International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Review on Analysis of User Behavior Based on Prediction Algorithm

Hemlata Kardas[1], Sulochana Sonkamble[2]

[1]Rajarshi Shahu College of Engineering and Research, JSPM Narhe, Technical Campus, Pune-India

Abstract: ***Various service provider systems or recommender systems have to manage the large amount of data regarding customer, service and information. This leads to Big Data analysis problem and its handling is also a challenging task. There are various organizations which serves user through service provider systems or recommender systems that provide efficient services in terms of set of good quality products or services to use by giving appropriate data as per their need. Hence the Decision Support System is used to determine the decision making activities in various organizations of recommender system which allows making certain planning, management and operations. This also helps to make decisions for structured and unstructured problem. It can be done in number of ways like descriptive, decision and prediction based. This paper will focus on Predictive Analysis of user behavior towards recommender system. The predictive analysis deals with the identification of relationship between specific performance of unit with respect to set of features or attributes defined for the system. Efficient service provider system have to determine preferences of user towards service through ratings, ranking or free-text given by them like reviews posted by user over internet. This is useful to get by understanding about user's behavior for particular system. This can be done through opinion mining, review analysis and sentiment analysis. The text mining allows mining of features and characteristics of product or service efficiently and further processing of features can be done by defining efficient predictive model for it. Hence, this data will be available in plenty of amount which must be processed by using new innovative framework called Hadoop. The Hadoop technology provides facility of Map Reduce mechanism to process large amount of data. The mined data will be processed through various techniques using prediction method. There are various existing methodologies that involve analysis based on probability model, product aspect ranking method, cross-domain sentiment classification, estimating helpfulness from set of data, and dynamic interaction using mashup tools. These methods can be implemented with Hadoop for efficient result.***

Keywords: Predictive Analysis, Opinion Mining, Semantic Analysis, Cross-Domain, Product Aspect, Map Reduce.

1. Introduction

Now-a-days there is increase in amount of customer, services and online information for any recommender system or service provider system. This causes Big Data formation which is difficult to analyze and process [14]. Hence there is need to provide scalable and efficient service to user by understanding their preferences for system. The decision support system has the predictive analysis mechanism for decision making activities of various organizations to improve their performance about the service they provide. Web usage mining also gives knowledge about user's behavior towards service by extraction of interesting patterns for analysis. There are various methodologies are mentioned by different authors to predict the behavior of user for service through internet. Many of the authors focus on the product aspect or features and sentimental behavior like their post on internet for analysis purpose. They have used text mining for mining the opinions, reviews or sentiments. The mining can be distinguished into number of ways like Domain-Driven data mining, review mining or ranking and recommender system. One of the recommender systems like Hotel Reservation System which analyzes user's preferences by collecting the reviews posted to get personalized list of hotels as per their requirement. This system is based on collaborative filtering algorithm where it recommend services to the user that users with similar tastes preferred in the past. It supports Hotel Reservation System through keyword based algorithm designed for it which extracts set of keywords commonly used while searching good hotels by user. Then it performs processing on keyword set by comparing with users ratings and ranking for it. This algorithm works on Hadoop technology for efficient result [1]. Some of the author did retrieval of product aspects and opinion through customer reviews generating product feature using unigram language model. Then they did the mapping of opinion to product feature [2].One of the method focused on ranking product features by designing product ranking framework to identify aspects and define semantic classifier with probabilistic model for comparing the opinion with overall opinion [3]. Some author defined the cross-domain thesaurus by using semi supervised data which is used to generate domain thesaurus randomly. [4]. Some authors have estimated sales impact performance based on reviews by computing helpfulness value using votes given by user for that product [5]. Specifically, some authors worked on generating mashup tool for processing user behavior using interaction, confidence, diagnostic, intention to analyze decision quality [6]. Some authors did multi criteria user modeling to improve the quality of service [7] and online review processing for movie domain to understand public sentiment and business intelligent [9].

2. Literature Review

The focus of the literature survey is to study and collect the information of user behavior from reviews or opinion based on semantics for service system and features of domain of service system. Shunmei Meng et al. focused on keyword based service recommender system which analyzes present and past user's behavior searching best hotel list as per their requirement through reviews posted by users. It actually dose preprocessing of HTML for collecting set of keywords like food, accommodation ,location etc. to form candidate set

which is fed to approximate and exact similarity computation algorithm along with preferences of current and past user reviews. This also defines the domain thesaurus of candidate set i.e. set of words having same domain like location having hill station, greenery or waterfall and accommodation having AC, NON-AC like. Here current user has to select the keywords from set along with importance value given to it. These algorithms does processing of given data using *Jaccard* and *sim* function comparing with some threshold value respectively to provide personalized list of hotels to users.

Here it also defines new model called Analytic Hierarchy Process Model to have exact matching computation. This AHP model used in exact similarity computation which constructs matrix for finding relative importance between two keywords and using weight vector function on the importance of keywords and current users preferred keyword apply TF/IDF computation for weight measurement of keywords. This paper work on Map Reduce framework for efficient processing of data collected using Keyword Based algorithm [1].

Lisette García-Moya et al. focused on identification of aspects or feature of product from customer's reviews about product features, semantic classification from opinion of customer and aspect ranking is identifying relevance of aspect and opinion. This system considers stochastic mappings between words to estimate a unigram language model of product features. It determines the probabilistic model for mapping opinion to product feature by retrieving words from reviews based on co-occurrence vale and refining them. Finally evaluation of retrieval is done using HITS method [2].

Zheng-Jun Zha et al. focused on extracting product feature and ranking them using probabilistic aspect ranking based on overall opinion rating by weighted aggregation of opinion on aspects. This model uses importance of aspect while aggregating words. This system uses *sim* function for product aspect identification for finding occurrence of words from noun and phrases. Also uses language model for scoring the aspect and semantic classifier which parses the opinion of user and generate set of aspect using lexicon method [3].

Danushka Bollegala et al. focused on automatic classification of semantic for various applications for opinion mining. This system constructs sentiment sensitive distributional thesaurus by labeling source data and unlabeled data in target domain for cross domain sentiment classification. It also expands the feature vector for enhancing the domain thesaurus. This system uses semi supervised method for classification of domain. The relatedness of reviews is computed using POS tagging and unigram, bigram model [4] [10] [11].

Anindya Ghose et al. focused on quality determination of reviews by text mining. Random Forest-based classifiers method is used for predicting reviews based on the decision process system which uses "*helpfulness*" value. Reviews basically have objective and subjective features used for computing *helpfulness*. Here objective features are nothing but characteristics or description of product by merchant and subjective features means reviews or personal opinion by customer. It then finds the probability of occurrence of text words in subjective and objective features which are used to identify sales rank based on reviews. The *helpfulness* value is computed by ratio of votes to total votes received for product [5].

Brandon A. Beemer et al. focused on dynamic interaction of user for determining the quality of service of product by revising and revisiting the inputs for changing decisions. The relationships between dynamic interactions, diagnostic, confidence, and intention are analyzed here using mashup tool. Post hoc analysis of decision quality suggests that increased levels of dynamic interaction also improve the overall quality of the decision made [6].

Kleanthi Lakiotaki et al. defined set of phases for user modeling is with first phase data acquisition gathers data in terms of numerical rating and ranking forming data matrix. Second phase multi criteria user modeling does aggregation of multi criteria data using UTA method providing weight vector with user modeling.

Third phase is clustering of data object and their relationship with each other using k-means algorithm. There set of clusters are fed to fourth phase called recommendation where it applies sim function on clusters to identify the similarity of users behavior using collaborative filtering algorithm. Here multi criteria similarity computation is based on multiple matrices like statistical or classification accuracy [7].

Xiaohui Yu et al. focused on mining of online reviews of movie domain based on prediction. This paper determines the sentiment expressed form reviews and quality of reviews for predicting impact on product sales performance. This system uses probabilistic approach to sentiment mining for number of blog records for box office information, user ratings. Then feature selection is done for collected data for finding relevance between numbers of occurrence of review data. This system also does the time series analysis of move domain based on Auto regression model by prediction method [9].

3. Problem Definition

The KASR (Keyword Aware Service Recommender System) method is used for processing ratings and ranking of user preferences for Hotel Reservation System to get personalized list of hotels [1]. It does not distinguish positive and negative preferences of user and do not process free text reviews posted by user also. KASR method is limited for Hotel Reservation system. It must work for other domain adaptation also like Mobile, movie domain etc.

4. Methodology

The proposed system will work on reviews posted by user for any recommender system. The reviews collected as dataset will be processed collectively under prediction algorithm like KASR (Keyword Aware Service Recommender System) method [1]. Here Domain adaptation is used which is

generation of domain thesaurus which works to deal with current user and previous user preferences by text mining. This methodology will work with domain thesaurus defined for different recommender systems and also work on distinguishing the positive and negative reviews of user. The implementation for dynamic review processing using Hadoop technology is used for efficient result of user behavior. The preferences of previous user posted in terms of ratings and ranking or free-text is used to get personalized list of required service and is compared with current user preferences. The map reduce method will perform the functionality of computations needed to process features of services and users reviews.

The set of computations for dataset collection and comparison are:

a. Text mining by preprocessing the HTML pages of previous user reviews to collect keyword dataset by mining words as per recommender system and here define domain thesaurus for different domain with set of negative keywords for distinguishing positive and negative preferences.
b. TF/IDF computation for frequency of occurrences of keyword set from current user and previous user preferences.
c. Weight computation of keywords by assigning importance to keywords based on words commonly used by user.
d. Cosine similarity function and Jaccard coefficient can be used for comparison of keyword set.
e. Based on set of keywords left after above comparison processing we can again perform the processing of dataset for efficient result using normalizing factor.

Above strategy can be used with Map Reduce framework of Hadoop for handling large review dataset.

5. Overview of Hadoop and Map Reduce Framework

The Hadoop environment supports for big data processing up to terabytes to petabytes. The analysis of such big data files must be reliable, efficient and fault tolerant with respect to handling data which is possible by Hadoop phenomenon, because it provides two major facilities like one is HDFS system. The Hadoop Distributed File System (HDFS) that has NameNode to store data, and multiple DataNodes (servers) and data blocks for storing detailed information that is metadata on Hadoop clusters. HDFS creates several replications of the data blocks and distributes them accordingly in the cluster in way that will be reliable and can be retrieved faster. A HDFS block size is 64MB. Every data block is replicated to multiple nodes across the cluster. Hadoop will internally make sure that any node failure will never results in a data loss. It has one NameNode that manages the file system metadata. There will be multiple DataNodes that will store the data blocks with actual set of data [13].

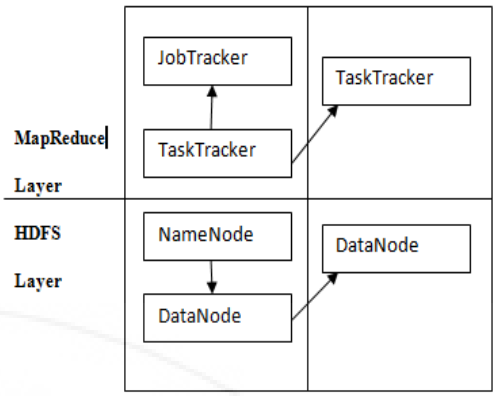

Figure 1: Hadoop technology

The facility is MapReduce: It is a parallel programming model that is used to retrieve the data from the Hadoop cluster to analyze it by performing processing on it. This isolates the data to different set of task to reduce overhead and executes on the various nodes parallel, thus speeding up the computation and retrieving required data from a huge dataset in a fast manner. They have to just implement two functions: map and reduce. The data are fed into the map function as key value pairs to produce intermediate key/value pairs. Once the mapping is done, all the intermediate results from various nodes are reduced to create the final output. JobTracker keeps track of all the MapReduce jobs that are running on various nodes. This schedules the jobs, keeps track of the entire map and reduces jobs running across the nodes and allows TaskTracker to performs the map and reduce tasks that are assigned [12].

6. Discussion

User behavior analysis using review processing is a need of various Organization or Enterprises so as to understand and fulfill the user's requirement about service providing system. Hence, extraction of reviews posted by number of users using web mining and retrieval of characteristics of service system and user behavior towards that system is a challenging task. There are various methodologies discussed in literature review for mining and processing the features or characteristics of particular product and user's behavior in terms of opinion or reviews, applying probabilistic and predictive algorithms, AHP model or HITS method on it based on parameters like computing TF/IDF frequency of terms in document, scoring terms, *helpfulness* term value etc. as per the analysis is determined. The system defined by some authors is limited for user perspective so as to satisfy their need like hotel reservation system based keyword based algorithm [1]. The proposed system will perform the same task for several other service providing systems like movie, mobile purchasing or etc. domain by defining different kinds of domain thesaurus and can distinguish positive & negative preferences of user for the same. This system have to define some sort of hypothesis for performing the task of analysis of user behavior by providing interface to user for easy searching of required data and database of domain thesaurus which will work efficiently to give efficient result. This system will use the Map Reduce technology for efficient result of user behavior analysis.

7. Conclusion

This literature review is useful to understand various methodologies used for user behavior analysis. The opinion mining or semantic behavior is done with the help of text mining method for retrieval if keywords or terms from reviews posted by user. Then with help of prediction model and probability algorithm the processing of reviews is done. The methodologies used to determine efficiency of review processing is done on fully supervised or semi supervised data collected and preferring domain thesaurus generation which static or dynamic. The proposed system is built for fully structured data processing and based on different domain thesaurus which is efficient with Map Reduce technology.

8. Future Scope

The future scope is we can try for user analysis based on different type of domain thesaurus merging with cross domain strategy and try to reduce complexity in the cross domain semi structured data with Hadoop technology for efficient result.

References

[1] Shunmei Meng, Wanchun Dou, Xuyun Zhang, Jinjun Chen, Senior Member, IEEE, "KASR: A Keyword-Aware Service Recommendation Method on MapReduce for Big Data Applications", IEEE Transaction on Parallel and Distributed Systems, TPDS-2013-12-1141.
[2] Lisette García-Moya, Henry Anaya-Sánchez, and Rafael Berlanga-Llavori Universitat Jaume I , "Retrieving Product Features and Opinions from Customer Reviews" , Published by the IEEE Computer Society ,2013.
[3] Zheng-Jun Zha ,Member, IEEE, Jianxing Yu, Jinhui Tang,Member, IEEE, Meng Wang, Member, IEEE, and Tat-Seng Chua , "Product Aspect Ranking and Its Applications" , IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 5, MAY 2014.
[4] Danushka Bollegala ,Member, IEEE, David Weir, and John Carroll , "Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus" , IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 8, AUGUST 2013.
[5] Anindya Ghose and Panagiotis G. Ipeirotis, Member, IEEE, "Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics", IEEE Transactions On Knowledge and Data Engineering, Vol. 23, no. 10, October 2011.
[6] Brandon A. Beemer and Dawn G. Gregg, "Dynamic Interaction in Decision Support: Effects on Perceived Diagnosticity and Confidence in Unstructured Domains", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, VOL .43, NO.1, JANUARY 2013.
[7] Kleanthi Lakiotaki and Nikolaos F. Matsatsinis, Technical University of Crete Alexis Tsoukiàs, Université Paris Dauphine, "Multicriteria User Modeling in Recommender Systems" , Published by the IEEE Computer Society, 2011.
[8] Giuseppe Di Fabbrizio, Ahmet Aker, and Robert Gaizauskas, University of Sheffield, "Summarizing Online Reviews Using Aspect Rating Distributions and Language Modeling" , Published by the IEEE Computer Society, 2013.
[9] Xiaohui Yu, Member, IEEE, Yang Liu, Member, IEEE, Jimmy Xiangji Huang, Member, IEEE, and Aijun An, Member, IEEE, "Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 4, APRIL 2012.
[10] Pravin Jambhulkar1, Smita Nirkhi 2, "A Survey Paper on Cross-Domain Sentiment Analysis", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 1, January 2014.
[11] N.Manjunathan, "Cross-Domain Opinion Mining Using a Thesaurus in Social Media Content", International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 2, Issue 5, May 2014.
[12] Hemlata Kardas, Dr. Akhil Khare,Analysis of log records by VDB and Log processor from web server, 2014 : presented to IEEE Xplore Conference 2014.
[13] http://en.wikipedia.org/wiki/Apache_Hadoop.
[14] http://en.wikipedia.org/wiki/Big_Data.

Author Profile

Miss. Hemlata Kardas is pursuing Masters of Engineering in Computer Science at Rajarshi Shahu College of Engineering and Research, JSPM NTC Pune-India. She obtained Bachelor of Engineering in Computer Science in 2011 at BMIT, Solapur-India. Her area of interest is Web Mining, Information retrieval, Cloud Computing and Hadoop technology. She has presented the conference paper for IEEE Explore on Cloud Computing and Hadoop in 2014 and also published IJMER Paper on Web mining and Information Retrieval.

Mrs. Sulochana Sonkamble is HOD of Computer Science Department at Rajarshi Shahu College of Engineering and Research, JSPM NTC Pune-India. She obtained Bachelor of Engineering, Masters of Engg. And PhD in Computer Science Shri Guru Gobind Singhji Institute of Engineering and Technology, Swami Ramand Tirth Marathwada University, Vishnupuri, Nanded-India.