ISSN: 0976-3104
SPECIAL ISSUE: Emerging Technologies in Networking and Security (ETNS)
Ajitha et al.

ARTICLE OPEN ACCESS

DESIGNING AN INTEREST SEARCH MODEL USING THE KEYWORD FROM THE CLUSTERED DATASETS

E. Ajitha, A. Nirmal Kumar*, A. Jayanthi, D. Daya Florance

Assistant Professor,Dept. of Information Technology,VeltechHightech Dr. Rangarajan Dr. Sakunthala Engineering College,Avadi, Chennai. INDIA

ABSTRACT

A social blogging service is one of the most abundant resources for data collection about one's personal interest on various things. Collection of these data will result in the explosion of short text messages. The analyzed and collected data contains enormous amount of noise and redundancy. The small scale data set along with a database is designed to collect and handle the text streams. These datasets are clustered based on time (Day wise). Keyword filtering is used to remove the noisy and outlier datasets. A user purchase model is designed which holds two search criteria Such as General Purchase and Profile Based Purchase. The interest is monitored in a continuous manner in social blog through a Hadoop Server.

Published on: 08th– August-2016

KEY WORDS

Clusters, Keyword filtering,Outlier, Recommendation,Hadoop

***Corresponding author: Email:** sa.nirmalkumar@gmail.com; **Tel.:** + 91 9443998771

INTRODUCTION

Social blogging has been so popular since its evolution in the world of internet and the amount of short text messages in the form of posts, likes, and chats range up to several millions per day. Inquiring social blog for a specific person's interest will be helpful in saving time over searching for his/her favorites over internet. Existing systems does not hold this property of displaying one's interest. Recommender systems, so far designed, deals only with the past browsed history of the user. Our proposed model discusses to deliver the user interest by monitoring one's personal interests over a social blog.

Hadoop server is used for continuous monitoring. Hadoop splits files into large blocks and saves them across cluster's nodes. To process the stored data, it transfers confined code to the nodes for parallel processing of data. This approach takes advantage of locality of the data nodes manipulatingthe data they have accessedto allow the dataset to be processed faster and more efficiently.
The increasing popularity has resulted in overwhelming of data even though they are informative. Retrieving those data is a tedious process even though filtering is used. Summarization along with clustering can be introduced to overcome these difficulties in retrieving a particular data.

RELATED WORKS

A clustering technique is used for handling real time data sets [1].The simple one-pass algorithms are too inefficient for the incoming data streams. From the application point of view, these algorithms are not enough to process the clusters. Since, they are large in size and these algorithms do not address the size. It is suggested to have micro cluster which refers to the data locality that points to the location of data inside a cluster.

A semi supervised co-clustering with side information [2] is used to process these clusters. This technique

www.iioab.org
THE IIOAB JOURNAL
www.iioab.webs.com
COMPUTER SCIENCE

THE IIOAB JOURNAL ISSN: 0976-3104

www.iioab.org

THE IIOAB JOURNAL

www.iioab.webs.com

describes to carry additional information about the main text such as author name, publication details and so on. Due to which the overhead in searching the cluster and its related co-clusters is increased. Side information is only necessary when the main context is damaged or corrupted due to error in the cluster. In those situations it is better to avoid collecting the side information since data corruption occurs very less in a cluster. Rootdatasets can only be concentrated for specific information to be mined.

BIRCH (Balanced Iteration Reducing and Clustering Using Hierarchies) is an improved way to cluster the datasets of related entities and related datasets of similar entities [3]. Balanced iteration additively and variably clusters the related datasets from various data localities and produces the best cluster based on the recent time frame. Best clusters are produced considering the available memory space. It will be more suitable for large clusters to hold on a certain threshold value in retrieving for particular dataset which can possibly eliminate outlier datasets.

Bursty feature representation along with the text mining technique is used on a classical static datasets. Bursty features are nothing but a part of text which is commonly repeated over and over in several other posts, the text may refer to a particular product or an event that are generated on a relatively short period of time[4]. It identifies the feature for clustering the datasets which in turn causes noisy datasets to be retrieved along with the queried information. This also results in exponential increase in cluster size.

Periodical time frame (say three sessions / event) and calculating the frequency of bursty texts over those time frames or session may improvise the system. The cluster with the highly populated bursty texts is selected for mining the useful information.

A different approach to overcome the difficulties faced by [5] in terms of previous time period cluster which combines online grouping algorithm with the existing scalable clustering techniques to produce fast and adaptive clusters of text streams[6]. Yet it fails to retain too old clusters due to memory constraints. It is suggested that sufficient weightage to the old clusters can be allotted and based upon which they can be dropped or kept for usage. The more historical data with less weightage can be neglected if more weightage clusters of the same period have evolved.

A framework for clustering hefty text and definite text streams [7]. So far [4] [5] [6] has only met with problem of time and space. But a real time handling and clustering along with segmentation for organizing documents in applications has been done. It uses statistical summarization methodology to cluster the data to individual text streams or as a categorical data streams that represents those respective topics. The drawback of this proposal is that the temporal locality of the clusters is not taken into account. Moreover repeated querying of different kinds cannot be answered as quickly as the incoming data streams.

It can be easily overcomed by having a continuous monitoring of social profiles so as to get the instant updates of each activity. Hadoop server can be deployed for such monitoring of data streams. Hadoop uses a special MapReduce function to count on each repeated activity of an individual user. It also implements a programming model for processing the huge amount of incoming text streams.

To identify the keyword in a sentence (post or status) a new conceptwhich constructs a lexical chain for text summarization [8] is introduced. Summarization by this technique causes only strong chains to be identified. This paper does not addresses the sentence granularity which extracts the central constituents from the text. It does not implement necessary keyword filtering techniques to control the length of the posts.

DISADVANTAGES OF EXISTING SYSTEM

1. More complex datasets cannot be handled.

2. Waiting time is increased.

3. Less accuracy.

THE IIOAB JOURNAL
ISSN: 0976-3104

www.iioab.org

THE IIOAB JOURNAL

4. Low data transmission rate.

5. Replication or requests due to inefficient offline grouping algorithm.

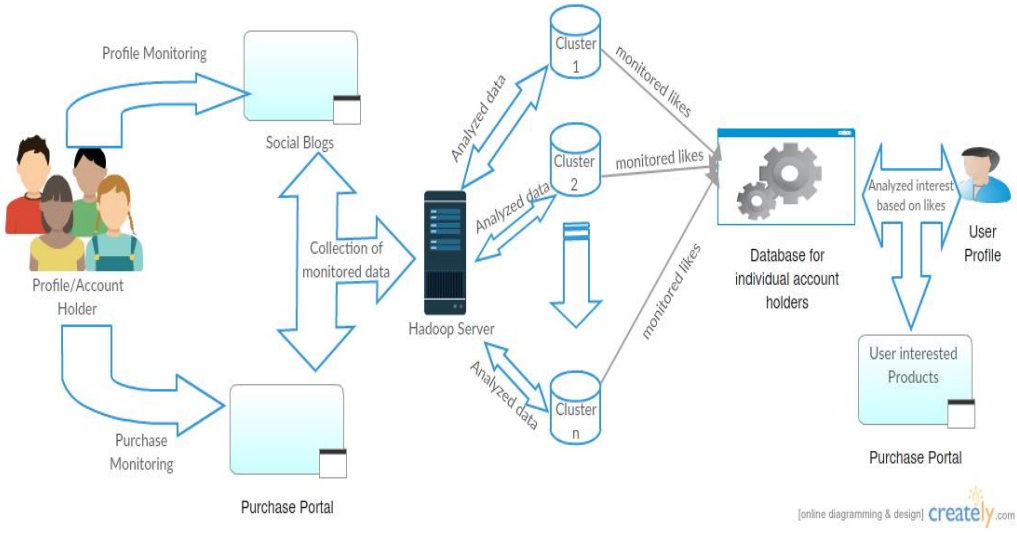
6. Datasets containing noise are also retrieved when mining important posts.

PROPOSED WORK

Designing a purchase model on user interest with the filtered keyword from the user posts and the frequency of posts on a particular product or a particular topic is implemented.Purchase Portal will have two options like General Purchase & Profile based Purchase. In General Purchase, usual items of a shopping site is recommended for purchase. In Profile based Purchase, Items are displayed based on the User's Interest. Related Items and Items which are purchased more often are also displayed to the user based on the User Interest.

Continuous monitoring of user profile maintained in social blog is carried. Following the user posts and likes, a Hadoop Server collects the data from user's profile and records it in a database. This database has bi-directional connectivity with both the user account on the blog and the purchase model designed for shopping. Updation of posts which user makes in the blogging site with minimal time delay is implied for the ease of shopping user's favorite product.

Summarization of users posts are produced by the means of continuous monitoring of user profile over days spanning to several weeks. This helps in predicting the user interest very closely with less error rate and achieving high accuracy as the time extends. Recommendation based on history is also available for inactive blog users.



ADVANTAGES OF PROPOSED SYSTEM

1. Reducing sampling rate consumes less time over computation of online clustering algorithm.

2. Accuracy is improved by mining sub clusters for additional filter.

3. Reliability over time frames.

4. Replica of requests is avoided.

5. Supports a range of data analysis tasks such as reports or historical survey.

www.iioab.webs.com

THE IIOAB JOURNAL
ISSN: 0976-3104

www.iioab.org

THE IIOAB JOURNAL

www.iioab.webs.com

OPERATIONS PERFORMED

1. DATA INSERTION &UPDATE.

2. CONSTRUCTION OF UP-TREE.

3. PURCHASE THE ITEM.

4. DISCARD UNPROMISING ITEMS.

5. FINDING THE PATTERN FROM UP TREE.

DATA INSERTION &UPDATE

In administrator side, one can add particular data and also update the data. Here, data refers to products with is associated values. Each item contains the profit and quantity. We can add and update the profit and quantity values for product list. All inserted values are added in the database. Also updated values are added in the database. We can also view the added and updated values. These values along with the items are will be displayed in the product list.

CONSTRUCTION OF UP-TREE

A compact tree structure is usedfor discovering immense utility item sets and maintaining the information aboutpatterns within databases. Immense utility item sets can be generated from UP-Tree efficiently with only two scans of original databases.

In the first scan, transaction unit of each transaction is computed. At the same time, transaction utility weight of each single item is also accumulated. An item and its supersets are unpromising to be immense utility if its weightis less than the defined threshold. An item is called a promising item if weight is greater than the minimum utility threshold. Otherwise it is called an unpromising item. Generally, an item is also called a promising item if its overestimated utility is no less than minimum utility. Otherwise it is called an unpromising item.

New transaction unit after discarding unpromising items is called reorganized transaction. By reorganizing the transactions not only less information is needed to be recorded in UP-Tree. Since the utilities of unpromising items are excluded, reorganized transaction must be no larger than utility weight. In the second scan, reorganized transactions are inserted into the UP Tree. Hence, the high potential utility item sets can be efficiently generated from the UP-Tree.

ELEMENTS IN UP-TREE

In a UP-Tree, each node N consists of name, count, nu, parent, hyperlink and a set of child nodes where name is the node's item name; count is the node's support count; nu is the node's node utility, i.e., falsehood utility of the node; parent records the parent node of N; hyperlink is a node link which points to a node whose item name is the same as node's name.

A table named header table is employed to enable the traversal of UP-Tree. In header table, each entry records an item name, afalsehood utility, and a hyperlink. The hyperlink points to the last referral of the node which has the same item as the entry in the tree. By following the hyplinks in header table, the nodes having the same name can be traversed efficiently.

THE IIOAB JOURNAL
ISSN: 0976-3104

www.iioab.org

THE IIOAB JOURNAL

www.iioab.webs.com

PURCHASE THE ITEM

In client side user can enter all details. Then user can login using particular username and password. All the inserted also updated items are added into the product list. Then select user wanted items then add all items into cart products with count of the each item. A warning message will display in dialogue box when the customer type the quantity above the constraint value mentioned in the database. All selected items are displayed in the cart product list. Then purchase the required items.

DISCARD UNPROMISING ITEMS

In Frequent Pattern growth mining, the minimum utility is entered. Based on the value of minimum utility we can find out the promising items and unpromising items. Transactional-weighted utility of an item set is the sum of the transaction utilities of all the transactions. If it's utility is less than a user specified minimum utility threshold. An items set is called a low utility item set. To select specific rules from the set of all rules, constraints on significance and interest can be used. The best known constraints are fixing the thresholdbased on support & confidence provided by trusted source. Then the low utility item sets are discarded at end of the transaction.

In Data Mining the task of finding the most hit pattern in large databases is very importantas it computationally more expensive, when a more number of patterns exist. These patterns which are mined during the various approaches make the user very difficult to identify the patterns which are very interesting for them. The goal of most hit itemset mining is to identify all frequent itemsets. Once the frequent itemsets are identified, association rules are generated for the identified itemsets.

In the real world, however, each item in the purchase portal has a different importance/price and single customer will be interested in buying a number of same products. Therefore, finding only classic frequent patterns in a database cannot fulfill the need of searching thevaluable itemsets that contribute the most to the profit in a retail business.

FINDING THE PATTERN FROM UP-TREE

Searching process for immense utility item set mining is difficult because a product or a item of a low utility item set may be a immense utility item set. If transactional-weighted utility is no less than a user specified threshold value. An item set is called a high utility item set. Based on the TWU we can find out promising items and unpromising items. Based on the threshold value we discard the unpromising items. Then find out the promising items. Candidate item sets are generated with the previously discussed database scans. Mining high utility item sets from database refers to the discovery of item sets with high utility like profit.

CONCLUSION

The proposed work is a prototype which supports continuous text stream summarization for enhanced blogging site with many facilities. It employs a text stream clustering algorithm to compress posts and texts into TCVs and maintains them online. Then, it uses a summarization algorithm for generating summaries contained in online as well as in history with random time durations. The topic progression can be observed automatically to produce varying timelines for text streams In future work, development of a multi-topic version of clustering algorithm in a distributed system can be introduced and to evaluate it on more complete and large-scale data sets. Monitoring of more than one account can be considered for future implementations

CONFLICT OF INTEREST
Authors declare no conflict of interest.

ACKNOWLEDGEMENT
None.

FINANCIAL DISCLOSURE
No financial support was received to carry out this project.

THE IIOAB JOURNAL
ISSN: 0976-3104

REFERENCES

[1] CC Aggarwal, J Han, J Wang, and PS.Yu. [2003] A framework for clustering evolving data streams, in Proc.29th *Int Conf Very Large Data Bases*, 81–92.

[2] RuchaBhutada and DA Borikar.[2015] An Approach to Semi-Supervised Co-Clustering With Side Information in TextMining, *International Journal of Engineering Trends and Technology (IJETT)* –19 (6)

[3] T Zhang, R Ramakrishnan, and M Livny.[1996] BIRCH: An efficient data clustering method for very large databases, in *Proc ACM SIGMOD Int Conf Manage Data*, 103–114

[4] Q He, K Chang, EP Lim, J Zhang.[2007] Bursty feature representation for clustering text streams, in *Proc SIAM Int. Conf Data Mining*, 491–496.

[5] J Zhang, Z Ghahramani, and Y Yang.[2014] A probabilistic model for online document clustering with application to novelty detection, in *Proc Adv Neural Inf Process Syst*, 1617–1624.

[6] Shi Zhong.. Efficient Streaming Text Clustering" Department of Computer Science and EngineeringFlorida Atlantic University, Boca Raton, FL 33431

[7] CC Aggarwal and PS Yu.[2010] On clustering massive text and categorical data streams, *Knowl Inf Syst*, 24(2): 171–196.

[8] R Barzilay and M Elhadad.[1997] Using lexical chains for text summarization, in *Proc. ACL Workshop Intell. Scalable Text Summarization*, 10–17.

www.iioab.org
THE IIOAB JOURNAL
www.iioab.webs.com