On the use of Clustering and the MeSH Controlled Vocabulary to Improve MEDLINE Abstract Search

Stephen Blott — Fabrice Camous — Cathal Gurrin — Gareth J. F. Jones — Alan F. Smeaton

School of Computing
Dublin City University
Glasnevin, Dublin 9, Ireland

{sblott, fcamous, cgurrin, gjones, asmeaton}@computing.dcu.ie

ABSTRACT: *Databases of genomic documents contain substantial amounts of structured information in addition to the texts of titles and abstracts. Unstructured information retrieval techniques fail to take advantage of the structured information available. This paper describes a technique to improve upon traditional retrieval methods by clustering the retrieval result set into two distinct clusters using additional structural information. Our hypothesis is that the relevant documents are to be found in the tightest cluster of the two, as suggested by van Rijsbergen's cluster hypothesis. We present an experimental evaluation of these ideas based on the relevance judgments of the 2004 TREC workshop Genomics track, and the CLUTO software clustering package.*

RÉSUMÉ: *Les bases de données génomiques contiennent de l' information structurée en plus de l'information textuelle que l'on trouve dans les titres et les résumés d'articles. Les techniques de recherche d'information non-structurée ne sont pas adaptées à l'exploitation de cette information structurée. Cet article décrit une technique d'amélioration des méthodes de recherche traditionnelles qui sépare un résultat initial de recherche en deux groupes à l'aide de l'information structurée disponible. L'hypothèse avancée est que les documents les plus pertinents se trouveront dans le groupe le plus densément peuplé, conformément à l'hypothèse de groupement de van Rijsbergen. Nous présentons une évaluation expérimentale de ces idées qui se base sur les documents jugés de l'atelier génomique de TREC 2004 et sur le logiciel de groupement CLUTO.*

KEYWORDS: *Genomic information retrieval, clustering, ontology, tree similarity measure.*

MOTS-CLÉS: *Recherche d'information génomique, groupement, ontology, mesure de similarité hiérarchique.*

1. Introduction

Databases of Genomic publications include textual information fields such as title and abstract as well as structured annotations. The MEDLINE bibliographic database employs human indexers to annotate each new entry with the Medical Subject Headings, or MeSH (National Library of Medicine, 2004), which is a controlled vocabulary thesaurus.

The specificity of the textual information available on MEDLINE in the title and abstract fields can limit the performance of text-based search methods. The terms of the abstract and title are only a part of the vocabulary that can be found in the full article. Also, the biomedical vocabulary, such as names given to genes and gene products, is not consistent and varies according to the area of research.

Controlled vocabulary thesauri such as MeSH can help us improve an initial text-based search. The fact that human indexers consistently use specific terms when annotating the records considerably reduces the ambiguity found in free text. Furthermore, the annotations are made with access to the full text of the article, so they cover more information than the title and abstract fields and are more reliable.

We propose to cluster MEDLINE documents resulting from an initial text-based search by using the MeSH terms they contain. More precisely, we want to cluster the result set in two distinct clusters and use the internal average cluster document similarity to measure the tightness of the clusters. Our assumption is that the relevant documents to the initial query will be found in the tightest cluster of the two, as suggested by van Rijsbergen's cluster hypothesis (van Rijsbergen, 1979).

To do this, we need to use an inter-document similarity measure that integrates the hierarchical nature of the MeSH vocabulary rather than just using textual inter-document similarity. The low average number of MeSH annotations per document and the small size of the MeSH vocabulary prevent us from representing documents as simple sets or vectors. Consequently it is necessary to use a measure that will take in account the relationships between the MeSH terms in the hierarchy.

This paper presents our experimental evaluation that is based on the relevance judgements of the 2004 TREC workshop Genomics track and the CLUTO clustering software package. It is organized as follow: Section 2 presents the background of the experiment, Section 3 describes our method and Section 4 shows the experimental results. Finally, Section 5 presents our conclusion and ideas for future work.

2. Background

2.1 *MEDLINE and the MeSH vocabulary Thesaurus*

MEDLINE is the U.S. National Library of Medicine's (NLM) premier bibliographic database that contains approximately 13 million references to journal articles in life sciences with a concentration on biomedicine.

Finding ways to improve searches of MEDLINE abstract is motivated by the fact that many biologists still use it as an entry point to the search of biological information despite the growing availability of full-text articles on the Internet.

A typical MEDLINE record contains textual fields and structured information such as MeSH annotations. The textual fields usually include a title and an abstract. Figure 1 shows an example of some of the fields that can be found in MEDLINE records.

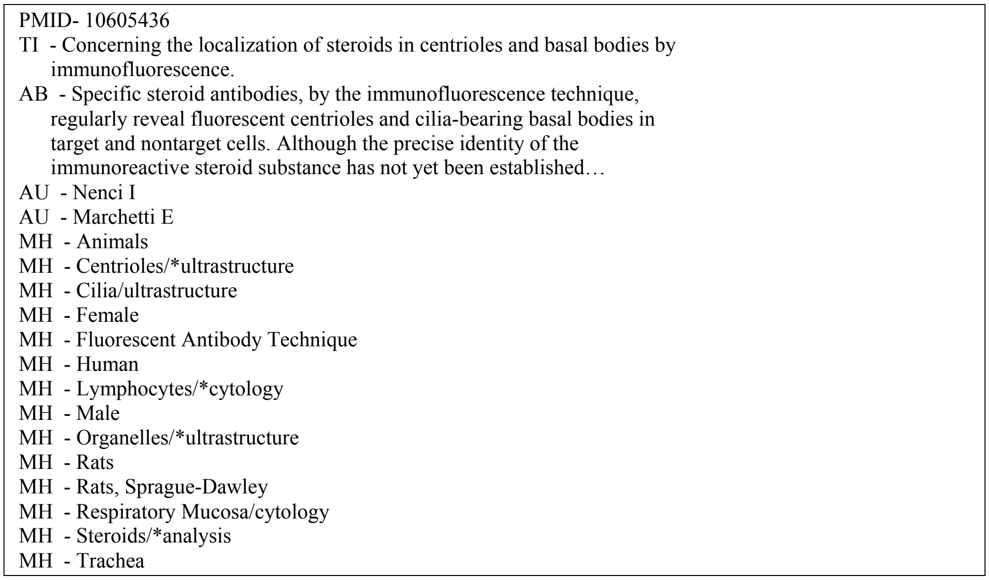
We hypothesize that the MeSH vocabulary thesaurus can be used efficiently to represent document content because of its consistency of use, as was shown in previous work (Funk and Reid, 1983), and the fact that the MeSH indexers have access to the full-text article.

There are a total of 22,568 unique terms or descriptors in the MeSH vocabulary that are organized into 15 hierarchies or trees. Each tree deals with a high-level medical class such as Anatomy, Diseases, Chemical and Drugs, or Geographic Locations, that are the roots of the trees and the highest ancestors of the hierarchy. Figure 2 shows a representation of the “Diseases” hierarchy where only a few nodes and levels are kept for the sake of clarity.

The MeSH hierarchies allow us to determine relationships between the descriptors contained in the documents. A descriptor has at least one tree number that indicates its position in the hierarchy. We use this position to compute a degree of similarity between this descriptor and others.

2.2. *The Generalized Cosine Similarity Measure (GCSM)*

Traditional similarity measures such as the Jaccard’s coefficient or the Cosine similarity (Wong et al., 1985. Ganesan, Garcia-Molina and Widom, 2001) do not integrate the hierarchical information and are not adapted to the size of the MeSH document representation and the size of the vocabulary. With an average of 12 descriptors per document in MEDLINE, a set-based or Vector Space-based similarity measure will yield very low scores unless perfect matches are found. However, two documents with a good degree of similarity may have few descriptors in common.



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

Wong et al. (1985) introduced the Generalized Vector Space Model as an extension of the Vector Space model that integrates the correlations that exist between the term vectors. The correlations are based on the co-occurrences of terms in documents.

Ganesan, Garcia-Molina and Widom (2001) explored the use of similarity measures that exploit hierarchical structures. One of them, the Generalized Cosine Similarity Measure (GCSM), is an evolution of the Cosine Similarity measure.

The GCSM model uses tree measures such as the depth of a particular node and the Lowest Common Ancestor (LCA) of two nodes. The depth of a node is the number of edges from that node to the root of the tree. The LCA of two nodes is the node of greatest depth that is an ancestor of both nodes.

Considering Figure 2 again, we can say that the depth of “Virus Diseases” is equal to 1 (1 edge from the root), and that the depth of “Pneumonia, Viral” is equal to 2 (2 edges from the root). Also, the LCA of “Pneumonia, Viral” and “Meningitis, Viral” is “Viral Diseases”, and the LCA of “Pneumonia, Viral” and “Precancerous Conditions” is “Diseases”, which is the root of this hierarchy.

In the Vector Space Model, two documents A and B are represented respectively by vectors

$$\vec{A}=\sum\nolimits_i a_i\vec{l}_i \;\; and \;\; \vec{B}=\sum\nolimits_j b_j\vec{l}_j \tag*{[1]}$$

where l_i and l_j are the descriptors contained in the documents and a_i and b_j their associated weights. The Cosine Similarity Measure (CSM) between and is given by the formula

$$sim\left(\vec{A},\vec{B}\right)=\frac{\vec{A}.\vec{B}}{\sqrt{\vec{A}.\vec{A}}\sqrt{\vec{B}.\vec{B}}} \tag*{[2]}$$

and the dot product between two vectors is determined by the formula

$$\vec{A}.\vec{B}=\sum_{i=1}^{n}\sum_{j=1}^{n}a_i b_j \vec{l}_i\vec{l}_j \tag*{[3]}$$

The GCSM measure uses the same formulae but differs in the calculation of the descriptors dot product $l_i.l_j$. In the CSM, $l_i.l_j$ is equal to one only if i=j. Otherwise the two vectors are orthogonal and their dot product is equal to zero. In the GCSM, $l_i.l_j$ is calculated with the formula

$$\vec{l}_i\vec{l}_j=\frac{2*depth\left(LCA\left(l_i,l_j\right)\right)}{depth(l_i)+depth(l_j)} \tag*{[4]}$$

Two descriptor vectors that are not identical are no longer considered perpendicular. If we look at Figure 2 again, we can calculate that the dot product of “Neoplastic Processes” and “Precancerous conditions” is equal to ½., and not 0. However, the dot product of “Pneumonia, Viral” and “Precancerous Conditions” is equal to zero, as their LCA is the root, which has depth zero. If the two nodes are identical, their dot product is still equal to 1.

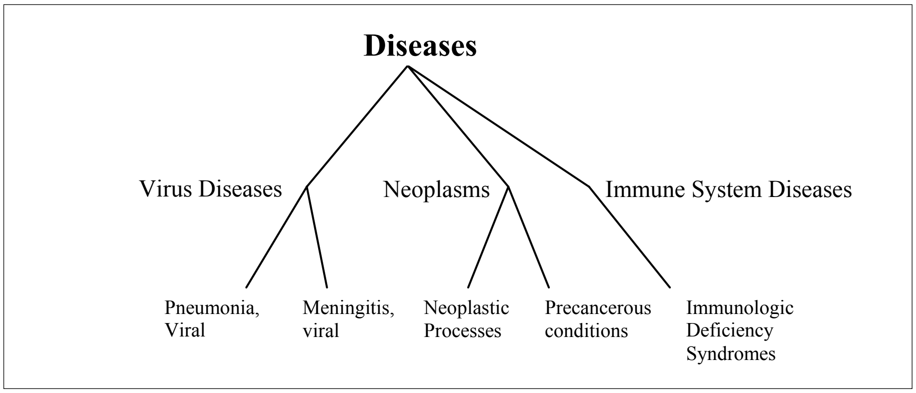


Figure 2. *A simplified representation of the "Diseases" hierarchy*

2.3. *The 2004 TREC Genomics Track Dataset*

The 13th Text REtrieval Conference (TREC) included for the second time a Genomics track in 2004. 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.

The TREC Genomics track 2004 (TrecGen04) ad hoc task consisted of a subset of the MEDLINE bibliographic database, a set of 50 topics and associated relevance judgments. The subset used for the track contained 10 years of completed citations from 1993 to 2004 inclusive, which amounted to a total of 4,591,008 documents.

We chose to consider the relevant judgments as the result of an initial text-based search on MEDLINE that we intended to improve using clustering. More precisely we wanted to show that the generation of two clusters would help us locating the relevant documents. The cluster hypothesis asserts that relevant documents are closely grouped together, i.e. they are more similar to each other than they are to the non-relevant documents.

2.4. *The CLUTO Software Package*

CLUTO is a software package for clustering high-dimensional dataset developed by George Karypis at the University of Minnesota, Minneapolis, MN, USA. We downloaded version 2.1.1 from the project homepage at http://www-users.cs.umn.edu/~karypis/cluto/index.html.

The CLUTO software approaches the clustering problem as the optimization of a criterion function. It implements several criterion functions. Some focus on maximizing the intra-cluster similarities, some focus on minimizing the inter-cluster similarities, and some are hybrids of the two previous types of functions. Others see the collection as a graph and try to minimize the number of edges that members of a cluster share with the rest of the collection. A detailed description of the criterion functions and their associated optimization methods is available in Zhao and Karypis (2003).

In our experiment, we used the direct optimization of a hybrid criterion function called H_2 and defined by the following formula

$$H_2 = \frac{\sum_{r=1}^{k} \left\| \vec{D}_r \right\|}{\sum_{r=1}^{k} n_r \vec{D}_r \vec{D} \Big/ \left\| \vec{D}_r \right\|} \qquad [5]$$

Where D_r is the composite vector of cluster r, n_r is the size of cluster r, k is the total number of clusters and D is the composite vector for all the documents in the collection. This choice of clustering method is in no way limitative and we intend to experiment in the future with other approaches to the clustering problem that work well in our domain.

3. Method

We selected a total of 10,335 judged documents covering 10 topics randomly selected from the TrecGen2004 ad hoc task 50 topics. The reason for this sampling was the time cost of computing similarity matrices for the 50 topics. Table 1 shows the distribution of the documents over the 10 topics selected.

For each topic, the total judged documents were interpreted as a result set from an initial text-based search that we intended to refine using clustering and MeSH descriptors. By generating two clusters, we want to show that the documents judged relevant will be found in the tightest cluster, in agreement with the cluster hypothesis. Figure 3 illustrates this approach.

MeSH descriptors can represent “central concepts” or “peripheral” ones in the document they are assigned to. The “central concept” descriptor corresponds to one of the principal concepts dealt with in the document whereas any other descriptors would characterize secondary concepts contained in the document. MEDLINE indexers use a star to distinguish the “central” descriptors from the others when annotating a document. In figure 1 the use of a star in “Centrioles/*ultrastructure” indicates that “Centrioles” is a central concept. The term coming after the forward slash, “ultrastructure” is called a qualifier and is used to specify a particular domain for the concept. We ignored the qualifiers in the experiment described here.

Topic	Total Judgments	Definitely Relevant	Possibly Relevant	Not Relevant	Definitely and Probably Relevant
1	879	38	41	800	79
10	1126	3	1	1122	4
11	742	87	24	631	111
12	810	166	90	554	256
13	1118	5	19	1094	24
14	948	13	8	927	21
15	1111	50	40	1021	90
16	1078	94	53	931	147
17	1150	2	1	1147	3
18	1392	0	1	1391	1

Table 1. *Distribution of the judged documents for 10 randomly selected topics (We simply picked the first 10 topics sorted by topic number, using a dictionary-order sort)*

We experimented with 3 ways to use the MeSH descriptors for document representation:

— Using the “central” descriptors only.

— Using all descriptors without distinction.

— Using all descriptors but giving more weight to “central” descriptors.

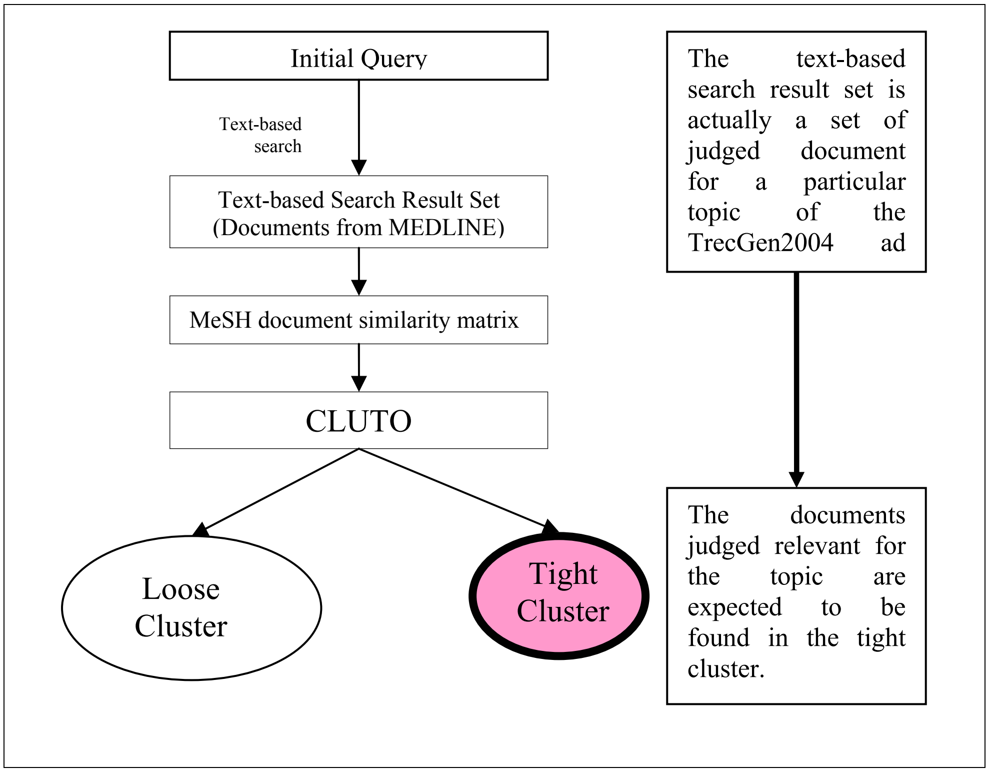

Initial Query
Text-based search
Text-based Search Result Set (Documents from MEDLINE)
MeSH document similarity matrix
CLUTO
Loose Cluster
Tight Cluster
The text-based search result set is actually a set of judged document for a particular topic of the TrecGen2004 ad
The documents judged relevant for the topic are expected to be found in the tight cluster.

Figure 3. *Overview of our method*

After selecting the MeSH descriptors for document representation, we used their tree numbers to build hierarchical description of the document. The tree number gives us the position of the term in the hierarchy and we trace the path back to the root of the hierarchy to create the sub-tree representation. Several solutions arise because a single descriptor can occur several times in the hierarchies. To solve this problem, we chose to use the method suggested by Ontrup (2003) that consists in keeping the "denser" tree, i.e. the one with the fewer edges between the leaves. Figure 4 shows the hierarchical representation of the MEDLINE record from Figure 1 when all the descriptors are considered.

For each topic, we computed a similarity matrix containing all the possible pair-wise tree similarity measures between documents. We chose a TF-IDF weighting scheme to calculate the similarity measure using equation [3] so that a_i=TF_i*IDF_i and b_j=TF_j*IDF_j

with $IDF_i=log_2(N/(n_i+1))$ (N is the collection size and n_i is the collection frequency of descriptor i). We tried several ways of calculating the TF-IDF weights. When only the "central" concepts were used or when all descriptors were used but without distinction, TF=1, which is always the descriptors document frequency value. However, when we used all descriptors with discrimination we experimented with values TF=2 and TF=3 for the "central" descriptors, leaving TF=1 for the others. We tested two ways for the calculation of the IDF value: First, N was chosen as the total amount of judged documents for each topic and n_i the collection frequency of descriptor i in the topic collection. Secondly, we considered the entire TrecGen2004 collection, i.e. N=4,591,008 and n_i the collection frequency of descriptor i in that collection regardless of the topic. The 6 different combinations of TF-IDF and MeSH descriptors selection provided us with a total of 6 similarity matrices for each topic. Table 2 sums up the characteristics of the 6 combinations.

Each similarity matrix was fed to the CLUTO software package choosing a direct 2-way cluster optimization of the hybrid H_2 criterion function and the content of the tightest cluster of the two was checked for relevant documents. The detailed evaluation of the experiment is available in the next section.

	Central Concepts MeSH only	All MeSH used	IDF: N, n_i calculated at topic level	IDF: N, n_i calculated at the TrecGen2004 collection level	TF=1 for all MeSH	TF=2 for Central Concepts, TF=1 for other MeSH	TF=3 for Central Concepts, TF=1 for other MeSH
Combination 1	X		X		X		
Combination 2	X			X	X		
Combination 3		X	X		X		
Combination 4		X		X	X		
Combination 5		X		X		X	
Combination 6		X		X			X

Table 2. *Summary of the characteristics of the 6 combinations*

4. Experimental Results

We can evaluate the result of our experiment for a topic i with the recall of the documents judged relevant in the tightest cluster. This recall for a given topic i is given by $R_{topic\ i}=N_{t,i}/R_i$ where $N_{t,i}$ is the number of judged relevant documents found in the tightest cluster and R_i is the total number of judged relevant documents for this topic i.

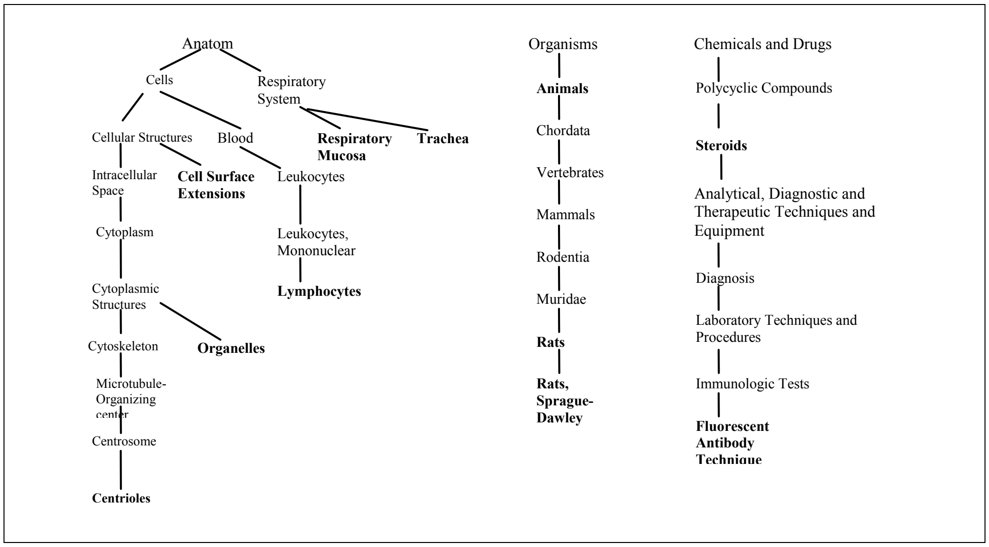


Figure 4. *Hierarchical representation of the MEDLINE record from Figure 1 when all descriptors are considered equally (the descriptors actually in the record are in bold)*

However we also need to have an idea of the proportion of relevant documents we manage to get in the tightest cluster which is the precision value $P_{topic\ i}= N_{t,i}/N_i$ where $N_{t,i}$ is the number of judged relevant documents found in the tightest cluster and N_i is the size of the tightest cluster for this topic i.

We can calculate an average recall $R_{average}$ and an average precision $P_{average}$ over the 10 topics. The values of $R_{average}$ and $P_{average}$ for the 6 combinations described in Section 3 are shown in Table 3. The initial average precision $IP_{average}$ is the average precision before any clustering is done (the average proportion of relevant documents in the initial result sets).

The best value amongst the 6 combinations for $R_{average}$ and $P_{average}$ respectively was 0.72 and 0.100388. It was obtained with combination 4 which is the only combination that is improving the initial average precision.

We can observe from the results above that both values of $R_{average}$ and $P_{average}$ for combination 2 are higher than for combination 1. Also, combination 4 improves the result of combination 3. Combination 1 and 3 compute the IDF of each descriptor using the topic-related judged documents only while combination 2 and 4 use the entire TrecGen2004 collection to calculate the descriptor IDF.

Combinations	$P_{average}$	$R_{average}$
Combination 1	0.077419	0.58
Combination 2	0.082484	0.67
Combination 3	0.077176	0.64
Combination 4	**0.100388**	**0.72**
Combination 5	0.078725	0.62
Combination 6	0.078218	0.61
Initial precision $IP_{average}$	0.082329	

Table 3. *$R_{average}$ and $P_{average}$ values for the 6 combinations described in Section 3*

The results for precision and recall do not allow us to conclude on the advantage of including all descriptors as opposed to keeping the central concepts only. When all descriptors are included, the use of weights (combination 5 and 6) seems to damage the performance of combination 5, which does not use higher weights for central concepts

Since our experiment covered 10 topics only, it is worth considering the recall and precision performance for each topic. Figure 5 and 6 illustrates respectively the recall and precision results for each combination and each topic.

Figure 6 shows that topics 10, 17 and 18 have a very low precision before and after the clustering, but they have a very high recall, as can be seen in figure 5, due to the small amount of relevant documents they contain initially. We can see in figures 5 and 6 how the good performance of combination 4 in both recall and precision is strongly correlated to its performance with topic 11 whereas all the other combinations fail to improve the initial precision and have very poor recall.

It is clear that the small amount of topics used in the experiment does not allow us to draw interesting and definite conclusions about the different combinations used and the experimental method in general.

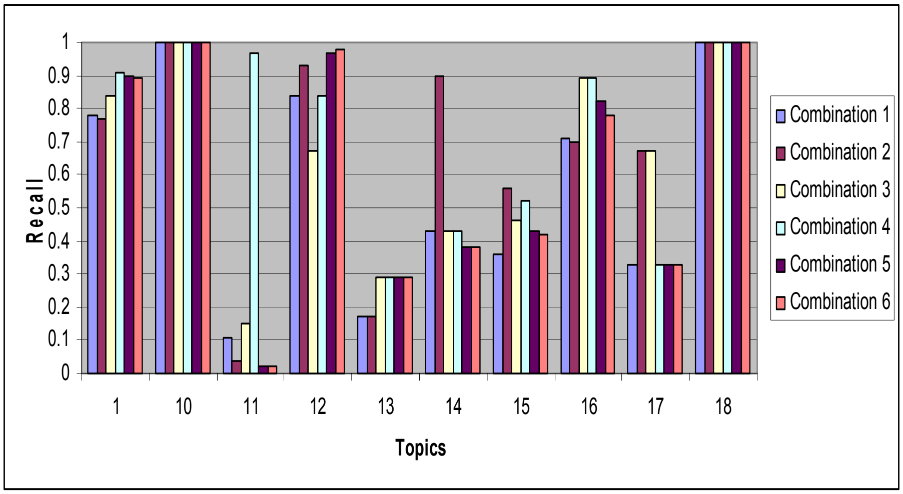


Figure 5. *Recall of the judged relevant documents in the tightest cluster by topic and combination*

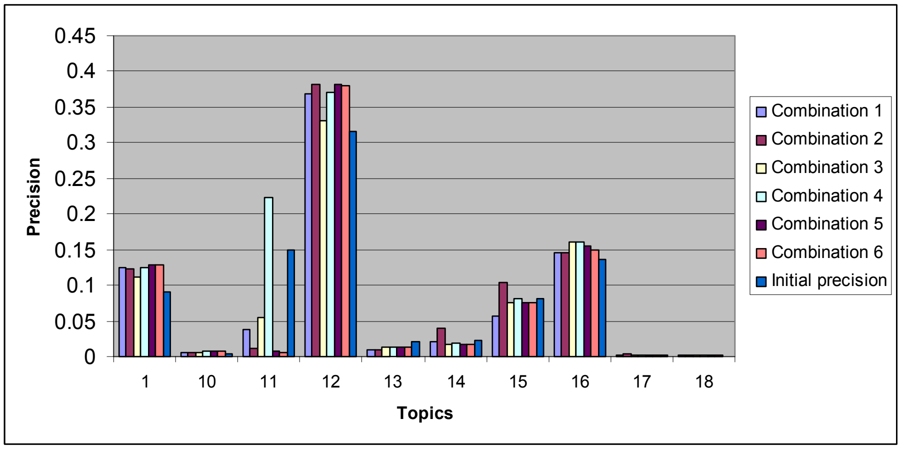


Figure 6. *Precision in the tightest cluster by topic and combination*

5. Conclusion and Future Work

The goal of our experiment was to simulate the improvement of an initial text-based search on MEDLINE using the MeSH vocabulary thesaurus and clustering. Applying the cluster hypothesis, we wanted to show that we could identify the documents relevant to a topic by looking for the documents tightly clustered together, i.e. the ones that are more similar to each other amongst a group of documents.

Since van Rijsbergen formulated the cluster hypothesis in 1979, much literature has referenced to it and used it as a basis for clustering experiments. Nonetheless many of the experiments were built with text-based inter-document similarity measures. Comparing text terms from documents in one of the great challenges of Information Retrieval partly because of the polysemious and synomymic characteristics of natural languages. Our approach innovates by using annotations from a controlled thesaurus, the MeSH vocabulary, and by using its hierarchical structure to give the documents a hierarchical representation.

The Generalized Cosine Similarity Measure that we used to compare the documents is derived from the traditional Cosine Similarity Measure and integrates the hierarchical representation of documents.

We extracted relevance judgments from 10 topics of the TrecGen2004 ad hoc task collection and used the CLUTO software package to obtain a 2-way clustering for each topic.

The best value for the average recall and precision of the documents judged relevant in the tightest cluster over the 10 topics was respectively 0.72 and 0.100388. It was obtained by using no distinction between MeSH descriptors and by using the entire TrecGen2004 ad hoc task collection to calculate the descriptor IDF values.

However, the small amount of topics used in the experiment prevented us from concluding on the real impact of any combination we used and of our general experimental method.

Therefore, extending our experiment to the entire judged documents collection, i.e. the documents judged for the 50 topics of the TrecGen 2004 ad hoc task, is the first thing we will focus on in future experimental explorations.

Secondly, we will refine the MeSH document representation. We mentioned earlier the occurrence of qualifiers in MEDLINE MeSH fields. They add more specificity to the choice of the human indexer and although they do not belong to the hierarchies, their integration in the document similarity measure might improve the performance.

We will also test more hierarchical similarity measures. The GCSM measure is a “first generation” evolution of the Cosine measure. We plan to experiment with “second generation” measures that can deal with the presence of many siblings in a document.

This is the case where many descriptors with the same parent are used to represent a document because it relates to topics close to each other and at the same level of specificity in the hierarchy. The GCSM measure will give a disproportionate high score when comparing this document with other similar documents with fewer descriptors. Additionally, we have yet to look at different clustering approaches in this experiment. We will investigate the various clustering solutions that are already available and evaluate these for the retrieval task at hand.

Finally, our work relates to any situations where alternatives to text can be used to compute inter-document similarity, such as annotations and indexing terms from ontologies. We plan to experiment with other controlled vocabularies, e.g. the Gene Ontology, that are available in the Genomic domain.

Acknowledgements

This work is funded by Enterprise Ireland under the Basic Research Grants Scheme, project number SC-2003-0047-Y.

References

Funk and Reid, Indexing consistency in MEDLINE, *Bull Med Libr Assoc*, 71(2): 176–183, 1983.

Ganesan, Garcia-Molina and Widom, Exploiting Hierarchical Domain Structure to Compute Similarity, Extended Technical Report, Stanford University, CA, 2001.

National Library of Medicine, Medical Subject Headings, MeSH. URL: http://www.nlm.nih.gov/mesh/meshhome.html, 2004.

Ontrup et al., A MeSH Term based Distance Measure for Document Retrieval and Labeling Assistance, Cancun, Mexico, *Proc. of EMBC2003 (25th Annual Int. Conf. of the IEEE Engineering in Med. and Biol. Soc.)*, 2003.

Van Rijsbergen, C. J., *Information Retrieval*, London: Butterworths, 1979.

Wong et al., Generalized vector spaces model in information retrieval, *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval*, Montreal, Quebec, Canada, pages: 18 – 25, 1985.

Zhao and Karypis, Criterion Functions for Document Clustering, Experiment and Analysis, University of Minnesota, MN, 2003.