

## AUTOMATIC COMPOSITION OF XML DOCUMENTS TO EXPRESS DESIGN INFORMATION NEEDS

Andy Dong, Shuang Song, Jialong Wu, and Alice M Agogino

*Keywords: Information analysis, classification and retrieval; information representation; design information management*

### 1 Introduction

Engineering design is an information intensive activity. It is reported that designers spend in excess of 50% of their time in handling information [7]. Thus, the efficiency and the quality of the design process depend considerably on how well designers are able to handle large amounts of information. One study [8] of the information required by design engineers to complete their jobs indicated that less than 50% of that information was actually available and only 20% could be provided by the existing specialized applications. Directing the right information to the right person at the right time is a complicated but crucial task.

Design information management has received increasing attention in recent years as a result of these findings and the recognition that lacking sufficient or missing key design information may lead to sub-optimal decision-making and design [2,4]. Much of the existing research has focused on the capture, storage, indexing and presentation of design information including informal information [3,9]. Less work has been done on information retrieval based on an understanding of individual designers, their experience, their skills and the ways in which they use information in the context of their design task [1]. One key step in finding the right information is expressing information needs in context. This paper presents a methodology to generate an XML ([eXtensible Markup Language](#)) document that expresses the information needs of a design engineer. XML documents and their underlying Document Type Definition (DTD) offer an efficient structure for the organization of design information [5] and representation of information needs. Through declarations of XML entities and the inherent structural hierarchy of XML documents, XML documents can express the designer's information needs while framing the design's structural hierarchies. For example, an element in the DTD may permit the design engineer to express a preference for formal company documents of past designs (e.g., technical memos) rather than informal design notes. The data in the XML document may be drawn from a repository of semi-structured or unstructured text documents that the design engineer has retrieved and placed into a personal information store by using an information retrieval system.

Our methodology draws from the computational linguistics techniques of natural language processing and latent semantic analysis (LSA). We assume there exists some underlying information needs that are expressed by the type of information the engineering designer wishes to view and download into a personal information store. The "type of information" will be distinguished primarily by subject but may include other identifiers such as the format of the information and the intended audience. By applying these linguistic techniques, we can

construct an XML document that is descriptive of this underlying need and contains information directly from the documents that is consistent with the major patterns of information preferences of the designer.

## 2 Methodology

### 2.1 Technical Approach

The methodology for automated composition of the XML documents proceeds along two axes: 1) *explicitly solicit information needs* from the designer through standard information retrieval means; 2) *implicitly monitor the information retrieval behaviour* of the designer to information sources including the type, quality and information contained in the documents the designer chose to retrieve. We test this methodology on access to unstructured engineering data, such as full-text, because unstructured, textual documents are the principal mode of communication by engineers. Before proceeding to a detailed discussion of our methodology, we discuss two core technologies to the methodology: mining of transaction logs to learn information needs implicitly and computational linguistics.

### 2.2 Learning Information Needs Implicitly

Many difficulties exist in determining what information a person wants to see as well as modelling user information needs. Most information retrieval systems require that people express information needs through a set of keywords or key phrases. However, ascertaining information needs simply by a word or two is inadequate and subject to loss of contextual information. For engineering design, studies have shown varying information needs of designers depending on level of expertise and stage of the design [1,6]. Our approach in modelling user information needs is based on a human-centred computing. Our methodology examines the user's document access patterns, that is, the user's personal information store and transaction history in an information retrieval system, for patterns of information preference. The basic approach is to examine the user's session information over all sessions while using the information management system. In learning information needs implicitly, we are primarily interested in discovering the similarity between documents that the user has downloaded into a personal information store. The assumption is that the user's particular choices of documents to store locally are indicative of information needs. Thus, instead of requiring the designer to *a priori* categorize the information, the system attempts to learn a similarity mapping using contextual clues such as project name, engineering discipline, and document format. Similar categorisation strategies have also been found in the classification of supplier information practised in industry [12].

To ascertain relevance, the system records each document downloaded into the user's personal information store. This is an accurate indicator of relevance because, in our system, before the user may download the document, the user has already read a brief description of the content of the document containing meta-information such as abstract, author, document type, and subject. The assumption is that during any single session utilizing the information retrieval system, the user has a dominant goal (information need) in mind that is expressed by the type(s) of documents the user decides to download. Similar assumptions exist in other studies [13] of information retrieval systems.

We use the vector space approach [14] for document and query representation. We analytically modelled information needs as a linear combination of the vectors representing

the query and the relevant document. The weighted vector average (arithmetic mean) of all combined vectors consisting of the query and relevant documents is called the “centroid” of the user’s intended information needs. The centroid is then used as a representation of the dominant information need of the user. The maximum angle between the document vectors or query string vector and the centroid is used as an indication of the variation in the user’s information needs.

## 2.3 Computational Linguistic Approaches

Our methodology employs two computational linguistics techniques, natural language processing and latent semantic analysis, to extract and summarize the primary topic of a set of similar documents.

### *Natural Language Processing*

While we use the designer’s past transactions as an indication of information need, the structured information contained in the documents that are referenced in the transactions needs to be discovered and refined before it can be utilised effectively. The key phrase retrieval process helps in crosschecking the indexed subjects and compacting the size of the aggregated XML document by representing paragraphs of text with just a few representative noun phrases. As the designer adds documents to a personal information store, we do not need to concatenate all the textual information about the document to the XML document, just the extracted noun phrases that are not already in there and plug them under the appropriate tags. By identifying the noun phrases in the document, we are able to find the corresponding contents for each XML tag in the text. Additional rules and procedures are needed other than standard noun phrase retrieval process to perform this task for all tags. This latter component forms a part of future research.

Extracting from the full-text content-bearing noun phrases documents that can be used in profiling and indexing involves 3 steps: tokenization, part-of-speech (POS) tagging and noun-phrase identification. Tokenization is a procedure that identifies sentence boundaries and removes extraneous punctuations. POS taggers then take the processed corpus and tag each word with their POS information. Taggers that operate following semantic rules or just statistical information were developed. After the text corpus has been tagged with POS information, we could use the contextual information to identify noun phrases. The extracted noun phrases are then attached to the corresponding DTD elements of the document.

### *Latent Semantic Analysis*

Latent Semantic Analysis (LSA) [9] is a statistical model of word usage that permits comparisons of semantic similarity between pieces of textual information. The idea is that the totality of information about all the word contexts in which a word does and does not appear provides a set of mutual constraints that largely determines the similarity of meaning of words and sets of words to each other. The primary assumption of LSA is that there exists an underlying or “latent” structure in the pattern of word usage across documents. LSA uses the matrix technique of singular value decomposition (SVD) to reflect the major associative patterns of words in the document and to ignore the smaller influences. The ability for LSA to remove the obscuring “noise” makes LSA useful as an analytical tool for discovering the primary conceptual content of documents. We use LSA to help us to categorize and group the documents by topic material that the user has downloaded and revealed as relevant and useful. Once these principal groupings are identified, we can then apply the rest of our methodology

to express information needs for each “centroid” of documents within a semantic locality through a single XML document.

## 2.4 Composing XML Documents

The system initiates by asking the user to specify an XML DTD containing metadata elements (XML entities) that express information needs. For this study, we asked that users express their information needs using a well-known metadata set, the IEEE Learning Object Metadata (LOM) [10], and in particular the core 20 elements. The LOM defines the information required to manage, locate and evaluate learning resources. Having a standards-based XML DTD allows us to assess our methodology’s completeness and efficacy and for comparison against other systems. Because of the analogous human cognitive processes of information processing and learning [11], the LOM serves as a reasonable model for describing information needs.

Once the user has specified an XML DTD, the user must now utilise the information management system, retrieving and downloading design documents. The system implicitly monitors the information retrieval transactions of the user, eventually formulating a seed XML document containing information from the user’s session. Typically, the XML document is seeded with the initial query the user posed to the system.

The implicit stage contains two phases. In phase one, the system applies latent semantic analysis to characterize the knowledge conveyed by the all documents the person chose to view. To perform this phase, we analysed the transaction logs containing the transactions of both the current user and all other users of the system. The user’s query and document(s) downloaded are recorded for each visit. The latter information identifies the relevant documents necessary for phase two. Using a similarity measurement, the system identifies topics represented by the documents the user viewed. By doing this step, the system ascertains the topic locality of the various documents the person viewed. This is a critical step because the user may have multiple and widely varying information needs. Then, for each topic locality, the system augments the XML document expressing the user’s information needs with the metadata elements from each relevant document. In practice, the information management system will contain most of the information required to complete the tagging such as Author, Title, Date, and Format. Subject and Description information are generated automatically from the second phase. This matching can be done by exact one-to-one correlation, i.e., both the tag and attribute match, or via a crosswalk between the information about the document contained in the information management system and the XML DTD. Both of these techniques will have required some prior means of tagging the documents in the information management system.

In the second phase, we apply natural language processing techniques to ascertain the principal subject of the documents within each topic as discussed in Section 2.3. The process repeats until all possible elements in the user’s original XML DTD are filled, resulting in a fully marked up XML document. The completed DTD is now an expression of the information needs of the designer, based upon the available information stores and the pattern of information retrieval undertaken.

In summary, a designer’s information needs can be found implicitly by looking at the set of documents the designer has deemed relevant and useful and stored in a personal information store. We use LSA to find signatures of similarity in this set of documentation. Once we find the signatures, we look at what the original information needs were to find the centroid of the

similarity. Finally, we construct a compound XML document that essentially reconstructs the full LSA space by combining information from all the similar documents, with some of the information filtered through NLP techniques to reduce the size of the document. Other information about the document indexed in the document database, such as document type, is added to the corresponding XML tag. This final XML document is then an expression of the designer's information needs.

### 3 Experimentation and Results

#### 3.1 Test Case

The experiment and prototype evaluation was conducted on a digital library project for science, math, engineering and technology education. Students and educators use the digital library to download courseware into their personal information stores. The documents used in the study discuss the design of engineering devices and related scientific theories. The users of the system typically search for material on engineering education. This is their primary information need.

#### 3.2 Results

First, we validated the ability of our methodology to discover information needs. We conducted this study by analysing for the known information needs of all users of the digital library. Based on the fairly homogeneous content of the digital library (courseware on engineering design) and the known profile of the audience of the digital library, we expected that our methodology would reveal one dominant information need, namely courseware related to engineering education. We would not expect for the system to reveal numerous distinct clusters of information needs. We ran the latent semantic analysis over the entire usage database. Figure 1 illustrates the distribution of all users' information needs over multiple sessions.

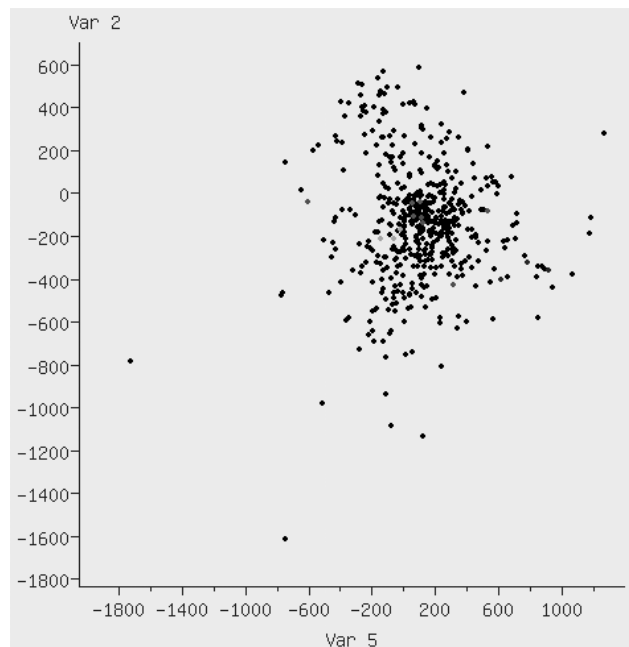


Figure 1. Information Needs Represented in LSA Space

Each dot represents in LSA space the combined vector of the user's query and a downloaded document for one session. By visual inspection, one can note one dominant information need. Specifically, the most commonly downloaded documents by all users of the system were case studies and courseware on the design of disk drives, a specific subset of engineering education. This result corroborates the known information needs of users of the database.

Second, we analysed for the information needs of individual users. Figure 2 illustrates the information needs of one sample user, specifically information on "control systems". The circle represents, in LSA space, the initial query to the information retrieval system whereas the boxed numbers indicate the documents downloaded by the user. One can then apply latent semantic analysis to ascertain the similarity between downloaded documents, the original query, and the documents themselves. Users may have multiple information needs despite using the same keyword to query the information retrieval system. In the example shown, document 165 is relevant to the user's query but not similar to the other documents, therefore potentially indicating different information needs. For this document set, we found that an angle of 71° between documents provided an adequate measurement of similarity. Based on the set of similar documents, we computed the centroid.

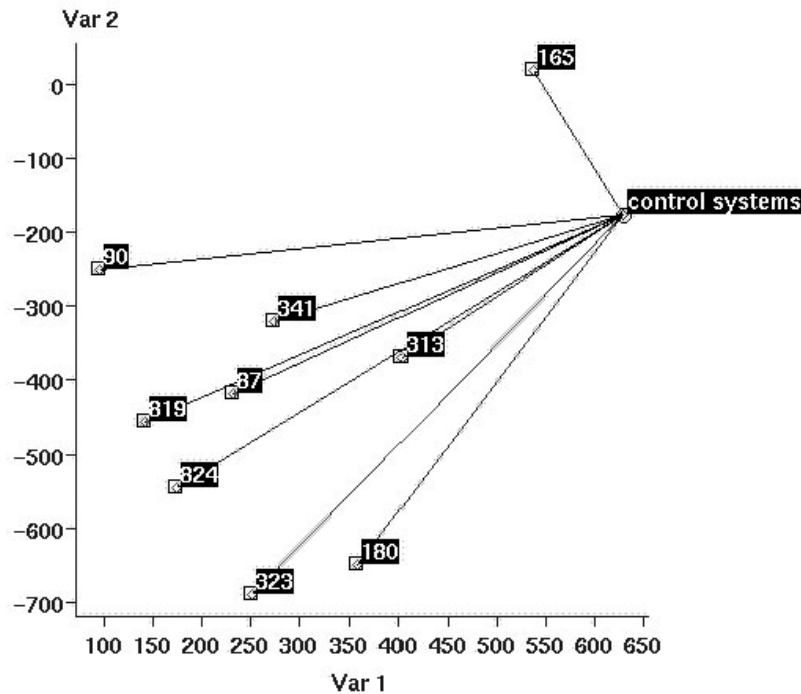


Figure 2. One User's Information Needs

Finally, we generated the XML documents to express the information needs. Portions of an XML document are illustrated in Figure 3.

```

<!-- The Core IMS Learning Object Metadata in XML, a subset of the IEEE LOM V3.5. -->
<metametadata>
  <metadatascheme>IEEELOM:1.0</metadatascheme>
  <language>en-US</language>
</metametadata>
<general>
  <title>
    <langstring>
      educational software
      engineering graphics tutorials
      engineering visual encyclopedia
  
```

```

        mechanics
        virtual disk drive design studio
    </langstring>
</title>
<language>en-US</language>
<description>
    <langstring>
        acme disk drive company
    </langstring>
</description>
<lifecycle>
    <contribute>
        <role>
            <langstring lang="en">Author</langstring>
        </role>
        <centity>
            BEGIN:vCard
        </centity>
    </contribute>
</lifecycle>
</document>
<?xml>

```

Figure 3. Sample XML Document

The XML document expresses out in human-readable format a summary of the user's information needs as an aggregation of the documents that the user found relevant and useful and that were related to each other.

## 4 Conclusions

This research has established a basic framework for identifying and modelling engineers' information needs using XML documents. We performed latent semantic analysis over a collection of engineering resources to construct information needs as vectors in LSA space based on usage analysis. We visualized different information needs in multi-dimensional space. Based on a cluster of similar documents representing an information need, we generated an XML document using natural language processing techniques to express the information need.

These results are encouraging. They show that latent semantic analysis can be applied to the task of ascertaining information needs by monitoring the information retrieval habits of a user. In order to assess the actual "truth" of the XML document in representing the user's information needs, we would need to have the user respond positively or negatively to suggested relevant information provided autonomously by the information retrieval system. We have projects in progress to incorporate this feedback. In addition, we are working on methods to incorporate reading time into the model of information needs and to predict the expected reading time of a document based on prior reading time.

We expect this methodology to impact the use of information in design in several ways. First, the XML documents can be used as information filters to direct critical pieces of information to the designer as others generate them. In addition, intelligent software agents might use the XML document as a guide to search document repositories for new, useful information. The methodology may provide insight into the cognitive states of the designer over various stages of design, offering a tool to study how changes in information needs relate to the designer's understanding of the design problem. We are currently analysing the effect of time on information needs, particularly the rate of change of information needs. In addition, we are investigating learning the information needs of design teams by analysing team communication. Our methodology presents a new means for learning information needs through a combination of LSA, natural language processing and a human-centred approach which places emphasis on understanding what it is that the user is doing.

## 5 References

- [1] Lowe, Alistair, McMahon, Chris, and Shah, Tulan, Culley, S., "A Method for The Study of Information Use Profiles for Design Engineers," Proceedings of the 1999 ASME Design Engineering Technical Conferences, September 12-15, 1999, Las Vegas, Nevada.
- [2] Court, A.W., Culley, S.J., and McMahon, C. A., "The Influence of IT in New Product Development: Observations of an Empirical Study of the Access of Engineering Design Information," International Journal of Information Management, 17(5), 1997, p359-375.
- [3] Dong, Andy and Agogino, Alice M., "Text analysis for constructing design representations," Artificial Intelligence in Engineering, 11, 1997, p65-75.
- [4] Rangan, R.M., and Fulton, R.E., "A data management strategy to control design and manufacturing information," Journal of Engineering with Computers, 7, 1991, p63-78.
- [5] Rezayat, M., "Knowledge-based product development using XML and KCs," Computer-Aided Design, 32, 2000, 299-309.
- [6] Ullman, David G., Dietterich, Thomas G., and Stauffer, Larry A., "A Model of the Mechanical Design Process Based on Empirical Data", Artificial Intelligence in Engineering Design and Manufacturing, 2(1), 1988, p33-52.
- [7] Williams, Ruth L., and Cothrel, Joseph, "Four smart ways to run online communities," Sloan Management Review, 41(4), Summer 2000, p81-91.
- [8] Wood, William H., Yang, Maria, et al., "Design information retrieval: improving access to the informal side of design," Proceedings of the ASME Design Engineering Technical Conferences, 1998.
- [9] Deerwester, Scott, Dumais, Susan T., Furnas, George W., Landauer, Thomas K., and Harshman, Richard, "Indexing by Latent Semantic Analysis," Journal Of The American Society For Information Science, September 1990, 41(6), p391-407.
- [10] Learning Object Metadata, [http://ltsc.ieee.org/doc/wg12/LOMdoc2\\_4.doc](http://ltsc.ieee.org/doc/wg12/LOMdoc2_4.doc).
- [11] In Klahr, David and Kotovsky, Kenneth, (Eds.), Complex Information Processing: The Impact of Herbert A. Simon, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1989.
- [12] Culley, Stephen J., Boston, Oliver P., and McMahon, Christopher A., "Suppliers in New Product Development: Their Information and Integration," Journal of Engineering Design, 10(1), 1999, 59-75.
- [13] Cooper, William S., 1976, "The Paradoxical Role of Unexamined Documents in the Evaluation of Retrieval Effectiveness," Information Processing and Management, **12**, 367-375.
- [14] Salton, Gerald and McGill, Michael J., 1983, Introduction to Modern Information Retrieval, New York: McGraw-Hill Book Company.