3 in 1: Meta-Search, Thesaurus, and GUI for Focused Web Information Retrieval ♣

© Pavel Braslavski
Institute of Engineering Sciences, UB RAS, Yekaterinburg
pb@imach.uran.ru

© Anton Shishkin
Institute of Engineering Sciences, UB RAS, Yekaterinburg
whoarym@imach.uran.ru

© Gleb Alshanski
Institute of Engineering Sciences, UB RAS, Yekaterinburg
alshansk@imp.uran.ru

Abstract

In this paper we introduce ProThes, a system for focused Web information retrieval. ProThes combines three approaches: meta-search, advanced graphical user interface (GUI) for query specification, and thesaurus-based query techniques. ProThes attempts to employ domain-specific knowledge, which is represented by a conceptual thesaurus. Moreover, ProThes uses domain-specific results ranking heuristics. Since the domain characteristics are separated from the system core, adjusting to a specific domain is trouble free. Thesaurus allows the user to build queries both manually and in automatic mode. The paper describes the system architecture, meta-search features, thesaurus representation format, query operations, GUI and provides a number of query examples. ProThes is implemented on Java 2 Enterprise Edition platform.

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 our research we address the problem of focused Web information retrieval, concentrating primarily on the query formulation in contrast to analyzing page contents or link structure (see e.g. [8]). As investigations show [2], query formulation, i.e. the transformation of a user's information need into a list of key-words, appears challenging for many searchers and remains an essential part of successful retrieval at the same time.

In this paper we introduce ProThes, a light and flexible solution that combines meta-search features, thesaurus-based query techniques, and graphical user interface (GUI) for query specification. An earlier version of ProThes was presented in [4], a short description of the current version can be found in [5].

Unlike many federated solutions for digital libraries (e.g. [10]), ProThes does not have own search index nor harvests metadata from the information sources. ProThes customization is achieved by means of a conceptual thesaurus and simple heuristics for partial results re-ranking. The separation of the domain characteristics from the system logic allows easily switching between different domains. We consider ProThes to be a tool for Internet/intranet search, as well as a useful digital library component.

The three mentioned approaches are not novel taken separately.

Meta-search is an established IR technique that allows delivering a transparent interface to numerous information sources and increasing overall recall (although the latter feature is not critical for the Web search in general, it can be quite useful in case of focused retrieval). Moreover, meta-search engines (MSE) can trawl so-called 'deep Net' [11] and conduct a thorough analysis of downloaded documents [13]. A comprehensive list of commercial MSEs can be found in [15].

Thesauri have been traditionally used for reducing the well-known vocabulary problem, i.e. the fact that one concept can be expressed through different terms. Thesaurus-based query expansion (QE) techniques have been extensively discussed in the literature; different effects have been reported. However, manually crafted thesauri employed in local information retrieval systems show good results [3], [9]. Recent studies [1] justify that QE techniques can significantly improve retrieval performance. However, there are many obstacles for using thesauri in the Web information retrieval. First of all, it is a challenging task to build and maintain a thesaurus that would meet the Web's scale and diversity. Second, this feature will highly probably be redundant for most SE users.

Nevertheless, there are some practical examples. AltaVista (www.altavista.com) has been offering

Proceedings of the 6th Russian Conference on Digital Libraries RCDL2004, Pushchino, Russia, 2004

‘AltaVista Refine’ feature in the late 90s. The user was able to decrease ambiguity of the terms in the initial query by selecting appropriate co-occurring terms (see [14] for a short description and screen-shots). Later, at midyear of 2002, AltaVista launched a similar service called ‘Prisma Query Refinement’. Now, Google (www.google.com) offers synonym search (e.g. the query ‘~cats’ will deliver results containing such words as ‘cat’, ‘dogs’, ‘pets’, and ‘kitten’). As opposed to these centralized and universal solutions we propose a stand-alone tool for restricted application domains.

Graphical user interface (GUI) for search applications is another option to improve query formulation. For example, representing Boolean queries in the form of Venn diagrams appears more intuitive for users and delivers more accurate searches [12]. Another approach is reported in [16]. Automatically extracted concepts of a document collection are represented in 2D space; their mutual positions reflect semantic closeness. The user can pick the terms to be ORed or ANDed from the conceptual map.

Bringing three mentioned approaches together, we aim for balancing out the universality of the Web search engines and the specificity of the user’s information needs. The following sections describe the three mentioned ProThes aspects in sequence. The fifth section presents some examples of ProThes use; section 6 concludes the paper and outlines the future work.

2 Meta-Search Engine

2.1 System Architecture

ProThes uses a client-server architecture. ProThes meta-search server is developed as a Web service using Java 2 Enterprise Edition (J2EE) platform. Server and client applications communicate via SOAP. Except for thesaurus component, whose functionalities are described in the next sections, the proposed structure is rather traditional (fig. 1). The server includes a thesaurus component (T), query and response dispatchers (QD and RD respectively), and search engine gates (they consist of query wrappers and response parsers in turn).

The thesaurus component performs initial thesaurus loading and subsequent thesaurus querying.

The gates reflect such SE peculiarities as protocols used, query syntax, and response format. Gates to Google and Yandex (www.yandex.ru) have already been implemented. The gates employ the APIs offered by the search engines that allow the application to exchange structured XML data.

Query wrappers translate internal treelike queries into plain strings. Both Google and Yandex have query length restrictions: Google limits a query to 10 terms (without taking ORs and ANDs into account), while Yandex limits the length of a single query to 255 characters (e.g. *&&* operator adds 10 extra characters as an HTML escape sequence *&amp;&amp;*). The wrappers analyze queries and split them along the OR operators when needed. Long queries containing exclusively AND operators can be only truncated.

The response parsers translate SE responses into internal response structures and send them to response dispatcher. The dispatcher merges the results, finds duplicates, and partially re-ranks the list according to the pre-defined preferences.

The client is a GUI application developed using Java Swing library. Description of the interface can be found in section 4. Moreover, the client application maintains the query-building operations.

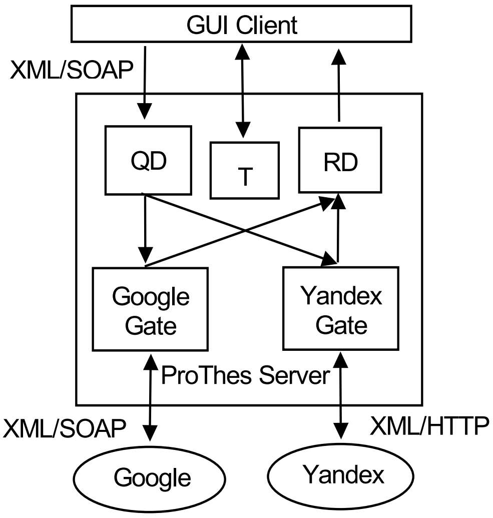


Figure 1. System architecture

2.2 Query Language

ProThes uses strict Boolean query syntax at the moment. The user can construct a query as an AND-OR-ANDNOT-tree (fig. 2) and define thereby the execution order precisely, which is important in some cases. Thus, Google interprets the query *mommy AND daddy OR son* as *mommy AND (daddy OR son)*, while Yandex understands the same query as *(mommy AND daddy) OR son*. ProThes supports phrase search (i.e. encloses multi-word terms in quotation marks).

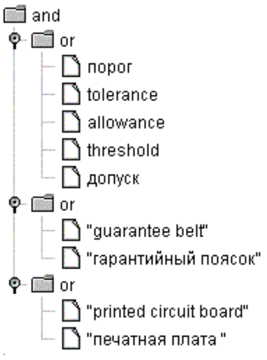


Figure 2. Query tree

2.3 Results Merging and Ranking

An important task ProThes has to execute is results merging and re-ranking.

While merging the results, ProThes resolves duplicates using simple heuristics based on URL, file name and size, as well as document title.

Domain-specific ranking preferences are expressed in an initialization XML file (fig. 3); so the final document position in the merged list depends on the position in the initial response, the SE confidence, file format, as well as domain name. In the latter case the URLs from the specific Web directory sections – such as Yahoo!Directory (http://dir.yahoo.com) or Open Directory Project (http://dmoz.org) – can be used.

Lacking for both global statistics and documents themselves we opted for performing only small shift-ups and shift-downs relative to the initial position in the SE response. Currently we use the following simple formula for calculating document ranks:

$$R = P_{initial} - a*sign(x+1)*ln(abs(x+1)),$$

where $P_{initial}$ is the initial document position in the SE response, a – a positive constant, and x – the additional points scored by the document according on the ranking preference settings. Since the ranks in the XML file (fig. 3) lie in the range –128 to 127, a document can score maximum 3*127=381 positive points and get up for approximately (6 time a) positions.

```
<extension rank="100">pdf</extension>
<extension rank="10">ps</extension>
<URL rank="5">http://www.aldsn.com</URL>
<URL rank="5">http://www.acae.com</URL>
<URL rank="5">http://www.alphacinc.com</URL>
<URL rank="5">http://www.cadesign.net</URL>
<URL rank="10">http://www.dvk.com</URL>
…
<SE rank="1">yandex</SE>
<SE rank="2">google</SE>
```

Figure 3. Ranking preferences sample

3 Thesaurus

3.1 Thesurus Model

Thesaurus is the basic means for representing domain-specific knowledge employed in ProThes. We aimed to produce a flexible solution, so we had to develop the thesaurus representation format especially carefully.

The basic element of the suggested structure is a concept rather than a term. A concept is defined purely through associated terms. By this approach, first, we 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 the set of link types 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.

As a representation format we have chosen XML, since it meets the requirements of openness, portability, flexibility, and extensibility. 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 entries*. On the bottom level lie *terms* along with associated *acronyms*, *cognates*, *variants*, 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 [7]. The developed core XML Schema is available at http://imach.uran.ru/pb/thesaurus/thesaurus.xsd

3.2 Thesaurus Sample

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 references to the sources used for thesaurus building can be found in [6]. The thesaurus contains approximately 200 concept entries, 800 bilingual terms, and 700 one-way links between concepts at the moment. Figure 4 represents a thesaurus fragment and gives an idea of proposed thesaurus representation.

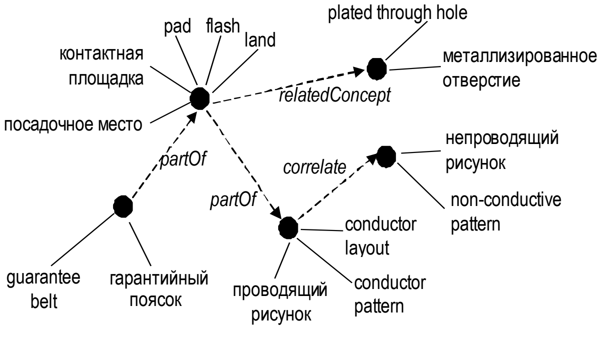


Figure 4. Thesaurus fragment

3.3 Thesaurus Operations

Besides providing the user with a ‘semantic map’ of the domain, thesaurus can be employed for automatic query transformations. We propose four kinds of them.

The first one is query translation, the simplest and possibly the most effective operation. Many professional terms are fixed multiword expressions that are hardly capable of being translated word-by-word. However, manual work must precede the automatic translation.

The second operation is template-based query expansion. A template defines term entry fields and link types to be used, appropriate operators (AND, OR), 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).

The third mode is the shortest path finding between two specified concepts. By this approach the path-composing concepts are ANDed, whereas the concept-related terms are ORed. An analogous method is presented in [3].

For queries built with the thesaurus appear frequently too strict, we propose the third kind of transformations – query loosening. A query can be loosened gradually by omitting quotation marks, adding quasi-synonyms, removing ANDNOT terms, replacing AND with OR. A similar technique was proposed in [9].

The latter two features have not been implemented yet.

Preliminary experiments with automatic query operations have shown significant spread in quality of the SE responses [6]. Moreover, template-based queries with depth more than one appeared to be very unhandy and loose. Despite the fact that some good results occur, the methods cannot deliver output of consistent quality. These outcomes confirm to the results reported in [1]. Therefore we consider automatically produced expressions rather as suggestions to user than as ready-to-send queries. The aim of the automatic query operations is to support the interactive manner of retrieval process and contribute to the search skills acquisition by the user.

4 Graphical User Interface

4.1 Thesaurus Browsing and Visualization

The current ProThes graphical user interface can be seen in figure 5. It consists of the alphabetically ordered searchable term list, a thesaurus visualization area, and a query constructor area. Search results are represented in separate windows.

Using the term list, the user can choose a concept. The chosen concept, its neighbor concepts and the associated terms are displayed is the right-hand area. The user can drag, enlarge and decrease the view. The hint windows submit additional information such as concept definition, term usage example, variants, etc.

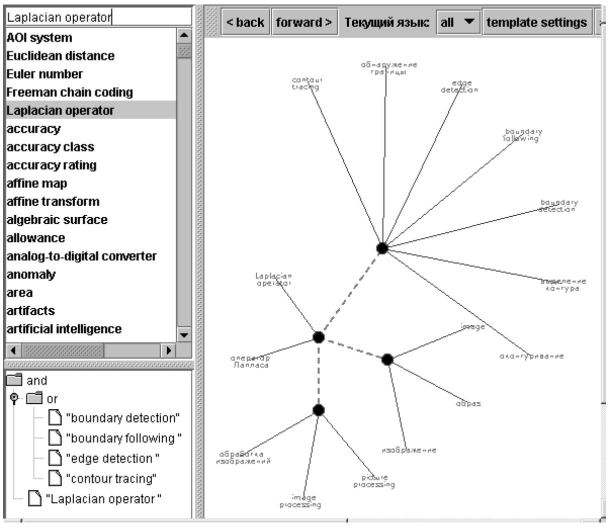
By clicking on the neighbor concepts one can make them 'central' and browse through thesaurus network in that way. 'Back' and 'forward' buttons allow the user to go through the session history.

The user can also restrict the thesaurus representation to one language.

4.2 Query Building

The user is able to specify a query as an AND-OR-ANDNOT-tree by adding new nodes (both terms and operators), deleting them, and changing operator type. The user can either pick terms from the thesaurus network or enter new ones. Mouse right-clicks add either a single term or the whole concept (associated terms are automatically ORed) to the current query node. Thus, ProThes maintains the pick-up metaphor of query building. The query tree is represented in the left-hand bottom area.

The second option is template-based query building. After defining template settings the user can apply the template and see the resulting query in the query constructor window. The user can edit the query when needed and send it to the search engines. The result will appear in a separate window.



Laplacian operator

AOI system
Euclidean distance
Euler number
Freeman chain coding
Laplacian operator
accuracy
accuracy class
accuracy rating
affine map
affine transform
algebraic surface
allowance
analog-to-digital converter
anomaly
area
artifacts
artificial intelligence

- and
  - or
    - "boundary detection"
    - "boundary following"
    - "edge detection"
    - "contour tracing"
  - "Laplacian operator"

Figure 5. Graphical User Interface

5 Examples

Four bilingual pairs of queries constructed using ProThes and number of respective SE responses are presented in table.

The results allow us to make at least two important remarks. First, the complex queries are easy to construct of terms presented in the thesaurus. Second, the user can benefit from automatic query translation combined with synchronous use of two different search engines.

However, the presented results do not substitute a thorough evaluation. Unfortunately, it is hard to evaluate the proposed meta-search and thesaurus features of the system in conventional IR terms of precision and recall using standard collections and methods, since we propose a solution for domain-specific retrieval.

Query	Google	Yandex
("automatic optical inspection" OR AOI) AND ("printed circuit board" OR PCB)	~11000	74
("автоматический оптический контроль" OR "автоматический визуальный контроль" OR АОК) AND "печатных плат"	15	112
("boundary detection" OR "boundary following" OR "edge detection" OR "contour tracing") AND "Laplacian operator")	581	2
("обнаружение границы" OR "выделение контура" OR оконтуривание) AND "оператор Лапласа"	1	4
("morphological operation" OR "morphological processing") AND (contraction OR shrinking) AND dilatation	5	1
"морфологическая операция" AND (дилатация OR расширение OR "сокращение фона") AND (контракция OR "сжатие объекта" OR "расширение фона")	0	0
("affine map" OR "affine transform") AND ("image processing" OR "picture processing")	996	24
("аффинное преобразование" OR "аффинное отображение") AND "обработка изображений"	1	142

Table. Queries built using thesaurus and number of respective responses

6 Conclusion and Future Work

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's information needs. We consider ProThes to be a light and flexible alternative for focused Web information retrieval. Although we focus on a solution for the Web, the proposed design can be seen as an adjustable search interface to heterogeneous information sources in different environments.

We recognize that the manual thesaurus building routine can be a bottleneck of the proposed framework, although we intend the approach for the application domains with manageable vocabulary (up to 1000 terms). Therefore we plan to address the problem of automatic lexical acquisition and to develop semiautomatic tools for thesaurus building.

There are many ambiguous result reported about usefulness of automated query expansion. Although the skepticism regards mainly the statistically built thesauri applied to the Web search, we do not make strong emphasis on automatic features. 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.

An important task we have to execute in the nearest future is a thorough study of the proposed GUI. Now we examine different approaches to the user interface evaluation.

In conclusion we would thank the anonymous reviewers, whose comments helped us to improve the paper.

References

[1] Alemayehu N. Analysis of Performance Variation Using Query Expansion. In *Journal of the American Society for Information Science and Technology*, volume 54(5), pages 379–391, 2003.

[2] Aula A. Query Formulation in Web Information Search. In Isaías, P. & Karmakar, N. (Eds.) In *Proceedings of IADIS International Conference WWW/Internet 2003*, volume I, pages 403-410, 2003. Available online: http://www.cs.uta.fi/~aula/questionnaire.pdf

[3] Bodner R., Song F. Knowledge-based approaches to query expansion in information retrieval. In *Advances in Artificial Intelligence.* McCalla, G. (Ed.), pages 146-158. New York: Springer, 1996.

[4] Braslavski P., Alshanski G., Titov P. Thesaurus-Based Query Building for the Web Search Engines: Semantic-Oriented Approach (in Russian). [Formirovanie informacionnyh zaprosov k mašinam poiska interneta na osnove tezaurusa: semantiko-orientirovannyj podhod]. In *Proceedings of the 8th International Conference on Electronic Publications 'Elpub-2003'*, Novosibirsk, October 8-10, 2003. Available online: http://www.ict.nsc.ru/ws/elpub2003/5964/

[5] Braslavski P., Alshanski G., Shishkin A. ProThes: Thesaurus-based Meta-Search Engine for a Specific Application Domain. In *Proceedings of the 13th International World Wide Web Conference,* May 17-22, 2004, New York, NY USA, volume 2, pages 222-223. Available online: http://www2004.org/proceedings/docs/2p222.pdf

[6] Braslavski P. Automatic Thesaurus-based Operations with Queries to the Web Search Engines: Approaches and Estimation (in Russian). [Avtomatičeskie operacii s zaprosami k mašinam poiska interneta na osnove tezaurusa: podhody i ocenki] *Proceedings of the International Conference "Dialogue-2004. Computational Linguistics and Intelligent Technologies",* pages 79-84. Moskva: Nauka, 2004.

[7] Braslavski P. Thesaurus for Query Expansion for the Web Search Engines: Structure and Functions (in Russian). [Tezaurus dlya rasšireniya zaprosov k masinam poiska Interneta: struktura i funkcii]. In *Proceedings of the International Conference*

"Dialogue-2003. Computational Linguistics and Intelligent Technologies", pages 95-100. Moskva: Nauka, 2003. Available online: http://www.dialog-21.ru/Archive/2003/Braslavskij.pdf

[8] Chakrabarti S., Berg M., Dom B. Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery. In *Proceedings of the 8th International World Wide Web Conference,* Toronto, Canada, May 11-14, 1999. Available online: http://www8.org/w8-papers/5a-search-query/crawling/index.html

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

[10] Gonçalves M. A., France R. K., Fox E. A. MARIAN: Flexible Interoperability for Federated Digital Libraries. In *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL2001),* Darmstadt, Germany, September 2001, pages 173-186.

[11] Hamilton N. The Mechanics of a Deep Net Metasearch Engine. In *Proceedings of the 12th International World Wide Web Conference,* May 20-24, 2003, Budapest, Hungary. Available online: http://www2003.org/cdrom/papers/poster/p170/poster/poster.html

[12] Jones S., McInnes S., Staveley M.S. A Graphical User Interface for Boolean Query Specification. In *International Journal on Digital Libraries Special Issue on User Interfaces for Digital Libraries*, volume 2(2/3), pages 207–223, 1999. Available online: http://www.cs.waikato.ac.nz/~stevej/Research/PAPERS/ijodlvquery.pdf

[13] Lawrence S., Giles C. L. Inquirus, the NECI meta search engine. In *Proceedings of the 7th International World Wide Web Conference,* April 14-18, 1998, Brisbane, Australia. Available online: http://www7.scu.edu.au/programme/fullpapers/1906/com1906.htm

[14] Schwarz C. Web Search Engines. In *Journal of the American Society for Information Science*, volume 49 (11), pages 973–982, 1998.

[15] Sherman Ch. Metacrawlers and Metasearch Engines (January 12, 2004) – http://searchenginewatch.com/links/print.php/34691_2156241

[16] Zhang J., Mostafa J., Tripathy H. Information Retrieval by Semantic Analysis and Visualization of the Concept Space of D-Lib® Magazine. In *D-Lib Magazine*, volume 8 (10), 2002. Available online: http://www.dlib.org/dlib/october02/zhang/10zhang.html

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