ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

Forum Reply Classification Using Clustering

P.Kalaiarasi [1], R.Rooba[2]

Assistant Professor, Kongu Arts and Science College, Erode, India[1,2]

Abstract: Due to richness of information in forums, researchers are increasingly interested in mining knowledge from forums. From this observation, the forum posts and replies are clustered and analyzed in order to improve the user knowledge in the field. To harvest knowledge from the forum the contents must be downloaded. Forum board or thread is usually divided into multiple pages which are linked by page flipping links. The forum sites contain different pages like entry pages, thread pages and page flipping. The forum mining have three phases: preprocessing, mining the data by applying various data mining strategies such as clustering and post processing. In preprocessing raw data is transformed into a usable format, mainly by parsing and cleaning. While preprocessing, the pages are downloaded as the html file and the files are invoked into parsing and assign attributes like forum id, forum title, thread count, post count. The parsing process is accomplished; data cleaning process is applied to the downloaded post sets and automatically remove noise data and irrelevant data. Clustering algorithm is applied for the preprocessed data to groups the forums into various clusters. The clustering is accomplished by using all topics and sub topics of the forum. The four dimensions of clustering are number of posts/topics, average sentiment values/topics, positive percentage of posts/topics and negative percentage of posts/topics. The posts/topics dimension are determined by number of replies for a post, the sentiment values of this topics are identified from user replies, it describe the user opinion, the positive and negative dimensions are determined from user replies, describe the user perception in the posts. The positive and negative dimensions are also used to identifying the user attitude and pros and cons of the specific topics are discussed in the particular forum. In the post processing stage numbers of clusters are obtained. The obtained final clusters are grouped based on the topics with similar sentiment values and user opinions. Based on the sentiment values, the positive and negative posts are clustered for each thread. Information seekers, decision makers can benefit from this clustering. It simplifies the decision making process.

Keywords: Clustering, Forum, Graph, Threads.

I. INTRODUCTION

Data mining, also popularly referred to as knowledge discovery from data (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored or catchable in large databases, data warehouses, the Web, other massive information repositories, or data streams.

A. Data Preprocessing

Incomplete, noisy and inconsistent data are commonplace properties of large real world databases and data warehouses. For this issue, the database is invoked into preprocessing using some data mining techniques[1]. Data preprocessing includes data cleaning, data integration, data transformation and data reduction[2].

Why preprocessing?

The real-world data that is to be analyzed by data mining techniques are:

i) ***Incomplete:*** lacking attribute values or certain attributes of interest, or containing only aggregate data. Missing data, particularly for tuples with missing values for some attributes, may need to be inferred.

ii) Noisy: containing errors, or outlier values that deviate from the expected. Incorrect data may also result from inconsistencies in naming conventions or data codes used or inconsistent formats for input fields, such as date. It is hence necessary to use some techniques to replace the noisy data.e.g. Salary=‖-10‖

iii) Inconsistent: containing discrepancies between different data items. Some attributes representing a given concept may have different names in different databases, causing inconsistencies and redundancies. Naming inconsistencies may also occur for attribute values. The inconsistency in data needs to be removed.e.g. age=‖12‖ and birthday=‖11/6/1990‖

iv) Aggregate Information: It would be useful to obtain aggregate information such as to the sales per customer region—something that is not part of any pre-computed data cube in the data warehouse.

v) Enhancing mining process: Large number of data sets may make the data mining process slow. Hence, reducing the number of data sets to enhance the performance of the mining process is important.

vi) Improve Data Quality: Data preprocessing techniques can improve the quality of the data, thereby helping to improve the accuracy and efficiency of the subsequent mining process. Data preprocessing is an important step in the knowledge discovery process, because quality decisions must be based on quality data. Detecting data anomalies, rectifying them early, and reducing the data to be analyzed can lead to huge payoffs for decision making.

B. Clustering

Clustering is the task of grouping a set of objects[3]. Objects in the same group are called as a cluster. It is a main task of exploratory data mining, and

DOI 10.17148/IJARCCE.2015.4207

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

C. Forums

An Internet forum, or message board, is an online discussion site where people can hold conversations in the form of posted messages. A discussion forum is hierarchical or tree-like in structure: a forum can contain a number of sub forums, each of which may have several topics.

II. A SURVEY ON FORUMS

Mining User Experiences

Valentin Jijkoun[4] and his research group discussed that users of online forums are interested in other people's experiences with concrete products and /or solutions for specific problems like opinion retrieval[5] and mood detection[6].

User Grouping Behavior

Xiaolin Shi and his research group focus to characterizing user grouping behavior in online social environments[7]. This characterizing not only help researchers to understand many of sociological problems of human behavior, also facilitates them to improve various applications in the online environment[8]. This type of environment have a rich complexity[9].

To identifying the relationship between these factors, users and communities were taken as a two set of nodes and construct a bipartite graph instead of decision trees [10], to encompass all the features and their relationships. Based on the bipartite graph, build a Bipartite Markov Random Field [Bimorph] [11][12][13] model to quantitatively evaluate how much each feature affects the grouping behavior in online forums, as well as their relationships with each other.

III.RELATED WORK

Clustering the forum threads based on positive and negative replies are analyzed and addressed in this paper. The proposed approach includes group the forums into various clusters using emotional polarity computation and integrated sentiment analysis based on K-means clustering. Also positive and negative replies are clustered. Using scalable learning the relationship among the topics are identified and represent it as a graph. Data are collected from forums.digitalpoint.com which includes a range of 75 different topic forums. Computation indicates that within the same time window, forecasting achieves highly consistent results with K-means clustering. Also the forum topics are represented using graphs[14]. In this graph the node represent the topics and the edge represent the similarity or relationship between the topics

IV. PROPOSED METHODOLOGY

A. Forum Topic Download

In Forum Topic Download, the source web page is keyed in (default: http://www.forums.digitalpoint.com) and the content is being downloaded[15]. The HTML content is displayed in a rich text box control.

B. Parse Forum Topic Text And Urls

In Parse Forum Topic Text And Urls, the downloaded source page web content is parsed and checked for forum links. The links are extracted and displayed in a list box control. Also the link text are extracted and displayed in another list box control.

C. Forum Sub Topic Download

In Forum Sub Topic Download, all the forum links pages in the source web page are downloaded. The HTML content is displayed in a rich text box control during each page download.

D. Parse Forum Sub Topic Text And Urls

In Parse Forum Sub Topic Text And Urls, the downloaded forum pages web content are parsed and checked for sub forum links. The links are extracted and displayed in a list box control. Also the link text are extracted and displayed in another list box control.

E. Preprocess

In Preprocess, the downloaded forum pages web content are preprocessed and assign the attributes like forumid, forum subid, forum topic, forumurl.

It has the following steps:

Step 1. Download the content from website "forums.digitalpoint.com". In this step, the information are extracted and downloaded from specified website. The content are parse in to different concept such as forum topics, forum subtopics, forum post as HTML file.

Step 2. After download the content, the html files are converted into text file for the purpose of obtaining different forum topic and subtopics which are reside in the html file.

Step 3. To convert text file data in to dataset.

Step 4. Data cleaning

When the parsing process is accomplished, data cleaning process is applied to the downloaded data sets. In this phase, automatically remove noise data and irrelevant data. Bag of words like stem word, stop word and synonym words are used to remove the noise and irrelevant data.

Identifying and removing Stop words: The raw data of documents contain non-informative words , conjunction since they are frequent and carry no information. e.g., _a', 'and', 'what', 'when', 'the' etc.

Stem words: In Stemming process root word will find out by removing prefixes and suffixes of the word. For example, work, working, works are all stemmed to work, and walker, walked, walking are all stemmed to walk.

DOI 10.17148/IJARCCE.2015.4207

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

Create Graph

In Create Graph, forum topics are created as a node. The name of the node is coined automatically. The name should be unique. Links between the nodes represent the relationship between the topics. The link can be created by selecting starting and ending node; a node is linked with a direction. The link name given cannot be repeated. The constructed graph is stored in database. Previous constructed graph can be retrieved when ever from the database.

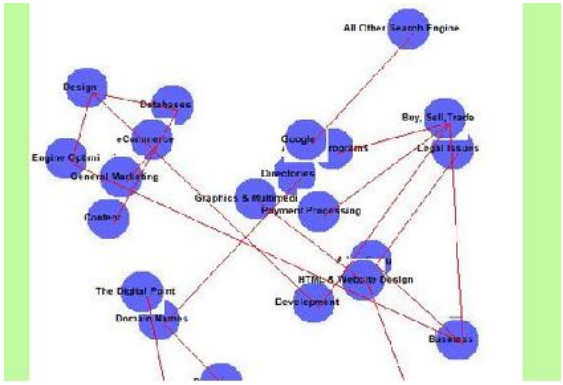

Fig -1Create graph using forum topics and its relatioship

Algorithm Of Scalable K-Means Variant

In Algorithm Of Scalable K-Means Variant, the data instances are given as input along with number of clusters, and clusters are retrieved as output. First it is required to construct a mapping from features to instances. Then cluster centroids are initialized. Then maximum similarity is given and looping is worked out. When the change is objective value falls above the 'Epsilon' value then the loop is terminated

Show Classification

In the Show Classification pre processed forum data is clustered using four different dimensions. The clustering is accomplished by using all topics and sub topics of the forum. The four dimensions of clustering are number of posts/topics, average sentiment values/topics, positive percentage of posts/topics and negative percentage of posts/topics. The posts/topics dimension are determined by number of replies for a post, the sentiment values of this topics are identified from user replies, it describe the user opinion, the positive and negative dimensions are determined from user replies, describe the user perception in the posts. After parsing positive and negative replies are determined. The positive and negative dimensions are also used to identifying the user attitude and pros and cons of the specific topics are discussed in the particular forum.

V. RESULTS

The K-means clustering is applied to group the posts and its replies. Posts are found to be valid only if they contain minimum three words. Posts are grouped and average value taken for each threads; likewise threads average sentiment values are taken and forum's average value is derived. It is found that more of the forums sentiment values are having more positive percent and less negative percent.

VI.CONCLUSION AND FUTURE ENCHANCEMENT

This paper has proposed relationships among the topics. It is determined using scalable learning and represented as a graph. This is used to send the same thread in related topics and get the replies in various dimensions based on the topics. This K-means clustering is applied to forums cluster analysis. Clustering algorithm is applied to group the forums into various clusters. The obtained final clusters grouped based on the topics with similar sentiment values and user opinions. Based on the sentiment values, the positive and negative posts are clustered for each thread. Information seekers, decision makers can benefit from this clustering. It simplifies the decision making process. At present, number of posts/forum, average sentiment values/forums, positive percentage of posts/forum and negative percentage of posts/forums are taken as feature spaces for K-Means clustering. In future, neutral replies, multiple-languages based replies can also be taken as dimensions for clustering purpose. In addition, currently forums are taken for hot spot detection. Live Text streams such as chatting messages can be tracked and classification can be adopted.

REFERENCES

[1] Guidelines: http://ptucse.loremate.com/dw/node/4
[2] Jiawei Han, Micheline Kambar and Jian Pei,‖Data mining concepts and techniques‖, Third Edition, Rajkamal Electric Press, Haryana, 2012.
[3] P.Berkhin,‖Survey of Clustering Data Mining Techniques‖. Technical report, Accrue Software, San Jose, Cailf, 2002.
[4] Valentin Jijkoun, Maarten de Rijke,Wouter Weerkamp, Paul Ackermans Gijs Geleijnse, "Mining User Experiences from Online Forums: An Exploration". In Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media.
[5] I. Ounis, C. Macdonald, M. de Rijke, G. Mishne, andI. Soboroff. "Overview of the TREC" 2006 Blog Track.In TREC 2006, 2007.
[6] K. Balog, G. Mishne, and M. de Rijke. "Why are theyexcited?: identifying and explaining spikes in blogmood levels". In EACL '06, pages 207–210, 2006.
[7] Xiaolin Shi, Jun Zhu, Rui Cai, Lei Zhang, "User Grouping Behavior in online forums", KDD'09.
[8] X.Song, B.L. Tseng, C-Y. and M.-T Sun, "Personalized recommendation driven by information flow" in SIGIR,2006.
[9] V. G_omez, A. Kaltenbrunner, and V. L_opez. "Statistical analysis of the social network and discussion threads in Slashdot". In WWW, 2008.
[10] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. "Group formation in large social networks: membership,growth, and evolution". In SIGKDD, 2006.
[11] O. Frank and D. Strauss. Markov graphs. "Journal of theAmerican Statistical Association", 81:832{842, 1986}.
[12] G. Robins, P. Elliott, and P. Pattison. "Network models for social selection processes. Social Networks", 23:1{30, 2001.
[13] S. Wasserman and P. Pattison. " Logit models and logisticregressions for social networks, I. an introduction toexponential random graphs and (p*)". Psychometrika,62:401{425, 1996.
[14] Lei Tang, Xufei Wang, and Huan Liu, " Scalable Learning of Collective Behavior".
[15] Jingtian Jiang, Xinying Song, Nenghai Yu, Chin – Yew Lin, " FOCUS : Learning to crawl web forums", IEEE Transactions on knowledge and Data Engineering.

DOI 10.17148/IJARCCE.2015.4207

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

BIOGRAPHIES

Ms.P.Kalaiarasi received M.C.A Ms.P.Kalaiarasi received M.C.A degree from Bharathiar University, Coimbatore,TN, India. She is currently working as a Assustant Professor in Kongu Arts and Science College,Erode, TN, India. She has 5 Years of teaching experience. Her Area of interest is Data Mining.

R.Rooba M.Sc(CT).,M.phil., Currently Working as a Assistant Professor in the Department of Computer Technology and Information Technology at Kongu Arts and Science College , Erode. Currently doing Ph.D in Bharathiar University, Coimbatore. Her area of interest is Data Mining, Semantic Web Mining, Big data and Cloud Computing. She presented 10 papers in various national and international conferences and published 3 papers in International journals.