IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 92-94

Web Mining Techniques for Query Log Analysis

Dr.V.Sangeetha
(Information Technology, Sankara College of science and commerce / Bharathiyar University, India)

***Abstract**: The objective of this thesis is to establish automatic content analysis methods and scalable graph-based models for query log analysis. One important aspect of this thesis is therefore to develop a framework to combine the content information and the graph information with the following two purposes: 1) analyzing Web contents with graph structures, more specifically, mining query logs; and 2) identifying high-level information needs, such as expertise retrieval, behind the contents.*
For the first purpose, a novel entropy-biased framework is proposed for modeling bipartite graphs, which is applied to the click graph for better query representation by treating heterogeneous query-URL pairs differently and diminishing the effect of noisy links. For the second purpose, a weighted language model is proposed to aggregate the expertise of a candidate from the associated documents. The model not only considers the relevance of documents against a given query, but also incorporates important factors of the documents in the form of document priors. Experimental results on the expert finding task demonstrate these methods can improve and enhance traditional the traditional expertise retrieval models with better performance.
***Keywords: -** bipartite graph, Expertise retrieval, novel entropy, Query log analysis, weighted language*

I. Introduction

The World Wide Web (Web) has been providing an important and indispensable platform for receiving information and disseminating information as well as interacting with society on the Internet Due to the properties of the Web data, we are currently drowning in information and facing information overload [90]. The information may consist of Web pages, images, people and other types of data. To help Web users find their infor-mation need, a critical issue is to understand what users want with respect to the given query by mining the query logs. On the other hand, it would be quite interesting and important to identify relevant experts with expertise for fur-ther consulting about the query topic, which is also called expertise retrieval. In order to achieve the above goals, Web mining has emerged as an impor-tant interdisciplinary research area by leveraging several disciplines such as information retrieval, data mining, machine learning, and database systems.

II. Web Mining Techniques

The Web search and mining research is a converging research area from several communities, such as information retrieval, link analysis, data mining, and machine learning, as well as others. Each of them has been separately studied in the past decades. Let us briefly introduce them as follows.

1.1 Traditional Information Retrieval

Web search and mining has its root in information retrieval. In general, information retrieval (IR) refers to the retrieval of unstructured data. Most often, it is related to Text Retrieval, i.e. the retrieval of textual docu-ments. Other types of retrieval include, for example, Image Retrieval, Video Retrieval, and Music Retrieval. Retrieving information simply means finding a set of documents that are relevant to the user query. The retrieval accuracy of an IR system is directly deter-mined by the quality of the scoring function. Thus, a major research challenge in information retrieval is to seek an optimal scoring function (retrieval func-tion), which is based on a retrieval model.

1.2 Link Analysis

The analysis of hyperlinks and the graph structure of the Web has been in-strumental in the development of Web search, Link analysis is one of many factors considered by Web search engines in computing a composite score for a Web page on any given query. Basically, link analysis for Web search has intellectual antecedents in the field of citation analysis , which seeks to quantify the influence of scholarly articles by analyzing the pattern of citations among them.

1.3 Machine Learning

There is a close relationship between machine learning and Web mining re-search areas. A major focus of machine learning research is to learn to recog-nize complex patterns and make decisions based on data. Machine learning has been applied to many applications of the Web search and mining, such as learning to rank

, text categorization , and Web query classification . Machine learning can be typically categorized as supervised learning, unsupervised learning, semi-supervised learning, as well as others. Supervised learning considers the problems of estimating certain functions from examples with label information, such as Support Vector Machines (SVM) Neural Network and naive Bayes classifier . Unsupervised learning considers the problem of learning from a collection of data instances without training labels. One of the most popular areas of study in unsupervised learning is data clustering techniques, which have been widely used for data mining applications . Semi-supervised learning has recently been proposed to take advantage of both labeled and unlabeled data, which has been demonstrated to be a promising approach.

III. Applications

Two main applications are studied. One is query log analysis, and the other is expertise retrieval. Although there are many differences between query log analysis and expertise retrieval, the key point is that all of these data, the query log data and the expertise retrieval data, can be viewed as the combination of the content and graph information. The objective of this work is to propose a general Web mining framework to combine the content and the graph information effectively, by leveraging Web mining techniques to boost the performance of these applications.

1.4 Query Log Analysis

Web query log analysis has been studied widely with different Web mining techniques for improving search engines' efficency and usability in recent years. Such studies mined the logs to improve numerous search engine's capabilities, such as query suggestion, query classification, ranking, targeted advertising, etc. The click graph , a bipartite graph between queries and URLs, is an important technique for describing the information contained in the query logs, in which edges connect a query with the URLs that were clicked by users as a result. As the edges of the click graph can capture some semantic relations between queries and URLs, it is useful to represent the query using the vector of documents when only considering the graph information. State-of-the-art approaches based on the raw click frequencies for modeling the click graph, however, are not noise-eliminated. Nor do they handle hetero-generous query-URL pairs well. To deal with these critical problems, a novel entropy-biased framework is proposed for query representation on the click graph.

1.5 Expertise Retrieval

Web mining and information retrieval tech-niques, many research efforts in this field have been made to address high-level information retrieval and not just the traditional document retrieval, such as expertise retrieval . Expertise retrieval has received increased interests since the introduction of an expert finding task.The task of expertise retrieval is to identify a set of persons with relevant expertise for the given query. Traditionally, the expertise of a person is characterized based on the documents that have been associated with the person. One of the state-of-the-art approaches [8, 37] is the document-based model using a statistical language model to rank experts. However, these methods only consider the documents associated with the experts. Actually, in addition to the associated documents, there is much other information that can be included, such as the importance of the documents, the graph information, and the community information. Therefore, how to utilize these informa-tion to model and enhance the expertise retrieval becomes an interesting and challenging problem.

IV. The Unified Framework And Its Contributions

This paper aims to propose a general Web mining framework to combine the content and the graph information effectively.

4.1 Novel Entropy-biased Framework for Query Representation on the Click Graph

Based on the click graph, the query can be represented by a vector of connected documents when only considering the graph information. A novel entropy-biased framework is proposed for better query representation by combining the inverse query frequency with the click frequency and user frequency information simultaneously. In the framework, a new notion, namely the inverse query frequency, is introduced to weigh the importance of a click on a certain URL, which can be extended and used for other bipartite graphs. The proposed entropy-biased model is the first formal model to distinguish the variation on different query-URL pairs on the click graph. In addition, a new source, called the user frequency, is identified for diminishing the manipulation of the malicious clicks. In our entropy-biased framework, Click Frequency-Inverse Query Frequency (CF-IQF) is a simplified version of the entropy-biased model. And this weighting scheme can be applied to other bipartite graphs.

4.2 A Weighted Language Model for Expertise Retrieval with Graph-based Regularization

In order to investigate the combination of more heterogeneous informa-tion, the high-level expertise retrieval task is addressed based on the large-scale DBLP bibliography and its supplemental data from Google Scholar. A novel expert finding framework is proposed to identify the relevant experts in the academic field. A weighted language model is formally defined to aggregate the expertise of a candidate from the as-sociated documents. The model takes into account not only the rele-vance between a query and documents, but also the importance of the documents. In the framework, the paper importance is interpreted by introducing a prior probability that the paper is written by an expert.Furthermore, a graph-based regularization method is integrated to boost the performance by refining the relevance scores of the documents.

4.3 Entropy Biased Model

The CF model only considers the raw click frequency, and treats diff erent query-URL pairs equally even if some URLs are very heavily clicked. More generally, a great variation in URL distribution is likely to appear, and it may thus cause the loss of important information since diff erent query-URL pairs are not sufficiently distinguished. It is predicted that the more general and highly ranked URL would be clicked and connected with more queries than the specific URLs. Thus the less spe-cific URLs would have a larger collection distribution than the more specific ones, which tends to increase the ambiguity and uncertainty of the URLs in the ordinary sense.

V. Experimental Evaluation

In the following experiments we compare our proposed models with other methods on the tasks of mining query logs through an empirical evaluation. We define the following task: Given a query and a click graph, the system has to identify a list of queries which are most similar or semantically relevant to the given query. In the rest of this section, we introduce the data collection, the assessments and evaluation metrics, and present the evaluation results.

5.1 Data Collection and Analysis

This dataset is the raw data recorded by the search engine, and contains a lot of noises. Hence, we conduct a similar method employed in to clean the raw data. We clean the data by removing the queries that appear less than 2 times, and by combining the near-duplicated queries which have the same terms without the stopwords and punctuation.

5.2 Assessments and Evaluation Metrics

It is difficult to evaluate the quality of query similarity/relevance rankings due to the scarcity of data that can be examined publicly. For an automatic evaluation, we utilize the same method used in [7] to evaluate the similarity of retrieved queries.

VI. Conclusion

This paper aims to develop a general framework to make use of the content and graph information eff ectively by leveraging information retrieval, machine learning, and knowledge discovery techniques for real-world applications, es-pecially query log analysis and expertise retrieval. To this purpose, we de-velop scalable automatic content analysis methods and graph-based models to analyze a huge amount of data resources including AOL query logs, on-line DBLP, Google Scholar, etc. and propose several approaches to tackle various challenging problems.

References

[1]. A. Agarwal, S. Chakrabarti, and S. Aggarwal. Learning to rank net-worked entities. In Proceedings of the Twelfth ACM SIGKDD Interna-tional Conference on Knowledge Discovery and Data Mining (KDD), pages 14–23, 2006.
[2]. J. Alpert and N. Hajaj. We knew the web was big... The Official Google Blog, July 25, 2008.
[3]. R. Baeza-Yates, B. Ribeiro-Neto, et al. Modern information retrieval. Addison-Wesley Harlow, England, 1999.
[4]. R. A. Baeza-Yates, C. A. Hurtado, and M. Mendoza. Query recommen-dation using query logs in search engines. In EDBT Workshops, pages 588–596, 2004.
[5]. R. A. Baeza-Yates and A. Tiberi. Extracting semantic relations from query logs. In KDD, pages 76–85, 2007.
[6]. K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora