Promoting Diversity in Top Hits for Biomedical Passage Retrieval

Bill Andreopoulos, Xiangji Huang, Aijun An, Dirk Labudde, and Qinmin Hu

Abstract. With the volume of biomedical literature exploding, such as BMC or PubMed, it is of paramount importance to have scalable passage retrieval systems that allow researchers to quickly find desired information. While topical relevance is the most important factor in biomedical text retrieval, an effective retrieval system needs to also cover diverse aspects of the topic. Aspect-level performance means that top-ranked passages for a topic should cover diverse aspects. Aspect-level retrieval methods often involve clustering the retrieved passages on the basis of textual similarity. We propose the HIERDENC text retrieval system that ranks the retrieved passages, achieving scalability and improved aspect-level performance over other clustering methods. HIERDENC runtimes scale on large datasets, such as PubMed and BMC. The HIERDENC aspect-level performance is consistently better than cosine similarity and Hamming Distance-based clustering methods. HIERDENC is comparable to biclustering separation of relevant passages, and improves on topics where many aspects are involved. Converting textual passages to GO/MeSH ontological terms improves the HIERDENC aspect-level performance.

1 Introduction

The body of biomedical literature is growing rapidly. PubMed, the main biomedical literature database, holds over 17 million abstracts and over 2000 new abstracts are added a day (1). Information retrieval (IR) technology

Bill Andreopoulos, Xiangji Huang, Aijun An, and Qinmin Hu
Dept. of Computer Science and Engineering, York University, Toronto, Canada
e-mail: {billa,jhuang}@cse.yorku.ca

Bill Andreopoulos
Biotechnological Centre, Technische Universität Dresden, Germany

Dirk Labudde
Bioinformatics group, University of Applied Sciences, Mittweida, Germany

Z.W. Ras and A. Dardzinska (Eds.): Advances in Data Management, SCI 223, pp. 371–393.
springerlink.com

plays a vital role in biomedical data management, especially for users who desire passages that are most relevant to a topic or question. A biomedical information retrieval system is a computer system for browsing, searching and retrieving passages from a large collection of biomedical literature. Methods of information retrieval may utilize some method of adding metadata to the text, such as ontological term annotations extracted via text mining, keywords or descriptions; then retrieval is performed over the textual annotations (2). Information retrieval is required to scale up efficiently to the quickly growing body of biomedical literature.

1.1 Aspect-Level Performance: Promoting Diversity in the Top Hits

For addressing users' questions on a topic in competitions, such as in the TREC 2007 Genomics Track (3; 4) or the ImageCLEF 2008 Photo Retrieval Task (5), one of the main tasks is to extract ranked textual snippets from documents (6; 7). The performance is considered better if the top-ranked textual snippets are not only relevant to the topic, but also cover diverse *aspects*. A biomedical researcher would like to avoid seeing similar or duplicated passages in the top hits, and redundant information is removed by covering diverse aspects. A search engine that retrieves a diverse, yet relevant set of textual passages at the top of a ranked list is more likely to satisfy its users. Another reason why it's a good idea to promote diversity is because often different people type in the same query but wish to see different results. Aspect-level retrieval performance was previously studied in the context of competitions such as TREC and ImageCLEF. Text-based clustering was used to group passages, consequently promoting diverse topics in the top retrieved hits.

The main difference of our work from previous work is to take a more practical approach to the problem of aspect-level retrieval, making it scalable to large and quickly expanding biomedical literature. Our system promotes diversity in the top hits through scalable text-based clustering. The main contributions of our work include: *a*. We propose scalable aspect-level retrieval that works with millions of documents as well as thousands of documents, and *b*. We convert passages to vectors of ontological terms, improving aspect-level retrieval performance. Further benefits of our methodology as far as the clustering method is concerned include: no re-clustering needed when new text is presented, no user-specified input parameters required, and insensitivity to ordering of passages.

This chapter is organised as follows. Section 2 discusses related work, including document clustering algorithms for separating the relevant passages, and ontology-based integration of biomedical information. Section 3 presents the HIERDENC algorithm that ranks the passages through text-based clustering. Section 4 presents the evaluation methods, experimental results and

the corresponding discussions, demonstrating scalable HIERDENC runtimes on large real world biomedical datasets. This section also demonstrates reasonable aspect-level performance with ranking and after converting passages to GO/MeSH ontological term vectors. Finally, section 5.1 gives our conclusions and Section 5.2 describes future research directions.

2 Related Work

2.1 Clustering in Information Retrieval

Clustering is a common technique for statistical data analysis, which has been used in biomedical question answering and aspect-level retrieval.

Goldberg et al. used a naive clustering for reranking passages; the results were discouraging, resulting in worse aspect-level performance than the original ranking, as well as lower document- and passage-level performance scores (6). In particular, they used bag-of-words vector representations and cosine-similarity based clustering. This differs from our work where we convert passages to ontological term vectors. While they interleaved results from clusters to achieve aspect diversity, our method ranks the clusters by their coverage in the entire dataset and keeps the most representative passage from each cluster. They also used random walks on a graph over passages to promote diversity, but our method considers which clusters are the most prominent in the dataset and likely to represent different aspects of the topic.

Si et al. derive the MeSH representations for the top-ranked passages for a user query, reflecting the topical aspects of passages. Then, they rerank the passage retrieval result to construct a new ranked list. A document is selected and added to the bottom of the current reranked list, by considering the novelty information of the topical aspects with respect to the current reranked list (8). While they extract representative MeSH terms for each passage, we also extract Gene Ontology terms. Another difference from our work is that they adopt a gradient-based search approach, while we consider globally significant clusters in the entire dataset. Therefore, while their method is sensitive to ordering of passage input, our method is not.

In this work, we will compare our HIERDENC clustering method to three other clustering methods, presented next. We will evaluate these methods' aspect-level retrieval performance: biclustering, cosine similarity and hamming distance-based clustering. We will cluster both the original text passages and the extracted ontological term vectors.

2.2 Biclustering of Passages

Biclustering allows simultaneous clustering of the rows and columns of a matrix, where the columns are textual passages and the rows correspond to

words (9; 10). In our biclustering approach, we produce only two clusters, since we want to separate the passages that are relevant to the topic from the rest. Another reason for producing only two clusters is that biclustering performance is known to deteriorate for more clusters. We select the smallest cluster as more likely to contain relevant passages, since usually the majority of passages retrieved are irrelevant. Biclustering differs from our proposed HIERDENC method that produces many clusters and then ranks them, keeping a representative passage from each cluster. Given an $m \times n$ word-by-document matrix, the biclustering algorithm generates biclusters - a subset of rows which exhibit similar behavior across a subset of columns, or vice versa. We find subgroups in a binary matrix where entries are one or zero.

Let A denote the $m \times n$ word-by-document matrix, and D_1 and D_2 denote diagonal matrices such that $D_1(i,i) = \Sigma_j A_{ij}$, $D_2(j,j) = \Sigma_i A_{ij}$. Then, the following equations define the singular value decomposition (SVD) of the normalized matrix $A_n = D_1^{-1/2} A D_2^{-1/2}$:

$$D_1^{-1/2} A D_2^{-1/2} v = (1-\lambda)u, \; and \; D_2^{-1/2} A^T D_1^{-1/2} u = (1-\lambda)v.$$

In particular, u and v are the left and right singular vectors respectively, while $(1-\lambda)$ is the corresponding singular value σ. We compute the left and right singular vectors corresponding to the second (largest) singular value of A_n, $A_n v_2 = \sigma_2 u_2$, $A_n^T u_2 = \sigma_2 v_2$, where $\sigma_2 = 1 - \lambda_2$. The right singular vector v_2 will give a bipartitioning of documents while the left singular vector u_2 will give a bipartitioning of the words. Given the singular vectors u_2 and v_2 the key task is to extract the optimal partition from these vectors. Biclustering looks for a bi-modal distribution in the values of u_2 and v_2. Let m_1 and m_2 denote the bi-modal values that we are looking for. The second eigenvector of the Laplacian matrix is given by $z_2 = \left(D_1^{-1/2} u_2 \;\; D_2^{-1/2} v_2\right)$. One way to approximate the optimal bipartitioning is by the assignment of $z_2(i)$ to the bi-modal values $m_j (j = 1, 2)$ via the classical k-means algorithm:

1. Given A, form $A_n = D_1^{-1/2} A D_2^{-1/2}$.
2. Compute the second singular vectors of A_n, u_2 and v_2; form the vector z_2.
3. Run the k-means algorithm on the 1-dimensional data z_2 to obtain the desired bipartitioning.

This algorithm runs k-means simultaneously on the reduced representations of both words and documents to get the co-clustering. Thus, the biclustering algorithm co-clusters words and documents.

2.3 Cosine Similarity Textual Clustering

Cosine similarity is a measure of similarity between two vectors of words by finding the angle between them, often used to compare documents (or passages)

in text mining (11). Cosine similarity is typically used for bags-of-words representations of textual passages, and is significantly slower than our proposed method depending on all-by-all comparisons. Given two vectors of words, A and B, the cosine similarity, θ, is represented using a dot product and magnitude as $\theta = arccos\frac{AB}{|A||B|}$. In the cosine similarity based approach that we used in our experiments, each passage is matched to its nearest passage according to θ; in graph terms this is conceptualized as a directed edge from the former passage to the latter. Then, every connected component is considered as a cluster. Passages that are not connected via a path are separate clusters.

2.4 Hamming Distance Textual Clustering

The Hamming Distance is used as a measure of dissimilarity between two vectors of words, by counting the number of words that are contained in one vector but not the other (12; 13). The HD-based clustering depends on the ordering of passage input, and exhibits quadratic complexity unlike our proposed method. Given two vectors of words, A and B, the Hamming Distance, HD, is computed as $HD = |(A - B) \cup (B - A)|$. The clustering iterates over all passages from the smallest to the largest; a passage π is matched to cluster c_π with which it has the most words in common, considering the union of all words appearing in the cluster. The passage π is clustered in c_π if the HD between them does not exceed a threshold ϕ. Threshold ϕ represents the maximum HD, determining if π is clustered or not; the ϕ value starts from 1 and is progressively relaxed, thus producing layers in clusters. Layered clusters have an "onion"-layered structure, such that the least dissimilar passages are placed in the initial-created layers and affect subsequent clustering decisions. The iteration through passages continues until all passages have been clustered.

2.5 Query Term Expansion

Query expansion is a popular and commonly used strategy to improve the passage-retrieval performance. Our ontological term extraction on retrieved passages resembles query term expansion, in the sense that ontological terms and potentially their ancestors are also associated with passages. In the past, expansion was done on queries, but nobody tried expansion on retrieved passages. Through extracting ontological terms from passages, our proposed method has potential to outperform methods that expand queries for improved retrieval performance. Moreover, previous work which applied query expansion based on hand-crafted thesaurus is often limited in improving the performance (14; 15). For example, Voorhees (15) expanded

queries with synonyms manually selected from WordNet and achieved only limited improvements (around −2% to +2%) on some queries. Recently a lot of work on biomedical information retrieval appeared in the TREC Genomics Track (16; 17; 3). Huang and others (18) achieved notable performance improvements by manually processing the gene name variants from gene databases. Zhou et al (19) proposed their effective conceptual retrieval model by incorporating five types of domain knowledge including synonyms.

2.6 Ontology-Based Data Integration in Bioinformatics

Individually developed ontologies often support the annotation of online databases for information retrieval purposes. Significant work has been done in the past two years to make the ontologies interoperable and support integration of information from different sources. These efforts aim to facilitate ontology interoperability and automated reasoning. We leverage our ontological term extraction for ontology linking, which differs from other ontology-based data integration methods through its automation and simplicity of use. We rely solely on the notion of term extraction from documents and co-occurring terms in the same passage (or image caption). Our ontology linking method is likely to appeal to biomedical practitioners and researchers better than RDF and semantic web-based methods that exhibit low usability and appeal.

Burek et al. (2006) (20) present a top-level ontological framework for representing knowledge about biological functions. This framework provides a means to capture existing functional knowledge in a principled way.

Garcia-Sanchez et al. (2008) (21) propose an ontology-based framework for seamlessly integrating intelligent agents and semantic web services. Agent technology can assist users in discovering services available on the Internet. This allows integrated access to biomedical information.

Smith et al. (2007) (22; 23) leverage the structure of the semantic web to enhance information retrieval for proteomics. They use an RDF graph that inter-relates documents through their associated biological identifiers (e.g., protein ID). In related work, they built a software system called LinkHub using semantic web RDF that manages the graph of identifier relationships. LinkHub facilitates cross-database queries and information retrieval in proteomics.

Ruttenberg et al. (2007) (24) discuss advancing translational research with the semantic web. They present a scenario that shows the value of the semantic web technologies for aiding biomedicine researchers. They conclude that semantic web technologies present promise and current tools and standards are already adequate for translational research.

3 Methods

The previous section discussed how clustering is used in the information retrieval process. In this section we will examine our HIERDENC system, which differs from previous work as follows: *a*. For aspect-level retrieval performance, it provides a scalable clustering method for ranking textual passages, and *b*. We use Go/MeSH ontological term vectors as caption-based term expansion. In this section we first present our textual retrieval system. Then, we present our test datasets, and the TREC evaluation measures used to compare performances of all methods.

3.1 Workflow of HIERDENC Text Retrieval System

Figure 1 shows the workflow of HIERDENC text retrieval. The objects to be clustered are the textual passages and snippets, which may be captions of images. HIERDENC applies text-based clustering in combination with ontological term extraction on text. Each passage is represented as a "word vector", whether it is the original passage or the one converted to ontological terms.

Automatic annotation of biomedical passages can be an important step when searching for information from a database. We used the GoPubMed term extraction algorithm for converting each passage to a vector of ontological terms. This vector describes each passage on an ontological basis.

Users search via keywords and the retrieved passages are clustered into groups of topics. Retrieved passages are clustered based on the original text, or extracted ontological term vectors. The HIERDENC retrieval system clusters the passages to achieve good *aspect-level* performance. The clustering imposes a ranking of retrieved passages, such that top ranked passages reflect different topics for the query. Top ranked passages are similar to many other passages in the database, but any two top ranked passages are likely to be different.

Text-based clustering can also be applied to biomedical image databanks, where images have textual captions or comments associated with them. Figure 2 shows a snapshot of HIERDENC retrieval as applied to image captions; caption-based clusters are represented on the right-hand side bar and clicking on a cluster takes the user to the corresponding cluster of images.

A final capability of our system is to link different ontologies (or vocabularies) if two extracted ontological terms co-occur in the same passage or caption. Linked ontologies support reasoning over the vast biomedical knowledge.

3.2 Ontological Term Extraction from Passages

After standard stop word removal in the data preparation step, we converted each passage to a vector of ontological terms extracted via the GOPubMed

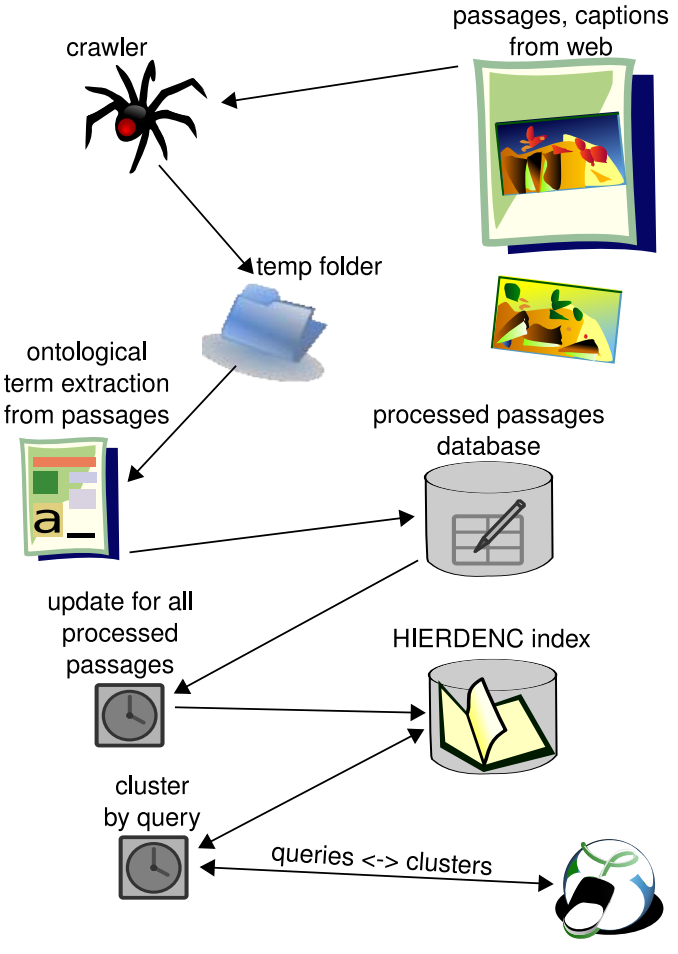
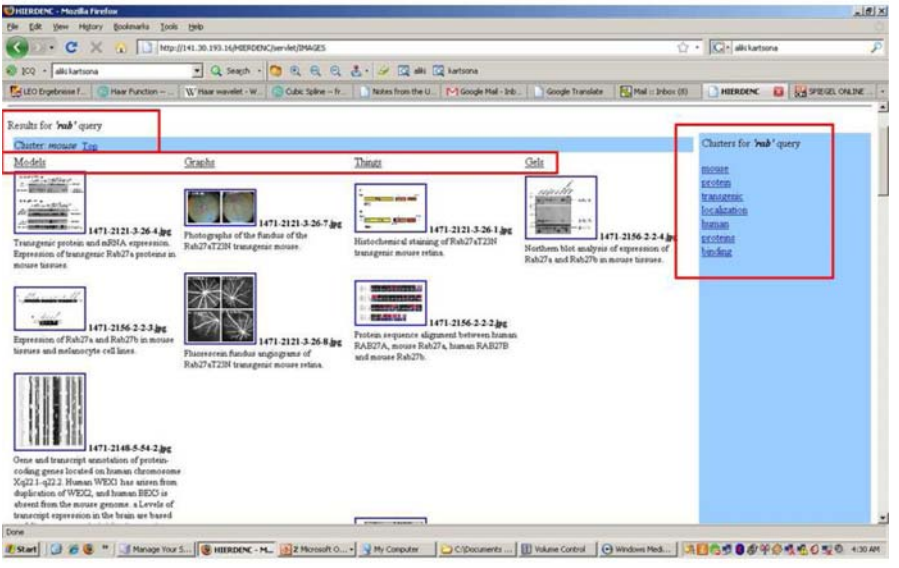


Fig. 1 The workflow of the HIERDENC image retrieval system. The system starts by converting passages (or image captions) to vectors of ontological terms. In continuation, the HIERDENC index is updated with the processed passages. User queries in the form of one or more keywords result in clusters of passages

text mining algorithms (2). The ontologies used for this purpose included MeSH and Gene Ontology (GO). The term extraction algorithm uses local sequence alignment of words of the passage and the words of GO terms. First we applied a tokenizer to the GO terms. The words of each term are then aligned against the passage text. Figure 3*a* shows an example of how we might extract ontological terms from a passage, in this case an image caption. The image caption contains terms from the Gene Ontology, MeSH, as well as a gene name. The ontological terms extracted from passages can be useful for

a)



b)

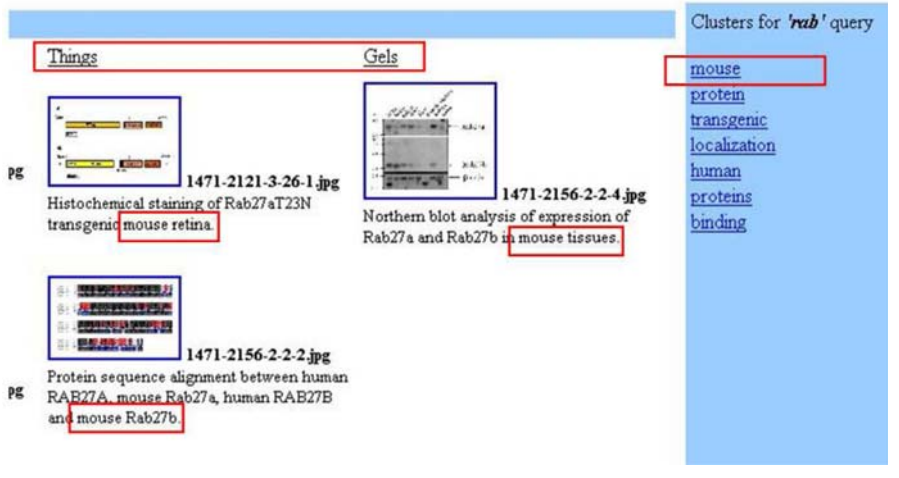


Fig. 2 *a*. A screenshot shows the results of passage retrieval for protein "rab" query. The HIERDENC system allows a user to search passages and retrieve clusters. In this example, the passages are image captions converted to ontological terms. The blue right-hand sidebar denotes clusters derived by text-based clustering, and clicking on a cluster name takes the user to the corresponding images. In this example, the clusters are defined by Gene Ontology and MeSH terms such as "mouse", "protein", "humans", "localization" and "binding". *b*. A zoom-in shows the clusters better. Columns denote classifications of images via image feature analysis done by an expert

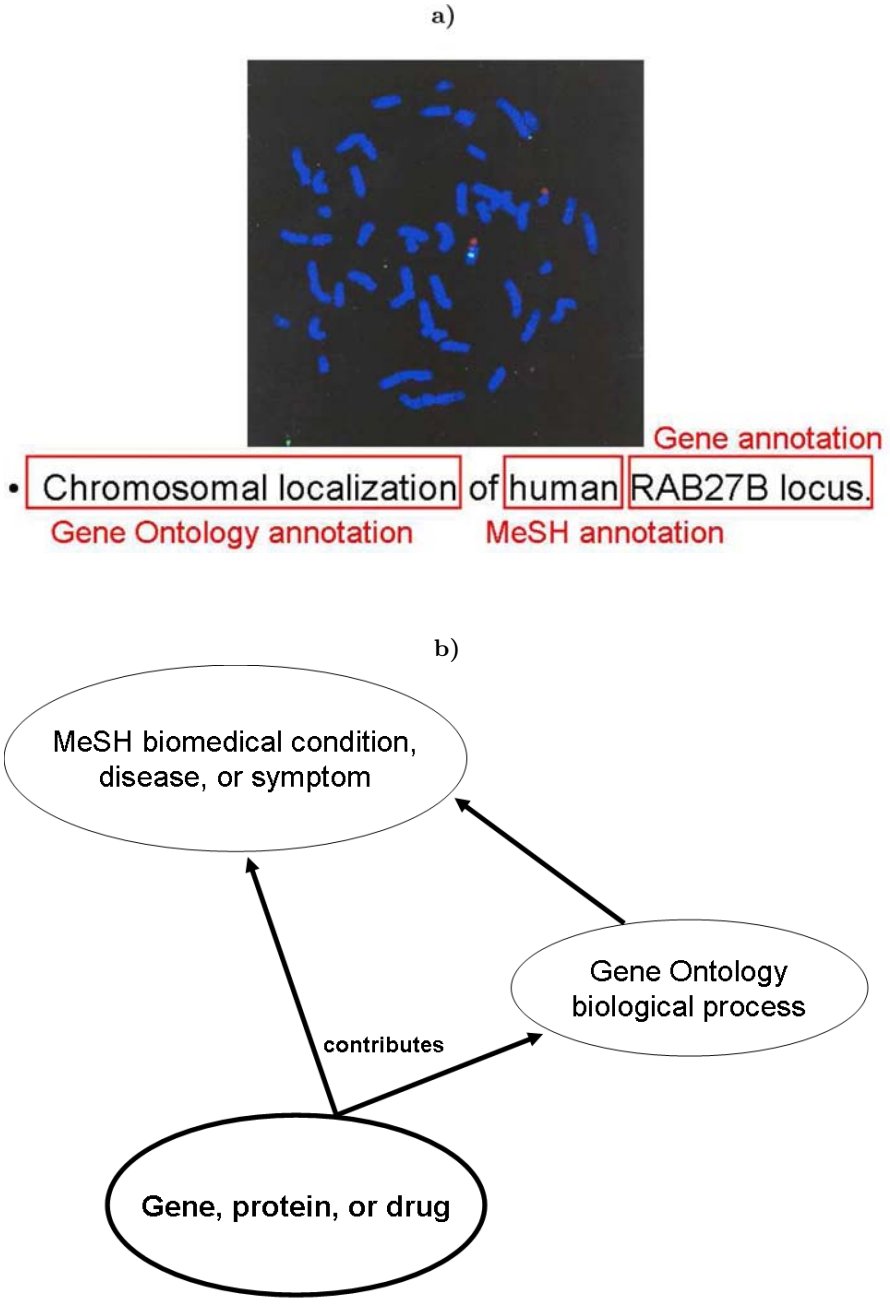


Fig. 3 Image caption ontological term extraction via text mining. *a*. An image caption on the Rab protein contains ontological terms from Gene Ontology, MeSH, and gene names. *b*. The co-occurrences of terms in the same caption can imply links between ontologies

clustering the passages more effectively and improving aspect-level retrieval performance. Figure 3*b* shows further how the ontological term extraction can be used to link different ontologies and vocabularies; terms that co-occur in the same passage imply a link between different ontologies, such as a gene contributing to a GO biological process which contributes to a MeSH medical condition.

For biclustering we used only the Gene Ontology/MeSH annotations of passages, transformed to boolean matrices, since the original passages would result in huge matrices. We tested the other clustering methods on both the original passages, as well as the ones converted to GO/MeSH ontological terms.

3.3 HIERDENC: Density-Based Clustering for Reranking Passages

We adapt the HIERDENC algorithm which was previously presented for *efficient density-based clustering of categorical data* (13; 25). The goal of this clustering adaptation is to rank the passages, such that: *a*. A highly ranked passage is representative of an aspect, identified in clustering as a relatively large group of similar passages. *b*. Top ranked passages cover many different aspects of the topic.

Basics. Let Π denote the set of all passages in our dataset. We define the cluster $\Pi_0(\pi_\mathbf{0}, \sigma) \subset \Pi$, centered at passage $\pi_\mathbf{0}$ with radius σ, as follows:

$$\Pi_0(\pi_\mathbf{0}, \sigma) = \{\pi : \pi \in \Pi \; and \; sim(\pi, \pi_\mathbf{0}) = \sigma\}.$$

The $sim(\cdot)$ is a similarity function representing the number of common words in two passages, defined as follows:

$$sim(\pi_\alpha, \pi_\beta) = |\pi_\alpha \cap \pi_\beta|$$

The *density* of a cluster $\Pi_X \subset \Pi$, where Π_X equals $\Pi_X(\pi_\mathbf{X}, \sigma) \subset \Pi$, involves the number of passages that are included in Π_X: $density(\Pi_X) = |\Pi_X|$, where $|\Pi_X|$ is the size of Π_X. This density can also be viewed as the likelihood that cluster $\Pi_X \subset \Pi$ contains a random passage from Π.

HIERDENC seeks the densest cluster $\Pi_0(\pi_\mathbf{0}, \sigma) \subset \Pi$. This is the cluster centered at $\pi_\mathbf{0}$ that has the most other passages from Π with a similarity of σ.

HIERDENC Ranking Algorithm and Discussion. The ranking is performed on the representative central passages of clusters; our goal is to rank higher passages that are centers of larger and more dense clusters, representing big distinct groups of passages. Every passage $\pi \in \Pi$ is the center

of a cluster with the maximum radius for which at least one other passage exists, $MaxSim_\pi$. We retrieve the clusters in order using the HIERDENC index, which supports finding the densest cluster of passages efficiently. The HIERDENC index is updated fast when a new passage is introduced.

For each passage π, the HIERDENC index stores three values determining the rank of the cluster centered at π: $MaxSim_\pi$ is the maximum similarity (cluster radius) found between π and any other passage; $PassSize_\pi$ is the length of π in terms of words; $NumSimPass_\pi$ is the number of passages that are cluster members with $MaxSim_\pi$ similarity to π, i.e., the size of the cluster centered at π. Figure 4 shows two clusters, which differ in terms of $MaxSim$, $PassSize$, and $NumSimPass$. The retrieved passages are ranked by decreasing $MaxSim$, increasing $PassSize$, and decreasing $NumSimPass$. The top-ranked passages are those retrieved for the highest value of radius $MaxSim$, the lowest $PassSize$ value, and the highest $NumSimPass$, capturing the centers of large clusters with similar passages. The decreasing $NumSimPass$ will give priority to larger clusters. The decreasing $MaxSim$ and increasing $PassSize$ are motivated by the Jaccard Index similarity measure; the Jaccard Index of two word vectors, π_α and π_β, results in a higher similarity for more common words and fewer overall words: $Jaccard\ Index(\pi_\alpha, \pi_\beta) = \frac{|\pi_\alpha \cap \pi_\beta|}{|\pi_\alpha \cup \pi_\beta|}$.

Figure 5 shows the pseudocode of the HIERDENC ranking process. For ranking the passages, we retrieve the passages using the HIERDENC index in the order described above. Then, we maintain a set Υ of the central and member passages of all clusters that were considered previously. We print the central passage under consideration if there is null intersection between its cluster members and Υ. Therefore, we print the central passages for the densest clusters, which are most likely to be representative of different aspects of the topic. If there is non-null intersection, then we put the central passage instead in a special ordered list Λ, which can be printed out after all passages have been iterated through if a user desires further results.

The HIERDENC index updating and cluster retrieval are efficient, achieving runtime scalability on the number of passages. For a new passage π, its most similar previous passages are found; this is done fast by maintaining each non-stop word's previous passage occurrences. Then the HIERDENC index is updated with the passages' $MaxSim$, $NumSimPass$, and $PassSize$ information. The first time the HIERDENC index is updated with N passages, the average runtime is $O(Nm)$, where m is the number of words (usually $m << N$). When n new passages are introduced, the updating of the index has a runtime of $O(nm)$. For the passages to be ranked by retrieving the centers of densest clusters, the worst-case runtime is $O(N)$; the ranking iterates until a maximum of N passages that are cluster centers are retrieved. The worst-case space complexity is $O(N^2)$, since for each passage information regarding the maximum similarity $MaxSim$ found to any other passage is stored; however, for large datasets, most pairs of passages have little similarity, significantly reducing the space requirement.

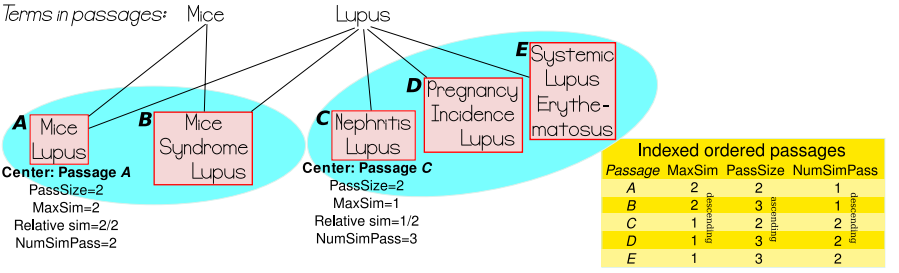


Fig. 4 HIERDENC indexing overview. Newly introduced passages involve updating the index. There are three fields involved in determining the rank for each passage: $MaxSim$ (descending), $PassSize$ (ascending), $NumSimPass$ (descending). The passages are records that are ordered accordingly in the HIERDENC index

```
function HIERDENC() {
    Π = retrieve list of ordered passages from HIERDENC index;
    Υ = empty set; //Holds passages already considered;
    Λ = empty list; //Holds remaining ordered passages;

    for passage π in Π:
        C = clusterCenteredAt(π, Π, MaxSim_π);
        if |C ∩ Υ| = 0:
            print π;
            Υ = Υ ∪ C;
        else: Λ = Λ ∪ π;
    print Λ;
}

function clusterCenteredAt(π, Π, MaxSim_π) {
    return {π_β ∈ Π|sim(π, π_β) = MaxSim_π};
}
```

Fig. 5 HIERDENC algorithm for retrieving passages in a ranked ordering, such that top-ranked passages include different aspects. Clusters C are retrieved in sequence based on their density, and the center π of each cluster C is printed if it does not overlap with any previous clusters. If there is overlap, then the center π is added to list Λ, which is printed out in the end after all passages have been considered

4 Evaluation: Results and Discussions

In the remainder of this chapter, we will discuss our evaluation: *a*. Passage retrieval scalability to very large datasets, *b*. Aspect-level retrieval performance, or how many aspects are represented by the top ranked passages.

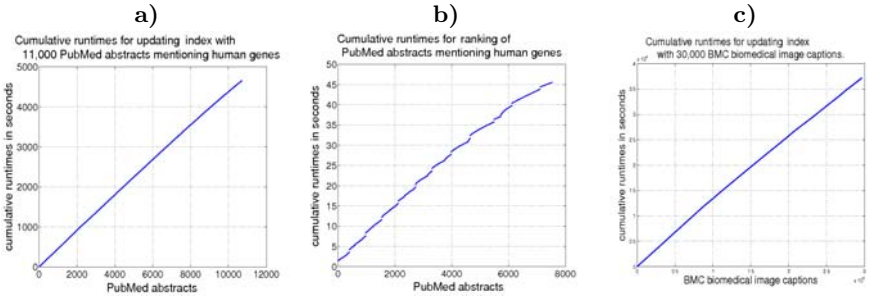


Fig. 6 HIERDENC index updating and passage reranking runtimes: *a*. HIERDENC index updating cumulative runtimes for a set of 11,000 PubMed passages containing human gene mentions, and *b*. Ranking cumulative runtimes, *c*. HIERDENC index updating cumulative runtimes for a set of 30,000 BMC biomedical image captions. In comparison, *k*-Means did not finish on clustering such a large image captions dataset

4.1 Scalability of HIERDENC to Large Image Datasets

Figure 6*a* shows the cumulative runtimes for updating HIERDENC with 11,000 documents that mention human genes in PubMed. Figure 6*b* shows the cumulative runtimes for clustering and returning all of the documents in a ranked ordering. To assess scalability further, we clustered the image captions of 30,000 BMC images published in the period 2000-2007. Figure 6*c* shows that the HIERDENC index updating runtimes scale with the number of image captions. These runtimes highlight the potential of HIERDENC as a text retrieval system that can deal with growing biomedical databanks.

4.2 Aspect-Level Retrieval on TREC 2007 Genomics Track

We evaluated HIERDENC's aspect-level retrieval performance on textual passages from the TREC 2007 Genomics track. To evaluate the HIERDENC aspect-level retrieval performance, we compare to separating the most relevant images using document clustering approaches: biclustering (9; 10), Hamming-distance (12; 13) and cosine-similarity (11) clustering.

TREC 2007 Genomics Track Topics and Evaluation. To evaluate aspect-level retrieval performance, we used the 36 topics from the TREC 2007 Genomics track. The topics are in the form of questions asking for lists of specific entities; these entities are based on controlled terminologies from different sources, with the source of the terms depending on the entity type.

Given a question, we initially retrieved 1,000 passages using the well-known OKAPI question answering system (26; 27; 28). Then, we updated the HIERDENC index with the passages and we retrieved clusters. Suppose that the information needed is: "What is the genetic component of alcoholism?" This is transformed into a question of the form: "What [GENES] are genetically linked to alcoholism?" Answers to this question will be passages that relate one or more entities of type GENE from MeSH terminology to alcoholism. For example, the following would be a relevant answer: "The DRD4 VNTR polymorphism moderates craving after alcohol consumption." The GENE entity supported by this statement would be DRD4.

The TREC results are evaluated based on how well they provide relevant information at three levels for a user trying to answer the given topic questions: passage retrieval, aspect retrieval, and document retrieval. The TREC statistic of Mean Average Precision (MAP) is the average precision at each point a relevant document or passage is retrieved. The evaluation measures for the TREC 2007 Genomics track are also called gold standard measures and have the following levels of retrieval performance (16):

Aspect-level MAP: A question could be addressed from different aspects. For example, the question "what is the role of gene PRNP in the Mad cow disease?" could be answered from aspects like "Diagnosis", "Neurologic manifestations", or "Prions/Genetics". This measure indicates how comprehensively the question is answered. Aspect retrieval was measured using the average precision for the aspects of a topic, averaged across all topics given their retrieved passages. The precision for the retrieval of each aspect was the fraction of relevant passages for the retrieved passages of a topic, up to the first passage in the ranked list that has the aspect assigned. These fractions at each point of first aspect retrieval were then averaged together to compute the average aspect precision. A relevant passage may have associated with it multiple aspects. Relevant passages that did not contribute any new aspects to the aspects of higher ranked passages were removed from the ranking, since the utility for a user of the same aspect occurring again further down the list is uncertain. Taking the mean over all topics produced the final aspect-based MAP (3).

Document-level MAP: This is the standard IR measure. The precision is measured at every point where a relevant document is obtained and then averaged over all relevant documents to obtain the average precision for a given query. For a set of queries, the mean of the average precision for all queries is the MAP of that IR system.

Passage-level MAP: As described in (17), this is a character-based precision calculated as follows: for each relevant retrieved passage, precision will be computed as the fraction of characters overlapping with the gold standard passages divided by the total number of characters included in all nominated passages from this system for the topic up until that point. Similar to regular MAP, relevant passages that were not retrieved will be

Table 1 2007 topic#1: "What serum [PROTEINS] change expression in association with high disease activity in lupus?". Frequent words and two-word phrases for the top 100 HIERDENC-ranked articles in increments of 10. These show different frequent contents for every 10 passages, and the contents indicate different aspects

	Top word frequency		**Two-word phrases frequency**	
Ranked	Word	Occ.	Expr.	Count
1-10	plasminogen	25	peptide elongation	10
11-20	purpura	9	anemia hemolytic	6
21-30	rickettsia	6	myocardial infarction	2
31-40	hepatitis	6	bone marrow	2
41-50	hepatitis	9	hepatitis evaluation	3
51-60	nervous	6	nervous system	6
61-70	thrombosis	7	thrombosis arteries	3
71-80	immune	10	immune response	10
81-90	contraceptive	3	postmenopause contraceptives	3
91-100	system	13	system development	2

added into the calculation as well, with precision set to 0 for relevant passages not retrieved. Then the mean of these average precisions over all topics will be calculated to compute the mean average passage precision".

Table 1 shows for the 2007 topic#1 the most significant words in the top 100 ranked passages, examined in increments of 10 passages. The significant words change between increments, showing that the top ranked passages cover several different aspects of the topic. Next, we compare HIERDENC's passage ranking to other clustering-based separation of relevant passages. To make the comparison meaningful, we evaluate all methods with the same TREC performance measures.

HIERDENC vs. Other Clustering Aspect-level Performance. We compare the aspect-level performance of HIERDENC on each topic to Hamming-distance and cosine-similarity based clustering methods that return two or more clusters. For Hamming-distance and cosine-similarity clustering, one can select the relevant passages based on cluster sizes: we prefer the smallest clusters (size ≤ 4) because they are more likely to correspond to diverse aspects of the topic in question. We produce results for the GO/MeSH ontology-converted passages and the original passages.

Table 2 shows the results. For all topics, HIERDENC ranking of retrieved passages improved the aspect-level performance over the Hamming-distance and cosine-similarity clusterings. The main reason for this is that HIERDENC considers all similarities found between passages in the dataset, and therefore can separate groups of passages considering whether their similarities are significant relative to the other similarities found. On the other hand, Hamming-distance and cosine-similarity clusterings do not consider the significance of a similarity relative to other similarities found elsewhere in the dataset. HIERDENC ranks all passages retrieved for a topic, while the latter

Table 2 Average aspect-level performance results over 36 TREC 2007 topics. We used HIERDENC clustering of passages for aspect-level performance, and three clustering methods to separate the relevant passages. We used the original passages as well as those converted to extracted GO/MeSH terms, except for biclustering where the original passages were too large for matrix computations. Small clusters have size ≤ 4 and are more likely to contain relevant passages than large clusters. HIERDENC ranking gives overall better aspect-level performance than Hamming-distance and cosine-similarity based clustering. Extracting ontological terms from passages results in improved apsect-level performance over using the original passages. Biclustering for separating the relevant passages gives better results on some topics. These results are also consistent across the other TREC evaluation measures that are summarised here: document-level, passage-level and passage2-level performance

Clustering	**Aspect**	**Document**	**Passage**	**Passage2**
HIERDENC GO/MeSH terms	**0.073**	**0.129**	**0.054**	**0.019**
HIERDENC original passages	0.034	0.127	0.049	0.013
Cosine-sim. GO/MeSH terms - LARGEST cluster	0.0167	0.0287	0.0012	0.00046
Cosine-sim. GO/MeSH terms - SMALL clusters	0.0449	0.0906	0.026	0.0098
Cosine-sim. original passages - LARGEST cluster	0.00562	0.01194	0.00084	0.00029
Cosine-sim. original passages - SMALL clusters	0.0458	0.1012	0.0303	0.0119
Hamming-distance GO/MeSH terms - SMALL clusters	0.02960	0.05902	0.01614	0.00565
Hamming-distance original passages - SMALL clusters	0.0231	0.07093	0.0206	0.00401
Biclustering GO/MeSH terms - SMALL clusters	**0.0749**	**0.1238**	**0.0591**	**0.0225**

consider the smallest clusters (size ≤ 4) to be the most relevant passages; however the smallest clusters may still exhibit insignificant similarity relative to the similarities found between other passages in the dataset. This suggests that one should consider a similarity in relation to the overall similarities found in a dataset.

Table 2 shows that for the Hamming-distance and cosine-similarity clusterings, taking the small clusters as relevant passages improves the results over taking the larger clusters. The reason is that in the smallest clusters the prominent words are more relevant to the topic than in the largest cluster. This especially holds true for topics with many aspects, where the smallest clusters are more likely to cover different aspects. For Hamming-distance clustering we notice a better result than for cosine similarity clustering, which can be explained by the gradually relaxing threshold making objects in small clusters to be similar to one another. Nevertheless, Hamming-distance clustering performance was not better than HIERDENC passage ranking.

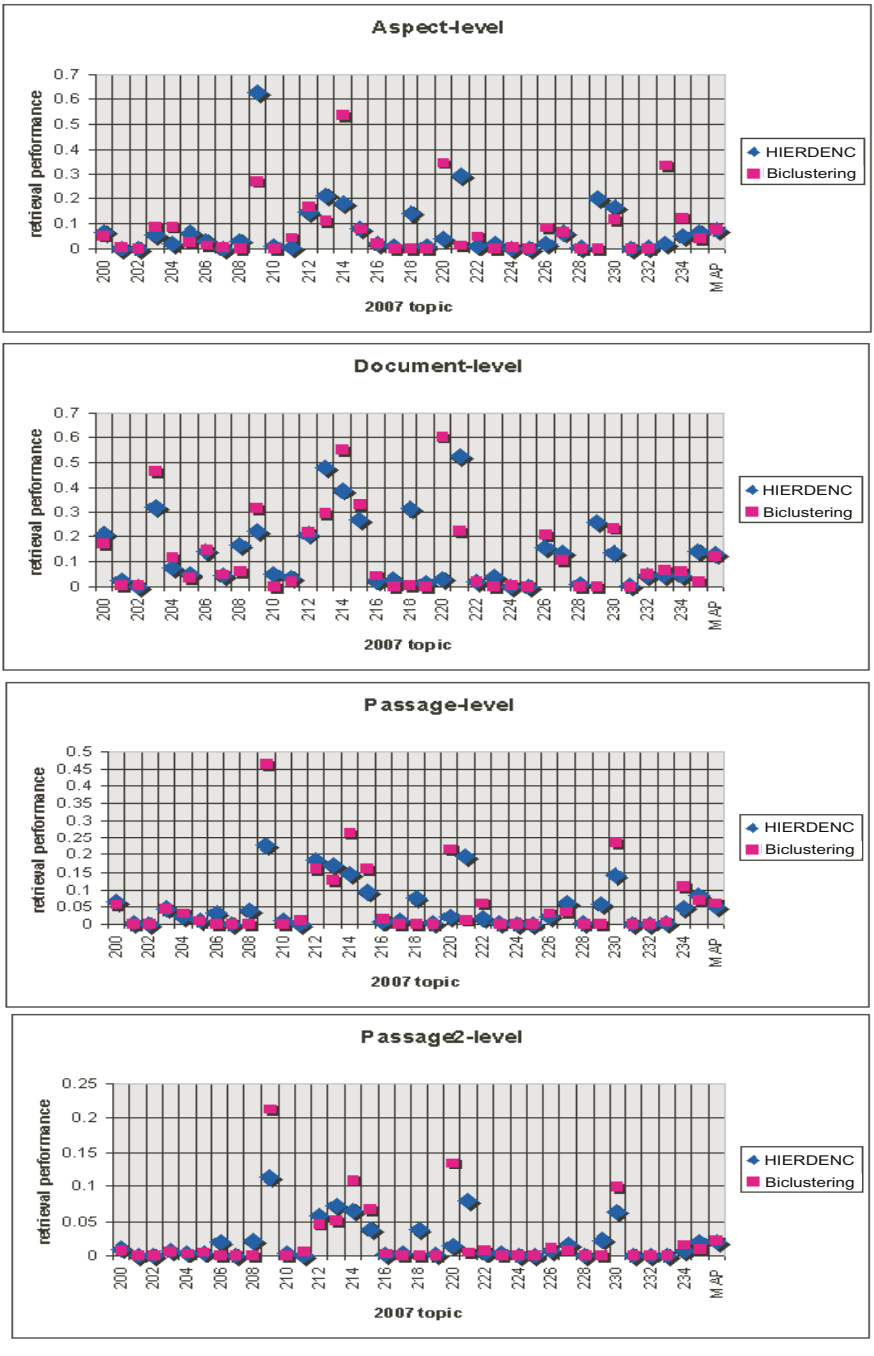


Fig. 7 All results for HIERDENC and biclustering across the 2007 topics (x-axis); *a*. Aspect-level, *b*. Document-level, *c*. Passage-level, and *d*. Passage2-level retrieval performance

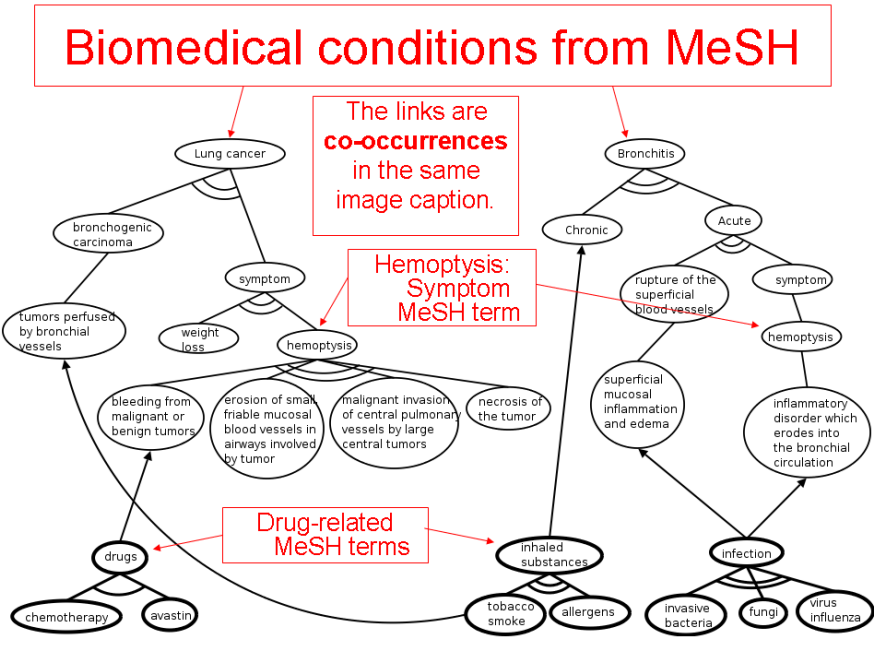


Fig. 8 Ontological terms that co-occur in the same passage (or image caption) indicate that both "Lung cancer" and "Bronchitis" involve the "Hemoptysis" symptom. Hemoptysis could be a symptom of lung cancer or bronchitis and in the former case there could be several causes (29). Serious hemoptysis often occurs in patients with lung cancer when treated with chemotherapy and Avastin. In this case, the incidence of hemoptysis is relatively high in patients receiving chemotherapy and Avastin, as compared to no cases in patients treated with chemotherapy alone. This figure shows that hemoptysis could also be a symptom of bronchitis; in this case it is mild and self-limited. Bronchitis is often a *viral* or *bacterial* disease which follows a cold or infection. A physician who diagnoses hemoptysis as a symptom of bronchitis may be wrong, since the patient could suffer from lung cancer instead. In fact, physicians who work long hours frequently make such errors. How can a physician tell which of all possible conditions holds for a patient with hemoptysis? With a unifying framework to integrate hemoptysis information online for fast lookup and analysis, a physician could make more informed decisions concerning the underlying cause of hemoptysis in a patient

Table 2 shows that in all cases converting passages to GO/MeSH ontological terms improved the result. Clustering the GO/MeSH terms extracted from passages has a significant effect, since stop word removal eliminated phrases like "of the" and "in the" that were considered prominent in the original top-ranked passages. Extracting ontological terms helps us to keep the semantically meaningful words for each passage that are more representative of the aspects.

We noticed more meaningful phrases in the top-ranked passages for GO/MeSH extracted terms than for the original passages.

HIERDENC vs. Biclustering Aspect-level Performance. We compare the performance of HIERDENC on each topic to biclustering that returns two clusters separating the relevant passages; for biclustering we use only extracted ontological terms from passages because of the huge sizes of the resulting matrices for the original passages.

Figure 7 shows all results for HIERDENC and biclustering on the various topics. The results often differ by topic. HIERDENC outperforms biclustering on topics with many aspects of retrieved passages. For topics with few (one or two) aspects, biclustering often outperforms HIERDENC, because of its focus on returning two clusters that are dissimilar. Therefore, biclustering succeeds in separating the relevant from the irrelevant passages, resulting in higher aspect-level retrieval performance values. The tradeoff between using HIERDENC vs. biclustering is that the former ranks all passages, while the latter returns only a subset of the passages predicted to be relevant; for some topics biclustering resulted in < 50 passages predicted to be relevant. For example, HIERDENC gave high aspect-level performance on the 2007 topic#10 (209) "What [BIOLOGICAL SUBSTANCES] have been used to measure toxicity in response to etidronate?". For this topic, the top 5 ranked passages by HIERDENC covered the diverse aspects of women, tumors, surgery, responses, and detoxification.

For some 2007 topics, HIERDENC gave significantly improved aspect-level performance over the original passage retrieval without using any clustering (4). An example is the topic#25 (224) "What [GENES] are involved in the melanogenesis of human lung cancers?". These topics request identifiers, allowing HIERDENC to use the surrounding terms of the identifiers to find the aspects.

5 Conclusions and Future Work

5.1 Conclusions

We complemented the HIERDENC clustering algorithm with an index that supports scalable ranking of textual passages, such as image captions. The top-ranked passages cover diverse aspects of a question on a topic. HIERDENC is useful for ranking passages for presentation to the user assuming several aspects exist, improving upon the results of other clustering methods. For topics with diverse aspects, HIERDENC results in improved aspect-level performance. HIERDENC is scalable to large and quickly growing datasets, such as the PubMed biomedical literature databank and biomedical image captions. Further benefits of HIERDENC include: no re-clustering needed when new text is presented, no user-specified input parameters required, and insensitivity to

ordering of passages. Other methods such as biclustering may be more useful for separating a subset of passages that are believed to be relevant for a topic. For all methods, using the GO/MeSH ontological term vectors extracted from text is likely to improve the aspect-level performance. This work provides guidelines for using clustering to improve aspect-level performance.

5.2 Future Work

Our methodology has potential to be extended to large biomedical image databanks, by using the textual image captions as passages. We are currently putting online a system for retrieving BMC biomedical images based on their captions `http://www.hierdenc.com` or `http://141.30.193.12/HIERDENC/images.html`.

Figure 3*b* showed that the ontological term extraction can also serve another purpose besides clustering and aspect-level performance: ontology linking if a pair of ontological terms from two different ontologies that co-occur in the same passage (or image caption). The main idea is to link ontological terms α and β from different ontologies or vocabularies if α and β co-occur in the same passage. We capture the following relations described in biomedical text, as shown in Figure 3*b*: a protein or drug contributes to a Gene Ontology biological process occurring over time; the GO biological process contributes to a MeSH medical condition; consequently the proteins contribute indirectly to the MeSH medical condition. The ultimate purpose of linking ontologies on the basis of co-occurring terms in passages is to reason over the information contained in biomedical text in a simpler manner than current semantic web-based integration frameworks would allow (22; 23; 24). Figure 8 shows a case of finding which medical condition is the most probable cause of a symptom. Suppose a patient is observed with the symptom *hemoptysis*, the act of coughing up blood (Figure 8). Hemoptysis is often a sign of lung cancer, but it may be caused by different underlying events in lung cancer patients. Hemoptysis also occurs in patients with acute or chronic bronchitis, as well as tuberculosis and pneumonia (29). Determining the cause of hemoptysis is often not a trivial matter. Figure 8 shows that a physician could use the linked ontologies over passages to find if the cause of hemoptysis in a patient is likely to be bronchitis or lung cancer. Linking ontologies is a step towards reasoning over existing medical knowledge, which may allow a physician to relate observed symptoms to a known medical condition, or find likely side-effects of a drug (30; 31; 32).

Acknowledgements. We are grateful for the financial support of the Natural Science and Engineering Research Council (NSERC), the Ontario Graduate Scholarship (OGS), the EU Sealife project, Dresden-exists, and the Nanobrain project.

References

[1] Vanteru, B.C., Shaik, J.S., Yeasin, M.: Semantically linking and browsing pubmed abstracts with gene ontology. BMC Genomics 9(suppl. 1), S10 (2008)

[2] Doms, A., Schroeder, M.: Gopubmed: exploring pubmed with the gene ontology. Nucleic Acids Res. 33(web server issue), W783–W786 (2005)

[3] Hersh, W., Cohen, A.M., Roberts, P.: TREC 2007 Genomics Track Overview. In: Proceedings of 16th Text REtrieval Conference. NIST Special Publication (2007)

[4] Huang, X., Hu, B., Rohian, H.: York University at TREC 2006: Genomics Track. In: Proceedings of 15th Text REtrieval Conference (2006)

[5] Deselaers, T., Mueller, H., Clogh, P., Ney, H., Lehmann, T.: The clef 2005 automatic medical image annotation task. International Journal of Computer Vision 74(1), 51–58 (2007)

[6] Goldberg, A., Andrzejewski, D., Van Gael, J., Settles, B., Zhu, X., Craven, M.: Ranking biomedical passages for relevance and diversity, uw-madison at trec genomics 2006. In: 15th Text Retrieval Conference, TREC 2006 (2006)

[7] Zhong, M., Huang, X.: Concept-based biomedical text retrieval. In: Proceedings of ACM SIGIR 2006 Conference (2006)

[8] Si, L., Lu, J., Callan, J.: Combining multiple resources, evidence and criteria for genomic information retrieval. In: 15th TREC Conference (2006)

[9] Dhillon, I.: Co-clustering documents and words using bipartite spectral graph partitioning. In: 7th ACM SIGKDD, pp. 269–274 (2001)

[10] Madeira, S., Oliveira, A.: Biclustering algorithms for biological data analysis: A survey. IEEE Transactions on Computational Biology and Bioinformatics 1(1), 24–45 (2004)

[11] Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco (1993)

[12] Andreopoulos, B., An, A., Wang, X., Faloutsos, M., Schroeder, M.: Clustering by common friends finds locally significant proteins mediating modules. Bioinformatics (Feburary 2007)

[13] Andreopoulos, B., An, A., Wang, X.: Hierarchical density-based clustering of categorical data and a simplification. In: Zhou, Z.-H., Li, H., Yang, Q. (eds.) PAKDD 2007. LNCS (LNAI), vol. 4426, pp. 11–22. Springer, Heidelberg (2007)

[14] Stairmand, M.A.: Textual Conext Analysis for Information Retrieval. In: Proceedings of the 1997 ACM SIGIR Conference (1997)

[15] Voorhees, E.M.: Query Expansion Using Lexical-Semantic Relations. In: Proceedings of the 1994 ACM SIGIR Conference (1994)

[16] Hersh, W., Cohen, A., Yang, J.: TREC 2005 Genomics Track Overview. In: Proceedings of 14th Text REtrieval Conference. NIST Special Publication (2005)

[17] Hersh, W., Cohen, A.M., Roberts, P.: TREC 2006 Genomics Track Overview. In: Proceedings of 15th Text REtrieval Conference. NIST Special Publication (2006)

[18] Huang, X., Zhong, M., Si, L.: York University at TREC 2005: Genomics Track.. In: Proceedings of the 14th Text Retrieval Conference (2005)

[19] Zhou, W., Yu, C., Neil, S., Vetle, T., Jie, H.: Knowledge-Intensive Conceptual Retrieval and Passage Extraction of Biomedical Literature. In: Proceedings of the 30th ACM SIGIR Conference (2007)

[20] Burek, P., Hoehndorf, R., Loebe, F., Visagie, J., Herre, H., Kelso, J.: A top-level ontology of functions and its application in the open biomedical ontologies. Bioinformatics 22(14), e66–e73 (2006)
[21] Garcia-Sanchez, F., Fernandez-Breis, J., Valencia-Garcia, R., Gomez, J., Martinez-Bejar, R.: Combining semantic web technologies with multi-agent systems for integrated access to biological resources. J. Biomed. Inform. 41(5), 848–859 (2008)
[22] Smith, A., Cheung, K., Krauthammer, M., Schultz, M., Gerstein, M.: Leveraging the structure of the semantic web to enhance information retrieval for proteomics. Bioinformatics 23(22), 3073–3079 (2007)
[23] Smith, A.K., Cheung, K.H., Yip, K.Y., Schultz, M., Gerstein, M.K.: Linkhub: a semantic web system that facilitates cross-database queries and information retrieval in proteomics. BMC Bioinformatics 8(suppl. 3), S5 (2007)
[24] Ruttenberg, A., Clark, T., Bug, W., Samwald, M., Bodenreider, O., Chen, H., Doherty, D., Forsberg, K., Gao, Y., Kashyap, V., Kinoshita, J., Luciano, J., Marshall, M.S., Ogbuji, C., Rees, J., Stephens, S., Wong, G.T., Wu, E., Zaccagnini, D., Hongsermeier, T., Neumann, E., Herman, I., Cheung, K.H.: Advancing translational research with the semantic web. BMC Bioinformatics 8(suppl. 3), S2 (2007)
[25] Andreopoulos, B., An, A., Wang, X., Labudde, D.: Efficient layered density-based clustering of categorical data. Elsevier Journal of Biomedical Informatics (2009) (in press)
[26] Beaulieu, M., Gatford, M., Huang, X., Robertson, S., Walker, S., Williams, P.: Okapi at TREC-5. In: Proc. of TREC-5. NIST Special Publication (1997)
[27] Huang, X., Peng, F., Schuurmans, D., Cercone, N., Robertson, S.: Applying machine learning to text segmentation for information retrieval. Information Retrieval Journal 6(4), 333–362 (2003)
[28] Huang, X., Huang, Y., Wen, M., Zhong, M.: York Univeristy at TREC 2004: Genomics and HARD Tracks. In: Proceedings of TREC-13. NIST Spec. Publ. (2004)
[29] Fartoukh, M., Khalil, A., Louis, L., Carette, M.F., Bazelly, B., Cadranel, J., Mayaud, C., Parrot, A.: An integrated approach to diagnosis and management of severe haemoptysis in patients admitted to the intensive care unit: a case series from a referral centre. Respir Res. 8, 11 (2007)
[30] Badea, L., Tilivea, D., Hotaran, A.: Semantic web reasoning for ontology-based integration of resources. In: Ohlbach, H.J., Schaffert, S. (eds.) PPSWR 2004. LNCS, vol. 3208, pp. 61–75. Springer, Heidelberg (2004)
[31] Brazhnik, O., Jones, J.: Anatomy of data integration. J. Biomed. Inform. 40(3), 252–269 (2007)
[32] Koehler, J., Philippi, S., Lange, M.: Semeda: ontology based semantic integration of biological databases. Bioinformatics 19(18), 2420–2427 (2003)