ProThes: Thesaurus-based Meta-Search Engine for a Specific Application Domain

Pavel Braslavski
Institute of Engineering Sciences
34 Komsomolskaya St.
620219 Ekaterinburg, Russia
+ 7 (343) 374 5953
pb@imach.uran.ru

Gleb Alshanski
Institute of Metal Physics
18 Sofia Kovalevskaya St.
620219 Ekaterinburg, Russia
+ 7 (343) 378 3563
alshansk@imp.uran.ru

Anton Shishkin
Institute of Engineering Sciences
34 Komsomolskaya St.
620219 Ekaterinburg, Russia
+ 7 (343) 374 5953
whoarym@imach.uran.ru

ABSTRACT

In this poster we introduce *ProThes*, a pilot meta-search engine (MSE) for a specific application domain. *ProThes* combines three approaches: meta-search, graphical user interface (GUI) for query specification, and thesaurus-based query techniques. *ProThes* attempts to employ domain-specific knowledge, which is represented by both a conceptual thesaurus and results ranking heuristics. Since the knowledge representation is separated from the MSE core, adjusting the system to a specific domain is trouble free. Thesaurus allows for manual query building and automatic query techniques. This poster outlines the overall system architecture, thesaurus representation format, and query operations. *ProThes* is implemented on J2EE platform as a Web service.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]. H.3.1: Content Analysis and Indexing – *thesauruses*; H.3.3: Information Search and Retrieval – *query formulation, information filtering;* H.3.5: Online Information Services – *Web-based services.*

General Terms

Experimentation, Design, Human Factors, Languages.

Keywords

Information retrieval, meta-search, user interface, thesaurus, query operations, Web services.

1. INTRODUCTION

The growth of the Web leads to high popularity of the online search services. Meeting the demand, Web search engines (SE) show superior productivity and extensive content coverage. Aiming for satisfying as many Web surfers as possible, search engines employ modest user interfaces in addition to simple query syntax by default and make strong assumptions about user behavior, preferences, etc. Searchers with specific information needs do not always benefit from this approach.

In this poster, we propose a solution for focused Web information retrieval, which emphasizes the query specification stage of the retrieval process (in contrast for example, to analyzing page contents or link structure [2]) and aims at employing the power of the all-purpose search engines. We introduce *ProThes*, a system that combines meta-search, graphical user interface for query specification, and thesaurus-based query techniques. *ProThes* customization is achieved by means of a conceptual thesaurus that is used for various query operations and simple heuristics for results merging and partial re-ranking. The separation of the domain-specific knowledge from the system logic allows easily switching between different domains.

WWW 2004, May 17-22, 2004, New York, NY USA.
ACM 1-58113-912-8/04/0005.

2. SYSTEM DESIGN

ProThes is a Web service developed using Java 2 Enterprise Edition (J2EE) platform. The server part includes a thesaurus component (T), query and response dispatchers (QD and RD respectively), and search engine gates (fig. 1). The gates to Google (www.google.com) and Yandex, Russian leading SE (www.yandex.ru) have already been implemented.

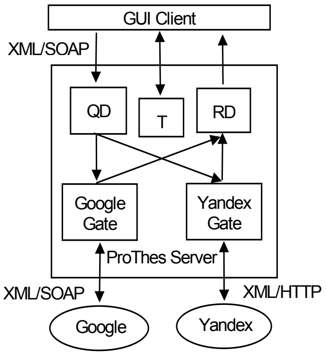


Figure 1. Overall System Architecture

The client is a GUI application developed using Java Swing library. It consists of a thesaurus visualization component, a query constructor, and a results representation area (in a separate window).

3. THESAURUS

The thesaurus is a key component of the proposed MSE. The basic element of the suggested thesaurus is a *concept* rather than a *term*. A concept is defined purely through associated terms. By this approach, we, first, gain a simple structure for describing various types of synonymy (including cross-language equivalents) and polysemy. Second, we can effectively choose the appropriate granularity of the knowledge representation. Third, we operate on a higher conceptual level than the lexical one.

Moreover, we assume that an accurate knowledge description can demand various semantic link types between concepts. Hence we would not limit link types set supposing that it must be adjusted to the specificity of each domain. However, as a singular case a thesaurus can be imagined in which each concept is presented by a single term and concepts are connected by no-named (e.g. statistically produced) links. The main idea is to let the developers choose thesaurus structure and link types freely.

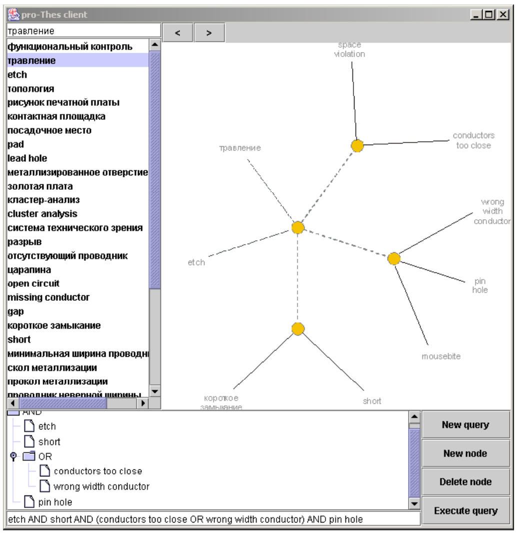
An XML Scheme for thesauri was developed. In general, the instance thesaurus consists of a header and a set of *concept entries*; each of them consists in its turn of *definition*, *links*, and a set of *term entrie*s. On the bottom level lie *term*s along with associated *acronym*s, *cognate*s, *variant*s, and usage *contexts*. Most of the thesaurus elements are optional. Developer of an instance thesaurus can expand the set of link types using the XML *redefine* mechanism.

Discussion on the thesaurus model and the format particularities can be found in [1]. The developed core XML Schema is available at http://imach.uran.ru/pb/thesaurus/thesaurus.xsd.

A Russian-English thesaurus of the domain "Automated Optical Inspection of the Printed Circuit Boards" was build manually from scratch. It consists mainly of PCB and computer vision related concepts. The thesaurus contains approximately 200 concepts, 800 terms, and 750 one-way links as of January 2004.

4. FUNCTIONALITY

Visualizing concept network along with definitions, related concepts, associated terms, terms usage etc., *ProThes* maintains the pick-up metaphor of manual query building. User can specify a query as an AND-OR-ANDNOT-tree, choosing appropriate terms from the thesaurus network (fig. 2).



pro-Thes client

травление

функциональный контроль
травление
etch
топология
рисунок печатной платы
контактная площадка
посадочное место
pad
lead hole
металлизированное отверстие
золотая плата
кластер-анализ
cluster analysis
система технического зрения
разрыв
отсутствующий проводник
царапина
open circuit
missing conductor
gap
короткое замыкание
short
минимальная ширина проводни
скол металлизации
прокол металлизации
проводник неверной ширины

AND
etch
short
OR
conductors too close
wrong width conductor
pin hole

etch AND short AND (conductors too close OR wrong width conductor) AND pin hole

Figure 2. ***ProThes'*** **GUI**

Moreover, we propose two kinds of automatic query transformations.

The first one is based on templates. A template defines *term entry* fields to be used, link types along with the appropriate operators (AND, OR, ANDNOT), expansion depth, and language options. Starting from the pointed pivot concept, *ProThes* builds a query using thesaurus breadth traversal. Selected elements within a concept are ORed; the resulting query can be translated and split between different search engines depending on language options (e.g. a Russian query is sent to Yandex, an English one – to Google).

For queries built with the thesaurus appear frequently too strict, the second kind of automatic transformations is query loosening and is similar to the one proposed in [3]. A query can be loosened gradually by omitting quotation marks, adding quasi-synonyms, replacing AND with OR.

An important task *ProThes* has to execute is results merging and re-ranking. Domain-specific preferences can be expressed in an initialization XML file, so the final document's position in the merged list depends on its position in the initial response list, the SE confidence, the file size, date and extension, as well as domain name. In the latter case the URLs from specific Web directory sections can be used. However, lacking for both global statistics and documents themselves, we can concern only partial re-ranking.

5. CONCLUSION

Combining three established techniques, – meta-search, graphical user interface for query specification, and thesaurus-based query operations, – we try to balance out the universality of the Web search engines and the specificity of the user information needs.

Our preliminary experiments have shown that automatic query techniques, although being very helpful in many cases, fail to deliver consistently good results. Hence, the automatically produced expressions should be considered rather as suggestions than as ready-to-send queries.

Manual procedures of thesaurus building can be a bottleneck of the proposed approach. In our future work we are going to address the problem of the automated lexical acquisition.

6. ACKNOWLEDGMENTS

The project was supported in part by the Russian Fund of Basic Research, grant # 03-07-90342.

Pavel Braslavski is especially grateful for the hospitality of the Computer Science Department at the Freie Universität (FU) Berlin, where part of his contribution to this project was made. His research at the FU Berlin was supported by the German Academic Exchange Service (DAAD).

7. REFERENCES

[1] Braslavski, P. Thesaurus for Query Expansion for the Web Search Engines: Structure and Functions (in Russian). [Tezaurus dlya rasširenija zaprosov k masinam poiska Interneta: struktura i funkcii]. In Proceedings of Dialogue'2003 (Protvino, Russia, June 2003), Nauka, 95-100.

[2] Chakrabarti, S., Berg, M., and Dom, B. Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery. In Proceedings of the WWW8 (Toronto, Canada, May 1999), http://www8.org/w8-papers/5a-search-query/crawling.

[3] Gauch, S., and Smith, J.B. An Expert System for Automatic Query Reformulation. In Journal of the American Society of Information Science, 1993, 44 (3), 124-136.