Semantic Deduplication in Databases

Anju K S, Sadhik M S, Surekha Mariam Varghese

Abstract:* *The presence of semantic duplicates imposes a challenge on the quality management of large datasets such as medical datasets and recommendation systems. A huge number of duplicates in large databases necessitate deduplication. Deduplication is a capacity optimization innovation that is being utilized to dramatically enhance storage efficiency. For this, it is required to identify the copies, with a quite solid approach to find as many copies as achievable and sufficiently ample to run in a sensible time. A similarity-based data deduplication is proposed by combining the methods of Content Defined Chunking (CDC) and bloom filter. These methods are exploited to look inside the files to check what portions of the data are duplicates for better storage space savings. Bloom filter is a probabilistic data structure and it is mainly used to decrease the search time. To enhance the performance of the system, methods like Locality Sensitive Hashing (LSH) and Word2Vec are also used. These two techniques are used to identify the semantic similarity between the chunks. In LSH, Levenshtein distance algorithm measures the similarity between the chunks in the repository. The deduplication performed based on semantic similarity checking improves the storage utilization and reduces the computation overhead effectively

Index Terms: Deduplication, Locality sensitive hashing, Bloom filter

I. INTRODUCTION

Today Database has pivotal job in IT industry. Precision of database is vital for do tasks. Database has huge dimension of nature of the data, it can huge cost suggestions to a framework that depends on data to capacity and direct business. In computerized media volume of data is expanding in quick way. It turn into a testing issue for information chairmen. Deduplication of record is the undertaking of recognizing information from an information store, records allude to a similar genuine element or article despite incorrect spelling words, types, diverse composition styles or even extraordinary construction portrayals or information types. For expelling reproductions from information archives, huge speculations are needed from private and government associations for creating techniques. Diverse characteristic are exist for two lines in a store then they deliver the equivalent physical reality, call them semantic copies. The procedure deduplication is the finished undertaking for evacuation of copies records. Storehouse copy activity is essential and fundamental paying little respect to the activity to be embraced on the information. In database copies are available all things considered 4% of the original information [1]. At the point when database estimate winds up bigger, recovery process is increasingly costly and troublesome. Because of copies numerous issues are existed.

For performing deduplication numerous techniques have been proposed and utilized. The drawback is no basic best arrangement is existed. Deduplication procedure has been produced with various structures by thinking about execution and overhead. Alternate qualities of informational collections like framework limit and deduplication time is likewise considered. In the proposed framework, emergency clinic databases is utilized. Understanding subtleties are put away as modified works. Theoretical contains the subtleties of the patients and their illnesses. Numerous patients have a comparative illness, so a probability of deduplication is occured. For copy recognition, techniques utilized is Locality Sensitive Hashing (LSH), lump level deduplication and sprout channel idea.LSH is existed in a different arrangement of scholarly research zones including advanced flag handling, data recovery, measurements, machine learning, and information mining.

A vital strategy for substitute portrayals is remove measure. There are a few techniques are likewise present like Euclidean separation measures[2], cosine remove[3], etc. It made complexities when working with vectors in higher measurements. Models is Levenshtein separation or Hamming separation. So in the proposed framework, LSH utilizes Levenshtein separate calculation for string correlation. Piece level deduplication is another strategy that is utilized in database deduplication. Here the given string that contains manifestations of the ailments that are isolated into settled size lumps. Check the closeness of lumps by utilizing the strategies for word2vec.The word2vec model [4] and its applications have as of late pulled in a lot of consideration from the machine learning network. These thick vector portrayals of words learned by word2vec have astoundingly been appeared to convey semantic implications and are valuable in a wide scope of utilization cases going from common language handling to arrange stream information investigation. Blossom channel is utilized for distinguishing similar pieces in the given database.

II. LITERATURE SURVEY

The programmed cancellation of copy information in an archive, regularly known as deduplication, is dynamically acknowledged as an effective strategy to decrease stockpiling costs. Hence, it has been ap-handled to various capacity types, including chronicles and reinforcements, essential stockpiling, inside strong state drives, and even to arbitrary access memory. Jinfeng Liu [5] propose a technique for secure closeness based information deduplication conspire by consolidating the strategies for sprout channel and substance characterized piecing, which can consequently lessen the calculation overhead by just performing deduplication tasks for comparative

Revised Manuscript Received on December 22, 2018.

Anju K S, Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, India.

Sadhik M S, Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, India.

Surekha Mariam Varghese, Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, India.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

documents. Sprout Filter is a probabilistic information structure that utilizes different hash capacities to store information in a substantial piece exhibit. It is basically utilized for enrollment inquiries when managing substantial informational collections.

Gan et al [8] sproposed a LSH calculation, Collision Counting LSH (C2LSH) that utilizes m-base LSH capacity to frame dynamic com-pound hash works rather than generally utilized static compound hash work. In C2LSH an impact limit is utilized for information item to expand the inquiry quality. Broder et al. [9] planned a LSH conspire dependent on Jaccard closeness of sets which utilizes Minhash capacities to take distinctive stages to figure Jaccard similitude for sets [7]. A few varieties have been proposed to enhance the execution of this LSH conspire in [10].

In 2005, Andrew McCallum [6] gives experimental proof that utilizing coverings for grouping can increment computational proficiency by a request of greatness without losing any bunching precision. The fundamental bunching approach use here is Greedy Agglomerative Clustering. So as to accomplish grouping in this area must give a separation metric to the space of bibliographic references. String alter remove is the best strategy for figuring the separation between the strings. It is dictated by powerful programming utilizing diverse expenses related with different change rules. One trouble with applying string alter separations to the space of references is that one can only with significant effort speak to handle transpositions as a nuclear cost activity in the dynamic programming. The string alter remove counts are moderately costly, since a dynamic genius gram must be comprehended to represent conceivable inclusions, cancellations, and transpositions.

III. PROBLEM DEFINITION

Before Present day medicinal innovation has turned out to be subject to the capacity to gather and store tolerant records, picture and video records, and a huge assortment of report types. There are additionally more subspecialties today, so a solitary patient may have many consideration suppliers concentrated on various parts of the patient's all out consideration. The unavoidable outcome is the expansion of information at a surprising rate. A lot of this information is unstructured and is overseen by an assortment of divergent applications and frameworks. The greatest difficulties in the changing Social insurance IT condition today are the exponential development of Medicinal services information, the growing access to that information crosswise over ventures, fortes, analysts, scholastics and the stricter authorization of administrative consistence estimates that clinics need to meet. Comport can enable the medical clinic to address these difficulties by advancing the emergency clinic's present frameworks to diminish unpredictability, decrease expenses and increment the accessibility of applicable information to those that need it to give enhanced patient consideration.

IV. PROPOSED METHODOLOGY

Technique for the automated deduplication of data and text files in the database is proposed. The proposed system is based on the methods like chunk level deduplication, Locality Sensitive Hashing (LSH), using the concept of bloom filter and wod2vec embedded method. The data that are related to the hospital database. That is, in hospitals there are many records related to patients. It contains the information about the patient details as well as their disease descriptions. So there are many patients have similar diseases. Based on their disease characteristics here perform the deduplication. The details that is stored in the databases, so deduplication can be performed in databases. The details that are represented in the form of abstracts. So to perform deduplication first partition the data into settled size chunks.

In this proposed system, the abstract data of the patient is newly entered into the database. The preprocessing module can take place. In this module, the given data can be tokenized and then the tokens that are lemmatized. In chunking module, the preprocessed data are decomposed into chunks and each chunk is fixed in size. The similarity checking module checks the similarity of each chunk by using different methods such as locality sensitive hashing, Bloom filter and Word2vec method. The similarity checking module also checks the semantics of each chunk. Here it will be performed based on the synonyms of each string in the chunk. For that use LSH algorithm and a dataset that contains the synonyms of the strings.

The proposed system has 4 stages – Preprocessing, chunking, similarity checking and Deduplication process. The proposed architecture is shown in Figure I. It shows how each of the phases is related to its predecessor phases. Evaluation of the example is performed in the preprocessing step. This examination is performed by looking over all the example information so as to distinguish the tasks of institutionalization and standardization to be made on the information to be standardize and furthermore to recognize "stop words". These tasks are vital for the thought of the syntactic contrasts essential for the genuine semantic copies. In preprocessing standardization and normalization is performed. In this step remove the stop words and create tokens. Chunking is a term alluding to the way toward taking individual snippets of data (pieces) and gathering them into bigger units. By collecting each chunk into a substantial entirety. Effective chunking is one of the key components that choose the general deduplication execution. The abstract data of patients that are divided into fixed size chunks and then it will be passed into next step. Each chunk will have fixed number of strings.

Similarity checking is the process of identifying the similarity between the disease description in the database. In the abstract format, the description of the same disease will be explained in different ways. But they are semantically similar. So in this module also check the semantic similarity between chunks. In similarity checking, several methods are used to identify the similarity between chunks. They are Locality sensitive hashing (LSH), Bloom filter and Word2vec.

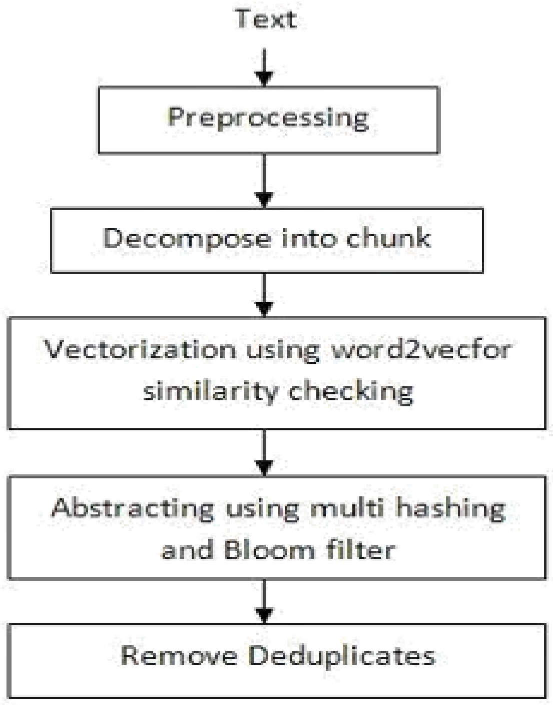
Text
Preprocessing
Decompose into chunk
Vectorization using word2vecfor similarity checking
Abstracting using multi hashing and Bloom filter
Remove Deduplicates

Fig. I. Proposed system architecture

In LSH, distance measure become important in this world of alternate representations of documents. There are Euclidean distance measures, cosine distance, and so on. On a basic two dimensional plot, this can be as simple as measuring the distance between two points. It can get more complex when working with vectors in higher dimensions. Alternatively, when working with strings or even just bits, other distance measures can be leveraged. Levenshtein distance (LD) is a proportion of the likeness between two strings, which we will allude to as the source string (s) and the objective string (t).

Bloom filter is another method that is used in the proposed system. Bloom Filter, a probabilistic data structure that utilizes various hash capacities to store information in a broad piece cluster. It is essentially utilized for membership questions when managing vast datasets. It decreases O (r) (r-number of elements) seek time to steady time. The seek exactness relies upon the extent of Bloom Filter (s) and a number of hash functions (p). The real preferred standpoint of different hash functions is that look exactness is made strides. The Bloom filters have the property that the false positives can be reduced. The larger the size of bloom filter, less are the false positives and smaller the size, more are false positives. The false positive rate is approximately (1-(*eprs⁄*)*p* with r number of elements which are entered; various values of m and k are used to construct the filter for the requisite application. Therefore, the more hash functions slow the bloom filter and it also fills up quickly. However, if few hash functions are used in bloom filter, it results in more false positives. Therefore, given the values of m and n, the function can be used to select the optimal value of p=s/r. ln2.

So, the steps used in choosing the size of bloom filter are:

1. For a value of r (number of items)
2. A value for s is selected
3. The suitable value of k is calculated
4. The error rate is calculated for the selected r, s and p values. If the error rate is not acceptable, return to step 2 and the value of m is changed.

Word2Vec is another method that is used to analyze the semantic similarity between the chunks. Words that we know to be equivalent words will in general have comparative vectors as far as cosine comparability and antonyms will in general have divergent vectors. Considerably more shockingly, word vectors will in general comply with the laws of relationship.

Deduplication is utilized to enhance stockpiling usage and generally execution of the framework. As soon as the data is uploaded, the data is divided into chunks in order to provide an efficient storage. Since there is a high chance for duplicate chunks occurrence in the repository. So it is mandatory for identifying and removing the duplicates in the repository. The identification of duplicate chunks are done using Levenshtein Distance Algorithm (LDA) and similarity check. For removing duplicates, first, the duplicates are identified by checking thresh-old value, which is obtained by performing LDA and semantics on the chunks. If the threshold value of the chunk is above 80% then it is considered to be duplicate chunk and it is removed else the new chunk is stored in the repository.

V. RESULT AND PERFORMANCE ANALYSIS

The performance measures for proposed system are evaluated relevant to the three performance parameters precision, recall, and accuracy[11]. The detected seizure is considered as positive class and detected normally is considered as a negative class. Therefore TN, TP, FN, FP are defined as follows:

TP – Number of correctly matched chunks
FP – Number of correctly matched chunks which are identified as incorrect
TN – Number of incorrect chunks
FN – Number of incorrect chunks which are identified as correct

The proposed system employs 40 recording from the dataset as test samples. The performance parameters are calculated by varying the threshold value as 55%, 60%, 65% etc up to 100%.

A. With Bloom Filter

The Bloom filter is a method for utilizing hash functions to decide set enrollment. Bloom filter discover application wherever quick set enrollment tests on vast informational indexes are required. Such applications incorporate spell checking, differential document refreshing, dispersed system stores, and textual examination. It is a probabilistic strategy with a set mistake rate. Utilizing bloom filter enhances the general execution of the framework by lessening the general inquiry time [12]. There is no chance of occurrence of false negatives. That is accuracy, precision and recall level of the system will be improved. Fig. II shows the accuracy, precision and recall with bloom filter.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

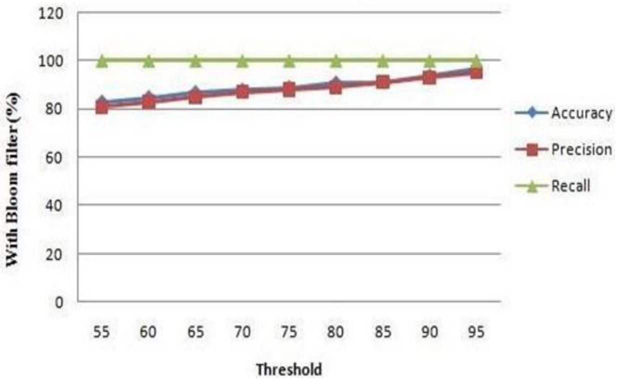

Fig. II. With bloom filter

B. Without Bloom Filter

Bloom Filter is a probabilistic data structure that utilizes multiple hash functions to store data in a large bit array. Without bloom filter there is a probability of occurring false negatives in the absence of bloom filter and it is lead to degradation of overall system performance. Fig. III shows the accuracy, precision and recall without bloom filter.

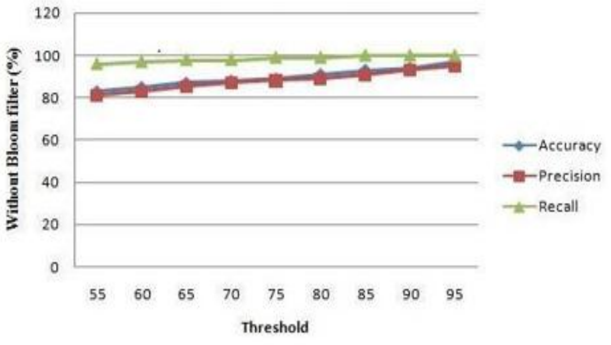

Fig. III. Without bloom filter

C. Without Semantics

The Semantic similarity of sentences depends on the implications of the words and the linguistic structure of the sentence. In the event that two sentences are comparable, auxiliary relations between words could conceivably be comparative. Basic relations incorporate relations among words and the separations between words. On the off chance that the structures of two sentences are comparable, they are increasingly conceivable to pass on comparative implications. The given two sentences as an input to this process; first the words of two sentences are compared. If the two words of the sentences are matched, its similarity score is calculated which are based on syntactic level. The probability of acquiring false positives is high in the absence of semantic analysis and this leads to the decreasing of overall system performance. Fig. IV shows the accuracy, precision and recall with semantics.

D. *With Semantics*

Bloom The overall performance f the system can be increased by using the semantic analysis method. Initially, the repository has the size more than 2MB. Using semantic analysis, the size of the repository compressed to less than 1MB. About more than 20% improvement in the overall performance of the system. Fig. V shows the accuracy, precision and recall with semantics.

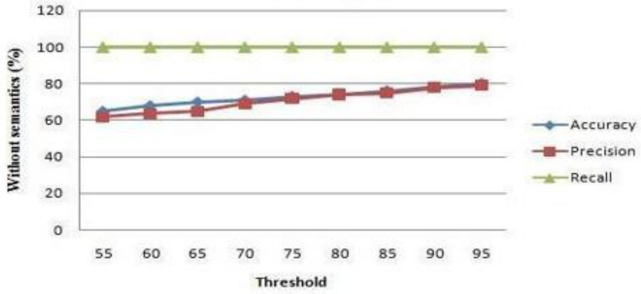

Fig. IV Without semantics

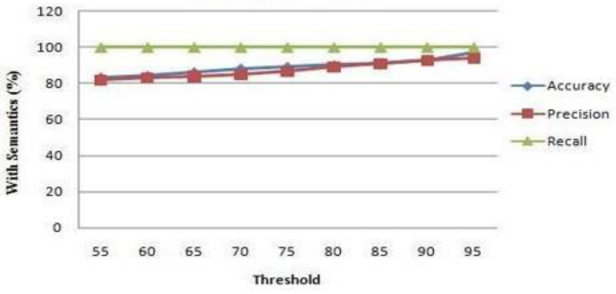

Fig. V. With semantics

VI. CONCLUSION

Despite the fact that different deduplication methods have been proposed and utilized, no single best arrangement has been produced to deal with a wide range of redundancies. The objective of this work is to improve the storage utilization in databases. From the experiments using the methods: LSH, bloom filter and semantic analysis improve the efficiency of the system. Chunk level deduplication divides data into chunks and bloom filter it reduces the search time. It can provide efficient deduplication in databases. The test results demonstrate that framework can achieve reduction in storage size and overall search time. From the observation, it could be concluded that the presence of bloom filter and semantic analysis greatly improves the system performance. The performance of the system can be improved by adding more hash functions in future.

REFERENCES

1. Ibrahim M N, Amolo-Makama Ophéli. Fast Semantic Duplicate Detection Techniques in Databases,Journal of Software Engineering and Applications, 2017, 10, 529-545
2. Coerjolli, D. and Montanvert, A. 2007. Optimal separable algorithm to compute the reverse Euclidean distance transformation and discrete medial axis in arbitrary dimension. IEEE Trans. Patt. Anal. Mach. Intell. 29, 3 (March), 437–448
3. Wolf, L., Hassner, T., Taigman, Y.: Similarity scores based on background samples. In: Zha, H., Taniguchi, R.-i., Maybank, S. (eds.) ACCV 2009. LNCS, vol. 5995, pp. 88–97. Springer, Heidelberg (2010)
4. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. CoRR, abs/1301.3781, 2013.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

5. Jinfeng Liu a, Jianfeng Wanga,b, Xiaoling Taoc and Jian Shend. (2017) Secure similarity-based cloud data deduplication in Ubiquitous city, Pervasive and mobile computing.
6. McCallum, Nigam, K. and Ungar, L.H. Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching. Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data mining, Boston, 20-23 August 2000, 169-178.
7. Bianco, G.D., Galante, R. and Heuser, C.A. A Fast Approach for Parallel Deduplication on Multicore Processors. 26th Symposium on Applied Computing SAC'11, TaiChung, 21-25 March 2011, 1027-1032.
8. J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer, "Reclaiming space from duplicate files in a serverless distributed file system," Proc. International Conference on Distributed Computing Systems (ICDCS), pp. 617–624, 2002
9. A. Broder et al., "Min-wise independent permutations", Proc. Theory of computing, ACM Symposium, New York, USA, pp. *327-336, 1998.*
10. J. Xu, E. Chang, and J. Zhou, "Leakage-resilient client-side deduplication of encrypted data in cloud storage," ePrint, IACR, http://eprint.iacr.org/2011/538.
11. J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves", in Proc. of the 23rd International Conference on Machine Learning, 2006, pp. 233-240.
12. Ken Christensen, Allen Roginsky, and Miguel Jimeno. A new analysis of the false positive rate of a Bloom filter. Information Processing Letters, 110(21):944 – 949, 2010.

AUTHORS PROFILE

Anju K S received Bachelor of Technology in Computer Science and Engineering from Government Engineering College Idukki in 2017 and currently pursuing Master of Technology in Computer Science and Engineering from Mar Athanasius College of Engineering, Kothamangalam affiliated to APJ Abdul Kalam Technological University. Her research interest is in Natural Language Processing, Deep Learning and Data Mining.

Sadhik M S received Bachelor of Technology in Computer Science and Engineering from Ammini College of Engineering, Palakkad in 2016 and Master of Technology in Computer Science and Engineering from Mar Athanasius College of Engineering, Kothamangalam in 2019. His research interest is in Deep Learning, Natural Language Processing and Data Mining.

Dr. Surekha Mariam Varghese is currently heading the Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India. She received her B-Tech Degree in Computer Science and Engineering in 1990 from College of Engineering, Trivandrum affiliated to Kerala University and M-Tech in Computer and Information Sciences from Cochin University of Science and Technology, Kochi in 1996. She obtained Ph.D in Computer Security from Cochin University of Science and Technology, Kochi in 2009. She has around 27 years of teaching and research experience in various institutions in India. Her research interests include Machine learning, Network Security, Database Management, Data Structures and Algorithms, Operating Systems and Distributed Computing.