MODELING INFORMATION NEEDS IN ENGINEERING DATABASES USING TACIT KNOWLEDGE

Shuang Song, Andy Dong and Alice Agogino
Department of Mechanical Engineering
University of California, Berkeley
5136 Etcheverry Hall
Berkeley, CA 94720-1740
{shuang,adong,aagogino}@me.berkeley.edu

ABSTRACT

Online resources of engineering design information are a critical resource for practicing engineers. These online resources often contain references and content associated with technical memos, journal articles and "white papers" of prior engineering projects. However, filtering this stream of information to find the right information appropriate to an engineering issue and the engineer is a time-consuming task. The focus of this research lies in ascertaining tacit knowledge to model the information needs of the users of an engineering information system. It is proposed that the combination of reading time and the semantics of documents accessed by users reflect their tacit knowledge. By combining the computational text analysis tool of Latent Semantic Analysis with analyses of on-line user transaction logs, we introduce the technique of Latent Interest Analysis (LIA) to model information needs based on tacit knowledge. Information needs are modeled by a vector equation consisting of a linear combination of the user's queries and prior documents downloaded, scaled by the reading time of each document to measure the degree of relevance. A validation study of the LIA model revealed a higher correlation between predicted and actual information needs for our model in comparison to models lacking scaling by reading time and a representation of the semantics of prior accessed documents. The technique was incorporated into a digital library to recommend engineering education materials to users.

INTRODUCTION

Engineering design is an information intensive activity. Engineers and engineering corporations are making wider adoption of corporate intranets and the Internet as online databases of design information and knowledge [1]. One study reported that designers spent in excess of 50% of their time handling information, e.g., retrieving, organizing, etc. [2]. Thus, the efficiency and the quality of the design process may depend considerably on how well designers are able to handle large amounts of information.

To date, much of the research in engineering information management systems has focused on the capture, storage, indexing and representation of design information [3-5] whereas industry has focused on product data management (PDM) systems. Fewer research has focused on modeling the information needs of engineering designers with respect to their skills and the ways in which background and supporting information in the context of their discipline, level of expertise and design context could be applied to specific design tasks.

This paper reports results on identifying and modeling information needs of users of an engineering database of education materials. The presumption is that the body of material that a user accesses illuminates the underlying tacit knowledge that arises from the education background and engineering enterprise in which the user practices. A design task, such as determining functional constraints, developing preliminary configurations, or refining a prototype against technical and economic criteria may trigger an information need to assist in the task. However, the selection of the material to satisfy the information need will be based not only on the relevancy of the material to the design task (e.g., similar design case) but also on its satisfaction of the informal, abstract knowledge of underlying engineering principles held by the engineer — that is, the tacit knowledge of the engineer.

We introduce Latent Interest Analysis (LIA) as a method for identifying information needs based on tacit knowledge by combining Latent Semantic Analysis (LSA) with the reading time associated with the materials accessed by a user of an engineering information system. Our results indicate that there is a strong correlation between time spent reading a document and the relevance of the document to the user. They also validate the proposition that users' historical patterns of information access behavior reflect, to a degree, their tacit knowledge, which can be used to model their information needs.

INFORMATION NEEDS AND TACIT KNOWLEDGE

The focus of this research is the study of a method to model information needs so as to address the issue of retrieving and filtering information from engineering information databases such as digital libraries, knowledge repositories, scholarly papers and industry reports. Vaughan [6] classifies information needs of users into two categories: (1) the need to locate specific documents for which the bibliographical references are known – referred to as a need for a *known item*; and (2) the need to locate documents relating to a particular theme – known as a *thematic need*. Examples of engineering information retrieval tasks relating to satisfying thematic information needs include updating general knowledge from professional development resources, consulting references on design cases and codes, and reviewing background technical literature relating to a specific design task. In the first category, the information needs are explicit, well-defined and readily satisfied. Information retrieval using structured Boolean queries on formal characteristics of the information, such as date or author, satisfies this category of information needs. In this paper, we only consider more complex thematic information need*s* which must not only satisfy task or subject relevance but also accord with the tacit knowledge of the engineer.

Michael Polanyi introduced the concept of "tacit knowledge" as knowledge that was incapable of full explicit expression, which underlies all explicit knowledge [7]. Tacit knowledge enters into the production of behavior and the constitution of mental states but is not ordinarily accessible to consciousness. In engineering design, tacit knowledge has been described as "the personal rendering of scientific beliefs and technical possibilities" by Bucciarelli [8]. These beliefs often arise from factors including the engineer's training and educational background, the community of professionals with which the engineer is part of, and the engineer's engineering enterprise (e.g., public, private, government, military, consumer goods). This tacit knowledge based on personal, experiential facets of the engineer's background shapes the engineer's perceptions towards not only whether a particular material is thematically relevant but also why it is relevant.

The engineering database of education material that is the basis for this study is an example of an engineering information source for which tacit knowledge would be a determinant for the type of materials users would choose to access. Compared to general public databases (e.g., Usenet), this database is thematically focused, uses technical terms, and contains material intended for audiences with specific backgrounds. The engineering database is designed for an engineering audience. The audience can be categorized as: researchers, educators, students, and practicing engineers. These sub-groups have different tacit knowledge. The researchers and the educators might prefer cutting-edge research and primary materials. Educators might be interested in materials to improve their curriculum that could be applied immediately to the classroom. The students may prefer secondary resources and supplemental materials. Finally, practicing engineers might only be interested in material that fits the special circumstances of their respective industry [9,10]. While these audiences may have similar thematic needs (e.g., search for information on "disk drive"), they differ in tacit knowledge and would likely prefer significantly different information sources even if each source were equivalently thematically relevant.

In summary, for the cognitive task of information seeking, tacit knowledge affects both the domain of inquiry (e.g., the subject area searched) and the procedures for searching the domain (e.g., selecting a set of known authors, subject headings, or keywords and appropriate documents from the result set). It has been proposed that tacit knowledge is implicitly revealed by the formation of the inquiry and the user's searching behavior such as reading time, saving, printing or downloading a document, or noting the citation. For example, Cooper [11] defines a binary judgment of document relevance in a Web-based library catalog based on whether or not during a search session the user saves, prints, mails, or downloads a citation. Cooper essentially assumes that when people engage in any sort of acquisition of information, their state of mind is to find only relevant resources that satisfy their information needs. The tacit knowledge may also implicitly lie in the materials themselves, characterized by, for example, the language (jargon) and accepted norms in document structure and format for presenting technical material and research. Landauer [12] suggests that the unconscious knowledge on which people rely manifests itself through the latent semantics of their written and spoken discourse. One study of practicing engineers describes how an engineer's personal data store reflects a personal preference for the types and sources of information, which vary depending on professional history (such as length of tenure at the company), education level, specialized knowledge, and computer skills [13]. In summary, the important distinction between the need for a known item and thematic information needs is that the satisfaction of thematic information needs bears heavily on the tacit knowledge of the information seeker. If one accepts the proposition that the tacit knowledge is tempered and shaped by an engineer's background and experiences, then the modeling of information needs must draw from available sources and expressions of tacit knowledge.

However, it is not necessarily straightforward for users to explicitly express their tacit knowledge vis à vis a query that expresses their information needs. One practical reason for this is that most information retrieval interfaces offer limited means for users to express their information needs aside from short phrases, pre-defined metadata fields, and relevance feedback, if at all [14]. Some studies have found that even when presented with enhanced features for expressing information needs, the "tyranny of the keyword" reigns; that is, users will often express information needs using just a few short phrases [15]. Thus, the ability to elicit tacit knowledge implicitly contained in the set of documents a user accesses and other behavioral patterns in accessing an engineering information database is key to successfully modeling information needs.

RELATED WORK

Identifying information needs with little or no direct interaction with the user is of interest to academic research and practitioners in information retrieval [16,17]. Various means have been explored for predicting information needs, particularly in the field of collaborative information filtering and recommender systems [18,19]. The primary distinction between information retrieval and information filtering is that information filtering deals with selecting information from a stream of data based upon a profile of information needs [20]. A key step in information filtering, then, is to develop a profile of the interests (information needs) of the user, either explicitly through active solicitation or implicitly. Of the numerous research results from research in collaborative filtering in the area of information needs modeling, two results apply directly to our methodology: (1) LSA has shown to be an effective tool for matching documents to information needs because it does not rely exclusively on keyword matching. (We discuss the technical details of LSA in the next section.) (2) "Document profiles," a method by which users indicate various information interests by grouping sets of documents they find relevant, are equally effective in explicitly profiling of interests through demographic data or keywords describing information needs [21]. (3) There exists some evidence of a correlation between reading time and the relevance or interest of the resource to the user [22,23]. This research extends these prior results to combine both LSA and reading time into a model of information needs. Our LIA methodology is demonstrated on a database of engineering courseware to show how to filter and recommend engineering materials.

LATENT SEMANTIC ANALYSIS

Existing full-text information retrieval techniques try to match query terms with words of documents literally. While these systems have been augmented with numerous enhancements in term weighting, authority linking, and heuristics to improve performance, these lexical matching methods can be inaccurate because users want to retrieve on the basis of conceptual content. There usually exist many ways to express a given concept. The literal words in a user's query or in the document may also have multiple meanings (polysemy), so individual words provided by the user may be unreliable evidence about the conceptual topic or meaning of a document and the user's information needs.

LSA (or Latent Semantic Indexing, LSI, in the information retrieval literature) [24] tries to overcome the problems of lexical matching by using statistically derived conceptual indices instead of individual words for retrieval. It assumes there is some underlying latent semantic structure in the data that is partially obscured by the randomness of word choice with respect to retrieval. It is a method for extracting the context-usage meaning of words using singular value decomposition (SVD) on a corpus of text represented in a word-by-document matrix. A truncated SVD is used to estimate the structure in word usage across documents. Retrieval is then performed using the database of singular values and vectors obtained from the truncated SVD. Performance data shows that these statistically derived vectors are more robust indicators of meaning than individual terms [25]. The underlying theory is that the totality of information about all the word contexts in which a given word does and does not appear provides a set of mutual constraints that determine the similarity of meaning of words and sets of words to each other [26]. SVD allows the arrangement of the space to reflect the major associative patterns in the data, and ignore the smaller, less important influences that are called "noise." As a result, words that did not actually appear in a document may still end up semantically "close to" the document, if they are consistent with the major patterns of word association in the data [24].

The LSA method has been applied to information retrieval [24] as well as to information filtering [21]. LSA was selected for our research as a tool for capturing tacit knowledge given its ability to find contextual similarity without relying on keyword or key phrase matching. One important empirical result from LSA research is that retaining up to the first 300 most significant singular values seems to sufficiently capture the "latent semantics" of the target document set [26]. How to choose the appropriate number of dimensions is still an open research issue. Ideally we want sufficient dimensions to capture all the real structure in the word-by-document matrix, but not too many, or we may start modeling noise or irrelevant detail in the data. In this research, we found that the first 200 singular vectors work best for our test data set.

LSA begins by forming a word-by-document matrix **X** in which the "t" rows of the matrix correspond to the unique words found in the documents and the "d" columns represent each document. The cells of the matrix represent the frequency that each word occurs in each document. A representation of this matrix is shown in Figure 1.

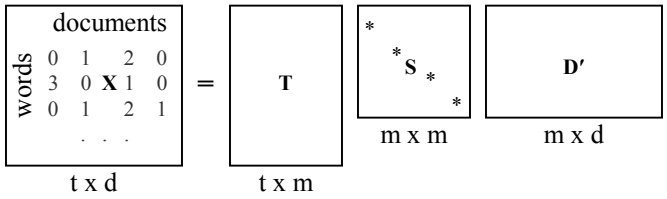

Figure 1 Schematic Representation of Singular Value Decomposition

The log-entropy weights are then computed for matrix X. The log entropy measurement is shown in Eq. (1) where $freq_{ij}$ is the frequency of occurrence of word *i* in document *j*.

$$\frac{\log(freq_{ij}+1)}{-\sum_{1-j}((\frac{freq_{ij}}{\sum_{1-j} freq_{ij}})*\log(\frac{freq_{ij}}{\sum_{1-j} freq_{ij}}))} \quad (1)$$

LSA then applies SVD to the log-entropy matrix **X** such that **X=TSD'** where **T** and **D** are the orthonormal left and right singular matrices and **S** is the diagonal matrix of singular values. It is possible to estimate the original word-by-document

matrix by truncating the SVD matrices into a set of k orthogonal factors, typically 100 to 300 as stated above. Instead of representing documents and words directly as vectors of independent words, LSA represents them as continuous values on each of the k orthogonal indexing dimensions, which we call the LSA dimensions. The dimension with the largest singular value is called 'LSA dimension one', the next largest is called 'LSA dimension two', and so on. It is then possible to visualize the placement of terms or documents in a two-dimensional geometric representation by plotting two dimensions from the rows of the **TS** and **DS** matrices, respectively.

A complete treatment of LSA, including an example of the dimensional reduction, may be found in Deerwester [24]; a more theoretical treatment by Laundauer [26]; and mathematical details for LSA are discussed by Berry [25].

INFORMATION NEEDS MODEL

The LIA approach in modeling information needs is based on both the query (the query used to retrieve a set of documents), the documents accessed during a user's session and user's historical search behavior. In essence, the methodology creates a space of documents that the user has previously accessed tempered by the length of time spent reading the documents. As discussed earlier, these documents and the user's behavior profile the users' tacit knowledge. The resulting information needs can then be modeled as a weighted graph shown in Figure 2 containing nodes for users, queries, information needs, and documents with relevance (derived from reading time) as edge weights. In this graph, a user node is linked with equal weights to all information needs s/he has in a series of search sessions. Each submitted query is defined as one search session. An information need node is associated with the respective query string and all the documents downloaded during that session, with the link weights reflecting the degree of relevance of the document to the query and information need.

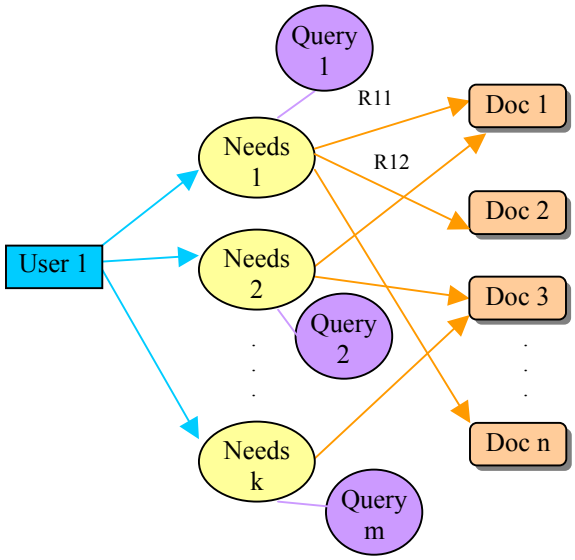

Figure 2 User Information Needs Model as Weighted Graph

The information need can then be expressed mathematically as a linear combination of the query vectors and document vectors in the k-dimensional LSA space as shown in Figure 3. In equation form, the information need **Q** is given by:

$$\mathbf{Q} = \omega_0 \mathbf{q} + \sum_{j=1}^{n} \omega_j \mathbf{d}_j \quad \text{in which} \quad \sum_{j=0}^{n} \omega_j = 1 \tag{2}$$

In Eq. (2), **q** is the query vector, which is simply the sum of the k-dimensional word vectors in the query string scaled by a factor of $\mathbf{S}^{1/2}$, where **S** is the decomposed singular value matrix (**S** is illustrated in Figure 1. For details, see Deerwester [24] and Berry [25]); $\mathbf{d}_j$ is the jth k-dimensional document vector among a total of n downloaded documents for that query and scaled by a factor of $\mathbf{S}^{1/2}$ as well, the degree of relevancy of each document to the information need is indicated by the weighting factor ω_j. The weight ω_0 and the sum of ω_j represent the degree of importance of the query string vector and the document vector, respectively, in reflecting the user's information need.

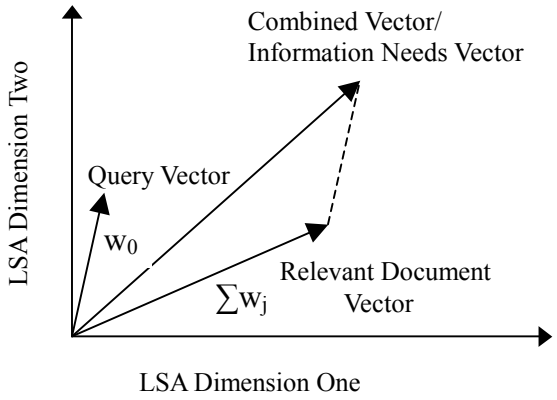

Figure 3 Vector Representation of Information Need

The value of each ω_j is defined according to the proportion of reading time spent on document j with respect to the total reading time over all n documents. For example, suppose a user downloaded 4 documents and spent 60, 25, 40 and 90 seconds reading each document, respectively. Then, each ω_j would be 60/215, 25/215, 40/215, and 90/215. However, since the sum of all the weights must equal 1.0, and no empirical or theoretical basis is available to assign ω_0, several validation experiments need to be conducted to investigate qualitatively the best ratio between the query string vector and the sum of document vectors. The results are discussed in the Information Needs Validation Study section.

This representation has several useful qualities. First, the model can be biased towards reading time or similarity to the query string by varying ω_j. This overcomes the issue of "cold start" of modeling information needs when the user has not accessed any documents. Another advantage is that users have long-term information needs as well as short-term ones. Long-term needs are usually stable while short-term needs may vary greatly even within a single user session. It is possible to observe long-term information needs by measuring the closeness of **Q** over time. If those vectors are close to each other, they may be considered as unchanged information needs.

That is, even though the query string might change over various sessions, the choice of documents accessed would indicate that the user still has the same "thematic" information need, though the user is now expressing that need in a slightly different manner.

THE LIA SYSTEM ARCHITECTURE

At the heart of LIA lies the search/recommendation engine that interacts with the user directly: accept user queries and provide relevant documents to the user by matching user's information needs with documents in the database. Pairwise similarity between individual document $\mathbf{d}_i$ in the database and user's information need $\mathbf{Q}$ is determined by a cosine measure, as shown by Eq. (3). The top N (specified by user) documents with greater scores are returned to users.

$$score(\mathbf{d}_i, \mathbf{Q}) = \frac{\mathbf{d}_i \cdot \mathbf{Q}}{\|\mathbf{d}_i\| \, \|\mathbf{Q}\|} \tag{3}$$

To implement the search/recommendation engine we have two modules as can be seen from Figure 4, one to build the LSA space, which is done offline and updated periodically due to the large computing requirements, and the other to construct user search sessions by data mining user transaction logs. These two modules function independently, but results from both are used to construct a user information needs vector in k-dimensional LSA space.

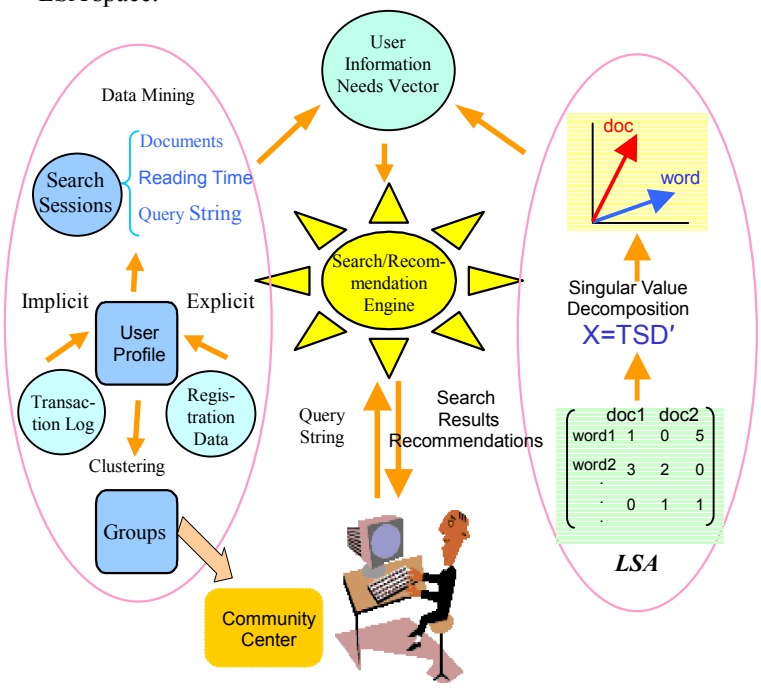

Figure 4 LIA Architecture Overview

TEST BED

Our LIA methodology for modeling information needs was applied to the NEEDS (National Engineering Education Delivery System) digital library (http://www.needs.org), an online repository containing digital education material related to engineering. NEEDS provides Web-based access to a database of learning resources where the user (whether they be learners or instructors) can search for, locate, download, and comment on resources to aid their learning or teaching process. There are a total of 1,889 engineering courseware modules in the database covering all the disciplines in engineering with no bias towards a particular discipline and 2,500 registered users as of the date the data for this study were collected. All user activity within NEEDS is automatically logged; each action is associated with unique transaction and session identifiers. Users may choose to be anonymous although quite a few of them are registered users, that is, they have a demographic profile stored. None of the demographic data was used in the information needs model described herein. The data utilized for this study were collected from May 1999 to December 2000. More than 17,000 user sessions were recorded and analyzed over this period. Reading time was extracted from a transaction database, which recorded the amount of time a user spent reading a summary of the document before downloading the document. As it is impossible to directly observe actual reading time in a distributed Web-based environment, we used this value as an approximation of the actual amount of time a user would spend reading the actual document.

We constructed the LSA space using the text from 1,889 courseware summaries in the NEEDS database. Using a custom-written text processing tool to remove stop words and words that occurred only in one summary, a 5951x1889 word-by-document matrix was formed. After applying SVD to this matrix, k-dimensional vectors in LSA space for both the word and document (courseware record) were generated. Finally information needs vectors were computed for all users.

INFORMATION NEEDS RESULTS

Query String	Number of Queries
Design	555
Electronics	136
Mechanical	104
Thermodynamics	90
Control	88
Heat transfer	76
Case study	76
Mechatronics	62
Electrical	57
Chemical Engineering	56

Table 1 Top 10 Query Strings in NEEDS Digital Library

Transaction log analysis as applied to NEEDS yielded the following aggregate data on users' usage pattern. It revealed that the average number of search sessions in a user visit is 1.8

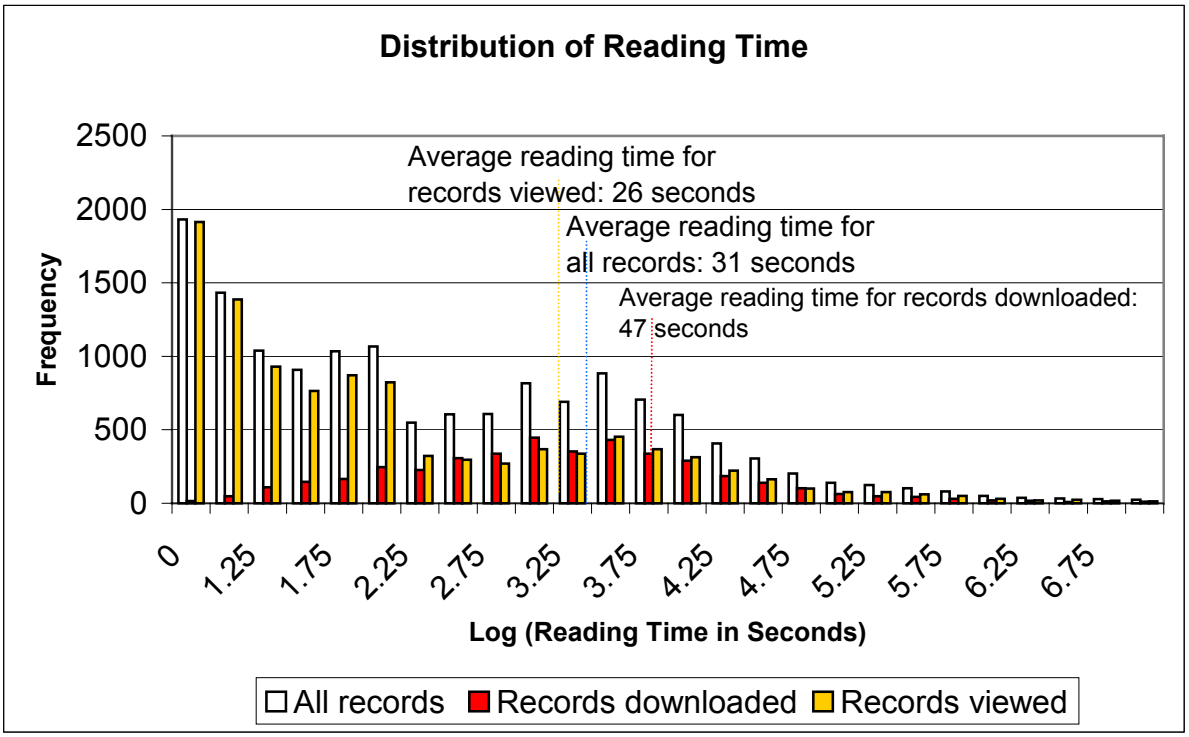

Figure 5 Histogram of Log Reading Time

(minimum 1, maximum 45, standard deviation 1.7). Users usually employed only one or two words in a query. Table 1 shows the top ten query strings. As expected, they viewed more material summaries than downloaded materials. Within a search session, the mean number of records viewed is 2.8 (minimum 1, maximum 131, standard deviation 4.4) and the average downloads is 1.6 (minimum 1, maximum 36, standard deviation 1.8). Only downloaded records were regarded as relevant to a user's information needs and were used to profile the tacit knowledge.

From the statistical analysis on the collected data, we observed that log reading distribution for records viewed and downloaded is quite different. Figure 5 shows the histogram of log reading time for downloaded records, records viewed only, and both over multiple user sessions. Comparing the first two distributions, the hypothesis of reading time as a measure of the degree of relevance of the document to information needs is generally valid in the NEEDS collection. Document reading time for downloaded records overall followed a normal distribution with mean of 47 seconds and standard deviation of 3 seconds. In contrast, the average reading time for records that the user viewed but did not download is only 25 seconds with a standard deviation of 4 seconds. A tailed distribution with a heavily populated starting point exists for the distribution of the reading time for records just viewed. We conclude that a major factor that influences the time spent for the summary of a record is the preference of a user to the record. Other factors played only a very small role in affecting the reading time. Our analysis shows there is a very low correlation of less than 0.03 between the length of the summary (number of words) and the time to read it in NEEDS dataset. This probably indicates that not all articles are completely read. Morita and Shinoda [23] reported similar results.

To interpret the analysis performed by LSA geometrically, Figure 6 shows a two-dimensional view of the information needs vector **Q** for all NEEDS users across multiple search sessions. The resulting information need is a k-dimensional vector in LSA space. The coordinates for the information needs vectors, represented by the dots, are plotted in a two dimensional space using the second and third LSA dimensions corresponding to the second and third largest singular values. The first dimension is the most dominant singular value and is typically regarded as representing the "average concept" whereas the second and third dimensions capture the meaningful distinctions among vectors represented in LSA space. Figure 6 plots the second and third LSA dimensions. The location of an information needs vector reflects the semantic position of an individual information need in LSA space. By calculating the cosine or dot product of a query or a document to the user's information needs vector, we can determine the similarity between them. This plot illustrates some interesting aspects about the NEEDS audience. First, while individual users searched for materials on various engineering subjects and disciplines, on the average, there appears to be a single information need shared by the entire audience, perhaps due to a tacit knowledge of going to NEEDS to find engineering education material or only to search for engineering material (rather than say sports scores). This is to be expected since the average document in the NEEDS collection deals with engineering. Second, a few outlying clusters suggest the

existence of information needs in other areas, possibly some of the smaller collections in NEEDS such as mathematics and life sciences.

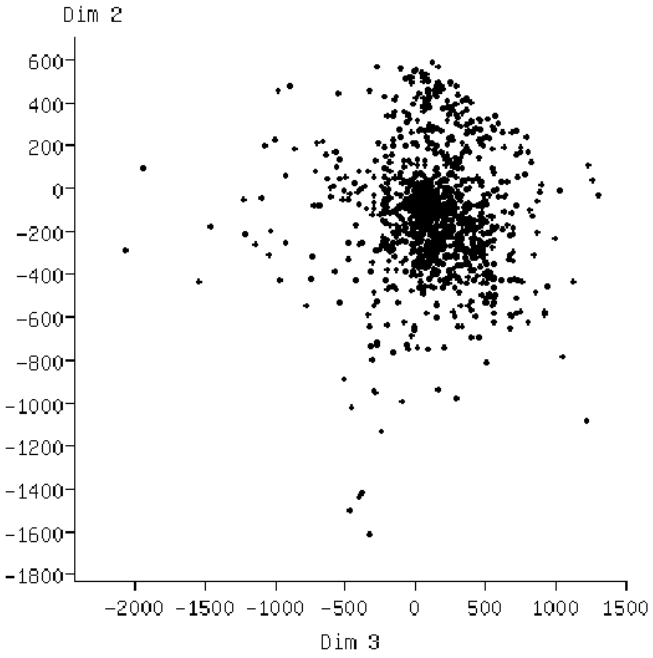

Figure 6 Vector Representations of Information Needs

We analyzed the components of the information needs vector for individual users. Figure 7 shows the relevant document vectors and query vectors for an arbitrary user across a series of search sessions. In this figure, filled circles represent the semantic location of query vectors in LSA space, filled rectangles represent the semantic location of courseware record vector, and lines link the query vector and the record vector the user downloaded for that particular query. The information needs vectors (not plotted) are merely the weighted sum of a query vector and its corresponding relevant record vectors.

One can observe apparently changing subject searches of the example user in Figure 7. This user had 11 search sessions; a total of 14 records were viewed across the 11 sessions, 9 records with records identification numbers d165, d90, d341, d313, d819, d87, d824, d323, and d180 were downloaded using the same query string (control system) and 2 other records d233 and d237 were downloaded using different query strings "quality" and "artificial intelligence" respectively. Among the 11 search sessions, it happened that this user downloaded one record per search session, 9 of them shared the same query string. While this user appears to have searched for resources about three separate concepts, if the queries are taken to be concepts, there is an underlying similarity in the documents downloaded. The documents this user chose to download dealt with intelligent control systems and their application to authentic design problems and were often of the case study format.

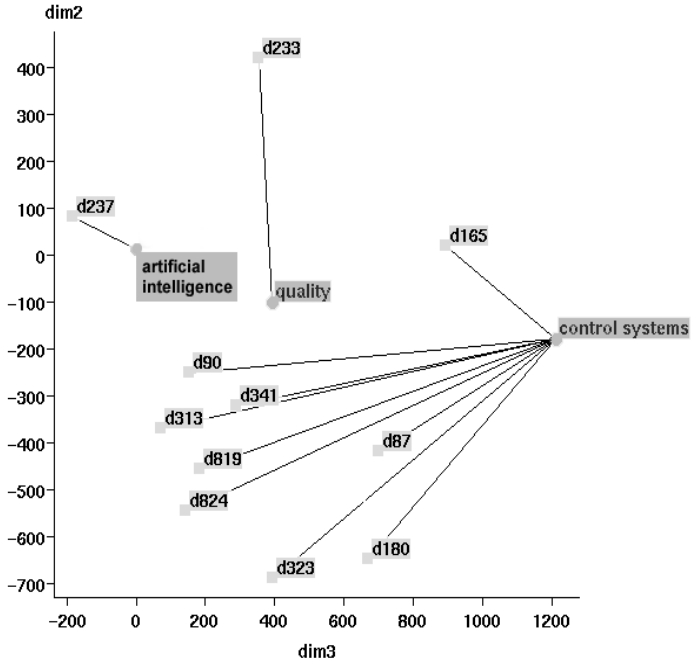

Figure 7 Information Needs of One User

We can see further evidence that although users might have used various query terms to search for documents, overall, they exhibited behavior to suggest a single information need. Figure 8 shows the information needs of five selected users of the NEEDS digital library. Users are represented by circles except the fifth user owning the scattered points without a label and a bounding circle. Points circled together show one user's information needs vector over multiple search sessions. The profiles of the fifth user's information needs were spread out (points without a bounding circle) with no clear center; others have a smaller radius. Clockwise from the top left, the thematic needs of these users, based on their query string and manual inspection of the downloaded records, were: *biophysics*, *stepper motors*, *printer design* and *fuel-injection systems*.

These results indicate that some users have well-defined information needs (and their needs do not seem to vary over time) whereas others have widely varying needs, as evidenced by the dispersion of points of information needs and the timestamp associated with each search session.

In general, user's information needs over a long period of time are hard to model accurately given their dynamic nature. We are currently working on identifying and modeling information needs shift. We are interested in observing information needs change for an engineering design team. Documents generated from a graduate level design course at UC Berkeley are used in this study. Our preliminary result shows there is a dominant information need in a design team, but in different design stages the team information needs vary from large deviation at the beginning (e.g. ideation stage) to a small deviation at the end from the dominant information needs. More detailed results will appear in Song [27]. The model

presented in this paper focuses on short-term information needs which are defined within one search session.

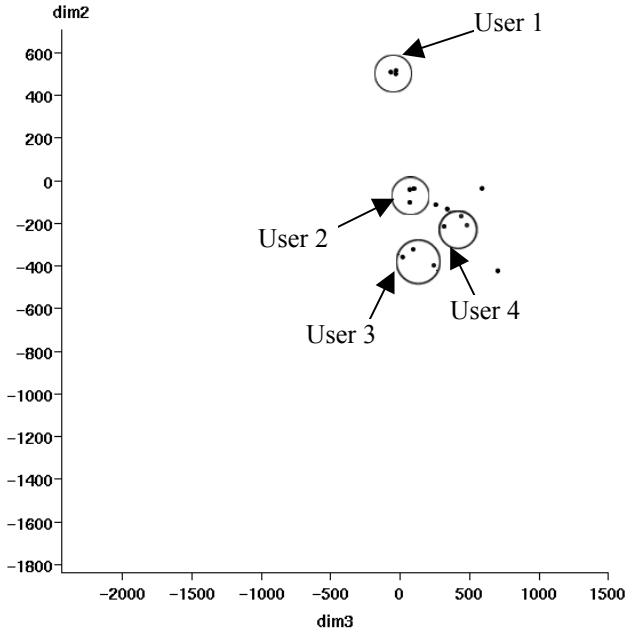

Figure 8 Information Needs of Multiple Users

INFORMATION NEEDS VALIDATION STUDY

A validation experiment was carried out to measure the accuracy of the LIA information needs model of Eq. 2 in predicting the observed reading time. Three comparison studies were conducted.

[1] Predicted reading time based on the query string only. This is the baseline case.

[2] Predicted reading time based on the information needs model of Eq. 2.

[3] Predicted reading time based on the relevant documents only. This is similar to the LSA document match profile described by Foltz and Dumais [21].

Sixty-three user sessions that had more than six courseware document records downloaded were used for the validation experiments. The documents in each session were then divided into two sets randomly, one training set and one validation set. The documents in the training set were used to model the users' information needs. That is, information needs vectors were generated for each user. Based on the LIA information needs model, the reading times corresponding to each document in the validation set were predicted. Both sets were derived from the user transaction logs.

Three weighting schemes for ω_0 and $\Sigma\omega_j$ were tested.

[4] $\omega_0/\Sigma\omega_j$ = 7:3 (query-biased)

[5] $\omega_0/\Sigma\omega_j$ = 1:1 (no bias)

[6] $\omega_0/\Sigma\omega_j$ = 3:7 (relevant document biased)

The primary goal in this validation study is to derive a relationship between the information needs and a given document in order to compute the predicted reading time for that document. We tested two techniques for deriving the relationship. The first technique is based on running a linear regression model to relate the information needs satisfaction (cosine of the information needs vector and the relevant document vector) and reading time. The second technique was based on comparing the probability distribution of the log reading time to the probability distribution for the information needs satisfaction.

Figure 9 shows the plot of log reading time and information needs satisfaction. A linear regression showed a linear relationship with a correlation coefficient of 71% and a standard error of estimation of 3%.

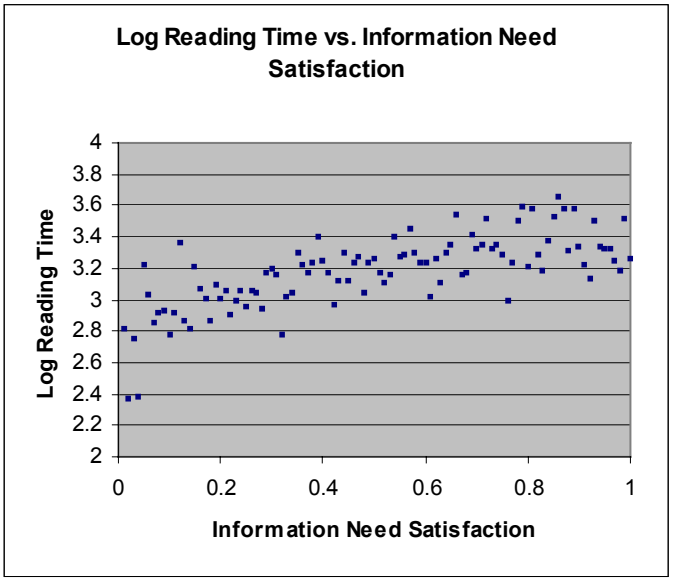

Figure 9 Scatter Plot for Linear Regression Model

The underlying assertion for the second technique is that the probability distribution for the log reading time is equivalent to the probability distribution for the satisfaction of a user's information need. To verify this assertion, a chi-squared test compared the distributions for log reading time and cosine of relevant document to the information needs vector. The test revealed a chi-squared value of 30.04 and a p-value of 0.09 with 21 Degrees Of Freedom (DOF), which tells us that the probability of observing a chi-square value at least as large as the observed value is significant at the 0.05 level. Thus, the two distributions are similar to a statistically significant degree.

The assertion of equivalence of distributions makes possible the prediction of reading time from the information needs vector. Given the equivalence of the two distributions, to predict reading time, one only needs to super-impose the two distributions upon each other such that the sample mean and deviation match. Then, given a probability of information needs satisfaction, the log reading time can be located.

The last step in this experiment is the calculation of correlations between the predicted and observed vectors using both prediction models. The Pearson correlation coefficient is used in this study to calculate the coefficient between predicted reading time and observed reading time.

Figure 10 shows the results for predicted reading time versus actual reading time using both techniques. The best correlation appeared (B) when both the query string vector and document vector were used to model information need. It illustrates there is a statistically significant positive relationship between the predicted reading time and actual reading time with moderate correlation coefficient around 0.36 and p-value less than 0.001. Similar results can be found in the Morita and Shinoda [23] study on correlating reading time and estimated document interest. To our knowledge, the Morita and Shinoda study is the only other thorough study on reading time correlations based on an estimation of document interest. A correlation coefficient of 0.49 is deemed a strong correlation in their paper. Their study was conducted under a well-controlled environment in which the reading times were recorded precisely and the subjects in the experiment read documents in their entirety without distractions, such as leaving the terminal or reading newly arrived e-mail. Thus, one would expect better correlation results than in our study. Another study by Konstan et al [19] also shows similar results; there is a high correlation between reading time and ratings of a Usenet new article, but the correlation is not calculated between reading time and estimated/predicted ratings, but reading time and explicit ratings. Given that, in our study, we could only estimate the reading time and the fact that the users of the digital library were free to behave in their habitual manner, we think that the correlation level found by our method is comparable in its ability to predict reading time based on an estimated document interest (information needs).

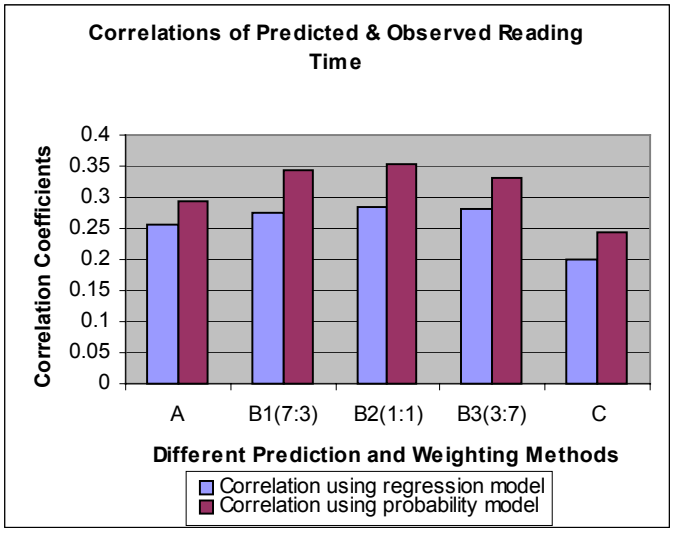

Figure 10 Results of Validation Studies

Figure 10 also shows the correlation was always slightly better than that obtained using only the query string (A) and substantially better than that obtained using only the relevant documents (C). Different choices of the weighting between the query vector and the relevant document vectors influenced the predictions of reading time. There is evidence to suggest that weighting the information needs biased towards the query string, but not comprised entirely of the query string, performs better than the converse. This test also reveals that the probabilistic model outperforms the regression model although the difference is not very large.

In practice, reading time depends on many factors including the difficulty of including the desired information within the document, the bandwidth of the user's Internet connection, and the constraints of the user's work environment [22]. Although these moderately positive correlations are encouraging, given what is known about reading time, there is a clear need for further study. Our results, however, provide important insights into the design of engineering information systems.

RECOMMENDER SYSTEM

Based on the information needs model, a recommender system was designed and implemented in the SMETE Digital Library, a next-generation version of NEEDS that includes educational material spanning science, mathematics, engineering and technology education (SMETE). The system attempts to address the contextual nature of user information needs by automatically recommending relevant records to users based on their past search sessions. They are generated by recommending the documents with the highest predicted reading time based on the information needs of the particular user. Figure 11 is a snapshot of the *My Interests* page from SMETE. Qualitative evaluation studies are underway to ascertain the usefulness of the recommendations to actual users.

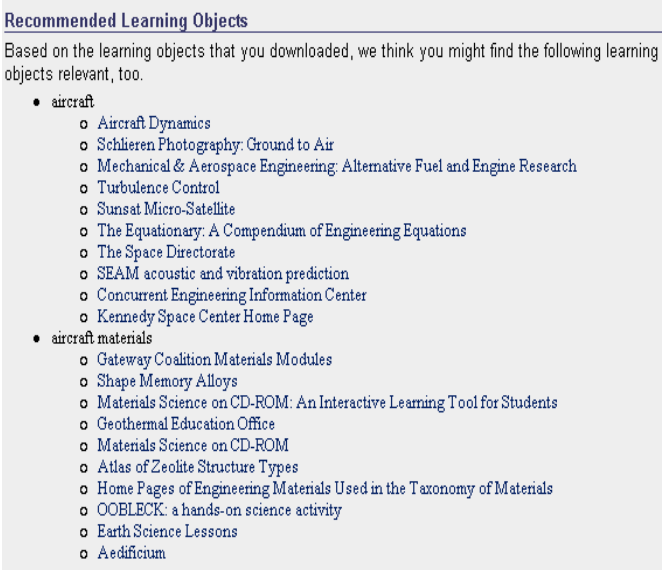

Figure 11 Recommended Resources

CONCLUSIONS

This is the first rigorous study in the field of modeling information needs using tacit knowledge in an engineering environment. It has established a basic methodology for identifying and modeling user information needs implicitly in online resources. The information needs model, LIA, based on LSA for modeling implicit needs, coupled with the analysis of user information-seeking behavior stored in user transaction

logs, is an effective method in discovering and modeling user information needs, especially in making "tacit" knowledge "explicit".

Results from the study on the NEEDS digital library, which has more than 4,000 users from the engineering community, show this information needs model is effective in modeling their information needs. A validation study, based on predicting the information needs of a target population given a training population, quantified the model's accuracy.

Archiving design documents in digital form is becoming of increasing importance to industry, e.g., Boeing's major investments in 100% digital design and documentation [28]. And not surprisingly, engineers tend to use internal technical reports more often than externally published materials [10]. However, engineers do not effectively utilize design documentation. According to Hertzum's study [29], when an engineering designer needs information or inspiration about how to accomplish a new design, for example about how to manipulate a stainless steel tube in a certain way, s/he turns to her/his close-by colleagues or walks up to the R&D lab and talks to the people there. The engineer does keep and consult a handful of textbooks and a select number of internal reports and other documents but finds that in general s/he makes more use of product specification sheets than actual text. The problem is engineering design is a complex task; as task complexity increases, so does the complexity of the information needed by the engineers. And design choices made by engineers depend, to a large extent, on their understanding of the context of the task and, consequently, on their success in obtaining information about this context. Current Product Data Management (PDM) products or other type of information management systems are not well tailored towards satisfying these particular information requirements of engineers/designers. A more flexible information retrieval system able to meet different information needs in various contexts is the key towards better serving an engineer's information needs and her/his use of digital documents. The methodology used in this paper could be useful to fulfill this purpose by integrating LIA into PDM systems. Given the immense practical importance of information retrieval in engineering design, engineering information management systems should include functionality to manage expertise and design knowledge rather than just purely recording engineering design data.

Future research will improve the methods described here including further investigation on long-term information needs shift, and the effect of explicitly identifying non-interesting documents into the information needs model. We are also working on evaluating our methods in an engineering design environment where designer's information needs might be influenced by different stages of the design life cycle.

ACKNOWLEDGMENTS

The authors would like to thank Prof. John Canny in the Computer Science Division and Eric Fixler, Jia-long Wu, Phillip Leal, Andrea Swafford, Brandon Maramatsu, and Flora McMartin at NEEDS/SMETE for their assistance and suggestions in this study. This research was funded, in part, by the National Science Foundation under Grant #DUE-0085878 for the National SMETE Digital Library program.

REFERENCES

[1] Court, A.W., Culley, S.J., and McMahon, C.A., 1997, "The Influence of Information Technology in New Product Development: Observations of an Empirical Study of the Access of Engineering Design Information," *International Journal of Information Management*, October 1997, Vol. 17, No. 5, pp. 359-375.

[2] Ahmed, S., Blessing, L.T.M., and Wallace, K.M., 1999, "The Relationships between Data, Information and Knowledge Based on A Preliminary Study of Engineering Designers," *Proceedings of ASME Design Engineering Technical Conferences*, Las Vegas, Nevada, September 12-15, pp. 121-130.

[3] Szykman, S., Bochenek, C., Racz, J. W., and Sriram, R., 2000, "Design Repositories: Next-Generation Engineering Design Databases," *IEEE Intelligent Systems and Their Applications*, May-June 2000, Vol. 15, No. 3, pp. 48-55.

[4] Dong, A., and Agogino, A.M., 1997, "Text Analysis for Constructing Design Representations," *Artificial Intelligence in Engineering*, Vol. 11, pp. 65-75.

[5] Wood III, W. H., and Agogino, A.M., 1996, "A Case–based Conceptual Design Information Server for Concurrent Engineering," *Journal of Computer Aided Design*, Vol. 28, No. 55, pp. 361–369.

[6] Vaughan, Anthony (Ed.), "International Reader in the Management of Library, Information and Archive Services," General Information Programme and UNISIST, UNESCO 1987, URL: http://www.unesco.org/webworld/ramp/html/r8722e/r8722e00.htm.

[7] Polanyi, M., 1962, *Personal Knowledge: Towards a Post – Critical Philosophy*. Chicago, University of Chicago Press.

[8] Bucciarelli, L., 1994, *Designing Engineers*, Cambridge, MA: The MIT Press.

[9] Gould, C., and Pearce K., 1991, "Information Needs in the Sciences: An Assessment". Mountain View, CA: The Research Libraries Group, Inc.

[10] King, D., Casto, J., and Jones, H., 1994, "Communication by Engineers: A Literature Review of Engineers' Information Needs, Seeking Processes, and Use". Washington: Council on Library Resources.

[11] Cooper, M.D., and Chen, H.M., 2001, "Predicting the Relevance of a Library Catalog Search," *Journal of the American Society for Information Science and Technology*, Vol. 52, No. 10, pp. 813-827.

[12] Landauer, T., Laham, D., and Foltz, P., 1998a, "Learning Human-like Kowledge by Singular Value Decomposition: A Progress Report," *Advances in Neural Information Processing Systems* 10, MIT Press: Cambridge MA, p. 45-51.

[13] Lowe, A., McMahon, C., Shah, T., and Culley, S., 1999, "A Method for the Study of Information Use Profiles for Design Engineers," *Proceedings of ASME Design*

Engineering Technical Conferences, Las Vegas, Nevada, September 12-15, pp. 109-119.

[14] Belkin, N.J., Cool, C., Kelly, D., Lin, S.J., Park, S. Y., Perez-Carballo, J., and Sikora, C., 2001, "Iterative Exploration, Design and Evaluation of Support for Query Reformulation in Interactive Information Retrieval," *Information Processing & Management*, Vol. 37, No. 3, May 2001, pp. 403-434.

[15] Yuan, W. J., 1997, "End-user Searching Behavior in Information Retrieval: A Longitudinal Study," *Journal of The American Society for Information Science*, March 1997, Vol. 48, No. 3, pp. 218-234.

[16] Nichols, D., 1997, "Implicit Ratings and Filtering," *Proceedigns of the 5th DELOS Workshop on Filtering and Collaborative Filtering, Budapest*, Hungary 10-12, ERCIM

[17] Kim, J., Oard, D., and Romanik, K., 2000, "Using Implicit Feedback for User Modeling in Internet and Intranet Searching," Technical Report, College of Library and Information Services, University of Maryland at College Park.

[18] Resnick, P., and Varian, H., 1997, "Recommender Systems," *Communications of the ACM*, Vol. 40, No. 3, pp 56-58.

[19] Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., and Riedl, J., 1997, "GroupLens," *Communications of the ACM*, Vol. 40, No. 3, pp. 77-87.

[20] Belkin, N.J., and Croft, W.B., 1992, "Information Filtering and Information Retrieval: Two Sides of the Same Coin?," *Communications of the ACM*, Vol. 35, No., 12, December 1992, pp. 29-38.

[21] Foltz, P., and Dumais, S., 1992, "Personalized Information Delivery: An Analysis of Information Filtering Methods," *Communications of the ACM*, Vol. 35, No. 12, pp. 51-60.

[22] Cheng, H, 2000, "A Probabilistic Model for Mining Tacit Knowledge for Information Retrieval," M.S. Thesis, Computer Science Department, University of California at Berkeley.

[23] Morita, M., and Shinoda, Y., 1994, "Information Filtering Based on User Behavior Analysis and Best Match Text Retrieval," *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY: ACM, pp. 230-237.

[24] Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R., 1990, "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*, Vol. 41, No. 6, pp. 391-407.

[25] Berry, M., Dumais, S., and O'Brien, G., 1994, "Using Linear Algebra for Intelligent Information Retrieval," Technical report, Computer Science Department, The University of Tennessee, Knoxville, TN.

[26] Landauer, T., Foltz, P., and Laham, D., 1998b, "An Introduction to Latent Semantic Analysis," *Discourse Processes*, Vol. 25, pp. 259-284

[27] Song, S., Dong A., and Agogino A., 2002, "Time Variant Analysis of Information Needs of Engineering Design Teams," Working paper #02-0901-1, Berkeley Expert Systems Technology Laboratory, UC Berkeley, 2002.

[28] Malhotra, A., Majchrzak, A., Carman, R., and Lott, V., 2001, "Radical Innovation Without Collocation: A Case Study at Boeing-Rocketdyne," *MIS Quarterly* Vol. 25, No. 2, pp. 229-249.

[29] Hertzum, M., 1999, "Managing Expertise: The Fundamental Importance of Trust in People's Assessment and Choice of Information Sources". Presented at the workshop on Beyond Knowledge Management: Managing Expertise (organised by M.S. Ackerman, A.L. Cohen, V. Pipek, and V. Wulf) held at ECSCW'99: *The Sixth European conference on Computer Supported Cooperative Work* (Copenhagen, Denmark, September 12-16).