International Journal of Engineering & Technology, 7 (3.1) (2018) 90-92

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Survey on Duplicate Data Filtering Methods in Big Data

S P Godlin Jasil[1]*, V Ulagamuthalvi[2]

[1]*Research Scholar, Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai.*
[2]*Professor, Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai.*
**Corresponding author E-mail: godlin_jasil@yahoo.co.in*

Abstract

Big Data analytics is the process of collecting heterogeneous huge sets of data for analyzing .The data are fetched from different sources and can be in heterogeneous form. Data arriving in the big data system will be in giga-bytes for every second. Since, the data are in huge volume, there is a possibility of redundant data that affect the network performance. This article presents the review of different filtering methods and algorithms that are used for duplicate elimination such as Bloom filter, Stable Bloom Filter, multi-layer bloom filter, Counting Bloom Filter with some disadvantages such as false positive and false negative. The aim of this paper is to propose an algorithm for eliminating the duplicate Data in a large data set by using big data analytics.

Keywords: *Bloom Filter, Duplicate, Data Filtering, False Positive, Multi-layer Bloom Filter.*

1. Introduction

As we are entering in to a more digital world, the data that are created and collected are in multiple gigabytes. Data are generated from social media, sensor networks; Internet enabled streaming data, mobile data, satellite data, and health data. Since the data are from different sources it can be a structured, unstructured and semi structured. This enormous volume of data is normally so big and it is difficult to process using traditional database [6]. In the recent years, Cloud Computing has emerged as an alternate to the acquisition and management of resources [10].

The data are in huge volume, velocity, variety and veracity. Volume represents data generated are large set of data which leads to big data. The second characteristic is velocity that represents the speed of data generated by the source. Big data collect data from different sources which produced data in different format called variety. Big data system should be able to process this different variety of data streams. Veracity represents the reliable and truthful sources so that the utility of the big data will be increased.

Big Data analytics help to organize the data which is useful to understand the information contained within the data. Big Data analytics can be performed by some software tools. Using highly developed analytics such as text analytics, machine learning, predictive analytics, data mining, statistics and natural language processing, business can analyze previously unused data sources with their existing enterprise data to gain new insights resulting in better and faster decisions [7].
Due to the huge volume of data there is a possibility of duplicate data. By removing this redundant data the efficiency of the network bandwidth can be widely improved.

2. Related Work

2.1. Bloom Filter

Bloom [1] proposed a model called bloom filter. A Bloom filter is an array of m size bit. Initially zero is set to all the elements in the array. When an element arrives, K bits in Bloom filter are set to 1 by a group of hash function $\{h_1(x).....h_k(x)\}$, assume all are independent. There is a possibility of setting the same one bit for multiple times, while the first changing operation from 0 to 1 can be done and the rest of the operation has no effect on that bit. To know they arrived new bit x_i is distinct, we have to check the bit with the hashing function $\{h_1(x).....h_k(x)\}$ and if any one bit is 0 then the new element is distinct and if all the bit in the array is set to 1 that element is considered as duplicate with some probability of error is shown in figure 1.

The bloom filter is useful if there is a proper memory space allocated for the new elements. When there is a large data stream, the allocated space can be small comparing with the stream of data, this leads to the decrease in number of zeros in the bloom filter and there is a chance of getting the false positive rate high. When the entire distinct element is considered as duplicate, Bloom filter becomes ineffective.

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

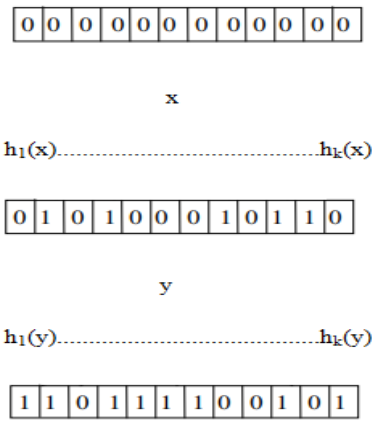


Fig. 1: Illustration of Bloom Filter

One solution to this problem is, a predefined threshold can be set and before the error rate reaches some elements can be dispossessed. This can be same as buffering process. There are several strategy for selecting a past element to set out. In Bloom filter the recent element and the past element cannot be distinguish because no time information is maintained [2,3,4,11].

2.2. Stable Bloom Filter

Stable Bloom Filter (SBF) is an enlargement of the regular bloom filter. Stable bloom filter is an array from s[1].....s[n] where the minimum value is initialize to 0 and the maximum value is max .Each element in the array is allocated d bits called cell in SBF. The relation between max and d is given by $max=2^d-1$. Each cell contains one or more bit.

Duplicate detection process: The entire cell is initializing to 0.By using some hashing function the recent element is mapped to K cell. Check whether the entire cell is non zero or not.

In SBF, to make the cells empty, generate a random number, decrement the corresponding P cells and p-1 cell by 1. This process is faster when comparing with generating random number for each element. Set the same K cells for detection process. LRU replacement policy is used for replacement of older element in the case of buffering [4].

Since the cells are decremented randomly there is a possibility of false positive rate.

2.3. Multi-layer Bloom Filter for Duplicated URL Detection

A web crawler download the web page by using its URL, it starts downloading the web site from the home page and take out all links in the page and gets web pages in the new layer by the new links. This process is repeated until a particular layer or when there is no new link. Consider there is a massive number of web pages and the links structures are complicated, it contains large number of duplicated links. If the web pages are collected randomly then it will leads to waste of large amount of network bandwidth .when web pages are collected, to improve the speed the multiple time downloading a page can be avoided. Web Crawler should register the page when it is collected and find out whether the next page is collected or not. In the process of web page collecting, the URL address of the page will be checked in the URL collection, if it present, this URL will be skipped, otherwise, this URL will be added to the waiting queue of collecting [5].

When a web crawler is working, duplicated URL detection is performed often. The algorithm used for duplicate URL detection should have less space and time complexity because duplicated URL detection should be done all pages. A multi-layer bloom filter algorithm divides a complete URL into some layers and stores them in multi-layer bloom filter. This multi-layer bloom filter lowers the false positive rate but the efficiency of the previous is almost the same as the new filter.

MLBF needs added memory space than classic bloom filter and false positive times of MLBF is much lesser than classic bloom filter, and the time taken to insert and query URL are same. MLBF runs only on single computer and not for a distributed environment.

2.4. Duplicate Elimination in RFID Data Streams

RFID technology produce massive amount of RFID data. To manage RFID data streams in a small memory space we use Bloom filter which is useful for static data. This paper proposed a Time Bloom Filters as an extension of Bloom filter in which the false negatives i.e. duplicate data contained in the result after filtering was avoided [8].

2.5. Counting Bloom Filter

In bloom filter, inserting an element is easy because it done by hashing the element by the hashing function and set the bit to 1.On the other hand, deleting an element is by reversing this procedure, i.e hashing the element and setting the bit to 0 is not possible because, there is a chance of making the location 0 the is hashed by some other element in the set. As a result bloom filter becomes is not correct for all set of elements.

Because of this disadvantage, counting Bloom Filter was introduced. As it is shown in figure 2 each entry is a small counter and not a bit. When an element is inserted, the appropriate location counter is incremented and when deletion happened the location counter is decremented [11]. Larger counters can be used to avoid counter overflow. From the analysis, 4 bit per counter can be sufficient for most applications [9].

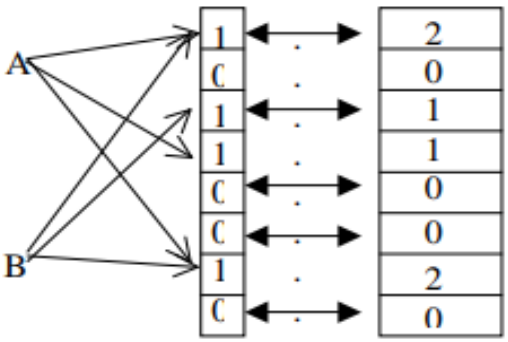


Fig. 2: Incrementing/decrementing Counters

3. Conclusion

In this paper, we discussed about the different existing methods, algorithms and the growth of data from different sources. It also brief about the redundant data that are stored in large dataset that effect the network efficiency. The presented literature review reveals that the existing algorithm such as Bloom filter, Stable Bloom filter and multilayer Bloom filter Counting Bloom Filter that are not able to handle the false positive and false negative rate. With the present survey an new algorithm can be proposed that solve the present problem.

References

[1] B. H. Bloom. Space/time trade-of in hash coding with allowable errors. In CACM, 1970.
[2] E. Cohen and M. Strauss. Maintaining time-decaying stream aggregates. In Proc. of PODS, 2003.

[3] T. Palpanas, M. Vlachos, E. J. Keogh, D. Gunopulos, and W. Truppel. Online amnesic approximation of streaming time series. In Proc. of ICDE, 2004.
[4] Fan Deng , Davood Rafiei . Approximately Detecting Duplicates for Streaming Data using Stable Bloom Filters In Proceedings of the 2006 ACM SIGMOD international conference on Management of data Pages 25-36 cross Ref Google Scholar
[5] Cen Zhiwang , Xu Jungang , Sun Jian (2010) A multi-layer bloom filter for duplicated URL detection In Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference .
[6] Muhammad Habib ur Rehma, Chee Sun Liew, Assad Abbas, Prem Prakash Jayaraman, Teh Ying Wah, Samee U. Khan(2016) Big Data Reduction Methods: A Survey cross Ref Google Scholar.
[7] https://www.ibm.com
[8] Chun-Hee Lee a, Chin-Wan Chung, An approximate duplicate elimination in RFID data streams. In Data & Knowledge Engineering 2011, pages 1070-1087.
[9] Fan, L., P. Cao, J. Almeida, & Broder, A. Z. (2000). Summary cache: a scalable wide-area web cache sharing protocol. IEEE/ACM Transactions on Networking, 8(3), pp. 281-293.
[10] S.P. Godlin Jasil , M. Deepa(2014). Efficient Utilization of Cloud Infrastructure by Handling Heterogeneous Workloads. International Journal of Applied Engineering Research. Volume 9, Number 23 (2014) pp. 13655-13666.
[11] Czerwinski, S., Zhao, B.Y., Hodes, T., Joseph, A.D., & Katz, R. (1999). An architecture for a secure service discovery service. Proceedings of MobiCom99, pp. 24-35.
[12] A Salman. Bloom's Filters: Their Types And Analysis. Dogu Universities Dergisi, 6 (2) 2005, 268-278.
[13] G. Ramprabu, S. Nagarajan, "Design and Analysis of Novel Modified Cross Layer Controller for WMSN", Indian Journal of Science and Technology, Vol 8(5), March 2015, pp.438-444.