

ProThes: Thesaurus-based Querying Middleware

Pavel Braslavski

*Institute of Engineering Sciences,
Ural Division RAS*



Key words: information retrieval, thesaurus, middleware



Samsung Young Scientist Day
April 26-27, 2004



Pavel Braslavski

- Date of birth: December 11, 1973
- Education:

1990 – 1997, Ural State Technical University, Yekaterinburg, Russia, specialty: Computer Systems. Diploma thesis: “Thesaurus System Development (Case Study on the Computational Linguistics Terminology)”

1997 – 2000: Ural State Technical University, Yekaterinburg, Russia, Ph.D. candidate. Ph.D. thesis: “Effective Methods of Scientific Information Retrieval on the Web”



Pavel Braslavski (2)

- Scientific interests:
Information Retrieval
Natural Language Processing
Digital Archives and Libraries
- Current positions:
Institute of Engineering Sciences, Ural Branch RAS,
Computer Systems Department, senior researcher;
Ural State Technical University, Computer Science Department,
part-time assistant professor
- Current projects:
ProThes
Russian Information Retrieval Evaluation Seminar
Virtual reality as a symbolic means of development of human
psyche



ProThes: Thesaurus-based Querying Middleware

- Defining a Task:

The problem referred frequently as *information overload* leads to high popularity of search services and makes information retrieval central to many professional activities. Most up-to-date search services employ key word queries. It is known that query formulation, i.e. the transformation of user's information need into a list of key words, appears challenging for many searchers.

Vocabulary problem, i.e. the fact that a concept can be expressed through different terms and a term can have different meanings, makes the task even more challenging. Moreover, user has to be acquainted with query syntax and system design.

As a partial solution of the problem we propose *ProThes*, a querying middleware, which is aimed at focusing on a specific application domain.

ProThes: Thesaurus-based Querying Middleware

Abstract

ProThes is a thesaurus-based middleware for querying heterogeneous information resources. Besides offering a transparent search interface through a set of wrappers, *ProThes* can be fine-tuned for a specific application domain. A conceptual thesaurus and results ranking heuristics represent domain-specific knowledge. *ProThes*' client provides a graphical user interface for query specification. The pilot version of *ProThes* is implemented on J2EE platform as a Web service. We consider *ProThes* to be a light and flexible tool for Internet/intranet search, as well as a useful digital library component.

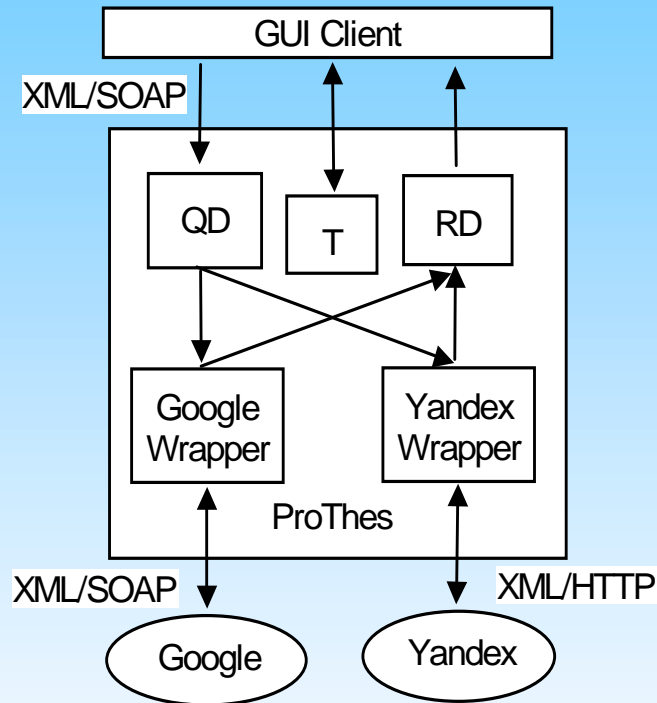
ProThes: Thesaurus-based Querying Middleware

- Project target:
To develop a flexible semantic-oriented querying middleware.
- Expected results:
The middleware will allow for more sophisticated querying heterogeneous information sources.
- Research objectives:
Knowledge representation in the form of a thesaurus.
Automatic techniques for thesaurus building.

ProThes: Thesaurus-based Querying Middleware

- Key technologies, methods and equipment:
Development platform: Java 2 Enterprise Edition (J2EE).
Implementation as a Web service (uses SOAP).
Server software: JSP/Servlet container (Apache Tomcat 4.1 or higher).
Server hardware: PIII 1 GHz, 256 Mb RAM
- Originality of Idea:
Each single feature of the application is not novel by itself. Bringing them together we hope to achieve a new quality.
- Output:
Querying middleware implemented as a Web service.
Methodology and tools for thesauri building.

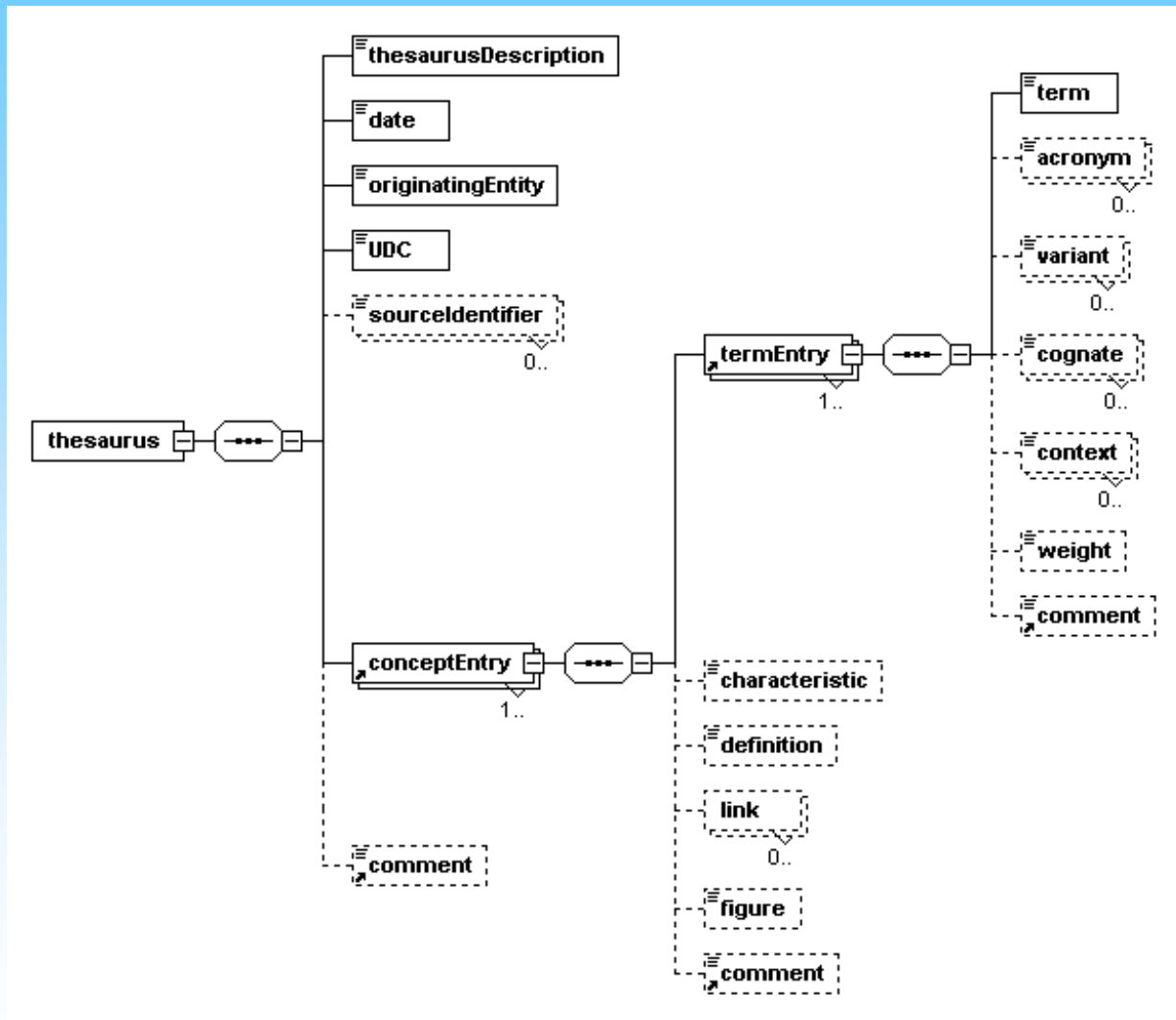
ProThes Architecture



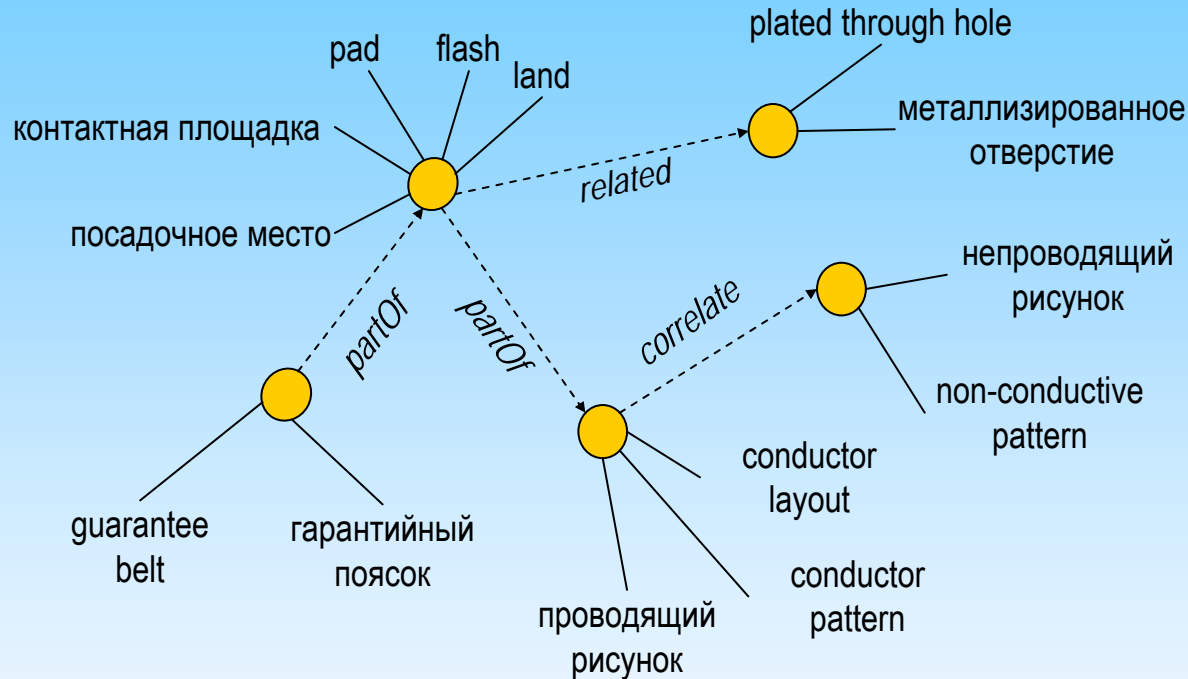
Legend:
T – thesaurus component;
QD – query dispatcher;
RD – response dispatcher.



Thesaurus XML Scheme



Thesaurus Fragment



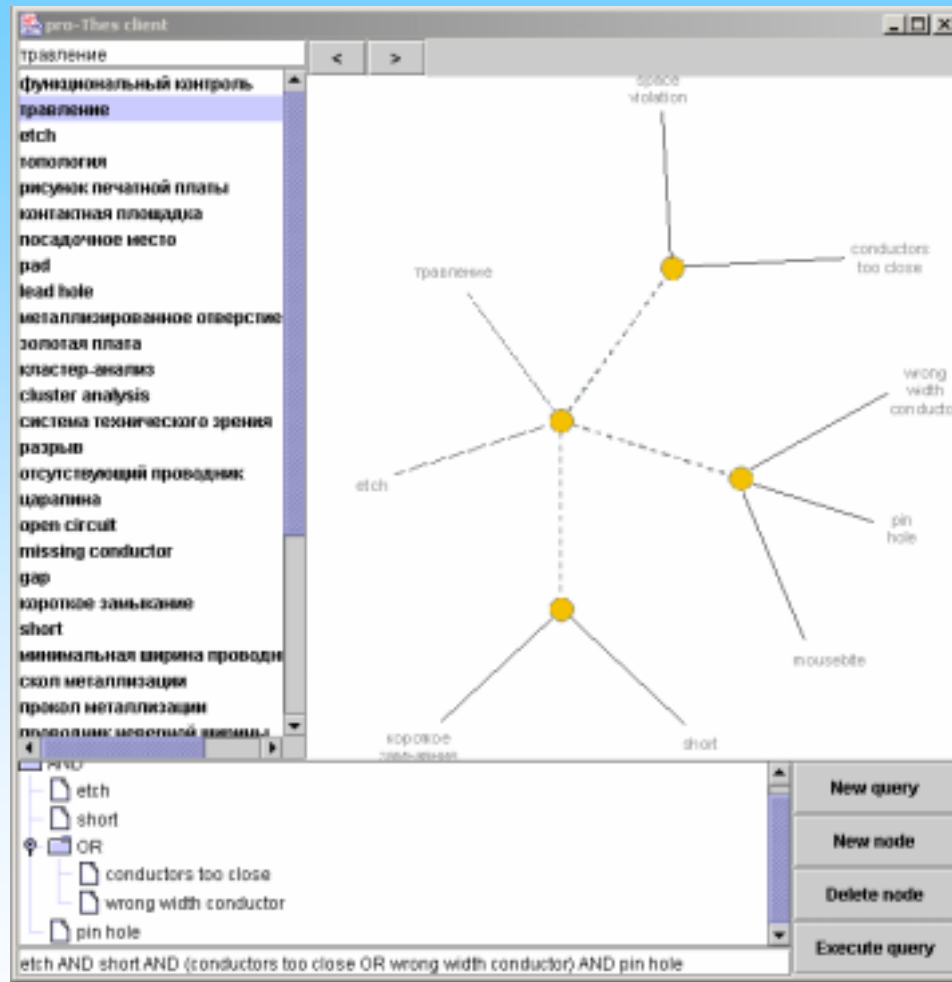
Russian-English thesaurus of the domain “Automated Optical Inspection of the Printed Circuit Boards”:
about 200 concepts, 800 terms

Automatic Query Operations

- Query translation.
- Template-based query expansion.
- Path finding between two concepts.
- Query loosening.



Graphical User Interface



Application Areas

- Internet/intranet search with advanced features: meta-search, graphical user interface, and thesaurus-based techniques
- Domain knowledge and terminology representation



Already Accomplished Tasks

- A pilot implementation as a Web service with wrappers for Google and Yandex.
- A pilot implementation of a client-side applet (thesaurus network visualization, query specification interface, and results representation)
- Thesaurus model and corresponding XML Schema.
- A sample Russian-English thesaurus of the domain “Automated Optical Inspection of the Printed Circuit Boards”.



Sample Queries Built Using ProTheS

Query	Google results
halftone "gray-level" grayscale monochrome	208
"automatic optical inspection" "image processing" (binarization OR thresholding)	15
("golden PCB" OR "golden board") ("reference comparison" OR "template matching") AOI	23
("golden PCB" OR "golden board") "reference comparison" (AOI OR „automatic optical inspection“)	11

Advantages of ProThes

- Transparent interface to heterogeneous information sources.
- Query assistance through conceptual thesaurus and graphical user interface.
- Adjusting to a specific domain is trouble free.



Personal Contribution to the Project

- Project leader;
- Problem definition;
- Overall system design and logic;
- Thesaurus-related issues;
- GUI appearance.



Scientific Plans

- Development of the automatic lexical acquisition techniques for the proposed thesaurus model.
- Development of the semi-automatic thesauri building techniques (concepts gathering and links designation).



Conclusion

- Main advantages:
Light and flexible solution for focused information retrieval.
- Goals:
Automatic and semi-automatic tools for thesauri building.
Design of a framework for more accurate evaluation of the approach.