Pacific Symposium on Biocomputing 7:386-397 (2002)

A THEMATIC ANALYSIS OF THE AIDS LITERATURE

W. JOHN WILBUR
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, U.S.A.

Faced with the need for human comprehension of any large collection of objects, a time honored approach has been to cluster the objects into groups of closely related objects. Individual groups are then summarized in some convenient manner to provide a more manageable view of the data. Such methods have been applied to document collections with mixed results. If a hard clustering of the data into mutually exclusive clusters is performed then documents are frequently forced into one cluster when they may contain important information that would also appropriately make them candidates for other clusters. If a soft clustering is used there still remains the problem of how to provide a useful summary of the data in a cluster. Here we introduce a new algorithm to produce a soft clustering of document collections that is based on the concept of a theme. A theme is conceptually a subject area that is discussed by multiple documents in the database. A theme has two potential representations that may be viewed as dual to each other. First it is represented by the set of documents that discuss the subject or theme and second it is also represented by the set of key terms that are typically used to discuss the theme. Our algorithm is an EM algorithm in which the term representation and the document representation are explicit components and each is used to refine the other in an alternating fashion. Upon convergence the term representation provides a natural summary of the document representation (the cluster). We describe how to optimize the themes produced by this process and give the results of applying the method to a database of over fifty thousand PubMed documents dealing with the subject of AIDS. How themes may improve access to a document collection is also discussed.

1 Introduction

There are at least two reasons for interest in clustering a set of documents. One is to improve retrieval efficiency and the other is to improve human understanding of the data in the collection. The first of these goals proved elusive historically because the quality of the retrieval degraded due to the clustering.[1, 2] With the much greater speed and memory of current computers the interest in clustering for efficiency has waned. However, the need for improved human understanding of large data sets has reached critical proportions with the advent of the Internet as well as the many large databases of documents that are now becoming available in different specialty areas.

Improved human understanding of data through clustering may consist of graphical aids in visualizing the data[3-5] as well as methods of examining textual summaries of cluster content. Given a predefined set of clusters there are many machine learning methods for inducing representations for the clusters.[6-9] These methods play an important role in the human comprehension of information from a variety of points of view. However, our interest is somewhat different from these in that we desire to find a rich representation of the topics or themes that occur in a database and view document clusters as a means to this end. For example, in the AIDS data that we study here "blood transfusion" is an important theme. This theme is described by a rich terminology and it is this terminology that we refer to in using the word theme. There is also the cluster of documents in the database that discuss this theme and this cluster plays an essential role in discovering the theme, but the theme and the cluster are treated as of equal importance each helping to define the other.

A number of methods have been proposed whereby topical groupings of terms are derived from a document collection in an effort to improve the representation of the documents. Such methods may or may not involve a clustering of the documents, but they are of interest since they address the problem of theme generation that is our interest. There are information bottleneck methods,[10, 11] probabilistic latent semantic indexing,[12] and mixture models.[13-15] The bottleneck approach produces term groupings that maximize the information relative to the document collection. This groups terms with a high mutual co-occurrence, but seems unsuited to produce the natural themes that occur in text. Theme generation as we conceive it will be almost certain to reduce the information relative to the documents because of the large number of terms that are grouped together in a theme. We believe this is simply the wrong paradigm for theme generation. Probabilistic latent semantic indexing and most mixture models assume in principle that a document arises from a single source even if that source is not determined. Again this is theoretically unsuitable for our purposes as a document does not arise from a single theme. Rather a document often contains multiple themes. The one approach in the literature that seems theoretically most consistent with our goal is the Multiple Cause Mixture Model.[15] While this approach solves the one document-multiple theme problem, it along with the other methods mentioned here has another unfortunate property. It requires that all terms that occur in documents must be forced into some topical word group even if they are function words, etc.

Aside from the theoretical problems mentioned here there is the practical problem that previous methods require the whole database to be processed before any result is obtained. This is very computationally expensive and out of reach for large collections (even the AIDS data we study here). Our approach is unique in that it produces one theme and one document cluster at a time. Because of this simplicity our method is readily applied to very large collections (even millions of documents). Indeed a very large collection may yet take a long time for complete analysis, but one can produce useable collections of themes without a complete analysis. The approach we present here is related to earlier work,[16, 17] but the model is different in its explicit treatment of themes, is simpler, and allows a much more efficient computation.

2. Preliminaries

Let D be a database of documents and let the set of all index terms that appear in at least one of the members of D be denoted by T. Let R denote the "occurs in" relationship between elements of T and D. Then it is customary to represent R as a subset of the product set of T and D, i.e., $R \subseteq T \times D$. The set R is known as a relation and if $(t,d) \in R$ we say t occurs in d and may also write tRd. If $U \subseteq T$ and $V \subseteq D$ it is standard usage to define

$$\begin{aligned} R[U] &= \{d \in D \mid \exists t \in U \ni (t,d) \in R\} \\ R^{-1}[V] &= \{t \in T \mid \exists d \in V \ni (t,d) \in R\}. \end{aligned} \tag{2.1}$$

For a single point t we write $R[t] = R[\{t\}]$ and for a single point d likewise $R^{-1}[d] = R^{-1}[\{d\}]$.

By a theme we mean a particular subject area that is discussed by some subset of the documents in the database D. Interestingly such a subject area is generally also characterized by a particular subset of index terms T that are used to describe that subject area. Intuitively then a theme means nonempty sets $U \subseteq T$ and $V \subseteq D$ with the property that all the elements of U have a high probability of occurring in all the elements of V. We require not only that this be true, but that it be true in some optimal sense which we will make explicit.

3. The Theme Generation Algorithm

In order to apply the EM algorithm we will follow the notation of Little and Rubin.[18] Our description will be in terms of the sets U and V which we have used to outline the concept of a theme. There is observed and missing data.

$$\begin{aligned} Y_{obs} &= R \\ Y_{miss} &= \{z_d\}_{d \in D} \end{aligned} \tag{3.1}$$

The observed data is the relation R. The missing data is a set of indicator variables that are defined by

$$z_d = \begin{cases} 1, \ d \in V \\ 0, \ d \notin V. \end{cases} \tag{3.2}$$

The parameters are

$$\Theta = U\left(\|U\| = n_U\right), \ \{p_t, q_t\}_{t \in U}, \ \{r_t\}_{t \in T}. \tag{3.3}$$

Here n_U is a constant positive integer and the size of the set U (number of elements). For any $t \in U$, p_t is the probability that for any $d \in V$, tRd, and q_t is the probability that for any $d \in D - V$, tRd . For any $t \in T$, r_t is the probability that for any $d \in D$, tRd . Constants in the process in addition to the integer n_U are the set of prior probabilities $\{pr_d\}_{d \in D}$ that are the prior probabilities that the elements d belong to V .

In order to develop the EM algorithm approach we will need to make an independence assumption about the statistical properties of R . This kind of assumption is common in many contexts for the purpose of facilitating the mathematical analysis of complicated data:

Independence Assumption. Within $T \times V \cap R$ all the atomic events tRd are independent of each other and likewise for $T \times (D - V) \cap R$.

Finally to facilitate the writing of mathematical formulas we will use the indicator variables $\{u_t\}_{t \in T}$ defined by

$$u_t = \begin{cases} 1, \ \mathrm{t} \in U \\ 0, \ \mathrm{t} \notin U \end{cases} \qquad (3.4)$$

and the delta notation

$$\delta_{td} = \begin{cases} 1, & tRd \\ 0, & \neg tRd. \end{cases} \qquad (3.5)$$

We must work with the quantity

$$p\left(R,\{z_d\} \mid \Theta\right) = p\left(R \mid \{z_d\},\Theta\right) p\left(\{z_d\} \mid \Theta\right). \qquad (3.6)$$

Computing from the right side we obtain

$$p\left(\{z_d\} \mid \Theta\right) = \prod\nolimits_{d \in D} pr_d^{z_d} \left(1 - pr_d\right)^{1 - z_d} \qquad (3.7)$$

$$p\left(R \mid \{z_d\},\Theta\right) =$$

$$\prod\nolimits_{t,d} \left\{ \left[\left(p_t^{\delta_{td}} \left(1 - p_t\right)^{1-\delta_{td}} \right)^{u_t} \left(q_t^{\delta_{td}} \left(1 - q_t\right)^{1-\delta_{td}} \right)^{1-u_t} \right]^{z_d} \left(r_t^{\delta_{td}} \left(1 - r_t\right)^{1-\delta_{td}} \right)^{1-z_d} \right\} \qquad (3.8)$$

It is next necessary to take the expectation of the log of (3.6) over the distribution $p\left(\{z_d\} \mid R,\Theta\right)$. In this we may ignore (3.7) because it will yield a constant and have no influence on the subsequent maximization. Thus we may compute

$$\begin{aligned} E\left(ln\, P\left(R \mid \{z_d\},\Theta\right)\right) = & \sum\nolimits_t u_t \sum\nolimits_d pz_d \left(\delta_{td}\, ln\, p_t + \left(1 - \delta_{td}\right) ln\left(1 - p_t\right)\right) + \\ & \sum\nolimits_t u_t \sum\nolimits_d \left(1 - pz_d\right)\left(\delta_{td}\, ln\, q_t + \left(1 - \delta_{td}\right) ln\left(1 - q_t\right)\right) + \\ & \sum\nolimits_t \left(1 - u_t\right) \sum\nolimits_d \delta_{td}\, ln\, r_t + \left(1 - \delta_{td}\right) ln\left(1 - r_t\right) \end{aligned}$$

(3.9)

In order to complete this calculation it is necessary to compute pz_d based on R and Θ . This we do from Bayes' theorem

$$pz_d = p\left(z_d = 1 \mid R,\Theta\right) = p\left(z_d = 1 \mid \left\{\delta_{td}\right\}_{t\in T},\Theta\right)$$
$$= \frac{p\left(\left\{\delta_{td}\right\}_{t\in T} \mid z_d = 1,\Theta\right) pr_d}{p\left(\left\{\delta_{td}\right\}_{t\in T} \mid z_d = 1,\Theta\right) pr_d + p\left(\left\{\delta_{td}\right\}_{t\in T} \mid z_d = 0,\Theta\right)\left(1 - pr_d\right)} \qquad (3.10)$$

Individual probabilities on the right side are given by

$$p\left(\left\{\delta_{td}\right\}_{t\in T} \mid z_d = 1,\Theta\right) = \prod\nolimits_t \left(p_t^{\delta_{td}}\left(1-p_t\right)^{1-\delta_{td}}\right)^{u_t}\left(r_t^{\delta_{td}}\left(1-r_t\right)^{1-\delta_{td}}\right)^{1-u_t}$$
$$p\left(\left\{\delta_{td}\right\}_{t\in T} \mid z_d = 0,\Theta\right) = \prod\nolimits_t \left(q_t^{\delta_{td}}\left(1-q_t\right)^{1-\delta_{td}}\right)^{u_t}\left(r_t^{\delta_{td}}\left(1-r_t\right)^{1-\delta_{td}}\right)^{1-u_t} \qquad (3.11)$$

Because of the common factor in these expressions it is convenient to write

$$pz_d = \frac{1}{1 + exp\left(-score_d + C\right)} \qquad (3.12)$$

where

$$C = \sum\nolimits_{t\in U} ln\left(\frac{1-p_t}{1-q_t}\right)$$
$$score_d = \sum\nolimits_{t\in U} \delta_{td}\, ln\left(\frac{p_t\left(1-q_t\right)}{q_t\left(1-p_t\right)}\right) + ln\left(\frac{pr_d}{1-pr_d}\right). \qquad (3.13)$$

The final step is to carry out the maximization of (3.9) over Θ. Too accomplish this we note that we may begin by choosing the values of p_t, q_t, and r_t so that the individual sums on the right in (3.9) are maximal if in the case of p_t and q_t, $u_t = 1$ and if in the case of r_t, $u_t = 0$. This is straightforward and yields

$$p_t = \sum\nolimits_d \delta_{td} pz_d \Big/ \sum\nolimits_d pz_d$$
$$q_t = \sum\nolimits_d \delta_{td}\left(1 - pz_d\right) \Big/ \sum\nolimits_d \left(1 - pz_d\right) \qquad (3.14)$$
$$r_t = n_t / N$$

Here we have defined

$$n_t = \left\|R\left[t\right]\right\| = \sum\nolimits_d \delta_{td}$$
$$N = \left\|D\right\|. \qquad (3.15)$$

Now for each t we define a quantity which is the difference between the contribution coming from t in the sum (3.9) depending on whether $u_t = 1$ or $u_t = 0$.

$$\begin{aligned}\alpha_t = n_{st}\,ln\left(\frac{p_t}{q_t}\right) + \left(n_s - n_{st}\right)ln\left(\frac{1-p_t}{1-q_t}\right) + \\ \left(n_t - n_{st}\right)ln\left(\frac{q_t}{r_t}\right) + \left(N - n_t - n_s + n_{st}\right)ln\left(\frac{1-q_t}{1-r_t}\right)\end{aligned} \qquad (3.16)$$

In addition to (3.15) we here employ the definitions

$$\begin{aligned} n_s &= \sum\nolimits_d pz_d \\ n_{st} &= \sum\nolimits_d \delta_{td}\, pz_d . \end{aligned} \qquad (3.17)$$

The maximization is completed by choosing the n_U largest α_t's and setting $u_t = 1$ for each of them and $u_t = 0$ for all others. If there is ambiguity due to equal α_t's choices are made arbitrarily to obtain the number n_U.

4. A Practical Algorithm

Our interest is in a practical algorithm for applications. With that objective we will outline here our approach first as a series of steps and then give more detail in how to begin the computation and how to control it.
Input: *R*, the number n_U, and the set of prior probabilities $\{pr_d\}_{d\in D}$.
Step 1: Compute the probabilities $\{pz_d\}_{d\in D}$ through the use of (3.13).
Step 2: Compute p_t, q_t, and r_t, all $t \in T$ from (3.14) and (3.15).
Step 3: Compute the α_t, all $t \in T$ from (3.16) and (3.17).
Step 4: Select the n_U points $t \in T$ for which α_t is the greatest to define the set U and the indicator values $\{u_t\}_{t\in T}$.
Step 5: Test for convergence and if not converged return to Step 1.
By examining the steps listed it is evident that if we can obtain the probabilities $\{pz_d\}_{d\in D}$ in Step 1, the remaining steps are relatively straightforward to perform (we will discuss convergence below). As a general approach we have found it quite satisfactory to restrict the values pz_d to either 0 or 1. This is simply accomplished by setting a cutoff value and using (3.13) to compute

$$pz_d = \begin{cases} 1, \ score_d > cutoff \\ 0, \ score_d \le cutoff \end{cases}. \qquad (4.1)$$

In practice we find that the number of *d*'s for which $pz_d = 1$ can have a large variation from one iteration to the next if we use a fixed cutoff. We have found improved stability by defining an integer we term the *stringency*. The stringency must be positive and not greater than n_U. We then set

$$cutoff = \frac{stringency}{n_U}\sum\nolimits_{t\in U} ln\left(\frac{p_t\left(1-q_t\right)}{q_t\left(1-p_t\right)}\right). \qquad (4.2)$$

When (4.1) and (4.2) are used to implement Step 1 we will refer to the result as the *binary* form of the algorithm. The algorithm begins by assigning the values pz_d to be 0 or 1 depending on some preliminary guess as to what *V* might be. This allows the first iteration through Steps 1-5. On the second and all subsequent iterations the equations (4.1) and (4.2) are used in Step 1. In Step 5 convergence is tested by observing when all quantities become fixed. In practice this is easily ascertained by observing when the value of *C* in (3.13) takes the same identical value on successive iterations.

Control of the algorithm is important in that it generally has the potential to converge to a local maximum in many different ways. Such control could be exerted through the choice of the values $\{pr_d\}_{d\in D}$. However our approach is generally to set these values all to 0.5 so that they have no influence in (3.13). We only exert control by the initial choice of the $\{pz_d\}_{d\in D}$ as binary values reflecting some estimate of *V*. Occasionally this is not satisfactory and we wish to force the algorithm to converge with certain d included in *V*. Then we set the values of the corresponding pr_d close to 1 or equivalently the values of $ln\left(pr_d/\left(1-pr_d\right)\right)$ large so that these particular d become locked into *V*.

5. Focusing a Theme

The binary form of the theme generation algorithm described in the foregoing works well in that it is successful in producing a large number of different themes on a database. The difficulty is that one must decide on the value n_U prior to generating a theme and this may not be optimal. If it is too small it will not allow the full theme to develop and if it is too large it will allow extraneous material to be pulled in to be part of the theme. In order to deal with this problem we have developed a method of focusing a theme to the optimal size. The method works by starting with a value of n_U that is too small and running the algorithm to stability or near stability and then increasing the size of n_U by a small amount to $n_{U'}$ and again running the algorithm close to stability. Let U and U' denote the two term sets corresponding to the two themes obtained at these two successive points. At each such step we check two things. First, are the two themes close together? To measure closeness let α_t represent the value from (3.16) corresponding to a $t\in U$ and likewise α'_t for $t\in U'$. We may then define a Dice coefficient of similarity between the two themes by

$$Dice\left(U,U'\right)=\left(\sum\nolimits_{t\in U\cap U'}\alpha_t+\alpha'_t\right)/\left(\sum\nolimits_{t\in U}\alpha_t+\sum\nolimits_{t\in U'}\alpha'_t\right). \qquad (5.1)$$

We require that

$$Dice\left(U,U'\right)>0.9 \qquad (5.2)$$

at each successive increment of n_U to $n_{U'}$ This is a continuity condition that is necessary because during expansion a theme my become unstable and suddenly in a single step metamorphose into a completely different theme or into one that is only distantly related to the theme from the previous step. If such a sudden change in the theme takes place we halt the process of focusing at the previous step. Our second concern is that the theme actually improves at each step. In order to measure improvement we require a fixed integer smaller than the number of terms in the theme. We will call this number the *focal size* of the theme and denote it by f. Then we define the *focus* of a theme U to be the average of the f largest α_t, $t \in U$. We denote the focus of a theme U by $\bar{\alpha}(U,f)$. Then we consider a step in focusing to be an improvement provided

$$\bar{\alpha}(U,f) \le \bar{\alpha}(U',f) . \qquad (5.3)$$

Thus if the process of focusing the theme does not end because of a violation of (5.2) it will eventually end because of a violation of (5.3) when we have reached at least a local maximum in the focus possible for that theme.

6. Themes from the AIDS data

In February of 2001 we extracted all documents in PubMed that had assigned the MeSH® term "Human Immunodeficiency Syndrome". This comprised 52,970 documents consisting of title, abstract (when present), and MeSH terms. This set is the database D for the thematic analysis presented here. We used as the index term set T all MeSH terms (with and without qualifiers assigned in the documents and with and without stars) as well as terms from the titles and abstracts. Title and abstract were broken into single word and two word terms and any term containing a stop word was discarded. No stemming was performed and no punctuation is allowed in the terms.

Our first step was to generate a set of themes with stringency 10 and n_U equal to 30. These parameter settings tend to give undersize themes. We attempted to generate such an initial theme for each document in D. For each $d \in D$ we used a vector document retrieval algorithm to obtain the 100 documents, $\{d_i\}_{i=1}^{100}$, in D most similar to d. We then set pz_d to 1 for all the $\{d_i\}_{i=1}^{100}$ and 0 for all other documents. With this initialization we then attempt to generate a theme. We succeeded in generating a theme in 42,395 of the 52,970 attempts. In some cases the set of documents $\{d_i\}_{i=1}^{100}$ has insufficient similarity within the set to produce a theme. When duplicates were removed the 42,395 themes resulted in a set of 7,311 unique themes. These themes provided the seeds for the focusing process described in the previous section.

For the focusing process we chose a focal size f of 10. Beginning with each of the 7,311 seed themes we carried out the focusing process and obtained a set of 5,236 unique focused themes. In a significant number of cases different seed themes produced the same focused theme. While the 5,236 themes are unique, there are many pairs of themes that are closely related to each other. We processed all pairs of the 5,236 themes and marked a pair (U,U') as equivalent if they satisfied

$$Dice(U,U') \geq 0.9 \ . \qquad (6.1)$$

We generated the equivalence relation based on the marked pairs and chose one of the largest themes from each class. This yielded a set of 1164 unique themes with a certain distance between any two themes in the set. These 1164 themes provide a picture of the AIDS literature in PubMed.

While one can view each of the 1164 themes, this is still a relatively large number of themes to examine. In order to facilitate browsing the data we performed single link clustering of the 1164 themes with different thresholds according to

$$Dice(U,U') \geq threshold \qquad (6.2)$$

to obtain clusters of themes that could be examined by a human. We performed the clustering at five levels beyond the baseline of 0.9 and obtained the numbers of clusters shown in Table 1.

Table 1. An analysis of the 1164 themes by single link clustering. As the threshold decreases there are fewer clusters of larger size.

Level	Threshold	Clusters
1	0.9	1164
2	0.8	772
3	0.7	477
4	0.6	287
5	0.5	171
6	0.4	92

We have developed a web interface to allow browsing of the 1164 themes. At level 1 this allows access to the individual themes. At higher levels one views the individual themes grouped into clusters and may still view an individual theme or may select a cluster and view its summary or differences between its members.

7. An Example theme

Here we give a partial listing of the term set for the theme developed on pneumocystis pneumonia in AIDS. This theme has 70 terms associated with it. This is of intermediate size. While some themes have only 30 terms many have over 70 and a few over 200 terms.

Table 2. The “pneumocystis pneumonia” theme in the AIDS database. The top twenty and the bottom twenty terms are listed. Terms ending with “!!t” or “!!p” are from the text (title or abstract). Terms ending with “!!T” or “!!P” are from the title. All other terms are MeSH terms.

$\propto_t$	weight	term
4203.16	7.67756	pneumocystis!!t
4131.06	7.45661	carinii!!t
4044.71	7.2371	pneumocystis carinii!!p
3451.11	6.22396	pneumonia, pneumocystis carinii!
3189.29	5.88545	pneumonia!!t
3148.41	8.89602	pneumocystis!!T
3128.89	6.16155	carinii pneumonia!!p
3076.95	11.3618	carinii!!T
3019.83	11.3184	pneumocystis carinii!!P
2056.77	6.119	pneumonia!!T
2034.68	10.5684	carinii pneumonia!!P
1334.72	4.80077	pneumonia, pneumocystis carinii!complications
1196.67	6.04187	pcp!!t
1166.19	7.33732	pneumonia pcp!!p
874.62	4.7418	pneumonia, pneumocystis carinii!drug therapy
860.536	2.73083	acquired immunodeficiency syndrome!complications
765.718	4.90711	pentamidine!
747.283	4.9847	pentamidine!!t
687.268	4.81072	pneumonia, pneumocystis carinii!diagnosis
644.147	4.36415	pneumonia, pneumocystis carinii!etiology
. . . .	. . . .	. . . .
288.126	5.02338	pneumocystis pneumonia!!p
279.602	5.41492	pneumocystis carinii!isolation & purification
279.038	3.85202	bronchoalveolar lavage fluid!
276.319	4.35901	trimethoprim sulfamethoxazole!!p
269.075	4.23562	bronchoscopy!
266.897	6.37945	pentamidine!administration & dosage*
258.448	4.98364	pneumonia, pneumocystis carinii!mortality
239.531	4.00696	trimethoprim-sulfamethoxazole combination!
239.136	1.70665	diagnosis!!t
236.596	7.98609	carinii infection!!P
231.826	4.30286	trimethoprim!
221.698	4.05579	prophylaxis!!T
220.914	2.85947	respiratory!!t
218.138	4.89353	trimethoprim!therapeutic use
217.447	4.23204	pentamidine!adverse effects
214.698	4.48168	transbronchial!!t
208.269	4.20944	trimethoprim-sulfamethoxazole combination!therapeutic use
199.971	4.20475	sulfamethoxazole!
193.587	4.89464	sulfamethoxazole!therapeutic use
190.416	3.63024	alveolar!!t

8. Future Plans

Our immediate plan is to extract the literature from MEDLINE® that deals with genetics (over one million documents) and produce a set of themes for this subject area. It is unclear whether such a large set of themes will lend itself to browsing, though we plan to experiment with browsing. We are more optimistic regarding a different strategy. We plan to treat the individual themes as documents and make them accessible through Boolean querying much as for documents. Because the terms in themes are rated by their associated α_t values, these values may be used to produce ranked retrieval. This is straightforward in the case of a single query term and for Booleans could make use of some kind of extended Boolean[19] or fuzzy logic. Once a user has selected a theme consistent with his interests he has the option of using it to produce ranked retrieval of the documents in the database. This is based on the weights associated with the terms in a theme (see Table 2) and is defined by (3.13). Another potential application of themes is to the problem of term disambiguation. If a term occurs in multiple themes and the term occurs in a document then one may compare the themes with the document to see which theme best fits the context and interpret the term accordingly. We hope to investigate the usefulness of this approach in future work.

References

1. P. Willett, "Recent trends in hierarchical document clustering: A critical review" *Information Processing & Management* **24**, 577-597(1988).

2. E.M. Voorhees, The cluster hypothesis revisited. Proceedings of the Eighth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (Association for Computing Machinery, New York, 1985) 188-196.

3. M.A. Hearst, C. Karadi, Cat-a-Cone: An interactive interface for specifiying searches and viewing retrieval results using a large category hierarchy. in N.J. Belkin, A.D. Narasimhalu, P. Willett, eds. ACM SIGIR'97, (ACM Press, Philadelphia, Pennsylvania, 1997)

4. J.A. Wise, "The ecological approach to text visualization" *Journal of the American Society for Information Science* **50**, 1224-1233(1999).

5. P. Au, M. Carey, S. Sweraz, Y. Gua, S.M. Ruger, New paradigms in information visualization. in N.J. Belkin, P. Ingwersen, M.-K. Leong, eds. ACM SIGIR2000, (ACM Press, Athens, Greece, 2000) 307-309.

6. P. Langley. *Elements of Machine Learning* (Morgan Kaufmann Publishers, Inc., San Francisco, 1996).

7. T.M. Mitchell. *Machine Learning* (WCB/McGraw-Hill, Boston, 1997).

8. A.D. Gordon. *Classification* (Chapman & Hall/CRC, New York, 1999).

9. R.O. Duda, P.E. Hart, D.G. Stork. *Pattern Classification* (Second ed.) (John Wiley & Sons, Inc., New York, 2000).

10. N. Slonim, N. Tishby, Document clustering using word clusters via the infomation bottleneck method. in N. Belkin, P. Ingwersen, M.-K. Leong, eds. ACM SIGIR2000, (ACM Press, Athens, Greece, 2000) 208-215.

11. L.D. Baker, A.K. McCallum, Distibutional clustering of words for text classification. in W.B. Croft, A. Moffat, C.J. van Rijsbergen, R. Wilkinson, J. Zobel, eds. ACM SIGIR'98, (ACM Press, Melbourne, Australia, 1998) 96-103.

12. T. Hofmann, Probabilistic latent semantic indexing. Twenty-second Annual International SIGIR Conference on Research and Development in Information Retrieval, (1999)

13. T. Hofmann, Learning and representing topic. Conference for Automated Learning & Discovery: Workshop on Learning from Text and the Web, (CMU, 1998)

14. H. Li, K. Yamanishi, Document classification using a finite mixture model. Conference of the Association for Computational Linguistics, (Madrid, Spain, 1997) 39-47.

15. M. Sahami, M. Hearst, E. Saund, Applying the multiple cause mixture model to text categorization. in L. Saitta, ed. Machine Learning: Proc. of the Thirteenth International converence, (Morgan Kaufmann, San Francisco, California, 1996) 435-443.

16. H. Shatkay, W.J. Wilbur, Finding themes in MEDLINE documents: Statistical similarity search. IEEE ADL2000, (Bethesda, Maryland, 2000) 183-192.

17. H. Shatkay, W.J. Wilbur, Genes, themes, and microarrays. ISMB2000, (San Diego, California, 2000) 317-328.

18. R.J.A. Little, D.B. Rubin. *Statistical Analysis with Missing Data* (John Wiley & Sons, New York, 1987).

19. G. Salton, E.A. Fox, H. Wu, "Extended boolean information retrieval" *Communications of the ACM* **26**, 1022-1036(1983).