Journal of Data Analysis and Information Processing, 2015, 3, 136-152
Published Online November 2015 in SciRes. http://www.scirp.org/journal/jdaip
http://dx.doi.org/10.4236/jdaip.2015.34014

Identifying Semantic in High-Dimensional Web Data Using Latent Semantic Manifold

Ajit Kumar[1], Sanjeev Maskara[2], I-Jen Chiang[3,4]

[1]Goa Institute of Management, Goa, India
[2]The Practice PLC, Buckinghamshire, UK
[3]School of Management, Taipei Medical University, Taiwan
[4]Institute of Biomedical Engineering, National Taiwan University, Taiwan
Email: ajitmaskara@gmail.com, sanjeevmaskara@gmail.com, ijchiang@ntu.edu.tw

Received 22 September 2015; accepted 16 November 2015; published 19 November 2015

Abstract

Latent Semantic Analysis involves natural language processing techniques for analyzing relationships between a set of documents and the terms they contain, by producing a set of concepts (related to the documents and terms) called semantic topics. These semantic topics assist search engine users by providing leads to the more relevant document. We develope a novel algorithm called Latent Semantic Manifold (LSM) that can identify the semantic topics in the high-dimensional web data. The LSM algorithm is established upon the concepts of topology and probability. Asearch tool is also developed using the LSM algorithm. This search tool is deployed for two years at two sites in Taiwan: 1) Taipei Medical University Library, Taipei, and 2) Biomedical Engineering Laboratory, Institute of Biomedical Engineering, National Taiwan University, Taipei. We evaluate the effectiveness and efficiency of the LSM algorithm by comparing with other contemporary algorithms. The results show that the LSM algorithm outperforms compared with others. This algorithm can be used to enhance the functionality of currently available search engines.

Keywords

Latent Semantic Manifold, Conditional Random Field, Hidden Markov Model, Graph-Based Tree-Width Decomposition

1. Introduction

In the traditional approach to data gathering, we collect data on a few well-chosen variables, and then manually perform various tasks, such as finding relevant information, analyzing them, making decisions, and so on [1].

How to cite this paper: Kumar, A., Maskara, S. and Chiang, I.-J. (2015) Identifying Semantic in High-Dimensional Web Data Using Latent Semantic Manifold. *Journal of Data Analysis and Information Processing*, **3**, 136-152.
http://dx.doi.org/10.4236/jdaip.2015.34014

However, in this high-tech era, the high volumes of data are generated with high velocity from a variety of resources (also known as 3 V—Volume, Velocity, and Variety) [2] [3]. The modern Information and Communication Technology (ICT) infrastructure, the advent of cloud computing, the cheaper availability of storage device, and the low cost computing power have made people capable of recording and storing an enormous amount of data [4]. As a result, gigantic repositories that include data, texts, and media have rapidly grown during recent years [5]-[9]. Nowadays, we create as much information in every two days as we have done since the dawn of civilization [10] [11]. Several huge repositories are freely available for the public use on the World Wide Web causing another problem—the relevant information is buried in the irrelevant ones.

To combat the problem to lose the relevant information in the overwhelming amount of data, a number of search engines have proliferated recently, which can aid users in searching contents which are relevant to them [8]. As the web pages are heterogeneous and consist of varying quality, they put limitations on search technologies [12] [13]. Moreover, the relationships among the words (polysemy, synonymy, and homophony) and sentences (paraphrase, entailment, and contradiction), and ambiguities (lexical and structural) diminish the search engines' power [14] [15]. Hence, the search engines often return inconsistent, uninteresting, and unorganized results [9] [12]. Web users have to devote substantial time and effort to differentiate meaningful items from the results returned by the search engines [9] [16] [17]. In order to facilitate and enhance relevant information access to the web users, it is essential for search engines to deal with ambiguities and imprecision [18] [19]. The need to enhance the search engines' capabilities has been felt such that the search engines can not only generate results of the web users' queried terms, but also can filter and organize meaningful items from the irrelevant ones [20]-[22].

Many effective search engines, such as MedEvi, EBIMed, MEDIE, PubNet, GoPubMed, Argo, and Vivisimo, have provided capabilities to fit search results to the users' intent. These search engines can discover latent semantic (relationships between a set of documents and the terms they contain) in the search engine generated documents and classify these documents into homogeneous semantic clusters [23]-[33]. In these search engines, each semantic cluster is considered as a topic, which indicates a summary of the generated documents. Later, the users can explore the topics that are relevant to their intent. For example, upon using these search engines, a query term, APC (Adenomatous Polyposis Coli), can yield abstracts of the relevant PubMed articles. In this case, the generated results will consist of not only abstracts 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 which are relevant to their intent (here Adenomatous Polyposis Coli) after going through the abstracts generated from the search. In summary, rather than providing huge number of web links related to the queried terms, search engines need to generate results relevant to users' intent.

In the past, many algorithms/techniques have been deployed to develop semantic search engines as described in the previous paragraph [25]. For instance, deterministic search techniques have provided metadata-enhanced search facility, where a user pre-selects different facets to generate more relevant search results [18] [19]. However, scaling metadata-enhanced search facility to the web is difficult and requires many experts to define controlled-vocabulary in order to create unique labels for the concepts having the same terminology [34] [35]. Luhn pointed out that the frequency of terms and their relative positions within a sentence in a document can be used to compute a relative measure of significance, first for the individual words and then for the sentences [36]. Word usage in a document collection tends to follow Zipf's distribution, in which a few words are used very frequently, but the vast majority only rarely [37]. Therefore, Salton and McGill addressed the $tf-idf$ scheme, which is a measure of each basic element (term) in a document collection to reveal the significance of elements within the collection [38]. For each document in the collection, the $tf-idf$ value of each term is determined by the term frequency, that is, the number of occurrences of each term in the document but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general. We may view each document as a vector with one component corresponding to each term together with a weight for each component. Thus, the $tf-idf$ scheme can reduce documents of arbitrary length to fixed-length lists of numbers. The $tf-idf$ weighting schemes are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query. In addition, $tf-idf$ can be successfully used for stop-words filtering in various subject fields including text summarization and classification. No doubt, the revolutionary change was realized in the information retrieval field with the introduction of $tf-idf$ scheme and its variants. However, in the $tf-idf$ scheme, the document collection is presented as a document-by-term matrix, which is usually enormously high-dimensional and sparse [38]-[40]. Often, for a single document, there

are more than thousands of terms in the matrix, and most of the entries are zero. The $tf-idf$ scheme can bring down some terms; yet, it provides a relatively small amount of reduction, which is not enough to reveal the statistical measures within a document or between documents.

In the last decades, some other dimension reduction techniques, such as Latent Semantic Indexing, Probabilistic Latent Semantic Indexing, and Latent Dirichlet Allocation models have been proposed to overcome the shortcomings of earlier search engines. But, all these are based on bag-of-words models. The bag-of-words models follow Aldous and de Finetti theorem of exchangeability where the order of terms in a document or order of documents in a corpus can be neglected [41]-[43]. As the spatial information conveyed by the terms in the document or documents in the corpus was highly neglected in these approaches, we found a statistical issue attached with these bags-of-words models [42]-[45]. In the 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 $F(z_1, z_2, \cdots, z_N) = F(t_1, t_2, \cdots, t_N)$ whenever $(z_1, z_2, \cdots, z_N)$ is a permutation of $(t_1, t_2, \cdots, t_N)$. However, these terms are exchangeable and the relationship between them can be established if the terms are located in proximity. For instance, we have a document describing products, such as laptops, mobile phones, and notepads. The appearance of the word "apple" can be associated with a company if it appears in proximity to words laptop, mobile phone, and notepad. However, in case, the word "apple" appears after several words or pages in the document, the relationship between "laptop, mobile phone or notepad" and "apple" weakens. Therefore, the criteria-the order of terms in a document can be neglected-should be modified to order of terms in a relationship of a document can be neglected. Likewise, 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 instance, a search term "network" would yield different topics if it occurs nearby to a term, such as computer, traffic, artificial neural, or biological neural; and hence, the order of in-relationship terms might be neglected [46].

As we can see from the literature review and our arguments that there is a need to enhance search engines' capabilities to reveal latent semantics in high-dimensional web data while preserving the relationship and order of term(s) or document(s). We proposed a novel algorithm called Latent Semantic Manifold (LSM), which identifies homogeneous groups in web data while preserving the spatial information about terms in a document or documents in the corpus. This paper aims to explain the Latent Semantic Manifold algorithm (from now on, LSM algorithm), its deployment, and performance evaluation.

2. Methods

This study consists of three key components: proposing and describing the LSM algorithm, its deployment, and evaluation. They are described in the following subsections.

2.1. Algorithm

The proposed LSM algorithm is based upon the concepts of probability and topology, which identifies the latent-semantic in data. **Figure 1** and **Table 1** provide the high-level view of the algorithm. The concepts deployed in the LSM algorithm are explained in the following four steps.

Step 1 (*Identifying relevant fragment from the user query generated documents*): A user can enter a query using a search engine, which generates a set of documents. The relevant fragments (paragraphs in the LSM) are identified from the generated documents. The identification of the fragments is handled by the "document preprocessor" of the search engine, which typically normalizes the document stream to a predefined format, breaks the document stream into desired retrievable unit, and isolates and metatags subdocument pieces.

Step 2 (*Recognizing named-entity and constructing heterogeneous manifold*): It is crucial to extract significant "terms" from the fragments (identified in Step 1) to construct heterogeneous manifolds. Notably, we can extract various types of terms with a large number of training documents. However, extracting different types of terms and calculating their marginal and conditional probabilities is highly computation-intensive [47]-[51]. Therefore, we stick to identifying nouns (words or phrases) or named-entities in the LSM framework. Hidden Markov Models (HMMs) are often used for part-of-speech tagging and sequential labeling [52] [53]. Yet, in the last decade, discriminative linear chain Conditional Random Field (CRF) models have been used for tagging and sequential labeling of features in the corpus because of its advantages over the HMMs [54]-[56]. The primary advantage of CRFs over HMMs is their conditional nature. A CRF is a simple framework for labeling and segmenting data that

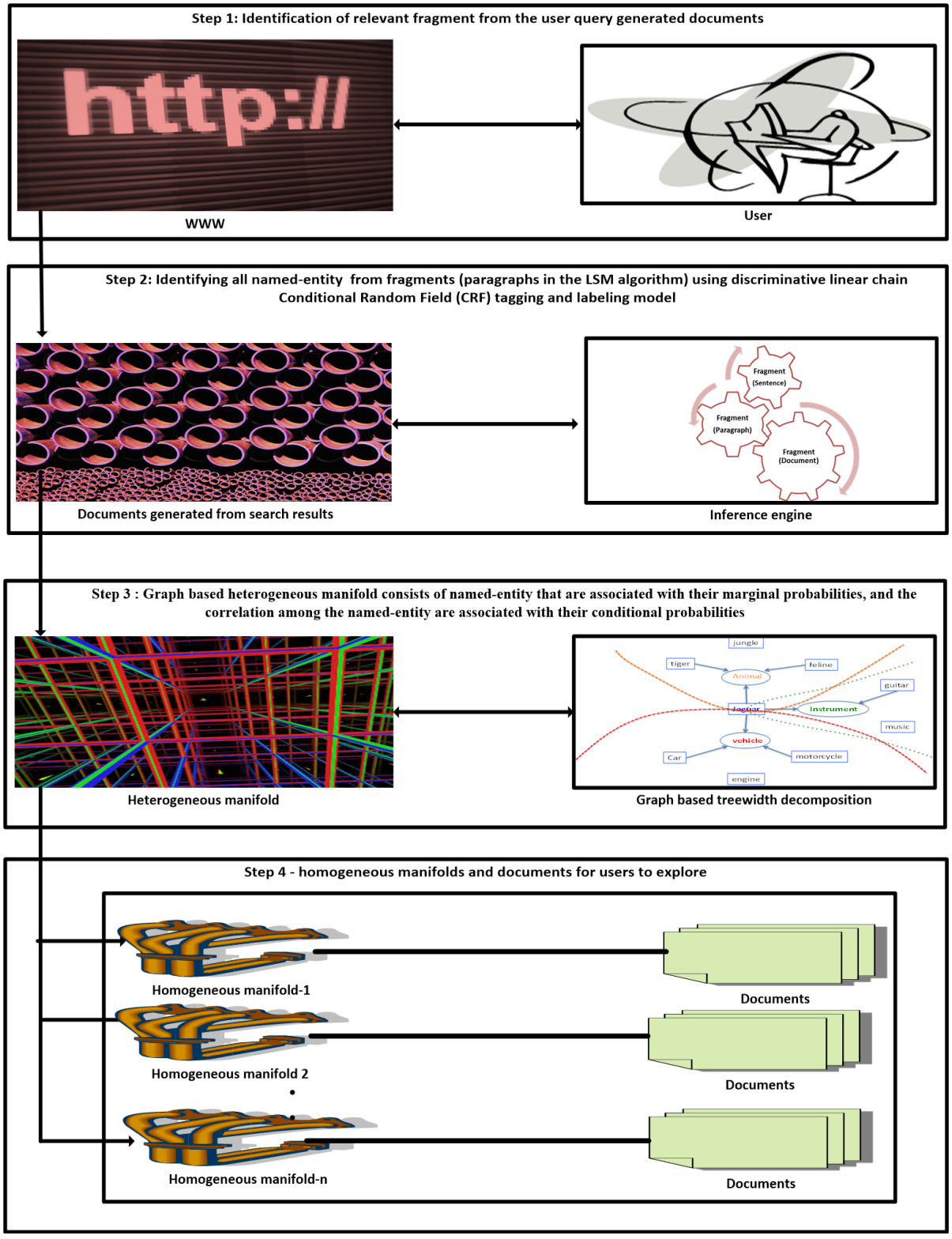

Figure 1. Illustration of LSM algorithm.

models a conditional distribution P($z|x$) by selecting the label sequence z, a named category, to label a novel observation sequence x with an associated undirected graph structure that obeys the Markov property. When

Table 1. LSM algorithm that construct semantic manifold.

Algorithm		
Require: A collection of returned documents from a search query. **Ensure:** A collection of semantic manifolds.		
Step 1	Perform feature extractions using discriminative linear chain Conditional Random Field method to generate named entities.	
Step 2	Construct a manifold from the set of named entities generated from the document collection. Classify the manifold into isomorphic (homogeneous) categories by using the Graph-based Tree-width Decomposition algorithm starting from a fixed dimension local manifold. **Require:** $V=\{t_1,t_2,\cdots,t_N\}$ is the vertex set of named entities that each t_i is associated with its named categories equipped with a weighted probability. **Ensure:** $M=\{M_1,M_2,\cdots,M_n\}$ is the set of isomorphic semantic manifolds. where $M_i=\{M_{ij}\mid \text{No } M_{ij} \text{ is a subset of } M_{ik}, j\neq k\}$.	
Step 3	**Step 3.1**	Let a semantic topic set: $C=\{Z_1,Z_2,\cdots,Z_m\}$. Let G = (V, E) be the undirected connected graph generated from the returned documents.
	Step 3.2	Given a tree-width d, find a semantic manifold M_j generated from single named entities for each semantic category z_i initially in which $\lvert M_j\rvert$ = d and the semantic mapping $f(M_{ij})\in C$ with a probability $P(M_{ij},z_k)\in[0,1]$, and quantity $f(M_{ij})=z_k$
	Step 3.3	Perform graph decompositions on G starting from M_j.

conditioned on the observations that are given in a particular observation sequence, the CRF defines a single log-linear distribution over the labeled sequence. The CRF model does not need explicitly to present the dependencies of input variables *x* affording the use of rich and global features of the input, thus allows relaxation of the strong independence assumptions required by HMMs in order to ensure tractable inference. The relationships among these named-entities construct a complex structure called a heterogeneous manifold.

The named-entities are indicated with their marginal probabilities, and the correlations among named-entities are indicated with their conditional probabilities. For example, the jaguar is considered as a named-entity, and it is assigned to the animal or vehicle type depending on the overall context of the fragment. The named-entities are indicated with their marginal probabilities, and the correlations among the named-entities are indicated with their conditional probabilities. As illustrated 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). Likewise, it has conditional probabilities, such as P (Jaguar, Car|Vehicle), P (Jaguar, Motorcycle|Vehicle).

Step 3 (*Decomposing a heterogeneous manifold into homogeneous manifolds*): As mentioned in Step 2, the heterogeneous manifold consists of a complex structure of named-entities, including estimates of marginal and conditional probabilities. A collection of fragment vectors lies on the heterogeneous manifolds, 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 *Rn*. In addition, all the points in the local spaces are strongly connected. As the heterogeneous manifold is overly complex, and the semantic is latent in local spaces; thus, instead of retaining just one heterogeneous manifold, we 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 Tree-width Decomposition algorithm is used to decompose a heterogeneous manifold into a collection of homogeneous manifolds [57]. As shown in **Figure 3**, assuming Jaguar as the heterogeneous manifold, we can decompose it into three homogeneous manifolds bounded by dotted lines in three different colors. In the Graph-based Tree-width Decomposition algorithm, we start selecting a random fixed dimension local manifold to be a separator as shown in **Figure 4** [58]. Afterward, the local manifold is decomposed into two local manifolds that are not adjacent. This decomposition is recursive until no further decomposition is possible. We can express the above concept formally,

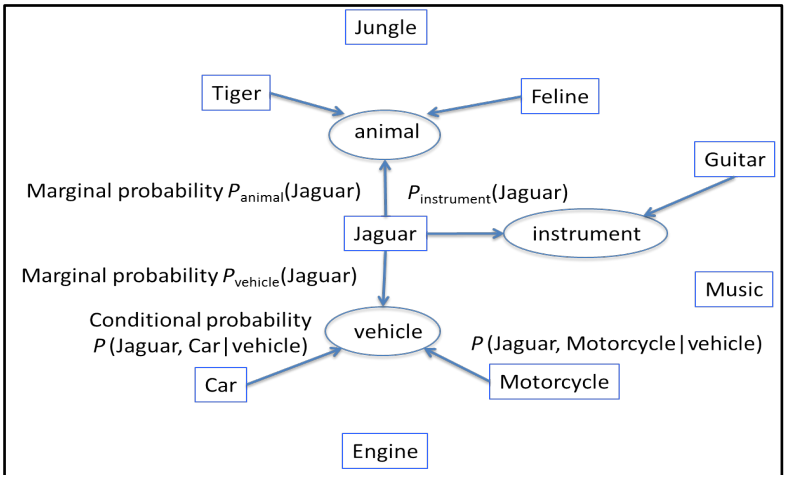

Figure 2. An example to demonstrate named-entities, its types, and associated marginal and conditional probabilities.

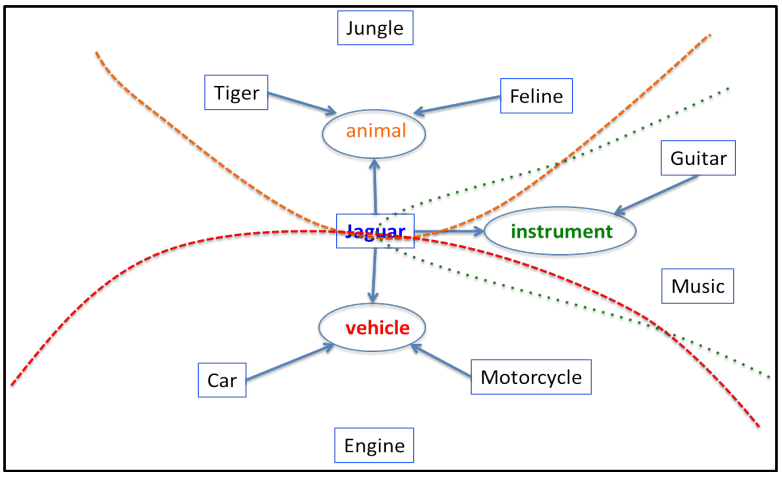

Figure 3. An example to demonstrate Graph-based Tree-width Decomposition.

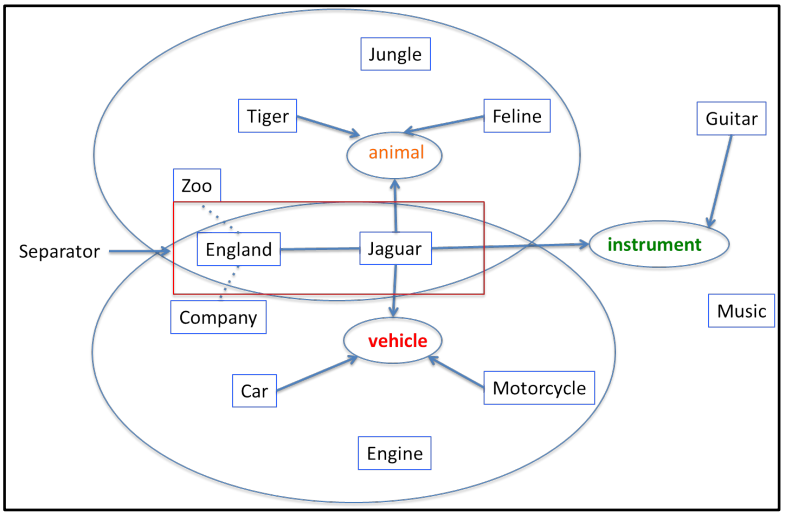

Figure 4. An example to demonstrate the concept of separator under Graph-based Tree-width decomposition.

let a heterogeneous manifold M_i for fragmenti be the set of homogeneous manifolds, such that $M_i = \{M_{ij} | \text{No } M_{ij} \text{ is a subset of } M_{ik}, j \neq k\}$. The semantics generated from fragment homogeneous manifolds are independent. In addition, a semantic topic set $C = \{Z_1, Z_2, \cdots, Z_m\}$. of the returned documents is associated with 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 the number of documents that are relevant to a homogeneous manifold 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 homogeneous manifold associated fragments, which they want.

Step 4 (*Exploring the homogeneous manifolds*): The relevant fragments cluster around their related homogeneous manifolds. For instance, a user query for the term APC, the fragments have aggregated into a collection of homogeneous manifolds as shown in **Figure 5**. Each fragment is assigned to a particular homogeneous manifold.

2.2. Deployment of the LSM Algorithm

The LSM algorithm was deployed to develop a search tool. A team of three researchers including an expert in the Java programming language developed the tool using the Eclipse Software Development Kit. The LSM tool was used for two years at two places in Taiwan: 1) Taipei Medical University Library, Taipei; and 2) Biomedical Engineering Laboratory, Institute of Biomedical Engineering, National Taiwan University, Taipei. The members of the library and lab used the LSM tool to perform semantic searches in the PubMed database.

2.3. Performance Evaluation of the LSM Algorithm

Data sets: Two data sets, Reuters-21578-Distribution-1 and OHSUMED, were used to evaluate the performance

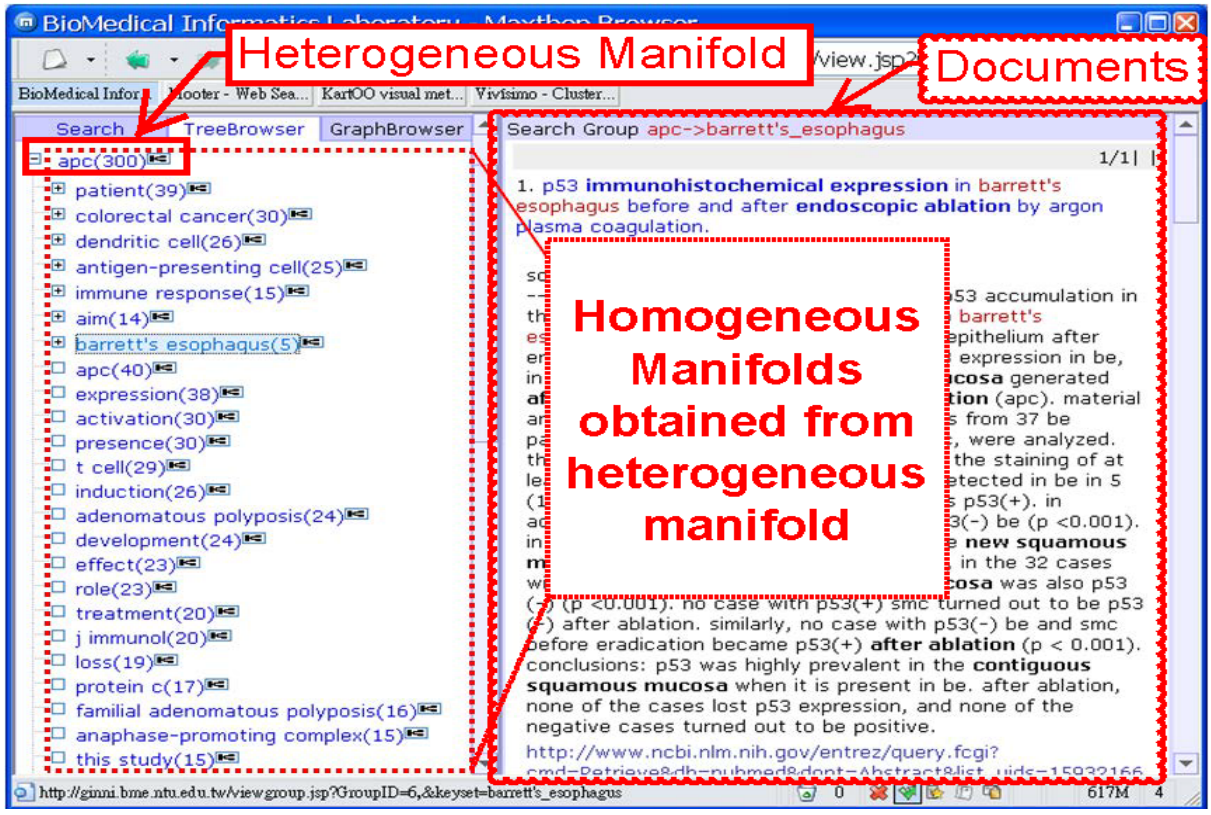

Figure 5. An example to demonstrate heterogeneous manifold, homogeneous manifolds and documents associated with homogeneous manifolds.

of the LSM algorithm. The Reuters-21578-Distribution-1 is a standard benchmark for the text categorization, which consists of Newswire articles classified into 135 topics [59]. In our tests, the documents with multiple topics (category labels) and single topic were separated. The topics that had less than five documents were removed. **Table 2** shows the summary of the Reuters-21578-Distribution-1 collection. OHSUMED is clinically oriented a Medline collection consisting of 348,566 references. It covers all the references from 270 medical journals belonging to 23 disease categories over a five-year period (1987-1991) [60].

Evaluation criteria: Effectiveness and efficiency were measured as an experimental evaluation of the LSM algorithm. Effectiveness is defined as the ability to identify the right cluster (collection of documents). As shown in **Table 3**, the generated clusters were verified by human experts to measure the effectiveness. The three measures of the effectiveness (Precision, Recall, and F_β) were calculated using the contingency in **Table 3**. Precision and Recall are respectively defined as follows:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}$$

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i}$$

Moreover, F_β measure, which combines Precision and Recall, is defined as follows:

$$F_\beta = \frac{\left(\beta^2 + 1\right) \times \text{Precision}_i \times \text{Recall}_i}{\beta^2 \times \text{Precision}_i + \text{Recall}_i}$$

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 in 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 (NMI) and overall *F*-measure, were also used [61]-[63]. Given the two sets of topics *C* and Cl, let *C* denote the topic set defined by experts, and Cl denotes the topic set generated by a clustering method, and both derived from the same corpora *X*. Let *N*(*X*) denote 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 metric *MI*(*C*, *C*′) is defined as:

$$MI\left(C, C'\right) = \sum_{z \in C, z' \in C'} P\left(Z, Z'\right) \text{Log}_2 \left(\frac{P\left(Z, Z'\right)}{P\left(Z\right) P\left(Z'\right)} \right)$$

$$\text{where } P\left(z\right) = \frac{N\left(z, X\right)}{N\left(X\right)}, \; P\left(z'\right) = \frac{N\left(z', X\right)}{N\left(X\right)}, \text{ and } P\left(z, z'\right) = \frac{N\left(Z, Z', X\right)}{N\left(X\right)}$$

Table 2. Statistics of reuters-21,578 corpora.

Statistics	Number of topics	Number of documents	Documents on a topic
Origin	135	21,578	0 - 3945
Single topic	65	8649	1 - 3945
Single topic (≥5 documents)	51	9494	5 - 3945

Table 3. Contingency table for category (c_i, where i = natural number)[a].

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

[a]TP: True Positive; FP: False Positive; FN: False Negative; TN: True Negative.

The Normalized Mutual Information metric *MI*(*C*, *C*′) will return a value between zero and $\max\left(H\left(C\right),H\left(C'\right)\right)$, where *H*(*C*) and *H*(*C*′) define the entropies of *C* and *C*′ respectively. A higher $MI\left(C,C'\right)$ value means that two topics are almost identical, whereas a lower value indicates the independence of topics. Therefore, the Normalized Mutual Information metric $MI\left(C,C'\right)$ is

$$MI\left(C,C'\right)=\frac{MI\left(C,C'\right)}{\max\left(H\left(C\right),H\left(C'\right)\right)}$$

Let F_i be an *F*-measure for each cluster Z_i defined above. The overall F-measure can be defined as

$$F^{*}=\sum_{z'\in C'}P\left(z'\right)\times\max_{z\in C}F\left(Z,Z'\right)$$

where *F*(*z*, *z*′) calculates the *F*-measure between *z* and *z*′.

Efficiency is the clustering time for a search query with a fixed number of features for each clustering scheme, where features set is fixed.

Experiments: The experiments were conducted using Reuters-21578-Distribution-1 and OHSUMED data sets. The clusters ranging from two to ten topics were randomly selected to evaluate the LSM with other clustering methods. For each clustering method, each test run was conducted on a selected topic, and Normalized Mutual Information of the topic and its corresponding cluster was calculated. After conducting fifty test runs on a fixed number of *k*'s topics, where $2\le k\le 10$, the final performance scores were obtained by averaging mutual information measures from these 50 test runs [61]. 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 **Table 4** and **Figure 6** in the result section. We also calculated the overall F-measure in combination of arbitrary *k* clusters by uniquely assigning to topics from the Reuters-21578-Distribution-1 data set where *k* was 3, 15, 30, and 60 [64]. Fifty test-runs were also performed using these LSM results to compare Frequent Itemset-based Hierarchical Clustering (FIHC) and bisecting *k*-means as shown **Table 5** and **Figure 7** in the Result section [64] [65].

Table 4. Normalized Mutual Information comparison of LSM algorithm with other sixteen methods using Reuters-21578-Distribution-1 dataset[b].

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

[b]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-NCL 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.

Table 5. 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

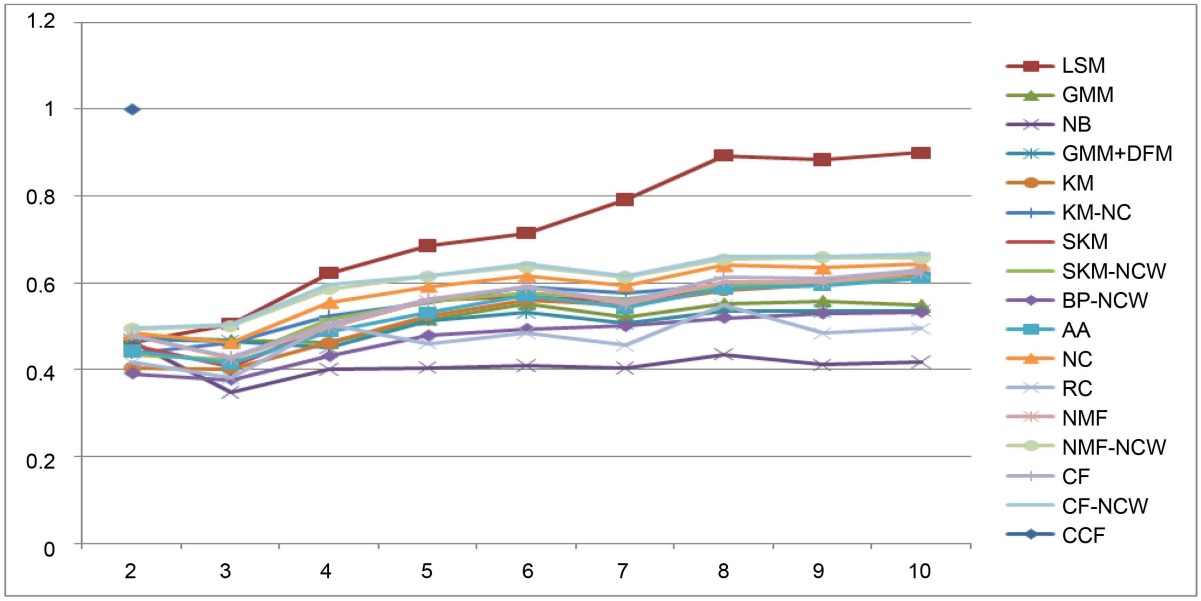

Figure 6. Mutual information values of 2 to 10 clusters built by LSM algorithm and other sixteen methods using Reuters-21578-Distribution-1 datasets[c]. c. LSM: Latent semantic manifold; GMM-Gaussian mixture model; NB: Naive Bayes clustering; GMM + DFM: Gaussian mixture model followed by the iterative 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.

The average precision, recall, overall F-measure, and Normalized Mutual Information of LSM, LST, PLSI, PLSI + AdaBoost, LDA, and CCF were evaluated using the Reuters-21578-Distribution-1 data set; and LSM, LST, and CCF were evaluated on an OHSUMED data set, as shown in **Table 6**, in the Result section [44] [66]-[69]. Besides the effectiveness, the efficiency tests of LSM, LST, and CCF were performed as shown in **Figure 8** in the Result section.

3. Results

Normalized Mutual Information comparison of the LSM algorithm with the other sixteen methods using Reuters-21578-Distribution-1 data setis shown in **Table 4** and **Figure 6** [61] [69]-[71]. The four metrics (precision, recall, overall *F*-measure, and Normalized Mutual Information) of LSM that used Reuters-21578-Distribution-1 data set for different *k* are listed in **Table 5**. In addition, the overall *F*-measure is compared with FIHC and bisecting *k*-means as shown in **Figure 7**. The average precision, recall, overall *F*-measure, and Normalized Mutual Information of 1) LSM, LST, PLSI, LDA, and CCF, which used Reuters-21578-Distribution-1 data set; 2) LSM, LST and CCF, which used OHSUMED data set are shown in **Table 6**. The efficiency tests results of the three methods, LSM, LST, and CCF, are shown in **Figure 8**.

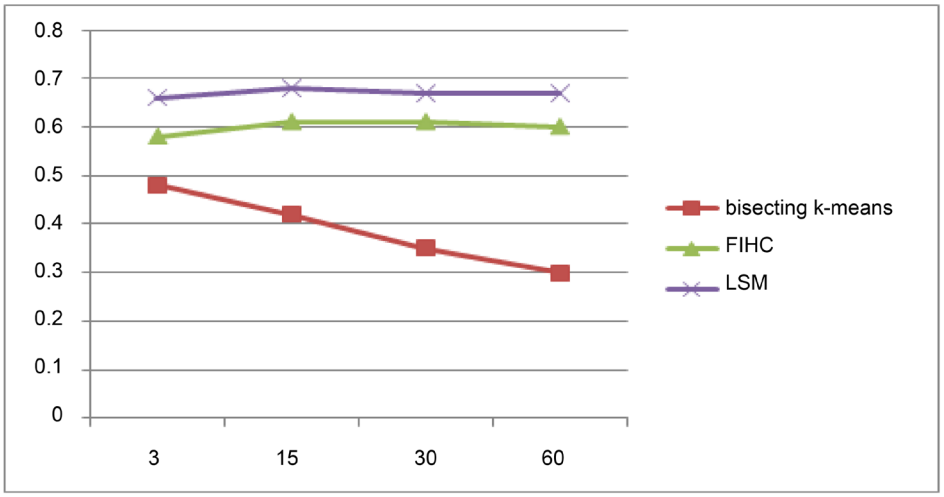

Figure 7. Overall *F*-measure of three methods, LSM, FIHC, and bisecting *k*-means, on Reuters-21578-Distribution-1 data set, where k (*x*-axis) is 3, 15, 30, 60 clusters.

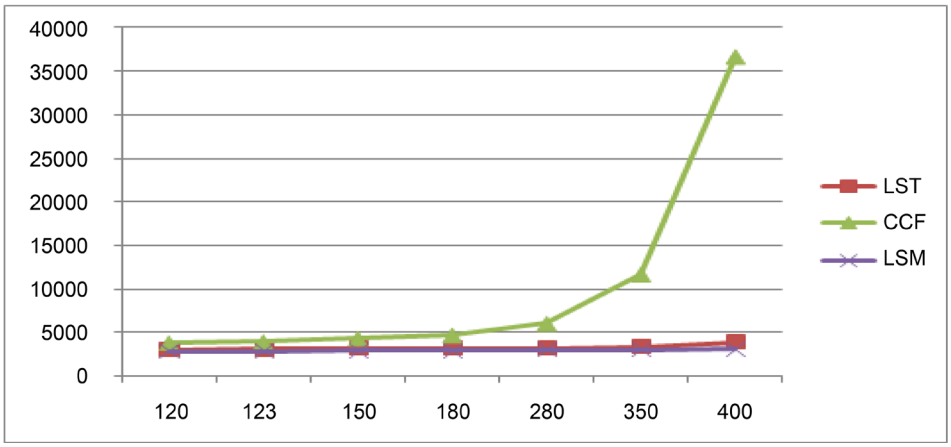

Figure 8. Efficiency of three clustering methods, wherein *x*-axis is the number of features and *y*-axis is run time in milliseconds (LSM: Latent semantic manifold; LST: Latent Semantic Topology; CCF: *k*-Clique Community Finding).

Table 6. 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[d].

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

[d]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.

4. Discussion

Our findings suggest that the LSM algorithm, which can discover the latent semantics in high-dimensional web data, might play an instrumental role in enhancing the search engine functionality. LSM carries out searches based on both keywords and meaning, which can assist researchers to perform semantic searches on databases. For example, a researcher can search APC with Adenomatous Polyposis Coli as his or her intended meaning in the PubMed database (the output of a user queried term APC is shown in **Figure 9**).

APC can also have other meanings, such as Antigen-Presenting Cells, Anaphase Promoting Complex, or Activated Protein C. Suppose, in a homogeneous manifold, we find APC, Colorectal Cancer, and gene-related documents are assembled, the homogeneous manifoldwould point out the meaning of APC as Adenomatous Polyposis Gene. Similarly, suppose APC, Major Histocompatibility Complex, and T-cells-related documents are assembled, it would indicate the meaning of APC as Antigen Presenting Cells. **Figure 9** shows that documents returned from the queried term APC can automatically associate to various homogeneous manifolds (semantic topics). In addition, the researcher can obtain a different vantage point based on the underlying data. For example, a search for the medical term NOD2 that was performed within the PubMed database retrieved almost 300 abstracts of published or in-press articles (**Figure 10** 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, Crohn's disease, and ulcerative colitis. Some evolving topics, such as the bacterial component were also discovered. However, the result was different when we searched NOD2 on Genia Corpus (**Figure 11**) which supports the argument the researcher can obtain a different meaningful vantage point based on the underlying data, using the "same" LSM algorithm [72].

We can see that results (**Figure 10** and **Figure 11**) are meaningfully structured with a possibility of semantic navigation in both databases. This indicates that the generalization capability of the LSM algorithm. We used concepts of topology in designing LSM algorithm. LSM has shown much better performance than the other sixteen clustering methods, especially when the number of clusters gets larger (**Table 4** and **Table 5**, and **Figure 6** and **Figure 7**). In general, we found that LSM could produce more accurate results than others could. We used paired t-test to assess the clustering results of the same topics by any two methods-LSM, LST, and CCF. The results of LSM were significantly better than the results of LST where we used 63 clusters in the experiments

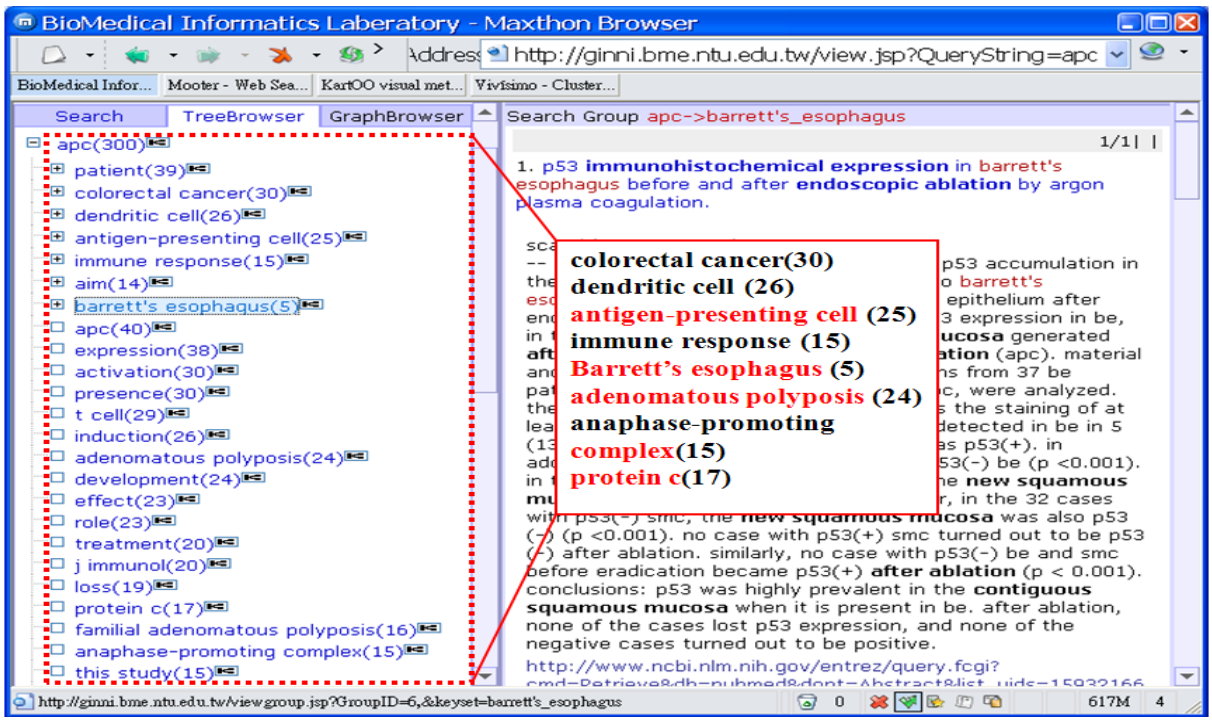

Figure 9. Result of query term, APC.

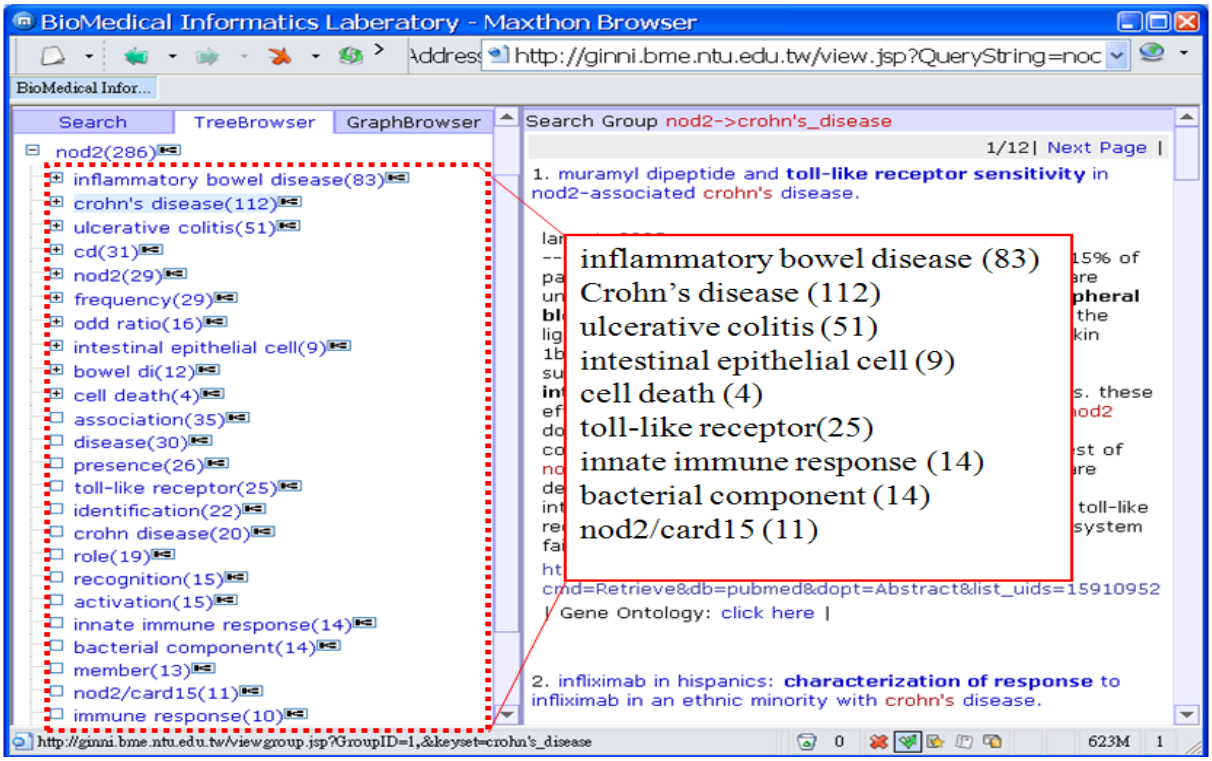

Figure 10. Clustering result of the query term, NOD2, retrieved from PubMed.

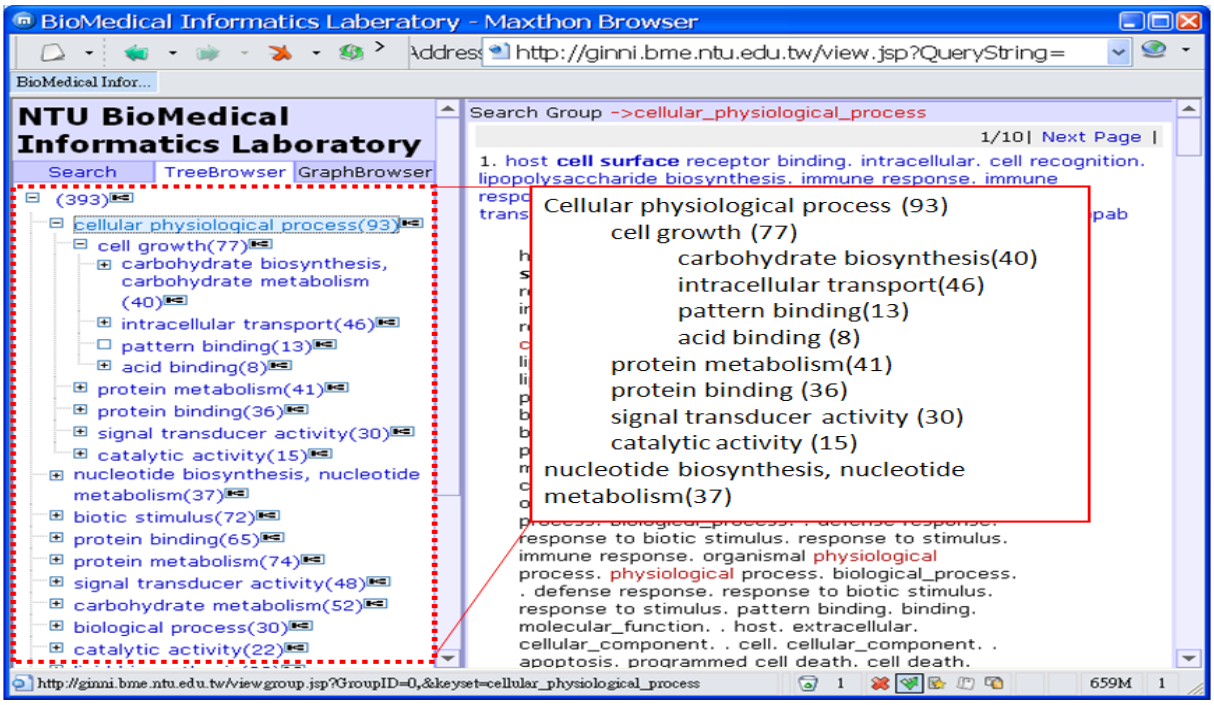

Figure 11. Clustering result of the query term, NOD2, retrieved from Genia Corpus.

(p-value < 0.05) (**Table 6**). 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 (**Table 6**). The efficiency evaluation of three methods, LSM, LST, and CCF, demonstrated that LSM performed better than the others did. In the case of

LSM, the time needed to build a latent semantic manifold does not increase significantly when the data became larger (**Figure 8**).

Limitation and future studies: This study has a few limitations that open up the scope of future studies. First, to identify and discriminate the correct topics in the collection of documents, a combination of features and their co-occurring relationships serve as clues, and probabilities display their significance. All features in documents comprise a topological probabilistic manifold, associate to probabilistic measures, and 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 the collection of documents. This process is a computation-intensive and time-consuming, which strongly depend on features and their identifications (named-entities). Second, some terms with similar meanings, such as anticipate, believe, estimate, expect, intend, and project, were separated into independent topics. Likewise, some terms were repeatedly specified in many topics. These issues might be addressed by utilizing thesauri and some other adaptive methods [73]. Third, some tools, such as MedEvi, EBIMed, MEDIE, PubNet, GoPubMed, Argo, and Vivisimo, also perform a latent semantic search in high dimensional web data [23]-[33]. However, in this study, we did not compare LSM algorithm based tool with others. Some further study is needed to compare the LSM algorithm based tool with already existing tools to find a space of synergy. Fourth, 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. Fifth, there are some already existing knowledge bases or resources in the biomedical domain, such as MeSH (Medical Subject Headings) [74] [75]. Some studies are needed to verify whether LSM algorithm based tool might be adapted to the existing knowledge bases or resources.

5. Conclusion

We found that the LSM algorithm can discover the latent semantics in high-dimensional web data and can organize them into several semantic topics. This algorithm can be used to enhance the functionality of currently available search engines.

Acknowledgements

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

References

[1] Donoho, D.L. (2000) High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality. AMS Math Challenges Lecture, 1-32. http://mlo.cs.manchester.ac.uk/resources/Curses.pdf

[2] Laney, D. (2001) 3D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note, 6.

[3] Hoehndorf, R., Rebholz-Schuhmann, D., Haendel, M. and Stevens, R. (2014) Thematic Series on Biomedical Ontologies in JBMS: Challenges and New Directions. *Journal of Biomedical Semantics*, **5**, 15. http://dx.doi.org/10.1186/2041-1480-5-15

[4] Raman, A.C. (2014) Storage Infrastructure for Big Data and Cloud. *Handbook of Research on Cloud Infrastructures for Big Data Analytics*, 110. http://dx.doi.org/10.4018/978-1-4666-5864-6.ch005

[5] Ranganathan, P. (2011) From Microprocessors to Nanostores: Rethinking Data-Centric Systems. *Computer*, **44**, 39-48.

[6] Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W. and Rhee, S.Y. (2008) Big Data: The Future of Biocuration. *Nature*, **455**, 47-50. http://dx.doi.org/10.1038/455047a

[7] Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P. and McCrae, J. (2012) Challenges for the Multilingual Web of Data. *Web Semantics*: *Science, Services and Agents on the World Wide Web*, **11**, 63-71. http://dx.doi.org/10.1016/j.websem.2011.09.001

[8] Croft, W.B., Metzler, D. and Strohman, T. (2010) Search Engines: Information Retrieval in Practice. Addison-Wesley, Reading, 88.

[9] Thomas, P., Starlinger, J., Vowinkel, A., Arzt, S. and Leser, U. (2012) Gene View: A Comprehensive Semantic Search Engine for PubMed. *Nucleic Acids Research*, **40**, W585-W591. http://dx.doi.org/10.1093/nar/gks563

[10] Every 2 Days We Create As Much Information As We Did up to 2003. http://techcrunch.com/2010/08/04/schmidt-data/

[11] Mayer-Schönberger, V. and Cukier, K. (2013) Big Data: A Revolution That Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt, Boston.

[12] Lingwal, S. and Gupta, B. (2012) A Comparative Study of Different Approaches for Improving Search Engine Performance. *International Journal of Emerging Trends & Technology in Computer Science*, **1**, 123-132.

[13] Freitas, A., Curry, E., Oliveira, J.G. and Riain, S.O. (2012) Querying Heterogeneous Datasets on the Linked Data Web: Challenges, Approaches, and Trends. *IEEE Internet Computing*, **16**, 24-33. http://dx.doi.org/10.1109/MIC.2011.141

[14] Dalal, M.K. and Zaveri, M.A. (2013) Automatic Classification of Unstructured Blog Text.

[15] Vercruysse, S. and Kuiper, M. (2012) Jointly Creating Digital Abstracts: Dealing with Synonymy and Polysemy. *BMC Research Notes*, **5**, 601. http://dx.doi.org/10.1186/1756-0500-5-601

[16] Singer, G., Norbisrath, U. and Lewandowski, D. (2012) Ordinary Search Engine Users Carrying out Complex Search Tasks. *Journal of Information Science*, **39**, 346-358.

[17] Brossard, D. and Scheufele, D.A. (2013) Science, New Media, and the Public. *Science*, **339**, 40-41. http://dx.doi.org/10.1126/science.1232329

[18] Beall, J. (2008) The Weaknesses of Full-Text Searching. *The Journal of Academic Librarianship*, **34**, 438-444. http://dx.doi.org/10.1016/j.acalib.2008.06.007

[19] Liu, L. and Feng, J. (2011) The Notion of "Meaning System" and Its Use for "Semantic Search". *Journal of Computations and Modelling*, **1**, 97-126.

[20] Stumme, G., Hotho, A. and Berendt, B. (2006) Semantic Web Mining: State of the Art and Future Directions. *Web Semantics*: *Science, Services and Agents on the World Wide Web*, **4**, 124-143. http://dx.doi.org/10.1016/j.websem.2006.02.001

[21] Blanco, R., Halpin, H., Herzig, D.M., Mika, P., Pound, J., Thompson, H.S. and Tran, T. (2013) Repeatable and Reliable Semantic Search Evaluation. *Web Semantics*: *Science, Services and Agents on the World Wide Web*, **21**, 14-29. http://dx.doi.org/10.1016/j.websem.2013.05.005

[22] Nessah, D. and Kazar, O. (2013) An Improved Semantic Information Searching Scheme Based Multi-Agent System and an Innovative Similarity Measure. *International Journal of Metadata, Semantics and Ontologies*, **8**, 282-297. http://dx.doi.org/10.1504/IJMSO.2013.058411

[23] Hogan, A., Harth, A., Umbrich, J., Kinsella, S., Polleres, A. and Decker, S. (2011) 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. http://dx.doi.org/10.1016/j.websem.2011.06.004

[24] Fazzinga, B., Gianforme, G., Gottlob, G. and Lukasiewicz, T. (2011) Semantic Web Search Based on Ontological Conjunctive Queries. *Web Semantics*: *Science, Services and Agents on the World Wide Web*, **9**, 453-473. http://dx.doi.org/10.1016/j.websem.2011.08.003

[25] Lu, Z.Y. (2011) PubMed and Beyond: A Survey of Web Tools for Searching Biomedical Literature. *Database*, **2011**, baq036. http://dx.doi.org/10.1093/database/baq036

[26] Kim, J.J., Pezik, P. and Rebholz-Schuhmann, D. (2008) MedEvi: Retrieving Textual Evidence of Relations between Biomedical Concepts from Medline. *Bioinformatics*, **24**, 1410-1412. http://dx.doi.org/10.1093/bioinformatics/btn117

[27] Rebholz-Schuhmann, D., Kirsch, H., Arregui, M., Gaudan, S., Riethoven, M. and Stoehr, P. (2007) EBIMed—Text Crunching to Gather Facts for Proteins from Medline. *Bioinformatics*, **23**, e237-e244. http://dx.doi.org/10.1093/bioinformatics/btl302

[28] Ohta, T., Tsuruoka, Y., Takeuchi, J., Kim, J.D., Miyao, Y., Yakushiji, A., *et al*. (2006) An Intelligent Search Engine and GUI-Based Efficient MEDLINE Search Tool Based on Deep Syntactic Parsing. *Proceedings of the COLING/ACL on Interactive Presentation Sessions*, Sydney, 17-21 July 2006, Association for Computational Linguistics, 17-20.

[29] Douglas, S.M., Montelione, G.T. and Gerstein, M. (2005) PubNet: A Flexible System for Visualizing Literature Derived Networks. *Genome Biology*, **6**, R80. http://dx.doi.org/10.1186/gb-2005-6-9-r80

[30] Doms, A. and Schroeder, M. (2005) GoPubMed: Exploring PubMed with the Gene Ontology. *Nucleic Acids Research*, **33**, W783-W786. http://dx.doi.org/10.1093/nar/gki470

[31] Argo: Genome Browser. http://www.broadinstitute.org/annotation/argo

[32] Engels, R., Yu, T., Burge, C., Mesirov, J.P., DeCaprio, D. and Galagan, J.E. (2006) Combo: A Whole Genome Comparative Browser. *Bioinformatics*, **22**, 1782-1783. http://dx.doi.org/10.1093/bioinformatics/btl193

[33] Koshman, S., Spink, A. and Jansen, B.J. (2006) Web Searching on the Vivisimo Search Engine. *Journal of the American Society for Information Science and Technology*, **57**, 1875-1887. http://dx.doi.org/10.1002/asi.20408

[34] Sah, M. and Wade, V. (2012) Automatic Metadata Mining from Multilingual Enterprise Content. *Web Semantics*: *Science, Services and Agents on the World Wide Web*, **11**, 41-62. http://dx.doi.org/10.1016/j.websem.2011.11.001

[35] Bergamaschi, S., Domnori, E., Guerra, F., TrilloLado, R. and Velegrakis, Y. (2011) Keyword Search Over Relational Databases: A Metadata Approach. *Proceedings of the* 2011 *ACM SIGMOD International Conference on Management of Data*, ACM, New York, 565-576. http://dx.doi.org/10.1145/1989323.1989383

[36] Luhn, H.P. (1958) The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, **2**, 159-165. http://dx.doi.org/10.1147/rd.22.0159

[37] Zipf, G.K. (1949) Human Behavior and the Principle of Least Effort.

[38] Salton, G. and McGill, M.J. (1983) Introduction to Modern Information Retrieval. McGraw-Hill Book Co., New York.

[39] Kupiec, J., Pedersen, J. and Chen, F. (1995) A Trainable Document Summarizer. *Proceedings of the* 18*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, New York, 68-73. http://dx.doi.org/10.1145/215206.215333

[40] Gabaix, X. (1999) Zipf's Law for Cities: An Explanation. *Quarterly Journal of Economics*, **114**, 739-767. http://dx.doi.org/10.1162/003355399556133

[41] Aldous, D.J. (1985) Exchangeability and Related Topics. Springer, Berlin, 1-198. http://dx.doi.org/10.1007/bfb0099421

[42] Warmuth, W. (1977) De Finetti, B.: Theory of Probability—A Critical Introductory Treatment, Volume 2. John Wiley and Sons, London-New York-Sydney-Toronto 1975. XIV, 375 S. *Biometrical Journal*, **19**, 382. http://dx.doi.org/10.1002/bimj.4710190515

[43] Reinhardt, H.E. (1978) Theory of Probability: A Critical Introductory Treatment, Vol. 2 (Bruno de Finetti). *SIAM Review*, **20**, 200-201. http://dx.doi.org/10.1137/1020030

[44] Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003) Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, **3**, 993-1022.

[45] Flores, J.G., Gillard, L., Ferret, O. and de Chandelar, G. (2008) Bag of Senses versus Bag of Words: Comparing Semantic and Lexical Approaches on Sentence Extraction. *TAC* 2008 *Workshop-Notebook Papers and Results*, Gaithersburg, 17-19 November 2008, 158-167.

[46] Chanlekha, H. and Collier, N. (2010) Analysis of Syntactic and Semantic Features for Fine-Grained Event-Spatial Understanding in Outbreak News Reports. *Journal of Biomedical Semantics*, **1**, 3. http://dx.doi.org/10.1186/2041-1480-1-3

[47] Juang, B.H. and Rabiner, L.R. (1991) Hidden Markov Models for Speech Recognition. *Technometrics*, **33**, 251-272. http://dx.doi.org/10.1080/00401706.1991.10484833

[48] Mooij, J.M. and Kappen, H.J. (2007) Sufficient Conditions for Convergence of the Sum-Product Algorithm. *IEEE Transactions on Information Theory*, **53**, 4422-4437. http://dx.doi.org/10.1109/TIT.2007.909166

[49] Yedidia, J.S., Freeman, W.T. and Weiss, Y. (2003) Understanding Belief Propagation and Its Generalizations. *Exploring Artificial Intelligence in the New Millennium*, **8**, 236-239.

[50] Yedidia, J.S., Freeman, W.T. and Weiss, Y. (2005) Constructing Free-Energy Approximations and Generalized Belief Propagation Algorithms. *IEEE Transactions on Information Theory*, **51**, 2282-2312. http://dx.doi.org/10.1109/TIT.2005.850085

[51] Wagholikar, K.B., Torii, M., Jonnalagadda, S. and Liu, H. (2013) Pooling Annotated Corpora for Clinical Concept Extraction. *Journal of Biomedical Semantics*, **4**, 3. http://dx.doi.org/10.1186/2041-1480-4-3

[52] Baum, L.E. and Petrie, T. (1966) Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, **37**, 1554-1563. http://dx.doi.org/10.1214/aoms/1177699147

[53] Rabiner, L.R. (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, **77**, 257-286. http://dx.doi.org/10.1109/5.18626

[54] Sutton, C. and McCallum, A. (2011) An Introduction to Conditional Random Fields. *Machine Learning*, **4**, 267-373. http://dx.doi.org/10.1561/2200000013

[55] Lafferty, J., McCallum, A. and Pereira, F.C. (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.

[56] Wallach, H.M. (2004) Conditional Random Fields: An Introduction. Technical Reports (CIS), 22.

[57] Srebro, N. and Jaakkola, T. (2003) Weighted Low-Rank Approximations. *Proceedings of the* 20*th International Conference on Machine Learning, ICML* 2003, **3**, 720-727.

[58] Diestel, R. (2005) Graph Theory. Springer-Verlag, New York.

[59] Rose, T., Stevenson, M. and Whitehead, M. (2002) The Reuters Corpus Volume 1—From Yesterday's News to Tomorrow's Language Resources. *Proceedings of the* 3*rd International Conference on Language Resources and Evaluation, LREC* 2002, **2**, 827-832.

[60] Hersh, W., Buckley, C., Leone, T.J. and Hickam, D. (1994) OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research. In: Croft, B.W. and van Rijsbergen, C.J., Eds., *SIGIR*'94, Springer, London, 192-201. http://dx.doi.org/10.1007/978-1-4471-2099-5_20

[61] Xu, W. and Gong, Y. (2004) Document Clustering by Concept Factorization. *Proceedings of the* 27*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, New York, 202-209. http://dx.doi.org/10.1145/1008992.1009029

[62] Dalli, A. (2003) Adaptation of the F-Measure to Cluster Based Lexicon Quality Evaluation. *Proceedings of the EACL* 2003 *Workshop on Evaluation Initiatives in Natural Language Processing*: *Are Evaluation Methods, Metrics and Resources Reusable*? Association for Computational Linguistics, Stroudsburg, 51-56.

[63] Kummamuru, K., Lotlikar, R., Roy, S., Singal, K. and Krishnapuram, R. (2004) A Hierarchical Monothetic Document Clustering Algorithm for Summarization and Browsing Search Results. *Proceedings of the* 13*th International Conference on World Wide Web*, ACM, New York, 658-665. http://dx.doi.org/10.1145/988672.988762

[64] Fung, B.C., Wang, K. and Ester, M. (2003) Hierarchical Document Clustering Using Frequent Itemsets. *Proceedings of the* 2003 *SIAM International Conference on Data Mining*, **3**, 59-70. http://dx.doi.org/10.1137/1.9781611972733.6

[65] Steinbach, M., Karypis, G. and Kumar, V. (2000) A Comparison of Document Clustering Techniques. *KDD Workshop on Text Mining*, **400**, 525-526.

[66] Cai, L. and Hofmann, T. (2003) Text Categorization by Boosting Automatically Extracted Concepts. *Proceedings of the* 26*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, New York, 182-189. http://dx.doi.org/10.1145/860435.860470

[67] Chiang, I.J. (2007) Discover the Semantic Topology in High-Dimensional Data. *Expert Systems with Applications*, **33**, 256-262. http://dx.doi.org/10.1016/j.eswa.2006.05.033

[68] Hofmann, T. (1999) Probabilistic Latent Semantic Indexing. *Proceedings of the* 22*nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, New York, 50-57. http://dx.doi.org/10.1145/312624.312649

[69] Palla, G., Derényi, I., Farkas, I. and Vicsek, T. (2005) Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society. *Nature*, **435**, 814-818. http://dx.doi.org/10.1038/nature03607

[70] Dhillon, I.S. and Modha, D.S. (2001) Concept Decompositions for Large Sparse Text Data Using Clustering. *Machine learning*, **42**, 143-175. http://dx.doi.org/10.1023/A:1007612920971

[71] Shi, J. and Malik, J. (2000) Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, 888-905. http://dx.doi.org/10.1109/34.868688

[72] Kim, J.D., Ohta, T., Tateisi, Y. and Tsujii, J.I. (2003) GENIA Corpus—A Semantically Annotated Corpus for Bio-Textmining. *Bioinformatics*, **19**, i180-i182. http://dx.doi.org/10.1093/bioinformatics/btg1023

[73] Cohen, W.W. and Richman, J. (2002) Learning to Match and Cluster Large High-Dimensional Data Sets for Data Integration. *Proceedings of the* 8*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, 475-480. http://dx.doi.org/10.1145/775047.775116

[74] Lipscomb, C.E. (2000) Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association*, **88**, 265-266.

[75] Lowe, H.J. and Barnett, G.O. (1994) Understanding and Using the Medical Subject Headings (MeSH) Vocabulary to Perform Literature Searches. *JAMA*, **271**, 1103-1108. http://dx.doi.org/10.1001/jama.1994.03510380059038