Web Page Clustering using Latent Semantic Analysis

Mr. Lalit A. Patil
Department of Computer Engineering
KKWIEER,Nasik
lalitpatil58@gmail.com

Prof. S M. Kamalapur
Department of Computer Engineering
KKWIEER,Nasik
snehal_kamalapur@yahoo.com

Mr.Dhananjay Kanade
Department of Computer Engineering
KKWIEER,Nasik
dmkanade@yahoo.com

ABSTRACT

Web mining techniques such as clustering help to organize the web content into appropriate subject based categories so that their efficient search and retrieval becomes manageable. Traditional WebPages clustering typically uses only the page content (usually the page text) in an appropriate feature vector representation such as Bags of words, term-frequency /inverse document frequency ,etc. and then applies standard clustering algorithms(e.g. K-means, Suffix tree, Query directed clustering). For example, Users can provide captions for images on the internet, provide tags to WebPages and other media content they regularly browse on the internet, etc. Therefore such user – generated content can provide useful information in various form such as meta-data or in more explicit ways such as tags. Typically, WebPages clustering algorithms only use feature extracted from the page text. However, the advent also social –bookmaking websites, such as StumbleUpon and Delicious has led to a huge amount of user-generated content such as the information that is associated with the WebPages. In multi-view learning, the feature can be split into two subset alone is sufficient for learning. Here as for, unsupervised learning algorithms, multiple views of the data can often help in extracting better features. Canonical Correlation Analysis (CCA) is an unsupervised feature extraction technique for finding dependencies between two (or more) views of the data by maximizing the correlations between the views in a shared subspace. But the drawbacks of the CCA is it gives The first approach is based on an annotation based probabilistic latent semantic analysis (LSA) over document-word and tag-word co-occurrence matrices

General Terms

Latent Semantic Analysis, Web clustering

Keywords

Canonical Correlation Analysis , probabilistic latent semantic analysis, term-frequency, Web page clustering

1. INTRODUCTION

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 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 and the evolution of Web 2.0. For example, results returned by search engines are clustered to help users quickly identify and focus on the relevant set of results. Customer comments are clustered in many online stores, such as Amazon.com, to provide collaborative recommendations. In collaborative bookmarking or tagging, clusters of users that share certain traits are identified by their annotations.

Modern web search engines are tasked with returning the few most relevant results based on an often ambiguous user query and billions of web documents. Over ten years, ranking techniques harnessing link, anchor text, and user click-through data as well as simply page text have been developed to address this challenge. However, a major challenge is the inherent ambiguity of the user query. These queries are rarely more than a few words in length and may represent many potential information needs. One of the most promising approaches to handle this ambiguity is through automatic clustering of web pages.

2. WEB MINING:

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. There are basically three types of web data mining given below

A) **Web usage mining**:

Web usage mining is the process of extracting useful information from server logs i.e users history. Web usage mining is the process of finding out what users are looking for on the internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data.

B) Web structure mining:

Web structure mining is the process of using graph theory to analyze the node and connection structure of a web

site. According to the type of web structural data, web structure mining can be divided into two kinds:
1. Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location.
2. Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage.

C) Web content mining:

Mining, extraction and integration of useful data, information and knowledge from Web page contents.

2.1 Web page clustering

A) Web page clustering using Single View:

In single view web page clustering we just consider the single view means text of page. In which only single feature subset use.

B) Web page clustering using Multi View:

In multi-view learning, the features can be split into two subsets such that each subset alone is sufficient for learning. By exploiting both views of the data, multi-view learning can result in improved performance on various learning tasks, both supervised and unsupervised Multi-view approaches help supervised learning algorithms by being able to leverage unlabeled data whereas, for unsupervised learning algorithms, multiple views of the data can often help in extracting better features

A) Webpage Clustering Using word

Traditional webpage clustering typically uses only the page content information, just the page text in an appropriate feature vector representation such as Bag of Words.

B) Webpage Clustering Using Tag

There are number of websites they have user define tag. such as Delicious or StumbleUpon. So, tag also provide the more accurate information about the page. The goal is to obtain a clustering of the WebPages using tag that provide more relevant information as compare to just using Word.

C) Webpage Clustering Using Tag

However, in multi view learning we can combine the both word + tag and make clustering. As compare to previous two method it is gives better performance.

3. CANANONICAL CORRELATION ANALYSIS (CCA)

Canonical Correlation Analysis (CCA) is a technique for modeling the relationships between two (or more) set of variables. CCA computes a low-dimensional shared embedding of both sets of variables such that the correlations among the variables between the two sets are maximized in the embedded space. CCA has been applied with great success in the past on a variety of learning problems dealing with multi-modal data .

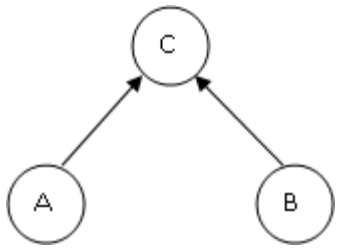

Fig. CCA view

In diagram 'A' represent features extracted from text-page subspace And 'B' represent features extracted from tag subspace and 'Z' represent common subspaces of both 'A' and 'B'.so ,that given a pair of datasets a $\in R^{D1\times N}$ and b $\in R^{D2\times N}$. CCA seeks to find linear projections wa $\in R^{D1}$ and wb $\in R^{D2}$ such that, after projecting, the corresponding examples in the two datasets are maximally correlated in the projected space. The CCA algorithm find the common coefficient between the two datasets in the following way

$$\rho=\frac{w_a^T AB^T w_b}{\sqrt{(w_a^T AA^T w_a)(w_b^T BB^T w_b)}}$$

the correlation is not affected by changing of the projections wa and wb, CCA is posed as a constrained optimization problem.

$$\max_{wa,wy} w_a^T AB^T w_b$$

$$w_b^T BB^T w_b = 1 \quad \text{and}$$
$$w_a^T AA^T w_a = 1$$

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. But there is some drawbacks of tagging web based clustering . So, how can resolve this problem?

4. DATASETS

Our dataset consists of a collection of 2000 tagged WebPages that we use for our webpage clustering task. All WebPages in our collection were downloaded from URLs that are present in both the Open Directory Project (ODP) web directory and Delicious social bookmarking website. The Delicious dataset of tags is available here: http://kmi.tugraz.at/staff/markus/datasets/each webpage that we crawled and downloaded was tagged by a number of users on Delicious. Therefore, for each webpage, we combine the tags assigned to it by all users who tagged that webpage.

5. PROPOSED SYSTEM

5.1 Need

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.

But there is some drawbacks of tagging web based clustering. So, how can resolve this problem?

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.

5.2 System Architecture:

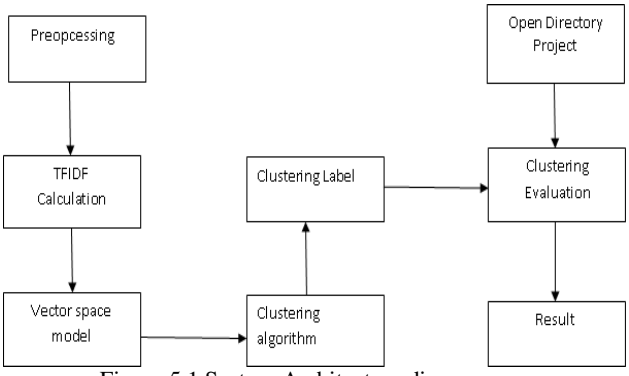

Figure 5.1 System Architecture diagram

Preparing the documents

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").

TFIDF Calculation

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

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. Recompute 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 same

5.3 Approach

What is LSA?

LSA (Latent Semantic Analysis) is a fully automatic mathematical/statistical technique for extracting and inferring relations of expected contextual usage of words in passages of discourse. It is not a traditional natural language processing or artificial intelligence program. It uses no humanly constructed dictionaries, knowledge bases, semantic networks, grammars, syntactic parsers, or morphologies, or the like, and takes as its input only raw text parsed into words defined as unique character strings and separated into meaningful passages or samples such as sentences or paragraphs.

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.

Note that the tag-word matrix is constructed using only the tagged WebPages Note that this is a more fine-granular association where we do not look for the associations between the tag and a webpage, but go a level further to consider the co-occurrences of tags with the actual words appearing in the WebPages (based on the tag-word co-occurrence matrix). Also, we would assume the same page-text vocabulary while constructing matrices A and B. To do this, we pool all N WebPages (with and without tag information) and construct a common vocabulary of size |W|. The vocabulary would not include tags (unless some tags, coincidentally, are words in some WebPages). Having constructed the document-word and word-tag co- occurrence matrices A and B, *joint* PLSA can be applied using A and B in a manner similar. A similar framework was applied in for the problem of clustering images on the social web using the image captions.

5.5 Applications

1) Medical Informatics, clustering patient records can be a difficult problem since these records often tend to be highly unstructured and noisy.
2) In web clustering
3) In conference paper classification

6. CONCLUSION

User generated content can be a very rich source of useful information for web-mining and information retrieval on the web. Often the usefulness of user-generated content is due to the fact that it is small but structured. In addition to being semantically precise, which can nicely complement the huge but unstructured information? Tag information can be exploited in numerous ways to improve webpage clustering, both when tags for available for all WebPages due to the discriminative information provided by the tags, the features extracted by our CCA based approach can also be useful for webpage classification. In the case when the tag information is available

only for a small subset of WebPages that condition the Latent Semantic Analysis is used to improve the performance.

7. REFERENCES

[1] 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.

[2] S. Poomagal, Dr. T. Hamsapriya, "K-means for Search Results clustering using URL and Tag contents "978-1-61284-764-1/11/$26.00 ©2011 IEEE.

[3] Lu, C., Chen, X., and Park, E. K. Exploit the tripartite network of social tagging for web clustering. In *CIKM '09 (2009), pp. 1545–1548.*

[4] Ramage, D., Heymann, P., Manning, C. D., and Garcia-Molina, H. Clustering the tagged web. In *WSDM '09 (2009)*

[5] Kakade, S. M., and Foster, D. P. Multi-view regression via canonical correlation analysis. In COLT'07 (2007)

[6] Ando, R. K., and Zhang, T. Two-view feature generation model for semi-supervised learning. In *ICML '07 (2007)*

[7] Bach, F. R., and Jordan, M. I. Kernel independent component analysis. *Journal of Machine Learning Research 3 (2003)*

[8] Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., and Su, Z. Optimizing web search using social annotations. In *WWW '07 (2007)*

[9] Bickel, S., and Scheffer, T. Multi-view clustering. In *ICDM '04 (Washington, DC, USA, 2004), IEEE* Computer Society,

[10] Blaschko, M. B., and Lampert, C. H. Correlational spectral clustering. In *CVPR (2008).*

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

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