International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 4, Issue 6, June 2015 **ISSN 2319 - 4847**

Analysis of Distributed Incomplete Pattern Matching

Miss Jayati S. Pande[1], Dr. J. W. Bakal[2]

[1]Student, M.E. Computer Engineering, Shree L. R. Tiwari College of Engineering,
Mumbai University, Thane, Maharashtra, 401107, India

[2] Principal, Shivajirao S. Jondhale College of Engineering,
Mumbai University, Thane, Maharashtra,421204, India

ABSTRACT

In this paper we present a novel Unsupervised Weighted Bloom Filter (UWBF) method, to perform incomplete pattern matching in distributed environment. Various pattern matching Techniques like Naive method; Bloom Filter and Weighted Bloom Filter perform pattern matching. In existing methods, global data from all the nodes is collected to the data centre, which conducts their aggregation and then applies the pattern matching. But this approach results into high communication-cost and traffic in network. In UWBF Data-Centre distributes data to various Base-Stations and then perform pattern matching. The problem of traffic in network is automatically solved. UWBF perform pattern matching by taking untrained patterns, so accuracy increases and it requires less time than previous methods. Compare and contrast of above techniques are discussed.

Keywords: Incomplete Pattern Matching, Unsupervised Weighted Bloom Filter, Security, Data-Centre

1. INTRODUCTION

In mobile environment the data is distributed at various Base-Stations and may be incomplete compare to global pattern. In this paper we discuss and compare various methods like Naïve, Bloom Filters (BF), Weighted Bloom Filters (WBF) and Unsupervised Weighted Bloom Filters(UWBF) to perform pattern matching in mobile environment. Existing system performs pattern matching by collecting all data to a data center from various base stations, and then performs pattern matching. By applying this type of solution, problem of communication channel bottleneck and huge cost may occur. An efficient solution is necessary to find matched patterns in mobile environment. Unsupervised Weighted Bloom Filter (UWBF), solves this problem by performing pattern matching individually at each Base-Station and then sending these matched patterns from various Base-Stations to Data-Centre. Aggregation and similarity ranking is then performed by Data-Centre. The UWBF ensures pattern matching and also save communication cost. Pattern matching is performed on un-trained patterns to find matched patterns. Blow-fish algorithm is used here, to make the network secure. Encryption and decryption algorithms are applied on patterns. Data centre and base station use these encryption decryption algorithms, when they exchange data among them. The proposed solution increases accuracy and it requires less time to perform pattern matching than existing solutions.

2. UNSUPERVISED WEIGHTED BLOOM FILTERS

There are various pattern matching algorithms like Naïve, Bloom Filters, and Weighted Bloom filters (WBF) etc. In this paper, we proposed Unsupervised Weighted Bloom Filter (UWBF) which is extension of WBF. UWBF also perform incomplete pattern matching like WBF. The main difference between WBF and UWBF is that, UWBF deals with untrained patterns also, where the previous methods can perform pattern matching on only trained patterns. UWBF also increases accuracy as compared to existing methods, as it deals with untrained patterns. The time required to perform pattern matching in UWBF is less as compared to WBF. In next section we study accuracy and time-cost graphs.

The flow chart of pattern matching process is shown in following fig 1. User will first input the query (pattern) to get the similar patterns. Then the pattern representation and encoding is performed. The weight of these encoded patterns is find out here, then only the ID of patterns along with their weights are send by Data-centre to various Base-Stations in encrypted format. The Base-Stations will decrypt the patterns and UWBF is applied here. The Base-Stations will send similar patterns in encrypted format to Data-Centre. The Data-Centre will decrypt the similar patterns, then it performs similarity ranking and aggregation and user will get matched patterns.

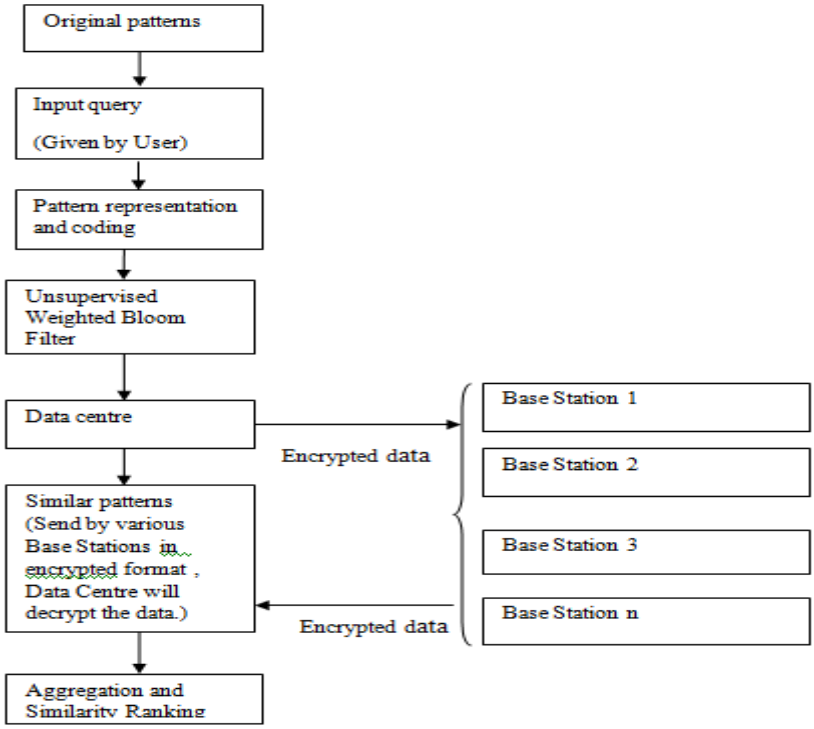

Figure1 Flow-Chart of Pattern matching process using UWBF

3. DATASET RESULTS AND ANALYSIS

The dataset is collected from the gedis-studio dataset repository. The dataset is of CDR mobile network .The data set consists of MSIDN, IMSI, IMEI, PLAN, Call_type, Corres_type, Corres_ISDN, Duration and Time-Date attributes. User gives input in three categories i.e. PLAN, Call_type and Duration to Data-Centre. As shown in above flow chart all steps are performed sequentially. In first phase, pattern representation and encoding is done on the local pattern. For eg. we take PLAN value as'2' i.e PLAN 2, Call_type as 'MTC' and duration=160 seconds therefore local pattern will be{2,2,160} then applying eq (1) on this local pattern we get {3,3,161}.After that data centre finds weight of this encoded pattern i.e. 160/161 in our example. Base stations match patterns on the basis of weights and IDs, it get from data centre. It send matched patterns to data centre, Data-Centre aggregates all similar patterns, then it apply similarity ranking on patterns and finally top k similar patterns are find out. UWBF gives more accurate and hence increases performance of system. Data privacy is also maintained in network by using Blowfish algorithm.

3.1 Graphical Comparison

In this section we will study and graphically compare the different factors like accuracy and time-cost of WBF and UWBF pattern matching methods. In Naive method the global data from all the nodes is collected to the data center, which conducts their aggregation and then applies the pattern matching. Here we take number of patterns at X-axis and precision, time-cost at Y-axis in their respective graphs.

3.1.1 Accuracy

Accuracy is defined by the precision, that is True positive/(True positive+False positive). It is the fraction matched patterns to that are relevant to the search. Following graphs shows accuracy of WBF and UWBF methods. In Weighted Bloom filter (WBF), accuracy decreases as the number of patterns increases, while UWBF can achieve much higher precision due to distinguishing the different patterns by their weight. Hence accuracy increases. Accuracy can be calculated by calculating its precision.

Precision can be calculated by following formula:-

Precision = No. of Match Patterns / Total No. of Records

Precision = 45 / 708

On the basis of above formula precision of WBF and UWBF is calculated and it is shown as graphically in following fig 3.

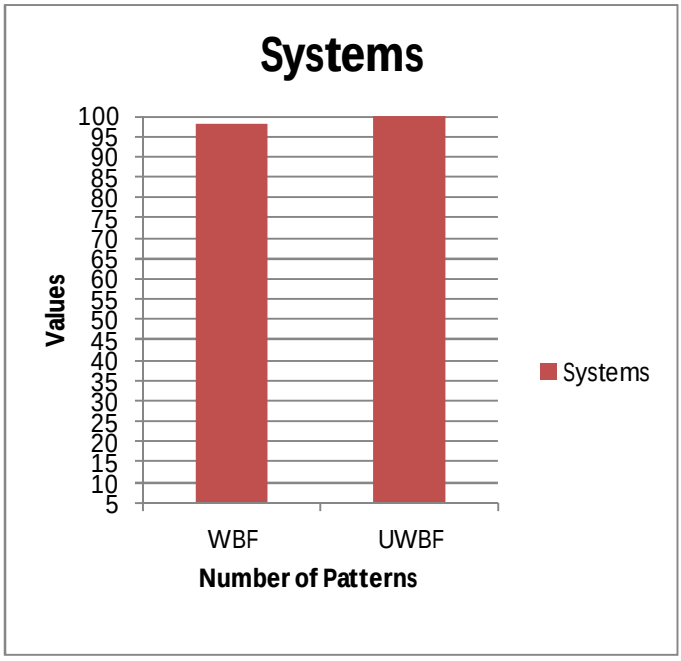

Figure2 Accuracy Graph

3.1.2 Time-Cost

Time cost is defined as time required for finding the target patterns. As the number of patterns increases, WBF required additional time to perform pattern matching. But this is not in case of UWBF. As the number of patterns increases UWBF require no additional time to execute pattern matching.

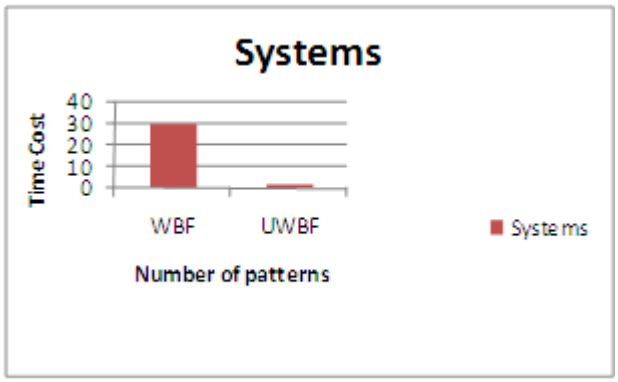

Figure 3 Time-Cost Graph

Success calculated on the basis of accuracy and time of UWBF is shown in following table.

TABLE I DIPM success rate compared to WBF and UWBF

Methods/Features	**Accuracy**	**Time-Cost**
WBF	95%	More
UWBF	100%	Constant

References

[1] Siyuan Liu, Member, IEEE , Lei Kang, Lei Chen," How to Conduct Distributed Incomplete Pattern Matching" Proc. IEEE Trans on Parallel and Distributed Systems, vol. 25, no. 4, Apr. 2014.

[2] H. Chen, H. Jin, L. Chen, Y. Liu, and L. Ni, "Optimizing Bloom Filter Settings in Peer-to-Peer Multi-keyword Searching," IEEE Trans. Knowledge and Data Eng., vol. 24, no. 4, pp. 692-706, Apr.2012.

[3] R. Ahmed and R. Boutaba, "Distributed Pattern Matching: A Key to Flexible and Efficient P2P search," IEEE J. Selected Areas in Communications, vol. 25, no. 1, pp. 73-83, Jan. 2007.

[4] C.C. Aggarwal, Y. Li, J. Wang, and J. Wang, "Frequent Pattern Mining with Uncertain Data," Proc. 15th ACM SIGKDD In'l Conf.Knowledge Discovery and Data Mining, 2009.
[5] J. Agrawal, Y. Diao, D. Gyllstrom, and N. Immerman, "Efficient Pattern Matching over Event Streams," Proc. ACM SIGMOD Int'l Conf.Management of Data, 2008.
[6] B. Babcock and C. Olston, "Distributed Top-k Monitoring," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2003.
[7] D. Guo, J. Wu, H. Chen, and X. Luo, "Applications of Bloom Filters in Peer-to-peer Systems: Issues and Questions," Proc. Int'l Conf. Networking, Architecture, and Storage (NAS '08), 2006.
[8] A. Bagnall, C.A. Ratanamahatana, E. Keogh, S. Lonardi, and G. Janacek, "A Bit Level Representation for Time Series Data Mining with Shape Based Similarity," Data Mining and Knowledge Discovery, vol. 13, no. 1, pp. 11-40, 2006.
[9] B.H. Bloom, "Space/Time Trade-Offs in Hash Coding with Allowable Errors," Comm. ACM, vol. 13, no. 7, pp. 422-426, 1970.
[10] F. Cuenca-acuna, C. Peery, R. Martin, and T. Nguyen, "PlanetP: Using Gossiping to Build Content Addressable Peer-to-Peer Information Sharing Communities," Proc. 12th IEEE Int'l Symp. High Performance Distributed Computing, 2003.
[11] Siyuan Liu , Lei Kang, Lei Chen, Lionel Ni," Distributed Incomplete Pattern Matching via a NovelWeighted Bloom Filter", IEEE 32 nd International Conference , Page(s): 122 – 131,2012.
[12] D. Guo, J. Wu, H. Chen, and X. Luo, "Theory and Network Applications of Dynamic Bloom Filters," Proc. IEEE INFOCOM, 2006.

AUTHOR

Jayati Pande received the B.E. degrees in Computer Engineering from S. S. V. P. S. College of Engineering, Dhule, North Maharashtra University in 2011. Now persuing M.E. from Shree L. R. Tiwari College of Engineering,Mumbai University, Thane, Maharashtra, 401107, India