International Journal of Modern Engineering Research (IJMER)
www.ijmer.com Vol.3, Issue.3, May-June. 2013 pp-1386-1388 ISSN: 2249-6645

A General Framework for Building Applications with Short and Sparse Documents

Syed Jani Basha, [1] Sayeed Yasin[2]
[1]*M.Tech, Nimra College of Engineering & Technology, Vijayawada, A.P., India*
[2]*Asst.Professor, Dept.of CSE, Nimra College of Engineering & Technology, Vijayawada, A.P., India*

Abstract: *with the explosion of e-commerce and online communication and publishing, texts become available in a variety of genres like Web search snippets, forum and chat messages, blogs, book and movie summaries, product descriptions, and customer reviews. Successfully processing them, therefore, becomes increasingly important in many Web applications. However, matching, classifying, and clustering these sorts of text and Web data pose new challenges. Unlike normal documents, these text and Web segments are usually noisier, less topic-focused, and much shorter, that is, they consist of from a dozen words to a few sentences. Because of the short length, they do not provide enough word co- occurrence or shared context for a good similarity measure. Therefore, normal machine learning methods usually fail to achieve the desire accuracy due to the data sparseness. To deal with these problems, we present a general framework that can discover the semantic relatedness between Web pages and ads by analyzing implicit or hidden topics for them.*

Keywords: *Classification, Clustering, Hidden topic, Sparse.*

I. INTRODUCTION

In contextual advertising, ad- messages are delivered based on the content of the Web pages that users are surfing. It can therefore provide Internet users with the information they are interested in and allow advertisers to reach their target customers in a non-intrusive way [1] [2]. In order to suggest the "right" ad- messages, contextual ad matching and ranking techniques are needed to be used. This has posed new challenges to the Web mining and IR researcher. Firstly, as words can have multiple meanings and some words in the target page are not important, they can lead to mismatch in the lexicon-based matching method. Moreover, a target page and an ad can still be a good match when they share no common terms or words but belong to the same topic.

To deal with these problems, we present a general framework that can discover the semantic relatedness between Web pages and ads by analyzing implicit or hidden topics for them. After that, both Web pages and the advertisements are expanded with their most relevant topics, which helps reduce the sparseness and make the data more topic-focused. The framework can therefore overcome the limitation of word choices, deal with a wide range of Web pages and ads, as well as processes future data, that is, previously unseen ads and Web pages, better. It is also easy to implement and general enough to be applied in different domains of advertising and in also different languages.

II. RELATED WORK

"Text categorization by boosting automatically extracted concepts" by Cai & Hoftmann in [3] is probably the study most related to our framework. This attempts to analyze topics from data using pLSA and uses both the original data and resulting topics to train two different weak classifiers for boosting. The difference is that they extracted topics only from the training and test data while we discover hidden topics from the external large-scale data collections. In addition, we aim at dealing with the short and sparse text and Web segments rather than normal text documents. Another related work is the use of topic features to improve the word sense disambiguation by Cai et al. [4].

In [5], the author Bollegala use search engines to get the semantic relatedness between words. Sahami & Heilman [8] also measure the relatedness between text snippets by using search engines and a similarity kernel function. Metzeler et al. [6] evaluated a wide range of similarity measures for short queries fromWeb search logs. Yih & Meek [7] considered this problem by improving Web-relevance similarity and the method in [8]. Gabrilovich & Markovitch [9] computing semantic relatedness for texts using Wikipedia concepts. Prior to recent topic analysis models, word clustering algorithms were introduced to improve text categorization in various different ways. Baker & McCallum [10] attempted to reduce dimensionality by class distribution-based clustering.

Bekkerman et al. [11] combined distributional clustering of words and SVMs. And Dhillon & Modha [12] introduced spherical *k*-means for clustering sparse text data. Clustering Web search has been becoming an active research topic during the past decade. Many clustering techniques were proposed to place search results into topic– oriented clusters [13].

III. PROPOSED FRAMEWORK

The proposed work consists of the document classification and the online contextual advertising. The first and foremost step is to analyze the hidden topics based on the semantic similarity. Once the topics are analyzed, then the classifier is built upon the hidden topics by integrating them with the available training data. For advertising, the web pages and the page ads will be matched and ranked based on their similarity.

A. Analysis with the Hidden Topics

Latent Dirichlet Allocation [14] [15] is a method to perform the latent (hidden) semantic analysis (LSA) to find the latent structure of topics and concepts in a text corpus. LSA is well known technique which partially addresses the synonymy and the polysemy issues. LDA is a probabilistic model for collection of discrete data and has been used in the text classification. The Latent Dirichlet Allocation (LDA) is similar to the Latent Semantic analysis (LSA) and Probabilistic LSA (pLSA), since they share some common assumptions such as, the documents having semantic structure, can infer topics from word-document and its co-occurences and the words related to the topic. In this classification of hidden topics process, the universal data set is collected and the topic analysis is done and then the training set data and the test set data are separated and then the training is performed on this set of data so that when the new data is inserted, then it could classify the given data under a specific domain or category.

B. Building Classifier with the Hidden Topics

Now- a- days, the continuous development of Internet has created a huge amount of documents which are difficult to manage, organize and navigate. As a result, the task of automatic classification, which is to categorize textual documents in to two or more predefined classes, has been received a lot of attentions. Several machine learning methods have been applied to text classification including decision trees, neural networks, support vector machines, etc. In the typical applications of machine learning methods, the training data is passed to a learning phrase. The result of the learning step is an appropriate classifier, which is capable of categorizing new documents. However, in the cases such as the training data is not as much as expected or the data to be classified is rare, learning with only training data can not provide us a satisfactory classifier. Inspired by this fact, we propose a general framework that enables us to enrich both training and new coming data with hidden topics from available large dataset so as to enhance the performance of text classification.

Classification with hidden topics is described in Figure 1. We first collect a very large external data collection called universal dataset. Next, a topic analysis technique such as pLSA, LDA, etc. is applied to the data set. The result of this step is an estimated topic model which consists of the hidden topics and the probability distributions of words over these topics. Upon this model, we can do topic inference for training dataset and the new data. For each document, the output of topic inference is a probability distribution of the hidden topics – the topics analyzed in the estimation phrase – given the document. The topic distributions of the training dataset are then combined with training dataset itself for learning classifier. In the similar way, the new documents, which need to be classified, are combined with their topic distributions to create the so called "new data with hidden topics" before passing to the learned classifier.

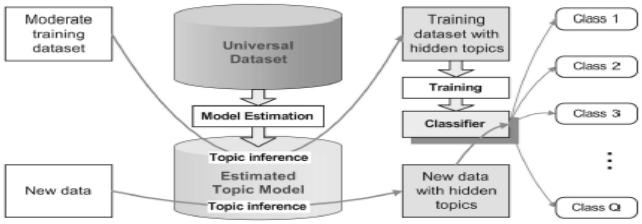

Figure 1: Classification with Hidden Topics

C. Building Clustering with the Hidden Topics

Text clustering is to automatically generate groups or clusters of documents based on the similarity or distance among documents. Unlike Classification, in clustering, the clusters are not known previously. Users can optionally give the requirement about the number of clusters. The documents will later be organized in to clusters, each of which contains "close" documents. Web clustering, which is a type of text clustering specific for the web pages, can be offline or online. Offline clustering means, it is to cluster the whole storage of available web documents and does not have the constraint of response time. In online clustering, the algorithms need to meet the real-time condition, i.e. the system need to perform clustering as fast as possible. For example, the algorithm should take the document snippets instead of the whole documents as input since the downloading of the original documents is time-consuming. The question here is how to enhance the quality of clustering for such document snippets in "online web clustering". Inspired by the fact those snippets are only small pieces of text (and thus poor in content) we propose the general framework to enrich them with hidden topics for clustering (Figure 2).

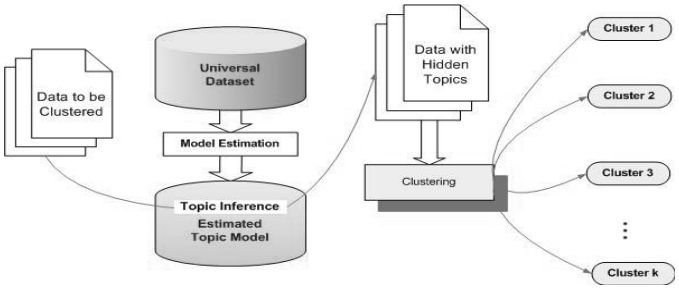

Figure 2: Clustering with Hidden Topics

D. Matching and Ranking of Contextual Advertisements

In matching and ranking of ads with the hidden topics, web pages ands the ads are matched based on their similarity. The similarities between those are measured using the cosine similarity. The ad- messages are arranged based on their similarity for each page. The keywords are also taken into consideration for ranking of the ads. The web pages and ad messages are considered and the topic inference is carried out for the both to identify under which category the web pages and the ad messages fall. The topic inference is very similar to the training process. Once the inference is done, then the new set of web pages and the ad messages are taken and then a contextual matching of those is done. The similarity is measured based on the context of the web pages and with ad messages. After identifying the contextual similarity, it is measured using the cosine similarity method, where the ranking process is done based on the similarity measure value. The web page related to the keyword that has the highest similarity value is ranked highest and given more preference while displaying the web search results. The similarity of the web page "*p"* and ads "*a"* is defined as follows:

sim_{AD} (p, a)=similarity (p, a) &

sim_{AD_kw} (p, a)=similarity (p, a U KWs)

Where KW is a set of keywords associated with the ad message "a".

IV. CONCLUSION

The proposed frame work presents a general framework for building classifiers that deal with short and sparse text & Web segments by making the most of hidden topics discovered from large scale data collections. The main motivation of this frame work is that many classification tasks working with short segments of text & Web, such as search snippets, forum & chat messages, blog & news feeds, product reviews, and book & movie summaries, fail to achieve high accuracy due to the data sparseness. We, therefore, come up with an idea of gaining external knowledge to make the data more related as well as expand the coverage of the classifiers to handle future data better. The underlying idea of the general framework is that for each classification task, we collect a large-scale external data collection called "universal dataset", and then build a classifier on both a (small) set of labeled training data and a rich set of hidden topics discovered from that data collection. The framework is general enough to be applied to different data domains and genres ranging from Web search results to the medical text. The advantages of the general framework are:

- The general framework is flexible and general enough to apply in any domain/language. Once we have trained a universal dataset, its hidden topics could be useful for several learning the tasks in the same domain.
- This is particularly useful in sparse data mining. Spare data like snippets returned from a search engine could be enriched with the hidden topics. Thus, enhanced performance can be achieved with this.
- Due to learning with the smaller data, the presented methods require less computational resources than semi-supervised learning.

REFERENCES

[1] P. Chatterjee, D. L. Hoffman, and T. P. Novak. Modeling the clickstream: Implications for web-based advertising efforts Marketing Science, 22(4):520–541, 2003
[2] R. Wang, P. Zhang, and M. Eredita. Understanding consumers attitude toward advertising. Proc. AMCIS, 2002.
[3] L. CAI and T. Hofmann. Text categorization by boosting automatically extracted concepts. Proc. ACM SIGIR, 2003.
[4] J. CAI, W. Lee, and Y. Teh. Improving WSD using topic features. Proc. EMNLP-CoNLL, 2007.
[5] D. Bollegala, Y. Matsuo, and M. Ishizuka. Measuring semantic similarity between words using Web search engines. Proc. WWW, 2007
[6] D. Metzler, S. Dumais, and C. Meek. Similarity measures for short segments of text. Proc. ECIR, 2007.
[7] W. Yih and C. Meek. Improving similarity measures for short segments of text. Proc. AAAI, 2007.
[8] M. Sahami and T. Heilman. A Webɲbased kernel function for measuring the similarity of short text snippets. Proc. WWW, 2006.
[9] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using Wikipedia-based explicit semantic analysis Proc. IJCAI, 2007.
[10] L. Baker and A. McCallum. Distributional clustering of words for text classification Proc. ACM SIGIR, 1998
[11] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter Distributional word clusters vs. words for text categorization. JMLR, 3:1183–1208, 2003.
[12] I. Dhillon and D. Modha. Concept decompositions for large sparse text data using clustering Machine Learning, 29(2–3):103–130, 2001.
[13] K. Kummamuru, R. Lotlikar, S. Roy, K. Singal, and R. Krishnapura A hierarchical monothetic document clustering algorithm for summarization and browsing search results Proc WWW, 2004
[14] Andrieu, C., Freitas, N.D., Doucet, A. and M.I. Jordan (2003), "An Introduction to MCMC for Machine Learning", Machine Learning Journal, pp. 5- 43.
[15] Bhattacharya, I. and Getoor, L. (2006), "A Latent Dirichlet Allocation Model for Entity Resolution", In Proceedings of 6th SIAM Conference on Data Mining, Maryland, USA.