RESEARCH ARTICLE OPEN ACCESS

CROSS-DOMAIN SENTIMENT CLASSIFICATION USING WEB USAGE MINING

[1]Mrs. Saraswathi.S, [2]Mrs. Anette Regina.I.

[1]M.phil Research Scholar, Department of computer Science Muthurangam Government Arts College(Autonomous), Vellore, Tamilnadu, India.

[2]Associate Prof, Department of Computer Science Muthurangam Government Arts College(Autonomous), Vellore.

----------**********----------

Abstract–

Web usage mining is crucial for the Cross-Domain Sentiment Classification (CDSC) as it can ensure customer satisfaction as far as the interaction between the customer and the organization is concerned. Web usage mining is also helpful for identifying or improving the visitors of a particular Website by accessing the log file of that site. In this paper the focus is on Web usage mining of Log data of an educational institution. Web usage mining (WUM) also known as Web Log Mining is the application of Data Mining. WUM techniques are applied on large volume of data to extract useful and interesting patterns from Web data, specifically from web logs, in order to improve web based applications. Web usage mining consists of four phases, data source, pre-processing, pattern discovery, and pattern analysis. After the completion of these four phases the user can find the required usage patterns and use this information for the specific needs in a variety of ways such as improvement of the Web application, identifying the visitor's behaviour, customer attraction, Customer retention etc..

Keywords **— Cross-Domain Sentiment Classification (CDSC) Web usage mining (WUM).Customer attraction, Data mining Techniques**

----------**********----------

I. INTRODUCTION

1.1 OVERVIEW OF THE SYSTEM

With the increase in internet based services, people express their opinions about products online. Such sentiment information obtained from the customers is growing exponentially. Thus making it difficult for the manufacturer to classify the nature of the reviews manually. An automatic sentiment classifiers classification of reviews into positive or negative based on the sentiment words expressed in documents which is necessary to be developed for the manufacturer and the customer in order to analyze the reviews of the customers. The goal of sentiment classification is to discover customer opinion on a product. Sentiment classification has been applied in various tasks such as opinion mining, market analysis, opinion summarization and contextual analysis.[1]

Specific domain is used in sentiment analysis to provide greater accuracy. Sentiment analysis uses feature vector that has a collection of words which are limited and specific to particular domain (domain can be consider as student, school etc.). However sentiments hold different meanings in different domains and it is costly to annotate data for each new domain in which we would like to apply a sentiment classifier. Cross domain sentiment analysis can be considered as the solution to this problem but the problem is that classifier trained in one domain may not work well when applied to other domain due to mismatch between domain specific words. So before applying trained classifier on target domain some techniques must be applied like feature vector expansion, finding relatedness among the words of source and target domain, etc. A different technique gives different analysis, result and accuracy which depend on the documents, domain taken into consideration for classification. Sentiment Classification is an important task in various applications such as Opinion Mining, Opinion Summarization and Contextual Advertisement. Sentiment Analysis has been used to help political strategies gauge public opinion on the Internet as Yahoo News shows (Weber).

1.1.1 Project Scope and Objectives

The objective of the project is to generalize the log file of a web site obtained from a Web Server using Attribute-Oriented Induction technique.

Context:

This can be used for identifying the frequent access pattern for any web site. Accordingly, website can be enhanced.

Web Mining:

*Web mining*can be broadly defined as the discovery and analysis of useful information from the World Wide Web. This broad definition on the one hand describes the automatic search and retrieval of information and resources available from millions of sites and on-line databases, i.e., *Web content mining*, and on the other hand, the discovery and analysis of user access patterns from one or more Web servers or on-line services, i.e., *Web usage mining*.

.

II. LITERATURE SURVEY

The automatic analysis of user generated contents such as online news, reviews, blogs and tweets can be extremely valuable for tasks such as mass opinion estimation, corporate reputation measurement, political orientation categorization, stock market prediction, customer preference and public opinion study. Also, new challenges raised by sentiment aware applications are addressed [9]. Sentiment classification systems can be broadly categorized into single domain and cross domain classifiers based upon the domains from which they are trained on and subsequently applied to.

In [2] 2013, DanushkaBollegala et al. [4] developed a technique which uses sentiment sensitive thesaurus (SST) for performing cross domain sentiment analysis. They proposed a cross-domain sentiment classifier using an automatically extracted sentiment sensitive thesaurus. To overcome the feature mismatch problem in cross domain sentiment classification, they use labeled data from multiple source domains and unlabeled data from source and target domains to compute the relatedness of features and construct a sentiment sensitive thesaurus. Then use the created thesaurus to extend feature vectors during train and test times for a binary classifier. Spectral feature alignment (SFA) method is first proposed by Pan et al. [5] in 2010. In this, features are classified as to domain specific or domain independent using the mutual information between a feature and a domain label. Both unigrams and bigrams are considered as features to represent a review. Next, a bipartite graph is constructed between domain specific and domain independent features. An edge is formed between a domain specific and a domain independent feature in the graph if those two features co occur in some feature vector. Spectral clustering is conducted to identify feature clusters. Finally, a binary classifier is trained using the feature clusters to classify positive and negative sentiment

SCLMI. This is the structural correspondence learning (SCL) method proposed by Blitzer et al. [6]. This method utilizes both labeled and unlabeled data in the benchmark data set. It selects pivots using the mutual information between a feature (unigrams or bigram
s) and the domain label. Next, binary classifiers are trained to predict the existence of those pivots. The learned weight vectors are arranged as rows in a matrix and singular value decomposition (SVD) is performed to reduce the dimensionality of this matrix. Finally, this lower dimensional matrix is used to project features to train a binary sentiment classifier.

In single domain sentiment classification, a classifier is trained using labeled data annotated from the domain in which it is applied. An investigation is d one to determinewhether it is sufficient to treat sentiment classification simply as a special case of topic based categorization or whether special sentimentcategorization methods need to be developed[8]This approach used three standard algorithms: Naive Bayes classification, maximum entropy

2.1 EXISTING SYSTEM

Supervised learning algorithms that require labeled data have been successfully used to build sentiment classifiers for a given domain [1]. However, sentiment is expressed differently in different domains, and it is costly to annotate data for each new domain in which we would like to apply a sentiment classifier. For example, in the electronics domain the words “durable” and “light” are used to express positive sentiment, whereas “expensive” and “short battery life” often indicates negative sentiment. On the other hand, if we consider the books domain the words “exciting” and “thriller” express positive sentiment, whereas the words “boring” and “lengthy” usually express negative sentiment. A classifier trained on one domain might not perform well on a different domain because it fails to learn the sentiment of the unseen words. Model the cross-domain sentiment classification problem as one of feature expansion, where we append additional related features to feature vectors that represent source and target domain reviews to reduce the mismatch of features between the two domains. Methods that use related features have been successfully used in numerous tasks such as query expansion [9] in information retrieval [10], and document classification [11]. For example, in query expansion, a user query containing the word car might be expanded to car OR automobile, thereby retrieving documents that contain either the term car or the term automobile. However, to the best of our knowledge, feature expansion techniques have not previously been applied to the task of cross domain sentiment classification. The proposed method can learn from a large amount of unlabeled data to leverage a robust cross domain sentiment classifier.

2.2 PROPOSED SYSTEM

Web usage mining mines web log records to discover web access pattern of web pages. Analyzing and exploring identifying potential customers for e-commerce enhance the quality and delivery of internet information services to end user and improve web server system performance.

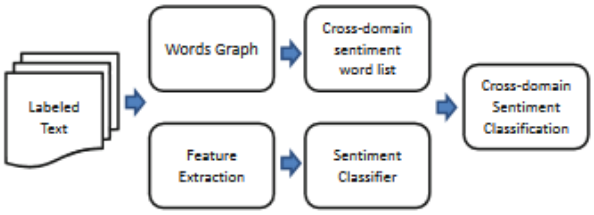

Fig : Cross Domain Classification

A general solution to sentiment classification is developed to address the cross domain problem [7]. In this problem, the systems do not have any labels in a target domain but have some labeled data in a different domain, regarded as source domain. In this cross domain sentiment Classification setting, to bridge the gap between the domains, a Spectral Feature Alignment (SFA) algorithm is proposed to align domain specific words from different Domains into unified clusters, with the help of domain independent words as a bridge. In this way, the clusters can be used to reduce the gap between domain specific words of the two domains, which can be used to train sentiment classifiers in the target domain accurately.The training time complexity of this classifier is linear to the number of training data and the space complexity is also linear to the number of features, thus it makes this learning technique both time and storage efficient.

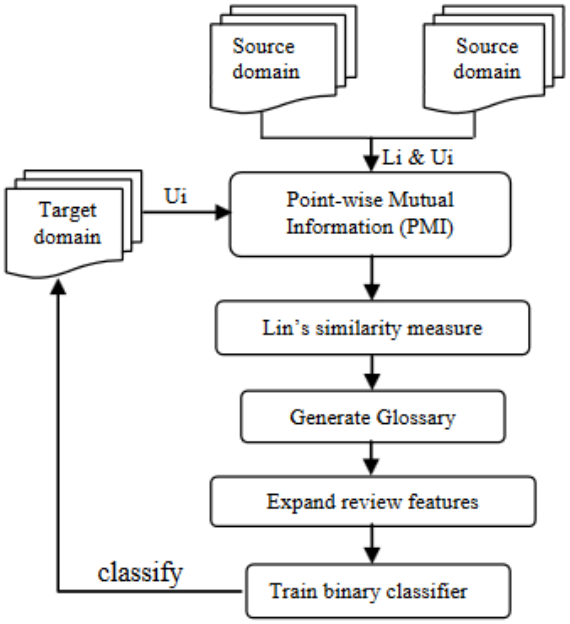

Figure 2.1: System Architecture

LOG FILE

Log files are files that contain a record of website activity. Every time a person visits the website, a log file is updated with the visitor's information by the web server. These log files can be downloaded and used to generate useful statistics.

An access of a web page or a file will generate a "Hit" on the web server. For example, if a web page contains 10 pictures, a visit on that page will generate 11 "hits" on the web server, one hit for the web page, 10 hits for the pictures. If a visitor viewed 5 web pages on the web site, each page contain 10 pictures, the web server will record:

WEBLOG FILES

Web Server log files are simple text files that are automatically generated every time someone accesses the Website. Every "hit" of the Web site, including each view of a HTML document, image or other object, is logged. The raw web log file format is essentially one line of text for each hit to the website. This contains information about who was visiting the site, where they came from, and exactly what they were doing on the particular Web site.

2.2.1 ADVANTAGES OF PROPOSED SYSTEM

- GUI Representation of Web Performance with the use of Web Charts
- GUI Representation of Network performance with the use of tables and charts.
- Advantage of save the graphs (Line Chart & Pie Chart).
- Monitoring the system performance using the tables.

V. IMPLEMENTATION OF WUM

DETAILED PROCESS OF WUM

Step 1: Data preprocessing

Data preprocessing has a fundamental role in Web Usage Mining applications. It has different tasks:

(a) **Data Cleaning**-This step consists of removing all the data tracked in web logs that are useless for mining purposes.

(b) **Session Identification and Reconstruction-**This step consists of (i) identifying the different users' sessions from the usually very poor information available in log files and (ii) reconstructing the users' navigation path within the identified sessions.

(c) **Content and Structure Retrieving-***Web content refers to the discovery of useful information from web contents including text, image, audio and video etc., structure retrieving gives the analysis of the out links of a webpage and it has been used for search engine result ranking.*

(d) **Data Formatting** - Once the previous phases have been successfully completed, data are properly formatted before applying mining techniques. So stored data extracted from web logs into a relational database.

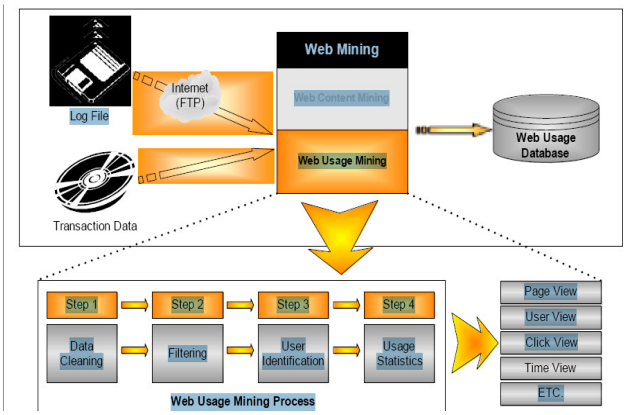

Figure : Phases of WUM

Step 2: Mining Algorithms

Process of mining algorithm or pattern discovery:

(a)**Statistical Analysis:** Statistical techniques are the most common method to extract knowledge about visitors to a Web site. By analyzing the session file, one can perform different kinds of descriptive statistical analyses (frequency, mean, median, etc.) on variables such as page views, viewing time and length of a navigational path. Many Web traffic analysis tools produce a periodic report containing statistical information such as the most frequently accessed pages, average view time of a page or average length of a path through a site.

(b)**Clustering:**Clustering is a technique which groups together a set of items having similar characteristics. In the Web Usage domain, there are two kinds of interesting clusters to discover. (i.e.) usage clusters and page clusters. Clustering of users tends to establish groups of users exhibiting similar browsing patterns. Such knowledge is especially useful for inferring user demographics in order to perform market segmentation in E-commerce applications or provide personalized Web content to the users.

(c)**Classification:** Classification is the task of mapping a data item into one of several predefined classes. In the Web domain, one is interested in developing a profile of users belonging to a particular class or category. This requires extraction and selection of features that best describe the properties of a given class or category.

(d)**Association Rules:** Association rule generation can be used to relate pages that are most often referenced together in a single server session. In the context of Web Usage Mining, association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold.

(e)**Sequential Patterns:** The technique of sequential pattern discovery attempts to find inter-session patterns such that the presence of a set of items is followed by another item in a time-ordered set of sessions or episodes. By using this approach, Web marketers can predict future visit patterns which will be helpful in placing advertisements aimed at certain user groups.

(f)**Dependency Modeling:** Dependency modeling is another useful pattern discovery task in Web Mining. The goal here is to develop a model capable of representing significant dependencies among the various variables in the Web domain.

Step 3: Pattern Analysis

Pattern analysis is the last step in the overall Web Usage mining process as described The motivation behind pattern analysis is to filter out uninteresting rules or Patterns from the set found in the pattern discovery phase. The exact analysis methodology is usually governed by the application for which Web mining is done. The most common form of pattern analysis consists of a knowledge query mechanism such as SQL.

A lot of work is done in sentiment analysis field using any social website like twitter and many techniques are devised to improve accuracy of classification of social sentiments. The future work gives us idea about improving the classification and accuracy of social data. Mainly good optimizations can be done in classification part using good classifiers. Topic modeling is one area where limited work is done and also it is not applied at a big scale on social data like twitter sentiments. Topic modeling is one area which can help to divide large social data into categories by building an intelligent system. Apart from this summarization of data is another field which can be explored at a bigger level. Both topic modeling and summarization can be applied together to generate a intelligent system that can be useful in giving useful information out of a huge bulk of data from social web.

CONCLUSION

Web Usage Mining is an active field for research and Web Usage Mining applications are being used in some famous Websites. This project presents an implementation of the Web Usage Mining. Web Server log files are mined in order to analyze the Web Usage pattern. The methodology employs *Data Preprocessing, Mining Algorithms and Pattern Analysis.* Data Processing phase for the Web Usage Mining is a challenging task. By applying mining algorithms to the Web log file, the relationship between the accessed pages can be mined. The Web usage patterns and user behavior are analyzed by using the

mining algorithms. The results from this project can be used by Web administrator and Web masters in order to improve Web services and performance through the improvement of Web sites, including their contents, structure, presentation and delivery. The current remote monitoring tools are based on the SNMP protocol. Most of the commercial network components have embedded SNMP agents. Because of the universality of the Internet with TCP/IP protocol, the transport of management information for SNMP management, which is TCP/IP based is resolved automatically. In addition, most of the popular host operating systems come with the TCP/IP suite and thus are amenable to SNMP management.

REFERENCES:

B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentimentclassification using machine learning techniques," in EMNLP2002, 2002, pp. 79–86.14

[2] P. D. Turney, "Thumbs up or thumbs down? semantic orientationapplied to unsupervised classification of reviews," in ACL 2002,2002, pp. 417–424.

[3] B. Pang and L. Lee, "Opinion mining and sentiment analysis,"Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp.1–135, 2008.

[4] Y. Lu, C. Zhai, and N. Sundaresan, "Rated aspect summarizationof short comments," in WWW 2009, 2009, pp. 131–140.

[5] T.-K. Fan and C.-H.Chang, "Sentiment-oriented contextual advertising,"Knowledge and Information Systems, vol. 23, no. 3, pp.321–344, 2010.

[6] M. Hu and B. Liu, "Mining and summarizing customer reviews,"in KDD 2004, 2004, pp. 168–177.

[7] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood,boom-boxes and blenders: Domain adaptation for sentiment classification,"in ACL 2007, 2007, pp. 440–447.

[8] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, "Cross-domainsentiment classification via spectral feature alignment," in WWW2010, 2010.

[9] H. Fang, "A re-examination of query expansion using lexical
resources," in ACL 2008, 2008, pp. 139–147.

[10] G. Salton and C. Buckley, Introduction to Modern Information Retreival.McGraw-Hill Book Company, 1983.