International Journal of Innovation and Scientific Research
ISSN 2351-8014 Vol. 11 No. 2 Nov. 2014, pp. 259-263
© 2014 Innovative Space of Scientific Research Journals
http://www.ijisr.issr-journals.org/

An Approach to Uncover Hidden Topics in Short & Sparse Web Documents

Sumayya Sameena[1] and Rehana[2]

[1]M.Tech, Department of CSE,
NimraCollege of Engg. & Tech, Vijayawada,
Andhra Pradesh., India

[2]Assistant Professor in CSE Dept,
NimraCollege of Engg. & Tech, Vijayawada,
Andhra Pradesh., India

Copyright © 2014 ISSR Journals. This is an open access article distributed under the ***Creative Commons Attribution License***, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT: My work introduces a hidden topic-based framework for processing short and sparse documents (e.g., search result snippets, product descriptions, book/movie summaries, and advertising messages) on the Web. The framework focuses on solving two main challenges posed by these kinds of documents: 1) data sparseness and 2) synonyms/homonyms. The former leads to the lack of shared words and contexts among documents while the latter are big linguistic obstacles in natural language processing (NLP) and information retrieval (IR). The underlying idea of the framework is that common hidden topics discovered from large external data sets (universal data sets), when included, can make short documents less sparse and more topic-oriented. Furthermore, hidden topics from universal data sets help handle unseen data better. The proposed framework can also be applied for different natural languages and data domains. We carefully evaluated the framework by carrying out two experiments for two important online applications (Web search result classification and matching/ranking for contextual advertising) with large-scale universal data sets and we achieved significant results.

KEYWORDS: Web mining, matching, Natural Language Processing, classification, clustering

1 INTRODUCTION

In this study, I developed the data diversity which has posed new challenges to Web Mining and IR research. Two main challenges we are going to address in this study are 1) short and sparse data problem and 2) synonyms and homonyms. In this we used on web search documents and e.g., search result snippets, product descriptions, book/movie summaries, and advertising messages) on the Web. We demonstrate that hidden topic-based approach can be a right solution to sparse data and synonym/ homonym problems.

We show that the framework is a suitable method to build online applications with limited resources. In this framework, universal data sets can be gathered easily because huge document collections are widely available on the Web. By incorporating hidden topics from universal data sets, we can significantly reduce the need of annotated data that are usually expensive and time-consuming to prepare. In this sense, our framework is an alternative to semi-supervised learning [2] because it also effectively takes advantage of external data to improve the performance.

2 PROBLEM STATEMENT

The main rule in this project search the data based on the search topic, display the data in short format or hidden the data. In the proposed system we are solving the two main challenges i.e., short and sparse data problem & synonyms and homonyms.

- Proposes the general framework of classification [1] and contextual matching with hidden topics.
- Describes the analysis of large-scale text/Web data collections that serve as universal data sets in the framework.

Describes how to build a matching and ranking model with hidden topics for online contextual advertising.

3 SCOPE

The main rule in this project hidden topic-based approach can be a right solution to sparse data and synonym/ homonym problems.

This section brings an in-detail description of hidden topic analysis of a large-scale Vietnamese news collection that serves as a "universal data set" in the general framework for contextual advertising mentioned earlier in With the purpose of using a large-scale data set for Vietnamese contextual advertising, we choose VnExpress as the universal data set for topic analysis. VnExpress is one of the highest ranking e-newspaper corporations in Vietnam, thus containing a large number of articles in many topics in daily life. For this reason, it is a suitable data collection for advertising areas.

4 OBJECTIVE

In this work data will be display the Tree view interconnected nodes. In this tree view data is display one tree view contain more than two nodes. Search Engine data display the Page Level but in this project data will display the Tree wise. A hidden topic-based approach is processing short and sparse documents in universal data set.

The proposed application can also be applied for different natural languages and data domains. We carefully evaluated the framework by carrying out two experiments for two important online applications (Web search result classification and matching/ranking for contextual advertising) with large-scale universal data sets. Area of work to data mining; data mining will allow us to retrieve data from huge amount of data, i.e. data mining techniques makes our work easy to get particular data.

By choosing this, I can provide the following features to the users:

- I can give reasons to the user why the particular getting for a particular query.
- I can provide dynamic Data search flow.
- I can improve the speed of the search.
- I can reduce the semantic web documents.

An e-commerce, online publishing, communication, and entertainment, Web data have become available in many different forms, genres, and formats which are much more diverse than ever before. Various kinds of data are generated everyday: queries and questions input by Web search users; Web snippets returned by search engines; Web logs generated by Web servers; chat messages by instant messengers blog posts and comments by users on a wide spectrum of online forums, e-communities, and social networks; product descriptions and customer reviews on a huge number of e-commercial sites; and online advertising messages from a large number of advertisers. However, this data diversity has posed new challenges to Web Mining and IR research[5]. Two main challenges we are going to address in this study are.

1) Short and sparse data problem and
2) Synonyms and homonyms.

5 A HIDDEN TOPIC BASED APPROACH

In this paper we discuss about the 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, we callW "universal data set" since it is large and diverse enough to cover a lot of information (e.g., words/topics) regarding the classification task. The whole framework of "learning to classify with hidden topics" is depicted. The framework consists of five subtasks.

- collecting universal data set W,
- carrying out topic analysis for W,
- preparing labeled training data,
- performing topic inference for training and test data, and
- Building the classifier.

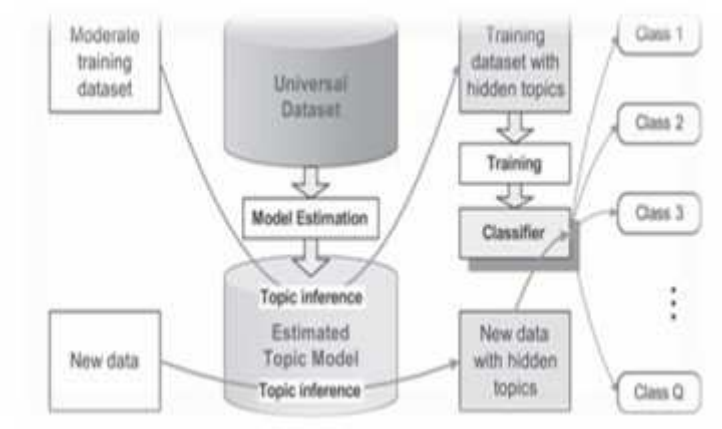

Fig-1: Framework of learning to classify sparse text/Web with hidden topics

(a) Choosing an appropriate " universal dataset"
(b) Doing topic analysis for the universal dataset
(c) Building a moderate size labeled training dataset
(d) Doing topic inference for training and futuer data
(e) Building the classifer

6 MATCHING/RANKING WITH HIDDEN TOPICS

A present our general framework for contextual page-ad matching and ranking with hidden topics discovered from external large-scale data collections.

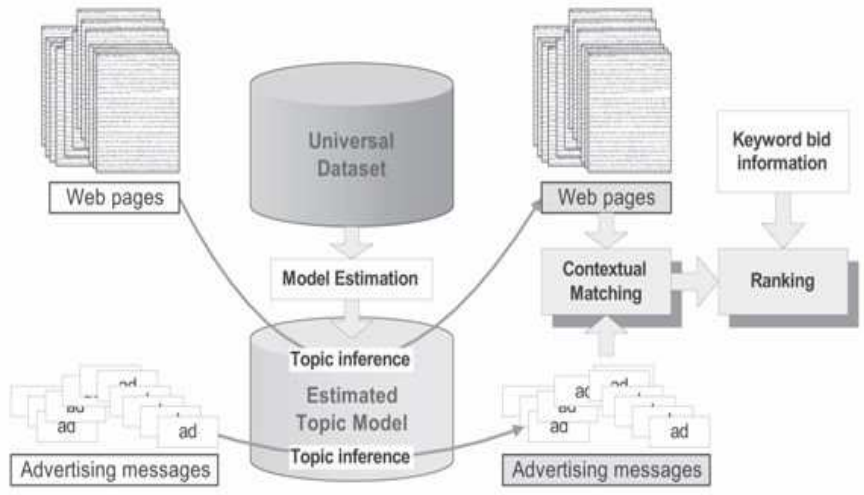

Fig-2: Framework of page-ad matching and ranking with hidden topics

(a) Choosing an appropriate " universal dataset"
(b) Doing topic analysis for the universal dataset
(c) Doing topic inference for web pages and ads
(d) Page-Ad matching and Ranking

7 GENERAL FRAMEWORK

We give a general description of the proposed framework: classifying, clustering, and matching with hidden topics discovered from external large-scale data collections. It is general enough to be applied to different tasks, and among

them we take two problems: document classification and online contextual advertising [3]as the demonstration. Document classification, also known as text categorization, has been studied intensively during the past decade.

Many learning methods, such as k nearest neighbors (k-NN), Naive Bayes, maximum entropy, and support vector machines (SVMs)[4], have been applied to a lot of classification problems with different benchmark collections and achieved satisfactory Results. However, our framework mainly focuses on text representation and how to enrich short and sparse texts to enhance classification accuracy.

8 APPROXIMATION METHODS

In this work, we show how we can efficiently calculate an approximation of the authority flow between two nodes (we generalize for multiple nodes as well) given the paths (trees for more than two nodes) with length up to M connecting the nodes. The latter problem has been extensively studied in prior work where efficient algorithms are presented to find all paths of a data graph (a graph of pages and hyperlinks for the web, or of data objects and their relationships for databases as we explain) connecting two nodes.

An analytical model is presented to estimate the quality of this approximation as a function of the maximum lengthM of the paths generated by the proximity algorithm. Also, we analytically calculate the error in ordering that this approximation imposes.

9 PAGE RANK ALGORITHM

A set of works has tackled the problem of improving the performance of the original Page Rank algorithm. Algorithms are presented in to improve the calculation of a global Page Rank. Jeh and Widom present a method to efficiently calculate the Page Rank values for multiple base sets, by precomputing a set of partial vectors which are used in runtime to calculate the Page Ranks. Tong et al. exploit structural properties of real graphs to efficiently perform random walk over large graphs and employ pre computation[6] to improve performance.

10 EXISTING SYSTEM

Due to the complexity of calculating the authority flow, current systems only use pre-computed authority flows in runtime. This limitation prohibits authority flow to be used more effectively as a ranking factor. The existing system have been developed to apply this principle

1. on the Web
   - Page rank
   - Topic Sensitive Page rank
2. Bibliographic data bases
   - Object rank
3. Biological data bases
   - Hubs of Knowles Project

The above system has some drawbacks

1. There is no way to explain to the user why a particular result is current score
2. The authority flow rates which have been shown to dramatically affect the results quantity in object ranks have to be set manually by a domain expert.
3. There is no query reformation methodology to refine the query results according to the user's preferences.

11 PROPOSED SYSTEM

The above system has some drawbacks.

1. There is no way to explain to the user why a particular result is current score
2. The authority flow rates which have been shown to dramatically affect the results quantity in Object ranks have to be set manually by a domain expert.
3. There is no query reformation methodology to refine the query results according to the user's preferences.

To resolve these problems in existing Proposal, we define authority flow between nodes in terms of the paths connecting the nodes. We initially focus on pair wise authority flows, that is, we approximate the authority flow between a base set comprised of a single node and a target node. We generalize to base sets with multiple nodes. To do so, we use the linearity theorem. We first estimate the relative error ES of the authority flow value calculated using the approximation method, and then we present how this error affects the ordering of the results of a query that uses authority flow as the ranking factor.

12 CONCLUSION

We presented a method to efficiently approximate the authority flow between a base set b and a node v. this method assumes no ieee transactions on knowledge and data engineering, Performance and quality experiments on performance with varying path length m. results' quality with varying path length performance and quality experiments on ds. Performance with varying path length. results' quality with varying path length m. prior knowledge of b or v, and hence, it is suitable for on- the-fly authority flow computation. Our work allows authority flow to be used in new real-time applications, like measuring the quality of a result tree in a proximity search system, or answering complex on the- fly queries. We also analytically prove the error of our approximation method and the error it imposes in the ordering of query results.

13 FUTURE ENHANCEMENTS

In my work, user enter any data will access Database but Future implementation using AJAX. Data will be access the client side only. I work is to develop the database only using sqlserver future implementation connect the different databases. Ever y request data access thought database future implementation data will access though XML or client side. This work does not maintain the historical data.

REFERENCES

[1] L. Baker and A. McCallum,"Distribuitional clustering of words for Text Classification,"Proc. ACM, SIGIR, 1998.

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

[3] A. Broder, M Fontoura, V. Josifovski, and L.Riedel," A Semantic Approach to Contextual Advertising, "Proc.ACM SIGIR, 2007.

[4] T.Joachims,"Text Categorization with SVMS: Learning With Many Relevant Feautures,"Proc.10th European Conf.Machine Learning (ECML), 1998

[5] D.Metzler, S.Dumais, And C. Meek," Similarity measures for Short Segments of Text," Proc.29th European Conf .IR Research (ECIR), 2007.

[6] A. Berger and A. Pietra and J. Pietra," A Maximum Entropy Approach To Natural Language Processing, "Computational Linguistics, vol.22, No.1.PP -39-71, 1996.

[7] 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

[8] R. Wang, P. Zhang, and M. Eredita. Understanding consumers attitude toward advertising. Proc. AMCIS, 2002.

[9] L. CAI and T. Hofmann. Text categorization by boosting automatically extracted concepts. Proc. ACM SIGIR, 2003.

[10] J. CAI, W. Lee, and Y. Teh. Improving WSD using topic features. Proc. EMNLP- CoNLL, 2007.

[11] D. Bollegala, Y. Matsuo, and M. Ishizuka. Measuring semantic similarity between words using Web search engines. Proc. WWW, 2007