Topics in 0-1 Data

Ella Bingham
Ella.Bingham@hut.fi

Heikki Mannila
Heikki.Mannila@hut.fi

Jouni K. Seppänen
Jouni.Seppanen@hut.fi

Helsinki University of Technology
Laboratory of Computer and Information Science and HIIT Basic Research Unit
P.O. Box 5400, FIN-02015 HUT, Finland

ABSTRACT

Large 0-1 datasets arise in various applications, such as market basket analysis and information retrieval. We concentrate on the study of topic models, aiming at results which indicate why certain methods succeed or fail. We describe simple algorithms for finding topic models from 0-1 data. We give theoretical results showing that the algorithms can discover the epsilon-separable topic models of Papadimitriou et al. We present empirical results showing that the algorithms find natural topics in real-world data sets. We also briefly discuss the connections to matrix approaches, including nonnegative matrix factorization and independent component analysis.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Contingency table analysis; H.2.8 [**Database Management**]: Database Applications—*Data mining*; I.5.1 [**Pattern Recognition**]: Models—*Structural*

General Terms

Algorithms, Theory

1. INTRODUCTION

Large 0-1 datasets occur in various applications, such as market basket analysis, information retrieval, and mobile service use analysis. Lots of research has been done in the data mining community on methods for analyzing such data sets. The techniques can be roughly divided into two classes: (i) the methods based on frequent sets and association rules, aiming at discovery of interesting patterns, and (ii) probabilistic modeling methods aimed at discovering global structure from the data set.

In this paper we consider methods that fall in between these classes of methods. We study the identification of *topics* from 0-1 data. Intuitively, a topic is a set of interconnected variables such that, the occurrence value 1 in one of them tends to increase the probability of seeing value 1 for the other variables. The term "topic" comes from information retrieval: if a document concerns a certain topic, then the occurrence of some words is more probable than in the case when the document does not concern that topic. A single document can discuss many topics, and all words belonging to a topic need not appear in a document about that topic.

The concept of a topic is not restricted to document data. For example, in market basket data one can consider the customers having different topics in mind when they enter the store. A customer might for example want to purchase beer; the actual brand is selected only later, and perhaps she/he buys more than one brand.

The problem of finding topics in data has been considered using various approaches. Examples of the approaches are identification of finite mixtures, latent semantic indexing, probabilistic latent semantic indexing, nonnegative matrix factorization, and independent component analysis (see, e.g., [5, 10, 8, 14, 11, 13, 12]). Related work is considered in some more detail in Section 6.

We describe a simple topic model, corresponding to a generative model of the observations. The model states that there is a number of topics in the data, and that the occurrences of topics are independent. Given that the topic occurs, the words belonging to that topic are also considered to be independent. Later, we consider an extension of the model where the probabilities of topics vary from document to document (as in, e.g., [11, 14]).

The first question to address is whether actual data sets can be considered to be results of such generative process. Our definition of topic models implies, e.g., that negative correlations between variables are absent. We show that this is indeed the case on real data sets: while there are negative correlations, they are typically quite weak and cannot be considered to be violations of the model.

Given the class of topic models, the problem is then whether the model parameters can be estimated from the data. Our model class is close to the class of finite mixtures of multivariate Bernoulli distributions, a nonidentifiable class [10]. However, while in Bernoulli distributions the information obtained from 0 and 1 are on an equal footing, in our model the values 0 and 1 are not symmetric. This implies that for models where the topics are almost disjoint (e.g., the ε-separability condition of Papadimitriou et al. [14]) we can efficiently identify the topics and their parameters. Our main focus is whether there are some simple theoretical arguments that imply that simple topic models can be estimated from

SIGKDD '02 Edmonton, Alberta, Canada

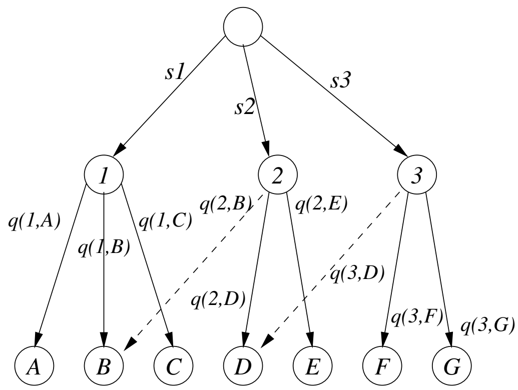

Figure 1: An example topic model

the data. We are able to show some first results in this direction, and support the results with empirical evidence.

The rest of this paper is organized as follows. In Section 2 we define the variants of topic models we consider. Section 3 describes the algorithms we use to find the topics. The theoretical results showing why the algorithms have a chance of working are given in Section 4. Some empirical results are described in Section 5. Related work is discussed in Section 6. Section 7 is a a short conclusion.

2. TOPIC MODELS

In this section we first introduce the simple topic model we consider, and then give an extension that corresponds to the model in [14]. We sometimes use the terminology of information retrieval, talking about documents instead of observations.

Given a set U of attributes, a *k-topic model* $\mathcal{T} = (\bar{s}, q)$ consists of k *topic probabilities* $\bar{s} = (s_1, \ldots, s_k)$ and a *topic-attribute probability matrix* q, giving for each $i = 1, \ldots, k$ and $A \in U$ an *attribute probability* $q(i, A)$ of A in topic i.

A document is sampled from $\mathcal{T}$ as follows. First, one selects independently which topics are on: topic i is on with probability s_i. All attributes are initially assigned value 0. Then, for each topic i that was selected, attribute A is assigned value 1 with probability $q(i, A)$.

Given a topic model $\mathcal{T} = (\bar{s}, q)$, the *weight* $w(i)$ of topic i in $\mathcal{T}$ is $\sum_{A \in U} q(i, A)$, i.e., the expected value of ones generated by topic i.

A topic model $\mathcal{T} = (\bar{s}, q)$ is *ε-separable*, if for each topic i there exists a set $U_i \subseteq U$ of attributes such that $U_i \cap U_j = \emptyset$ for $i \neq j$ and $\sum_{A \notin U_i} q(i, A) \leq \varepsilon w(i)$. That is, each topic i concentrates most of its mass to entries in U_i, and the overlap between these sets gets at most mass ε. We call U_i the *primary set of attributes* of topic i, and for $A \in U_i$ we say that i is the topic of A. If $A, B \in U_i$ for some i, we say that A and B belong to the same topic. Thus a 0-separable topic model is one where for each attribute A there is at most one topic i such that $q(i, A) > 0$.

Figure 1 illustrates an ε-separable topic model. The attribute set is $U = \{A, B, \ldots, G\}$, there are 3 topics, and the attribute subsets corresponding to the topics are $U_1 = \{A, B, C\}$, $U_2 = \{D, E\}$, and $U_3 = \{F, G\}$. The dashed arrows are examples of relationships that are possible in an ε-separable model with $\varepsilon > 0$.

A possible drawback with the above model for the generation of observations is that the topic probabilities s_i are considered to be constant: this could be considered unrealistic. Next we describe a variant, the *varying-probability topic model* in which they are also allowed to vary. Such a topic model is described as $\mathcal{T} = (\mathcal{S}, q)$, where $\mathcal{S}$ is a finite set of topic probability vectors $\bar{s}$.

A document is sampled from a varying-probability topic model by sampling first the topic probabilities $\bar{s}$ from $\mathcal{S}$, and then using the resulting topic model $(\bar{s}, q)$ as above. Thus this model is quite similar to the ones described in [14, 11]. The weight of a topic in such a model is defined to be the expected weight of topic under the sampling of the probability vector $\bar{s}$.

The condition of ε-separability is defined for varying-probability topic models in the same way as for normal topic models: at most a fraction of ε of the weight of each topic goes outside the primary attributes of that topic.

Given an 0-1 table over attributes U, denote for $A, B \in U$ by $p(A)$ the probability in the data of the event $A = 1$ and by $p(AB)$ the probability of $A = 1 \wedge B = 1$. Then the conditional probability $p(A|B)$ of A given B is of course $p(AB)/p(B)$. In practice, the probabilities are estimated as frequencies in the data.

There are certain degenerate cases in which the identification of topics does not succeed. For example, if there is one topic with one attribute, then different combinations of topic and attribute probabilities give the same observed frequency.

3. ALGORITHMS FOR FINDING TOPICS

In this section we describe two simple algorithms for finding topics. The first algorithm is applicable only to the basic model, while the second works also for varying-probability topic models.

Ratio algorithm. Consider first a k-topic 0-separable model $\mathcal{T} = (\bar{s}, q)$. Given two attributes A and B belonging to the same topic i, we have $p(A) = s_i q(i, A)$ and $p(B) = s_i q(i, B)$. Furthermore, $p(AB) = s_i q(i, A) q(i, B)$. Thus we have

$$\frac{p(A)p(B)}{p(AB)} = s_i.$$

If, however, A and B belong to different topics i and j, we have $p(A) = s_i q(i, A)$ and $p(B) = s_j q(j, B)$, and $p(AB) = s_i s_j q(i, A) q(j, B)$. Hence

$$\frac{p(A)p(B)}{p(AB)} = 1.$$

In the ε-separable case, any attribute may in principle be generated by any topic, and so $p(A) = \sum_i s_i q(i, A)$ and $p(AB) = \sum_i s_i q(i, A) q(i, B) + \sum_i \sum_{k \neq i} s_i s_k q(i, A) q(k, B)$.

Thus the algorithm for finding topics is simple. Compute the ratio $r(A, B) = p(A)p(B)/p(AB)$ for all pairs A and B; if the ratio is about 1, the attributes belong to different topics, if it is less than 1, the attributes might belong to the same topic.

Finding the topics from these ratios can be formalized as follows. We search for a partition of the set of attributes U into subsets so that within subsets most of the ratios $r(A, B)$ are close to a constant, and between subsets most of the ratios are close to 1. That is, given the matrix $r(A, B)$, where $A, B \in U$, and an integer k, find the partition of U to subsets U_i for $i = 1, \ldots, k$, minimizing the score

$$\alpha \sum_{i=1}^{k} \sum_{A,B \in U_i} (r(A,B) - \gamma_i)^2 + \beta \sum_{i=1}^{k} \sum_{j=1,\ldots,k, j \neq i} \sum_{A \in U_i} \sum_{B \in U_j} (r(A,B) - 1)^2,$$

where α and β are constants and γ_i is the average of the ratios $r(A,B)$ within block U_i. This is a typical clustering problem, NP-complete in its general form, but lots of good approximate solutions exist.

This almost trivial method actually works quite nicely on some artificial and real data sets. However, it fails whenever the observations are generated using the varying-probability topic model. Thus we need more refined techniques.

Probe algorithm. Our second method is still quite simple. It is based on the method for finding similar attributes in 0-1 data described by Das et al. [7]. The basic intuition behind the algorithm is as follows. If two attributes A and B belong to the same topic, then the information that the occurrence of A (meaning the event $A = 1$) gives is about the same as the information given by the occurrence of B. Thus, if we have a measure for the similarity of the information given by two attributes, we can use that to find topics.

The *probe distance* $d(A,B)$ of two attributes is defined by

$$d(A,B) = \sum_{C \in U \setminus \{A,B\}} |p(C|A) - p(C|B)|.$$

The intuition here is that attributes A and B are similar if the distributions of the other attributes in the rows with $A = 1$ and in the rows with $B = 1$ are about the same. The attributes C serve as probes which are used to measure how similar the sets of rows are.

Our algorithm is as follows. Compute distances $d(A,B)$ for all pairs of attributes. (For a data set of n rows and p attributes, this can be done in time $O(np^2)$.) Again, find a partition of the set U of all attributes to subsets U_i minimizing the within-cluster distance and maximizing the distances between clusters. This can, of course, be solved using any clustering method. The details of the clustering are not our main focus; rather, we aim at giving results indicating why the method works. This is done in the next section.

4. PROPERTIES OF THE PROBE ALGORITHM

In this section we consider the properties of the probe algorithm given in the previous section. We first consider the case of 0-separable models, which naturally are quite simple. We show that for large sample sizes the distance between two attributes in the same topic tends to 0, and that the expected distance between two attributes belonging to different topics is quite large. We then consider the case of ε-separable models, and show that the same results continue to hold under some additional conditions. Most of the results are formulated under the assumption of no sample effects, i.e., by assuming infinite sample size.

We start with a lemma showing that for 0-separable models the distance between two attributes in the same topic goes to 0 as the sample size grows.

LEMMA 1. *Let r be a table of n rows over attributes U generated by a 0-separable topic model $\mathcal{T} = (\bar{s}, q)$. If A and B belong to the same topic U_i, then $\lim_{n\to\infty} d(A,B) = 0$.*

The next proposition extends this result to varying-probability topic models.

THEOREM 1. *Let r be a table of n rows over attributes U generated by a 0-separable varying-probability topic model $\mathcal{T} = (\mathcal{S}, q)$. Then, if A and B belong to the same topic U_i, then $\lim_{n\to\infty} d(A,B) = 0$.*

PROOF. Consider each probability vector $\bar{s} \in \mathcal{S}$. For the observations generated using the topic model $(\bar{s}, q)$ the lemma holds. As the statement of the lemma is independent of the actual topic probabilities s_i, the claim follows. □

LEMMA 2. *Let r be a table of n rows over attributes U generated by a 0-separable topic model $\mathcal{T} = (\bar{s}, q)$. If attribute A belongs to topic i, and attribute D belongs to topic j with $j \neq i$, then $E(d(A,D)) = (1 - s_i)(w(\mathcal{T}, i) - q(i,A)) + (1 - s_i)(w(\mathcal{T}, j) - q(j,D))$.*

THEOREM 2. *Let r be a table of n rows over attributes U generated by a 0-separable varying-probability topic model $\mathcal{T} = (\mathcal{S}, q)$. If attribute A belongs to topic i, and attribute D belongs to topic j with $j \neq i$, then $E(d(A,D)) = (1 - s_i)(w(\mathcal{T}, i) - q(i,A)) + (1 - s_i)(w(\mathcal{T}, j) - q(j,D))$.*

The proof is the same as for Theorem 1.

The above results show that the probe distances have a meaningful relationship to the topics of a 0-separable topic model. The details for general ε-separable models are far messier, but we give here an analogue of Lemma 1. The intuition is that when we add some weak links to a 0-separable model, the conditional probabilities are not perturbed too much, and thus the probe distances within a single topic will remain small. However, there are pathological ε-separable models: for example, consider a model where all attribute probabilities are much less than ε. Then, changes of the order of ε will naturally have a significant impact on the model. Of course, there is little hope of finding the topics in this kind of a model.

To rule out this kind of cases, there are several possibilities. For example, we can define the *distinctiveness* of an ε-separable topic model $\mathcal{T} = (\bar{s}, q)$ as the smallest value of the probability of an attribute being generated in the context of its primary topic:

$$\Delta(\mathcal{T}) = \min_{i, A \in U_i} s_i q(i,A),$$

where the minimum is taken over all topics i and all attributes $A \in U_i$. Thus, if a model has high distinctiveness ($\Delta(\mathcal{T}) \gg \varepsilon$), the generated attributes should usually reflect the topics they belong to.

An alternative restriction would be to say that the ε-separable topic model $\mathcal{T}$ has *θ-bounded conspiracy*, if for all attributes A with topic i we have $\sum_{j \neq i} q(j,A) \leq \theta$, i.e., the model $\mathcal{T}$ assigns at most a mass of θ to any attribute from topics other than its main topic. That is, the other topics do not conspire against a single attribute in a topic. Similar results as the one below can be proved for that case.

LEMMA 3. *Let r be a table of n rows over attributes U generated by a ε-separable topic model $\mathcal{T} = (\bar{s}, q)$. If attributes A and B belong to the same topic i, then $E(d(A,B)) \leq 2|U|k\varepsilon/\Delta(\mathcal{T})$.*

5. EMPIRICAL RESULTS

5.1 Experiments on simulated data

To evaluate how well do our algorithms perform, we generated artificial data according to our topic models described in Section 2. The data consisted of 100 attributes and 10 topics, each topic having a random number of primary attributes, and the number of observations was 100000. We performed tests on a ε-separable model with $\varepsilon = 0$, 0.01 and 0.1. In all experiments with the first (constant topic probabilities) model, the topic probabilities s_i were the same, so that we were able to test the effect of ε in model estimation accuracy.

Ratio algorithm. First we considered the ratios $r(A, B) = p(A)p(B)/p(AB)$. Recall that this should yield s_i, probability of topic i if A and B belong to the same topic i, and 1 otherwise, as then A and B are independent and their joint probability is separable. By listing these ratios in a matrix one can easily distinguish which topics belong to the same topic, as all of them have approximately the same ratio. In this way we can estimate the topic structure of the data, and also the topic probabilities s_i and topic-attribute probabilities $q(i, A)$ of A in topic i. Comparing to the true probabilities, the mean squared errors (MSE) of topic probabilities and the MSEs of topic-attribute probabilities are listed in Table 1 for $\varepsilon = 0$, 0.01 and 0.1. These figures are averages of 10 experiments. The variance between experiments was very small.

ε	MSE of topic probs.	MSE of topic-attr. probs.
0	$0.92 \cdot 10^{-4}$	$1.00 \cdot 10^{-3}$
0.01	$1.04 \cdot 10^{-4}$	$1.02 \cdot 10^{-3}$
0.1	$1.01 \cdot 10^{-4}$	$1.03 \cdot 10^{-3}$

Table 1: Mean squared errors of estimated topic and topic-attribute probabilities in the ratio algorithm.

In our varying-probability topic model, the topic probabilities s_i are randomly drawn for each document, and the ratio algorithm is not applicable.

Probe algorithm. Sammon mapping [17] is a convenient way to visualize how the attributes are grouped into distict topics. Figure 2 shows the Sammon map of the probe distances of the attributes in the 0-separable model. We can see that the attributes are nicely grouped into about 10 clusters, most of which are clear in shape. The clusters are not of equal size, as each topic has a random number of primary attributes. In the case of $\varepsilon = 0.01$, the clusters are a bit more vague in shape but still visible; with $\varepsilon = 0.1$, no clear clusters are seen anymore. The probe algorithm is quite resistant to the extension of varying topic probabilities: the Sammon maps are remarkably similar to those obtained for the nonvarying-probability topic models.

Maximum entropy model. We also considered whether the maximum entropy method described in e.g. [16, 15] might be useful in finding topics. The method is used to answer queries about the data as follows: first, one mines frequent sets with some threshold [1, 2], and then finds the maximum entropy distribution [3, 9] consistent with the frequent sets. We performed experiments using simulated data to see whether the results are consistent with the topic models used to generate the data. The results (not shown) indicate that this method does find results consistent with topic models quite satisfactorily but not perfectly. However, the performance is comparable only when the method is given roughly as much input as the simpler probe algorithm, and degrades badly when the frequency threshold increases and the input size decreases.

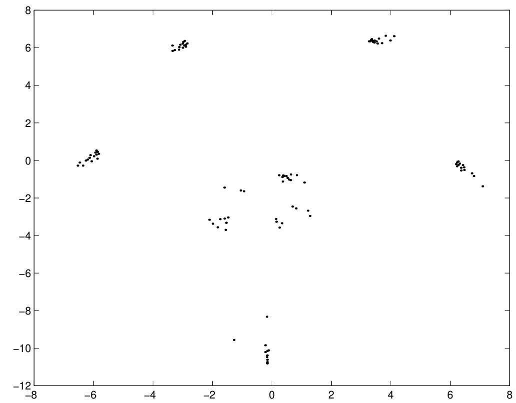

Figure 2: Sammon map of probe distances of attributes in artificial data; $\varepsilon = 0$.

5.2 Experiments on real data

Correlations. To determine the validity of the model assumptions on real data, we performed some trials on a collection of bibliographical data on computer science available on the WWW[1]. We call this the "Theory" dataset. As a preprocessing step, we removed all words occurring in fewer than 20 documents in the database. This reduced the number of words to 4227; the number of documents is 67066.

After preprocessing, we determined the probabilities $p(A)$ and $p(AB)$ for all words A, B (using word frequencies) and computed the covariances $\mathrm{cov}(A, B) = p(AB) - p(A)p(B)$. We can derive from the theoretical model in Section 2 that $\mathrm{cov}(A, B) \geq 0$ for all words A, B. This is not true in the dataset; indeed, more of the covariances are negative than positive. However, the distributions of the positive and negative covariances are very different. Figure 3 displays logarithmic histograms of the covariances in the Theory data. The histograms have been scaled to have equal areas. A short vertical line marks the position corresponding to one line in the database; covariances that are (absolutely) much smaller than this aren't usually very interesting, since they tend to reflect small-sample effects in cases where $p(AB)$ is very small (perhaps 0 or 1 lines) and $p(A)p(B)$ is nonzero but small.

Probe algorithm. We studied the behavior of the probe algorithm on the Theory bibliography. As a preprocessing step, we removed a small set of stop words and all numbers in the data, and then selected the 200 most frequent terms.

The probe distances of the terms were computed, and the term pairs with minimum probe distance are listed in Table 2. The table lists all pairs whose probe distance is under 1, in increasing order; the mean distance was about 2.7 and maximum distance about 6.2. The term pairs, most of which are abbreviations, are quite meaningful: e.g. 'stoc' is ACM Symp. on Theory of Computing and 'focs' is Symp.

[1] `http://liinwww.ira.uka.de/bibliography/Theory/Seiferas/`

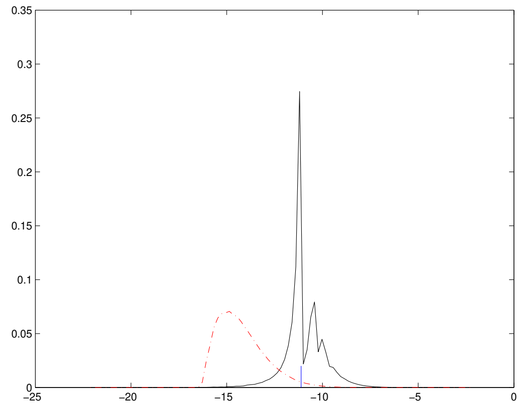

Figure 3: **Histogram of $\ln(|\operatorname{cov}(A,B)|)$ for positive (solid) and negative (dashdotted) covariances for words A, B in Theory. A short vertical line marks $\ln(1/67066) = -11.1$.**

dist.	terms		dist.	terms	
0.50	stoc	focs	0.91	jacm	libtr
0.63	infctrl	tcs	0.92	extended	abstract
0.63	tr	libtr	0.93	stacs	icalp
0.67	icalp	tcs	0.94	actainf	tcs
0.75	infctrl	icalp	0.95	fct	jcss
0.76	eurocrypt	crypto	0.95	fct	mfcs
0.79	mfcs	tcs	0.96	stacs	jcss
0.81	infctrl	jcss	0.96	jacm	tr
0.81	mfcs	icalp	0.96	sijdm	damath
0.81	jcss	tcs	0.97	ipps	jpdc
0.84	mfcs	infctrl	0.98	stoc	tr
0.86	mfcs	jcss	0.98	icpp	jpdc
0.88	jcss	icalp	0.99	sicomp	libtr
0.88	ipps	icpp	0.99	stacs	infctrl
0.89	mst	jcss	0.99	stacs	tcs

Table 2: Term pairs with minimum probe distance in the Theory data set

on Foundations of Computer Science; 'infctrl' is Information and Computation (formerly Information and Control) and 'tcs' is Theoretical Computer Science. For each term pair, the pair members belong to the same topical field, be it theoretical computer science, technical reports, cryptography, parallel processing, discrete mathematics etc. All these terms appear quite often in the data base, which makes the estimation of their probe distances reliable.

Does the method find topics? For example, listing the 10 terms with minimum probe distance to 'stoc' we get 'focs', 'tr', 'sicomp', 'libtr', 'stacs', 'jacm', 'jcss', 'icalp', 'infctrl', and 'ipl'. Computing the average distances of every term in this list to all other terms in the list, and taking the average of these averages, we get a distance of 1.17. On the other hand, computing the average distances of every term in this list to all other terms in the vocabulary, and again taking the average, yields 2.30. So the terms close to 'stoc' are also very close to one another but less close to other terms, and can thus be seen as forming a sort of topic. A similar comparison can be done to the closest neighbors of 'focs', giving a similar term list as above with similar average distances.

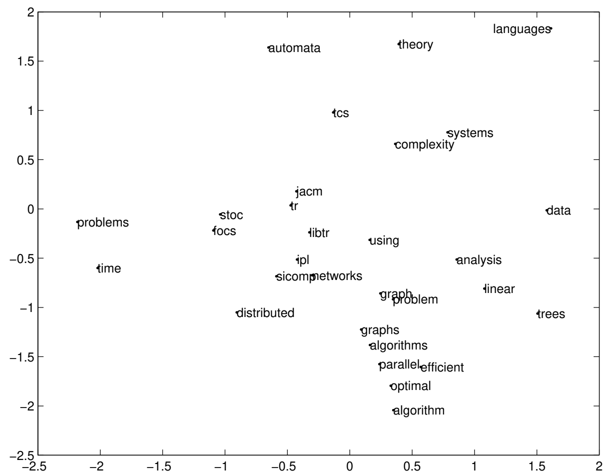

Figure 4: Sammon map of the probe distances of the 30 most common terms in the Theory data set.

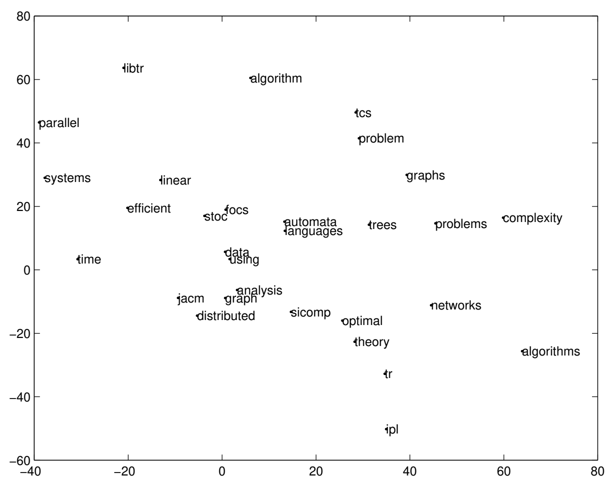

Figure 5: Sammon map of the LSI projections of the 30 most common terms in the Theory data set.

We used Sammon's mapping to project the data into two dimensions; Figure 4 shows how the 30 most common terms are located. There is clear evidence of clustering of related terms.

For comparison, we also projected the data into its 20-dimensional LSI [8] space. The Sammon map of the 30 most common terms is seen in Figure 5. In interpreting the figures, one should bear in mind that a two-dimensional Sammon map may not truly represent the locations of high-dimensional vectors.

6. RELATED WORK

The idea of looking at topics in 0-1 data (or other discrete data) has been considered in various contexts. The latent semantic indexing (LSI) method [8] uses singular-value decomposition (SVD) to obtain good choices of topics. This method works quite nicely in practice; the reason for this remains unclear. In a seminal paper [14], Papadimitriou et al. gave some arguments justifying the performance of LSI. Their basic model is quite general and we have adopted their basic formalism; to obtain the results on LSI they have to restrict the documents to stem from a single topic. Of course, some restrictions are unavoidable.

Hofmann [11] has considered the case of probabilistic LSI. His formal model is close to ours, having the form $P(w|d) = \sum_z P(z|d)P(w|z)$, where the z's are topics, d refers to a document, and w to a word. Hofmann's main interest is in good estimation of all the parameters using the EM algorithm, while we are interested in having some reasoning explaining why the methods would find topics.

Cadez et al. [4] have considered the estimation of topic-like market-basket data, with the added complication that the same customer has multiple transactions, leading to the need of individual weights.

Our topic models are fairly close to the class of finite mixtures of multivariate Bernoulli distributions, a nonidentifiable class [10] (see also [5]). However, for those models, the values 0 and 1 have symmetric status, while for the topic models defined above this is not the case. We conjecture that the class of topic models is essentially identifiable provided that the topics are almost disjoint in, e.g., the ε-separability sense.

In nonnegative matrix factorization (NMF), an observed data matrix V is presented as a product of two unknown matrices: $V = WH$. All three matrices have nonnegative entries. Lee and Seung [13] give two practical algorithms for finding the matrices W and H given V. Restriction to binary variables is not straightforward in these algorithms.

Independent component analysis (ICA) [6, 12] is a statistical method that expresses a set of observed multidimensional sequences as a combination of unknown latent variables that are more or less statistically independent. Topic identification in 0-1 data can be interpreted in the ICA terminology as finding latent binary sequences, unions of which form the observed binary data. ICA in its original form relies heavily on matrix operations; for sparse data, union is roughly equivalent to summation, so methods for ICA could be considered for the problem at hand. Nevertheless, most existing ICA algorithms are suitable for continuosly distributed data with Gaussian noise — the case of 0-1 variables and Bernoulli noise is quite different, and practical ICA algorithms tend to fail in this case.

7. CONCLUSIONS

We have considered the problem of finding topics in 0-1 data. We gave a formal description of topic models, and showed that relatively simple algorithms can be used to find topics from data generated using such models. We showed that the probe algorithm works reasonably well in practice.

Lots of open issues remain, both on the theoretical and on the practical side. The detailed relationship of our model compared to, e.g., Hofmann's model remain to be studied. We conjecture that the topic models are identifiable, in contrast with general mixtures of multivariate Bernoulli distributions. Understanding the behavior of LSI is still open. Similarly, seeing how nonnegative matrix factorization is connected to the other approaches is open, as are the ways of extending ICA to the Bernoulli case.

8. REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *SIGMOD '93*, pages 207–216, 1993.

[2] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo. Fast discovery of association rules. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, chapter 12, pages 307–328. AAAI Press, 1996.

[3] A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

[4] I. V. Cadez, P. Smyth, and H. Mannila. Probabilistic modeling of transaction data with applications to profiling, visualization, and prediction. In *KDD 2001*, pages 37–46, San Fransisco, CA, Aug. 2001.

[5] M. A. Carreira-Perpinan and S. Renals. Practical identifiability of finite mixtures of multivariate Bernoulli distributions. *Neural Computation*, 12:141–152, 2000.

[6] P. Comon. Independent component analysis — a new concept? *Signal Processing*, 36:287–314, 1994.

[7] G. Das, H. Mannila, and P. Ronkainen. Similarity of attributes by external probes. In *Knowledge Discovery and Data Mining*, pages 23–29, 1998.

[8] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[9] S. Della Pietra, V. J. Della Pietra, and J. D. Lafferty. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):380–393, 1997.

[10] M. Gyllenberg, T. Koski, E. Reilink, and M. Verlaan. Non-uniqueness in probabilistic numerical identification of bacteria. *Journal of Applied Probability*, 31:542–548, 1994.

[11] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99*, pages 50–57, Berkeley, CA, 1999.

[12] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

[13] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, Oct. 1999.

[14] C. H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala. Latent semantic indexing: A probabilistic analysis. In *PODS '98*, pages 159–168, June 1998.

[15] D. Pavlov, H. Mannila, and P. Smyth. Probabilistic models for query approximation with large sparse binary datasets. In *UAI-2000*, 2000.

[16] D. Pavlov and P. Smyth. Probabilistic query models for transaction data. In *KDD 2001*, 2001.

[17] J. W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 18(5):401–409, May 1969.