share within this library developed, and usage patterns of thousands of users spread across the Net evaluated.
- Investigation of fundamental **research** issues in information systems, information science, computer science, sociology and economics that will address the scalable organization of a large digital collection to provide transparent access for a broad spectrum of users across national networks.

The testbed centers around the new Granger Engineering Library Information Center at the University of Illinois in Urbana-Champaign (UIUC). The $26M Center is intended as a showcase for state-of-the-art digital libraries and electronic information distribution. Construction of this national digital library testbed is possible through the active participation of two major institutions at UIUC, the University Library and the National Center for Supercomputing Applications (NCSA).

The digital library itself will be centered around a collection of engineering journals and magazines, obtained through collaboration with a range of major professional and commercial publishers. The intention is to attract a broad range of usage from a broad range of users. All documents will be structured and complete, that is, encoded in SGML and containing all pictorial material. The documents will include general engineering magazines (e.g., computer science from IEEE), specific engineering journals (e.g., aerospace engineering from AIAA), and specific scientific journals (e.g., physics from APS). Finally, articles from commercial engineering publishers (e.g., Wiley & Sons) will be collected for users in our economics (charging) study.

We plan to gather a significant new digital collection of structured documents in the engineering literature and combine this with existing sources available from our front end (Mosaic) and back end (BRS) software (discussed below). For example, these full-text materials will be integrated into an expanded on-line

catalog including access to major periodical indexes in science and engineering (Current Contents, Engineering Compendex, INSPEC) which will be linked to SGML documents. Collections on the Internet will also be made transparently available, e.g., the physics preprints at Los Alamos, the Unified Computer Science Technical Reports at Indiana University, and the international collection of on-line library catalogs.

The testbed software will go through two primary phases within the proposal period (September, 1994-August, 1995). The goal of version 1 is to leverage off our substantial existing resources to build a functional digital library with a large collection used by a substantial user population. Concurrently during this period, the technology research will be developing significant new functionality (semantic retrieval and customized retrieval will be described below) and sociology research will be observing the significant usage patterns of the existing functionality. Together, these efforts will enable us to develop and deploy scalable digital library technology on a national testbed. The goal of version 2 is to demonstrate the advanced technical feasibility of a full functional Interspace system.

The version 1 software will evolve from two of our existing projects. The first is the existing information retrieval system in the current Grainger Library developed by co-PI Mischo. This is based on a PC front end to a full-text retrieval search from the major commercial vendor BRS. The frond end on this search engine will be the NCSA Mosaic software developed under the supervision of co-PI Hardin. In essence, version 1 will exhibit the browsing and searching capabilities currently available on several Mosaic-based servers, e.g., EiNet Galaxy, the World Wide Wed Worm, the JumpStation, NorthStar, and so on. However, the Mosaic-BRS software will allow access to a large collection of well-formatted and recent engineering literature.

This paper focuses on the proposed Information Science research, which centers around semantic retrieval and user customization, supervised by co-PI Chen. the semantic retrieval supports a higher level of abstraction in user search which can

help overcome the vocabulary problem for information retrieval. Rather than searching for words within the object space, the search is for terms with in a concept space. Co-occurrence graphs seem to provide good suggestive power in specialized domains, such as biology. The research questions resolve around their effectiveness in the more general engineering domains. Using the same sort of techniques, it is possible to infer terms of interest to the users from the objects that have been retrieved. These techniques will be used to provide a form of customized retrieval, where a user profile consisting of terms and demographics specified by the users orients the search matching towards more preferred objects. In this project, the semantic retrieval and user customization will be used to supplement the full-text search and browsing in the testbed. Research plan for semantic retrieval and user customization are presented below.

B. Semantic Retrieval:

Based on our extensive experience in creating domain-specific concept space and supporting semantic, concept-based retrieval, we have found that the proposed techniques are robust and domain-independent and have shown great promise for supporting information retrieval in a large information space. We believe we are ready to employ the techniques experimentally on a larger and more general testbed collection and with a more diverse user population. Several of the proposed algorithms will be parallelized and implemented more efficiently on the NCSA machines (e.g., CM-5, SGI's Power Challenge, and Convex's Exemplar).

Our information science research plan will be based on an incremental and scale-up approach, starting from a few selected, focused scientific communities including molecular biology and physics and proceeding to testing in other general engineering and popular science domains. The research effort will coincide with the testbed collection process.

In the more restricted areas of molecular biology and selected physics domains, we will be able evaluate the concept spaces generated in detailed, controlled experiments. The effects of including concept spaces and the semantic retrieval

functionalities in the UIUC digital library environment will be studied through ethnography and user surveys of a larger user population.

Several crucial research questions in the context of large-scale digital libraries will be addressed in this project. First, we need to examine the feasibility of the proposed techniques in more general and diverse domains. While the concept space approach has been shown to be useful in relatively restricted scientific domains and with a somewhat more uniform user population, i.e., research scientists, will the concept spaces generated remain robust and useful for more general application domains and can they be used by searchers of varying backgrounds (e.g., professors and school children)? Second, we need to address the scalability issue by testing our techniques' ability to support semantic retrieval in an even larger (terabyte) information space. We plan to parallelize selected algorithms with the assistance of NCSA and are already designing algorithms for incremental update and generation of concept space.

We believe, with the realistic testbed collection and large user population proposed in this research, we will be able to examine critically the issues surrounding ``intelligent,'' semantic retrieval in a genuine, large-scale digital libraries environment and develop scalable technology to help alleviate the information overload and vocabulary problems in information retrieval.

C. Customized Retrieval:

In the digital libraries environment, there is a critical need to create ``user models'' which could aid in providing more customized information service and more focused and useful information sources and documents to individual users (e.g., a customized magazine for user X) and there is also a pressing need to know the retrieval patterns of different groups of users (e.g., what types of magazines and what subject areas are of most interest to the group of double-income, professional, suburban users?). The conventional manual approach to generating a user modeling component is infeasible because of the difficulty of achieving a complete and up-to-date knowledge base. However, the availability of large amounts of regular usage information (in the search logs) and the power

of selected statistical and machine learning algorithms to analyze usage patterns may be able to provide a more robust and algorithmic solution to creating user models for IR.

The availability of a large testbed collection and user population presents a unique opportunity to research a knowledge discovery approach to user modeling in digital libraries. After the completion of a significant portion of the testbed collection (i.e., molecular biology, physics, and engineering) and the selection and identification of the testbed users (e.g., engineering faculty and students at UIUC), we can proceed to collect (log) the usage data and statistics of selected user group (200-500 target users) over a period of several months. Each logged search session will include information such as date searched, magazines browsed, articles retrieved, search terms used, search options selected, and so on. Each user also will be requested during their first log-in to provide detailed demographic information.

Upon completion of usage data collection, we will proceed to analyze individual usage patterns. For example, by analyzing the articles retrieved during numerous IR sessions conducted by the same user using the concept space approach described earlier, we will be able to generate a smaller but more user-specific concept space that best represents that user's interests. Such a concept space could be invoked for future retrieval sessions (to match with the system's underlying, bigger concept space) or be stored as the user interest profile and used to selectively route relevant new information to the user. Other usage statistics such as types of magazines browsed, months/days of heaviest retrieval activity, etc. can also be used to provide more customized service to the individual searcher in the future.

Following the individual usage analysis, a user group analysis will be based on the information provided in the demographic surveys, the usage patterns shown by the entire group of users, and statistical (e.g., regressions and discriminant analysis) and/or machine learning (e.g., ID3) based analyses in an effort to determine the critical information needs of different user groups. Results of such

analysis may have a major impact on the practices of information providers (e.g., what types of advertisements should be placed in what kinds of magazines in order to attract the interest of targeted dual-career, suburban families?). A better profile and understanding of their main audiences could help individual information sources plan their marketing strategy.

In summary, we believe the digital libraries testbed collection and users proposed to be incorporated in this research present both a unique challenge and opportunity to study semantic retrieval and user customization in digital libraries. The results will have a potential impact on the practices of electronic publishers (e.g., IEEE, McGraw-Hill) and information retrieval service providers (e.g., UIUC engineering libraries or internet resource discovery).

Author Biographies

Hsinchun Chen received the Ph.D. degree in Information Systems from the Leonard N. Stern School of Business, New York University, New York, NY, in 1989. He is an Assistant Professor of Management Information Systems at the Karl Eller Graduate School of Business, University of Arizona. His research interests include CSCW, human-computer interactions, text-based information management and retrieval, multilingual information retrieval, internet resource discovery, knowledge acquisition and knowledge discovery, machine learning, and neural network modeling and classification. He received an NSF Research Initiation Award in 1992 and was awarded a Digital Library Initiative grant by NSF/NASA/ARPA (1994-1998) recently. Dr. Chen has published more than 20 articles in publications such as *Communications of the ACM*, *IEEE COMPUTER*, *IEEE Transactions on Systems, Man, and Cybernetics*, *IEEE EXPERT*, *Journal of the American Society for Information Science, Information Processing and Management, International Journal of Man-Machine Studies,* and *Advances in Computers.* He is a member of *IEEE*, *ACM*, *AAAI* and *TIMS*. EMAIL:

hchen@bpa.arizona.edu. For access to Chen's recent publications and work: http://bpaosf.bpa.arizona.edu:8000/

Bruce Schatz is associate professor of the Graduate School of Library and Information Science (joint in Computer Science Department) at the University of Illinois, Urbana-Champaign and research scientist at the Illinois National Center for Supercomputer Applications. He is the PI of a major NSF national Collaboratories project, ``the Worm Community Systems," 1990-1995, and a major NSF/ARPA/NASA Digital Library Project, ``Building the Interspace," 1994-1998. Both projects aimed to create large-scale, heterogeneous digital library for scientific and biology community users. Dr. Schatz won the NSF Young Investigator Award while he was at the University of Arizona (1992-1997). He has served as a member of the National Research Council and Internet Activities Board. EMAIL: bschatz@ncsa.uiuc.edu.

References

1

M. J. Bates. Subject access in online catalogs: a design model. *Journal of the American Society for Information Science*, 37(6):357-376, November 1986.

2

H. Chen. Collaborative systems: solving the vocabulary problem. *IEEE COMPUTER*, 27(5):58-66, Special Issue on Computer-Supported Cooperative Work (CSCW), May 1994.

3

H. Chen, P. Hsu, R. Orwig, L. Hoopes, and J. F. Nunamaker. Automatic concept classification of text from electronic meetings. *Communications of the ACM*, 37(10), October 1994.

4

H. Chen and K. J. Lynch. Automatic construction of networks of concepts characterizing document databases. *IEEE Transactions on Systems, Man and Cybernetics*, 22(5):885-902, September/October 1992.

5

H. Chen, K. J. Lynch, K. Basu, and T. Ng. Generating, integrating, and activating thesauri for concept-based document retrieval. *IEEE EXPERT, Special Series on Artificial Intelligence in Text-Based Information Systems*, 8(2):25-34, April 1993.

6

H. Chen and T. Ng. An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): symbolic branch-and-bound vs. connectionist Hopfield net activation. In *Journal of the American Society for Information Science*, 1994, in press.

7

H. Chen, B. Schatz, T. Yim, and D. Fye. Automatic thesaurus generation for an electronic community system. In *Journal of the American Society for Information Science*, 1994, forthcoming.

8

G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964-971, November 1987.