

# Collaborative Systems:

## Solving the Vocabulary Problem

Hsinchun Chen, University of Arizona

**Can on-line information retrieval systems negotiate the diverse vocabularies of different users? This article suggests a robust algorithmic solution to the vocabulary problem in collaborative systems.**

Research in information science and human-computer interactions has shown that people tend to use different terms to describe a similar concept, depending on their backgrounds, training, and experiences. Vocabulary differences have created difficulties for on-line information retrieval systems and are even more of a problem in computer-supported cooperative work (CSCW), where collaborators with different backgrounds engage in the exchange of ideas and information.

Our research group at the University of Arizona has investigated two questions related to the vocabulary problem in CSCW. First, what are the nature and characteristics of the vocabulary problem in collaboration, and are they different from those observed in information retrieval or in human-computer interactions research? Second, how can computer technologies and information systems be designed to help alleviate the vocabulary problem and foster seamless collaboration? We examine the vocabulary problem in CSCW and suggest a robust algorithmic solution to the problem.

### Vocabulary differences

Vocabulary differences in human-system interactions have been studied extensively in recent years. In almost all computer applications, users must enter correct words for the desired objects or information. Furnas et al. found that in spontaneous word choice for objects in five domains, two people favored the same term with less than 20 percent probability.<sup>1</sup> This fundamental property of language limits the success of various design methodologies for vocabulary-driven interaction. The designers' chosen vocabularies, which are often quite different from the users' preferred terms, can cause serious communication breakdown and an interaction bottleneck.

In information science, indexing and searching uncertainty have been recognized as the primary sources of information retrieval problems. Research has shown that

different indexers, even though well trained in an indexing scheme, might assign different index terms for a given document.<sup>2</sup> Research also shows that indexers will use different terms for the same document at different times (possibly because of increased familiarity with the material or because their state of mind while indexing has changed). Search terms also carry a high degree of uncertainty because searchers tend to use different terms for the same information.

Because of the indeterminism involved in indexing and searching, an exact match between the searcher's terms and those of the indexer is unlikely. Bates argues that for a successful match, the searcher must somehow generate as much variety in the search as is produced in indexing.<sup>2</sup> Further, to increase the chances of a successful match, there should be a number of indexes for each document, and searchers should articulate their needs clearly. This allows searchers to *dock* onto the system more easily.

However, the variety produced by an indexer can also be viewed as redundancy in the sense that it consists of partially overlapping classifications applied to a document, and, in practice, information science discourages indexing redundancy and favors conciseness and precision.

**Collaborative differences.** Collaboration involves multiple participants with shared goals and requires reciprocal exchange of ideas and extensive information sharing. The differences in the participants' backgrounds, knowledge, and expertise are likely to cause vocabulary differences during information sharing, which may adversely affect the process and outcome of collaborative activities. The geographic distribution of collaborators and the temporal aspect of collaboration (synchronous versus asynchronous) further complicate the vocabulary problem, as the concepts/ideas and their associated vocabularies may evolve and change over time.

In the following, we present two collaboration scenarios. One involves distributed, asynchronous information sharing and retrieval among molecular biologists. The other involves the synchronous generation and consolidation of ideas in a system-supported group meeting environment.

*Scientific collaboration.* The Worm Community System (WCS)<sup>3,4</sup> was recently developed as part of the National

## Status report

The *concept space* approach is an algorithmic solution for creating a vocabulary-rich dictionary/thesaurus. A concept space is generated by extracting concepts (terms) automatically from the texts produced during collaboration. Similar concepts are then linked through co-occurrence analysis. This approach guarantees that any terms brought out by group members will be captured and that terms with similar meanings will be linked (associated) by the system. The concept space represents the collective vocabularies used during collaboration, as well as the similarity probabilities between these vocabularies.

Two applications have been developed using the concept space approach:

- An automatic indexing tool for on-line scientific information retrieval in the National Laboratories environment. This system was implemented in X Windows.
- A concept generator tool for categorizing electronic brainstorming comments generated during electronic meetings in the GroupSystems environment. This application was implemented in MS Windows.

Future research will explore the feasibility of the concept space approach for seamless and *intelligent* Internet information access.

Science Foundation's National Laboratories effort for a community of molecular biologists who study the nematode worm, *Caenorhabditis elegans*. Molecular biology is a largely data-driven experimental science, and due to such efforts as the Human Genome Initiative, data is rapidly being accumulated and stored in databases.

Despite the usefulness of database technologies for community systems, the diverse vocabulary associated with collaborative information sharing poses a problem. For an electronic community system like the WCS, potential users vary

from expert worm biologists to novices, from senior worm insiders to community outsiders (for example, fly biologists). These users often do not share the same vocabularies, and they may experience difficulties using system-specific query terms.

Figure 1 shows several molecular biology-specific concepts related to worm and fly genome research. For example, both fly and worm biologists are familiar with the common concepts of *spermiogenesis*, *vulva*, and *egg*. But the specific functions, structures, and proteins of these three model organisms, as well as their associated terminologies, are very different and

<p>Fly-specific concepts in sperm motility:  {"flagella" a whip-like motility appendage present on the surface of flies}  {"actin" a fly muscle protein}</p> <p>Common fly-worm concepts in sperm motility:  {"spermiogenesis"}  {"vulva"}  {"egg"}</p> <p>Worm-specific concepts in sperm motility:  {"pseudopod" sperm of nematode worm}  {"MSPs" Major Sperm Proteins, worm-specific proteins}</p>
---

**Figure 1. Although molecular biologists would be familiar with general concepts (such as spermiogenesis, vulva, and egg), the functions and structures of specific organisms, as well as their associated terminologies, are very different and can cause problems for researchers accessing data from another research community.**

What are the problems our company needs to address to improve the product development process and engineering/support interface?

1.1 Vision

1.2 Selection of markets

1.3 Selection of products

1.4 Individuals who should be focused on market or product selection full time are burdened with day to day activities that greatly reduce their effectiveness.

...

2.2 Accountability a serious problem. Commitment to a customer or to critical market niche entry timing seems to drive us better.

2.3 When a development is undertaken, it must be accompanied by commitment not only from the design engineer, but also from test, fab, qa, marketing for PDSs, etc. As it is only the design engineer has accountability.

...

2.5 re 2.3 — everybody has accountability, but to whom, and with what measures/rewards/punishments for not attaining goals — what goals, they're seldom defined or stated clearly

...

3.2 Organization

3.3 Priorities for markets and products

...

**Figure 2. Electronic brainstorming promotes significant meeting productivity by letting group members enter comments or ideas simultaneously and anonymously. This sample output generated by a group of 15 manufacturing executives clearly illustrates that participants often use different vocabularies to convey similar ideas, creating a potential bottleneck during the subsequent idea-organization process.**

can cause researchers in one community great difficulty when accessing information from another community.

*Electronic brainstorming and idea organization.* While the above example shows the vocabulary problems associ-

ated with asynchronous scientific information sharing, the following example addresses the vocabulary problem posed by synchronous electronic brainstorming processes. An electronic brainstorming session helps achieve significant meeting productivity, especially for

divergent tasks, by letting group members enter comments or ideas simultaneously and anonymously.<sup>5</sup> During electronic brainstorming, the goal is to generate creative, uncensored ideas. Vocabulary differences often exist in the brainstorming comments because different participants use different vocabularies to convey similar ideas. During the subsequent idea organization process, which is considered a convergent task, the vocabulary differences tend to create a bottleneck. Very little system support has been provided to meeting participants during this cognitively demanding process (see the above sidebar on system-based solutions).<sup>6</sup>

Figure 2 presents sample electronic brainstorming comments generated by a group of 15 manufacturing firm executives. (Their "raw" quality illustrates the amount of "noise" generated in an EBS session.) Although the wording was different, comments 1.2 and 1.3 about "selection of markets" and "selection of products" are similar to comment 3.3 about "priorities for markets and products." Similarly, comment 1.4 focused on the role of individuals responsible for market and product selection, whereas comments 2.2, 2.3, and 2.5 addressed the "accountability" issue from different angles. For the 300-plus comments generated by this group (a typical group of 10-20 participants can generate 300-500 comments), the task of browsing the comments and consolidating ideas was overwhelming.

## CSCW research on system-based solutions

Based on our review of the CSCW literature, we found that many researchers recognize the existence and importance of the vocabulary problem but do not suggest any algorithmic or system-based solution to it. Johansen<sup>7</sup> discussed the organizational and system development issues involved in synchronous and asynchronous collaborations. He commented that asynchronous communication capabilities in particular will be one of the most compelling features of many groupware products.

We echo his view and further postulate that the evolution and change of concepts and ideas over time, as perceived by different collaborators, could cause asynchronous collaboration to become extremely complex and dynamic. As concepts and ideas evolve, a group member's vocabulary may change, and different group members' vocabularies may differ significantly; this may cause serious communica-

tion breakdown. The temporal features of group work and how those temporal matters are affected when technological tools are added are discussed extensively in McGrath<sup>8</sup> from a primarily organizational rather than a system development perspective.

Previous CSCW field research has demonstrated the importance of collaborative information sharing and retrieval and has pinpointed problems associated with current system-supported information processing functions in various scientific, engineering, and business domains.<sup>3,6,7</sup> A major difficulty in accessing pertinent information stems from current systems' lack of support for concept-based information retrieval. Searchers able to express their concepts using their own vocabularies are unlikely to find relevant information because of the vocabulary differences between different collaborators and between the system and the searchers.

## Solving the problem: A concept space approach

To solve the vocabulary problem, researchers in human-computer interactions and information science have suggested both expanding the vocabularies for objects and linking vocabularies of similar meanings. For example, Furnas et al.<sup>1</sup> showed that creating an unlimited number of aliases for objects helps alleviate the vocabulary problem. In information science, Bates<sup>2</sup> proposed using a domain-specific dictionary to expand user vocabularies and let users “dock” onto the system more easily. However, even though the usefulness of rich vocabularies has been verified, the manual process of creating different vocabularies (aliases) and linking similar or synonymous ones often creates a bottleneck.

Based on our experiences with several collaboration applications, we have developed the following algorithmic approach for 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 produced during collaboration. Similar concepts are then linked through the co-occurrence analysis of concepts in texts. This approach guarantees that any terms brought out by group members will be captured and that terms with similar meanings will be linked (associated) by the system. The concept space represents the collective vocabularies used during collaboration, as well as the similarity probabilities between these vocabularies. We present the blueprint of our concept space approach below.

**Vocabulary identification.** Despite the increasing availability of other presentation media such as image, voice, animation, and video, the most natural and popular means of communication remains natural language. In system-supported collaboration, on-line textual output reveals the vocabularies used by different members and can be used to create a shared concept space for all group members.

The first task for concept space creation is to identify the vocabularies used in textual collaboration. AI-based natural language processing (NLP) techniques such as the Augmented Transition Net-

work (ATN) parsing, case grammar, and semantic grammar have been used for creating unambiguous internal representation of English statements. However, such techniques are either too computationally intensive or are domain-dependent and therefore inappropriate for identifying content descriptors (terms, vocabularies) from texts. A simple and domain-independent alternative for content identification is the automatic indexing method, often used in information science for indexing literature.

**Linking similar vocabularies.** While automatic indexing identifies vocabularies used in texts by different group members, the relative importance of each term for representing the group members' concepts may vary. That is, some of the

### Vocabulary is a problem in the scientific community because of the specialized domains and the fluid nature of scientific discovery.

vocabularies used may be more important than others in conveying meanings. Salton's Vector Space Model<sup>9</sup> assigns each term a weight to represent its descriptive power (a measure of importance). Among the many probabilistic techniques that have been developed by information science researchers, those that incorporate *term frequency* and *inverse document frequency* have been found to be quite useful.<sup>9</sup> The basic rationales underlying these two measures are: (1) terms that appear more times in a specific text should be assigned higher weights (term frequency), and (2) more specific or unique terms that appear in fewer texts should also have higher weights (inverse document frequency).

The Vector Space Model can be extended for concept space generation by using *cluster analysis*. The first stage in cluster analysis is to convert the raw data (terms and weights) into a matrix of similarity measures between any pair of terms. The similarity measure computa-

tion is mainly based on the probabilities of terms co-occurring in a text produced during collaboration. The probabilistic weights between terms indicate their strengths of relevance or association. For example, if two terms (such as “manufacturing process” and “product fabrication”) appear in many texts produced during collaboration, their high degree of co-occurrence will cause cluster analysis to produce a strong probabilistic weight between them.

**Traversing the concept space.** When a group member encounters a vocabulary problem during collaboration, it would be helpful if he or she could consult (browse) the concept space and identify other relevant vocabularies for use. This is in fact what professional librarians do when assisting patrons in finding relevant information (a collaborative information searching process).

The “unlimited aliasing” proposed by Furnas et al.<sup>1</sup> lets users consult a manually created concept space of synonymous terms. (Cluster analysis creates *similar* links, not *synonymous* links, although in practice many synonymous terms will have a high similarity probability between them.) An alternative method can be based on system-aided, multiple-link searching algorithms.<sup>10</sup> For example, a Hopfield network search could traverse the concept space in a parallel mode and combine evidence from multiple links until the search algorithm converges.<sup>11</sup>

To illustrate our approach to the vocabulary problem, we apply it to the two collaborative applications discussed earlier. In our first example, we describe a system that builds a concept space for a community of worm biologists engaged in asynchronous information retrieval and information sharing. In our second example, we describe a concept classification tool for synchronous electronic meetings. The tool helps extract vocabularies used in the brainstorming comments and assists in consolidating ideas into a list of consensus topics. Both applications rely on our techniques for identifying vocabulary similarity.

## Asynchronous information sharing

The Worm Community System is considered a model electronic community system. It offers traditional database functionalities, along with literature, in-

formal information, research lore, mapping programs, graphics, and the ability for users to browse, share, and filter a large amount of timely worm community knowledge.

Vocabulary is problematic in the molecular biology community because of the diversity of specialized domains and the process of scientific discovery, especially in genome research. According to Frenkel, "Biology... involves concepts that are dynamic, or fluid, meaning that the phenomena under study, and the scientists' understanding of them, keep changing."<sup>12</sup> Experimental errors or approximations are common occurrences, and definitions for concepts will evolve and "become better understood as more knowledge is accumulated and integrated."<sup>12</sup>

Our research aims to develop a worm concept space that captures the unique vocabularies used in the worm genome research (for example, genes, functions, and subjects). We used four main sources of textual documents in the WCS for vocabulary identification and concept space generation: *The Worm Book*, the *Worm Breeder's Gazette*, journal articles, and conference proceedings abstracts. They comprised 4,714 documents and 8 Mbytes of textual information. (Future research will apply the same techniques to build a fly concept space, a human concept space, and so on.)

**Building a worm concept space.** To identify candidate descriptors in each document, we performed object filtering and automatic indexing. Several object filters were created for genes, researchers, experimental methods, and subjects. For example, 1,520 gene names were identified from the WCS gene list and the conference proceedings articles. Automatic indexing was implemented based on the procedure reported in Salton.<sup>9</sup>

After the concept descriptors for each document were identified, we performed term co-occurrence analysis for all documents in the WCS. Here is the procedure for generating such a concept space:

- (1) Compute the term and document frequency for each term. Term frequency,  $tf_{ij}$ , represents the number of occurrences of term  $j$  in document  $i$ . Document frequency,  $df_j$ , represents the number of documents in a collection of  $N$  documents in which

term  $j$  occurs. High term frequency indicates that a term is highly related to a document. High document frequency, on the other hand, indicates that a term is too general to be useful as a descriptor and has no descriptive power.

- (2) Compute the combined weight of term  $j$  in document  $i$ ,  $d_{ij}$ , based on the product of "term frequency" and "inverse document frequency" as follows:

$$d_{ij} = tf_{ij} \times \log\left(\frac{N}{df_j} \times w_j\right)$$

where  $N$  represents the total number of documents in the WCS, and  $w_j$  represents the number of words in descriptor  $T_j$ . Multiple-word terms are assigned heavier weights than single-word terms because multiple-

word terms usually convey more precise semantic meaning than single-word terms.

$$d_{ijk} = tf_{ijk} \times \log\left(\frac{N}{df_{jk}} \times w_j\right)$$

where  $tf_{ijk}$  represents the smaller number of occurrences of either term  $j$  or term  $k$  in a given document  $i$ . Terms  $j$  and  $k$  do not need to appear in the same word span. The expression  $df_{jk}$  represents the number of documents (in a collection of  $N$  documents) in which terms  $j$  and  $k$  occur together. The term  $w_j$  represents the number of words of descriptor  $T_j$ .

Figure 3 shows sample entries in the system-generated co-occurrence tables for the WCS. The complete worm concept space consisted of about 8,000 worm-specific terms and 1.7 million probabilistic links. For example, using *msp* and *oocyte* as query terms, the thesaurus window displays related terms in ranked order (*spermatogenesis*, *male*, *hermaphrodite*, etc.). The searcher can add thesaurus terms to a query and retrieve relevant objects from the WCS (for example, gene descriptions, articles, and so forth).

## The worm concept space consists of about 8,000 worm-specific terms and 1.7 million probabilistic links.

- (3) Generate term co-occurrence tables based on the asymmetric Cluster Function we developed.<sup>10</sup> (We have shown that this asymmetric similarity function represents term association better than the popular cosine function because it often generates more specifically related terms.)

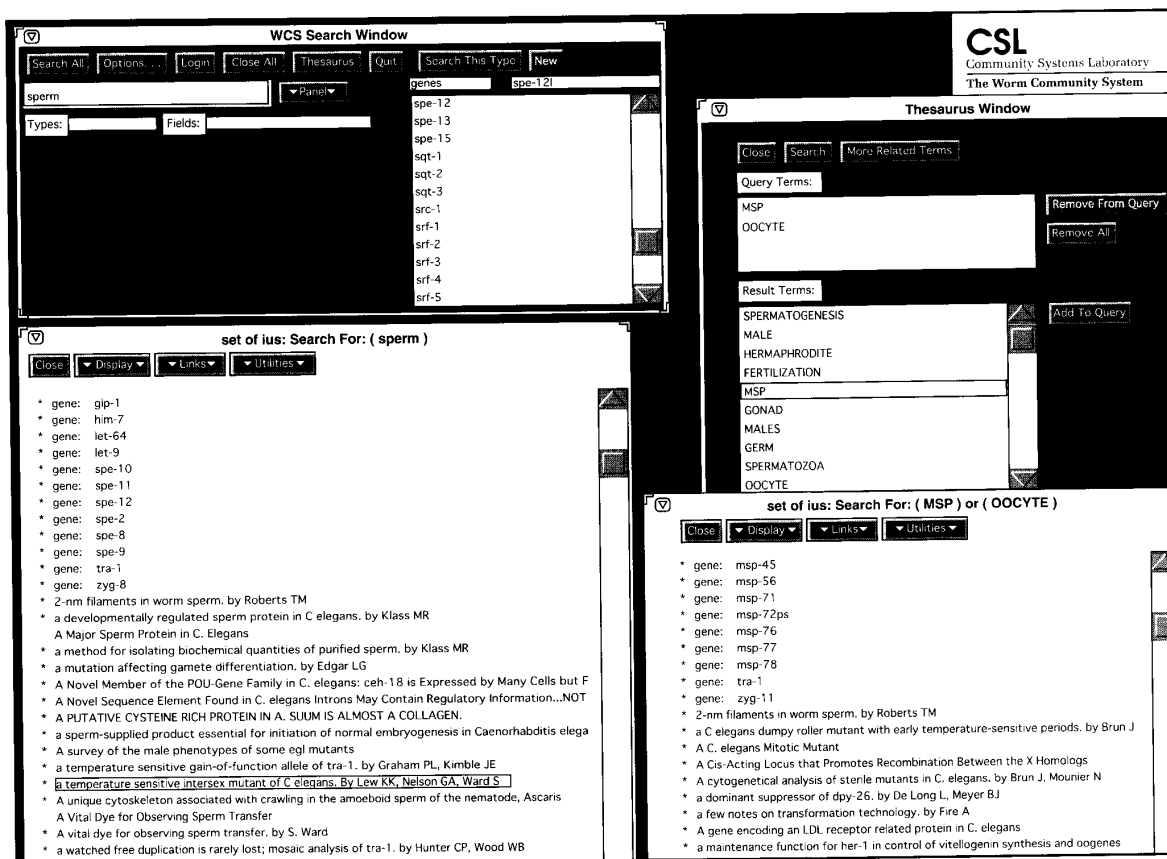
$$ClusterWeight(T_j, T_k) = \frac{\sum_{i=1}^n d_{ijk}}{\sum_{i=1}^n d_{ij}}$$

$$ClusterWeight(T_k, T_j) = \frac{\sum_{i=1}^n d_{ikj}}{\sum_{i=1}^n d_{ik}}$$

These two equations indicate the similarity weights from term  $T_j$  to term  $T_k$  (the first equation) and from term  $T_k$  to term  $T_j$  (the second equa-

**Worm concept space evaluation.** To evaluate the usefulness of the worm concept space, we conducted an experiment in the Winter of 1992 with subjects from the Arizona Worm Laboratory. The experiment consisted of two parts: a term-association experiment and a searcher-browsing experiment. We selected six subjects with different backgrounds to determine the effects of a searcher's expertise on thesaurus usage. Subjects 1 and 2 were considered experts in molecular and cellular biology (MCB), having worked in the Arizona worm lab for several years and having published papers in this field. Subjects 3 and 4 were considered novices; one was a master's student in MCB and the other an undergraduate senior. Subjects 5 and 6 were outsiders with respect to worm research. Subject 5 was a new master's student in MCB who had extensive working experience in fly biology. Subject 6 was a graduate student in ecology and evolutionary biology at the University of Arizona.

The first step of the term-association



**Figure 3.** The worm concept space for the Worm Community System contains more than 8,000 terms and 1.7 million probabilistic links. Here, a searcher is using *msp* and *oocyte* as query terms, and the thesaurus window displays related terms in ranked order (spermatogenesis, male, hermaphrodite, and so forth). Searchers can add thesaurus terms to a query and retrieve additional objects from the WCS (for example, gene descriptions or articles).

experiment was to give each subject a sequence of 16 terms chosen with the help of several worm researchers. Terms included researchers' names, gene names, and subject descriptors. The subjects were asked to write down concepts (genes, researchers, methods, and subject descriptors) related to each preselected term. Then, the subjects were asked to rank system-generated terms according to their relevance. After the term-association experiment, subjects were asked to browse the on-line worm thesaurus freely, using any terms they preferred and exploring as they wished. During browsing, subjects were asked to think aloud and give specific comments, observations, or suggestions. We summarize the results below.

- *The worm concept space helped suggest more relevant terms.* By counting

the numbers of terms generated by the subjects themselves and the system-suggested terms marked relevant by the subjects, we were able to tabulate and analyze whether the concept space was able to contribute relevant terms during a retrieval process. A quantitative analysis revealed that the worm concept space was able to suggest 8.5 terms for each concept, which was significantly higher than the average number of terms produced by the six subjects (6.1). In particular, the worm concept space helped identify more relevant terms for novices and outsiders.

- *For the most part, the system appeared to list the more relevant terms first.* According to Subject 5, for example, "The first 11 are great, but after that, they're not so good. The first five are quite relevant, six and seven are way

too general. I'd say it's very good initially, and then there's a fairly low frequency of relevance." The system's ability to list more relevant terms first is important, especially when it comes to designing an effective thesaurus-browsing interface.

- *Learning, serendipity browsing, and memory-jogging occurred frequently during experiments.* Many subjects found something interesting or unexpected that would help them in their queries. This was particularly evident with novices and outsiders, who were often amazed by the thesaurus' ability to relate genes, researchers, and subject topics. For all subjects, the thesaurus also served as an excellent tool to remind them of something they previously had forgotten. For example, Subject 5 said, "This is doing extremely well.

This is potentially very useful to me, I mean, I'm the novice." Searching the term *longevity*, Subject 6 said, "I am not familiar with people who do stuff about this . . . . There might be some paper . . . . Oh yeah, spermio pumping."

The experiment's results were very encouraging. The worm concept space suggested relevant terms and concepts that would not only be helpful for different users, but useful in spurring users' acquisition of knowledge. Without the assistance of a system-generated concept space, searchers of a large scientific database would have to perform a trial-and-error process of generating various

search terms themselves, using their mental model of the subject domain (a painstaking and cognitively demanding process). In contrast, a concept space can serve as an on-line search aid and can be invoked by searchers for query refinement and concept exploration.

## Synchronous collaboration

Our second example is based on the GroupSystems electronic meeting system, a pioneering example of meeting software technology developed at the University of Arizona<sup>5</sup> and installed at

more than 300 organizational sites, including businesses, government agencies, and universities. Many group meetings follow a common sequence, typically beginning with participants generating ideas, organizing those ideas into a list of key issues, and prioritizing them into a short list. The group then generates ideas for action plans to address the important issues, following the steps just described.

Electronic brainstorming (EBS) and idea organizer (IO) tools have been used frequently in electronic meetings. Electronic brainstorming allows group members to enter comments or ideas simultaneously and to share them anonymously. An idea organizer allows participants to identify and consolidate ideas, typically by separately suggesting topics or ideas that merit further consideration by the group. During the consolidation process, group members can browse the list of EBS comments and interact verbally with each other and the facilitator to condense the topic list to a manageable size by eliminating redundant or extraneous topics. While the EBS process is often productive, the IO process organizing the EBS comments can be problematic.

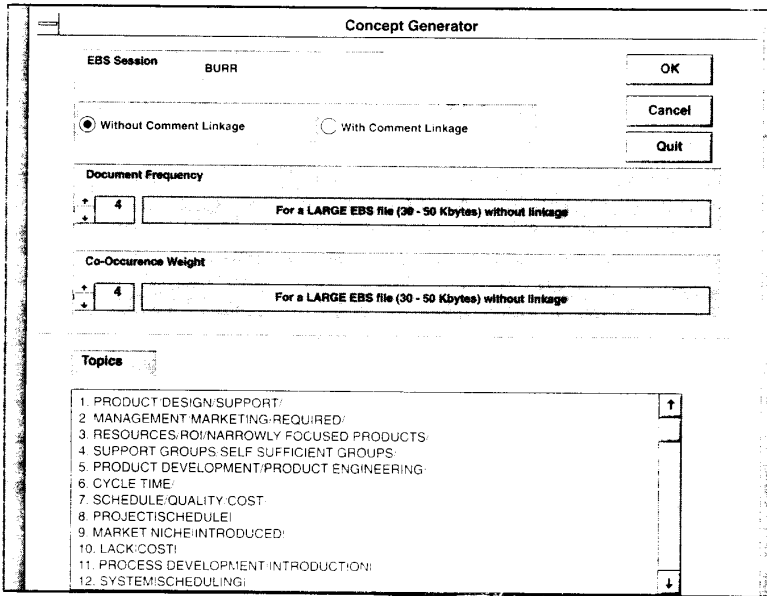
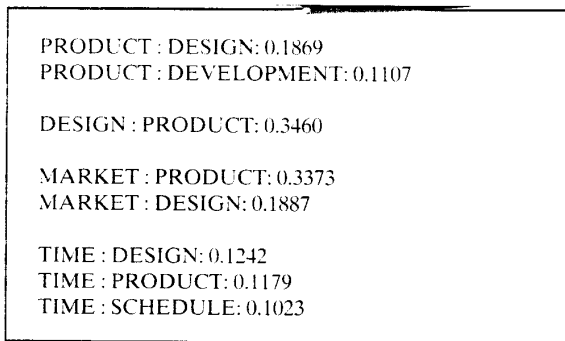
Using our concept-space approach, we designed an on-line tool to extract the vocabularies used in the EBS comments, identify the similarities among vocabularies (in the context of the meeting agenda and topics), and cluster similar vocabularies into unique topics. Lessons learned from our preliminary evaluation are described below.

### Building a meeting concept space.

Throughout, we used automatic indexing to identify terms (single words and multiple words) from the EBS comments. Cluster analysis was then adopted to identify co-occurrence probabilities between any two terms. Finally, we used the Hopfield net algorithm to group similar terms representing similar ideas. Each group of terms then represented a unique topic. We present sample output from the process below. (Chen et al.<sup>6</sup> contains details.)

To improve the granularity of analysis, we treated each meeting comment as a document — the basic information unit for our automatic indexing procedure. The EBS output, after automatic indexing, consisted of a list of terms (indexes). After indexes were assigned to each comment, we used the *cluster function* to identify the co-occurrence pattern of vocabularies that appeared in all comments.

**Figure 4.** The concept space approach, when applied to an EBS session, outputs a network of disjunctors and their weighted relationships.



**Figure 5.** The concept generator tool for GroupSystems provides a textual representation of concepts generated during EBS sessions. The tool lets users generate topics for EBS comments on the fly.

Figure 4 shows some partial co-occurrence tables generated by our system for the manufacturing session described earlier.

The final output was a network of descriptors and their weighted relationships, akin to a neural network of nodes and weighted links. Our system adopted a variant of the Hopfield network activation procedure to identify clusters of relevant descriptors in the concept space through their weighted links.<sup>6</sup> Figure 5 shows our concept generator tool for EBS sessions, as well as a textual representation of the topics (clusters of relevant descriptors) identified by our system for the manufacturing session. The tool lets users generate topics for EBS comments on the fly.

**Topics list evaluation.** Many similar algorithms developed earlier for the WCS application were adopted for this system. The EBS comments were extracted from the GroupSystems output files. For a typical EBS output of several hundred comments, the complete process of automatic indexing, concept space generation, and Hopfield net classification lasted about 4 minutes.

To determine the performance and usefulness of the system's classification results and to pinpoint directions for future research, we conducted an EBS comment classification experiment that compared the system's suggested list of topics with those generated by professional meeting facilitators.

Four facilitators were chosen for their expertise in guiding groups through the idea organization process. We felt that their experience prepared them to develop some criteria as to what would make a good list. The experiment consisted of two stages: a categorization stage and a list-evaluation stage.

In the first stage, each facilitator was presented with a complete set of brainstorming comments generated by an actual group. Each facilitator was asked to create a list of topics that pertained to those comments.

In the second stage, each facilitator was asked to critique five lists: three were generated by the other facilitators, one was generated by the actual group, and one was generated using our system. Each facilitator was first asked to rank the five lists. Following the ranking, we asked the subjects to add topics that they felt were missing or delete topics that they felt were inappropriate. We then tallied the results for each list. We did not al-

low facilitators to access their own lists during the list evaluation stage.

There was approximately a week of lag time between the two stages for each facilitator. The experiment revealed that

- The system's output list was comparable to those of two facilitators but not as complete as the output of the other two facilitators. Lists generated by facilitators 1 and 2 were consistently ranked higher by all facilitators and were considered to have more relevant concepts than those generated by the other facilitators or the system. Facilitators 1 and 2 appeared to be better at capturing the meeting ideas at the appropriate ab-

---

## As information space grows, information access and sharing will become increasingly more complex and challenging.

---

straction level — neither too general nor too specific. This observation was confirmed in the verbal protocols stated by several facilitators during the list evaluation stage.

- Facilitators spent more time (50-168 minutes) evaluating and classifying EBS comments, and their performances varied widely. This variation of performance may have been due in part to differences in the facilitators' areas of expertise. Despite our effort to choose a session that was nontechnical in nature, facilitators 3 and 4 still expressed discomfort with the topics discussed in the manufacturing session. We can continue to expect a wide variation in facilitator performance because facilitators will inevitably have to monitor sessions on topics about which they have little knowledge, and because of the time constraints and the extreme cognitive demand of real-life meeting processes (session facilitation, software and hardware control, group dynamics monitoring, and so forth).
- The system's list needs to be more

precise and should be presented within proper context. Based on the comments supplied by the subjects, we were able to identify some directions for improving the system's performance and suggestions for using the system's analysis. Terms on the system list need further expansion in order to provide clearer meaning. One immediate solution would be to let facilitators or meeting participants browse the comments associated with the topics on the list and let them make necessary refinements. Because all EBS comments are indexed by the system, our bottom-up approach of generating meeting topics from EBS comments provides the added benefit of letting humans trace the justification of the topics suggested.

In summary, the proposed concept-space approach to the automatic classification of electronic brainstorming output presents several unique advantages over the conventional manual approach. As discussed earlier, meeting participants often felt overwhelmed by the large number of EBS comments and were discouraged by the convergent task of generating a list of consensus topics. The cognitive demand of the manual convergence process and the varying quality of facilitator support often made the idea organization stage a less than satisfying experience. Our proposed approach provides an efficient and algorithmic (domain-independent) alternative for the analysis of the EBS comments. The system's topic list can be used as a "straw man" for further group discussion and refinement. The refinement process will then be significantly less cognitively demanding and more efficient. We believe the proposed approach can help alleviate the group vocabulary problem and assist in converging ideas in electronic meetings.

**C**ollaboration, a process that involves multiple collaborators working jointly for the same goal, is severely constrained by the backgrounds, experiences, and expertise of its collaborators. Like the information retrieval and the human-computer interaction environments in which people may use different terms to describe the same object or concept, vocabulary differences may create a significant bottleneck for both synchronous and asynchronous collaborations. We have presented an algo-

rhythmic concept space approach that relies on various information science, statistical, and artificial intelligence techniques, including object filtering, automatic indexing, cluster analysis, and search/classification algorithms. In contrast to the previous manual techniques for solving the vocabulary problem (for example, unlimited aliasing), our techniques automatically extract vocabularies from text, identify vocabulary similarities, and group similar vocabularies together. This computationally intensive approach helps alleviate some cognitive burden of collaborators while consolidating different vocabularies.

Our current research effort involves testing the concept space approach for seamless and *intelligent* Internet information access. As information space continues to grow due to the recent growth of Internet resource discovery and the national information infrastructure, information access and sharing will become increasingly more complex and challenging. We believe the difficulties associated with keyword searching and user browsing can also be partially alleviated by adopting the proposed concept space approach. ■

## Acknowledgments

This project was supported mainly by NSF grant No. IRI-9211418, 1992-1994. I thank J.F. Nunamaker, director, Center for the Management of Information, University of Arizona, for making available the GroupSystems output and facilities, and B. Schatz, director, Community Systems Laboratory, for his support in building a worm concept space for the Worm Community System.

## References

1. G.W. Furnas et al., "The Vocabulary Problem in Human-System Communication," *Comm. ACM*, Vol. 30, No. 11, Nov. 1987, pp. 964-971.
2. M.J. Bates, "Subject Access in On-line Catalogs: A Design Model," *J. Am. Soc. for Information Science*, Vol. 37, No. 6, Nov. 1986, pp. 357-376.
3. B. Schatz, "Building an Electronic Community System," *J. Management Information Systems*, Vol. 8, No. 3, Winter 1991-92, pp. 87-107.
4. R. Pool, "Beyond Database and E-mail," *Science*, Vol. 261, Aug. 1993, pp. 841-843.
5. J.F. Nunamaker et al., "Electronic Meeting Systems to Support Group Work,"

*Comm. ACM*, Vol. 34, No. 7, July 1991, pp. 40-61.

6. H. Chen et al., "Automatic Concept Classification of Text from Electronic Meetings," *Comm. ACM*, to be published in 1994.
7. R. Johansen, *Groupware: Computer Support for Business Teams*, The Free Press, New York, 1988.
8. M.C. McGrath, "Time Matters in Groups," in *Intellectual Teamwork: Social and Technological Foundations of Cooperative Work*, J. Galegher, R.E. Kraut, and C. Egidio, eds., Lawrence Erlbaum Associates, Hillsdale, N.J., 1990.
9. G. Salton, *Automatic Text Processing*, Addison-Wesley Publishing, Reading, Mass., 1989.
10. H. Chen and K.J. Lynch, "Automatic Construction of Networks of Concepts Characterizing Document Databases," *IEEE Trans. Systems, Man and Cybernetics*, Vol. 22, No. 5, Sept.-Oct. 1992, pp. 885-902.
11. H. Chen et al., "Generating, Integrating, and Activating Thesauri for Concept-Based Document Retrieval," *IEEE Expert*, Special Series on Artificial Intelligence in Text-Based Information Systems, Vol. 8, No. 2, Apr. 1993, pp. 25-34.
12. K.A. Frenkel, "The Human Genome Project and Informatics," *Comm. ACM*, Vol. 34, No. 11, Nov. 1991, pp. 41-51.

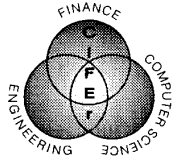
## Call for Papers

### IEEE/IAFE Conference on Computational Intelligence for Financial Engineering

April 9-11, 1995, New York City, Crowne Plaza Manhattan

The IEEE/IAFE CIFE Conference is the first major collaboration between the professional engineering and financial communities, and will be the leading forum for new technologies and applications in the intersection of computational intelligence and financial engineering. Intelligent computational systems have become indispensable in virtually all financial applications, from portfolio selection to proprietary trading to risk management. Topics in which papers, panel sessions, and tutorial proposals are invited include, but are not limited to, the following:

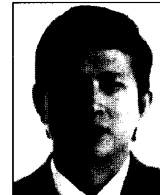
**Financial Engineering Applications**  
 Asset Allocation  
 Trading Systems  
 Corporate Financing  
 Forecasting  
 Hedging Strategies  
 Options and Futures  
 Risk Arbitrage  
 Risk Management  
 Complex Derivatives  
 Currency Models  
 Technical Analysis  
 Portfolio Management  
 Standards Discussions



**Computer & Engineering Applications & Models**  
 Neural Networks  
 Machine Intelligence  
 Probabilistic Reasoning  
 Fuzzy Systems  
 Parallel Computing  
 Pattern Analysis  
 Genetic Algorithms  
 Stochastic Processes  
 Dynamic Optimization  
 Knowledge & Data Engineering  
 Time Series Analysis  
 Harmonic Analysis  
 Signal Processing  
 Non-Linear Dynamics



For more information contact:  
 Meeting Management 2603 Main Street, Suite 690, Irvine, CA 92714  
 (714) 752-8205 Fax (714) 752-7444



**Hsinchun Chen** is an assistant professor of management information systems at the Karl Eller Graduate School of Management, University of Arizona. His research interests include CSCW, human-computer interactions, text-based information management and retrieval, Internet resource discovery, knowledge acquisition and knowledge discovery, and neural-network modeling and classification.

He received his PhD in information systems from New York University in 1989. He is a member of IEEE, ACM, AAAI, and The Institute of Management Sciences. He received an NSF Research Initiation Award in 1992 and the HICSS Conference Best Paper Award in 1994.

Readers can contact Chen at the Karl Eller Graduate School of Management, University of Arizona, McClelland Hall 430Z, Tucson, Arizona 85721; e-mail hchen@bpa.arizona.edu