International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 2 | ISSN : 2456-3307

Fast Memory Access in Data Streams Using Pagerank Algorithm

R. Ramya
Computer Science Department, IFET College of Engineering, Villupuram, Tamil Nadu , India

ABSTRACT

The memory efficient incremental local outlier (MiLOF) detection algorithm is used in data streams, and it is more flexible algorithm ,and Incremental LOF is used within a static reminiscence bound its similar to(MiLOF).By using the pageRank algorithm the content is reduced in data stream according to the sentence or word. The proposed algorithm have better memory and time complexity than memory efficient incremental local outlier (MiLOF).In addition, we show that PageRank algorithm is dynamic to changes in the number of data points, the number of essential clusters and the number of dimensions in the data stream.

Keywords : MiLOF, LOF, pageRank, DSMS

I. INTRODUCTION

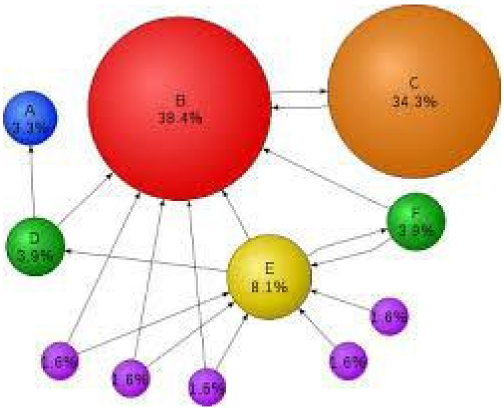
Data mining, in general, deals with the discovery of non-trivial, hidden and interesting knowledge from different types of data. With the development of information technologies, the number of databases, as well as their dimension and complexity grow rapidly. It is necessary what we need automated analysis of great amount of information. The main difference between a traditional database and a data stream management system (DSMS) is that instead of relations, we have unbounded data streams. Applications, such as fraud detection, network flow monitoring, telecommunications ,data management, etc., where the data arrival is continuous and it is either unnecessary or impractical to store all incoming objects.

II. METHODS AND MATERIAL

A. PageRank Algorithm

Page Rank is a numeric value that represents the importance of a page present on the document.

Moreover, implies more importance. Importance of the page that is casting the vote determines the importance of the vote. Several computation circles based on the sentence determined by the PageRank algorithm A page is important if it is pointed to other important pages. Importance of each vote is taken into account when a page's Page Rank is calculated. It matters because it is one of the factors that determines a page's ranking in the search results. In practice, consists of billions of documents and it is not possible to find a solution by inspection. Because of the size of the actual form, the Google search engine uses an approximate, iterative computation of PageRank values.



This means that each page is assigned an initial starting value and the PageRank's of all pages are then calculated.

B. Literature Survey

1. **Incremental Local Outlier Detection For Data Streams**

Author: D. Pokrajac, A. Lazarevic

Outlier detection has lately become an imperative problem in many industrial and financial applications. This problem is further grim by the fact that in many cases, outliers have to be detected from data streams that attain at an mammoth pace.An incremental LOF (local outlier factor) algorithm, appropriate for detecting outliers in data streams, is proposed. The proposed incremental LOF algorithm provides equivalent detection performance as the reiterated static LOF algorithm (applied after insertion of each data record), while requiring significantly less computational time.

In addition, the incremental LOF algorithm also dynamically updates the profiles of data points. This is a very important property, since data profiles may change over time.

2. **Loop : Local Outlier Probabilities**

Author: Hp. kriegal, P. kroger

Many outlier detection methods do not merely provide the decision for a single data object being or not being an outlier but give also an outlier score or "outlier factor" signaling "how much" the respective data object is an outlier.A major problem for any user not very acquainted with the outlier detection method in question is how to interpret this "factor" in order to decide for the numeric score again whether or not the data object indeed is an outlier. Here, we formulate a local density based outlier detection method providing an outlier "score" in the range of [0, 1] that is directly interpretable as a probability of a data object for being an outlier.

3. **Anomaly Detection : A Survey**

Author : V.Chandola A. Banerjee V. Kumar

Anomaly detection is an important problem that has been researched within diverse research areas and application domains. Many anomaly detection techniques have been specifically developed for certain application domains, while others are more generic. This survey tries to provide a structured and comprehensive overview of the research on anomaly detection. We have grouped existing techniques into different categories kbased on the underlying approach adopted by each technique. For each category we have identified key assumptions, which are used by the techniques to differentiate between normal and anomalous behavior.

III. RESULTS AND DISCUSSION

A. Proposed System

- ✓ By proposed a more efficient approach which is an PageRank. This approach is capable of dynamically finding the number of summaries to keep in memory.
- ✓ In addition, it is more stable in terms of changes to the available memory size and more accurate in detecting outliers by using the pageRank algorithm in proposed system.
- ✓ The sentence are ranked by using PageRank algorithm and stored into the memory.And it provides less memory space.

B. Module Description

1. Data Streams

Data Stream Mining is the process of extracting knowledge structures from continuous, rapid data records. A data stream is an ordered sequence of instances that in many applications of data stream mining is read one time or few times by using limited computing and storage capabilities**.**

2. Document Summarization

Every time the number of data in memory reaches the limit the algorithm invokes the summarization phase. This phase includes building a summary over the past data points along with their corresponding values and deleting them from memory**.**

3. Document Merging

Since the summarization is performed every time new data points are received, a new set of cluster centers will be generated after each step. The cluster centers in step are merged with the data cluster from the document so that there is only a single set of cluster centers maintained by the detection framework.

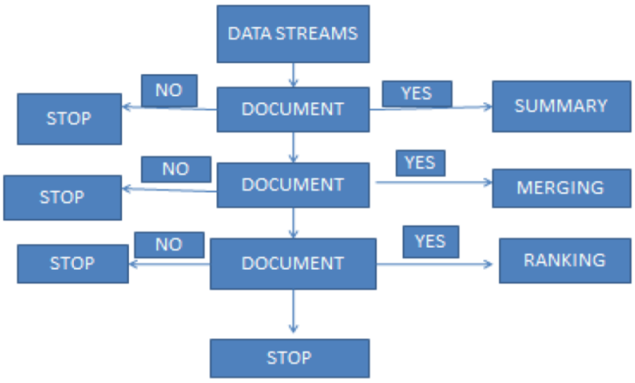
4. Document Ranking

Systems must be designed to aid users in determining which documents of those retrieved are most likely to be relevant to given queries. Therefore document ranking is very important part. Most commercial text retrieval systems employ inverted files to improve retrieval speed. The inverted file specifies a document identification number for each document in which the word occurs**.**

IV. FUTURE WORK

In our approach PageRank, a more effective and flexible algorithm reduces the density of outliers in the previous data points to increase the detection accuracy for future data points, while remaining fast and memory efficient. PageRank is more stable than MiLOF, it is less sensitive to the size of available memory and independent of number of data summaries. In addition we showed that PageRank is more scalable with respect to the dataset size, the number of data dimensions and the number of clusters. In futher this process proceed by the new alorithms for better performance.

Architecture Diagram



V. CONCLUSION

PageRank algorithm is an important requirement in the meadow of outlier detection for data streams. In this paper, the proposed approach to tackle the problem of identifying local outliers. And the pageRank algorithm has extensive practical in improving widely based on reducing content. It requires for saving all previous data points to compute local outlier factors. This method avoids this difficulty by summarizing subsets of the previous data and accumulating an evolutionary history of the streaming information and ranking the datas. In this way PageRank algorithm saves computation time as well as memory. For the data sets used in datastreams, it required substantially less computation time and memory than (MILOF),while achieving comparable accuracy

VI. REFERENCES

[1]. S. Sadik and L. Gruenwald, "Research issues in outlier detection for data streams," ACM SIGKDD Explorations Newsletter, vol. 15, no. 1, pp. 33–40, 2014.

[2]. D. Pokrajac, A. Lazarevic, and L. J. Latecki, "Incremental local outlier detection for data streams," in Computational Intelligence and Data Mining, 2007, pp. 504–515.

[3]. M. Salehi, C. Leckie, J. C. Bezdek, and T. Vaithianathan, "Local outlier detection for data streams in sensor networks: Revisiting the utility problem,"

[4]. S. Papadimitriou, H. Kitagawa, P. B.Gibbons, and C. Faloutsos,"Loci: Fast outlier detection using the local correlation integral,"in International Conference on Data Engineering, 2003, pp. 315–326.

[5]. H.-P. Kriegel, P. Kr¨oger, E. Schubert, and A. Zimek, "LoOP: localoutlier probabilities," in ACM Conference on Information and KnowledgeManagement, 2009, pp. 1649–1652.

[6]. S. Rajasegarar, C. Leckie, and M. Palaniswami, "Anomaly detectionin wireless sensor networks," IEEE Wireless Communications,vol. 15, no. 4, pp. 34–40, 2008.

[7]. V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Computing Surveys, vol. 41, no. 3, pp. 1–58, 2009.

[8]. M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier Detectionfor Temporal Data: A Survey," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 1, pp.1–20, 2013.

[9]. C. C. Aggarwal, "Outlier Analysis," 2013 .

[10]. K. Yamanishi, J.-I. Takeuchi, Williams, and P.Milne, "Online unsupervised outlier detection using finite mixtures with discounting learning algorithms," in SIGKDD, 2000, pp. 320–324.

[11]. K. Yamanishi and J.-i. Takeuchi, "A unifying framework for detecting outliers and change points from non-stationary time series data," in SIGKDD, 2002, pp. 676–681.

[12]. C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," in VLDB, 2003, pp. 81–92.

[13]. F. Cao, M. Ester, W. Qian, and A. Zhou, "Density-based clustering over an evolving data stream with noise," in SIAM Conference on Data Mining, 2006, pp. 328–339.

[14]. S. Guha, A. Meyerson, N. Mishra, R. Motwani, andL. O'Callaghan, "Clustering data streams: Theory and practice,"IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 3,pp. 515–528, 2003.

[15]. C. C. Aggarwal, J. Han, J. Wang, and P. S.Yu, "A framework for projected clustering of high dimensional data streams," in International Conference on Very Large Data Bases-Volume 30, 2004,pp. 852–863.