31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Exploitation of ontological resources for scientific literature analysis: searching genes and related diseases

Antonio Jimeno-Yepes, Rafael Berlanga-Llavori, and Dietrich Rebholz-Schuhmann

Abstract— **Ontological resources such as controlled vocabularies, taxonomies and ontologies from the OBO foundry are used to represent biomedical domain knowledge. The development of such resources is a time consuming task. Once they are finished they contribute to standardization of information representation, interoperability of IT solutions, literature analysis and knowledge discovery.**

Text mining comprises IT solutions for information retrieval (IR) and information extraction (IE). IR technology exploits ontological resources to select documents that fit best to the processed query, for example, through indexing of the literature content with concept ids or through disambiguation of terms in the query. IE solutions make use of the ontological labels to identify concepts in the text. The text passages that denote conceptual entries are then used either to annotate named entities or to relate the named entities to each other.

For knowledge discovery (KD) solutions the identified concepts in the scientific literature are used to relate entities to each other, e.g. to identify gene-disease relations based on shared molecular functions.

I. Introduction

DIFFERENT types of ontological resources are available in the biomedical domain. Controlled vocabularies are collections of terms, e.g., the Medical Subject Headings (MeSH terms) that are part of UMLS, and often include a taxonomic structure. More advanced ontological resources make use of further relation types, provide a meaningful definition of the concepts and ensure consistency parameters across the ontology (see OBO foundry). All resources provide concept ids as reference for every integrated term.

The development of ontologies is a time-consuming task. The formal representation of domain knowledge is one important step and the acquisition and confirmation of terms representing relevant concepts is another one. It can be expected that work efficiency and development progress can be improved, if textual and semantic resources such as the scientific literature can be exploited for the design and development work.

Large document collections like the World Wide Web or the biomedical scientific literature (Medline) are readily available, however neither one is currently available in a semantically structured representation. The extraction of information from both sources requires text-mining solutions. This information could be beneficial for the ontology development since literature resources provide a significant portion of the domain knowledge.

Text mining contributes to the generation of ontological resources, but also text mining profits from the use of ontological resources. For example, the use of an ontology can support the disambiguation of terms that represent different concepts and furthermore, the use of synonyms linked to individual concepts can enlarge the coverage of the text mining solution. The most basic text mining tasks that integrate ontologies into text mining solutions are the mapping of concept labels to terms in textual sources (e.g. named entity recognition) and the expansion of query terms in information retrieval solutions, which is a specialty of text mining. As has been shown, the combination of ontologies with text mining solutions leads to benefits in different IT approaches and their combined exploitation is developing into a dedicated research topic.

Examples of text mining tasks are text categorization, document retrieval and fact extraction from documents. In principle, the use of ontological knowledge can improve any of these tasks, since the integration of explicit semantics from the ontology supports one of the basic text mining tasks that is the mapping of concepts to terms in any type of document (e.g. named entity recognition).

II. Overview on Text Mining

A. *Information retrieval and information extraction*

In text mining systems, information retrieval (IR) and information extraction (IE) are usually interlinked (e.g. Figure 1). IR is used to retrieve relevant documents or parts of the document (e.g., paragraphs or sentences) to be possibly further processed by IE methods. The other way around, IE may feed identified results into an IR system to produce better results.

Any IR system pre-processes the documents: all documents are tokenized and the tokens are normalized. After this step, the document is represented with the

Manuscript received April 23, 2009. This work was supported in part by the EC STREP project BOOTStrep [FP6-028099].

A. Jimeno-Yepes is with the European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, U.K. (e-mail: yepes@ebi.ac.uk).

R. Berlanga-Llavori is with the Dept. of Computer Systems and Languages, Universitat Jaume I, Av. de Vicent Sos Baynat, s/n. 12071 Castelló de la Plana, Spain (e-mail: berlanga@lsi.uji.es).

D. Rebholz-Schuhmann is with the European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, U.K. (phone: +44-1223-492594; fax: +44-1223-494458; e-mail: rebholz@ebi.ac.uk).

978-1-4244-3296-7/09/$25.00 ©2009 IEEE

selection of tokens ("words") that have been identified in the text (called bag of words, BoW). The words are used to index the documents. Any query submitted to the retrieval engine is again decomposed into a BoW and is resolved against the index to generate a list of documents. The use of terminological and ontological resources improves the document retrieval, since ambiguities due to polysemy can be resolved before the indexing process.

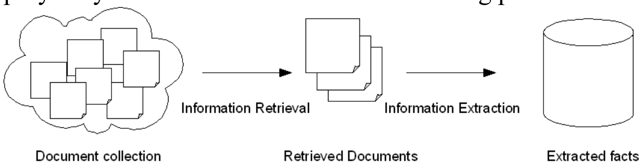

Figure 1: Information Retrieval and Information Extraction interaction

Information extraction deals with the identification of facts (e.g. entities, events, relations) from textual sources (see above), e.g. variants of the gene BRCA1 being involved in causing breast cancer. Information extraction components deliver detailed information that can be reused in information retrieval solutions. Combined with ontological resources, IE solutions enable better identification of concepts in the documents.

Even though the requirements for information extraction depend on the application, IE systems are usually composed of the same components [1] that can be combined in a pipeline of modules as shown in Figure 2. This pipeline approach modularizes the IE systems allowing the interchange of several components. As an engineering artifact, a common representation will guarantee the inter-operability of the components. Typical components in an IE system are:

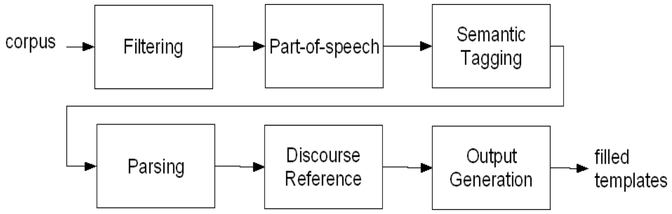

Figure 2: Information Extraction components

Very important IE tasks are the identification of named entities (e.g., genes), terms representing concepts and novel terms from the scientific literature.

B. Named entity recognition and normalization

In the identification of named entities, we can distinguish the identification of an entity mention from the normalization of the named entity. In the first case, the IE system delivers the boundaries of the named entities (e.g., PGHS-2 vs. PGH Synthase 2). In the latter case, the named entity has to be mapped to a concept id (e.g., PGHS-2 vs. UniProtKb:O62698, called "normalisation").

The normalization of named entities from the text to concepts in the ontology, i.e. the mapping of the potential surface form of a concept label in the text to the concept identifier in the ontological resource, has to tackle the following two basic problems. The first one is that a given concept is represented using different surface forms in the text, for example different representations of a given term (morpho-syntactic variation; e.g., PGHS-2 vs. PHS II) or simply different terms for the same concept (synonymy; e.g., COX-2 vs. PGHS-2). Both cases require that different surface forms be linked to the same concept id. The other basic problem arises from the situation that the same term may denote more than one concept (polysemy). In this case, the context of the term is further analyzed to resolve any ambiguity of the term in the text to derive the appropriate concept.

Resolving the ambiguity of polysemous terms requires special solutions. A large lexical resource, which proposes semantic types to given terms, can directly contribute to the resolution of ambiguous cases [2]. In other cases it is necessary to process the contextual information in the documents. Recently, special approaches take the topology of the ontology into consideration to disambiguate terms [3].

Contextual information of a term is commonly used to enable disambiguation reaching 98% on Medline abstracts [4]. Several disambiguation algorithms have been proposed that exploit the ontology topology and the context of the co-occurring terms to estimate the conceptual distance between the associated terms [5]. The contextual information is compared with a model of the concept based on its terminology and relations as expressed in the ontology.

III. TEXT MINING FOR POPULATING THE ONTOLOGY

Different techniques can be applied to normalize given surface forms to a concept label. In the case of yet unknown terms that have a high similarity to existing concept labels natural language processing techniques can be applied to include the missing entries [6]. It may happen that the terminological or ontological resources in the biomedical domain lack the required terminology for the mapping since not all terms for the concepts have been made available [7,8]. This gap in the resources is solved by adding the missing terms. This process is supported by term extraction tools, which process large amounts of text to identify statistically overrepresented terms that are assumed relevant for the domain [9]. In the case of the carotenoid pathway, 37 new and relevant concepts could be identified from 89,086 terms [10]. The other way around, mapping of existing terms (e.g., from the gene ontology) to text passages supports the identification of related terms [11, 12].

It has to be kept in mind that this task is constrained by the fact that content (e.g., facts) from the scientific literature is, in many cases, not factual but hypothetical and requires confirmation in the future [13]. For example, experimental results give first evidence on drug-gene or gene-diseases relations but require further confirmation through future experiments. Even names such as “CREB-binding Protein” convey that the true function and nature of the protein still has to be identified.

A. *Term and concept extraction*

Hierarchical clustering can be applied to generate groupings of terms based on the contextual information of the terms. The resulting structures can be verified by an ontology engineer or domain expert. [14] used hierarchical clustering to group gene-products (proteins) from the literature. The result was a collection of disjoint trees that were then merged by the knowledge expert. Term composition, like head-modifier, can give additional input that helps to add a taxonomical structure to the term repository. For instance, *colon cancer* is more specific than *cancer*, i.e. it is the sub-specification of the cancer to the organ type labeled the colon.

The context of the terms can be used to identify similar and related terms. [15] collected additional tokens from the context of terms. These tokens were integrated into a vector space model that was thereafter reduced to its main components based on principal component analysis (PCA). This approach led to the identification of hierarchical relations at 58 % precision. A similar solution from [16] added the POS information to discriminate better the words defining the context.

Other solutions to generate content for ontologies are based on language patterns. [17] and [18] have build dictionaries from extracted terms and have automatically assigned the terms to general categories. [19, 15] initiated his solution with an initial set of reliable extraction patterns that then have been combined with bootstrapping methods to identify additional extraction patterns. Their objective was the extraction of hyponym and hypernym relations without using a dictionary.

B. *Ontology Refinement*

Ontology refinement has the objective to transform an existing ontological resource to perform better in a task that requires a more specialized ontological resource [20]. Different techniques based on IE have been proposed, which typically identify statistically overrepresented terms with indication of domain relevance in contrast to frequently occurring common terms.

Proposed methods exploit term co-occurrence to identify novel information supported by statistical means. [21] selected association rules that specify relations between concepts based on the analysis of a set of documents describing hotels. [22] have compared GO to different other ontologies to identify conflicting concepts (e.g. circular definitions) and new synonyms that are then presented to the ontologists.

[23] extracted new terms from text and then placed them in the taxonomy or identifying taxonomic relations between existing concepts. In the biomedical domain, [24] propose an automated method to refine the Gene Ontology. The idea is to find rules based on GO terms variations for automatic expansion that is validated with the literature.

IV. Ontologies and Information Retrieval

Altogether, ontologies have been used in information retrieval to support query reformulation, semantic indexing and improved navigation of the search results.

A. *Query reformulation*

In Query Reformulation (QR) the query is transformed into a new representation (query expansion and refinement). QR refers to the different operations that are applied to the original user query in order to improve its performance. A user facing an information retrieval system has to consider how to transform his information need into a query representation in the system's query language in such a way that it is effective in terms of retrieval performance [25].

1) *Query Expansion*

Query expansion (QE) uses ontological resources as one source to gather expansion terms [26]. [27] tested manual query expansion and found that query expansion improves the performance in the case of short queries. However, she also measured that in principle queries that contain a larger number of terms can better specify the information need, but they usually lead to a decrease in the retrieval performance. This is mainly due to query drift, i.e. added terms contribute to the overall ambiguity induced by every single term and thus increase the likelihood for the inclusion of documents that do not properly fit to the query. In short, she noticed that each query term contributes its specific peculiarities linked to its inherent ambiguity and thus influences the performance of the IR problem.

[28] also could not demonstrate that the addition of the expansion terms into the original query improves retrieval performance because the added terms introduce too much emphasis in the query and the retrieval. This lead to the final result that the Boolean “OR” operator has to be used in combination with the added terms to improve the performance because the added terms have to act as alternatives to the original query terms. Only this approach does not put significant bias on the user needs through the introduction of the expansion terms.

For the biomedical domain, [29] and [30] explored on retrieval improvements for documents that refer to genes and proteins. They made use of UMLS and the content from the LocusLink database to achieve their goal. [31] have improved their query expansion through the integration of terms from UMLS and from the OSHUMED document collection into the queries. They have proposed to fit user queries to template specific queries in which the expansion terms are selected according to their relation to the original query terms.

Such relations between terms have been identified either in UMLS or in the document collection. [11] could show using the TREC Genomics data that the mapping of the user queries to MeSH concepts improves IR.

2) *Query Refinement*

The retrieval of a large number of documents is preferred by users of an IR system. However, in the case of large document collections like Medline or the Web this leads to large and huge numbers of documents, in particular if the system is queried with polysemous terms. This induces the interest in increasing the precision of the IR system to avoid an *information overload*. One particular case is the resolution of acronyms to several long-forms, i.e. theirs different expansions. For instance, if there is an interest in documents from Medline with reference to APC, the IR system will produce documents that report on the disease (or gene) called *adenomatous polyposis coli* (*APC*) or on the biological structure *anaphase-promoting complex* (*APC*). In both cases, the user query does not specify clearly, which result is the expected one.

In query refinement, the query is modified in order to filter out irrelevant documents and thus to increase the precision. Different techniques exist that make assumptions on the appropriate interpretation of the user query taking into consideration content of the documents in the underlying collection.

In all these cases, QR solutions make suggestions for improvements to the query. One suggestion is to select a subpart of the terms in the query or to specify better the meaning of given terms in the query. Proposed techniques used in IR solutions include the one from Scatter/Gather [32] where the retrieved documents are clustered and an informative label is attached to each cluster. The user then selects a label and the associated cluster as his query refinement and receives the documents in the cluster as the retrieval result. [33, 34] have investigated into research on the classification of the user queries and on the expected clustering results according to the different categories. For the biomedical domain the IR solution called SOPHIA represents a similar solution [35]. Other systems that rely on the integration of ontological resources, e.g. OntoRefiner [36], post-process the retrieved documents to display them in an alignment along a given lattice.

B. *Semantic indexing*

Several examples of semantic indexing have been proposed, i.e. [37] for general document collections, [38] for technical documents and [39] for biomedical scientific literature.

The semantic index that refers to concepts from an ontological resource in addition to lexical tokens (e.g. words and phrases) should enable disambiguation and normalization of terms to concepts. The ontological and textual sources are combined during the indexing process to improve the specificity of the tokens and to normalize terms to the same concept if the identified terms are synonymous.

Information extraction components like named entity recognition (NER) are used to identify the concepts in the documents. The concept ids are then integrated into the semantic index. Now the query has to be transformed in a similar way, which means that query terms have to be mapped to a conceptual representation, i.e. the concept id. This last step can be problematic if the query context is not sufficient to find the appropriate id or if several alternatives have to be resolved (ambiguity).

Finally yet importantly, semantic indexing enables querying the indexing engine with types only. In MedEvi for example, the query "[disease] and [protein]" is resolved to all combinations of any known gene and known disease that can be found in single sentences delivering gene-disease associations through the retrieval engine [40]. This approach avoids extensive query expansion techniques and query drift that would be a side effect.

C. *Organization of search results*

The large amount of documents returned by a retrieval system can be organized using a categorization scheme for the documents according to available taxonomic resources. The categorization scheme enables improved navigation of the search results based on the underlying taxonomic resources (e.g., ontologies), for example, the user can address directly and explore the documents attributed to the different subtopics and can match them to his information need. Two examples of an application that post-processes the retrieval results based on ontological and terminological resources are EBIMed and Facta [39, 41]. These solutions analyze the complete set of retrieved documents and identify the associations between the concepts contained in the documents (co-occurrence). All associations are then delivered in a table and for every association it is possible to recover the documents that support the evidence. Facta delivers documents that contain collocations of genes and diseases, whereas EBIMed filters out sentences where proteins are annotated with GO concepts, drugs and species.

In another solution, GoPubMed [42], retrieved documents are categorized into sets that all have been labeled with concepts that best represent the set. All labels come from existing taxonomic resources such as MeSH and the Gene Ontology (GO). This categorization enables the user to navigate to his preferred topics referring to a set of retrieved documents and to explore the used taxonomy overall to identify other subtopics of interest.

V. Ontologies for Information Extraction

Information extraction (IE) is the engineering science leading to solutions that gather facts from unstructured textual sources (e.g. documents). The information extraction need is expressed as a template that has to be filled with content from the document, i.e. the template slots have to be filled with pieces of text from the document by the IE system (cf. Figure 3). Several steps in the analysis have to be performed to produce a structured output expressed by the template.

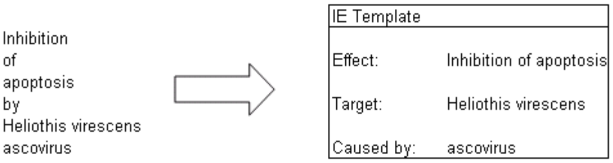

Figure 3 Information Extraction example

IE solutions are difficult to compare and to assess through comparison since each one deals with a different extraction need that is reflected in the available data sets [43]. Furthermore, only a few of the solutions are publicly accessible to measure performances.

In the biomedical domain, several data sets are now freely available for selected and standardized IE tasks: Biocreative I and II corpus, GENIA corpus, BioInfer dataset, AIMed corpus, Prodiser and the weakly annotated dataset for diseases [44, 45, 43, 46, 47, 48]. All these datasets cover only part of the information needs in the biomedical domain and furthermore, their focus is mainly limited to the identification of protein and gene names (PGN), the functional annotation of proteins and the interaction between proteins (protein-protein interaction, PPI).

The extraction of relations between entities is usually based on a set of rules applied on annotated text (e.g. based on part-of-speech and named entities). This set of rules can be improved by applying inference using domain knowledge producing simpler rules with similar or higher efficiency and easier to maintain.

The gene ontology (GO) has been most widely used to identify hidden knowledge in the scientific literature. For example, the annotation of proteins with GO terms from the scientific literature was efficient, if additional bioinformatics data resources contributed to the annotation of the proteins [49]. The annotation of genes with GO terms from the scientific literature and the same annotation for diseases lead to the generation of concept profiles based on GO terms. If these profiles were matched, then gene-disease associations could be identified (knowledge discovery) [50].

More advanced information extraction solutions make use of the compositional structure of the event representation in the Gene Regulation Ontology [7]. The IE solution then identifies features in the text that can be fitted to the compositional structure of the events. For example, the representation of an event has to include the involvement of the agent (e.g., the transcription factor, TF), the patient (the TF binding site) and additional features such as the binding of both involved entities and the directionality of the event [51]. The inference of the event is based on the ontological representation of the GRO. The inference helps to identify facts that are underspecified concerning the event representation in the ontological resource.

VI. CONCLUSION

Text mining supports researcher working on a new ontological resource in several aspects. Terms can be identified in the scientific literature and thus help to increase the coverage of relevant terms in the ontology (concept extraction). In addition, other methods support the refinement of the ontology.

The use of ontologies and thesauri is mandatory to achieve named entity normalization and semantic indexing. Both resources provide the concept ids and the IE task has to achieve the appropriate mapping to the semantic resource.

The integration of ontologies in information retrieval enables better categorization of the results. Improvements to the query performance are difficult to prove. Ontologies used in the information extraction process are still work in progress. The structure of existing ontological resources is not yet optimized to the needs of IE solutions, i.e. the ontological resource has to be combined with a lexical resource to efficiently support IE solutions.

REFERENCES

[1] J. Cowie and W. Lehnert., „Information extraction". *Communication ACM*, 39(1):80–91, 1996.

[2] P. Pezik, A. Jimeno-Yepes, V. Lee, and D. Rebholz-Schuhmann, "Static dictionary features for term polysemy identification," Building and evaluating resources for biomedical text mining," *LREC Workshop*, 2008.

[3] I. Spasic, S. Ananiadou, J. McNaught, and A. Kumar, "Text mining and ontologies in biomedicine: Making sense of raw text," *Briefings in Bioinformatics*, 6(3):239–251, 2005.

[4] S. Gaudan, H. Kirsch, and D. Rebholz-Schuhmann, "Resolving Abbreviations to Their Senses in Medline," *Bioinformatics* 21, vol. 18 (2005): 3658-64.

[5] E. Agirre and G. Rigau, "Word sense disambiguation using conceptual density," In Proceedings of the 16th conference on Computational linguistics, August, pp. 05–09, 1996.

[6] C. Jacquemin, "Spotting and discovering terms through natural language processing," *The MIT Press*, 2001.

[7] E. Beisswanger, M. Poprat, and U. Hahn, "Lexical Properties of OBO Ontology Class Names and Synonyms," In 3rd International Symposium on Semantic Mining in Biomedicine, 2008.

[8] A. Jimeno-Yepes, E. Jimenez-Ruiz, R. Berlanga-Llavori, D. Rebholz-Schuhmann, Use of shared lexical resources for efficient ontological engineering," Workshop "Semantic Web Applications and Tools for Life Sciences", Edinburgh, 2008.

[9] Spasic, I., D. Schober, S. A. Sansone, D. Rebholz-Schuhmann, D. B. Kell, and N. W. Paton. "Facilitating the Development of Controlled Vocabularies for Metabolomics Technologies with Text Mining," *BMC Bioinformatics* 9, vol. SUPPL, 5 (2008): Article S5.

[10] A. Waagmeester, P. Pezik, S. Coort, F. Tourniaire , C. Evelo, and D. Rebholz-Schuhmann. "Pathway enrichment based on text mining and its validation on carotenoid and vitamin A metabolism." OMICS (To appear)

[11] S. Gaudan, A. Jimeno Yepes, V. Lee, and D. Rebholz-Schuhmann, “Combining Evidence, Specificity, and Proximity Towards the Normalization of Gene Ontology Terms in Text,” *Eurasip Journal on Bioinformatics and Systems Biology* 2008 (2008): Article No 342746.

[12] D. Trieschnigg; P. Pezik; V. Lee; F. de Jong; W. Kraaij; and D. Rebholz-Schuhmann, “MeSH Up: Effective MeSH Text Classification for Improved Document Retrieval,” *Bioinformatics* 2009; doi: 10.1093/bioinformatics/btp249.

[13] J.I. Tsujii and S. Ananiadou, “Thesaurus or logical ontology, which do we need for mining text?” *Language Resources and Evaluation*, 39(1):77–90, September 2005.

[14] C. Blaschke and A. Valencia, “Automatic Ontology Construction from the Literature,” *Genome Informatics Series*, pp. 201–213, 2002.

[15] M.A. Hearst and J. O. Pedersen, “Reexamining the cluster hypothesis: Scatter/gather on retrieval results,” In SIGIR, pp. 76–84, 1996.

[16] D. Widdows, “Unsupervised methods for developing taxonomies by combining syntactic and statistical information,” In HLT-NAACL, 2003.

[17] E. Riloff and J. Shepherd, “A corpus-based approach for building semantic lexicons,” In Claire Cardie and Ralph Weischedel, editors, Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, pp. 117–124, ACL, Somerset, New Jersey, 1997.

[18] B. Roark and E. Charniak, “Noun-phrase co-occurence statistics for semiautomatic semantic lexicon construction,” In COLING-ACL, pp. 1110–1116, 1998.

[19] M.A. Hearst, “Automatic acquisition of hyponyms from large text corpora,” Technical Report S2K-92-09, 1992.

[20] A. Jimeno-Yepes, R. Berlanga-Llavori, D. Rebholz-Schuhmann, “Ontology refinement for improved information retrieval,” Information Processing & Management: Special Issue on Semantic Annotations in Information Retrieval, 2009 (to appear).

[21] A. Maedche and S. Staab, “Ontology learning for the semantic web,” IEEE Intelligent Systems, 16(2):72–79, 2001.

[22] J. Köhler, K. Munn, A. Ruegg, A. Skusa, and B. Smith, “Quality control for terms and definitions in ontologies and taxonomies,” *BMC Bioinformatics*, 7:212, 2006.

[23] R. Navigli and P. Velardi, “Automatic adaptation of wordnet to domains,” In Proceedings of 3rd International Conference on Language Resources and Evaluation, 2002.

[24] J.B. Lee, J.J. Kim, and J.C. Park, “Automatic extension of gene ontology with flexible identification of candidate terms,” *Bioinformatics*, 22(6):665–670, 2006.

[25] J.J. Kim, and D. Rebholz-Schuhmann, “Categorization of Services for Seeking Information in Biomedical Literature: A Typology for Improvement of Practice,” *Brief Bioinform* (2008a), 9(6):452-465.

[26] E. Efthimiadis, “Query expansion,” In: Williams, Martha E., ed Annual Review of Information Systems and Technologies (ARIST), v31, pp. 121–187, 1996.

[27] E.M. Voorhees, “Query expansion using lexical-semantic relations,” In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 61–69, Springer-Verlag New York, Inc., 1994.

[28] J.Y. Nie and F. Jin, “Integrating logical operators in query expansion in vector space model,” In Workshop on Mathematical/Formal Methods in Information Retrieval, 25thACM-SIGIR, Tampere, Finland, volume 8, 2002.

[29] A.R. Aronson and T. C. Rindflesch, “Query expansion using the UMLS Metathesaurus,” In Amia, 1997.

[30] Z. Liu and W.W. Chu, “Knowledge-based query expansion to support scenario-specific retrieval of medical free text,” In SAC ’05: Proceedings of the 2005 ACM symposium on Applied computing, pp. 1076–1083, New York, NY, USA, 2005.

[31] W.W. Chu, Z. Liu, and W. Mao, “Textual document indexing and retrieval via knowledge sources and data mining,” In Communication of the Institute of Information and Computing Machinery(CIICM), Taiwan, 5(2), 2002.

[32] M.A. Hearst, D.R. Karger, and J.O. Pedersen, “Scatter/gather as a tool for the navigation of retrieval results,” In Working Notes AAAI Fall Symp, AI Applications in Knowledge Navigation, 1995.

[33] W. Pratt, M.A. Hearst, and L.M. Fagan, “A knowledge-based approach to organizing retrieved documents,” In AAAI/IAAI, pp. 80–85, 1999.

[34] W. Pratt and H. Wasserman, “QueryCat: Automatic Categorization of MEDLINE Queries,” *Journal-American Medical Informatics Association*, 7:655–659, 2000.

[35] D. Patterson, N. Rooney, V. Dobrynin, and M. Galushka, “Sophia: A novel approach for textual case-based reasoning,” In IJCAI, pp. 15–20, 2005.

[36] B. Safar, H. Kefi, and C. Reynaud, “OntoRefiner, a user query refinement interface usable for Semantic Web Portals ,” In Applications of Semantic Web technologies to web communities, Workshop ECAI, 2004.

[37] M. Baziz, M. Boughanem, and N. Aussenac-Gilles, „The Use of Ontology for Semantic Representation of Documents,” In proceedings of Semantic Web and Information Retrieval Workshop, SIGIR, pp. 38-45, 2004.

[38] J. Ambroziak, and W.A. Woods, “Natural Language Technology in Precision Content Retrieval,” Technical report, Sun Microsystems, Inc., 1998

[39] D. Rebholz-Schuhmann, H. Kirsch, M. Arregui, S. Gaudan, M. Riethoven, P. and Stoehr, “EBIMed – Text crunching to gather facts for proteins from Medline,” *Bioinformatics* 23, vol. 2: e237-e244, 2007.

[40] J.J. Kim, P. Pezik, and D. Rebholz-Schuhmann, "Medevi: Retrieving Textual Evidence of Relations between Biomedical Concepts from Medline,” *Bioinformatics* 24, vol. 11 (2008b): 1410-12.

[41] Y. Tsuruoka, J. Tsujii, and S. Ananiadou, “FACTA: a text search engine for finding associated biomedical concepts,” *Bioinformatics*, 2008 Nov 1;24(21):2559-60, Epub 2008 Sep 4.

[42] A. Doms and M. Schroeder, “GoPubMed: exploring PubMed with the gene ontology,” *Nucleic Acids Research*, vol 33, 2005.

[43] R. Bunescu, R. Ge, R.J. Kate, E.M. Marcotte, R.J. Mooney, A.K. Ramani, and Y.W. Wong, “Comparative experiments on learning information extractors for proteins and their interactions,” In Comparative Experiments on Learning Information Extractors for Proteins and their Interactions. Journal of Artificial Intelligence In Medicine, 33 (2): 139-155.

[44] J. Kim, et al., “GENIA corpus–semantically annotated corpus for bio-text mining,” *Bioinformatics* 2003, 19(Suppl 1):i180-i182.

[45] S. Pyysalo, F. Ginter, J. Heimonen, J. Björne, J. Boberg, J. Järvinen, and T. Salakoski, “BioInfer: a corpus for information extrac-tion in the biomedical domain,” *BMC Bioinformatics*, 9;8:50.

[46] M. Krallinger, R. Malik, and A. Valencia, “Text mining and protein annotations: the construction and use of protein description sentences,” *Genome Inform* 2006, 17(2):121-130.

[47] M. Craven, and J. Kumlien, “Constructing Biological Knowledge Bases by Extracting Information from Text Sources,” In Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, AAAI Press; 1999:77-86.

[48] S. Ray, and M. Craven, “Representing Sentence Structure in Hidden Markov Models for Information Extraction,” In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, Seattle, WA: Morgan Kaufmann; 2001:1273-1279.

[49] S. Jaeger, S. Gaudan, U. Leser, and D. Rebholz-Schuhmann, "Integrating Protein-Protein Interactions and Text Mining for Protein Function Prediction,” *BMC Bioinformatics* 9, vol. SUPPL, 8 (2008): Article S2.

[50] C. Grabmüller, A. Jimeno-Yepes, V. Lee, and D. Rebholz-Schuhmann, “Identification of novel human disease candidate genes with pathogenetic implications through large-scale literature analysis,” (In review).

[51] U. Hahn, K.Tomanek, E. Buyko, J.J. Kim, and D. Rebholz-Schuhmann, “How Feasible and Robust is the Automatic Extraction of Gene Regulation Events? A Cross-Method Evaluation under Lab and Real-Life Conditions,” *BioNLP* 2009.