CS229 Project Report
Document Retrieval by Similarity: An Application of Probabilistic Latent Semantic Analysis (PLSA)

Firouzeh Jalilian Haiyan Liu Jia Pu

December 10, 2009

1 Introduction

In this project, we explored using Probabilistic Latent Semantic Analysis (PLSA) technique to model a large collection of documents. Based on such a model, we also investigated the possibility of an interface that allows a user to 1) observe and explore the document collection on macroscopic level; 2) conduct specific search based on document similarity. PLSA's performance on standard information retrieval (IR) tasks has been well documented [2]. The emphasis of this project is to build an application based on PLSA model to help end-users to explore and search in large collection of documents. Many practical issues arise during this attempt, such as computational efficiency and interpretation of learned model parameters. This report summarizes experience gained and lessons learned from building such an application.

2 Motivations

There are two main motivations for providing similarity based document search facility to end users. First, as storage capacity keeps growing, large repository of documents exists not only on enterprise databases or internet, but also on personal computers. For instance, it's quite common to have tens of thousands of email messages on a personal email account. Given such large number of documents, a user not only expects a good precision-recall rate, but also expects to see most relevant results first. Secondly, a keyword based search system requires a user to know the exact term used in the document he wants to retrieve. The value of a similarity based search system is to retrieve other related documents based on what the user has found using the keyword method.

3 PLSA

Detail of PLSA can be found in [2]. PLSA model can be viewed as unsupervised variation of naive Bayes model. In PLSA, the topic (i.e. class label) of each document example is unknown. So it is necessary to learn the topics from training data. Since the number of topics is given (often determined empirically), PLSA model assigns probability distributions over the topics to documents. Intuitively, such a "soft" classification agrees with the fact that a document often contains more than one topic.

Formally, given a set of documents $\mathcal{D}$ with terms from a vocabulary $\mathcal{W}$, we assume there are a set of latent topics, $\mathcal{Z}$, so that:

$$P(d, w) = P(d) \sum_{z \in \mathcal{Z}} P(w|z)P(z|d)$$

Let $f(d, w)$ denote the number of occurrence of word w in document d. The parameters, $P(w|z)$,

and $P(z|d)$, are learned by maximizing the log-likelihood:

$$\mathcal{L} = \sum_{d\in\mathcal{D}}\sum_{w\in d} f(d,w) log P(d) + \sum_{d\in\mathcal{D}}\sum_{w\in d} f(d,w) log P(w|d)$$

which can be maximized with EM algorithm as following:

E-step:

$$P(z|d,w) = \frac{P(z|d)P(w|z)}{\sum_{z'\in\mathcal{Z}} P(z'|d)P(w|z')}$$

M-step:

$$P(w|z) = \frac{\sum_{d\in\mathcal{D}} f(d,w)P(z|d,w)}{\sum_{d\in\mathcal{D}}\sum_{w'\in\mathcal{W}} f(d,w')P(z|d,w')}$$
$$P(z|d) = \frac{\sum_{w\in\mathcal{W}} f(d,w)P(z|d,w)}{\sum_{w\in\mathcal{W}} f(d,w)}$$

4 Definition of Similarity

In order to put the learned PLSA model into use in our final application, we need to obtain a definition of similarity.

Assume we have a model using K latent topics, i.e. $|\mathcal{Z}| = K$. We use $P(\mathcal{Z}|d)$, a multinomial distribution, to denote the probabilities of document d being on each topic of $\mathcal{Z}$. Since $\sum_{z\in\mathcal{Z}} P(z|d) = 1$, document d can be represented as a point in the (K-1)-simplex defined by $\sum_{i=1}^{K} x_i = 1$. A possible geometric interpretation of similarity is Euclidean distance in this simplex:

$$D_{Euclidean}(d_1,d_2) = \| P(\mathcal{Z}|d_1) - P(\mathcal{Z}|d_2) \|_2$$

However, a major drawback of this definition of similarity is that it does not take any advantage of information theory. For example, if we have a two-topic PLSA model, and four documents d_1, d_2, d_3 and d_4. Also, if these documents have following probabilities on the two topics:

$$P(\mathcal{Z}|d_1) = \{0.0, 1.0\}$$
$$P(\mathcal{Z}|d_2) = \{0.1, 0.9\}$$
$$P(\mathcal{Z}|d_3) = \{0.5, 0.5\}$$
$$P(\mathcal{Z}|d_4) = \{0.6, 0.4\}$$

In this case, $D_{Euclidean}(d_1,d_2)$ is same as $D_{Euclidean}(d_3,d_4)$, although from information theory perspective, the difference between having probability 0 and 0.1 is much more significant than the difference between 0.5 and 0.6.

Kullback-Leibler (KL) divergence is another measure of the difference between two probability distributions. It has been used to measure difference between two multinomial distributions in similar context [5] as document similarity in this report. Using KL divergence as document similarity, it is defined as:

$$D_{KL}(d_1,d_2) = \sum_{i=1}^{K} P(z_i|d_1) \log \frac{P(z_i|d_1)}{P(z_i|d_2)}$$

But one issue with KL divergence is that it is non-symmetric, hence difficult to interpret in this application. After all, what do we mean when we say that A is similar to B, but B is not similar to A? Also, this definition of similarity does not utilize $P(z)$, the probability of each topic on the whole training corpus. Intuitively the fact that $d1$ and $d2$ both have high probability on topic z is more significant if z has overall low probability.

So we need a definition of similarity that is based on sound statistic theory, and has the properties of well defined metric. In [4], the authors proposed a Fisher kernel for generative statistical model. For two examples x_1 and x_2 generated by a model which is parameterized by θ, Fisher kernel is defined as:

$$\mathcal{K}(x_1,x_2) \propto U_{x_1}^T I^{-1} U_{x_i}$$

where $U_x = \nabla_\theta \log P(x|\theta)$. The value of Fisher kernel is that it defines a metric relationship directly on generative model. In [3], the authors derived the Fisher kernel for PLSA model. In

their derivation, the kernel consists of two components:

$$\mathcal{K}_1(d_1, d_2) = \sum_k \frac{P(z_k|d_1)P(z_k|d_2)}{P(z_k)},$$

and

$$\mathcal{K}_2(d_1, d_2) = \sum_j \Big[tf(w_j|d_1) tf(w_j|d_2) \cdot \sum_k \frac{P(z_k|d_1, w_j)P(z_k|d_2, w_j)}{P(w_j|z_k)} \Big]$$

where $tf(w|d)$ is the term-frequency defined as $f(d, w)/\sum_w f(d, w)$. The two components emphasize two aspects of PLSA model. $\mathcal{K}_1$ compares two documents based on their topics, while $\mathcal{K}_2$ compares two documents based their shared words. Since the computation of $\mathcal{K}_2(d_1, d_2)$ is expensive, due to the limited time, we only use $\mathcal{K}_1(d_1, d_2)$ as similarity measure in our implementation, which alone shows improved precision in certain scenario (detail in section 5). To summarize, in our final application, similarity between two documents is defined as :

$$D(d_1, d_2) = \frac{1}{\mathcal{K}_1(d_1, d_2)} = 1/\sum_k \frac{P(z_k|d_1)P(z_k|d_2)}{P(z_k)}$$

The inversion is just to make smaller distance to indicate higher degree of similarity, so that it is consistent with other definitions of similarity we studied in this project.

5 Experiments

People have done many precision-recall evaluations on various corpus using PLSA model. We repeated some of the experiments during our project to 1) find better parameters; 2) verify that our PLSA implementation is correct.

In this project, we used a subset of 6432 documents from Reuters-21578 dataset. In this dataset, each document d has been given a set of tags T_d by human annotators. To evaluate, a random document q is chosen as a query, and up to N most similar documents, $D = \{d_1, \cdots, d_N\}$, are returned. For each d_i in D, if it shares at least one tag with q, i.e. $T_{d_i} \cap T_q \neq \emptyset$, then d_i is a successful retrieval. Our dataset contains 101 unique tags.

Although this type of evaluation is standard practice in IR research, we need to view it with a grain of salt, because matching tags is rather a narrow view of document similarity. For example, in Reuter corpus, document 06114 is on the subject of China's winter crop production. This document is tagged with "grain", "wheat" and "rice". When using this document as query, one of the returned similar documents, 00229, is tagged with "soybean", "red-bean" and "oilseed". For some users, this would be appropriate search result in that they are both clearly on the topic of agriculture. But if we only look at the tags, this is considered a false positive.

For text processing, we only used rather trivial methods. Each document is tokenized using white spaces. Then tokens that do not contain any alphabet, and tokens that are only one character long are removed. After that, all tokens, except acronyms that consist of only upper-case letters, are converted to lower case. And a Porter stemmer is applied. Finally, we removed top words that are identified using normalized entropy [1].

Using evaluation described above, Fig 1 shows the precision-recall rate of using three different definitions of similarity. As it is shown, Fisher kernel generally achieves better precision-recall accuracy than other two similarity measures. Also, empirically we determined to use 30 latent topics in our final model. Fig 2 shows the precision-recall curves using different number of topics.

6 "Spotlight"

The second component of this project is a GUI based application that allows one to explore a collection of documents. It also helped us to subjectively evaluate PLSA model. This application is designed with following goals:

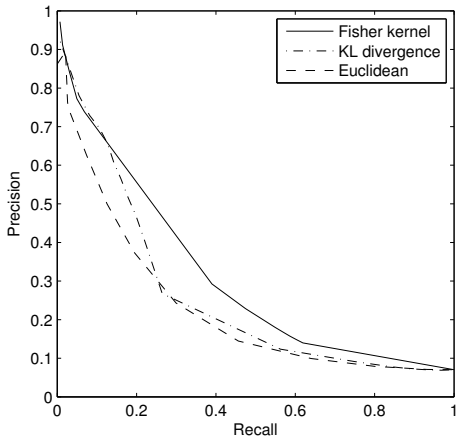

Figure 1: Precision-recall with 30 latent topics

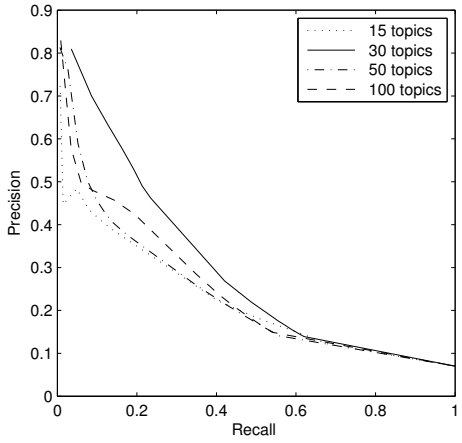

Figure 2: Precision-recall of different number of topics

- It shows similar documents given a query document. This is done by showing the nodes connected by edge. From subjective evaluation, it seems PLSA is quite accurate on identifying similar documents, at least on this dataset.
- Given a pair of similar documents, it shows words from the two documents that contribute most to their common topic. However, sometimes these words are not really essential to the subject of the documents. For instance, for documents about grain production forecast, word such as "December", or "February" often show up.
- By expanding nodes that are connected, it allows a user to identify a set of documents that are similar to each other, but are not similar to any other documents. In fact, these documents form a connected component in the graph built on the whole corpus.

  We set off on this project to investigate the possibility of using PLSA on personal computers. From implementing it, we realized that it is computationally very expensive when the number of latent topics is high, due to the iterative method of learning PLSA model. In our implementation, we processed up to eight data partitions simultaneously. Even with this optimization, it takes more than 5 hours to run EM training for 100 iterations on 100 topics. The demand on computation resource makes it a less attractive solution on today's personal computer. But it still could be useful for moderate dataset.

7 Conclusion

This project demonstrated that PLSA is a powerful technique for modeling document similarity. Combined with keyword search, it provides more flexibility on searching in large collection of documents. In our implementation, we didn't utilize $P(w|z)$. It deserves further investigation on how we can take advantage of $P(w|z)$ to provide more useful information.

References

[1] J.R. Bellegarda. Latent semantic mapping: dimensionality reduction via globally opti-

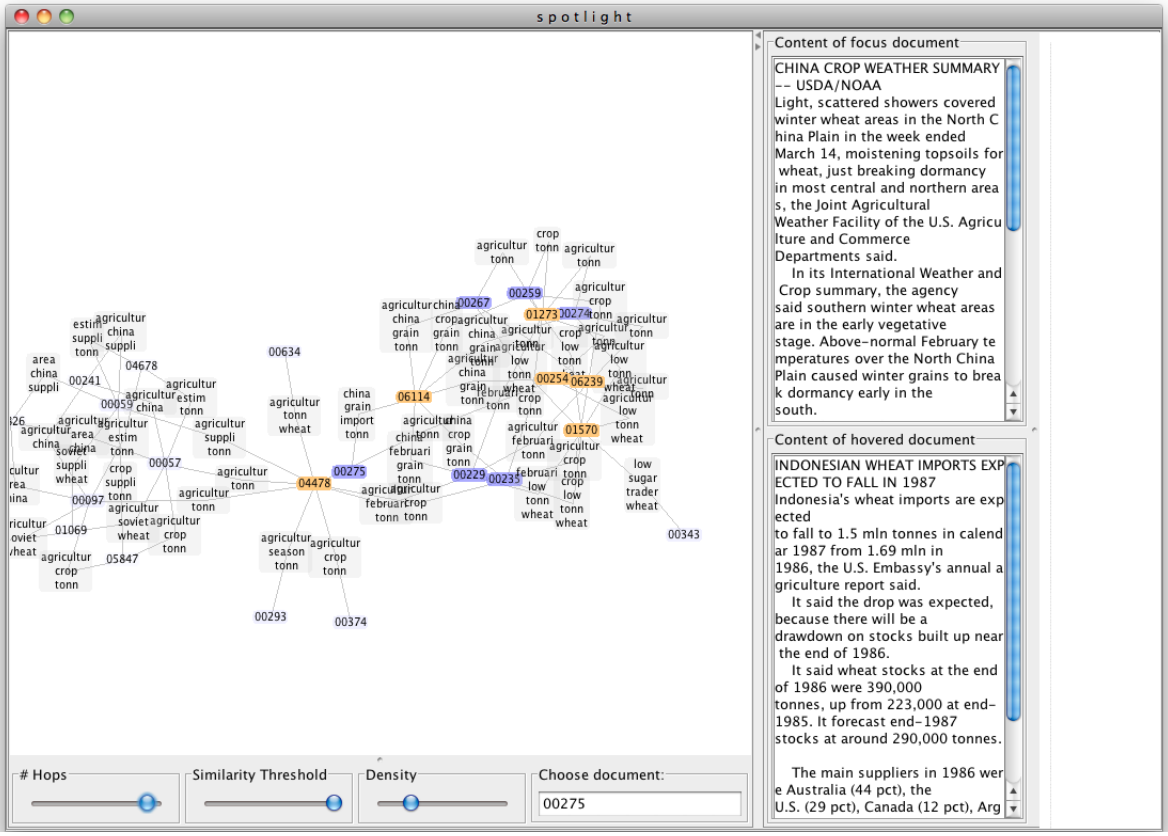

Figure 3: Screen capture of "Spotlight" application

mal continuous parameter modeling. In *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*, pages 127–132, Nov. 2005.

[2] Thorsten Brants. Test data likelihood for plsa models. *Inf. Retr.*, 8(2):181–196, 2005.

[3] Thomas Hofmann. Learning the similarity of documents: An information-geometric approach to document retrieval and categorization, 2000.

[4] Tommi S. Jaakkola and David Haussler. Exploiting generative models in discriminative classifiers. In *Proceedings of the 1998 conference on Advances in neural information processing systems II*, pages 487–493, Cambridge, MA, USA, 1999. MIT Press.

[5] Dong Zhang, Daniel G. Perez, Samy Bengio, and Deb Roy. Learning influence among interacting Markov chains. IDIAP-RR 48, IDIAP, Martigny, Switzerland, 2005.