Improved Web Page Clustering using Words and Tags

M.A. Shelke **K.C. Sadavarte** **R.K. Dhurjad** **N.P. Pandit**

ABSTRACT - Automatically clustering WebPages into semantically relevant classes, results into improved search and browsing on the web. Typically, webpage clustering algorithms only use features that are extracted from the web page-text. However, the advent of social-bookmarking websites like Stumble Upon and Delicious, has led to a huge amount of user-generated content such as the tag information that is associated with the WebPages. In our project, we use a subspace based feature extraction approach which leverages tag information to complement the page-contents of a webpage to extract highly discriminative features, with the goal of improved clustering performance. We consider page-text and users generated tags as two separate views of the data, and learn a shared subspace that maximizes the correlation between these two views. K-means clustering algorithm can then be applied using this subspace. We compare our subspace based approach with existing clustering method i.e. word only. And also we will show that the subspace based approach leads to improved performance on the webpage clustering task.

Keywords: Social Tagging, term-frequency, Webpage Clustering.

General Terms
K-means Algorithm, Web clustering

I. INTRODUCTION

Internet has become increasingly important as a medium for life, work and study as well as for dissemination of information. Information retrieval is common activity of internet users and for that purpose search engines are used. Modern search engines are tasked with returning the few most relevant results based on often ambiguous user query and billion of web documents. One of the common approaches to handle this ambiguity is through automatic clustering of web pages.

Web mining is the mining of data related to the World Wide Web. Clustering, a web mining techniques help to organize the web content into appropriate subject based

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. Traditional webpage clustering typically uses only features extracted from page-text in an appropriate vector representation such as bag of words, Term –Frequency/Inverse Document Frequency, etc. And then applies standard clustering algorithms. Other approach somewhat related to clustering is to mine topic information from documents collections (e.g., Latent Dirichlet Allocation), which can be seen as clustering words occurring in each document. Social bookmarking websites such as delicious and Stumble Upon enable users to tag any webpage with short free-form text strings, collecting hundreds of thousands of keyword annotations per day. The set of tags applied to document is an explicit set of keywords that users have found appropriate for categorizing that document. The increasing amounts of user generated content now a day nicely complements this information and can help in an effective mining of data present on the web. Therefore user generated content can provide useful information in various form such as meta-data, or in more explicit ways such as tags. User generated tags have proven to be extremely effective in browsing, organizing, and indexing of WebPages. So, the web mining task, namely webpage clustering will definitely improve by using these social tags.

The main question is how tag data can be used to improve web document clustering? Since user generated tags for web pages can often be very discriminative, we want to exploit them by treating the tag information as an alternate view of data.Multiview learning algorithm use to extract highly discriminative features and the standard clustering algorithm applied on these features. We use these two views of data (page text and tags) to extract features and perform clustering using these features. K-means algorithm is very popular vector space model for clustering which works iteratively by assigning each data point to its nearest cluster center, recomputing the cluster centers, and repeating the process until convergence. Here, we use K-mean algorithm, instead of it we can also use any vector space clustering algorithm.

II.WEBPAGE CLUSTERING

Clustering is one of the most important text mining methods that help users to navigate, summarize and organize documents. Web page clustering can be used, to organize the results returned by a search engine in a better

way. K-means clustering clusters the documents using the traditional vector space model and then it assigns labels to each cluster by identifying the frequent phrases. Among all the existing clustering algorithms, K-means algorithm performs very well since it does not assign any document to more than one cluster. It first sets the initial parameters needed for clustering. Initial parameters are set of centroids and the number of clusters (K).The distance from the centroids to all documents is assigned to the nearest cluster. The process continues until the convergence criterion is met.

Formally, for our clustering task, we are given a collection of N WebPages, with each webpage consisting of a bag of words from a word vocabulary W, and a bag of tags from a tag vocabulary T. The goal is to cluster the WebPages in K clusters where K is the desired number of clusters. There are number of ways in which the vector space algorithms such as K-means can exploit the tag information to improve clustering of WebPages. Some of the common choices are:

A. Clustering By Words Only

It will give clustering on basis of words only. These words also called as features will get by preprocessing of webpages.After extraction of features, the TF/IDF value for each feature will calculate, and represent in feature vector. And the feature vector will be given as input for K-mean algorithm. This option will work in all conditions, i.e. whether the page will tag or untagged. It is the most common approach for clustering until now.

B. Clustering By Tags Only

It will give clustering on basis of tags only. The user generated tags will be used as features. We will get tag information i.e. tag name and how many users gave the same tag (frequency) from Delicious datasets. These tags will represent as feature vector and will be used for further clustering process.

C. Clustering by Word + Tag Vector

User will get clustering on basis of both words and tags. System will preprocess WebPages, extract unique words from the input pages, and also extract the tag information for those pages from Delicious dataset. The correlation between word feature vector and tag feature vector will be get by applying CCA [Canonical Correlation Analysis]. The result of CCA will be given to Clustering algorithm and result will display to user .User will get improved results compare to other two approaches.

The goal is to obtain a clustering of the WebPages into semantically relevant categories. To assess the relevance and coherency of the discovered clusters, one can use hierarchical web directories such as the Open Directory Project (ODP) as the standard. The concatenation of word and tag feature vectors outperforms approaches that use feature vectors derived from the word vocabulary, the tag vocabulary, or vocabulary derived from a union of words and tags.

III.ODP [Open Directory Project]

ODP is free, user maintained hierarchical web directory. Each node in the ODP hierarchy has a label like “Arts” “Business”, etc and a set of related documents. A gold standard clustering using ODP is thus defined by a particular node’s k’ children. When we give the clustering algorithm a value K, this is equal to k’ children of selected node.

IV.MULTI VIEW LEARNING

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. 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. This property makes CCA a suitable choice for multi-view learning algorithms. In our settings, the two views are words in the page-text, and the set of tags for each webpage. CCA is then applied as a projection technique to extract features from webpage data, with projection direction guided by the tag information. Final clustering is then performed using the features extracted by CCA.

A. Canonical Correlation Analysis

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.

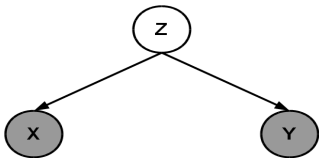

Figure 1: The dependency view of CCA: datasets X and Y, and their shared subspace defined by Z. In our webpage clustering, X corresponds to the features extracted from the page-text and Y corresponds to the features extracted from the tags. Z represents the semantic subspace shared by both words and tags.

More formally, given a pair of datasets $X \in R^{D1\times N}$ and $Y \in R^{D2\times N}$. CCA seeks to find linear projections $w_x \in R^{D1}$ and $w_y \in R^{D2}$ such that, after projecting, the corresponding examples in the two datasets are maximally correlated in the projected space. The correlation coefficient between the two datasets in the embedded space is given by

$$\rho = \frac{w_x^T X Y^T w_y}{\sqrt{(w_x^T X X^T w_x)(w_y^T Y Y^T w_y)}}$$

Since the correlation is not affected by rescaling of the projections w_x and w_y, CCA is posed as a constrained optimization problem

Subject to:

$$\max_{w_x, w_y} w_x^T X Y^T w_y$$

$$w_x^T X X^T w_x = 1, w_y^T Y Y^T w_y = 1$$

It can be shown that the above formulation is equivalent to solving the following generalized eigen-value problem:

$$\begin{pmatrix} 0 & \Sigma_{xy} \\ \Sigma_{yx} & 0 \end{pmatrix} \begin{pmatrix} w_x \\ w_y \end{pmatrix} = \lambda \begin{pmatrix} \Sigma_{xx} & 0 \\ 0 & \Sigma_{yy} \end{pmatrix} \begin{pmatrix} w_x \\ w_y \end{pmatrix}$$

Where Σ_{xx} and Σ_{yy} denotes the covariance's of data samples $X = [x_1, \ldots, x_n]$ and $Y = [y_1, \ldots, y_n]$ respectively, and Σ_{xy} denotes the cross-covariance between X and Y

V. ALGORITHM

Kmeans is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

The algorithm is:

1: The number of clusters is provided as input: k parameter.

2: K points are randomly selected. They will be the first "cluster centers": $C_1, \ldots\ldots, C_k$

3: The algorithm will place each instance in the cluster to which centre is closest according to the distance measure used.

4: After each instance is distributed, for each cluster, the cluster "centroid" is recalculated using all the instances that are now part of that cluster.

5: The new centre of the cluster is calculated.

6: Repeat step 2, with the new centers. The clustering process is stopped when the same instance is repeatedly distributed to the same cluster. At this point we say the clusters are stable.

VI. DATASETS

Our tagged document collection is a subset of the Stanford Tag Crawl Dataset. The Tag Crawl consists of one contiguous month of the recent feed on delicious, a popular social bookmarking website. Our dataset consists of a collection of 2000 tagged web pages that we use for our webpage clustering task. All web pages in our collection were downloaded from URLs that are present in both the Open Directory Project (ODP) web directory (so that their ground-truth clustering are available) and Delicious social bookmarking website (so that their tag information is available).

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. 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).

VII. EXPERIMENTAL EVALUATION

To assess the efficiency of the inclusion of tag information for webpage clustering, we compare the following approaches in our experiments:

1. **Word feature vector only**: For this, we only consider the words appearing in the WebPages. We construct feature vector for each webpage using the bag of words representation, using the words extracted from the page-text.

2. **Tag feature vector only**: For this, we only consider the tags associated with each webpage, and construct feature vector for each webpage using the bag of tags representation. The tag set for each webpage consists of the tags applied to it by *all* users in the Delicious dataset.

3. **Word feature vector + Tag feature vector**: For this, we created an augmented feature vector by con-catena ting the tag feature vector with the word feature vector and normalized appropriately.

VIII. 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, this can nicely complement the huge but unstructured information. Tag information can be exploited in numerous ways to improve webpage clustering, both when tags available for all web pages 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.

ACKNOWLEDGMENT

We would like to thank our project supervisor, *Prof. Snehal Kamalapur*. For highlighting the idea of *Improved Webpage Clustering Using Words and Tags and her* able guidance & enlightening comments throughout the project work. We would also like to thank her for her helpful suggestions & numerous discussions.

We gladly take this opportunity to thank **Prof. Dr. S. S. Sane**, Head of Dept., Computer Engineering, for providing facilities during progress of the project.

We are thankful to all those who helped us directly or indirectly.

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

[13]http://www.demoz.com

[14] Foster, D. P., Kakade, S. M., and Zhang, T. 2008. Multi-view dimensionality reduction via canonical correlation analysis. *Technical Report TTI-TR-2008-4*. Gestel, T. V., Suykens, J. A. K., Brabanter, J. D., Moor, B. D

AUTHOR'S PROFILE

Miss. Kavita Sadavarte

She is student of BE Computer Engineering in Pune University.She is implementing this paper as her academic project.She is interested in Data Mining.

kkavita43@gmail.com

Miss.Reshma Dhurjad

She is student of BE Computer Engineering in Pune University. She is implementing this paper as her academic project. She is interested in Neural Network, Artificial Intelligence.

rdhurjad@gmail.com

Miss.Nilima Pandit

She is student of BE Computer Engineering in Pune University. She is implementing this paper as her academic project. She is interested in Statistics.

neelimapanditp@gmail.com

Miss.Megha Shelke

She is student of BE Computer Engineering in Pune University. She is implementing this paper as her academic project. She is interested in Web Mining.

shelkemegha@gmail.com