# Semantic Based Real-Time Clustering for PubMed Literatures

Ruey-Ling Yeh[1], Ching Liu[1], Ben-Chang Shia[2], I-Jen Chiang[3,4],Wen-Wen Yang[5], and Hsiang-Chun Tsai[4]

[1] Division of Biometrics, Graduate Institute of Agronomy, National Taiwan University, Taipei, Taiwan
[2] Department of Statistics and Information Science, Fu Jen Catholic University, Taipei, Taiwan
[3] Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei, Taiwan
[4] Institute of Biomedical Engineering, National Taiwan University, Taipei, Taiwan
[5] Graduate Institute of Medical Sciences, Taipei Medical University, Taipei, Taiwan
{d90621202, m485} @ntu.edu.tw, stat1001@mails.fju.edu.tw, {ijchiang, d102094022}@tmu.edu.tw, and ginnitsai@gmail.com

**Abstract.** This paper addresses to use the latent semantic topology to real-time cluster the literatures retrieved by PubMed in response to clinical queries and evaluates its performance by professional experts. The result shows that semantic clusters properly offer an exploratory view on the returned search results, which saves users' time to understand them. Besides, most experts conceive that the documents assigned to the identical cluster are similar and the concepts of clusters are appropriate.

**Keywords:** real-time, semantic clustering, combinatorial topology, Web Mining.

## 1 Introduction

An overwhelming amount of biomedical literatures stored in PubMed grows rapidly and becomes quickly diverse. Online Mendelian Inheritance in Man (OMIM) classifies varied PubMed literatures base on a biomedical taxonomy ontology. XplorMed [1] and GoPubMed [2] use the predefined classes from the MeSH or GeneOntology to classify biomedical literatures. However, the taxonomy needs a pile of laborious maintenance work and is unable to satisfy medical specialists' requests. It is necessary to classify the immensely retrieved literatures from PubMed immediately. Therefore, latent semantic clustering is considered to be one predominant approach [3] to automatically cluster data into meaningful groups.

Document clustering has been contemplated as one of the most pivotal techniques for dealing with the diverse and enormous amount of information on the World Wide Web. Traditional methods based on k-means, hierarchical clustering, and nearest neighbor clustering select a set of key terms or phrases to organize the feature vectors corresponding to different documents. Zamir *et al.* [4] presented a suffix-tree clustering (STC), which identified the sets of documents that share common phrases and formed

V. Corruble, M. Takeda, and E. Suzuki (Eds.): DS 2007, LNAI 4755, pp. 291–295, 2007.

document clusters depending on the similarity between documents. Ertoz *et al.*[5] proposed a clustering approach that found out the nearest neighbors of each data point and then identified core points and then built clusters.

The semantic topology-based method [3,6] yielded better results than the k-means, AutoClass, HCA, and PDDP [7] on classifying the high-dimensional data, such as the Web pages from [7], the newswire articles from the Reuters-21578, and so forth. In this paper, we apply it to produce a real-time clustering on a vast amount of biomedical literatures retrieved by PubMed in response to clinical queries. In our framework, documents are represented as a topology of features, e.g., keywords. An agglomerative clustering algorithm to construct a semantic hierarchy based on the combination of those features is in use to discover latent semantics behind those documents.

This paper is organized as follows. We briefly review feature selections and the latent semantic topology method in the next section. Section 3 discusses some experimental results and evaluations, followed by the conclusion.

## 2 The Method

We use Hidden Markov Models (HMMs) to generate a part-of-speech tagger [8] for biomedical literatures to extract noun phrases in a sentence. All the collections of noun phases are considered to be the key features in a document. Then *tfidf* indexing is applied to weight features. Those features with higher *tfidf* values are selected and put in the feature list of the document collection. We believe that the set of co-occurred key features (within in a short distance, e.g., a sentence or a paragraph; in our paper, we use a paragraph) reflects latent semantics in the collection. According to topological property it naturally organized a hierarchical lattice of the co-occurred feature sets, which is called semantic topology. The upper level hierarchy filters the verbose terms contained in the lower lever hierarchy, therefore, it induces more concrete and kernel information brings from some lower sets in the topology.

A latent semantic topology illustrates a hierarchy of concept disciplines associated with the extracting features. Basically, the algorithm is divided into three main parts (referring [6] in detail): first, to construct an undirected connected graph, i.e., a skeleton $S_0^1$ of the simplicial complex, from a data set; second, to generate the concepts from graph recursively; third, to cluster the data based on generated concepts.

We built a Web-based clustering search engine, it consists three layers as follows.

1. Presentation layer: This layer presents the search results and their hierarchical semantic structures.
2. Business layer: This layer contains the main processing logic of clustering. The statistical mechanism, Hidden Markov Model (HMM), is used for feature extraction. The features are extracted from the returned search results using the HMM-based part-of-speech tagger [8] to generate simplical complex to make an agglomerative clustering of the search results.
3. Data layer: The data layer stores metadata of the returned PubMed literatures, such as title, abstract, authors, and so on. Different parts of PubMed documents will be assigned different weights for document clustering.

## 3 PubMed Experiments and Results

We conducted an expert evaluation of total twenty-six volunteers who are pharmacists, medical engineers, physicians and public health experts. Three types of test queries, ambiguous, entity, and general terms showed in Table 1 were selected by them. Ambiguous terms yield multiple interpretations in biomedical fields. General terms cause common concepts in biomedical fields. Besides, they provided three entity terms in their respective fields for the evaluation. Some qualitative parameters [9-10] are chosen to evaluate the comprehension of auto-generated hierarchies.

**Table 1.** Three types of query terms

| Type | Query terms |
|---|---|
| Ambiguous terms | cure, drug, NEC, peer, neglect, response, channel, quality control, prime, order |
| Entity terms | NTG, celecoxib, metformin, vaccine, hospital, addiction, parkinson's disease, schizophrenia, spinal cord |
| General terms | NSAIDs, antibiotics, sex hormones, POCT, video game, medical devices, Speech Recognition, patient safety, X-ray, immunity |

The definition of each parameter used in the experiment is described in the following:

1.Summarization: whether the clusters at top level are enough summarization.
2.Missing concepts: whether the clusters at top level have any missing concepts.
3.Redundancy: whether the concepts of clusters at top level have redundancy.
4.Cohesiveness:whether the documents assigned to the identical cluster are similar.
5.Isolation: whether the clusters at the same level are discriminating and their concepts do not subsume one another.
6.General to specific concept: whether the generated concept hierarchy is traversed from broader concepts at the higher levels to narrower concepts at the lower levels.
7.Navigation balance: whether the fan-out at each level of the hierarchy is appropriate.
8.Readability:whether the concepts of clusters are appropriate.
9.Search Time: whether our online system compared with PubMed really helps in reducing time to locate information.

For each parameter, the evaluators were asked to rate each type on a scale of 1-10: a higher value indicates a higher agreement. Table 2 shows the results, median and quartile deviation for all terms. To evaluate reliability, internal consistency methods are widely used and Cronbach $\alpha$ is used to evaluate it. The 0.67 of Cronbach $\alpha$ implies a good reliability and credibility [11]. We find that the expert's opinions have a wide discrepancy even though they have the same discipline such as Physician. The results show a high diversity of “redundancy at top level” in ambiguous terms and entity terms for physician. One of the reasons is just like the feedback from some evaluators that the ambiguous terms have their innate equivocation, especially in biomedical domain. Besides, 3 evaluators replied in support of the cluster-based information retrieval that the clusters from the retrieved search results aroused them to concretize their information needs. The results (table 2 and fig.1) show that almost all evaluators agree our clustering search engine can reduce search time as compared with PubMed (V9).In

addition, most experts conceive that the documents assigned to the identical cluster are similar (V4) and the concepts of clusters are appropriate (V8).

**Table 2.** Expert study for evaluating the concept hierarchies

| Item | Ambiguous terms | | Entity terms for pharmacist | | Entity terms for physician | | Entity terms for public health | | General terms | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Me | Q.D. | Me | Q.D. | Me | Q.D. | Me | Q.D. | Me | Q.D. |
| Summarization at top level(V1) | 7.0 | 1.5 | 7.0 | 0.5 | 4.0 | 2.1 | 7.0 | 0.5 | 7.0 | 0.6 |
| Missing concepts at top level(V2) | 7.0 | 1.0 | 8.0 | 1.5 | 4.0 | 1.4 | 6.0 | 0.5 | 7.0 | 0.6 |
| Redundancy at top level(V3) | 3.0 | 2.5 | 6.0 | 2.0 | 6.5 | 3.4 | 5.0 | 0.0 | 6.0 | 1.4 |
| Overall cohesiveness(V4) | 8.0 | 1.0 | 7.0 | 1.5 | 6.5 | 1.3 | 8.0 | 0.0 | 7.0 | 1.0 |
| Overall isolation(V5) | 5.0 | 2.0 | 7.0 | 0.5 | 6.0 | 1.8 | 7.0 | 0.5 | 6.0 | 0.5 |
| Overall general to specific concept(V6) | 7.0 | 2.5 | 7.0 | 0.5 | 6.0 | 1.8 | 6.0 | 0.0 | 6.0 | 0.6 |
| Overall navigation balance(V7) | 5.0 | 1.0 | 6.0 | 1.0 | 6.0 | 1.8 | 7.0 | 0.5 | 6.5 | 0.6 |
| Overall readability(V8) | 6.0 | 0.5 | 9.0 | 1.0 | 6.0 | 1.0 | 7.0 | 0.5 | 7.0 | 1.0 |
| Overall search Time(V9) | 7.0 | 2.0 | 7.0 | 1.0 | 7.0 | 1.1 | 7.0 | 0.0 | 7.0 | 0.1 |

Note:1. Me indicates Median; 2. Q.D. indicates quartile deviation.

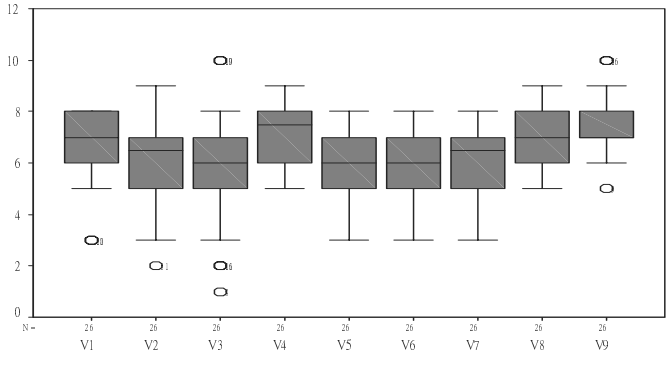


**Fig. 1.** The boxplots of items response

## 4 Conclusion

This paper applies the combinatorial topology-based semantic clustering method to real-time cluster search results from PubMed. Although the real-time clustering is not easy to be objectively evaluated, we attempt to built several measures as a tool of overall appraisal. The results demonstrate that building meaningful clustering search results from PubMed is useful for health professionals to save their time. Besides, most experts are in agreement on that the documents assigned to the identical cluster are similar and the concepts of clusters are appropriate.

**Acknowledgments.** We wish to thank the anonymous referees for their valuable comments which helped us to improve the paper.

## References

1. Perez-Iratxeta, C., Perez, A.J., Bork, P., Andrade, M.A.: Update on XplorMed: a web server for exploring scientific literature. Nucleic Acids Research 31, 3866–3868 (2003)
2. Doms, A., Schroeder, M.: GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Research 33, W783–W786 (2005)
3. Chiang, I.-J.: Discover the Semantic Topology in High- Dimensional Data. Expert Systems with Applications 33, 256–262 (2007)
4. Zamir, O., Etzioni, O.: Web document clustering: a feasibility demonstration. In: Proc. 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 98), pp. 46–54 (1998)
5. Ertoz, L., Steinbach, M., Kumar, V.: Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data. In: Proc. 2003 SIAM International Conference on Data Mining (SDM'03), San Francisco. CA, pp. 59–70 (2003)
6. Lin, T.Y., Chiang, I.-J.: A simplicial complex, a hypergraph, structure in the latent semantic space of document clustering. International Journal of Approximate Reasoning 40, 55–80 (2005)
7. Boley, D., Gini, M., Gross, R., Han, E.-H., Hastings, K., Karypis, G., et al.: Document categorization and query generation on the world wide web using webace. Artificial Intelligence Review 13(5–6), 365–391 (1999)
8. Dias Guilloré, S., Lopes, J.G.P.: Extracting Textual Associations from Part-Of-Speech Tagged Corpora. In: European Association for Machine Translation Workshop on Harvesting Existing Resources, Ljubljana, Slovenia (2000)
9. Chuang, S.-L., Chien, L.-F.: A practical web-based approach to generating topic hierarchy for text segments. In: Proc. ACM Conference on Information and Knowledge Management (CIKM'04), pp. 127–136 (2004)
10. Zheng, H.-J., He, Q.-C., Chen, Z., Ma, W.-Y., Ma, J.: Learning to cluster Web search results. In: Proc. SIGIR 2004, pp. 210–217 (2004)
11. Cronbach, L.J.: Coefficient Alpha and the Internal Structure of tests. Psychometrika 16(3), 297–334 (1951)