Identifying Latent Semantics in High-Dimensional Web Data

Ajit Kumar[1], Sanjeev Maskara[2], Jau-Min Wong [3], I-Jen Chiang [1, 3, *]

[1] Graduate Institute of Biomedical Informatics, Taipei Medical University, Taiwan
[2] Ovens and King Community Health Services, Wangaratta, Victoria, Australia
[3] Institute of Biomedical Engineering, National Taiwan University, Taipei, Taiwan
{d110099005,ijchiang}@tmu.edu.tw

Abstract. Search engines have become an indispensable tool for obtaining relevant information on the Web. The search engine often generates a large number of results, including several irrelevant items that obscure the comprehension of the generated results. Therefore, the search engines need to be enhanced to discover the latent semantics in high-dimensional web data. This paper purports to explain a novel framework, including its implementation and evaluation. To discover the latent semantics in high-dimensional web data, we proposed a framework named Latent Semantic Manifold (LSM). LSM is a mixture model based on the concepts of topology and probability. The framework can find the latent semantics in web data and represent them in homogeneous groups. The framework will be evaluated by experiments. The LSM framework outperformed compared to other frameworks. In addition, we deployed the framework to develop a tool. The tool was deployed for two years at two places - library and one biomedical engineering laboratory of Taiwan. The tool assisted the researchers to do semantic searches of the PubMed database. LSM framework evaluation and deployment suggest that the framework could be used to enhance the functionalities of currently available search engines by discovering latent semantics in high-dimensional web data.

Keywords: latent semantic manifold; semantic cluster; conditional random field; hidden Markov models; graph-based tree-width decomposition

1 Introduction

Gigantic repositories, including data, texts, and media have grown rapidly [1-5]. These are made available on the World Wide Web for the public use. The search engine tools assist users in searching contents relevant to them quickly [4]. However, the search engines often return inconsistent, uninteresting, and disorganized results due to various reasons [5, 6]. First, the web pages are heterogeneous and consist of varying quality [6, 7]. Second, the relationships among the words (polysemy, synon-

* Address: 250, Wuxing Street, Taipei -11031, Taiwan
Phone: +886 -2-27361661 Ext. 3343 Fax: +886-2-27392914

ymy, and homophony), sentences (paraphrase, entailment, and contradiction), and ambiguities (lexical and structural) put a limitation on search technologies that diminish the power of the search engines [8, 9]. Users have to devote substantial time to differentiate amongst meaningful items from the generated results [5, 10, 11]. Thus, the users felt a need that search engines should be enhanced to filter and organize meaningful items from the irrelevant results generated from the search queries [12, 13]. An effective search approach advocate to fit search results to the users' intent by discovering latent semantic in the generated documents, and then, classify documents into 'homogeneous semantic clusters' [14, 15]. In this approach, each semantic cluster is seen as a 'topic' that indicates a summary of the generated documents. Later, the users can explore the topics that are relevant to their intent. For example, a query term APC (Adenomatous Polyposis Coli) can be used to retrieve articles' abstract from the PubMed. However, the generated results would consist of not only articles about Adenomatous Polyposis Coli, but also others such as Antigen Presenting Cells (APC), Anaphase Promoting Complex (APC), and Activated Protein C (APC). The users need to find articles relevant to their intent (here Adenomatous Polyposis Coli) after going through the abstracts generated from the search. Similarly, a query term 'network' might generate different topics if it occurs near to a term such as computer, traffic, artificial neural, and biological neural in the context of searched documents. The generated results are desired to be relevant, not just outbound links pertaining to the query terms. In order to facilitate and enhance relevant information access to the web users, it is essential for search engines to deal with ambiguity, elusiveness, and impreciseness of the users' request [16].

Several researchers had made efforts towards semantic search of giant repositories. For example, a deterministic search provided metadata-enhanced search facility, wherein a user preselects different facets to generate more relevant search results [17]. However, scaling the metadata-enhanced search facility to the web is difficult and requires many experts to define controlled-vocabulary to create unique labels for concepts having the same terminology [18, 19]. A revolutionary change in information retrieval was realized by the introduction of the $\mathrm{tf}-\mathrm{idf}$ scheme [20-22]. In this scheme, the document collection is presented as a document-by-term matrix, which is usually enormously high dimensional and sparse. Often, for a single document, there are more than thousands of terms in a matrix, and most of the entries are zero. The $\mathrm{tf}-\mathrm{idf}$ scheme can reduce some terms; however, it provides the relatively small amount of reduction, which is not enough to reveal the statistical measures within or between document(s). In the last decades, some other dimension reduction techniques such as Latent Semantic Indexing, Probabilistic Latent Semantic Indexing, and Latent Dirichlet Allocation models were proposed to overcome some of these shortcomings. However, all these were bag-of-words models. These bag-of-words models follow Aldous and de Finetti theorem of exchangeability, wherein 'order of terms in a document' or 'order of documents in a corpus' can be neglected [23-25]. As the spatial information conveyed by the 'terms in the document' or 'documents in a corpus' was highly neglected, a statistical issue was found to be attached with these bags-of-words models [24-27]. In probability theory, the random variables (here referred as terms) $t_1, t_2, \cdots, t_N$, are said to be exchangeable if the joint distribution $F(t_1, t_2, \cdots, t_N)$ is

invariant under permutation of its arguments, so that $F(z_1, z_2, \cdots, z_N) = F(t_1, t_2, \cdots, t_N)$ where $z_1, z_2, \cdots, z_N$ is a permutation of $t_1, t_2, \cdots, t_N$. Thus, a semantic generates from somewhat co-occurring 'in relationships terms' and 'in the limited number of terms'. The criterion that 'order of terms in a document can be neglected' should be modified to 'the order of terms in a relationship of a document can be neglected'. Similarly, 'the order of documents in a corpus can be neglected' should be modified to 'the ordering documents in relationships of a corpus can be neglected'. For example, a query term 'network' would yield different 'topics' if it occurs nearby to a term such as 'computer', 'traffic', 'artificial neural' or 'biological neural,' and the 'order of terms in a relationship' might be neglected.

As we can see from the literature and arguments mentioned above, there was a need to enhance search engines to reveal latent semantics in high dimensional web data, while preserving the relationship and order of term(s) or document(s). Therefore, we proposed a Latent Semantic Manifold (LSM) framework that identifies homogeneous groups in web data, while preserving the spatial information of terms in a document, or documents in the corpus. This paper aims to explain the Latent Semantic Manifold framework (hereinafter, LSM framework), including its implementation and evaluation.

2 Materials and Methods

This study consists of three key components – proposal of a novel theoretical framework, implementation, and evaluation. They are explained in the following subsections.

2.1 Theoretical framework

The proposed Latent Semantic Manifold (LSM) framework is a mixture model based on the concepts of probability and topology, which identifies the latent semantic in data. The concepts deployed in LSM framework are explained in the following four steps. Figure 1 shows the high-level view of the framework.

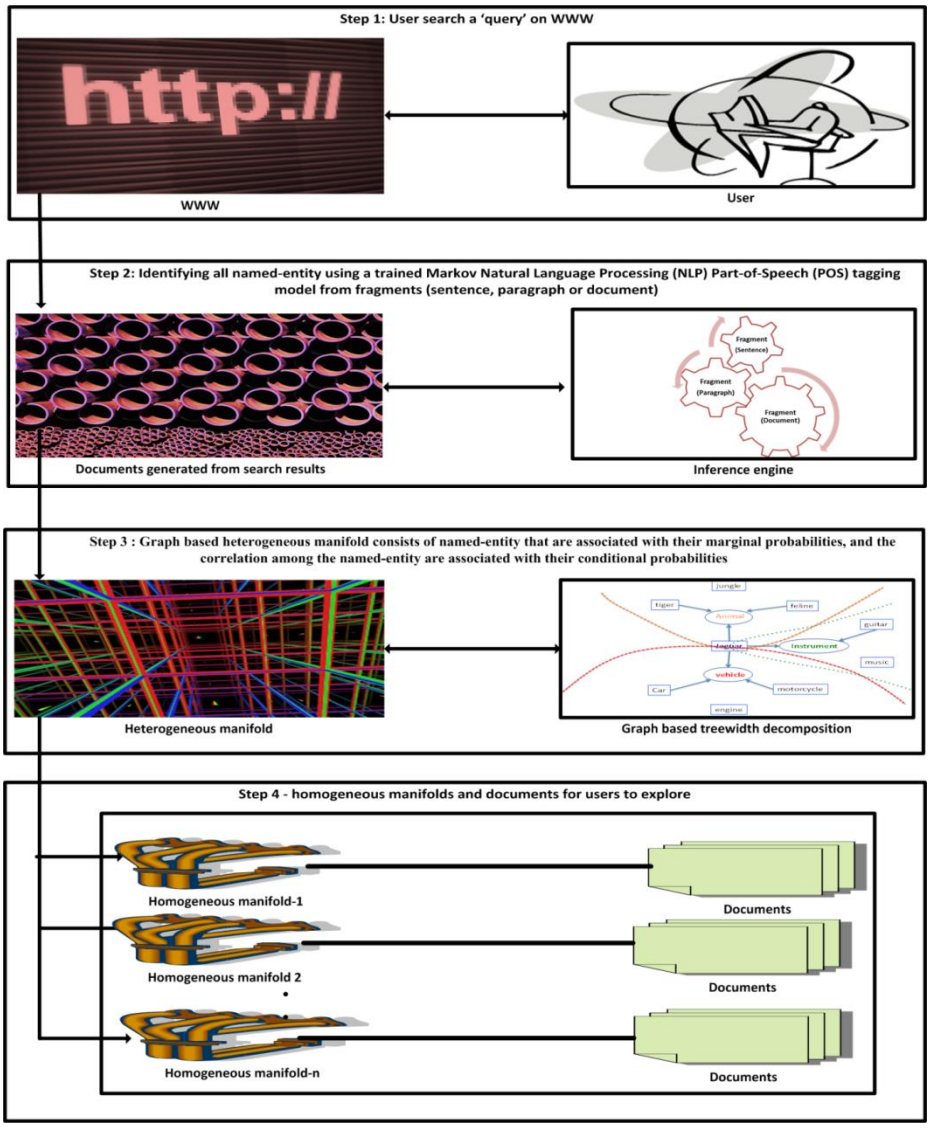

Fig. 1. Theoretical framework

Step 1: A 'query' entry for searching the high-dimensional web data

The user can enter the 'query' using a search engine that generates a set of documents. The generated documents need to be processed to get semantics, which can be a sentence, paragraph, section, or even a whole document. The generated documents are referred as 'fragments' in the following Step 2 and 3. For example, the sentence 'Jaguar is an animal living in Jungle' can be considered as a 'fragment' in Step 2 and 3. At times, 'fragment' has another meaning - context, which we have mentioned explicitly at the appropriate place.

Step 2: Named-entity recognition and heterogeneous manifold construction

The significant noun terms are identified from the 'fragments'. For example, if the sentence 'Jaguar is an animal living in Jungle' is considered to be fragmented; 'Jaguar,' 'animal,' and 'Jungle,' are significant 'noun terms'. Some natural language processing methods, called as named-entity recognition, are used to select named-entity (noun terms) and its 'type' [28]. The named-entity recognition and classification algorithms extract the named-entities (noun terms) from fragments, and then, classify those entities by 'type' such as person, organization, and location. For example, the 'jaguar' is considered as a named-entity, and it is assigned to the animal or vehicle 'type' depending on the fragment (context). The named-entities are indicated with their marginal probabilities, and the correlations among the named-entities are indicated with their conditional probabilities. As shown in Figure 2, Jaguar is a named-entity with three possible types – animal, vehicle, and instrument. It has marginal probabilities such as Panimal(Jaguar), Pvehicle(Jaguar), and Pinstrument(Jaguar). Similarly, it has conditional probabilities such as P (Jaguar, Car | Vehicle), P (Jaguar, Motorcycle | Vehicle).

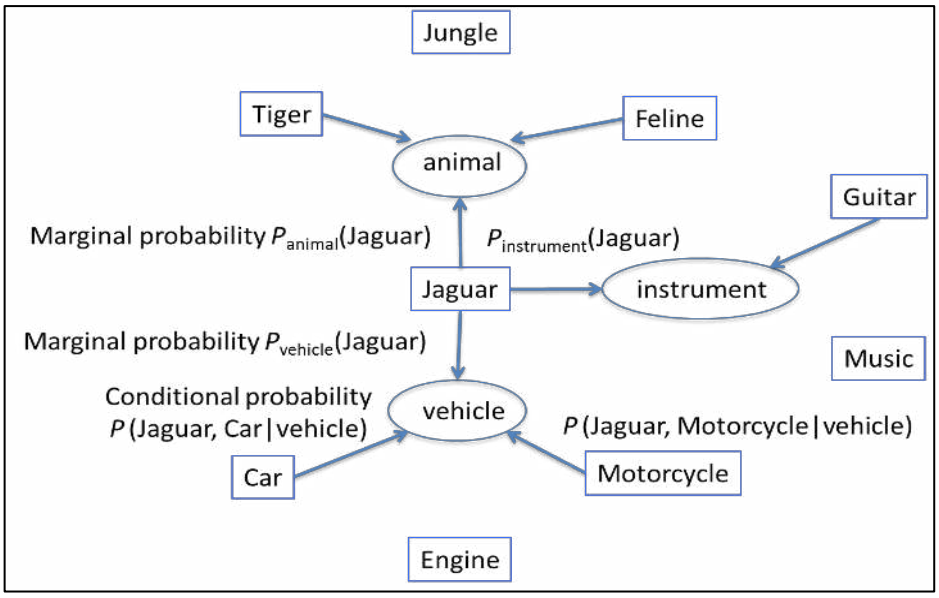

Fig. 2. An example to demonstrate named-entities, its types, and associated marginal and conditional probabilities

Although, we can enumerate all possible types of terms including their marginal and conditional probabilities using a large number of training documents; however, it is highly computational. Therefore, only nouns (words or phrases) are kept in reserve instead of identifying all types of terms and their probabilities [29-32]. The Hidden Markov Models (HMMs) were often used to draw 'terms' and their 'relationships' [33, 34]. In the last decade, a discriminative linear chain Conditional Random Field

(CRF) was also used to extract 'terms' in the corpus [35-39]. In this study, we used a trained Markov Natural Language Processing (NLP) Part-of-Speech (POS) tagging models to extract all named-entities (noun terms and its types) by the inferences of 'fragments' [31, 32]. The relationships among those named-entities construct a complex structure manifold. As the complex structure manifold is heterogeneous; therefore, we call it 'heterogeneous manifold' hereinafter.

Step 3: Decomposing a heterogeneous manifold into homogeneous manifolds

As mentioned in Step 2, the heterogeneous manifold consists of the complex structure of named-entities including estimates of marginal and conditional probabilities. A collection of fragment vectors lie on heterogeneous manifold, which contains some local spaces resembling Euclidean spaces of a fixed number of dimensions. Every point of the n-dimensional heterogeneous manifold has a neighborhood homeomorphic to the n-dimensional Euclidean space R^n. In addition, all points in the 'local spaces' are strongly connected. As the heterogeneous manifold is overly complex, and semantic is latent in 'local spaces'; therefore, instead of retaining just one heterogeneous manifold, we can break it into a collection of 'homogeneous manifolds'. The topological and geometrical concepts can be used to represent the latent semantics of a heterogeneous manifold as a collection of homogeneous manifolds. A graph-based treewidth decomposition algorithm is involved to decompose the a heterogeneous manifold into the collection of homogeneous manifolds [40]. As shown in Figure 3, assuming 'Jaguar' as heterogeneous manifold, we can decompose it into three 'homogeneous manifolds' bounded by dotted lines of three different colors.

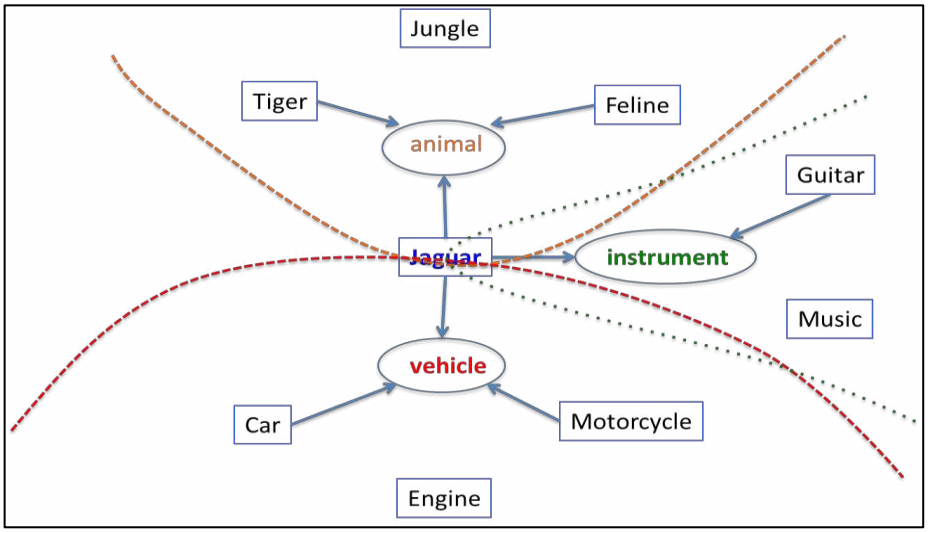

Fig. 3. An example to demonstrate Graph-based treewidth decomposition

In the graph-based treewidth decomposition algorithm, we can start by selecting a random 'fixed dimension local manifold' to be a separator as shown in Figure 4 [41].

Later, the local manifold is decomposed into two local manifolds that are not adjacent. This decomposition is recursive until no further decomposition is possible.

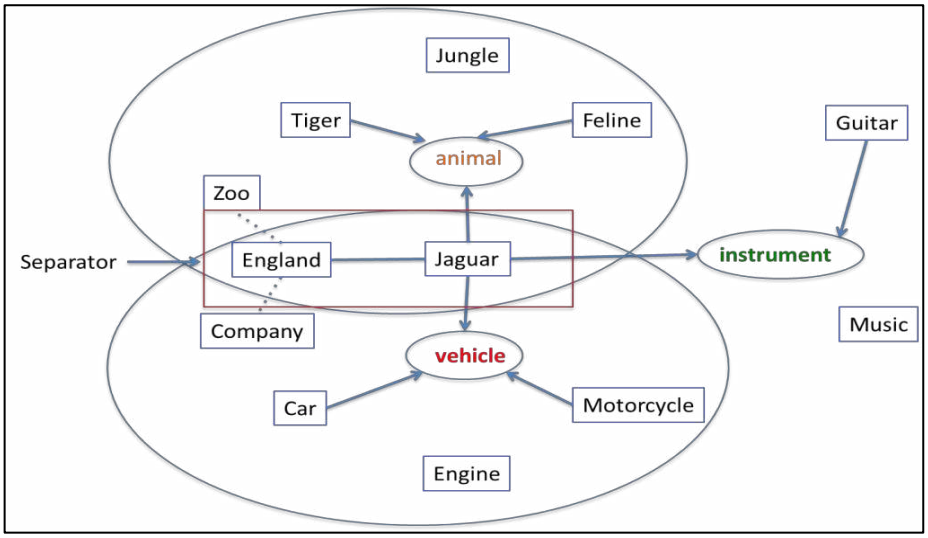

Fig. 4. An example to demonstrate the concept of separator under Graph-based treewidth decomposition

We can express the above concept formally - let a heterogeneous manifold M_i for fragment i be the set of homogeneous manifolds such that $M_i = \{M_{ij} \mid \text{No } M_{ij} \text{ is a subset of } M_{ik}, j = k\}$. The semantics generated from local homogeneous manifolds, which are equipped with fragments, are independent. In addition, a semantic topic set $C = \{z_1, z_2, \cdots, z_m\}$ of the returned documents is associated with a semantic mapping $f(M_{ij}) \in C$ with a probability $P(M_{ij}, z_k) \in [0, 1]$, and quantity $f(M_{ij}) = z_k$. The probabilities indicate how many documents pertaining to a homogeneous manifold are relevant and match the user's intent. To induce homogeneous manifolds, it is crucial to extract significant 'terms' from fragments. In addition, we should demonstrate the relevance of each fragment to the homogeneous manifold. The users can refer only those homogeneous manifolds' associated fragments, which they want.

Step 4: Exploring homogeneous manifold

The search-generated documents (referred as fragments in step 2 and 3) are clustered to their related homogeneous manifolds. For example, a query by the user for the term APC, the documents returned from the queried term aggregated into a collection of homogeneous manifolds as shown in Figure 5. Each document is assigned to a particular homogeneous manifold. The occurrence of a particular document in the whole set of documents denotes its significance in homogeneous manifold.

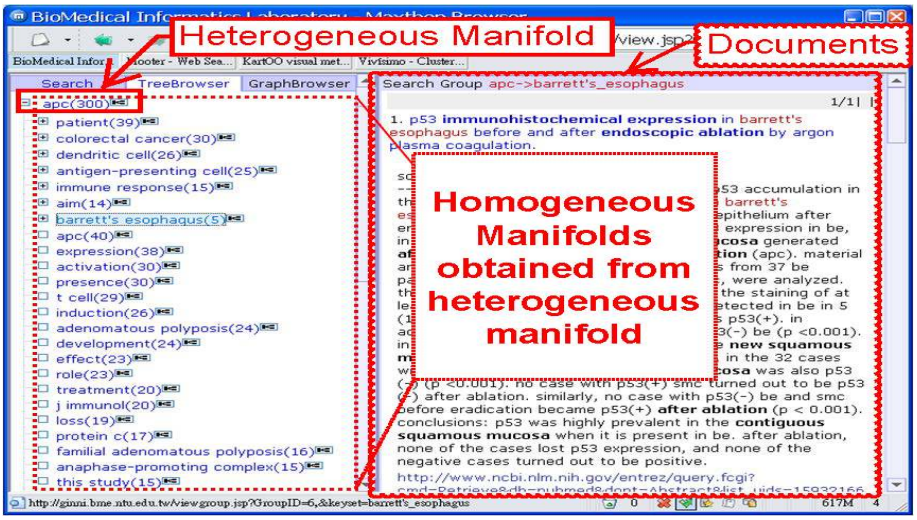

Fig. 5. An example to demonstrate heterogeneous manifold, homogeneous manifolds, and documents associated with homogeneous manifolds

2.2 Implementation

The LSM framework was implemented using the Eclipse Software Development Kit. A team of three researchers, who were expert in the Java programming language, developed the entire system. The development took almost 11 months. We provided a straightforward search interface facility as shown in Figure 6 in the Result section. The output of a user queried term (for example, APC) is shown in Figure 7 in the Result section. The system was deployed for two years at two places - library and one biomedical engineering laboratory of Taiwan. This system assisted the researchers to perform semantic searches of the PubMed database. For example, a researcher can search APC with Adenomatous Polyposis Coli as his or her intended meaning. However, APC can also have meaning such as Antigen-Presenting Cells, Anaphase Promoting Complex, or Activated Protein C among others. For instance, in a homogeneous manifold, if APC, Colorectal Cancer, and Gene related documents are assembled, homogeneous manifold would point out the APC as Adenomatous Polyposis Gene. Similarly, an APC, Major Histocompatibility Complex and T-cells related documents are assembled; it would indicate APC as Antigen Presenting Cells. In the Result section, the Figure 8 shows that documents returned from the queried term APC can automatically associate to homogeneous manifolds (semantic topics). In addition, the researchers can obtain a different 'vantage point' based on the underlying data. For example, a PubMed search retrieved 300 randomly selected published or in-press articles' abstracts for a medical term NOD2. Figure 9 shows latent semantic topics as a clustering result. According to the result, inflammatory bowel disease and its type (Crohn's disease and ulcerative colitis) are associated with gene NOD2. The term NOD2 was found to be evenly spread over these three topics - inflammatory bowel

disease and its type. Some evolving topics such as 'bacterial component' were also discovered. However, when we searched NOD2 on Genia corpus[†], the result was different as shown in Figure 10 [42].

2.3 Experiment

Data Sets. Two data sets, Reuters-21578-Distribution-1 and OHSUMED, were used to evaluate performance of the LSM framework and its implementation. The Reuters-21578-Distribution-1 collection consists of Newswire articles. The articles were classified into 135 topics, which were used to affirm the clustering results. In the test, the documents with multiple topics (category labels) and single topic were separated. The topics, which had less than five documents, were removed. Table 1 shows the summary of the Reuters-21578-Distribution-1 collection.

Table 1. Statistics of Reuters-21578 corpora

Statistics	Number of topics	Number of documents	Documents on a topic
Origin	135	21578	0-3945
Single Topic	65	8649	1-3945
Single Topic (>=5 documents)	51	9494	5-3945

OHSUMED is a clinically oriented Medline collection, consisting of 348,566 references. It covers all the references from 270 medical journals of 23 disease categories over a five-year period (1987-1991) [43].

Evaluation criteria. The experimental evaluation of the LSM framework measured both effectiveness and efficiency. Effectiveness is defined as an ability to identify the right cluster (collection of documents). In order to measure the effectiveness, the clusters generated were verified by human experts as shown in Table 2.

Table 2. Contingency table for category (c_i, *where* i = *natural number)*[‡]

Category		Clustering results	
		Yes	No
Expert Judgment	Yes	TP_i	FN_i
	No	FP_i	TN_i

[†] Genia corpus contains 1,999 Medline abstracts, selected using a PubMed query for the three MeSH terms 'human,' 'blood cells,' and 'transcription factors'.

[‡] TP – True Positive; FP – False Positive; FN – False Negative; TN – True Negative

The three measures of the effectiveness of clustering methods (Precision, Recall, and F_β) were calculated using the contingency Table 1. The Precision and Recall are defined respectively as follows.

$$\begin{aligned} \text{Precision}_\text{i} &= \frac{\text{TP}_\text{i}}{\text{TP}_\text{i} + \text{FP}_\text{i}} \\ \text{Recall}_\text{i} &= \frac{\text{TP}_\text{i}}{\text{TP}_\text{i} + \text{FN}_\text{i}} \end{aligned} \tag{1}$$

The F_β measure, which combines Precision and Recall, is defined as follows.

$$\text{F}_\beta = \frac{\left(\beta^2 + 1\right) \times \text{Precision}_\text{i} \times \text{Recall}_\text{i}}{\beta^2 \times \text{Precision}_\text{i} + \text{Recall}_\text{i}} \tag{2}$$

F_1 measure is used in this paper, which is obtained assigning β to be 1, which means that precision and recall have equal weight for evaluating the performance. In case, many categories are generated and compared, the overall precision and recall are calculated as the average of all precisions and recalls belonging to various categories. F_1 is calculated as the mean of all results, which is a macro-average of the categories.

In addition, two other evaluation metrics, normalized mutual information and overall F-measure were also used [44-46]. Given the two sets of topics C and Cl, let C denote the topic set defined by experts and Cl denote the topic set generated by a clustering method, and both derived from the same corpora X. Let N(X) denotes the number of total documents, N(z, X) denotes the number of documents in topic z, and N(z, z', X) denotes the number of documents both in topic z and topic z', for any topics in C. The normalized mutual information (NMI) metric MI (C, C') is defined as follows.

$$MI(C,C') = \sum_{z \in C, z' \in C'} P(z,z') \log_2 \left(\frac{P(z,z')}{P(z)P(z')} \right) \tag{3}$$

Where $P(z) = N(z,X)/N(X)$, $P(z') = N(z',X)/N(X)$, and $P(z,z') = N(z,z',X)/N(X)$. The normalized mutual information metric $MI(C,C')$ will return a value between zero and $max(H(C),H(C'))$, where H(C) and H(C') define the entropies of C and C' respectively. The higher $MI(C,C')$ value means that two topics are almost identical, otherwise more independent. The normalized mutual information metric $MI(C,C')$ is, therefore, transferred to be

$$\text{MI}(C,C') = \frac{MI(C,C')}{\max(H(C),H(C'))} \tag{4}$$

Let Fi be an F-measure for each cluster z_i defined above. The overall F-measure can be defined as:

$$F^* = \sum_{z' \in C'} P(z') \times \max_{z \in C} F(z, z') \quad (5)$$

Where F (z, z') calculates the F-measure between z and z'.

2.4 Evaluation

The experiments were conducted on Reuters-21578-Distribution-1 and OHSUMED dataset. The clusters, from two to ten, were selected randomly to evaluate LSM and other clustering methods. Fifty test runs were conducted for the randomly chosen clusters from the corpus, and the final performance scores were obtained by averaging the scores from the 50 test runs [44]. The t-test assessed whether homogeneous clusters generated by the two methods (LSM vs. Other methods) were statistically different from each other as shown in as Table 3 and Figure 11 in the Result section. We also calculated the overall F-measure in combination of arbitrary 'k' clusters that were uniquely assigned to topics from the Reuters-21578-Distribution-1 dataset, where k was 3, 15, 30, and 60 [47]. Fifty test runs were also performed using these LSM results to compare 'Frequent Item set-based Hierarchical Clustering (FIHC)' and 'bisecting k-means' as shown Table 4 and Figure 12 in the Result section [47, 48]. The average precision, recall, overall F-measure, and normalized mutual information of LSM, LST, PLSI, PLSI + Ada Boost, LDA, and CCF was evaluated on the Reuters-21578-Distribution-1 dataset; and LSM, LST, and CCF were evaluated on an OHSUMED dataset as shown in Table 5 in the Result section [26, 49-52]. Besides the effectiveness, an efficiency testing was performed on LSM, LST, and CCF as shown in Figure 13 in the Result section.

3 Results

3.1 LSM implementation results

Figures 6 and 7 are input and output interface of the system. Figure 8 shows the results of query term 'APC' and explains potential functionalities that can be enhanced in the search engines using the LSM framework. Figure 9 and 10 different views of a query term NOD2 due to different underlying data sets.

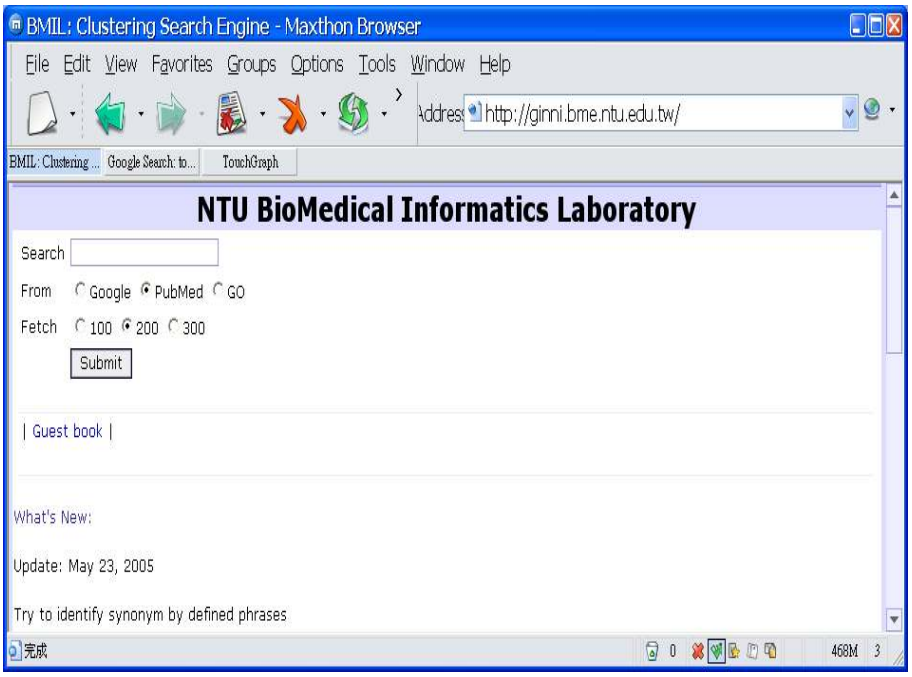

Fig. 6. Users search 'APC' in the above input interface

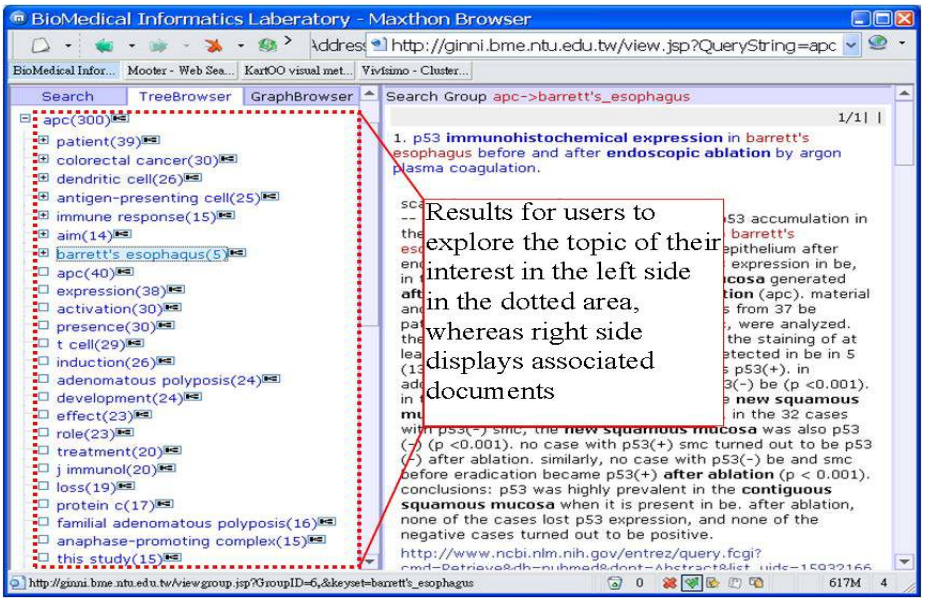

Fig. 7. The clustering result of the query term 'APC' retrieved from PubMed as output

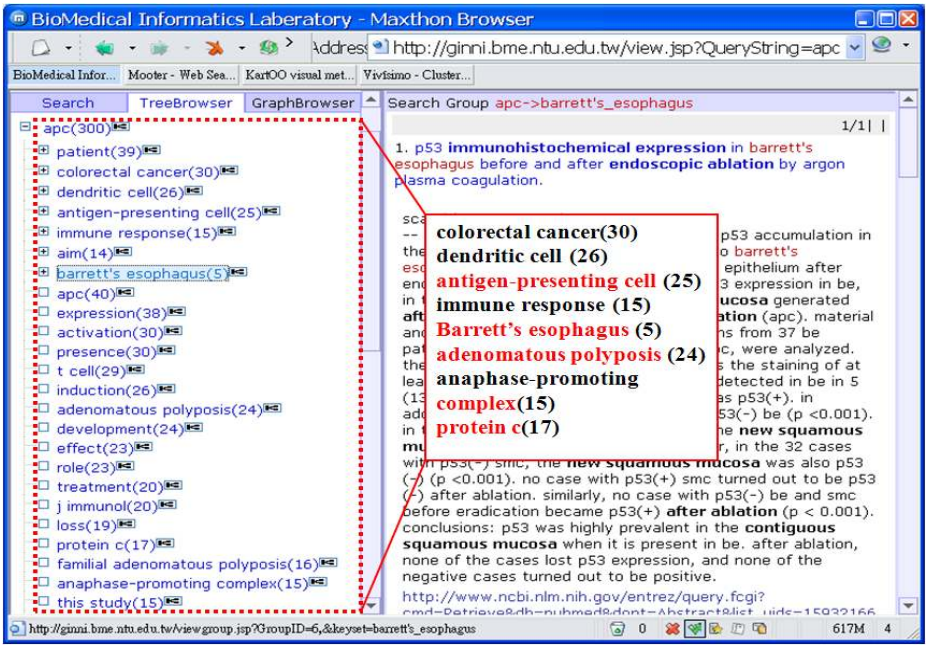

Fig. 8. Result from 'APC' term query

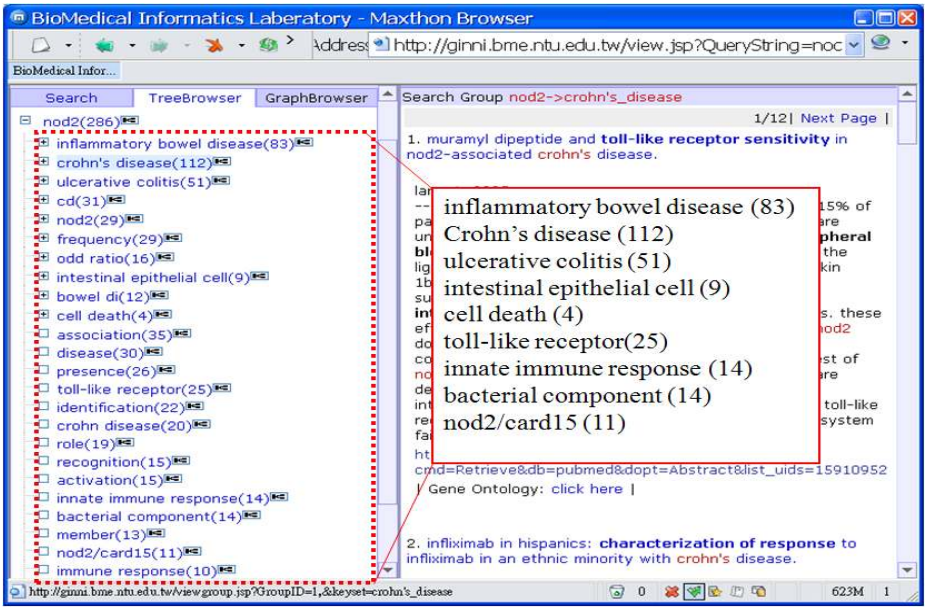

Fig. 9. The clustering result of the query term 'NOD2' retrieved from PubMed

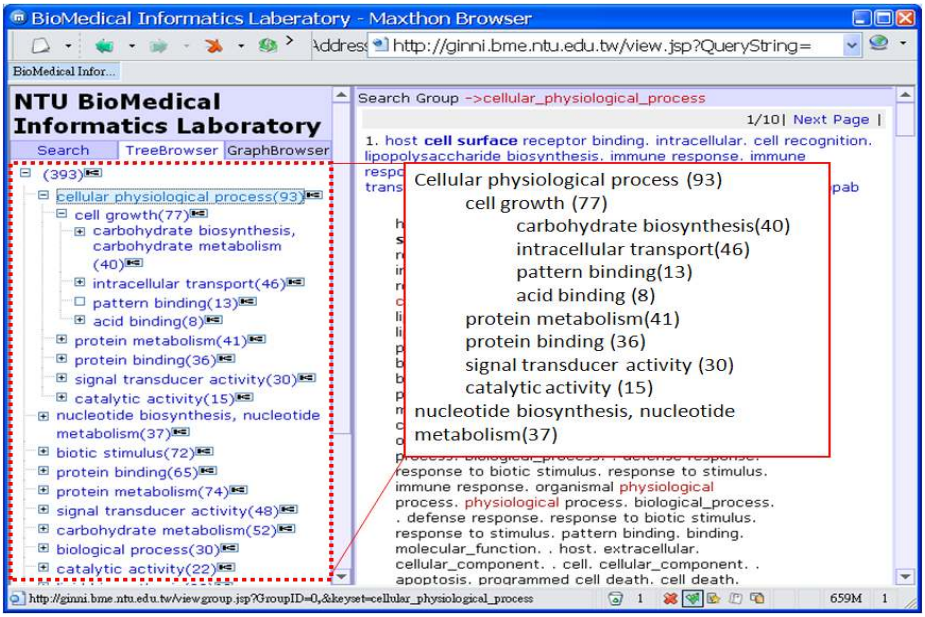

Fig. 10. Clustering result of the query term 'NOD2' retrieved from Genia corpus parsed with biological Conditional Random Fields (CRFs)

3.2 Experiment results

Normalized Mutual Information comparison of the LSM framework with the other sixteen methods using Reuters-21578-Distribution-1 dataset is shown in Table 3 and Figure 11 [44, 52-54].

Table 3. Normalized Mutual Information comparison of LSM framework with other sixteen methods using Reuters-21578-Distribution-1 dataset[§]

k	2	3	4	5	6	7	8	9	10	Average
LSM	0.461	0.505	0.622	0.686	0.714	0.792	0.893	0.884	0.9	0.717
CCF	0.569	0.563	0.607	0.62	0.605	0.624	0.633	0.647	0.676	0.616
GMM	0.475	0.468	0.462	0.516	0.551	0.522	0.551	0.557	0.548	0.517
NB	0.466	0.348	0.401	0.405	0.409	0.404	0.435	0.411	0.418	0.411
GMM + DFM	0.47	0.466	0.45	0.513	0.531	0.506	0.535	0.535	0.536	0.505
KM	0.404	0.402	0.461	0.525	0.561	0.548	0.583	0.597	0.618	0.522
KM-NC	0.438	0.462	0.525	0.554	0.592	0.577	0.594	0.607	0.618	0.552
SKM	0.458	0.407	0.499	0.561	0.567	0.558	0.591	0.598	0.619	0.54
SKM-NCW	0.434	0.423	0.515	0.556	0.577	0.563	0.593	0.602	0.612	0.542
BP-NCW	0.391	0.377	0.431	0.478	0.493	0.5	0.519	0.529	0.532	0.472
AA	0.443	0.415	0.488	0.531	0.571	0.542	0.587	0.594	0.611	0.531
NC	0.484	0.461	0.555	0.592	0.617	0.594	0.64	0.634	0.643	0.58
RC	0.417	0.381	0.505	0.46	0.485	0.456	0.548	0.484	0.495	0.47
NMF	0.48	0.426	0.498	0.559	0.591	0.552	0.603	0.601	0.623	0.548
NMF-NCW	0.494	0.5	0.586	0.615	0.637	0.613	0.654	0.659	0.658	0.602
CF	0.48	0.429	0.503	0.563	0.592	0.556	0.613	0.609	0.629	0.553
CF-NCW	0.496	0.505	0.595	0.616	0.644	0.615	0.66	0.66	0.665	0.606

[§] LSM – Latent semantic manifold; CCF – k-clique community finding algorithm; GMM – Gaussian mixture model; NB – Naive Bayes clustering; GMM + DFM – Gaussian mixture model followed by the iterative cluster refinement method; KM – Traditional k-means; KM-NC – Traditional k-means and Spectral clustering algorithm based on normalized cut criterion; SKM – Spherical k-means; SKM-NCW – Normalized-cut weighted form; BP-NCW – Spectral clustering based bipartite normalized cut; AA – Average association criterion; NC – Normalized cut criterion; RC – Spectral clustering based on ratio cut criterion; NMF – Non-negative matrix factorization; NMF-NCW – Nonnegative Matrix Factorization-based clustering; CF – Concept factorization; CF-NCW – Clustering by concept factorization

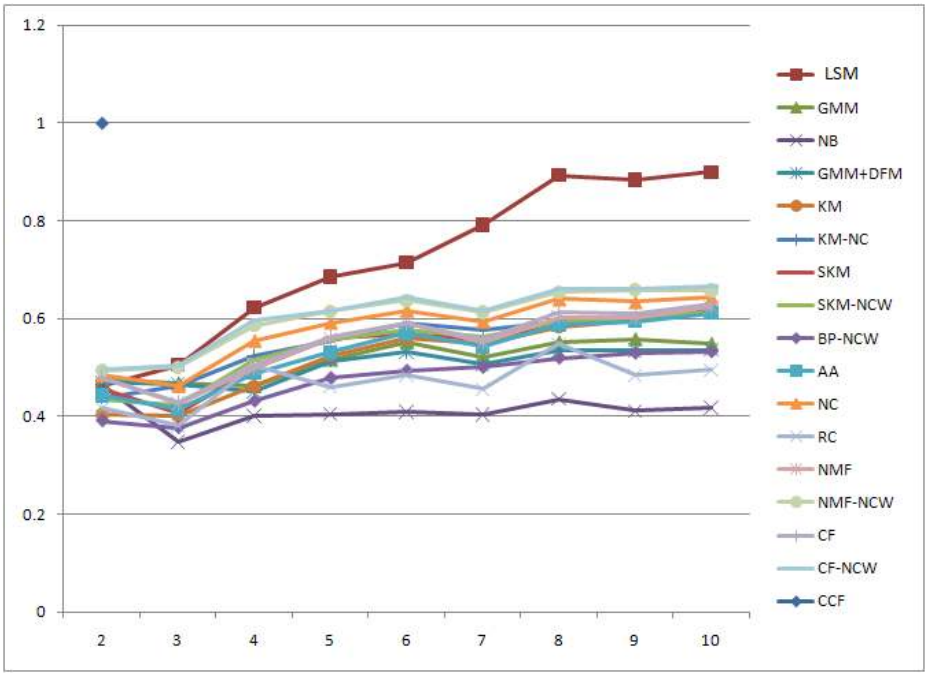

Fig. 11. The Mutual information values of 2 to 10 clusters built by LSM framework and other sixteen methods using Reuters-21578-Distribution-1 datasets[**]

The four metrics (precision, recall, overall F-measure, normalized mutual information) of LSM on Reuters-21578-Distribution-1 dataset for different k are listed in Table 4. In addition, overall F-measure is compared with FIHC and bisecting k-means as shown in Figure 12.

[**] LSM –Latent semantic manifold; GMM – Gaussian mixture model; NB – Naive Bayes clustering; GMM + DFM – Gaussian mixture model followed by the iterative cluster refinement method; KM –Traditional k-means; KM-NC – Traditional k-means and spectral clustering algorithm based on normalized cut criterion; SKM – Spherical k-means; SKM-NCW – Normalized-cut weighted form; BP-NCW – Spectral clustering based bipartite normalized cut; AA – Average association criterion; NC – Normalized cut criterion; RC – Spectral clustering based on ratio cut criterion; NMF – Non-negative matrix factorization; NMF-NCW – Nonnegative Matrix Factorization-based clustering; CF – Concept factorization; CF-NCW – Clustering by concept factorization ; CCF – k-clique community finding algorithm

Table 4. Precision, Recall, Overall F-measure, and Normalized Mutual Information (NMI) of Latent Semantic Manifold on Reuters-21578-Distribution-1 dataset

k	2	3	4	5	6	7	8	9	10
Precision	0.9845	0.9579	0.9385	0.9352	0.8909	0.9013	0.9148	0.8913	0.8859
Recall	0.7085	0.6384	0.6453	0.6056	0.5916	0.6543	0.6822	0.6688	0.6805
Overall F-measure	0.7988	0.7297	0.7399	0.6986	0.6822	0.7329	0.7562	0.7343	0.7472
NMI	0.4617	0.5051	0.6221	0.6866	0.7148	0.7925	0.8936	0.8848	0.9006

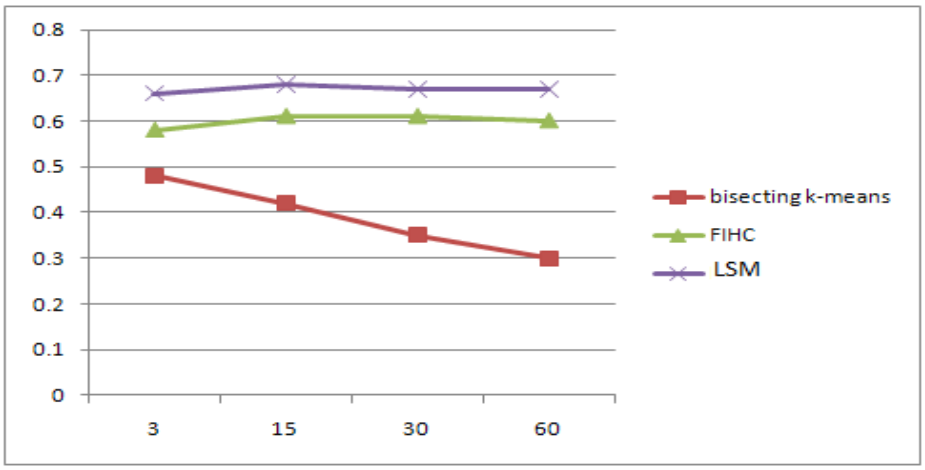

Fig. 12. The overall F-measure of three methods (LSM, FIHC, and bisecting k-means) on Reuters-21578-Distribution-1 dataset, where FIHC – Frequent itemset-based hierarchical clustering††

The average precision, recall, overall F-measure, and normalized mutual information of LSM, LST, PLSI, LDA, and CCF on Reuters-21578-Distribution-1 dataset; and LSM, LST and CCF on OHSUMED are shown in Table 5. The efficiency testing results of the three methods LSM, LST, and CCF are shown in Figure 13.

†† LSM – Latent semantic manifold; FIHC – Frequent itemset-based hierarchical clustering

Table 5. The average Precision, Recall, Overall F-measure, and Normalized Mutual Information (NMI) of LSM, LST, PLSI, PLSI + AdaBoost, LDA, and CCF on Reuters-21578-Distribution-1 dataset; and LSM, LST and CCF on OHSUMED[‡‡]

Dataset	Method	Precision	Recall	Overall F-measure	NMI
Reuters	LSM	0.81	0.773	0.786	0.717
	LST	0.779	0.745	0.754	0.633
	PLSI	0.649	0.627	0.636	0.54
	PLSI + AdaBoost	0.772	0.812	0.697	N/A
	LDA	0.66	0.714	0.686	0.61
	CCF	0.727	0.73	0.723	0.616
OHSUMED	LSM	0.59	0.479	0.522	0.315
	LST	0.586	0.388	0.456	0.257
	CCF	0.514	0.54	0.513	0.214

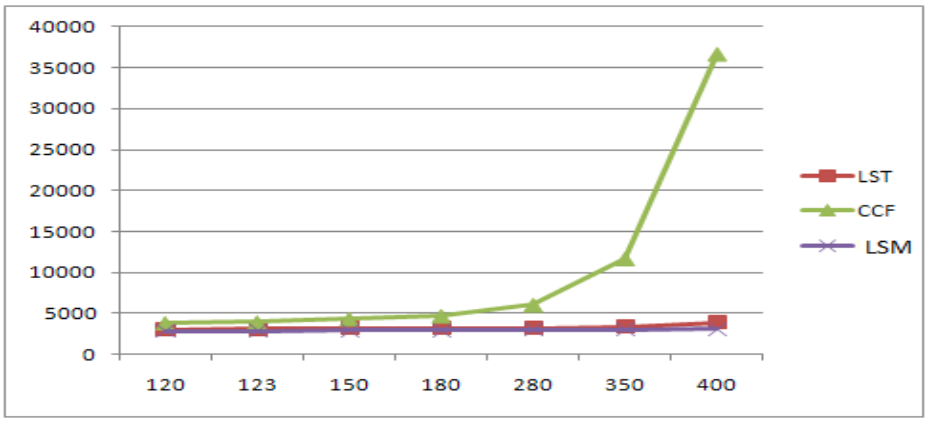

Fig. 13. The efficiency of three clustering methods, wherein x-axis identified the number of features and y-axis denoted the run time (milliseconds)[§§]

[‡‡] LSM – Latent semantic manifold; LST – Latent semantic topology; PLSI – Probabilistic latent semantic indexing; PLSI + AdaBoost – Probabilistic latent semantic indexing + additive boosting methods; LDA – Latent Dirichlet allocation; CCF – k-clique community finding algorithm

[§§] LSM – Latent semantic manifold; LST – Latent semantic topology; CCF – k-clique community finding algorithm

4 Discussion

4.1 Primary findings

Our findings suggest that the LSM framework might play an instrumental role to enhance the search engine functionalities by discovering the latent semantics in high-dimensional web data (Figure 6 -10).

4.2 Secondary findings

LSM has much better performance than the other sixteen clustering methods, especially when the number of clusters got larger on Reuters-21578-Distribution-1 and OHSUMED dataset (Table 3-4 and Figure 11-12).In general, LSM can produce more accurate results than others. The paired t-test assessed the clustering results of the same topics by any two methods - LSM, LST, and CCF. With a p-value less than 0.05, the results of LSM were significantly better than the results of LST, wherein we used 63 clusters in the experiments. Similarly, with a p-value less than 0.05, the results of LSM were significantly better than the results of the CCF in 48 randomly selected clusters out of 72, in the experiments (Table 5). The efficiency of three methods LSM, LST, and CCF with a number of features also demonstrated that LSM is better than the others. The time needed to build latent semantics does not increase significantly when the data became larger (Figure 13).

4.3 Limitation and future studies

This study had a few limitations that open up the scope of future studies. First, to identify and discriminate the correct topics in a collection of documents, the combinations of features and their co-occurring relationships serve as clues, and the probabilities display their significance. All features in documents compose a 'topological probabilistic manifold' associated with 'probabilistic measures' to denote the underlying structure. This complex structure is decomposed into inseparable components at various levels (in various levels of skeletons) so that each component corresponds to topics in a collection of documents. However, it is a time-consuming process and strongly dependent on features and their identifications (named-entities). Second, some terms with similar meanings such as 'anticipate,' 'believe,' 'estimate,' 'expect,' 'intend,' and 'project' could be separated into several independent topics; however, those topics could have a same meaning. Some data of a 'single topic' might be specified in several clusters. These issues would be considered in the further research by utilizing thesauri and some other adaptive methods [55]. Third, in this study, the evaluation was carried out mainly by comparing with other latent semantic indexing (LSI) algorithms. However, many alternative approaches for searching, clustering, and categorization exist. Further study is needed to compare this approach with alternatives. Fourth, some tools, such as GOPUBMED, ARGO, Vivisimo, also perform latent semantics search of high dimensional web data. Some further study is needed to compare LSM-based tool proposed in this study with already existing tools to find some

space for synergy. Fifth, there are some already existing knowledge bases or resources in biomedical domain, such as (Medical Subject Headings). We already perfomed some studies using Genia corpus, which contains 1,999 Medline abstracts, selected using a PubMed query for the three MeSH terms 'human,' 'blood cells,' and 'transcription factors (Figure 10).' Some more studies need to be carried to verify if this approach might be easily adapted to knowledge bases or resources.

5 Conclusion

We found that LSM framework could discover the latent semantics in high-dimensional web data and organize those into several semantic topics. This framework could be used to enhance the functionalities of currently available search engines.

6 Acknowledgement

The National Science Foundation (NSC 98-2221-E-038-012) supported this work.

Reference

1. Ranganathan, P.: The data explosion. (2011)
2. Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D.P., Kania, R., Schaeffer, M., St Pierre, S.: Big data: The future of biocuration. Nature 455, 47-50 (2008)
3. Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., McCrae, J.: Challenges for the multilingual Web of Data. Web Semantics: Science, Services and Agents on the World Wide Web 11, 63-71 (2012)
4. Croft, W.B., Metzler, D., Strohman, T.: Search engines: Information retrieval in practice. Addison-Wesley Reading (2010)
5. Thomas, P., Starlinger, J., Vowinkel, A., Arzt, S., Leser, U.: GeneView: a comprehensive semantic search engine for PubMed. Nucleic acids research 40, W585-W591 (2012)
6. Lingwal, S., Gupta, B.: A Comparative Study Of Different Approaches For Improving Search Engine Performance. International Journal (2012)
7. Freitas, A., Curry, E., Oliveira, J.G., O'Riain, S.: Querying heterogeneous datasets on the linked data web: Challenges, approaches, and trends. Internet Computing, IEEE 16, 24-33 (2012)
8. Dalal, M.K., Zaveri, M.A.: Automatic Classification of Unstructured Blog Text. Journal of Intelligent Learning Systems and Applications 5, 108-114 (2013)
9. Vercruysse, S., Kuiper, M.: Jointly creating digital abstracts: dealing with synonymy and polysemy. BMC research notes 5, 601 (2012)
10. Singer, G., Norbisrath, U., Lewandowski, D.: Ordinary search engine users carrying out complex search tasks. Journal of Information Science (2012)

11. Brossard, D., Scheufele, D.A.: Science, New Media, and the Public. Science 339, 40-41 (2013)
12. Stumme, G., Hotho, A., Berendt, B.: Semantic Web Mining: State of the art and future directions. Web Semantics: Science, Services and Agents on the World Wide Web 4, 124-143 (2006)
13. Blanco, R., Halpin, H., Herzig, D.M., Mika, P., Pound, J., Thompson, H.S., Tran, T.: Repeatable and Reliable Semantic Search Evaluation. Web Semantics: Science, Services and Agents on the World Wide Web (2013)
14. Hogan, A., Harth, A., Umbrich, J., Kinsella, S., Polleres, A., Decker, S.: Searching and browsing Linked Data with SWSE: The Semantic Web Search Engine. Web Semantics: Science, Services and Agents on the World Wide Web 9, 365-401 (2011)
15. Fazzinga, B., Gianforme, G., Gottlob, G., Lukasiewicz, T.: Semantic Web search based on ontological conjunctive queries. Web Semantics: Science, Services and Agents on the World Wide Web 9, 453-473 (2011)
16. Liu, L., Feng, J.: The Notion of "Meaning System" and its use for "Semantic Search". Journal of Computations & Modelling 1, 97-126 (2011)
17. Beall, J.: The weaknesses of full-text searching. The Journal of Academic Librarianship 34, 438-444 (2008)
18. Şah, M., Wade, V.: Automatic metadata mining from multilingual enterprise content. Web Semantics: Science, Services and Agents on the World Wide Web 11, 41-62 (2012)
19. Bergamaschi, S., Domnori, E., Guerra, F., Trillo Lado, R., Velegrakis, Y.: Keyword search over relational databases: a metadata approach. In: Proceedings of the 2011 international conference on Management of data, pp. 565-576. ACM, (Year)
20. Luhn, H.P.: The automatic creation of literature abstracts. IBM Journal of research and development 2, 159-165 (1958)
21. Salton, G., McGill, M.J.: Introduction to modern information retrieval. (1986)
22. Zipf, G.K.: {Human Behaviour and the Principle of Least-Effort}. (1949)
23. Aldous, D.: Exchangeability and related topics. École d'Été de Probabilités de Saint-Flour XIII—1983 1-198 (1985)
24. De Finetti, B.: Theory of Probability: A critical introductory treatment. Vol. 1. Wiley (1974)
25. De Finetti, B.: Theory of probability: a critical introductory treatment (1993)
26. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research 3, 993-1022 (2003)
27. Flores, J.G., Gillard, L., Ferret, O., de Chandelar, G.: Bag of senses versus bag of words: comparing semantic and lexical approaches on sentence extraction. In: TAC 2008 Workshop-Notebook papers and results, pp. 158-167. (Year)
28. Grishman, R., Sundheim, B.: Design of the MUC-6 evaluation. In: Proceedings of a workshop on held at Vienna, Virginia: May 6-8, 1996, pp. 413-422. Association for Computational Linguistics, (Year)
29. Juang, B.-H., Rabiner, L.R.: Hidden Markov models for speech recognition. Technometrics 33, 251-272 (1991)
30. Mooij, J.M., Kappen, H.J.: Sufficient conditions for convergence of the sum–product algorithm. Information Theory, IEEE Transactions on 53, 4422-4437 (2007)

31. Yedidia, J.S., Freeman, W.T., Weiss, Y.: Understanding belief propagation and its generalizations. Exploring artificial intelligence in the new millennium 8, 236-239 (2003)
32. Yedidia, J.S., Freeman, W.T., Weiss, Y.: Constructing free-energy approximations and generalized belief propagation algorithms. Information Theory, IEEE Transactions on 51, 2282-2312 (2005)
33. Borkar, V., Deshmukh, K., Sarawagi, S.: Automatic segmentation of text into structured records. In: ACM SIGMOD Record, pp. 175-186. ACM, (Year)
34. Bunescu, R., Mooney, R.J.: Relational markov networks for collective information extraction. In: ICML-2004 Workshop on Statistical Relational Learning. (Year)
35. Grimmett, G., Welsh, D.J.: Disorder in physical systems: a volume in honour of John M. Hammersley on the occasion of his 70th birthday. Oxford University Press, USA (1990)
36. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. (2001)
37. Peng, F., Feng, F., McCallum, A.: Chinese segmentation and new word detection using conditional random fields. In: Proceedings of the 20th international conference on Computational Linguistics, pp. 562. Association for Computational Linguistics, (Year)
38. Settles, B.: ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics 21, 3191-3192 (2005)
39. Taskar, B., Abbeel, P., Koller, D.: Discriminative probabilistic models for relational data. In: Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence, pp. 485-492. Morgan Kaufmann Publishers Inc., (Year)
40. Srebro, N., Jaakkola, T.: Weighted low-rank approximations. In: Machine Learning International Workshop and Conference, pp. 720. (Year)
41. Diestel, R.: Graph theory. 2005. Springer-Verlag (2005)
42. Kim, J.-D., Ohta, T., Tateisi, Y., Tsujii, J.i.: GENIA corpus—a semantically annotated corpus for bio-textmining. Bioinformatics 19, i180-i182 (2003)
43. Hersh, W., Buckley, C., Leone, T., Hickam, D.: OHSUMED: an interactive retrieval evaluation and new large test collection for research. In: SIGIR'94, pp. 192-201. Springer, (Year)
44. Xu, W., Gong, Y.: Document clustering by concept factorization. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 202-209. ACM, (Year)
45. Dalli, A.: Adaptation of the F-measure to cluster based lexicon quality evaluation. In: Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable?, pp. 51-56. Association for Computational Linguistics, (Year)
46. Kummamuru, K., Lotlikar, R., Roy, S., Singal, K., Krishnapuram, R.: A hierarchical monothetic document clustering algorithm for summarization and browsing search results. In: Proceedings of the 13th international conference on World Wide Web, pp. 658-665. ACM, (Year)
47. Fung, B.C., Wang, K., Ester, M.: Hierarchical document clustering using frequent itemsets. In: Proceedings of the SIAM international conference on data mining, pp. 59-70. (Year)
48. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: KDD workshop on text mining, pp. 525-526. Boston, (Year)

49. Cai, L., Hofmann, T.: Text categorization by boosting automatically extracted concepts. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 182-189. ACM, (Year)
50. Chiang, I.-J.: Discover the semantic topology in high-dimensional data. Expert Systems with Applications 33, 256-262 (2007)
51. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 50-57. ACM, (Year)
52. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature 435, 814-818 (2005)
53. Dhillon, I.S., Modha, D.S.: Concept decompositions for large sparse text data using clustering. Machine learning 42, 143-175 (2001)
54. Shi, J., Malik, J.: Normalized cuts and image segmentation. Pattern Analysis and Machine Intelligence, IEEE Transactions on 22, 888-905 (2000)
55. Cohen, W.W., Richman, J.: Learning to match and cluster large high-dimensional data sets for data integration. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 475-480. ACM, (Year)