Ontology-based MEDLINE Document Classification

Fabrice Camous[1], Stephen Blott[1], and Alan F. Smeaton[2]

[1] School of Computing,
fcamous@computing.dcu.ie
sblott@computing.dcu.ie
[2] Centre for Digital Video Processing & Adaptive Information Cluster,
asmeaton@computing.dcu.ie
Dublin City University,
Glasnevin, Dublin 9, Ireland

Abstract. An increasing and overwhelming amount of biomedical information is available in the research literature mainly in the form of free-text. Biologists need tools that automate their information search and deal with the high volume and ambiguity of free-text. Ontologies can help automatic information processing by providing standard concepts and information about the relationships between concepts. The Medical Subject Headings (MeSH) ontology is already available and used by MEDLINE indexers to annotate the conceptual content of biomedical articles. This paper presents a domain-independent method that uses the MeSH ontology inter-concept relationships to extend the existing MeSH-based representation of MEDLINE documents. The extension method is evaluated within a document triage task organized by the Genomics track of the 2005 Text REtrieval Conference (TREC). Our method for extending the representation of documents leads to an improvement of 18.3% over a non-extended baseline in terms of normalized utility, the metric defined for the task.

1 Introduction

In the current era of fast mapping of entire genomes, genomic information is becoming extremely difficult to search and process. Biologists spend a considerable part of their time searching the research literature. An important task for Genomic database curators is to locate experimental evidence in the literature to annotate gene records.

The increasing and overwhelming volume to search through, coupled with the ambiguity [19] of unstructured information found in free-text make manual information processing prohibitively expensive. Biomedical ontologies, when available, can help disambiguate the information expressed with natural languages: they offer standard terms that only relate to specific concepts and therefore eliminate the occurrence of synonyms and polysems. They also contain information about the relationships between concepts and this information could be used to express semantic similarities between concepts.

In the biomedical literature, the Medical Subject Headings (MeSH)[3] is used to annotate the conceptual content of the MEDLINE database records. The MeSH ontology is organized in several hierarchies that indicate the level of specificity of the MeSH concepts.

We hypothesize that the semantic information contained in the MeSH network can benefit the representation of documents. Our approach extends initial MeSH-based document representations with additional concepts that are semantically close within the MeSH semantic network. The document representation extension method described in this paper is domain-independent and can be applied to other ontologies.

Our approach is evaluated in the context of a binary classification or triage of documents. The triage corresponds to a stage in the information retrieval process where the possibly relevant documents are selected from the mass of non-relevant documents before being thoroughly examined later on. In particular, our method is assessed with a document triage task organized by the Genomics track of the 2005 Text REtrieval Conference (TREC)[4].

The paper is organized in the following manner: section 2 describes our ontology-based document representation extension method. Section 3 presents the evaluation framework of our approach, including related work and the results we have obtained. Finally section 4 concludes with future research directions.

2 Methodology

This section describes our ontology-based document representation extension method. Our extension approach can apply to any ontology-based document representation. In this paper, however, we focus on MEDLINE records and the MeSH-based content descriptions they contain. Some background information about the MEDLINE database and the MeSH ontology is given in section 2.1. Our method includes comparing concepts semantically with the MeSH ontology hierarchy. Therefore, some background about network-based semantic measures is also available in section 2.2.

2.1 MEDLINE and MeSH

MEDLINE, the U.S. National Library of Medicine (NLM)[5] bio-medical abstract repository, contains approximately 14 million reference articles from around 4,800 journals. Around 400,000 new records are added to it each year. Despite the growing availability of full-text articles on the Web, MEDLINE remains in practice a central point of access to bio-medical research [7, 8].

The MEDLINE record fields include text-based fields, the title and abstract fields, and ontology-based fields: the MeSH fields. Most MEDLINE records contain 10-12 MeSH term fields. MeSH is a biomedical controlled vocabulary pro-

[3] http://www.nlm.nih.gov/mesh/meshhome.html
[4] http://trec.nist.gov/
[5] http://www.nlm.nih.gov/

duced by the U.S. NLM and used since 1960. MeSH 2004 includes 22,430 descriptors, 83 qualifiers, and 141,455 supplementary concepts. Descriptors are the main elements of the vocabulary. Qualifiers are assigned to descriptors inside the MeSH fields to express a special aspect of the concept. Both descriptors and qualifiers are organized in several hierarchies. Figure 1 shows a simplified representation of the descriptor "Diseases" hierarchy. The 83 qualifiers are or-

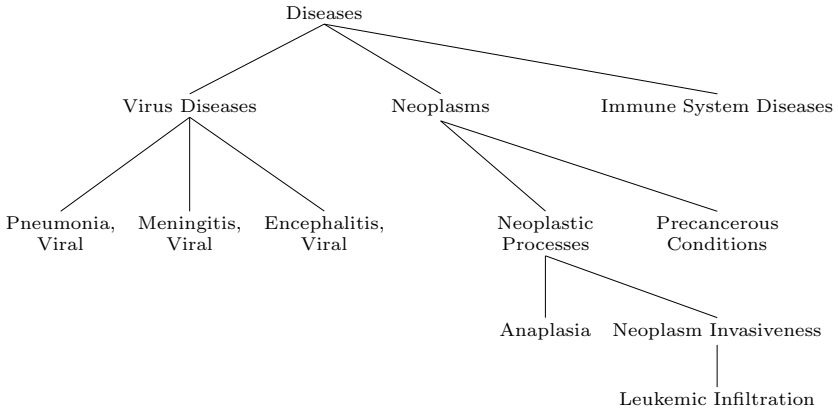

Fig. 1. A simplified representation of the Diseases hierarchy.

ganized in shallow hierarchies with the most general qualifiers at the top of the hierarchies.

The relationships between descriptor or qualifier nodes in the MeSH network are of the "broader/narrower than" type [2]. The "narrower than" relationship is close to the hypernymy (is a) relationship, but it can also include a meronymy (part of) relationship. Inversely, the "broader than" relationship is close to the hyponymy (has instance) relationship, and can also include a holonymy (has a) relationship. In MeSH, one term is narrower than another if the documents it retrieves in a search are contained in the set of documents retrieved by the broader term.

An example of MEDLINE record, describing a full-text article, is shown in Figure 2. It includes textual fields, such as title and abstract, as well as MeSH fields (denoted MH). The MeSH fields present several advantages over textual fields: Unlike the free-text content of the title/abstract fields, the MeSH fields unambiguously associate a single term to a single concept. In addition, MeSH terms are assigned to the records after the examination of the entire research article by human indexers. Consequently, it covers more conceptual ground than the title/abstract free text. Finally, previous work by [5] showed that indexing consistency was high amongst NLM human indexers: the authors reported a mean agreement in index terms between 33%-74% for different experts.

PMID - 10605436
TI - Concerning the localization of steroids in centrioles and basal bodies by immunofluorescence.
AB - Specific steroid antibodies, by the immunofluorescence technique, regularly reveal fluorescent centrioles and cilia-bearing basal bodies in ...
AU - Nenci I
AU - Marchetti E
MH - Animals
MH - Centrioles/*ultrastructure
MH - Cilia/ultrastructure
MH - Female
MH - Fluorescent Antibody Technique
MH - Human
MH - Lymphocytes/*cytology
MH - Male
MH - Organelles/*ultrastructure
MH - Rats
MH - Rats, Sprague-Dawley
MH - Respiratory Mucosa/cytology
MH - Steroids/*analysis
MH - Trachea

Fig. 2. A MEDLINE record example (PMID: PubMed ID, TI: title, AB: abstract, AU: author, MH: MeSH term)

A MEDLINE MeSH field is a combination of a MeSH descriptor with zero or more MeSH qualifiers. In Figure 2, "Centrioles/*ultrastructure" is the combination of descriptor "Centrioles" with qualifier "ultrastructure". MeSH fields can describe major themes of the article (a concept that is central to the article) or minor themes (a peripheral concept). A star is used to distinguish the major themes from the minor ones. Therefore the association "Centrioles/*ultrastructure" is a major theme of the MEDLINE record of figure 2, along with "Organelles /*ultrastructure" and "Steroids/*analysis".

2.2 Network-based Semantic Measures

Network-based measures are usually classified into two groups in the literature: edge-based and information-based measures [2]. Edge-based measures rely mainly on the information contained in the network. For example, the position of a given concept in the network and the number of links from this concept to other concepts provide information about the semantic proximity of this concept to others. The depth of a concept (distance to the root concept), in the case of a hierarchical network, also gives information about the level of specificity of this concept. On top of network information, information-based measures introduce external information about the nodes of the network from their distribution in a

corpus. They are also called node-based measures, as the additional information is about the nodes, and hybrid measures, as they combine network and corpus knowledge. Network-based measures can also be analyzed at the inter-concept level (comparing two concepts) and at the inter-document level (comparing two groups of concepts). Some measures work at both levels but others are limited to the inter-concept level.

When using the network to compare two concepts, a simple edge-based approach is to count the number of links, or edges, that separate them in the ontology. As there can be several possible paths between two concepts in the network, a further step is to decide that the shortest path between the two concepts gives us a measure of the semantic distance between them. Rada et al. [13] uses such an approach while comparing two concepts with their "Distance" inter-concept measure. With the "Distance" measure, the semantic distance between two concepts A and B, Distance (A, B), is the minimum number of edges separating A and B in the network. For example, we can calculate a semantic distance for pairs of concepts in the simplified "Diseases" hierarchy of Figure 1:
Distance ("Pneumonia, Viral", "Meningitis, Viral") = 2.
Distance ("Pneumonia, Viral", "Neoplatic Processes") = 4.
Distance ("Neoplatic Processes", "Precancerous Conditions") = 2.
Distance ("Anaplasia", "Neoplasm Invasiveness") = 2.
According to Distance, the second pair of concepts contains two concepts, "Pneumonia, Viral" and "Neoplatic Processes", that are more semantically distant (Distance=4) than the concepts contained in the other three pairs.

To compare groups of concepts, Rada et al. [13] defines an extended version of the inter-concept "Distance" measure:

$$Distance\left(X_1 \wedge \cdots \wedge X_k, Y_1 \wedge \cdots \wedge Y_m\right) = \frac{1}{km} \sum_{i=1}^{k} \sum_{j=1}^{m} Distance\left(X_i, Y_j\right)$$

Rada et al.'s extended measure gives the semantic distance between two "conjunctive concepts" or groups of concepts that contain k X_i and m Y_j "elementary" concepts respectively.

Another method considers only the best semantic match amongst concepts of group B for each concept in group A. This method gives an asymmetrical measure expressing the semantic contribution of A concepts in relation to B. By switching A and B, we can combine the two asymmetrical measures into a symmetrical one. Azuaje et al. in [1] uses such a measure:

$$Dist_{azu}\left(A, B\right) = \frac{1}{k+m} \times \left(\sum_{i=1}^{k} \min_{j} \left(Distance\left(X_i, Y_j\right)\right) + \sum_{j=1}^{m} \min_{i} \left(Distance\left(X_i, Y_j\right)\right) \right)$$

where A and B are two groups of k and m concepts respectively, and Distance is Rada et al.'s inter-concept semantic distance.

2.3 Our Method

MeSH-only document representations are the starting point of our extension method. Descriptors and qualifiers are chosen as the minimal units of information or features. Associations between descriptors and qualifiers found in the MeSH fields (see section 2.1) are split and a qualifier appearing in several associations is considered to appear only once in the document. MeSH field distinctions between major and minor themes (see again section 2.1) are also ignored. Keeping descriptors and qualifiers as minimal information units allows to keep track of the concepts they represent and their locations in the MeSH hierarchies.

MeSH-based document representations are expressed with vectors containing 22,513 elements (22,430 descriptors and 83 qualifiers). Most MEDLINE documents contain 10-12 MeSH fields that each contain one descriptor associated optionally to one or more qualifiers. This means that originally the MeSH-based document vectors contain essentially zero values.

The MeSH network can be used to add new MeSH terms to the original document MeSH content. Descriptors and qualifiers have locations in the MeSH hierarchies from which an evaluation of their semantic similarity can be derived. For example, if "Neoplastic Processes" is found in a MeSH field, it can be assumed that the document is also about "Neoplasms" or "Anaplasia", to some degree (see Figure 1).

In the vector representation, a weighting scheme can be used to distinguish between the original MeSH elements and the ones derived from the extension process. In our experiment, the original MeSH elements get a weight $w_o = 1$ whereas derived MeSH elements get a weight w_d, $0 \leq w_d < 1$, depending on how semantically close they are to the original MeSH representation.

The MeSH network is separated into several descriptor and qualifier hierarchies. In order to be able to compare all MeSH terms with each other, an artificial "MeSH" root node is placed at the top of all the hierarchies as indicated in figure 3.

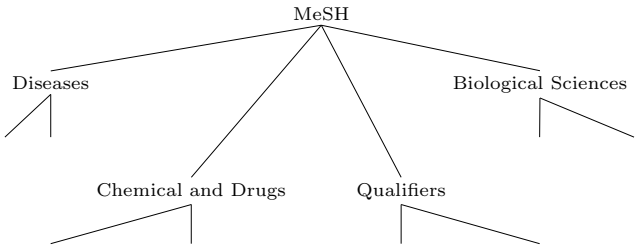

Fig. 3. MeSH root node add-up (only a few child nodes depicted for clarity)

Given that a MeSH term m can have k tree locations $tree\_loc_i$ (1.8 on average), and an original MeSH-based document representation can contain elements

corresponding to p tree locations $tree\_loc_j$, the derived weight w_dm of m is calculated with the following formula:

$$w_dm = 1 - \left(\min_k \left(\min_p \left(Distance \left(tree\_loc_i, tree\_loc_j \right) \right) \right) / max\_dist \right)$$

The maximum distance in the MeSH network between two tree locations, max_dist, is 23 edges. Distance is Rada et al.'s [13] inter-concept edge-based distance presented in section 2.2.

3 Evaluation

Our document representation extension method is evaluated in the context of document binary classification or triage. Our evaluation framework uses a MEDLINE triage task organized by the Genomics track of the 2005 Text REtrieval Conference (TREC). This task is described in section 3.1. Related work is reviewed in section 3.2. The document classification is done with the SVM^{light} software [10] (section 3.3), and the results are presented in section 3.4.

3.1 TREC 2005 Genomics Track GO Triage task

The Text REtrieval Conference (TREC) includes a Genomics track since 2003. The TREC guidelines and common evaluation procedures allow research groups from all over the world to evaluate their progress in developing and enhancing information retrieval systems.

One of the tasks of the 2004 and 2005 Genomics track was a biomedical document triage task for gene annotation with the Gene Ontology (GO) [6]. GO is used by several model organism databases curators in order to standardize the description of genes and gene products. The triage task simulated one of the activities of the curators of the Mouse Genome Informatics (MGI) [4]. MGI curators manually select biomedical documents that are likely to give experimental evidence for the annotation of a gene with one or more GO terms.

For both 2004 and 2005 GO triage tasks, the same subset of three journals from 2002 and 2003 was used. The subset contained documents that had been selected or not for providing evidence supporting GO annotation. The 5837 documents from 2002 were chosen as training documents, and the 6043 from 2003 as test documents. In 2004, they contained 375 and 420 positive examples (documents labeled relevant), respectively. In 2005, the number of positive examples for the training and test documents was updated to 462 and 518, respectively.

The triage task was evaluated with a normalized utility measure $U_{norm} = U_{raw}/U_{max}$ with U_{max} being the best possible score. U_{raw} was calculated with the following formula:

$$U_{raw} = (u_r \times relevant\_docs\_retrieved) + (u_{nr} \times nonrelevant\_docs\_retrieved)$$

where u_r is the relative utility of a relevant document and u_{nr} the relative utility of a non-relevant document. With u_{nr} set at -1 and u_r assumed positive, u_r

was determined by preferred values for U_{norm} in 4 boundary cases: completely perfect prediction, all documents judged positive (triage everything), all documents judged negative (triage nothing), and completely imperfect prediction. With different numbers of positive examples for classifcation between 2004 and 2005, $u_r = 20$ and $u_r = 11$ were chosen in 2004 and 2005, respectively.

Our experiment used the updated number of positive examples of 2005 for the classification, and the values $u_r = 11$ and $u_{nr} = -1$. A detailed description of TREC 2004 and 2005 Genomics track tasks can be found in [7, 8].

3.2 Related Work

Most approaches used in the 2004 and 2005 GO triage task extract features from various text fields along with the MeSH fields. In contrast and similarly to us, Seki et al. [17] and Lee et al. [11] experiment with MeSH-only document representations in the 2004 and 2005 GO triage task respectively.

Seki et al. extracts features from the MeSH fields to train a Naive Bayes Classifier. A gene name filter is then used to eliminated false positive. The gene name list includes gene names appearing only in negative examples and names appearing in a certain percentage of negative examples. A normalized utility of 0.434 is obtained with a gene filter at 10% (gene appearing in at least 10% of negative documents added to the filter list) and of 0.342 with a gene filter at 5%. Lee et al. uses a SVM classifier and experiments with several feature sets including title/abstract terms, MeSH terms, figure/table captions, and combinations of the former three. The MeSH-based features yields the best normalized utility (0.4968) out of the various feature sets. Seki el al.'s and Lee et al.'s MeSH-only representations both correspond to our MeSH-only un-extended document representations (our baseline).

However, the best results for the 2004 and 2005 triage tasks used methods that extract features from other fields than MeSH (title, abstract, full-text) and use domain-dependant techniques (term extraction, gene filtering). Dayanik et al. [3] obtains the best normalized utility (0.6512) of the 2004 GO triage task. Documents are represented with features extracted from the title, abstract and MeSH fields of the MEDLINE document format. Niu et al. [12] achieves the best normalized utility (0.5870) of the 2005 GO triage task using the SVM^{light} software [10]. Features are first extracted from full-text articles. The extraction is followed by porter stemming, stopwording, and a domain-specific term extraction method that is using corpus comparison.

3.3 Text Categorization and SVM^{light}

A discussion on text categorization and machine learning techniques is beyond the scope of this paper and is found elsewhere [16]. Our experiment focuses on evaluating an ontology-based document representation extension method and does not aim at comparing several text categorization techniques.

We use the SVM^{light} software which is an implementation of the Support Vector Machine method by Joachims [10]. The SVM learner defines a decision

surface between the positive and negative examples of the training set. SVM training leads to a quadratic optimization problem and learning from large training sets can quickly become computationally expensive. The SVM^{light} implementation allows for large-scale SVM learning at lower computational cost.

We use the default settings for the learning module (svm_learn) and the classification module (svm_classify) of SVM^{light}. For svm_learn, default settings include the use of a linear kernel. The only modification is the setting of parameter j (cost-factor by which training errors on positive examples out-weight errors on negative examples, default being 1) to 11 similarly to Subramaniam et al. [18]. The j parameter allows to tune the classifier to the difference between u_r ($= 11$), the relative utility of a relevant document, and u_{nr} ($= -1$), the relative utility of a non-relevant document (see section 3.1).

3.4 Results

We experimented with 3 threshold values for w_dm to add new MeSH terms to the original MeSH-based document representation: 0.5, 0.3, and 0.2. Lower threshold values produced files that were too large for the SVM^{light} software to process. Table 1 shows the results of our MeSH-network-based document representation extension method in terms of Precision, Recall, F-Score and Normalized Utility [12]. The results are compared to the simple use of the original MeSH content for document representation (Ori. MeSH Rep.) and to the best result from the 2005 GO triage task in terms of Normalized Utility. The results correspond to the classification of the test set documents after the learning on the training set documents with SVM^{light}.

Table 1. Result for the network-based MeSH document representation extension

	Precision	Recall	F-Score	Norm. Utility
Ori. MeSH Rep.	0.2980	0.6139	0.4013	0.4824
Ext. 0.5	0.3006	0.6197	0.4048	0.4886 (+1.3%)
Ext. 0.3	0.2643	0.7027	0.3842	0.5249(+8.8%)
Ext. 0.2	0.2034	0.8861	0.3308	0.5706(+18.3%)
Best 2005	0.2122	0.8861	0.3424	0.5870

The following formulae are used for Precision, Recall, and F-Score:

$$Precision = \frac{relevant\,documents\,retrieved}{documents\,retrieved}$$

$$Recall = \frac{relevant\,documents\,retrieved}{relevant\,documents}$$

$$F - Score = \frac{2 \times Precision \times Recall}{Precison + Recall}$$

Norm. Utility corresponds to the Normalized Utility defined in section 3.1.

The extension of MeSH-based document representations with the MeSH network leads to a drop in Precision but an increase in Recall. However, because of the importance of Recall in this particular triage task and therefore in the utility function, our document representation extension has a positive impact on the Normalized Utility. With a threshold at 0.2 for the introduction of new elements, the Normalized Utility goes up by 18.3% compared to our simple non-extended approach (Ori. MeSH Rep. in table 1). Our result with extension threshold at 0.2 is also a 14.9% improvement on the best MeSH-only document representation approach [11] of the GO triage task (see section 3.2).

Moreover, with a value of 0.5706, the Normalized Utility approaches the best value for the 2005 GO triage task (0.5870) and positions our method inside the top 5 runs from a total of 47. However, the best run of the track used a domain-specific technique, and the free-text fields of MEDLINE records [12] (see section 3.2). This suggests that the results of our domain-independent method could be further improved by the use of other fields than MeSH and domain-specific knowledge.

Most importantly, the improvements over non-extended MeSH-only document representations show that the hierarchical structure of the MeSH ontology can be used with little modification (MeSH root node add-up) to build extended MeSH-based document representations that are beneficial to the GO triage task. Finally, our extension method is based on a simple edge count distance that estimates the semantic distance of the MeSH tree locations. More refined semantic measures integrating more information from the network and the domain could yield better results.

4 Conclusion and Future Work

With the growing availability of biomedical information mainly in the form of free-text biologists need tools to process information automatically. In the early stage of information retrieval it is useful to select the probably relevant information from the mass of non-relevant information. Such a selection can be done with binary classification (also called triage) techniques.

However free-text is an ambiguous information representation. It can contain synonyms and polysems that require external knowledge for disambiguation. An alternative is to use ontologies to represent information. Ontologies provides standard terms for naming concepts and explicitly define relationships between concepts.

In this paper we proposed a method that extend ontology-based representations of biomedical documents. The method used in our experiment included the use of the Medical Subject Headings (MeSH) for biomedical document representation. The initial MeSH-only representations were then extended with MeSH concepts that were semantically close within the MeSH hierarchy. A simple edge count distance measure was used to evaluate semantic proximity.

Our method was evaluated on a document triage task that consisted in selecteing documents containing experimental evidence for the annotation of genes

in a model organism database. The triage task was organized by the Genomics track of the 2005 Text REtrieval Conference (TREC). Our document representation extension method led to an increase of 18.3% of Normalized Utility, the metric defined for the triage task. The Utility value obtained, 0.5706, positions our method amongst the top 5 runs out of 47 for the 2005 task, without the use of domain-specific techniques and by relying only on the MeSH document representation. This suggests that our results could be improved by integrating other MEDLINE fields and using domain-specific knowledge.

In future work we will evaluate our domain-independent method with other ontologies and in other contexts. The Gene Ontology is an example of taxonomy that provide ontology-based description of gene records in model organism databases. Ontology-based representation extension can also be applied to document ad hoc retrieval and document clustering. We also want to experiment with other measures in order to evaluate inter-concept semantic proximity. Some measures integrate more information from the hierarchy of the ontology, such as the depth and density of the concept nodes [9]. Others use statistical information about concepts generated from a corpus [14, 15]. In contrast the measure used in this paper only counts the edges separating two concepts to evaluate their semantic proximity. More sophisticated measures could bring further improvements to our method.

5 Acknowledgments

This research was supported by Enterprise Ireland under the Basic Research Grants Scheme project number SC-2003-0047-Y and by the European Commission under contract FP6-027026-K-SPACE.

References

1. Francisco Azuaje, Haiying Wang and Olivier Bodenreider (2005): Ontology-driven similarity approaches to supporting gene functional assessment. In proceedings of the ISMB'2005 SIG meeting on Bio-ontologies 2005, p9-10, 2005.
2. A. Budanitsky (1999): Lexical Semantic Relatedness and its Application in Natural Language Processing. Technical Report CSRG-390, Department of Computer Science, University of Toronto, August 1999.
3. Aynur Dayanik, Dmitriy Fradkin, Alex Genkin, Paul Kantor, David D. Lewis, David Madigan and Vladimir Menkov (2004): DIMACS at the TREC 2004 Genomics Track. Proceedings of the Thirteenth Text Retrieval Conference (TREC 2004), November, Gaithersburg, Maryland.
4. Eppig JT, Bult CJ, Kadin JA, Richardson JE and Blake JA (2005): The Mouse Genome Database (MGD): from genes to mice - a community resource for mouse biology. Nucleic Acids Res 2005; 33: D471-D475.
5. Mark E. Funk, Carolyn Anne Reid and Leon S. McGoogan (1983): Indexing consistency in MEDLINE. Bull Med Libr Assoc., Vol. 71(2), p176-183, 1983.
6. Gene Ontology Consortium (2004): The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research, Vol. 32, 2004.

7. Willian R. Hersh, Ravi Teja Bhuptiraju, Laura Ross, Phoebe Johnson, Aaron M. Cohen and Dale F. Kraemer (2004): TREC 2004 Genomics Track Overview. Proceedings of the Thirteenth Text Retrieval Conference (TREC 2004), November, Gaithersburg, Maryland, 2004.
8. Willian R. Hersh, Aaron M. Cohen, J Yang, Ravi Teja Bhuptiraju, P. M. Roberts and Marty A. Hearst (2005): TREC 2005 Genomics Track Overview. Proceedings of the Fourteenth Text Retrieval Conference (TREC 2005), November, Gaithersburg, Maryland, 2005.
9. Jay J. Jiang and David W. Conrath (1997): Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. Proceedings of ROCLING X, Taiwan, 1997.
10. T. Joachims (1999): Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schlkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.
11. C. Lee, W.-J. Hou and H.-H. Chen (2005): Identifying Relevant Full-Text Articles for Database Curation. Proceedings of the Fourteenth Text Retrieval Conference (TREC 2005), November, Gaithersburg, Maryland, 2005.
12. J. Niu, L. Sun, L. Lou, F. Deng, C. Lin, H. Zheng and X. Huang (2005): WIM at TREC 2005. Proceedings of the Fourteenth Text Retrieval Conference (TREC 2005), November, Gaithersburg, Maryland, 2005.
13. R. Rada, H. Mili, E. Bicknell and M. Blettner (1989): Development and application of a metric on semantic nets. In IEEE Transaction on Systems, Man, and Cybernetics, Vol 19(1), p17-30, 1989.
14. Philip Resnik (1995): Using Information Content to Evaluate Semantic Similarity in a Taxonomy. Proceedings of the 14th International Joint Conference on Artificial Intelligence, p448-453, 1995.
15. R. Richardson and A. F. Smeaton (1995): Using Wordnet in a Knowledge-Based Approach to Information Retrieval. Working paper CA-0395, School of Computer Applications, Dublin City University, Dublin, 1995.
16. Fabrizio Sebastiani (2002): Machine learning in automated text categorization. ACM Comput. Surv., Vol. 34(1), 2002.
17. Kazuhiro Seki, James C. Costello, Vasanth R. Singan and Javed Mostafa (2004): TREC 2004 Genomics Track experiments at IUB. Proceedings of the Thirteenth Text Retrieval Conference (TREC 2004), November, Gaithersburg, Maryland, 2004.
18. L.V. Subramaniam, S. Mukherjea and D. Punjani (2005): Biomedical Document Triage: Automatic Classification Exploiting Category Specific Knowledge. Proceedings of the Fourteenth Text Retrieval Conference (TREC 2005), November, Gaithersburg, Maryland, 2005.
19. M. Sussna (1993): Word sense disambiguation for free-text indexing using a massive semantic network. Proceedings of the second international conference on Information and knowledge management, p67-74, Washington, D.C., United States, 1993.