International Journal of Advanced Technology in Engineering and Science
Vol. No.5, Issue No. 03, March 2017
www.ijates.com
ijates
ISSN 2348 - 7550

A SURVEY ON WEB MINING APPROACHES FOR INFORMATION EXTRACTION FROM WORLD WIDE WEB

Sunil Kumar Thota[1], Srikanth Lakumarapu[2], Thallamudi Pavan Kumar[3]

[1,2,3]*Asst. Prof., Computer Science and Engineering, Sphoorthy Engineering College, JNTUH (India)*

ABSTRACT

With the advent of information technology, a user is able to obtain relevant information from the World Wide Web, which contains a huge amount of information, simply and quickly by entering search queries [3][4]. In response to the queries, the database servers generate the information and deliver it directly to the user. Due to the rapid growth of digital data made available in recent years, knowledge discovery and data mining have attracted a great deal of attention with an imminent need for turning such data into useful information and knowledge. Many applications, such as market analysis and business management, can benefit by the use of the information and knowledge extracted from a large amount of data. Knowledge discovery can be viewed as the process of nontrivial extraction of information from large databases, information that is implicitly presented in the data, previously unknown and potentially useful for users. Data mining is therefore an essential step in the process of knowledge discovery in databases. Most of the existing web data mining techniques are proposed for the purpose of developing efficient mining algorithms to find particular patterns within a reasonable and acceptable time frame. With a large number of patterns generated by using data mining approaches, how to effectively use and update these patterns is still an open research issue.

I. INTRODUCTION

Web text mining is the discovery of interesting knowledge in text documents [1]. It is a challenging issue to find accurate knowledge or features in text documents to help users to find what they want. In the beginning, Information Extraction (IE) provided many term-based methods to solve this challenge, such as Rocchio and probabilistic models [7], rough set models, BM25 and support vector machine (SVM) based filtering models. The advantages of term based methods include efficient computational performance as well as mature theories for term weighting, which have emerged over the last couple of decades from the IE and machine learning communities. However, term based methods suffer from the problems of polysemy and synonymy, where polysemy means a word has multiple meanings, and synonymy is multiple words having the same meaning. The semantic meaning of many discovered terms is uncertain for answering what users want.

II. SURVEY OF WEB MINING TECHNIQUES

Over the years, people have often held the hypothesis that phrase-based approaches could perform better than the term based ones, as phrases may carry more “semantics” like information.

Phrase-based classification (PBC) can achieve extremely high precision (95%) with reasonable coverage (80%, improvable) on a large-scale text collection. The class of data is characterized on which Phrase-based classification can be successful i.e., the data that satisfies the near-sufficiency property, and prove that PBC will be feasible on data of virtually any size. This hypothesis has not fared too well in the history of IE [13]. Although phrases are less ambiguous and more discriminative than individual terms, the likely reasons for the discouraging performance include: 1) phrases have inferior statistical properties to terms, 2) they have low frequency of occurrence, and 3) there are large numbers of redundant and noisy phrases among them [13].

The paper *High-Precision Phrase-Based Document Classification on a Modern Scale* [9] presents a document classification system that employs lazy learning from labeled phrases, and argues that the system can be highly effective when most of information on document labels is captured in phrases. The research quantifies the near sufficiency property using the Information Bottleneck principle . The natural language restricts the number of common phrases composed of a vocabulary to grow *linearly* with the size of the vocabulary. PBC explains the classification results excellently, as well as cost is low in development maintenance. Hence Phrase-based Classification is most useful in *multilabel* classification.

Many types of text representations have been proposed in the past. A well-known one is the bag of words that uses keywords (terms) as elements in the vector of the feature space. In [5], the *tf*idf* weighting scheme is used for text representation in Rocchio classifiers. The paper concentrated on *semi-supervised text classification*. The problem is regarded as a two-class (*positive* and *negative*) classification problem, where there are only labeled *positive* training data, but no labeled *negative* training data. Due to the lack of negative training data, the classifier building is thus *semi-supervised*.

It is possible to manually label some negative examples but it is very time consuming. The paper builds a classifier using only a set of positive examples and a set of unlabeled examples. Collecting unlabeled examples or documents is normally easy and inexpensive in many text or Web page domains. With the growing volume of text documents on the Web and digital libraries, the documents that are related to one's interest are to be found. The ability to build classifiers without negative training data is particularly useful if one needs to find positive documents from many text collections or sources. The paper proposes a more effective and robust technique to solve the problem. It is based on the Rocchio method and SVM. The idea is to first use Rocchio to extract some reliable negative documents from the unlabeled set and then apply SVM iteratively to build and to select a classifier.

In addition to *TFIDF*, the global *IDF* and entropy weighting scheme is proposed in [8] and improves performance by an average of 30 percent. The paper concentrates on Text categorization which plays an important role in applications where information is filtered, monitored, personalized, categorized, organized or searched. Feature selection remains as an effective and efficient technique in text categorization. Feature selection metrics are commonly based on term frequency or document frequency of a word. The paper focuses on relative importance of these frequencies for feature selection metrics. The document frequency based metrics of discriminative power measure and GINI index were examined with term frequency for this purpose. Experimental results show that the term frequency based metrics may be useful especially for smaller feature sets. Two characteristics of term frequency based metrics were observed by analyzing the scatter of features

among classes and the rate at which information in data was covered. These characteristics may contribute toward their superior performance for smaller feature sets.

Various weighting schemes for the bag of words representation approach were given in [14]. The main function of a term-weighting scheme is the enhancement of retrieval effectiveness. Effective retrieval depends on items likely to be relevant to the user's needs must be retrieved and items likely to be extraneous must be rejected. Two measures are normally used to assess the ability of a system to retrieve the relevant and reject the non-relevant items of a collection, known as *recall* and *precision*, respectively. Recall is the proportion of relevant items retrieved, measured by the ratio of the number of relevant retrieved items to the total number of relevant items in the collection. Precision is the proportion of retrieved items that are relevant, measured by the ratio of the number of relevant retrieved items to the total number of retrieved. Three main considerations appear important in this connection.

First, terms that are frequently mentioned in individual documents, or document excerpts, appear to be useful as recall enhancing devices. This suggests that a *term frequency* (tf) factor be used as part of the term-weighting system measuring the frequency of occurrence of the terms in the document or query texts.

Second, term frequency factors alone cannot ensure acceptable retrieval performance. Specifically, when the high frequency terms are not concentrated in a few particular documents, but instead are prevalent in the whole collection, all documents tend to be retrieved, and this affects the search precision. Hence a new collection-dependent factor must be introduced that favors terms concentrated in a few documents of a collection. *Term discrimination* considerations suggest that the best terms for document content identification are those able to distinguish certain individual documents from the remainder of the collection. This implies that the best terms should have high term frequencies but low overall collection frequencies. A reasonable measure of term importance may then be obtained by using the product of the term frequency and the inverse document frequency (tf x idf).

A third term-weighting factor, in addition to the term frequency and the inverse document frequency, appears useful in systems with widely varying vector lengths. In many situations, short documents tend to be represented by short-term vectors, whereas much larger-term sets are assigned to the longer documents. When a large number of terms are used for document representation, the chance of term matches between queries and documents is high, and hence the larger documents have a better chance of being retrieved than the short ones. Normally, all relevant documents should be treated as equally important for retrieval purposes.

The problem of the bag of words approach is how to select a limited number of features among an enormous set of words or terms in order to increase the system's efficiency and avoid over fitting [13]. In order to reduce the number of features, many dimensionality reduction approaches have been conducted by the use of feature selection techniques, such as Information Gain, Mutual Information, Chi-Square, Odds ratio, and so on. Dimensionality Reduction is also beneficial since it tends to reduce *overfitting*, that is, the phenomenon by which a classifier is tuned also to the *contingent* characteristics of the training data rather than just the *constitutive* characteristics of the categories. Classifiers that overfit the training data are good at reclassifying the data they have been trained on, but much worse at classifying previously unseen data. Experiments have shown that, in order to avoid overfitting a number of training examples roughly proportional to the number of terms used is needed. If Dimensionality Reduction is performed, overfitting may be avoided even if a smaller amount

of training examples is used. However, in removing terms the risk is to remove potentially useful information on the meaning of the documents. It is then clear that, in order to obtain optimal (cost-) effectiveness, the reduction process must be performed with care. Various DR methods have been proposed, either from the information theory or from the linear algebra literature, and their relative merits have been tested by experimentally evaluating the variation. There are two distinct ways of viewing DR, depending on whether the task is performed locally or globally: A second, orthogonal distinction may be drawn in terms of the nature of the resulting terms:

—*DR by term selection*

—*DR by term extraction*

The choice of a representation depended on what one regards as the meaningful units of text and the meaningful natural language rules for the combination of these units [13]. With respect to the representation of the content of documents, some research works have used phrases rather than individual words.

In [15], the combination of unigram and bigrams was chosen for document indexing in text categorization (TC) and evaluated on a variety of feature evaluation functions (FEF). Negative relevance feedback is very useful for information filtering. However, whether negative feedback can largely improve filtering accuracy is still an open question. This paper presents a pattern mining based approach for this open question. It introduces a method to select negative documents (or called offenders) that are close to the extracted features in the positive documents. It also proposes an approach to classify extracted terms into three groups: positive specific terms, general terms and negative specific terms. In this perspective, it presents an iterative algorithm to revise extracted features. The research provides a promising methodology for evaluating term weights based on discovered patterns (rather than documents) in both positive and negative relevance feedback. The paper proposes a pattern mining based approach to select some offenders from the negative documents, where an offender can be used to reduce the side effects of noisy features. It also classifies extracted features (i.e., terms) into three categories: positive specific terms, general terms, and negative specific terms. In this way, multiple revising strategies can be used to update extracted features.

In the paper *Learning to Classify Texts Using Positive and Unlabeled Data* [5] **,** a set *P* of documents of a particular class (called *positive class*) and a set *U* of unlabeled documents that contains documents from class *P* and also other types of documents (called *negative class documents*), a classifier is built to classify the documents in *U* into documents from *P* and documents not from *P*. The key feature of this problem is that there is no labeled negative document, which makes traditional text classification techniques inapplicable. It combines the Rocchio method and the SVM technique for classifier building. Experimental results show that the new method outperforms existing methods significantly.

A phrase-based text representation for Web document management was also proposed in [10]. The World Wide Web has provided the facility of bringing information to the fingertips of its users. Since most of the documents available on the web are machine-readable but not machine-understandable, ensuring the retrieval of relevant information continues to be a difficult task. The drawbacks in this approach are lack of direct relationship between word frequency and its importance, and the effect of the word ambiguities. Considering these shortcomings of the keyword-based method, this paper presents a phrase-based text representation approach that uses rule-based natural language processing (NLP) techniques. Extraction of key-phrases from text documents is

based on a process of partial parsing. By making the indexing terms more meaningful through reduction of the ambiguity in words considered in isolation, improvement in retrieval effectiveness is achieved. This paper proposes an attempt to use a phrase-based text representation approach for improving the effectiveness of searching a document on the web. Since users have a tendency of providing input queries in the form of phrases, phrasal indexing aims to improve the retrieval effectiveness. From experiments, it is observed that natural language processing has a limited but a definite positive impact on retrieval effectiveness. Query preprocessing techniques aimed at producing higher quality queries appear to be very effective.

Term-based ontology mining methods also provided some thoughts for text representation

For example, hierarchical clustering was used to determine synonymy and hyponymy relations between keywords. Also, the pattern evolution technique was introduced in [6] in order to improve the performance of term-based ontology mining. There is no doubt that numerous discovered patterns can be found from the Web data using data mining techniques. However, it is ineffective to use the discovered patterns in Web user profile mining due to the ambiguities in the data values (terms). The consequent result is that some inappropriate discovered patterns and many discovered patterns include uncertainties. In this paper, an ontology mining technique is developed to provide a solution for this challenge. The discovered ontology consists of the top backbone which illustrates the linkage between compound classes of the ontology and the base backbone which illustrates the linkage between primitive classes and compound classes. A mathematical model is set to represent discovered knowledge on the ontology. A novel method is also presented for capturing evolving patterns in order to refine the discovered ontology. In addition, a formal method is established for learning how to assess relevance in the ontology in order to effectively use discovered knowledge. The echnique not only gains a better performance on both precision and recall, it also decreases the burden of online training. The research is significant for WI since it makes a breakthrough by effectively synthesizing taxonomic relation and nontaxonomic relation in a mathematical model. It is also significant for data mining because it provides an approach for representation, application, and maintenance of discovered knowledge for solving real problems.

A hybrid recommendation method with reduced data for large-scale application was proposed in [3]. Internet users search the Web to find information or products of interest. Users are faced with terminologies, such as personalized search, retrieval, filtering, intelligent agent, etc. The frequency of these terminologies reflects the efforts of commercial sites and researchers to suggest information or products that meet customers' needs. Such recommendations are an essential part of attracting customers. Online companies have the capability to acquire customers' preferences, and then, use them to recommend products on a one-to-one basis in real time and, more importantly, at a much lower cost to company.

The software that makes such customized responses possibleis commonly called recommendation systems. According to the type of information used to form their responses to customers, recommendation systems can be further categorized as content based (CB) filtering, collaborative filtering (CF), or hybrid ones. Personalized recommendation algorithms, such as CF and CB filtering, have been extensively researched. CB recommendation algorithms operate by matching customer interests with product features that describe the characteristics of items. This approach suffers because of the difficulty of finding a few common features to represent items, and then, getting users' feedback on their own preferences concerning the features. CF algorithms, which are the most successful recommendation technique, have been suggested to overcome this

problem. The CF technique relies on the rating of items or on the transactional data of each specific user. CF identifies customers or neighbors, whose purchasing patterns are similar to those of the target user, and recommend items based on the information of similar customers. The CF approach nevertheless has been reported as having several major limitations, including scalability, sparsity, and new item problems [14]. The research paper proposes a hybrid algorithm combining a modified Pearson's correlation coefficient-based CF and distance-to-boundary (DTB) based CB. HYRED, the hybrid method proposed can be used effectively and efficiently for three reasons. HYRED proposes the concept of neighborhood in CF to efficiently analyze the transaction data. The use of the nearest and farthest neighbors of a target customer yields a reduced dataset of useful information for solving the scalability problem. Fewer training and testing datasets enables not only to lessen the computing effort, but also to improve the performance of recommendations. The processes of filtering irrelevant data by using the neighborhood concept of CF make it possible to consider the items that are likely to be purchased by a target user. It proposed the generalized rating system based on the distance of an item to the decision boundary of a classifier. In this the item closer to the class of purchased items may have a higher probability of being sold. A generalized hybrid recommendation algorithm by using a weighted coefficient in which the DTB and CF methods are special cases of our generalized algorithm is proposed. The weighting scheme makes the algorithm adequate for generalized applications, and HYRED is flexible enough for application with any available datasets. Moreover, HYRED, when weighting is properly valued, has yielded better results than pure DTB, pure CF, and simple combined hybrid method. Content-based (CB) filtering uses the features of items, whereas collaborative filtering (CF) relies on the opinions of similar customers to recommend items. In addition to these techniques, hybrid methods have also been suggested to improve the performance of recommendation algorithms. However, even though recent hybrid methods have helped to avoid certain limitations of CB and CF, scalability and sparsity are still major problems in large-scale recommendation systems. In order to overcome these problems, this paper proposes a novel hybrid recommendation algorithm HYRED, which combines CF using the modified Pearson's binary correlation coefficients with CB filtering using the generalized distance-to-boundary-based rating. In the proposed recommendation system, the nearest and farthest neighbours of a target customer are utilized to yield a reduced dataset of useful information by avoiding scalability and sparsity problem when confronted by tremendous volumes of data. The use of reduced datasets enables us not only to lessen the computing effort, but also to improve the performance of recommendations. In addition, a generalized method to combine CF and CB system into a hybrid recommendation system is proposed by developing on the normalization metric. The experiments have shown that the use of reduced datasets saves computational time, and neighbour information improves performance.

Constructing a single text classifier that excels in any given application is a rather inviable goal. As a result, ensemble systems are becoming an important resource, since they permit the use of simpler classifiers and the integration of different knowledge in the learning process. However, many text-classification ensemble approaches have an extremely high computational burden, which poses limitations in applications in real environments. Moreover, state-of-the-art kernel-based classifiers, such as support vector machines and relevance vector machines, demand large resources when applied to large databases. Therefore in paper Distributed Text Classification With an Ensemble Kernel-Based Learning Approach[4] , the use of a new systematic distributed ensemble framework was proposed, based on a generic deployment strategy in a cluster

distributed environment. It employs a combination of both task and data decomposition of the text-classification system, based on partitioning, communication, agglomeration, and mapping to define and optimize a graph of dependent tasks. The framework also includes an ensemble system where diverse patterns of errors are exploited and gain from the synergies between the ensemble classifiers. The experimental results show that the performance of the proposed framework outperforms standard methods both in speed and classification.

Pattern mining has been extensively studied in data mining communities for many years. A variety of efficient algorithms such as Apriori-like algorithms [2], PrefixSpan [11], FP-tree [12], SPADE[16], SLPMiner [17], and GST [12] have been proposed.

IV. CONCLUSION

These research works have mainly focused on developing efficient mining algorithms for discovering patterns from a large data collection. However, searching for useful and interesting patterns and rules was still an open problem [13], [14], [15]. In the field of text mining[1], pattern mining techniques can be used to find various text patterns, such as sequential patterns, frequent itemsets, co-occurring terms and multiple grams, for building up a representation with these new types of features. Many data mining techniques have been proposed for mining useful patterns in text documents. However, how to effectively use and update discovered patterns is still an open research issue, especially in the domain of web text mining. Since most existing text mining methods adopted term-based approaches, they all suffer from the problems of polysemy and synonymy. Over the years, people have often held the hypothesis that pattern (or phrase)-based approaches should perform better than the term-based ones, but many experiments do not support this hypothesis.

REFERENCES

[1]. Yuefeng Li, A. Algarni, M. Albathan, Yan Shen, and M. A. Bijaksana, "Relevance Feature Discovery for Text Mining", IEEE Transactions on Knowledge And Data Engineering, VOL. 27, NO. 6, JUNE 2015

[2]. NingZhong, Yuefeng Li, and Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining", IEEE Transactions on Knowledge And Data Engineering, VOL. 24, NO. 1, JANUARY-2012

[3]. S. H. Choi, Y.-S. Jeong, and M. K. Jeong, "A hybrid recommendation method with reduced data for large-scale application," IEEE Trans. Syst., Man, Cybern., vol. 40, no. 5, pp. 557–599, Sep. 2010.

[4]. C. Silva, U. Lotric, B. Ribeiro, and A. Dobnikar, "Distributed text classification with an ensemble kernel-based learning approach," IEEE Trans. Syst., Man, Cybern., vol. 40, no. 3, pp. 287–297, May 2010.

[5]. X. Li and B. Liu, "Learning to Classify Texts Using Positive and Unlabeled Data," Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI '03), pp. 587-594, 2003.

[6]. Y. Li and N. Zhong, "Mining Ontology for Automatically Acquiring Web User Information Needs, "IEEETrans. Knowledge and Data Eng., vol. 18, no. 4, pp. 554-568, Apr. 2006

[7]. T. Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with tfidf for Text Categorization," Proc. 14th Int'l Conf. Machine Learning (ICML '97), pp. 143-151, 1997.

[8]. N. Azam and J. Yao, "Comparison of term frequency and document frequency based feature selection metrics in text categorization," Expert Syst. Appl., vol. 39, no. 5, pp. 4760–4768, 2012.

[9]. R. Bekkerman and M. Gavish, "High-precision phrase-based document classification on a modern scale," in Proc. 11th ACM SIGKDD Knowl. Discovery Data Mining, 2011, pp. 231–239.

[10]. R. Sharma and S. Raman, "Phrase-Based Text Representation for Managing the Web Document," Proc. Int'l Conf. Information Technology: Computers and Comm. (ITCC), pp. 165-169, 2003.

[11]. X. Yan, J. Han, and R. Afshar, "Clospan: Mining Closed Sequential Patterns in Large Datasets," Proc. SIAM Int'l Conf. Data Mining (SDM '03), pp. 166-177, 2003.

[12]. Y. Huang and S. Lin, "Mining Sequential Patterns Using Graph Search Techniques," Proc. 27th Ann. Int'l Computer Software and Applications Conf., pp. 4-9, 2003

[13]. F. Sebastiani, "Machine Learning in Automated Text Categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.

[14]. G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," Information Processing and Management: An Int'l J., vol. 24, no. 5, pp. 513-523, 1988.

[15]. Y. Li, A. Algarni, and Y. Xu, "A pattern mining approach for information filtering systems," in Inf. Retrieval, vol. 14, pp. 237– 256, 2011.

[16]. M. Zaki, "Spade: An Efficient Algorithm for Mining Frequent Sequences," Machine Learning, vol. 40, pp. 31-60, 2001

[17]. M. Seno and G. Karypis, "Slpminer: An Algorithm for Finding Frequent Sequential Patterns Using Length-Decreasing Support Constraint," Proc. IEEE Second Int'l Conf. Data Mining (ICDM '02), pp. 418-425, 2002.