INTERNATIONAL CONFERENCE ON ENGINEERING DESIGN
ICED 97 TAMPERE, AUGUST 19-21, 1997

# AUTOMATED INDEXING OF DESIGN CONCEPTS FOR INFORMATION MANAGEMENT

Maria C. Yang and Mark R. Cutkosky

*Information Retrieval, Electronic Design Notebooks, Design Capture*

## 1 Introduction

This study explores ways to increase both the usability and accuracy of design information retrieval and reuse. The DedalAI system attempts to achieve these goals by automatically indexing design concepts in electronic notebooks. DedalAI relies on a two part query structure: a descriptor, which is a generic design concept, and a subject, which is a specific part of a design model. DedalAI builds on the Dedal system [Baudin, *et al* 1993] by automating the indexing of the descriptor half of a query. It is assumed that a descriptor concept can be signaled by words in text related to the descriptor. This paper discusses the development of thesauri for descriptor detection and its application to two types of design documentation: electronic design notebooks and patents. This useful but simple approach can increase both precision and recall over Boolean keyword searches.

## 2 Objectives

### 2.1 Previous work in design information retrieval

Many traditional information retrieval methods rely on word frequency [Faloutsos and Oard 1995] to construct indices of text. However, such indices may not take into account word meaning or intent. For example, if the word “trunk” appears many times in a document, that document will likely be indexed and retrieved for a query containing “trunk.” However, no distinction will be made about whether “trunk” means tree trunks, car trunks, or elephant trunks.

While augmenting word frequency approaches with other methods, such as case based reasoning, has proved promising in alleviating the above problem [Wood 1996], the Dedal system focuses on the search of design concepts, rather than words. Work began by observing the discussion of designers at work [Baya and Leifer 1994]. Over time, generic design concepts like <*Rationale*> and <*Assumptions*> emerged and were codified into a set of descriptors. A companion subject specified what components of an is-a, part-of style device model was being searched for (Fig. 1).

| **Descriptor** | + | **Subject** |
|---|---|---|
| *<Alternative>* | | *Part of an object model* |
| *<Assumption>* | | *Ex:* |
| *<Comparison>* | | |
| *<Construction>* | | |
| *<Location>* | | |
| *<Operation>* | | |
| *<Performance>* | | |
| *<Rationale>* | | |
| *<Relation>* | | |
| *<Requirement>* | | |

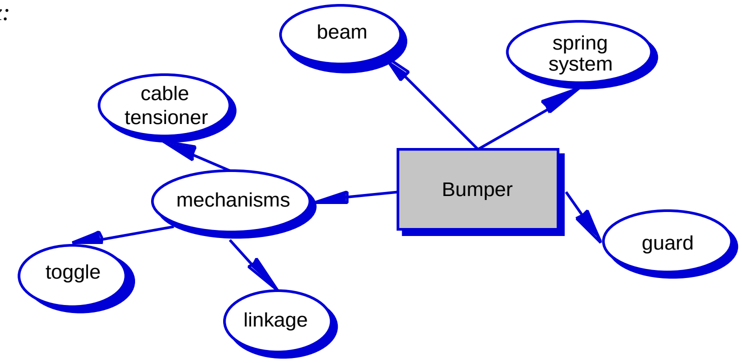


Figure 1: Dedal descriptors and subjects

Using standard information retrieval measures of precision and recall [van Rijsbergen 1979], it has been demonstrated that Dedal can improve precision of design information retrieval by approximately 70% over Boolean searches [Baudin, *et al* 1993]. Precision is the number of relevant retrieved responses divided by all retrieved responses. Recall is the number of relevant retrieved responses divided by all relevant responses in a collection (Equations 1 and 2). Precision is an accuracy measure, while recall is a measure of how much good information is retrieved. Poor precision means users may have to wade through mountains of bad information, while poor recall means that much good information is missed.

$$\text{Precision} = \frac{\text{Correct retrieved}}{\text{Total retrieved}} \qquad (1)$$

$$\text{Recall} = \frac{\text{Correct retrieved}}{\text{Total correct in system}} \qquad (2)$$

## 2.2 Design Opportunity

An obstacle to Dedal's usability, however, is that documentation must be indexed by hand. According to Baudin, *et al* [1991], the main factors in efficient indexing are the familiarity of an indexer with both the material being indexed and the process of indexing. However, even if indexers understand the topic being indexed and have experience indexing, they may be resistant to the additional work required to process information, especially if it is felt that the processing does not benefit them adequately [Grudin 1994]. The goal of DedalAI is to reduce such overhead for designers and engineers while retaining the benefits of conceptual information retrieval.

# 3 Methods

## 3.1 The descriptor thesaurus

In DedalAI, the first step in automatically indexing text was to find ways of indexing the descriptor half of Dedal. This was done using a thesaurus of descriptor synonyms, applicable to a range of electromechanical design documentation. In order to focus on the design of the descriptor thesaurus, Dedal subjects were hand indexed for these experiments.

Instead of searching for pairs of hand indexed descriptors and subjects, a DedalAI retrieval match is made when a descriptor synonym and a subject are detected in text. For example, if the word *possibilities* is a synonym for the descriptor <*Alternative*>, then the DedalAI query <*Alternative*> of <beam> will retrieve the match, "There are many *possibilities* for the design of the <beam>."

## 3.2 Electronic design notebooks

Descriptor thesauri were derived from projects documented in electronic design notebooks. Electronic design notebooks are an emerging genre of software tools for design [Gwidzka, *et al* 1996] and are an electronic analog to the traditional engineer's logbook. The notebooks documented industry-sponsored projects from a year-long graduate course in electromechanical design. Topics included car bumpers, personal digital assistants, and fountains. These notebooks were created with PENS, Personal Electronic Notebook with Sharing [Hong, *et al* 1995], a collaborative, Web-based tool that allows teams of designers to share their work.

For testing, several distinct descriptor thesauri were created, tailored to each project. Every notebook was queried using its own tailored thesaurus, along with combinations of the thesauri of other notebooks. It was surmised that testing the effectiveness of other thesauri on a notebook might indicate how a generalized thesaurus would work on unexamined texts.

Dedal queries themselves began as natural language questions and were then structured as descriptor-subject pairs. For example, the question, "How was the device designed to simulate the human leg?" became the Dedal query <*Construction*> of <device>. Note that there may be a loss of meaning in translation from natural language to Dedal. Therefore, the relevance of a retrieved match was judged on whether it answered the natural language question, rather than the Dedal query.

## 3.3 Role of document structure

The contribution of document structure was also considered in testing. Could a descriptor thesaurus that effectively retrieved information from PENS notebooks perform as well for design information in a different format? To address this, a set of patents covering topics similar to the PENS notebooks was examined. As was done for the PENS notebooks, patents were tested using thesauri tailored to them, as well as with combinations of thesauri from other patents.

From studying patents, it became clear that their organization differed from PENS notebooks. Material in PENS notebooks varied widely from entry to entry, from detailed formal reports to fragmentary lists. Often Dedal subjects were absent or represented by synonyms. In comparison, patents followed a consistent format with its own formal, highly stylized language. Many of the patent terms, used for legal purposes, are rarely used in the everyday lexicon of engineers.

A caveat of patents was that they did not chronicle the design process with the same richness as a notebook covering a year's worth of work. Furthermore, patents did not detail the performance of devices as critically as PENS notebooks, which were part of a graded design project. Descriptor terms about the evolution of a design, like <*Alternative*> or <*Construction*>, or its performance, such as <*Performance*>, did not occur in patents as frequently as they did in design notebooks.

# 4 Results

## 4.1 PENS descriptor thesaurus

Testing demonstrated that a descriptor thesaurus can significantly increase precision compared to Boolean search. Boolean search is regular expression matching of a subject. If a query is <*Rationale*> of <damper>, Boolean search will retrieve only the word "damper," rather than the indexed <damper>. The use of a tailored thesaurus on the notebook it was designed for showed the best results, with recall and precision increasing from 20% to about 70%. Even using thesauri from other notebooks increased precision from about 20% to over 50% (Figure 2).

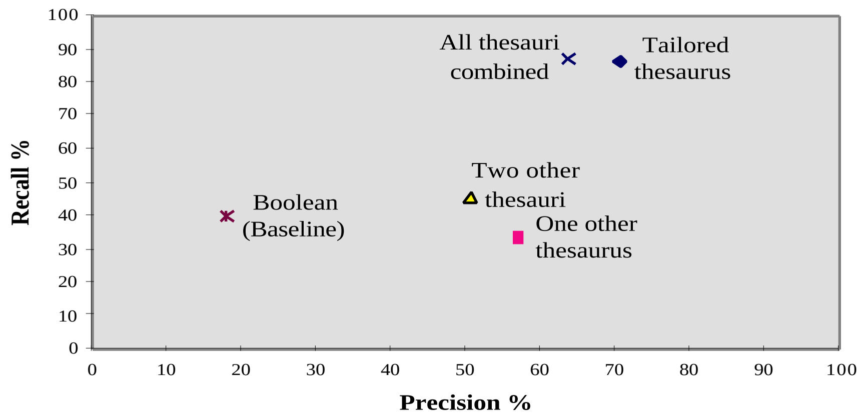


Figure 2: Results of using thesauri to detect Dedal descriptors

## 4.2 Patent descriptor thesaurus

Tests were made of PENS notebooks using descriptor thesauri created from the patents and combinations of PENS notebooks. Precision rose from 20% to 40% and 50% over Boolean search (Figure 3). However, the patent thesaurus showed a drop in recall, from 40% to 20%.

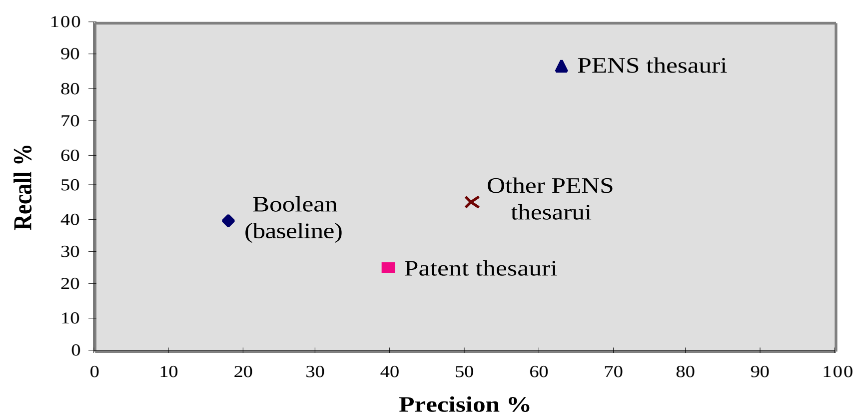


Figure 3: Results of PENS indexing using patent and PENS descriptor thesauri

Tests were also made of patents using descriptor thesauri created from patents and combinations of PENS notebooks. It should be noted that Boolean search alone in patents produced 80% recall and 30% precision, far better than Boolean search in PENS. Although there may be many reasons for this, the main factor is probably the explicit writing style of patents. In PENS notebooks, writing may be informal and fragmentary. Patents, however, need to spell out what is being patented, so subjects are clearly stated, making Boolean search of subjects easier (Figure 4). Still, the PENS thesaurus, and combinations of patent thesauri did increase precision. However, only a patent thesaurus that contained terms tailored to that patent increased recall over Boolean search.

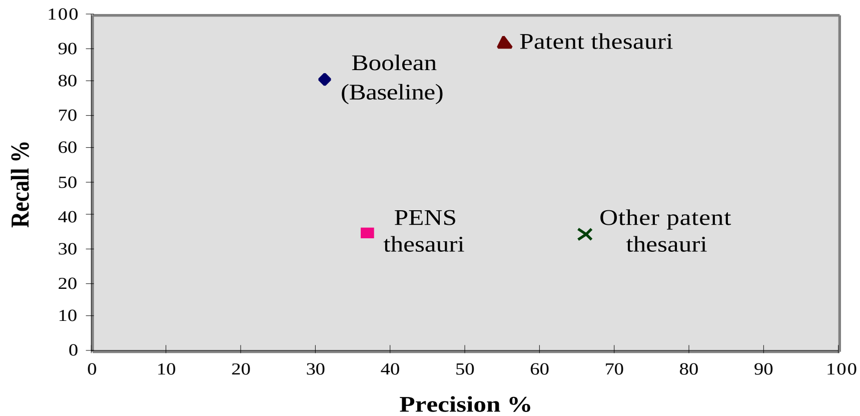


Figure 4: Results of patent indexing using patent and PENS descriptor thesauri

# 5 Conclusions and Future Work

Testing of descriptor thesauri for PENS notebooks suggests that the DedalAI approach can significantly increase precision and recall over Boolean search. Even using the descriptor thesauri of other notebooks showed increases in precision. Results also showed that the use of patent thesauri on PENS notebooks and PENS thesauri on patents only slightly improved precision but decreased recall compared to Boolean search. It is proposed that the underlying reason for this is the difference in language between the two types of documents, which makes interchange of descriptor thesauri difficult. Patents utilize the stylized language of legal documents, contrasted with the informal language of a PENS notebook. Clearly, many issues remain to be explored in DedalAI before stronger conclusions can be made:

• Subject detection

The experiments described in this paper deal only with automatic descriptor detection, but in a truly usable system, both descriptors and subjects would be automatically indexed. An attempt at automatically detecting Dedal subjects using part-of-speech analysis has already been made [Mabogunje and Leifer 1996]. However, results of this work are inconclusive due to the natural language phenomena of anaphora, in which nouns are represented by other nouns or only implied.

• Testing on other documents

Results of the tests described in this paper indicate that writing style, more than document structure, affect the applicability of thesauri to different documents. Future testing with thesauri will involve documentation using similar language, preferably another electronic design notebook.

• User Testing
Outside user testing will be necessary to validate the results of these and future experiments. It is one matter for a single user to judge the relevance of retrieved information, but those results may be quite different from the judgments of other users.

Acknowledgments: The support of Defense Advanced Research Projects Agency and the Office of Naval Research is gratefully acknowledged. Thanks also to Dr. Marti Hearst of XeroxPARC for her insights and advice.

**References:**

Baudin, C., Underwood, J. G., and Baya, V., "Using Device Models to Facilitate the Retrieval of Multimedia Design Information", Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, Chambery, France, August 1993, pp 1237-1243.

Baudin, C., Gevins, J., Baya, V., and Mabogunje, A., "Dedal: Using Domain Concepts to Index Engineering Design Information", Proceedings of the Ninth National Conference on Artificial Intelligence, American Assoc. for Artificial Intelligence, Anaheim, CA, July 1991, pp 702-707.

Baya, V. and Leifer, L. J., "A Study of the Information Handling Behavior of Designers During Conceptual Design", Proceedings of the Sixth International Conference on Design Theory and Methodology, American Society of Mechanical Engineers, Minneapolis, MN, 1994, pp 153-160.

Faloutsos, C. and Oard, D. W., "A Survey of Information Retrieval and Filtering Methods", Technical Report CS-TR-3514, University of Maryland, August 1995.

Grudin, J., "Groupware and Social Dynamics: Eight Challenges for Developers", Communications of the ACM, Vol. 37, No. 1, 1994, pp 92-105.

Gwidzka, J., Louie, J., and Fox, M. S., "EEN: A Pen-based Electronic Notebook for Unintrusive Acquisition of Engineering Knowledge", Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises WETICE '96, Stanford, CA, 1996.

Hong, J., Toye, G., and Leifer, L., "Using the WWW for a Team-Based Engineering Design Class", Electronic Proceedings of the Second WWW Conference, Chicago, IL, October 1994.

Mabogunje, A. and Leifer, L., "210-NP: Measuring the Mechanical Engineering Design Process", Proceedings of the Twenty-Sixth Annual Frontiers in Education Conference on Technology-Based Re-Engineering Engineering Education, Salt Lake City, UT. November 1996.

van Rijsbergen, C. J., "Information Retrieval", Butterworths, London, 1979.

Wood, W. H., "Supplying Concurrent Engineering Information to the Designer: The Conceptual Design Information Server", Ph.D. Dissertation, University of California, Berkeley, 1996.

Maria C. Yang, Graduate Student Researcher
Center for Design Research, Stanford University
560 Panama Street, Stanford, CA 94305 USA
(415) 723-7909, (415) 725-8475 FAX, *mcyang@cdr.stanford.edu*