Improving web page clustering using Probabilistic Latent Semantic Analysis

Lalit A. Patil
Department of
Computer Engineering
KKWIEER,Nasik

S M. Kamalapur
Department of
Computer Engineering
KKWIEER,Nasik

ABSTRACT

Traditional clustering algorithms are usually based on the bag-of-words (BOW) approach. A notorious disadvantage of the BOW model is that it ignores the semantic relationship among words. As a result, if two documents use different collections of core words to represent the same topic, they may be assigned to different clusters, even though the core words they use are probably synonyms or semantically associated in other form and other disadvantage of conventional web page clustering technique is often utilized to reveal the functional similarity of web pages. Tagging can be beneficial to improve the clustering performance. Several efforts have been made to explore social tagging for clustering. But there is some drawbacks of tagging web based clustering. To our knowledge, all the existing approaches exploiting tag information for webpage clustering assume that all the WebPages are tagged, which is a somewhat restrictive assumption. In a more realistic setting, one can only expect that the tags will be available for only a small number of WebPages. In this paper, we propose a new web page grouping approach based on Probabilistic Latent Semantic Analysis (PLSA) model. An iterative algorithm based on maximum likelihood principle is employed to overcome the aforementioned computational shortcoming.

Keywords

Probabilistic latent semantic analysis, Singular Value Decomposition , term-frequency, Web page clustering

1. INTRODUCTION

The main goal of this paper is to present a joint probabilistic model of document to find the discriminant feature of webpages and find the actual relation between word and page. Huge repositories of textual data have become available to a large public. Today, it is one of the great challenges in the information sciences to develop intelligent interfaces for human machine interaction which support computer users in their quest for relevant information. Although the use of elaborate ergonomic elements like computer graphics and visualization has proven to be extremely fruitful to facilitate and enhance information access, progress on the more fundamental question of machine intelligence is ultimately necessary to ensure substantial progress on this issue. In order for computers to interact more naturally with humans, one has to deal with the potential ambivalence, impreciseness, or even vagueness of user requests, and has to recognize the difference between what a user might say or do and what she or he actually meant or intended. One typical scenario of human machine interaction in information retrieval is by natural language queries: the user formulates a request, e.g., by providing a number of keywords or some free-form text, and expects the system to return the relevant data in some amenable representation, e.g., in form of a ranked list of relevant documents. Many retrieval methods are based on simple word matching strategies to determine the rank of relevance of a document with respect to a query. Yet, it is well known that literal term matching has severe drawbacks, mainly due to the ambivalence of words and their unavoidable lack of precision as well as due to personal style and individual differences in word usage. Latent Semantic Analysis (LSA) [1] is an approach to automatic indexing and information retrieval that attempts to overcome these problems by mapping documents as well as terms to a representation in the so called latent semantic space. LSA usually takes the (high dimensional) vector space representation of documents based on term frequencies as a starting point and applies a dimension reducing linear projection. The specific form of this mapping is determined by a given document collection and is based on a Singular Value Decomposition (SVD) of the corresponding term/document matrix. The general claim is that similarities between documents or between documents and queries can be more reliably estimated in the reduced latent space representation than in the original representation. The rationale is that documents which share frequently co-occurring terms will have a similar representation in the latent space, even if they have no terms in common. LSA thus performs some sort of noise reduction and has the potential benefit to detect synonyms as well as words that refer to the same topic. In many applications this has proven to result in more robust word processing. Although LSA has been applied with remarkable success in different domains including automatic indexing (Latent Semantic Indexing, LSI) [1, 3], it has a number of deficits, mainly due to its unsatisfactory statistical foundation. The primary goal of this paper is to present a novel approach to LSA and factor analysis called Probabilistic Latent Semantic Analysis (PLSA) that has a solid statistical foundation, since it is based on the likelihood principle and defines a proper generative model of the data. This implies in particular that standard techniques from statistics can be applied for questions like model fitting, model combination, and complexity control. In addition, the factor representation obtained by PLSA allows to deal with polysemous words and to explicitly distinguish between different meanings and different types of word usage. In Section 2, we describe different types of web page clustering. Section 3 briefly describes proposed system. We discuss related work in Section 4. Our system experiment is described in Section 5. In Section 6, specifies the required dataset and conclude in Section 7.

2. WEB PAGE CLUSTERING

Traditional webpage clustering typically uses only the page content information usually, just the page text in an appropriate feature vector representation such as Bag of Words, Term-Frequency/Inverse-Document-Frequency, etc., and then applies standard clustering algorithms e.g., K-means algorithm , spectral clustering, etc. Traditional webpage clustering has some drawbacks that reducing the speed of searching make navigation problem and Quality of clustering is not good. We are facing an ever increasing volume of text documents. The abundant texts flowing over the Internet, huge collections of documents in digital libraries and repositories, and digitized personal information such as blog articles and emails are piling up quickly everyday. These have brought challenges for the effective and efficient organization of data. These have brought challenges for the effective and efficient organization of text documents. Clustering in general is an important and useful technique that automatically organizes a collection with a substantial number of data objects into a much smaller number of coherent groups. In the particular scenario of text documents, clustering has proven to be an effective approach for quite some time and an interesting research problem as well. It is becoming even more interesting and demanding with the development of the World Wide Web.

What is clustering?

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. We can show this with a simple graphical example:

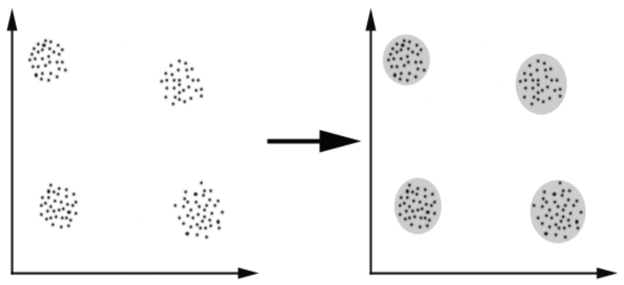


Figure 1: Graphical example of Clustering

In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance*:* two or more objects belong to the same cluster if they are "close" according to a given distance . This is called distance-based clustering. Another kind of clustering is conceptual clustering*:* two or more objects belong to the same cluster if this one defines a concept *common* to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

3. PROPOSED SYSTEM

In this section, we suggest some alternatives which will make it possible to exploit tag information even when the tag information in available for only a small number of WebPages. The drawback of above mentioned traditional web page clustering solved by web page tagging: How can tagging data be used to improve web document clustering? This is part of a major trend in information retrieval to make more and better use of the increasingly prevalent user-provided data. Social bookmarking websites such as del.icio.us and Stumble Upon enable users to tag any web page with short free-form text strings, collecting hundreds of thousands of keyword annotations per day. The set of tags applied to a document is an explicit set of keywords that users have found appropriate for categorizing that document within their own filing system. Thus tags promise a uniquely well suited source of information on the similarity between web documents. While others have argued that tags hold promise for ranked retrieval, including at least one approach that uses clustering, High quality clustering based on user-contributed tags has the potential to improve all of the previously stated applications of the cluster hypothesis, from user interfaces to topic-driven language models to increasing diversity of results.

Therefore such user generated content can provide useful information in various form such as meta-data, or in more explicit ways such as tags. User specified tags, in particular, have proven to be extremely effective in browsing, organizing, and indexing of webpage. Various social bookmarking websites such as StumbleUpon and Delicious allow users to tag webpages with keywords or short text snippets that can provide a description of the webpages. Users can collaboratively tag webpages and this has made organizing, sharing, navigating, and retrieving web content much easier than ever before. The aim to exploit the tag information for a webmining task, namely webpage clustering. Since user provided tags can often be very discriminative for webpages we want to exploit them by treating the tag information as an alternate *view* of the data. Motivated by the success of multi-view learning algorithms in various machine learning tasks, we use two views of the data to extract highly discriminative features and perform clustering using these features.

The feature extraction amounts to performing clustering in a lower dimensional subspace which is also effective in dealing with the problem of over fitting when we only have a small number of documents having a very large number of features. In particular, we use a regularized variant of the Kernel Canonical Correlation Analysis algorithm [9] to learn this subspace. has received tremendous attention due to its ability for effectively extracting useful features from heterogeneous or parallel data sources, such as images and text , or features and labels supervised dimensionality reduction. Therefore such an approach is expected to be useful for extracting useful features in the case of webpage clustering as well since the data often does have multiple views.

Advantages:

1) Tagging has made organizing, sharing, navigating, and retrieving web content much easier than ever before.
2) High quality clustering based on user-contributed tags has the potential to improve all of the previously stated applications of the cluster.
3) Reduced the searching time.

4. PROBABILISTIC LATENT SEMANTIC ANALYSIS

A] Latent Semantic Analysis

Latent semantic Analysis is to map documents and by symmetry terms to a vector space of reduced dimensionality, the latent semantic space. This mapping is computed by decomposing the term/document matrix N with SVD, $N = X\sum Y^t$, where X and Y are orthogonal matrices $X^tX = Y^tY = I$ and the diagonal matrix $\sum$ contains the singular values of N. The LSA approximation of N is computed by thresholding all but the largest K singular values in $\sum$ to zero ($=\sum$) which is rank K optimal in the sense of the L2-matrix norm as is well-known from linear algebra.

B] Geometry of the Aspect Model

Now consider the class-conditional multinomial distributions P(.|c) over the vocabulary in the aspect model which can be represented as points on the M-1 dimensional simplex of all possible multinomial's. Via its convex hull, this set of K points defines a K -1 dimensional sub-simplex. The modeling assumption expressed is that all conditional distributions P(.|d) are approximated by a multinomial represent able as a convex combination of the class-conditionals P(.|c). In this geometrical view, the mixing weights P(c|d) correspond exactly to the coordinates of a document in that sub-simplex. This demonstrates that despite of the discreteness of the latent variables introduced in the aspect model, a continuous latent space is obtained within the space of all multinomial distributions. Since the dimensionality of the sub-simplex is K- 1 as opposed to M -1 for the complete probability simplex, this can also be thought of in terms of dimensionality reduction and the sub-simplex can be identified with a probabilistic latent semantic space.

The core of PLSA is a statistical model which has been called a aspect model . The latter is a latent variable model for general co-occurrence data which associates an unobserved class variable $c \in C = \{c1 \ldots\ldots c_k\}$ with each observation, i.e., with each occurrence of a word $w \in W = \{w1 \ldots\ldots w_m\}$ in a document $d \in D=\{d1 \ldots\ldots d_N\}$. In terms of a generative model it can be defined in the following way:

- select a document d with probability P(d),
- pick a latent class c with probability P(c|d),
- generate a word w with probability P(w|c).

As a result one obtains an observed pair (d,w), while the latent class variable c is discarded. Translating this process into a joint probability model results in the expression

$$P(d,w) = P(d)P(w|d) \text{ , where} \qquad \text{(i)}$$

$$P(w|d) = \sum_{c\in C} P(w|c)P(c|d) \qquad \text{(ii)}$$

(ii) one has to sum over the possible choices of z which could have generated the observation. The aspect model is a statistical mixture model which is based on two independence assumptions: First, observation pairs (d,w) are assumed to be generated independently; this essentially corresponds to the 'bag{of{words' approach. Secondly, the conditional independence assumption is made that conditioned on the latent class z, words w are generated independently of the specific document identity d. Given that the number of states is smaller than the number of documents (K ‹‹ N), c acts as a bottleneck variable in predicting w conditioned on d. Notice that in contrast to document clustering models document {specific word distributions P(w|d) are obtained by a convex combination of the aspects or factors P(w|c). Documents are not assigned to clusters, they are characterized by a specific mixture of factors with weights P(c|d).These mixing weights over more modeling power and are conceptually very different from posterior probabilities in clustering models and (unsupervised) naive Bayes models. Following the likelihood principle, one determines P(d), P(c|d), and P(w|c) by maximization of the log likelihood function

$$\ell = \sum_{d\in D} \sum_{w\in W} n(d,w) \log P(w,d) \qquad \text{(iii)}$$

where n(d,w) denotes the term frequency, i.e., the number of times w occurred in d. This is a symmetric model P(z|d) with the help of Bayes' rule, which results in

$$P(d|w) = \sum_{c\in C} P(c)P(w|c)P(d|c) \qquad \text{(iv)}$$

(iv) is a re-parameterized version of the generative model that described by (i), (ii).

5. SYSTEM ARCHITECTURE

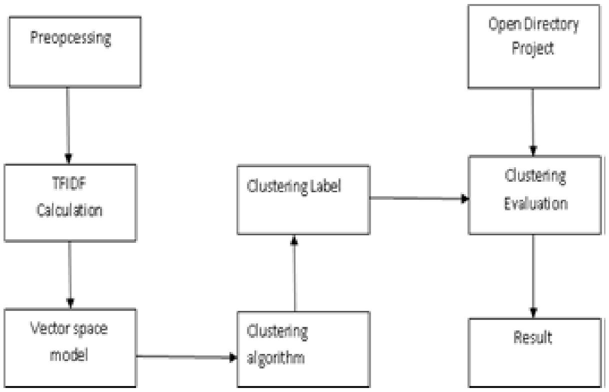


Fig.1. System Architecture Diagram

1. Preprocessing

Retrieving the document snippets from Google and parsing and stemming the results. Delete HTML tags, non-letter characters such as "$", "%" or "#". For example, the words: connected, connecting, interconnection should be transformed to the word connect. In third step Clear the Stop words Natural candidates for stop words are articles (e.g. "the"), prepositions (e.g. "for", "of") and pronouns (e.g. "I", "his").

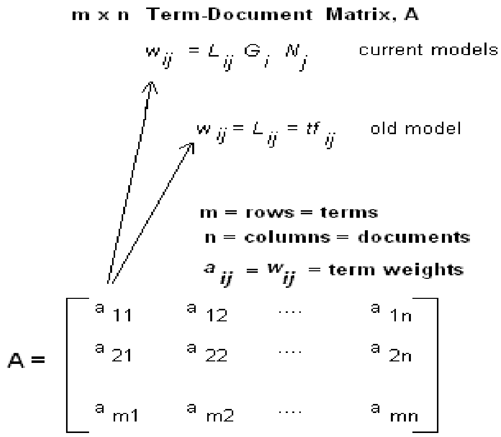
2. TFIDF Calculation

This assigns to term i a weight in document j given by
TF IDFi,j = TFi,j * IDFi

3. Annotation based Probabilistic LSA

Assume that we are given two sets of WebPages - one set T is tagged and the other set U is non-tagged. Further, |T | «|U|, and N = |T |+|U| is the total number of WebPages. The goal is to obtain a clustering of all N WebPages. We define the following:

A = document-word co-occurrence matrix (bag-of-words representation) of size N ×|W| where N is the number of documents (WebPages) in the corpus, and |W| is the page-text vocabulary size. Aij denotes the frequency of the word j appearing in document i. Note that the document-word co-occurrence matrix is constructed using both tagged and non-tagged WebPages. B = tag-word co-occurrence matrix (bag-of-words representation) of size |T| × |W| where |T| is the total number of tags in the corpus, and |W| is the page-text vocabulary size. Bij denotes the number of times tag i is associated with word j. Having constructed the document-word and word-tag co- occurrence matrices A and B, *joint* PLSA can be applied using A and B .


m x n Term-Document Matrix, A
$w_{ij} = L_{ij} G_i N_j$ current models
$w_{ij} = L_{ij} = tf_{ij}$ old model
m = rows = terms
n = columns = documents
$a_{ij} = w_{ij}$ = term weights
A =
a_{11} a_{12} .... a_{1n}
a_{21} a_{22} .... a_{2n}
a_{m1} a_{m2} .... a_{mn}

row= Tag

Columns= Documents

$$B = \begin{bmatrix} b_{11} & b_{12} & \ldots\ldots & b_{1n} \\ b_{21} & b_{22} & \ldots.. & b_{2n} \\ b_{m1} & b_{12} & \ldots... & b_{mn} \end{bmatrix}$$

4. K-means Clustering algorithm

1. Choose cluster centroids to coincide with K randomly selected documents from the document set.
2. Assign each document to its closest cluster.
3. Recomputed the cluster centroids using the current cluster memberships.
4. If there is a reassignment of documents to the new cluster, go to step 2. Typical stopping criteria is : Groups formed by the subsequent iterations must be

6. DATASETS

The dataset required for this project is a dataset that consist of user generated data. This is available on Delicious social bookmarking website. For evaluation purpose the directory structure that is required is available on the ODP(Open Directory Project) site. We used the bag-of-words representation for the feature vectors. Our approach can however also be applied with other feature representations such as the term-frequency/inverse-document-frequency (TF/IDF). A number of techniques have been proposed in the past to improve information retrieval tasks using auxiliary sources information, e.g., anchor text for web search interconnectivity of WebPages captions for image retrieval etc. Recent works on exploiting social annotations

7. CONCLUSION

Probabilistic Semantic Analysis has important theoretical advantages over standard LSA, since it is based on the likelihood principle, defines a generative data model, and directly minimizes word perplexity. Probabilistic Latent Semantic Analysis (PLSA) that has a solid statistical foundation, since it is based on the likelihood principle and defines a proper generative model of the data. we propose a new web page grouping approach based on Probabilistic Latent Semantic Analysis (PLSA) model. An iterative algorithm based on maximum likelihood principle is employed to overcome the aforementioned computational shortcoming.

8. REFERENCES

[1] Thomas Hofmann, “Probabilistic Latent Semantic Indexing”, Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval.

[2] Dempster, A., Laird, N., and Rubin, D. “Maximum likelihood from incomplete data via the EM algorithm.” J. Royal Statist. Soc. B 39 (1977), 138.

[3] Dumais, S. T. Latent semantic indexing”, Trec-3 report. In Proceedings of the Text Retrieval Conference (TREC-3) (1995), D. Harman, Ed., pp. 219.

[4] Gildea, D., and Hofmann, T. Topic-based language models using em. In Proceedings of the 6th European Conference on Speech Communication and Technology(EUROSPEECH) (1999).

[5] Hofmann, T. Latent class models for collaborative filtering. In Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI) (1999).

[6] Hofmann, T. Probabilistic latent semantic analysis. In Proceedings of the 15th Conference on Uncertainty in AI (1999).

[7] Hofmann, T., Puzicha, J., and Jordan, M. I. Unsupervised learning from dyadic data. In Advances in Neural Information Processing Systems (1999),vol. 11.

[8] Michael Tipping and Christopher Bishop. 1999. Probabilistic principal component analysis. Journal of the Royal Statistical Society, Series B, 61(3):611-622.

[9] Anusua Trivedi, Piyush Rai, Scott L. DuVall “Exploiting Tag and Word Correlations for Improved Webpage Clustering “SMUC’10, October 30,2010, Toronto, Ontario, Canada. Copyright 2010 ACM.

[10] http://www.stumbleupon.com

[11] http://www.delicious.com

[12] Open Directory Project (http://www.dmoz.org/)