Semantic Search and Analytics over Large Repository of Scientific Articles

Hung Son Nguyen[1], Dominik Ślęzak[1,2], Andrzej Skowron[1], and Jan G. Bazan[3,1]

[1] Institute of Mathematics, University of Warsaw
ul. Banacha 2, 02-097 Warsaw, Poland
[2] Infobright Inc.
ul. Krzywickiego 34 lok. 219, 02-078 Warsaw, Poland
[3] Chair of Computer Science, University of Rzeszów
ul. Rejtana 16A, 35-310 Rzeszów, Poland

> *Good mathematicians see analogies between theorems or theories. The very best ones see analogies between analogies.*
>
> – Stefan Banach [36]

Abstract. We present the architecture of the system aimed at search and synthesis of information within document repositories originating from different sources, with documents provided not necessarily in the same format and the same level of detail. The system is expected to provide domain knowledge interfaces enabling the internally implemented algorithms to identify relationships between documents (as well as authors, institutions et cetera) and concepts (such as, e.g., areas of science) extracted from various types of knowledge bases. The system should be scalable by means of scientific content storage, performance of analytic processes, and speed of search. In case of compound computational tasks (such as production of richer semantic indexes for the search improvements), it should follow the paradigms of hierarchical modeling and computing, designed as an interaction between domain experts, system experts, and appropriately implemented intelligent modules.

Keywords: semantic search, semantic information retrieval and synthesis, document analytics, document repositories, interactive and hierarchical computing, decision support, behavioral patterns, wisdom technology.

1 Introduction

This article outlines the proposed architecture, major requirements, assumptions and ideas related to the engine for semantic search and analytics developed within the "Interdisciplinary System for Interactive Scientific and Scientific-Technical Information" project financed by Polish Government in 2010-2013[1] The goal is to extend capabilities of the existing (usually Web-based, sometimes enterprise-based)

[1] `http://ismis2011.ii.pw.edu.pl/post_conference_event.php`

R. Bembenik et al. (Eds.): Intelligent Tools for Building a Scient. Info. Plat., SCI 390, pp. 1–8.
springerlink.com

semantic search engines by deeper analysis of the semantics of user requests. Given such assumptions, we will refer to our engine as *SONCA*, which stands for *S*earch based on *ON*tologies and *C*ompound *A*nalytics.

The engines such as SONCA should support a dialog of users with the text sources gathered within available repositories. It should lead not only to the search of significant documents, including their rankings [6,21], but also to intelligent systems helping users to specify and solve their problems [2,16].

An example of required functionality may be related to user's needs to understand the state of the art in a given domain of science. Surely, we can imagine various implementations addressing such requirement. The final answer to such understood user's query may summarize and synthesize various aspects, such as the most meaningful parts of representative documents related to the given area, characterization of the leading scientific groups and research initiatives, characterization of the most significant concepts and open problems, historical trends of progress in similar domains et cetera. Such framework needs to combine standard search capabilities based on keywords, dictionaries, glossaries, and ontologies [11,14,15] with hierarchical knowledge representation and reasoning [4,17,30], based on domain knowledge acquired from experts and users.

The idea of SONCA can be explained using the principles of interactive calculi on compound objects called information granules [18,29]. The user provides specification (in form of a query) in a language that is most often the natural language or its simplified fragment [40]. Then, the goal of the engine is to construct a compound information granule satisfying the formulated specification to a satisfactory degree. Establishing methods and algorithms for constructing such granules and measures of satisfiability of specifications by granules is crucial. Background for such computations may take a form of information systems [28] or relational database models [10]. Another aspect is to develop methods based on a dialog with users in order to understand their expectations (for example, a degree of advancement in a given scientific area).

Certainly, our engine should be scalable with respect to the volumes of data. It should be also at least partially scalable regarding the number of users, who, usually, would work rather with pre-computed semantic indexes, partially precomputed results and patterns, but with some fraction of users wanting to run more ad-hoc queries against the entire data. Thus, scalability should also refer to diversity of users' needs and kinds of usage of the stored data. Scalability can be secured at the algorithmic level, e.g. by adapting various forms of approximate reasoning [8,24], as well as at the level of data representation, data processing and data structures, by using database architectures that are aimed at data analysis and compound object handling [1,34].

2 System's Assumptions and Components

SONCA is aimed at extracting and constructing information based on text repositories originating from various libraries and publishers. Documents may be in different formats, such as XML, PDF, and scans of different quality. The system

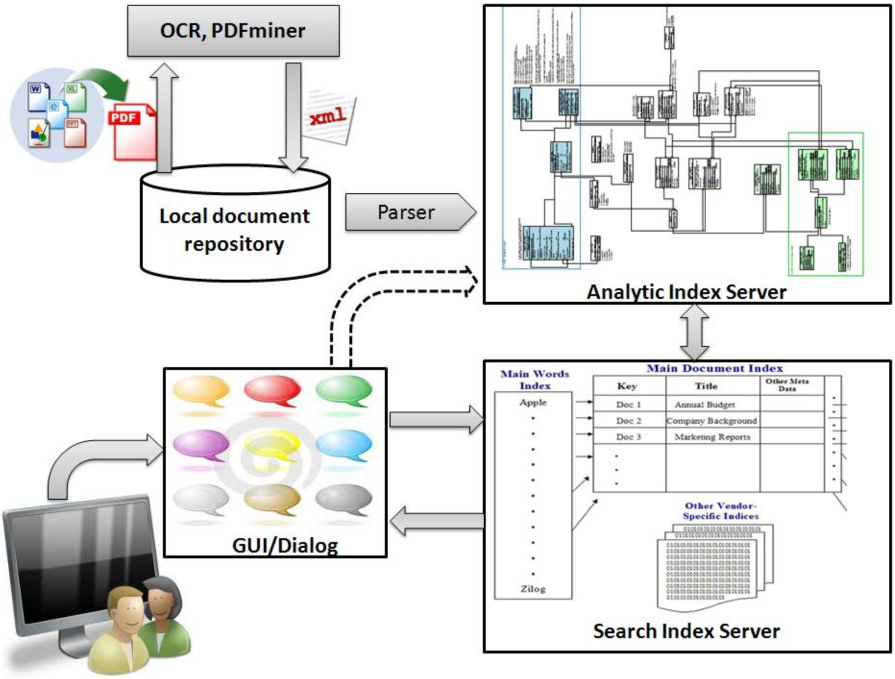

Fig. 1. The proposed architecture of SONCA

should be able to take into account various knowledge bases about the considered areas. There may be also independent sources of information about the analyzed objects, e.g., information about scientists who can be later identified as the authors of documents stored in a repository. Knowledge bases can be employed in multiple ways:

- As a means for communicating with users.
- As a means for injecting domain knowledge.
- As a specification of structures for objects and their relations.
- As a basis for discovering patterns useful for indexing objects.

We refer to [25] for further analysis how to use various sources of knowledge at various stages of our system.

Comparing to engines where documents are the main search process target, the proposed methodology has some additional features, e.g.:

- Ability to represent and search for various objects: concepts, authors, conferences, organizations, results, images, et cetera.
- Ability to use documents and knowledge bases to produce semantic indexes represented as information systems, further applicable in scalable search and information synthesis.

We refer to [26,35] for some examples of algorithms that prepare a background for semantic indexes from the above-mentioned combination of knowledge and information or use pre-computed semantic indexes for further processing.

Figure 1 outlines the system's architecture with its four major modules that can be integrated in multiple ways, also with their completely external counterparts in a more general framework:

- Local document repository of articles, including acquired information about them. They can occur in various forms (see below). They may be also given in metadata-only form.
- Analytic index server, including a framework for developing analytic operations that compute various intermediate structures leading eventually to the output semantic indexes.
- Search index server aimed at providing scalable external access to the latest available versions of semantic indexes.
- User interfaces aimed at decomposing user requests along the lines of domain-knowledge-driven hierarchical modeling.

Local document repository needs to be prepared for truly diversified forms of incomplete information. It is also important to equip it with interfaces to methods that can extract meaningful information and structures from the original files [5,31]. Below there are some observations that can influence local document repository and its interactions with other components:

- It can acquire various types of data, including:
  - *i.* Scanned articles.
  - *ii.* Digitalized articles (digitalized PDFs).
  - *iii.* Metadata – structured (but potentially heterogeneous, given different origins of data) descriptions of articles.
- Metadata can take various forms, including:
  - *i.* Input articles with no metadata assigned.
  - *ii.* Input articles with complete metadata.
  - *iii.* Metadata of articles that are not (yet) present in the database.
  - *iv.* Information about other objects (authors, institutes, domains).
- At the stage of receiving data we assume redundancy, e.g.: some article may be already present in the database, may be loaded to the database in future or never. Even metadata with no files assigned may be relevant for analysis, so they should be treated as articles with incomplete information.

Figure 2 illustrates our motivation to choose the MongoDB software [9] to establish the local document repository. It also outlines our choice of the Infobright [33,34] and Lucene [22] software in order to represent information about articles and other entities under consideration in two synchronized forms: repository-oriented and analytics-oriented. As already discussed in Section 1, our investigations led us to applying standard RDBMS data schemas to represent available data in a form that would be useful for more compound analytic processes. We refer to [20] for our further studies in this area.

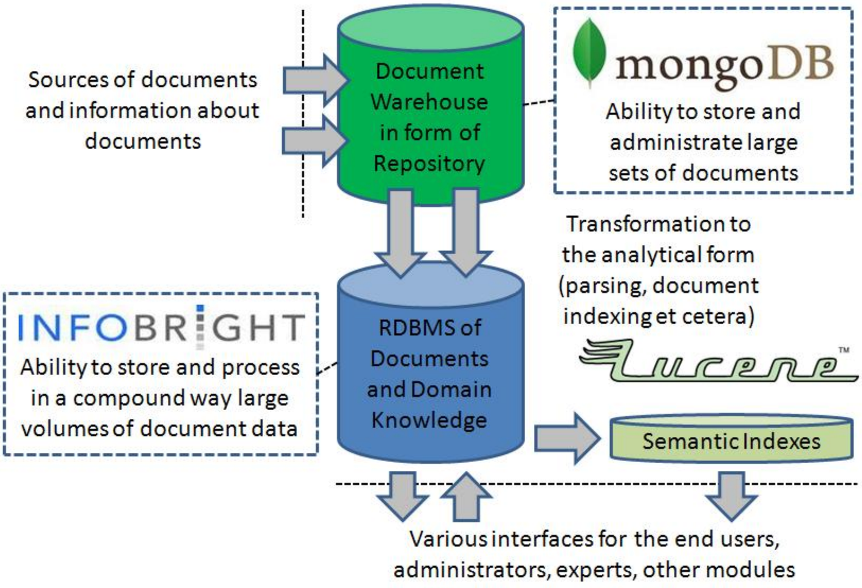
Sources of documents and information about documents
Document Warehouse in form of Repository
mongoDB
Ability to store and administrate large sets of documents
Transformation to the analytical form (parsing, document indexing et cetera)
INFOBRIGHT
Ability to store and process in a compound way large volumes of document data
RDBMS of Documents and Domain Knowledge
Lucene
Semantic Indexes
Various interfaces for the end users, administrators, experts, other modules

Fig. 2. The proposed software components used to implement local document repository, analytic index server, as well as data flow modules and mechanisms.

Certainly, an important aspect is also to create a framework for (possibly at least partially interactive) evaluation of each of the above components and the whole system with respect to the quality of results and efficiency of providing them to end users. While efficiency may be understood by quite standard means of speed and scalability of the search and analytic processes, evaluation of the quality may seek for analogies in information retrieval [7]. We refer to [39] for some detailed investigation in this area.

3 Further Perspectives

One of our motivations for developing the SONCA system is to extend functionality of the current enterprise and Web search engines towards the support for problem solving via enhanced search capabilities over various types of entities, information synthesis leading to answers that do not correspond to any single original entities, as well as letting advanced users omit some external interface layers and step down to the level of intermediate and core structures. While designing SONCA, we have been seeking for inspiration in many other projects and approaches, related to such domains as, e.g., social networks [23] and heterogeneous information networks [16]. Certainly there are plenty of details to be further discussed, e.g., how and in what form the results of search and analysis should be transmitted between modules and eventually reported to end users. With this respect, we can refer to, e.g., enriching original contents [2], approximate querying [33], or linguistic summaries of query results [19].

Practically all above aspects require a good means for employing domain knowledge, e.g., by developing methods similar to those for semantic Web [3,12]. In particular, one can consider learning behavioral patterns used by domain experts while solving problems [4,38]. Some hints for the research in this direction may follow from our experience with ontology-based approximation of compound concepts and identifying behavioral patterns in different applications. Interactions between Web/enterprise/repository resources and domain experts/users play an important role in learning such patterns. Thus, we plan to provide a framework for dialog between SONCA and its users, e.g., basing on interactive rough granular computations designed within the Wistech framework [18,32], where combination of personalization, interaction, and wisdom is claimed to lead to significant semantic search engine extensions.

In a broader sense the discussed challenges lead to such fundamental issues of mathematics as understanding of the concepts as proof, similarity of proofs, analogies between theorems, analogies between strategies of proofs used in different domains et cetera [36]. In real-life applications one should be ready to deal with even more compound situations. Due to uncertainty and/or necessity of overcoming infeasibility caused by computational complexity, we are forced to deal with interactive approximate reasoning schemes (plans, networks) over vague concepts instead of crisp reasoning schemes [37].

Further extensions of capabilities of engines similar to SONCA should be related to evolution of languages [13,27] applied to constructing and describing information granules related to articles and other entities that queries refer to. Therefore, we should take into account the strategies of automatic adaptive evolution of the language of patterns and granules. We should also expect evolution of domain knowledge and behavioral patterns, even if the language aimed at expressing them does not change. Thus, we should provide the means for storing new automatically learned concepts and behavioral patterns in knowledge bases that influence the processes of search and analytics.

Finally, our investigations should be continued also at a more technical level, ensuring sufficient scalability, performance, and interaction characteristics of the overall solution. For example, in [20] we outline some preliminary observations with respect to the usage of Infobright RDBMS software [34]. Although the outcomes are pretty optimistic, we will keep running scalability and performance tests over multiple platforms, paying a special attention to integration of different technologies supporting different parts of SONCA.

Acknowledgments. The authors are supported by the grant N N516 077837 from the Ministry of Science and Higher Education of the Republic of Poland and by the National Centre for Research and Development (NCBiR) under the grant SP/I/1/77065/10 by the Strategic scientific research and experimental development program: "Interdisciplinary System for Interactive Scientific and Scientific-Technical Information".

References

1. Agrawal, R., Ailamaki, A., Bernstein, P.A., Brewer, E.A., Carey, M.J., Chaudhuri, S., Doan, A., Florescu, D., Franklin, M.J., Garcia-Molina, H., Gehrke, J., Gruenwald, L., Haas, L.M., Halevy, A.Y., Hellerstein, J.M., Ioannidis, Y.E., Korth, H.F., Kossmann, D., Madden, S., Magoulas, R., Ooi, B.C., O'Reilly, T., Ramakrishnan, R., Sarawagi, S., Stonebraker, M., Szalay, A.S., Weikum, G.: The Claremont Report on Database Research. Commun. ACM 52(6), 56–65 (2009)
2. Agrawal, R., Gollapudi, S., Kannan, A., Kenthapadi, K.: Enriching Education through Data Mining. In: Kuznetsov, S.O., Mandal, D.P., Kundu, M.K., Pal, S.K. (eds.) PReMI 2011. LNCS, vol. 6744, pp. 1–2. Springer, Heidelberg (2011)
3. Badr, Y., Chbeir, R., Abraham, A., Hassanien, A.: Emergent Web Intelligence: Advanced Semantic Technologies. Springer, Heidelberg (2010)
4. Bazan, J.G.: Hierarchical Classifiers for Complex Spatio-temporal Concepts. T. Rough Sets 9, 474–750 (2008)
5. Betliński, P., Gora, P., Herba, K., Nguyen, T.T., Stawicki, S.: Semantic Recognition of Digital Documents. In: Bembenik, R., Skonieczny, Ł., Rybiński, H., Niezgódka, M. (eds.) Intelligent Tools for Building a Scientific Information Platform. Springer, Heidelberg (2011)
6. Breitman, K., Casanova, M., Truszkowski, W.: Semantic Web: Concepts, Technologies and Applications. Springer, Heidelberg (2007)
7. Butcher, S., Clarke, C., Cormack, G.: Information Retrieval: Implementing and Evaluating Search Engines. MIT Press (2010)
8. Cao, L.: Data Mining and Multiagent Integration. Springer, Heidelberg (2009)
9. Chodorow, K., Dirolf, M.: MongoDB: The Definitive Guide: Powerful and Scalable Data Storage. O'Reilly Media (2010)
10. Codd, E.: Derivability, Redundancy and Consistency of Relations Stored in Large Data Banks. SIGMOD Record 38(1), 17–36 (2009)
11. Colomb, R.: Ontology and the Semantic Web. IOS Press (2007)
12. Davies, J., Grobelnik, M., Mladenic, D.: Semantic Knowledge Management: Integrating Ontology Management, Knowledge Discovery, and Human Language Technologies. Springer, Heidelberg (2009)
13. Feldman, J.A.: From Molecule to Metaphor: A Neural Theory of Language (A Bradford Book). MIT Press (2006)
14. Gasevic, D., Djuric, D., Devedzic, V.: Model Driven Engineering and Ontology Development. Springer, Heidelberg (2009)
15. Gómez-Pérez, A., Fernández-López, M., Corcho, O.: Ontological Engineering with Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web. Springer, Heidelberg (2004)
16. Han, J.: Construction and Analysis of Web-Based Computer Science Information Networks. In: Kuznetsov, S.O., Ślęzak, D., Hepting, D.H., Mirkin, B.G. (eds.) RSFDGrC 2011. LNCS (LNAI), vol. 6743, pp. 1–2. Springer, Heidelberg (2011)
17. Helbig, H.: Knowledge Representation and the Semantics of Natural Language. Springer, Heidelberg (2006)
18. Jankowski, A., Skowron, A.: A Wistech Paradigm for Intelligent Systems. T. Rough Sets 6, 94–132 (2007)
19. Kacprzyk, J., Zadrożny, S.: Computing with words is an implementable paradigm: Fuzzy queries, linguistic data summaries, and natural-language generation. IEEE T. Fuzzy Systems 18(3), 461–472 (2010)

20. Kowalski, M., Ślęzak, D., Stencel, K., Pardel, P., Grzegorowski, M., Kijowski, M.: RDBMS Model for Scientific Articles Analytics. In: Bembenik, R., Skonieczny, Ł., Rybiński, H., Niezgódka, M. (eds.) Intelligent Tools for Building a Scientific Information Platform. Springer, Heidelberg (2011)
21. Ledford, J.L.: Search Engine Optimization Bible. Wiley (2009)
22. McCandless, M., Hatcher, E., Gospodnetić, O.: Lucene in Action, 2nd edn. Manning Publications (2010)
23. Mika, P.: Social Networks and the Semantic Web. In: Proc. of Int. Conf. on Web Intelligence (WI), pp. 285–291 (2004)
24. Nguyen, H.S.: Approximate Boolean Reasoning: Foundations and Applications in Data Mining, pp. 334–506 (2006)
25. Nguyen, L.A., Nguyen, H.S.: On Designing the SONCA System. In: Bembenik, R., Skonieczny, Ł., Rybiński, H., Niezgódka, M. (eds.) Intelligent Tools for Building a Scientific Information Platform. Springer, Heidelberg (2011)
26. Nguyen, S.H., Świeboda, W., Jaśkiewicz, G.: Extended Document Representation for Search Result Clustering. In: Bembenik, R., Skonieczny, Ł., Rybiński, H., Niezgódka, M. (eds.) Intelligent Tools for Building a Scientific Information Platform. Springer, Heidelberg (2011)
27. Nolfi, S., Mirolli, M.: Evolution of Communication and Language in Embodied Agents. Springer, Heidelberg (2010)
28. Pawlak, Z.: Information Systems Theoretical Foundations. Inf. Syst. 6(3), 205–218 (1981)
29. Pedrycz, W., Skowron, A., Kreinovich, V. (eds.): Handbook of Granular Computing. Wiley (2008)
30. Poggio, T., Smale, S.: The Mathematics of Learning: Dealing with Data. Notices of the AMS 50(5), 537–544 (2003)
31. Shinyama, Y.: PDFMiner: Python PDF Parser and Analyzer (2010), `http://www.unixuser.org/~euske/python/pdfminer/`
32. Skowron, A., Stepaniuk, J., Świniarski, R.W.: Approximation Spaces in Rough-Granular Computing. Fundam. Inform. 100(1-4), 141–157 (2010)
33. Ślęzak, D., Kowalski, M.: Towards Approximate SQL – Infobright's Approach. In: Szczuka, M., Kryszkiewicz, M., Ramanna, S., Jensen, R., Hu, Q. (eds.) RSCTC 2010. LNCS, vol. 6086, pp. 630–639. Springer, Heidelberg (2010)
34. Ślęzak, D., Wróblewski, J., Eastwood, V., Synak, P.: Brighthouse: An Analytic Data Warehouse for Ad-hoc Queries. Proc. VLDB Endow. 1(2), 1337–1345 (2008)
35. Szczuka, M., Janusz, A., Herba, K.: Clustering of Rough Set Related Documents with Use of Knowledge from DBpedia. In: Yao, J. (ed.) RSKT 2011. LNCS (LNAI), vol. 6954, pp. 394–403. Springer, Heidelberg (2011)
36. Ulam, S.: Analogies Between Analogies: The Mathematical Reports of S. M. Ulam and His Los Alamos Collaborators. University of California Press (1990)
37. Valiant, L.G.: Robust Logics. Artif. Intell. 117(2), 231–253 (2000)
38. Vapnik, V.: Learning Has Just Started (An interview with Vladimir Vapnik by Ran Gilad-Bachrach) (2008), `http://seed.ucsd.edu/joomla/index.php/articles/12-interviews/9-qlearning-has-just-startedq-an-interview-with-prof-vladimir-vapnik`
39. Wasilewski, P.: Towards Semantic Evaluation of Information Retrieval. In: Bembenik, R., Skonieczny, Ł., Rybiński, H., Niezgódka, M. (eds.) Intelligent Tools for Building a Scientific Information Platform. Springer, Heidelberg (2011)
40. Zadeh, L.A.: Computing with Words and Perceptions - A Paradigm Shift. In: Proc. of Int. Conf. on Parallel and Distributed Processing Techniques and Applications (PDPTA), pp. 3–5 (2010)