# On Combining Text and MeSH Searches to Improve the Retrieval of MEDLINE documents

**Fabrice Camous\* — Stephen Blott\* — Alan F. Smeaton\*\***

*\*School of Computing / \*\*Centre for Digital Video Processing*
*Dublin City University*
*Glasnevin, Dublin 9, Ireland*

*{fcamous,sblott, asmeaton}@computing.dcu.ie*

RÉSUMÉ. *MEDLINE est le plus grand répertoire au monde de résumés biomédicaux. Il demeure le point de départ de toute recherche d'information pour beaucoup de biologistes malgré la disponibilité croissante de l'intégralité des articles sur le Web. Chaque entrée MEDLINE est indexée manuellement avec des termes MeSH et afin d'améliorer la recherche, les champs MeSH ont déjà été utilisés avec succès dans des techniques de «pseudo relevance feedback » et d'expansion de requête. Néanmoins, ces expériences ont peu exploité la structure des champs MeSH. Cet article étudie l'impact de la structure des champs MeSH dans une méthode qui combine des recherches portant sur les champs de texte libre et les champs MeSH contenus dans les entrées d'un vaste sous-ensemble de MEDLINE. Les requêtes MeSH sont générées par la technique « Offer Weight » de Robertson. La méthode utilisée est évaluée avec l'ensemble standard de l'épreuve ad hoc de l'atelier génomique de TREC 2005. Les résultats montrent que notre approche améliore la recherche de manière significative.*

ABSTRACT. *The MEDLINE database is the world largest repository of bio-medical abstracts. It is a central information entry point for most biologists despite the growing availability of full-text articles on the WWW. Each article is manually annotated by MeSH terms to allow easy access and in order to improve retrieval, the MeSH fields of MEDLINE records were successfully used in the past with pseudo-relevance feedback and MeSH query expansion. However, previous experiments often ignored the MeSH field structure information. This paper investigates the impact of the MEDLINE MeSH field structure on a method that combines text and MeSH searches on a large subset of the MEDLINE database. Robertson's Offer Weight technique is used to generate MeSH queries. Our method is evaluated within the TREC 2005 Genomics Track on the ad hoc task collection and our results show that this approach does significantly improve retrieval performance.*

MOTS-CLÉS : *information biomédicale, ontologie, « pseudo relevance feedback », combinaison de recherches.*

KEYWORDS: *biomedical information, ontology, pseudo-relevance feedback, search combination.*

## 1. Introduction

MEDLINE, the U.S. National Library of Medicine (NLM)[1] bio-medical abstract repository, contains approximately 13 million reference articles from around 4,800 journals. 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.

The huge volume of biomedical literature is not the only difficulty facing biologists searching for information. The field suffers from a lack of naming convention which leads to natural language-specific ambiguity problems. Fortunately, much effort has been put into creating controlled vocabularies, or ontologies, in order to bring some consistency into the naming and description of concepts. Ontologies can greatly help the retrieval of genomic information by solving problems that are inherent to natural languages. MEDLINE records, on top of textual information fields such as title and abstract, include annotations from the Medical Subject Headings (MeSH)[2], the NLM controlled vocabulary thesaurus. MeSH offers standard terms for medical concepts and relationships between the terms represented as hierarchies.

Amongst previous work, Srinivasan (1996a & 1996b) and Shin et al. (2004) successfully availed of the MeSH fields of MEDLINE records in IR techniques such as query expansion and pseudo-relevance feedback or retrieval feedback to improve retrieval performance. However, their results were reported on rather small collections of a few thousand MEDLINE documents.

In other work (de Bruijn & Martin, 2003, Fujita, 2004, Abdou et al., 2005), the MeSH fields were integrated to improve the retrieval on large collections such as TrecGen 2003 (525,938 documents), TrecGen 2004 (4.5 million documents) and TrecGen 2005 (identical to TrecGen 2004 but with a different set of 50 topics). These collections were developed by the Genomics track committee of the Text Retrieval Conference (Hersh & Bhupatiraju, 2003, Hersh et al., 2004, 2005). Our method differs from the work cited by the way it isolates the MeSH fields and focuses on analysing the significance of document MeSH representations. Specifically, our experiment investigates the impact of integrating the combinations of MeSH terms that occur in the record MeSH fields. The TrecGen 2005 collection is used for evaluation.

Our approach merges a text-field search with a MeSH-field search. The text queries are processed from the TREC 2005 Genomics track ad hoc task topics. The MeSH queries are generated by pseudo-relevance feedback with Robertson's Offer Weight method (Robertson & Sparck Jones, 1997). Text-based and MeSH-based document rankings are produced with the Físreál (Ferguson et al., 2005) search engine. This search engine implements the BM25 algorithm and was developed at Dublin City University.

---

[1] http://www.nlm.nih.gov/
[2] http://www.nlm.nih.gov/mesh/meshhome.html

The paper is organized in the following way: Section 2 introduces some background on the structure of the MEDLINE record MeSH fields and pseudo-relevance feedback methods used. Section 3 describes our experimental method and its analysis. Section 4 reviews related work and section 5 concludes on future work and experiments.

## 2. Background

### 2.1. *The TREC 2005 Genomics Track Collection*

The Text REtrieval Conference (TREC)[3] guidelines and common evaluation procedures allow research groups from all over the world to evaluate their progress in developing and enhancing information retrieval systems. TREC has included a Genomics track since 2003.

The TREC 2005 Genomics track ad hoc search task collection (TrecGen05) consists of a subset of the MEDLINE bibliographic database, a set of 50 topics, and associated relevance judgments (Hersh et al., 2005). The collection contains 10 years of completed citations from 1993 to 2004 inclusive, which amounts to a total of 4,591,008 documents. All records have a title, 75.8% contain an abstract and 99% of records contain MeSH fields.

The 50 topics consist of 10 instances of template topics. An example of a template topic is: "Find articles describing the role of a <u>gene</u> involved in a given <u>disease</u>". Instances of a template replace the two generic underlined terms, gene and disease, by specific names of genes and diseases. An example of an instance of the template mentioned above is "Provide information about the role of the gene Interferon-beta in the disease Multiple Sclerosis". Text queries used in our experiment are derived from the topic by selecting the terms that are specific to the instance. In the previous example, only "Interferon-beta" and "Multiple Sclerosis" are kept to generate the text query.

### 2.2. *MEDLINE MeSH Field Structure*

The MeSH ontology contains 22568 descriptors and 83 qualifiers in the 2004 version. Descriptors are the main elements of the vocabulary. They cover medical concepts from categories such as diseases and anatomy at various levels of specificity. Qualifiers are assigned to descriptors to express a special aspect of the concept. Descriptors and qualifiers can be phrases. When it is the case we use the descriptor/qualifier phrase as the minimal information token.

A MEDLINE record includes 10-12 MeSH fields. Each field contains one descriptor and zero or more qualifiers associated with it. Figure 1 shows examples of MeSH fields in a MEDLINE record, along with title, abstract, and author fields.

---

[3] http://trec.nist.gov/

MeSH fields are listed in alphabetical order. The importance of a descriptor or a descriptor/qualifier association in terms of their relevance to the record is given by a star "*"sign placed in front of the descriptor or qualifier. This distinction was not used in this experiment.

### 2.3. *Robertson's Offer Weight Method for Term Weighting*

Given an initial document ranking, the top N documents from a search system output can be assumed to be relevant. Terms contained in the N documents can be selected to create a new query. The Robertson Offer Weight method is used to score and rank candidate terms. The score function is:

$$term\_score = r.log\left(\frac{(r+0.5)(N-n-R+r+0.5)}{(n-r+0.5)(R-r+0.5)}\right)$$

In the above formula, r is the number of relevant documents containing the term, n the number of documents containing the term, R is the total number of relevant documents for the topic, and N is the total number of documents in the collection (N= 4,591,008).

```
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
      target and nontarget cells. Although the precise identity of the
      immunoreactive steroid substance has not yet been established…
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
```

**Figure 1.** *Examples of MeSH (MH) fields in a MEDLINE (PMID: PubMed ID, TI: title, AB: abstract, AU: author)*

In the work presented in this paper pseudo-relevance feedback is used to evaluate the impact of the descriptor/qualifier associations on retrieval performance. The next section describes our experiments.

## 3. Experiment and analysis

### 3.1. *Indexing*

The text fields (title and abstract) and the MeSH fields in the document collection are indexed separately. For the text index, the text tokens are stemmed with the Porter algorithm and MEDLINE-specific stop words obtained from PubMed help[4] are removed from the index. Two different MeSH indices are created in order to evaluate the impact of the descriptor/qualifier associations. In the first MeSH index, MH1, all associations are removed so that qualifiers are considered as independent MeSH terms. In the second MeSH index, MH2, all associations are kept. For both MH1 and MH2, MEDLINE Check Tags, which are very frequent descriptors, are removed from the vocabulary. The size of the MH1 vocabulary is 21,999 (21,916 descriptors and 83 qualifiers present in the TrecGen2005 collection). The size of MH2 vocabulary is 308,333 (all descriptor/qualifier allowed combination are considered as unique terms).

### 3.2. *MeSH query generation*

A set of 50 queries derived from the 50 TrecGen2005 topics were submitted to the Físreál search engine to obtain a set of 50 document rankings. For each ranking, the top 5 documents are assumed to be relevant and we refer to these as *pseudo-relevant*. The MeSH terms from these pseudo-relevant documents are extracted and ranked with the Offer Weight method described in section 2.3. In the experiment, 8 query lengths were used for the MeSH query: top 5, 10, 15, up to the top 40 MeSH terms.

### 3.3. *Text and MeSH searches combination*

For each topic, a text ranking can be combined with a MeSH ranking obtained from a MH1 MeSH query or a MH2 MeSH query. The fist step of the ranking combination is to normalize document scores in each ranking by dividing each score by the highest score in each separate ranking. Scores from the text ranking are given a weight $\alpha$ and scores from the MeSH ranking are given a weight (1- $\alpha$). $\alpha$ varies from 0 to 1. The documents from both ranking are then combined and scores of

---

[4]

http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helppubmed.table.pubmedhelp.T42

documents present in both rankings are added up. The document scoring function is given by the following formula:

$$combi\_doc\_score = \alpha \, (text\_doc\_score / max\_text\_score)$$
$$+ (1 - \alpha) \, (mesh\_doc\_score / max\_mesh\_score)$$

Once the document lists are combined scores are added up, they are re-ranked and the top 1,000 documents for each topic are evaluated with the trec_eval program.

Performance evaluations for the 50 topics with different values of $\alpha$ were carried out. The best results were obtained with $\alpha = 0.9$. The best $\alpha$ value was determined for the set of all 50 topics. However, it is possible to determine $\alpha$ for individual topics or classes of topics when they can be identified. The 50 topics of TrecGen2005 are organized into 5 sets of 10 instances of a template topic, i.e. a query class. In future work, $\alpha$ values will be examined for each separate query class. The query class can be known in advance if the query interface allows the user to express their query in pre-determined formats (gene role in a disease, gene role in a biological process, gene interaction). An unformatted query can also be classified so that the system selects the best combination weights. Previous work in Video Retrieval (Yan et al, 2004), where several "aspects" such as text and image need to be combined, showed the benefits of using query-class dependent combination weights.

### 3.4. *Result of MeSH query expansion*

The results of the MeSH query expansion are shown in terms of Mean Average Precision (MAP) and Average Recall in figure 2 and 3, respectively. The MAP is calculated over the 50 topics. The Average Recall is simply the total number of relevant documents retrieved for the 50 topics divided by the total number of relevant documents for the 50 topics.
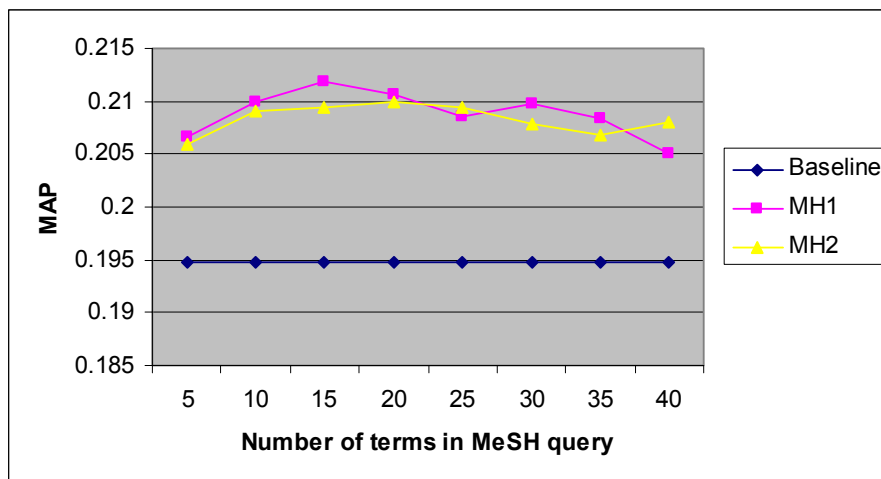
**Figure 2.** *MAP over 50 topics for baseline text search and for combinations with MeSH searches on indices MH1 and MH2*

Combinations using both MH1 and MH2 indices yield better MAP values than the original text-only search for all MeSH query lengths. All combinations improve MAP values for 29 to 34 queries out of 50. The best improvement of 8.8% is obtained by combining the text search with a 15-MH1-term search. This result corresponds to an MAP improvement for 30 queries, a null impact for 4 queries, and a negative impact on 16. In terms of Average recall, combinations using both MH1 and MH2 indices also yield better values than the original text-only search for all various MeSH query lengths. A combination with a 5-MH1-term search gives the best improvement of Average Recall at 5% (improvement for 22 queries, null impact for 23, and negative impact for 4 queries).

The results obtained show that keeping the descriptor/qualifier relationships present in the MeSH fields, as in the index MH2, does not have a significant impact within our pseudo-relevance feedback method.
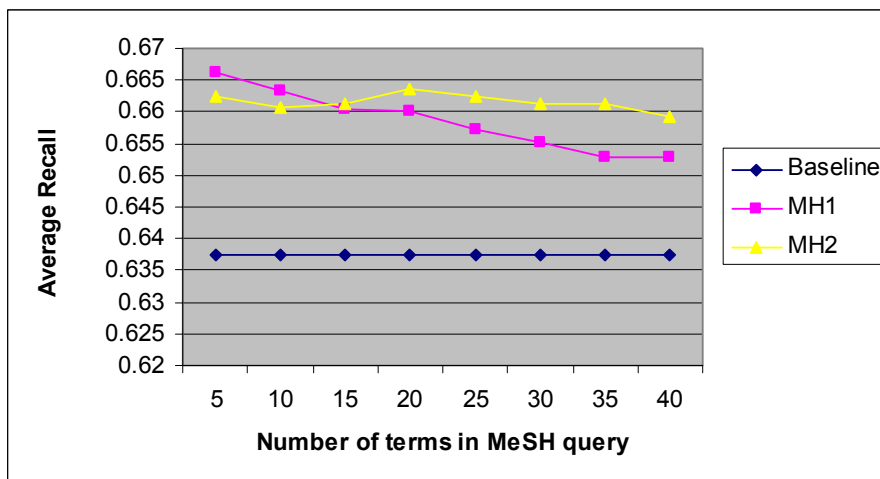
**Figure 3.** *Average Recall over 50 topics for baseline text search and for combinations with MeSH searches on indices MH1 and MH2*

## 4. Related work

Srinivasan (1996a) evaluated several strategies for MeSH query expansion in order to improve a baseline performance. Cornell's SMART retrieval system was used on a collection of 2,334 MEDLINE citations with 75 queries. The MeSH query expansion consisted of generating a MeSH query vector beside the original text query vector. The text query and the MeSH query search a text index and a MeSH index respectively. The vocabulary of the MeSH index contains the non-trivial words of the MeSH concepts. The document's relevance to the initial information need is evaluated by the following similarity function:

$$\text{Similarity}(D, Q) = \text{delta} * \text{similarity(text vector)} + \text{similarity(MeSH vector)}$$

The parameter delta allows varying the weight of the text vector in relation to the MeSH vector.

Three expansion strategies are evaluated: Expansion with an inter-field statistical thesaurus, expansion via pseudo-relevance feedback, and expansion using a combined approach. The combined approach gives the best improvement with 17% over the baseline performance of 0.5169 11-point average precision. However, Retrieval feedback alone gives almost the same improvement with 16.4%.

Our approach of combining a text and a MeSH search is inspired by Srinivasan's method but we differ by the way we tokenize MEDLINE record MeSH content. Our minimal token is a MeSH concept, which can be a phrase, such as "Multiple

Sclerosis". In the case of the index MH2, the minimal token can even be a combination of phrases, such as "Multiple Sclerosis/drug therapy". This allows us to evaluate the impact of keeping the information present in the MEDLINE record about the relationships between descriptors and qualifiers.

More recently, Shin et al. (2004) evaluated a query expansion strategy using relevance feedback on a collection including 1,239 MEDLINE records and 100 queries with their associated relevance judgements. The authors modify the method developed by Srinivasan by introducing a distinction between MeSH terms representing major or central concepts in a document and MeSH terms representing minor or peripheral concepts in a document. The authors use relevance information in a MeSH expansion method that assigns more weight to major concepts. They report a 16% improvement on Srinivasan's technique in term of R-precision.

Shin et al. also use the MeSH concept, or phrase, as a minimal token but it is not clear whether they consider the qualifiers or not. It can be assumed that the relationships between descriptors and qualifiers are ignored during their experiment.

The collections used in Srinivasan (1996a) and Shin et al. (2004) are rather small, 2,334 and 1,239 MEDLINE records respectively. In contrast, other experiments used larger subsets of MEDLINE such as TrecGen 2003 and 2004 (Hersh & Bhupatiraju, 2003, Hersh et al., 2004).

de Bruijn and Martin (2003) used the TrecGen 2003 collection in order to evaluate a modular system which is based on LitMiner, a tool box for literature access developed by the National Research Council (NRC) of Canada. TrecGen2003 consists of 525,938 MEDLINE records with 50 topics and their associated relevance judgments. When producing the index, the authors mixed the text fields (title and abstract), the MeSH fields and the RN (EC/RN Number) fields. The RN fields contain "Number assigned by the Enzyme Commission to designate a particular enzyme or by the Chemical Abstracts Service for Registry Numbers". Pseudo-relevance feedback is used on the RN fields: RN terms found in at least 20% of the previously retrieved documents are considered, between 5 and 10 terms are used per iteration. Two iterations are done. The parameters are set on the training material provided with the TrecGen 2003 collection. The pseudo-relevance feedback improves the baseline precision (0.2051 Mean Average Precision) by 8.2% and the baseline recall (0.765) by 14.4%.

Contrary to our experiment, MeSH fields are not isolated but mixed with other fields when building the document representations. Also, the impact of adding the MeSH field content is not clearly stated.

Fujita (2004) evaluated a system based on the Lemur toolkit 2.0.1 on the TrecGen 2004 collection. Two approaches were used: BM25 TF*IDF, which is an implementation of TF*IDF with Okapi BM25 as TF, and KL-divergence of probabilistic language models with Dirichlet prior smoothing. The MeSH ontology was used along with the LocusLink database (Pruitt & Maglott, 2001) in a "Reference Database Feedback" strategy that aimed at expanding query terms with synonyms. Pseudo-relevance feedback strategies are also used: Rocchio feedback

279

(Rocchio, 1971) for BM25 TF*IDF and the mixture model query update method for KL-divergence retrieval model. The top 7 documents are assumed relevant and the top 30 terms are added to the original query with coefficient 0.1. The parameters are determined on the partial training data provided with the TrecGen 2004 collection. The author does not specify whether the pseudo-relevance feedback is done on text fields only or also on other fields such as MeSH. Results show that the pseudo-relevant feedback performs better than the Reference Database Feedback but only improves average precision from 2% to 4.39%. The authors explain that this low improvement is due to the long TrecGen 2004 queries: The queries include title, need and context fields. When only the titles are used as queries, the benefits of the feedback procedures are higher (maximum of 13.3%).

MeSH is not used in isolation but along with another reference database, LocusLink. Also, the vocabulary is used for expansion independently from the document MeSH representations. Our long-term goal, however, is to exploit and improve ontology-based document representation.

Abdou et al. (2005) considered several vector-space and probabilistic models for evaluation on the TrecGen 2005 collection. A domain-specific query expansion method was used. It included spelling variation rules (lower/upper cases, Roman/Greek characters) and synonyms dictionaries. The authors also used blind query expansion with Rocchio's scheme (Rocchio, 1971). Four stemmers, Porter, Lovins, SMART and S-stemmer (minimal algorithm for plural forms), were tested. Surprisingly, results showed that stemming does not improve MAP performance and that minimal stemming (S-stemmer) gives the best performance. Also, both the domain-specific expansion and the Rocchio expansion methods decreased the performance (in average and for a majority of queries). However, mixing MeSH with text for indexing and improved the performance with statistical significance (average of 9%).

Similarly to de Bruijn and Martin (2003), Abdou et al. (2005) mixed MeSH fields with text fields and showed that this is a good strategy for the TrecGen 2004 collection. In our experiment, MeSH-based document representations are considered in isolation from the text representation as our objective is to create ontology-based links between document representations.

## 5. Conclusion and future work

This paper described a method involving the combination of text and MeSH searches on TrecGen2005, a large collection of MEDLINE records. The MeSH queries were generated by pseudo-relevance feedback from initial text searches. Robertson's Offer Weight technique was used to select the MeSH terms extracted from the documents assumed to be relevant. The results showed that the

combinations of searches consistently improved the Mean Average Precision and the Average Recall of the initial text searches.

The paper also investigated the impact of MEDLINE record MeSH field structure in the MeSH query generation process. Results showed that keeping the relationships between descriptors and qualifiers, the main types of MeSH terms, made no significant difference.

The methods used in our experiments are not new and were used with success in the past. Our contribution resides in the exploitation of structural information present in the MeSH fields and its evaluation on a large collection. The experiment presented in this paper is a starting point for future work involving MEDLINE record MeSH representations. The goal is to examine the nature of the MeSH-based links that can be created between records. Ontologies such as MeSH are usually organized semantic networks. The MeSH link generation process can benefit from using similarity measures that integrate the MeSH network.

Descriptors and qualifiers belong to separate networks and it is crucial to know how to use the relationships they represent in MeSH fields. Future experiments involving the semantic networks will help us to understand the impact of the descriptor/qualifier relationships on the quality of MeSH links.

Several approaches exist that evaluate the similarity of two terms within a semantic network (Budanitsky & Hirst, 2001) but few of them actually tackled the problem of using the network to compare groups of terms or phrases, such as MEDLINE record MeSH representations. Moreover, only some evaluations used domain-specific ontology like the MeSH vocabulary semantic network. Future work will involve a detailed examination of network-based similarity measure in the context of inter-document link generation. Improving the links in the biomedical literature is vital for research. It can potentially lead to a better identification of "themes" and to the generation of new hypotheses.

## Acknowledgements

## References

Abdou S., Savoy J., Ruck P. (2005), "Evaluation of Stemming, Query Expansion and Manual Indexing Approaches for the Genomic Task", *in Proceedings of TREC 2005,* Gaithersburg, MD.

de Bruijn B., Martin J. (2003), "Finding Gene Function using LitMiner", *in Proceedings of TREC 2003,* Gaithersburg, MD.

Budanitsky A., Hirst G. (2001), "Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures", *in the proceedings of the Workshop on WordNet and Other Lexical Resources,* Second meeting of the North American Chapter of the Association for Computational Linguistics, 2001, Pittsburgh.

Fujita S. (2004), "Revisiting again document length hypotheses - TREC 2004 Genomics Track experiments at Patolis", *in Proceedings of TREC 2004,* Gaithersburg, MD.

Ferguson P., Gurrin C., Wilkins P., Smeaton A. F. (2005), "Físréal: A Low Cost Terabyte Search Engine", *in Proceedings of the European Conference in Information Retrieval (ECIR 2005),* Santiago de Compostela, Spain.

Ganesan P., Garcia-Molina H., Widom J. (2003), "Exploiting Hierarchical Domain Structure to Compute Similarity", *ACM Transactions on Information Systems(TOIS)*, Jan 2003.

Hersh W., Bhupatiraju R. T. (2003), "TREC Genomics Track Overview", *in Proceedings of TREC 2003,* Gaithersburg, MD.

Hersh W. et al. (2004), "TREC 2004 Genomics Track Overview", *in Proceedings of TREC 2004,* Gaithersburg, MD.

Hersh W., Cohen A., Yang J., Bhupatiraju R. T., Roberts P., Hearst M. (2005), "TREC 2005 Genomics Track Overview", *in Proceedings of TREC 2005,* Gaithersburg, MD.

Pruitt KD, Maglott DR. (2001), "RefSeq and LocusLink: NCBI gene-centered resources", *Nucleic Acids Res. 2001* Jan 1;29(1):137-40.

Robertson S. E., Sparck Jones K. (1997), "Simple, proven approaches to text retrieval", Technical Report TR356, Cambridge University Computer Laboratory, 1997.

Rocchio J. J. (1971), "Relevance feedback in information retrieval", *in The SMART Retrieval System: Experiments in Automatic Document Processing,* G. Salton ed. Prentice-Hall, Englewood Cliffs, NJ, 313-323.

Shin K., Han S.-Y., Gelbukh A., Park J. (2004), "Advanced Relevance Feedback Query Expansion Strategy for Information Retrieval in MEDLINE", *Progress in Pattern Recognition, Image Analysis and Applications (CIARP 2004)*, Lecture Notes in Computer Science, N 3287, Springer-Verlag, 2004, p. 425-431.

Srinivasan P. (1996a), "Retrieval Feedback in MEDLINE", *Journal of the American Medical Informatics Association*, 1996; 3(2): 157-167.

Srinivasan P. (1996b), "Query expansion and MEDLINE", *Information Processing and Management,* 1996; 32(4): pp. 431-443.

Yan R., Yang J., Hauptmann A. G., "Learning Query-Class Dependent Weights in Automatic Video Retrieval", *in Proceedings of ACM Multimedia 2004,* New York, New York, October 10-16, 2004.