International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Volume 7, Issue 03 (March 2020) www.ijirae.com
Special Issue – 6th International National Level Conference- "MEEMIC-2020"

NON MOTIONLESS MODEL FOR CRIME RATE DEDUCTION USING CURRENT URBAN DATA

M.Karthikeyan[1],.Amsaveni R[2] , Deivakani@Brindha M[3], Rubin Raj S[4]
Assistant Professor[1], UG Scholar [2][3][4]
Department of Information Technology,
Sengunthar College of Engineering, Tiruchengode, Tamil Nadu, India
[1]karthickamsec@gmail.com, [2]amsit825@gmail.com, [3]deivakanibrindha@gmail.com, [4]rubinraj4@gmail.com

Manuscript History
Number: **IJIRAE/RS/Vol.07/Issue03/Special Issue/32.MRITSCE10111**
Received: 15, February 2020
Final Correction: 27, February 2020
Final Accepted: 10, March 2020
Published: **14, March 2020**
Editor: Dr.A.Arul Lawrence selvakumar, Chief Editor, IJIRAE, AM Publications, India

Abstract: Shortest profile-random-projection locality-sensitive hashing (SPR-LSH) using BFS Model is a probabilistic dimension reduction method which provides an unbiased estimate of angular similarity, yet suffers from the large variance of its estimation. We present a BFS (Breadth First Search) Redundant Blocking Framework that relies on the Locality-Sensitive Hashing technique for identifying candidate Crime record pairs, which have undergone an anonymization transformation. In this context, we demonstrate the usage and evaluate the performance of a variety of families of hash functions used for blocking. The parameters, of the blocking scheme, are optimally selected so that we achieve the highest possible accuracy in the least possible running time. We also introduce an LSH-based protocol (Hamming, jaccard, Euclidean distance) in order to compare the formulated Crime record pair's homomorphically, without running the risk of breaching the privacy of the underlying records.

Keywords: SPR-LSH,BFS,LSH-based protocol

INTRODUCTION

Data Mining is the process of finding relevant and useful information from databases. Although data mining is still in its infancy, companies in a wide range of industries - including retail, finance, health care, manufacturing transportation, and aerospace - are already using data mining tools and techniques to take advantage of historical data. By using pattern recognition technologies, statistical and mathematical techniques to sift through warehoused information, data mining helps analysts recognize significant facts, relationships, trends, patterns, exceptions etc., Data mining is becoming increasingly common in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce cost, enhance research, and increase sales. In the public sector, data mining applications initially were used as means to detect fraud and waste, but also have grown to be used for purpose such as measuring and improving program performance.

LITERATURE REVIEW

Maintaining an online bibliographical database, the problem of data quality

In this work, CiteSeer and Google-Scholar are huge digital libraries which provide access to (computer) science publications. Both collections are operated like specialized search engines, they crawl the web with little human intervention and analyze the documents to classify them and to extract some metadata from the full texts. On the other hand there are traditional bibliographic data bases like INSPEC for engineering and PubMed for medicine. For the field of computer science the DBLP service evolved from a small specialized bibliography to a digital library covering most subfields of computer science. The collections of the second group are maintained with massive human effort.

On the long term this investment is only justified if data quality of the manually maintained collections remains much higher than that of the search engine style collections. In this paper we discuss management and algorithmic issues of data quality. We focus on the special problem of person names.[2]

The igrid index, reversing the dimensionality curse for similarity indexing in high dimensional space

In this work, the similarity search and indexing problem is well known to be a difficult one for high dimensional applications. Most indexing structures show a rapid degradation with increasing dimensionality which leads to an access of the entire database for each query. Furthermore, recent research results show that in high dimensional space, even the concept of similarity may not be very meaningful. In this paper, we propose the IGrid-index; a method for similarity indexing which uses a distance function whose meaningfulness is retained with increasing dimensionality. In addition, this technique shows performance which is unique to all known index structures; the percentage of data accessed is inversely proportional to the overall data dimensionality. Thus, this technique relies on the dimensionality to be high in order to provide performance efficient similarity results. The IGrid index can also support a special kind of query which we refer to as projected range queries; a query which is increasingly relevant for very high dimensional data mining applications.[3]

Blocking-aware private crime record linkage

In this paper, the problem of quickly matching records (i.e., Crime record linkage problem) from two autonomous sources without revealing privacy to the other parties is considered. In particular, our focus is to devise secure blocking scheme to improve the performance of Crime record linkage significantly while being secure. Although there have been works on private Crime record linkage, none has considered adopting the blocking framework. Therefore, our proposed blocking-aware private Crime record linkage can perform large-scale Crime record linkage without revealing privacy. Preliminary experimental results showing the potential of the proposal are reported.[4]

Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions

In this work, We present an algorithm for the c-approximate nearest neighbor problem in a d-dimensional Euclidean space, achieving query time of O(dn1/c2+o(1)) and space O(dn + n 1+1/c2+o(1)). This almost matches the lower bound for hashing-based algorithm recently obtained in [27]. We also obtain a space-efficient version of the algorithm, which uses dn + n logO(1) n space, with a query time of dnO(1/c2). Finally, we discuss practical variants of the algorithms that utilize fast bounded-distance decoders for the Leech Lattice.[5]

Privacy preserving crime record linkage via grams projections

In this work, Crime record linkage has been extensively used in various data mining applications involving sharing data. While the amount of available data is growing, the concern of disclosing sensitive information poses the problem of utility vs privacy. In this paper, we study the problem of private Crime record linkage via secure data transformations. In contrast to the existing techniques in this area, we propose a novel approach that provides strong privacy guarantees under the formal framework of differential privacy. We develop an embedding strategy based on frequent variable length grams mine in a private way from the original data. We also introduce personalized threshold for matching individual records in the embedded space which achieves better linkage accuracy than the existing global threshold approach. Compared with the state-of-the-art secure matching schema [23], our approach provides formal, provable privacy guarantees and achieves better scalability while providing comparable utility.[6]

A de-randomization using min-wise independent permutations:

In this work, Min-wise independence is a recently introduced notion of limited independence, similar in spirit to pairwise independence. The later has proven essential for the derandomization of many algorithms. Here we show that approximate min-wise independence allows similar uses, by presenting a derandomization of the RNC algorithm for approximate set cover due to S. Rajagopalan and V. Vazirani. We also discuss how to derandomize their set multi-cover and multi-set multi-cover algorithms in restricted cases. The multi-cover case leads us to discuss the concept of k-minima-wise independence, a natural counterpart to k-wise independence.[7]

Investigation of techniques for efficient & accurate indexing for scalable crime record linkage & deduplication:

In this work, Crime record linkage is the process of matching records from several databases that refer to the same entities. When applied on a single database, this process is known as de-duplication. Increasingly, matched data are becoming important in many applications areas, because they can contain information that is not available otherwise, or that is too costly to acquire. Removing duplicate records in a single database is a crucial step in the data cleaning process, because duplicates can severely influence the outcomes of any subsequent data processing or data mining.

With the increasing size of today's databases, the complexity of the matching process becomes one of the major challenges for Crime record linkage and deduplication. In recent years, various indexing techniques have been developed for Crime record linkage and deduplication. They are aimed at reducing the number of Crime record pairs to be compared in the matching process by removing obvious non matching pairs, while at the same time maintaining high matching quality. This paper presents a survey of variations of six indexing techniques. Their complexity is analyzed, and their performance and scalability is evaluated within an experimental framework using both synthetic and real data sets. These experiments highlight that one of the most important factors for efficient and accurate indexing for Crime record linkage and deduplication is the proper definition of blocking keys.[9]

Some methods for blindfolded crime record linkage

In this work, a simple but effective algorithm for matching adult patients seen at more than one site in a multi-site de identified registry is described. In a data set of 19,000 records a derived match variable consisting of a 2-character prefix from both first and last name combined with date of birth has a 97% sensitivity; by contrast, an anonym zed identifier based on the patients' full names and date of birth has sensitivity of only 87%.[10]

Learning to match and cluster large high-dimensional data sets for data integration

In this work, Part of the process of data integration is determining which sets of identifiers refer to the same real-world entities. In integrating databases found on the Web or obtained by using information extraction methods, it is often possible to solve this problem by exploiting similarities in the textual names used for objects in different databases. In this paper we describe techniques for clustering and matching identifier names that are both scalable and adaptive, in the sense that they can be trained to obtain better performance in a particular domain. An experimental evaluation on a number of sample datasets shows that the adaptive method sometimes performs much better than either of two non-adaptive baseline systems, and is nearly always competitive with the best baseline system.[11]

Efficient robust private set intersection

In this work, Computing Set Intersection privately and efficiently between two mutually mistrusting parties is an important basic procedure in the area of private data mining. Assuring robustness, namely, coping with potentially arbitrarily misbehaving (i.e., malicious) parties, while retaining protocol efficiency (rather than employing costly generic techniques) is an open problem. In this work the first solution to this problem is presented.[12]

Fast locality-sensitive hashing

In this work, Locality-sensitive hashing (LSH) is a basic primitive in several large-scale data processing applications, including nearest-neighbor search, de-duplication, clustering, etc. In this paper we propose a new and simple method to speed up the widely-used Euclidean realization of LSH. At the heart of our method is a fast way to estimate the Euclidean distance between two d-dimensional vectors; this is achieved by the use of randomized Hadamard transforms in a non-linear setting. This decreases the running time of a (k,L)- parameterized LSH from O(dkL) to O(d log d + kL). Our experiments show that using the new LSH in nearest-neighbor applications can improve their running times by significant amounts. To the best of our knowledge, this is the first running time improvement to LSH that is both provable and practical.[13]

Advanced crime record linkage methods and privacy aspects for population re construction

In this work, recent times have seen an increased interest into techniques that allow the linking of records across databases. The main challenges of Crime record linkage are (1) scalability to the increasingly large databases common today; (2) accurate and efficient classification of compared records into matches and non-matches in the presence of variations and errors in the data; and (3) privacy issues that occur when the linking of records is based on sensitive personal information about individuals. The first challenge has been addressed by the development of scalable indexing techniques, the second through advanced classification techniques that either employ machine learning or graph based methods, and the third challenge is investigated by research into privacy-preserving Crime record linkage. In this paper, we describe these major challenges of Crime record linkage in the context of population reconstruction, outline recent developments of advanced Crime record linkage methods, and provide directions for future research.[14]

PROPOSED SYSTEM

Our proposed work could save many computation cycles and thus allow accurate information provided to the right people at the right time. Two considerations when forming a data warehouse are data cleansing (including entity resolution) and with schema integration (including Crime record linkage). Uncleansed and fragmented data requires time to decipher and may lead to increased costs for an organization, so data cleansing and schema integration can save a great many (human) computation cycles and can lead to higher organizational efficiency. In this work based on our previous methodologies proposed or developed for entity resolution and Crime record linkage. This survey provides a foundation for solving many problems in data Crime record linkage analysis.

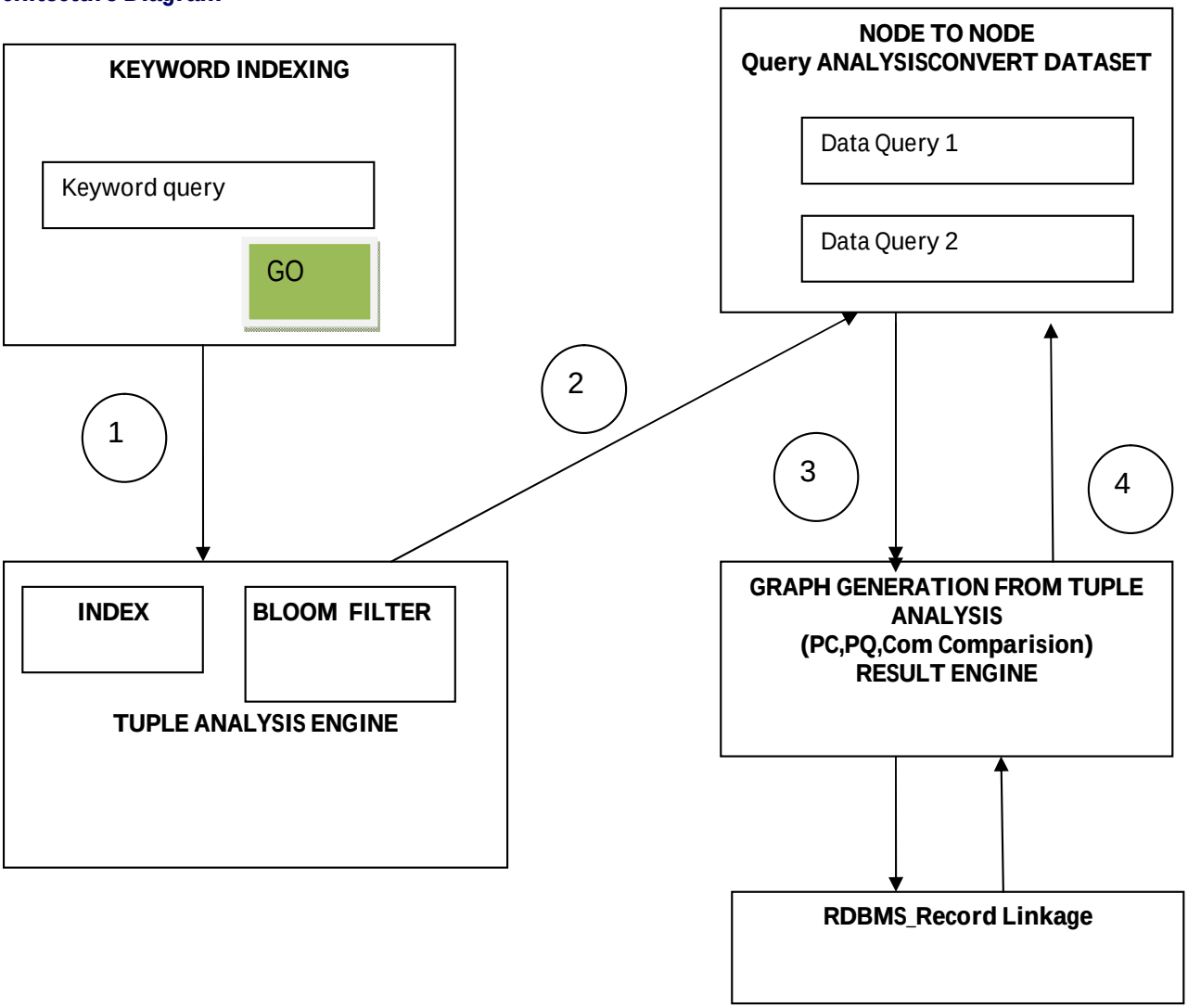
For instance, little or no research has been directed at the problem of maintenance of cleansed and linked relations. Our proposed work used an BFS (Breadth First Search) algorithm is an iterative method for finding maximum likelihood or maximum a posterior (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent entities. Crime record linkage identifies matching Crime record pairs in two separate data files. The Crime record linkage results in a classification of pairs of records as links and non links. Pairs of records which represent identical observational units are called match. Our fuzzy BFS which work based on two modules, exact match, and distance match.

SYSTEM DESIGN

Input Design

The system design is an interactive process through which requirements are translated into a blue print or a representation of software that can be accessed for quality before code generation begins. Input design is a part of overall system design, which requires careful attention. Input of data as designed as user-friendly and easier. Input design is a process of converting the user- oriented description of the input to the computer based information system into programmer oriented specification. The objective of the input design is to create an input layout that is easy to follow and prevent operator errors algorithm is used to find all pairs of shortest Crime record linkage Profiles, i.e. P. Each Crime record linkage Profile pi consists of sequence of vertices from source to destination. The graph traversal module produces set of shortest Crime record linkage Profiles between all pair of source and destination as intermediate results. All the shortest Crime record linkage Profiles are computed using well-known BFS algorithm. Secondly, the overlapped regions of shortest Crime record linkage Profiles are identified through pattern mining approach.

Architecture Diagram



We limit the number of BFS execution up to K. Therefore, instead all pair shortest Crime record linkage Profiles, K ∗ N number of shortest Crime record linkage Profiles are computed where K << N. These sample shortest Crime record linkage Profiles are further utilized for identifying the overlapped regions. The social networks are usually dense, follows the power law distribution, so even small number of shortest Crime record linkage Profiles can lead us to better or acceptable analysis.

Output Design

The output design refers to the results and information that are generated by the system for many end users. Efficient and intelligent output design improves the system relationships with the user and help in decision making. We randomly generate 100 shortest Crime record linkage Profile queries with constraint Crime record linkage Profiles for comparing the average time cost. We use the single directional BFS approach [3] using the index table for searching shortest Crime record linkage Profile. The proposed method (BFS) requires about 40K rows and BFS requires about 1,300K rows. We also randomly generate 100 shortest Crime record linkage Profile queries with constraint Crime record linkage Profile for comparing the average time cost. In Figure 1(b), LSH consumes 0.75 seconds and BFS consumes 0.72 seconds. These methods show similar time cost. From the experimental results, LSH shows higher space efficiency than BFS with similar execution time. Improves the classification accuracy. Filtering imbalance is overcome by fine turning. It can provide to very close to the class boundary and are sensitive to small changes in attribute values. Best accuracy to classify nugget data information's.

CONCLUSION

In this work, we empirically analyze the shortest Crime record linkage Profiles to anticipate the behavior of BFS algorithm on real life networks. A set of shortest Crime record linkage Profiles are evaluated using pattern mining approach. We have found that the nodes with very high degree are retained in majority shortest Crime record linkage Profiles. However, nodes with average degree are not considered by the traversal algorithm. The statistical analysis also shows the similar behavior in terms of network properties, including clustering coefficient, average shortest Crime record linkage Profile, and betweens centrality, on various types of networks. The influence of edge weights and directional information on shortest Crime record linkage Profile traversal is still an interesting area of research. It achieves over 30% mean squared error reduction over BFS-LSH in estimating angular similarity, when the Super-Bit depth N is close to the data dimension d. Moreover, BFS-LSH performs best among several widely used data-independent LSH methods in approximate nearest neighbor retrieval experiments.

REFERENCES

1. (2013).North Carolina voter registration database [Online]. Available: https://www.app.sboe.state.nc.us/data
2. (2014). The dblp computer science bibliography [Online]. Available: https://dblp.uni-trier.de/xml/
3. C.Aggarwal and P. Yu, "The igrid index: Reversing the dimensionality curse for similarity indexing in high dimensional space," in Proc. 6th ACM Int. Conf. Knowl. Discov. Data Mining, 2000, pp. 119–129.
4. A.Al-Lawati, D. Lee, and P. McDaniel, "Blocking-aware private Crime record linkage," in Proc. 2nd Int. Workshop Inf. Quality Inf. Syst., 2005, pp. 59–68.
5. A.Andoni and P. Indyk, "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions," Commun. ACM, vol. 51, no. 1, pp. 117–122, 2008.
6. L. Bonomi, L. Xiong, R. Chen, and B. C. M. Fung, "Frequent grams based embedding for privacy preserving Crime record linkage," in Proc. 21st ACM Int. Conf. Inf. Knowl. Manage., 2012, pp. 1597–1601.
7. A.Z. Broder, M. Charikar, A. Frieze, and M. Mitzenmacher, "Minwise independent permutations," in Proc. 30th ACM Symp. Theory Comput., 1998, pp. 327–336.
8. P.Christen, Data Matching-Concepts and Techniques for Crime record Linkage, Entity Resolution, and Duplicate Detection (Data-centric systems and applications). New York, NY, USA: Springer, 2012.
9. P.Christen, "A survey of indexing techniques for scalable Crime record linkage and deduplication," IEEE Trans. Knowl. Data Eng., vol. 12, no. 9, pp. 1537–1555, Sep. 2012.
10. T.Churches and P. Christen, "Some methods for blindfolded Crime record linkage," BMC Informat. Decision Making, vol. 4, p. 9, 2004.
11. W.Cohen and J. Richman, "Learning to match and cluster large high-dimensional datasets for data integration," in Proc. ACM Int. Conf. Knowl. Discov. Data Mining, 2002, pp. 475–480.
12. D.Dachman-Soled, T. Malkin, M. Raykova, and M. Yung, "Efficient robust private set intersection," in Proc. 7th Int. Conf. Appl. Cryptography Netw. Security, 2009, pp. 125–142.

13. M.Datar, N. Immorlica, P. Indyk, and V. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in Proc. 20th Symp. Comput. Geometry, 2004, pp. 253–262.
14. E.Durham, "A framework for accurate efficient private Crime record linkage," Ph.D. dissertation, Faculty of the Graduate School, Vanderbilt University, Nashville, TN, USA, 2012.
15. C.Dwork, "Differential privacy," in Proc. 33rd Int. Colloquium, 2006, pp. 1–12.
16. Z.Erkin, M. Franz, J. Guajardo, S. Katzenbeisser, I. Lagendijk, and T. Toft, "Privacy-preserving face recognition," in Proc. 9th Int. Symp. Privacy Enhancing Technol., 2009, pp. 235–253.
17. C.Faloutsos and K. Lin, "Fastmap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets," in Proc. ACM Int. Conf. Manage. Data, 1995, pp. 163–174.
18. H.Gabriela and F. Martin, "Cluster preserving embedding of proteins," Dept. Comput. Sci., Center Discrete Math. Theor. Comput. Sci., Rutgers Univ., Piscataway, NJ, USA, Tech. Rep. 99-50, 1999.
19. Satish Kumar.R. and Sanavullah, M.Y.," An Improved Intelligence Approach for FDTD Modeling and simulation of Microwave Heating for Egg Pasteurization", Archives Des Sciences, ISSN 1661-464X, Vol.65, p.10, Aug.2012.
20. Satish Kumar.R. and Sanavullah, M.Y.," Optimization using Genetic Algorithm for FDTD Modeling and simulation of Microwave Heating for Egg Pasteurization", European Journal of Science Research , ISSN 1450-216X, Vol.84, No.1.pp.81-90,2012.
21. Satish Kumar.R. and Sanavullah, M.Y.," Theoretical and Experimental identification of cooking spot for shell eggs without explosions in a domestic Microwave Oven", Canadian Journal on Electrical and Electronics Engineering, Vol. 1, No. 4, pp.71-78, June 2010.
22. Satish Kumar.R, K.Uma Devi, and Sanavullah, M.Y.," Performance Analysis of using Exterior Rotor Permanent Magnet Brushless DC (ERPMBLDC)Motor", Improvement by a Novel Peak Torque Excitation Technique", International journal of Innovative research in Advanced Engineering, Vol. 1, 2012, pp. 1-7 (Impact factor 1.311).
23. Raja, G P& Mangai, S 2018, 'Investigation On Optimization, Prioritizing and Weight Allocation Techniques for Load Balancing and Controlling Multimedia Traffic in Wireless Mesh Network', International Journal of Business Information Systems, SCOPUS Indexed Journal (Inderscience) - (P ISSN No: 1746-0972). Published Online: 10th Feb 2020, DOI: 10.1504/IJBIS.2020.105161.IF: 0.72.
24. Raja, G P& Mangai, S 2017, 'Firefly Load Balancing Based Energy Optimized Routing for Multimedia Data Delivery in Wireless Mesh Network', Cluster Computing-The Journal of Networks Software Tools and Applications, SCOPUS Indexed Journal (Springer) - (E ISSN No: 1573-7543).Published Online: 27th Dec 2017, https://doi.org/10.1007/s10586-017-1557-1. IF: 2.040.
25. Geetha. E & Nagarajan. C , 2019, 'Stochastic Rule Control Algorithm Based Enlistment of Induction Motor Parameters Monitoring in IoT Applications', Springer, Wireless Personal Communications. October 2018, Volume 102, Issue 4, pp 3629 - 3645.

