ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 3, March 2013

A Framework for Building Applications Based on Hidden Topics with Short and Sparse Web Documents

Kanimozhiveena E, D. Ramya Dorai

***Abstract* – The main aim of this paper is to provide an approach for resolving two major issues in the web such as (1) data sparseness and (2) synonymy of the data. This paper provides a model that could reduce the data sparseness and the synonymy issues. To attain this objective, here the external data from users is taken. This external data helps to reduce both the mentioned issues. The external data is taken into consideration along with the dataset to reduce the data sparseness. It is because if a document that has more relevant content in it but, with very few sentences present in it, related to the keyword given in the query space, then the classification is not likely to be done perfectly. In this case, to classify such sparse and short documents more accurately, we use external data where the document may contain very few sentences and very fewer keywords present it and then enhance classification. In advertising, the ad messages and web pages are considered. Semantic similarity is measured between the ad messages and the web pages for their matching and ranking.**

Index terms – classification, data sparseness, matching/ranking, text categorization, semantic similarity, web mining

I. INTRODUCTION

With the huge amount of information available online, the World Wide Web is a fertile area for data mining research. The World Wide Web is a collection of electronic documents whose size is growing exponentially. This makes it difficult for users to find useful information. The Web provides enormous resource for many informational topics but does not provide a good means to find the information relevant to specific user interests. Now-a-days web has turned to be the largest information source available. In today's society there is an increasing need for automated systems providing personalized recommendations to a user faced with a large number of choices. Automated methods are needed to provide a large number of users with the ability to efficiently locate and retrieve information according to their preferences. The task of collaborative filtering is to predict preferences of an active user on unseen items, given preferences of other users, typically expressed as numerical ratings.

Manuscript received Feb, 2013.

Kanimozhiveena E, *Department of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur, TamilNadu, India.*

D. Ramya Dorai, *Department of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur, TamilNadu, India.*

Web mining is a very huge research topic which combines two research areas i.e. Data mining and World Wide Web. This web mining research relates to several research communities such as Database, information retrieval and artificial intelligence. The term of applying data mining techniques to automatically discover and extract useful information from the World Wide Web documents and services is generally said to Web mining. It is the technology that analyzes less structured data having the goal to extract knowledge from web log files. It is capable of finding, extracting and structuring information related to a particular domain from the web documents.

In this paper, classification and clustering plays a major role in reducing data sparseness. *Classification* consists of building a model for each given class based features in the web log data and generating classification rules from such models. The models are constructed by analysing a training web log data set whose class label is known. The classification rules can be used to develop a better understanding of each class in the web database and also customize answers to user requests based on the classes of requests. *Clustering* is the process of creating partition so that all the members of each set of the partition are similar according to some metric. In web usage mining, it is said as a technique to group information into clusters based on their common characteristics. The latent dirichlet allocation is used for the text categorization or classification which enables more hidden topics to be classified appropriately. The maximum entropy classifier is used to build the classifier based on the text categorization performed on the training data. The cosine similarity is used to match and rank the web pages and ads by identifying the similarity of the web pages and the ad messages based on the keyword used for searching the information.

The two main challenges posed in this paper is (1) data sparseness (2) synonymy in the documents. The short and sparse documents are those that are usually noisier, less topic focused and have only few words or sentences in it. Due to the short length, it does not provide enough word co-occurrence for a good similarity measure. Another issue is synonymy where, two or more words in the document have the same meaning. Thus sometimes, unintentionally the advertisements may be put in a different web page that may not be much relevant to it. This becomes the main issue for problems in classification, clustering and matching the ad messages with the relevant web pages.

II. RELATED WORKS

There have been a considerable number of related studies that focused on short and sparse data to find a suitable method of representation for the data in order to get a better classification, clustering and matching performance. In this section, a short introduction of related studies is given that are relevant to this work.

Previous works focused on similarity between the short texts. Bollegala et al. [6] used web search engines for measuring the semantic similarity between words. In document classification, the target concept is class label. Thus the word similarity is measured by the distributions of class labels associated with the words in question. So, Baker and McCallum [2] used distribution based clustering to measure the word similarity for text classification which will sensibly cluster words that are indicative more than one class. Text categorization is carried out for classifying the documents into fixed number of predefined categories with SVMs by T.Joachims [8].

Sahami and Heilman [11] measured the similarity between the short text snippets by a web-based kernel function. In many machine learning settings, unlabeled examples are significantly easier to come by than labeled ones. To train a system to automatically classify the web pages, one would typically rely on hand labeled web pages. Thus, Blum and Mitchell [5] combined both labeled and unlabeled data with the Co-Training algorithm.

Subscribers to the popular news or blog feeds often face the problem of information overhead as these feed sources usually deliver a large number of items periodically. One way to deal with this problem is clustering. To improve accuracy of clustering short texts is enriching their representation with additional features from wikipedia. Banerjee, Ramanathan and Gupta [3] used this method for clustering short texts. Zeng et al. proposed ranking algorithm, salient phrases extraction and their ranking for clustering the web search results. Hoffman [9] used the Probabilistic latent semantic analysis for the analysis of two-mode and co-occurrence data. Cai et al. [7] used the Latent dirichlet allocation method to improve the word sense disambiguation.

III. OUTLINE OF THE WORK

A general framework for building applications on short web documents can be done by utilizing external data. The external data could help in the better classification of future unseen data. The overall work of this paper is to enhance the classification accuracy even if the document has sparse and short data in it. To achieve this, a framework is built where even if the data which is given as input has few keywords and fewer sentences related to the search query, it should be able to classify it and cluster it into a specified category. These categories are determined by a prior training which is done on the basis of the previously available data. In advertising, the main issue is that we need to put right ad messages in the right web pages, if this is not done, then the ad messages will be irrelevant to the web pages. This is done in order to attract the users' attention. The web pages and ad messages are trained with the already available data based on the keywords present in them. So that, when a new ad message is given, for it to be placed in the appropriate web page, the keyword of the ad message and the web page are verified for their similarity matching. If it matches then, the ad messages could be placed in that specific web page. Then based on the keyword match, the ad messages are given priority and ranked in the web page. Thus, the classification could be enhanced by reducing the classification error and the ad messages could be placed in the appropriate web pages for the users' could easily view them.

A. Classification with Hidden Topics

Given a small training data set **D**={ $(d_1,c_1,),(d_2,c_2),\ldots,(d_n,c_n)$ } that consists of n short and sparse documents d_i and their class labels $c_i(i=1\ldots n)$; and **W**={$w_1,w_2,\ldots,w_m$} be a large scale data collection containing m unlabeled documents $w_i(i=1\ldots m)$. This approach provides a framework to gain additional knowledge from **W** in terms of hidden topics to modify and enrich the training set **D** in order to build a better classification model. Here, **W** is universal data set since it is large and diverse enough to cover a lot of information regarding the classification task. The classification of hidden topics model is given in Figure 1.

The universal data set should be large enough that it should cover a large amount of words, topics and concepts and should also be consistent that even in the future the classifier should be able to work with.

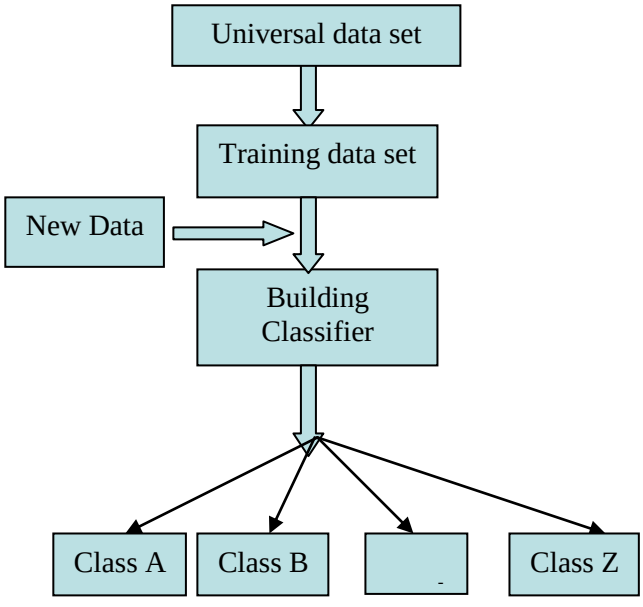

Figure 1 General workflow to classify short and sparse text with hidden topics

The topic analysis is done using a hidden topic analysis model called Latent dirichlet allocation because; it has more document generation assumption. For training data, only a moderate size labeled data is to be considered to avoid time consumption when using a large amount of data. Topic inference for training the data depends on the technique for training the classifier. The Maximum entropy method is used to build the classifier.

B. Matching/Ranking of Advertisements with Hidden Topics

Given a set of *n* target Web pages **P**= {p_1, p2,…,p_n}, and a set of *m* ad messages (ads) $\mathbf{A_i}$={a_{i1},a_{i2},…,a_{im}}. For each web page p_i, find a corresponding ranking in the list of ads $\mathbf{A_i}$, i=1…n such that more relevant ads will be placed higher in the list. These ads are ranked based on their relevance to the target page and the keyword information. The matching and ranking of page ads with hidden topic is given in Figure 2.

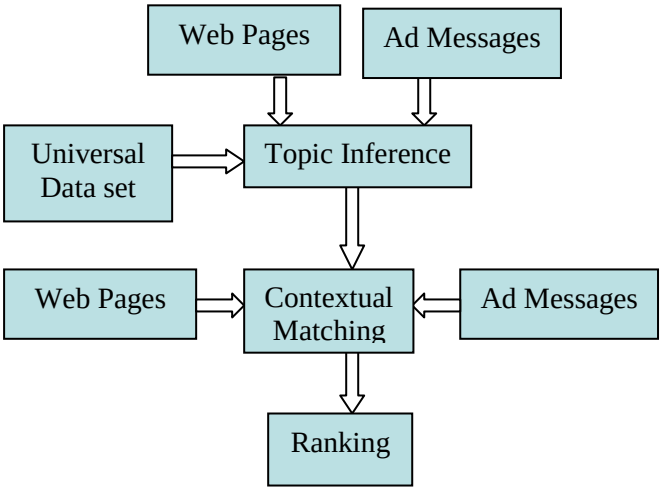

Figure 2 Workflow model for page-ad matching and ranking with hidden topics

An external document collection should be done such that a large words, topics and domains of web pages and ads are covered. The hidden topic analysis will yield an estimated topic model that includes hidden topics discovered from universal data set and the distribution of topics over the terms. The topic inference is to be done for both the web pages and ads to discover the meanings and the topic focus. This information will be integrated into corresponding web pages or ads for matching and ranking them based on their relevance.

IV. PROPOSED APPROACH

The proposed work consists of the document classification and online contextual advertising. The first and foremost step is to analyse the hidden topics based on the semantic similarity. Once the topics are analysed, 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.

1. Analysis with the Hidden Topics

Latent Dirichlet Allocation [2], [4] 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 polysemy and the synonymy issues. LDA is a probabilistic model for collection of discrete data and has been used in text classification. The Latent Dirichlet Allocation (LDA) is similar to 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, it could classify the given data under a specific domain or category.

2. Building Classifier with the Hidden Topics

After topic analysis is performed on the universal data set, then the hidden topics has to be integrated with it for training. For building the classifier, Maximum Entropy (MaxEnt) [10] method is used. It is used to estimate the probability distribution from the available data. The maximum entropy is used for text classification by estimating the conditional distribution of the class variable given the document. Maximum Entropy utilizes the word count of the document for text classification. The labeled training data is used to estimate the expected value of these word counts on a class-by-class basis. It even performs better than SVMs for classifying sparse data. After the data is trained, the new set of data is given to build the classifier. The classifier is built for each and every specific category and domain. To check the classifier accuracy, the test data set are given as input and check whether the classifier built classifies it under its specific domain with the help of the keyword of the document.

3. Matching and Ranking of Contextual Advertisements

In matching and ranking of ads with the hidden topics, web pages ands ads are matched based on their similarity. The similarities between those are measured using cosine similarity [1]. The ad messages are arranged based on their similarity for each page. The keywords are also taken into consideration for ranking 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 page and the ad messages fall. The topic inference is similar to the training process. Once the inference is done, then the new set of web pages and 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 ad messages. After identifying the contextual similarity, it is measured using the cosine similarity method, where the ranking 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 priority while displaying the web search results. The similarity of the web page *p* and ads *a* is defined as follows:

$$sim_{AD}(p, a)=similarity(p, a) \;\&$$
$$sim_{AD\_KW}(p, a)=similarity(p, a \cup KWs)$$

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

V. EVALUATION RESULTS

This work is implemented using java, with a web framework where it behaves like a search engine and the search documents could be viewed when a search query is given in the search field. The web framework developed consists of two parts namely for short texts and contextual advertisements. For the experimental data, we consider training dataset and the test dataset under various domains for both. The training dataset consists of already available predefined data in it and the test data are given by the user to check the classification accuracy. The framework is built in such a way that even if the test data has a very few sentences or words, it would be able to classify it under the domain that it belongs to. The training dataset and the test dataset are totally exclusive. When a new data, whish is said to the test data is entered, it automatically classifies it in its recommended category and could be viewed by the user under that classifier label.

Figure 3 The short texts classified under the categories predefined

For the experimental data in advertisements matching and ranking, a set of web pages and the ad messages are provided. The web pages and the ad messages contain the description about them respectively. The keywords in the ad messages are mapped with the keywords of the web pages and if the set of keywords are matched, then the ad messages are placed in that particular web page.

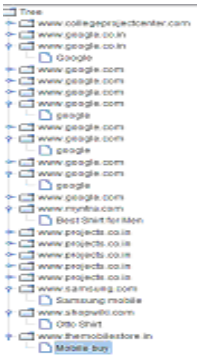

Figure 4 Matching the ad messages into corresponding web pages category

VI. CONCLUSION

The above two general approaches aims at, (1) improving the classification accuracy for short text/web documents, (2) improving the matching/ranking performance for online contextual advertising. Improvising the classification accuracy and the matching/ranking of documents is done by taking the advantage of the hidden topics from the external data. This approach provides a way to make sparse documents more related and topic focused by performing topic inference with a rich source of information about the topics, concepts, words and terms. To overcome the issues such as synonymy and vocabulary mismatch, the integration of hidden topics is done for obtaining a better classification and matching/ranking over the topics.

FUTURE ENHANCEMENT

The future studies could be focused on improving the framework in a number of ways such as focusing more on homonyms, find more fine-grained topic analysis to meet more sophisticated data and applications, could pay more attention to the consistency between the universal data set and the data we need to work with and incorporate the keyword information into ad ranking to achieve a full solution to matching and ranking for online contextual advertising.

ACKNOWLEDGEMENT

I take this opportunity to express my profound gratitude and deep regards to my Professor and guide Ms. D. RamyaDorai for her exemplary guidance, monitoring and constant encouragement throughout the course of this work. Lastly, I thank almighty, my beloved family members and my friends for their blessings, timely help and support for accomplishing this work successfully.

REFERENCES

[1] X.-H. Phan, C.-T. Nguyen, D.-T Le, L.-M. Nguyen, S. Horiguchi, "A Hidden Topic Based Framework toward Building Applications with Short Web Documents", IEEE Transactions on Knowledge and Data Engineering, VOL. 23, NO. 7, JULY 2011

[2] L. Baker and A. McCallum, "Distributional Clustering of Words for Text Classification," Proc. ACM SIGIR, 1998.

[3] S. Banerjee, K. Ramanathan, and A. Gupta, "Clustering Short Texts Using Wikipedia," Proc. ACM SIGIR, 2007.

[4] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," J. Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[5] A. Blum and T. Mitchell, "Combining Labeled and Unlabeled Data with Co-Training," Proc. 11th Ann. Conf. Computational Learning Theory (COLT), 1998.

[6] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Measuring Semantic Similarity between Words Using Web Search

Engines," Proc. 16th Int'l Conf. World Wide Web (WWW), 2007.

[7] J. Cai, W. Lee, and Y. Teh, "Improving WSD Using Topic Features," Proc. Joint Conf. Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLPCoNLL), 2007.

[8] T. Joachims, "Text Categorization with SVMs: Learning with Many Relevant Features," Proc. 10th European Conf. Machine Learning (ECML), 1998.

[9] T. Hofmann, "Probabilistic LSA," Proc. Fifteenth Ann. Conf. Uncertainty in Artificial Intelligence (UAI), 1999.

[10] Kamal Nigam, John Lafferty, Andrew McCallum, "Maximum Entropy for Text Classification".

[11] M. Sahami and T. Heilman, "A Web-Based Kernel Function for Measuring the Similarity of Short Text Snippets," Proc. 15th Int'l Conf. World Wide Web (WWW), 2006.

BIOGRAPHY

Kanimozhiveena E received her B.E degree in Computer Science and Engineering from Anna University, Chennai, TamilNadu, in 2010. Currently she is pursuing her M.E degree at Adhiyamaan College of Engineering, Hosur, TamilNadu, India, affiliated to Anna university, Chennai.

D. Ramya Dorai received her B.E degree in Computer Science and Engineering from Bangalore University, Karnataka, in 2003. She received her M.E degree in Computer Science and Engineering from Anna University, in 2006. Currently she is working as associate professor, in Department of Computer Science and Engineering at Adhiyamaan College of Engineering, Hosur, Tamilnadu, India.