Review on Record LINKAGE and Deduplication based on Suffix Array Indexing

Warke Yamini
Dr.D.Y. Patil SOET, Pune
Savitribai Phule Pune University
Pune, India

Arti Mohanpurkar
Dr.D.Y. Patil SOET,Pune
Savitribai Phule Pune University
Pune, India

ABSTRACT

Record linkage is a momentous process in data soundness which is used in combining, matching and duplicate removal from more than two databases that refer to the same entities. Deduplication is the process of taking off duplicate records in a united database. Now a day, data cleaning and standardization becomes a pompous process. Due to yielding capacity of today’s database, discovering matching records in united database is a crucial one. Indexing technique specifically suffix array is used to efficiently implement record linkage and deduplication.

Keywords

Record linkage, suffix array, blocking

1. INTRODUCTION

As various government agencies, business, and research projects assemble exceptionally large amounts of data, skill that permit productive processing, examining and mining of large databases have in recent years admire both academy and industry for holding the attention. Linking or matching records which related to same entity from more than two database become grater task in the phase of assembling data of many data mining project. The aim of such linkages is to match and make concrete of all records relating to the same entity, such as sick person, a purchaser, enterprise, a client product, a copyright citation. To permit further use of existing data sources for new studies and minimize the cost and determined attempt in data acquisition ,record linkage and deduplication can be used.That is why removing duplicate records in a single database is important one. In motor servicing station ,refer the example given in table 1. The first name refers to Business name and its residential location, the second is the name of the holder of the business with his home address. Third is the address of accountant who does the books for the company. The name ‘ P A S.Inc’ is an abbreviation of the actual name of the business ‘Patil A Sumit’ which is the holder of motor servicing station. It is possible that different list Associated with the set of businesses may have entries corresponding to anyone of the listed forms of the entity which is the motor servicing station. In such case there may be duplicate Entries found, that duplications are corrected when that particular individual return the form. but it is very tedious task if we want [1]that information after some years, as that person may be not at the corresponding address. Table 1.illustrates this example.

We can take other example of banking system ,one person may have more than one account in different banks. and that person may use certain different name in each bank. for example. suppose In IDBI bank he has kept name like Bhirud Sparsh P and in CANARA bank has kept name as B Sparsh p and in HDFC like Bhirud S P. All these names are referred to same entity that is (Bhirud Sparsh P).In order to find out that whether that all names are referred to same person, record linkage is used. As the amount of digital information is rapidly increasing all over the world and most of the data is unstructured one such as image,audio,video &document files. This rapid growth of data size causes several problems such as storage limitation, increasing cost. We can overcome this problem by using deduplication technique.

Table 1. Examples of Names and Addresses Referring to the Same Business Entity

Associated Address	Description
SR.#81/7 Near Yogi Hotel Tathwade, Mulashi,Pune,Maharashtra.	Residential location of business
Patil A sumit 345 Shri ram park Dhanori Toad No.7	Residential location of holder of business.
P A S,Inc C/o sunil pegonkar Dhanori road no. Vishrantvadi chowk .Pune.	Incorporated name of business accountant does books and government forms.

One familiar example, when any faculty in our college send us mail at that time faculty has to send same mail to all students so there are too many duplicate copies of same mail in data server. In this case we can use deduplication .we can keep only reference of that mail on server instead keeping whole copy. Suffix array is used in pattern searching problem in large database. here we can take example suppose there are two people sunny & joy who are playing the very uninteresting game, sunny has very large string and joy asked sunny that ‘ does the following substring is substring of yours’? joy had asked too many questions to sunny ,sunny has to give the answer as early as possible. Sunny is programmer so he think that it would be better to know all the substrings that appear in joys string .before doing all this work sunny is wondering about how many substring will be there. in joys string. Solution to this is that suppose we assume that sunny has string‘’babc’’ hassubstringb’’,ba’’,bab’’,babc’’,bc’’,a’’,ab’’,abc’’,and c’’. determined by the path starting from the root and going toward nodes 2, 3, 4, 5, 6, 7, 8 and 9 in this order. because building the suffix tree is not always a pleasant job and has a quadratic complexity, an approach using suffix arrays would be much more useful[2][13]. Suffix array is useful for pattern matching using reduced space on disk suffix array. In our computer lots of data is present..we cant store our whole data in main memory we need secondary storage.like hard disk,cd,dvds.so here we can go by one way that we keep only important data or reference to main data in main memory and remaining one at secondary storage. so available space get minimized .[3][12]

There are some advantages &limitation of suffix array. When suffix array is used for pattern matching in disk, save the no of disk access and space.[3] [14]To find longest common substring suffix array is useful. Limitation is that suffix array is costly construction process.

2. EXISTING SYSTEM

In existing suffix array based indexing only suffixes down to minimum length lm are inserted into suffix array. for example ,for BKV pitambar and lm=5,thevalues‘pitambar’,’itambar’,’tamb’,’ambar’, will be generated ‘and identifiers of all records that have this BKV will be inserted into corresponding four inverted index list.

Identifires	BKVs (Givenname)	Suffixes
R1	Yamini	Yamini,amini ,mini,Ini
R2	Damini	Damini,amini, Mini,Ini
R3	Kamini	Kamini,amini ,mini,Ini
R4	Saudamini	Saudamini,audamini,udamini,damini,amini, mini, Ini

Fig .1. suffix array based indexing with given name used as BKVs, a minimum suffix length lm=3 and a maximum block size bm=2.the table on right hand side shows the resulting sorted suffix array. The block with suffix value ‘amini’ and ‘mini’ will be removed because it contains more than bm record identifiers.

To limit the maximum size of blocks a second parameter, bm,permit the maximum number of record identifiers in block to be set .Blocks which contain more than bm record identifiers will be removed from suffix array. For example in fig 1.,block with bm=2 having suffix value ‘amini’, ‘mini’ and ‘ini’will be removed since it contains four record identifiers[4]. As can be seen in fig.1, one problem with suffix array based indexing is that errors and variations at the end of BKVs will result in records being inserted into unusual blocks, and true matches get lost.

Suffix	Identifiers
Yamini	R1
Amini	R1,R2,R3,R4
Mini	R1,R2,R3,R4
Damini	R2,R4
Kamini	R3
Suadamini	R4
Udamini	R4
Audamini	R4
Ini	R1,R2,R3,R4

3. LITERATURE REVIEW

In [1946] H.L.Dunn published why record linkage is necessary .he had given that no one has distinguished post in the Book of our Life than the registrar. The registrar's basic responsibility has always been,To obtain fullness and exactness of registration, to maintain records, to approve from records . The exactness of important records would be enhanced because of unsettled that would show up. The fullness of important records would give moral benefit because subsequent documents would show that previous records which should have been filed had not been placed on file. Approval would become more secure from fraud. For Example, birth records of dead people could not be approved for illegal purposes. In [1962] H. B. Newcomb and J. M. Kennedy proposed that IN searching and updating any large files of documents some difficulties meet unexpectedly. This can be identified basically on the basis of names and other personal information. Generally, name retrieval systems have been developed. In [1969] I. P. Fellegi and A. B. Sunter proposed that, mathematical model is important for identifying records in two file which represent same person, comparison should be made between that record values and decision should be taken whether that comparison pair represent same person. On the basis of that three decision will be taken link ,nonlink,possible link.In [1993] L. Gill, M. Goldacre, H. Simmons, G. Bettley, and M. Griffith described how medical records are linked in computer.they had concluded that using array of idetifiers than matching record character by character comparisons we can achieve much correct matching. In [2000] L. Jin, C. Li, and S. Mehrotra Introduced different methods with which we can extract beneficial knowledge from data. As day by day online data growing swiftly because of internet and widely distributed use of database, considerable need of KDD is required. In [2003] L. Jin, C. Li, and S. Mehrotra. Describes a novel approach, For each attribute of records, first mapping values to a multidimensional Euclidean space that preserves domain-specific similarity. Many mapping algorithms can be applied and taking Fast Map approach as an example. In[2003]R. Baxter, P. Christen, and T. Churches proposed the comparison of two new blocking methods, bigram indexing and canopy clustering with TFIDF (Term Frequency/Inverse Document Frequency), With this two older methods of standard traditional blocking and sorted neighborhood blocking, recent blocking methods such as bigram indexing and canopy clustering provide scalable blocking methods while maintaining or improving upon record linkage accuracy. In[2005] Akiko Aizawa, Keizo Oyama proposes a rapid and capable method for linkage detection. The features of the proposed approach are: first, it utilizes a suffix array structure make possible linkage detection using variable length n-grams. Second, it dynamically produces blocks of possibly corresponding records using ‘blocking keys’ extracted from already known reliable linkages. In experiments the proposed method was applied to the integration of four bibliographic databases, which scale up to more than 10 million records. In [2006] William E. Winkler provides an overview of record linkage. Record linkage is also referred to as data cleaning or objects identification. It gives background on how record linkage has been applied in matching lists of businesses. It points out directions of research for improving the linkage methods. In [2007]Su Yan, Dongwon Lee, Min-Yen Kany, C.

Lee validate hypothesis, by taking a classical record linkage algorithm, the sorted neighborhood method (SNM), and demonstrate how we can achieve improved accuracy and performance by adaptively changing its fixed sliding window size. Also analytically and empirically validated algorithm, using both real and synthetic data sets of digital libraries and other domains fixed during the execution. In [2008] L. Huang, L. Wang, and X. Li. proposed a framework for implementing the longest common Subsequence (LCS) as a similarity measurement in reasonable computing time, which leads to both high precision and recall. Second, they had presented an algorithm to get a trustable partition from the LCS to reduce the negative impact from templates used in web page design. A inclusive experiment was conducted to evaluate our method in terms of its impressiveness, capability, and quality of result. More specifically, the method has been successfully used to partition a set of 430 million web pages into 68 million subsets of similar pages, which describes its effectiveness. For quality, they had compared method with simhash and a Cosine-based method through a sampling process (Cosine is compared to LCS as an alternative similarity measurement). The result showed that algorithm reached an overall precision of 0.95 while simhash was 0.71 and Cosine was 0.82. At the same time method obtains 1.86 times as much recall as simhash and 1.56 times as much recall as Cosine.In [2009] C. R. Arvind Arasu and D. Suciu presented a declarative framework for collective deduplication of entity references in the presence of constraints. Constraints occur naturally in numerous data cleaning domains and can improve the quality of deduplication. An example of a constraint is “each paper has a unique publication venue”; if two paper references are duplicates, then their associated conference references must be duplicates as well. Framework supports collective deduplication, meaning that we can dedupe both paper references and conference references collectively in the example above. Framework is based on a simple declarative Data log style language with precise semantics. Most previous work on deduplication either ignores constraints or uses them in an ad-hoc domain-specific manner. Also, using a prototype implementation, algorithms scale to very large datasets. In [2010] Thomas Bernecker, Hans-Peter Kriegel, Nikos Mamoulis, Matthias Renz, and Andreas Zuefle proposed a framework which incrementally calculates example, ranking position and the probability of the object falling at that ranking position. The resulting rank probability distribution can take as input for several state-of-the-art probabilistic ranking models. Framework reduces this to a linear-time complexity while having the same memory requirements, facilitated by incremental accessing of the uncertain vector instances in increasing order of their distance to the reference object.

4. CONCLUSION AND FUTURE SCOPE

Suffix array blocking is highly capable and relevant to outperform traditional methods in scalability, at the cost of indicative amount of accuracy, depending on the attributes of the data used. Our improvement derives these qualities, but significantly improves the accuracy at the cost of very small amount of extra processing.

In future work we can use link list instead of using suffix array. As in suffix array we have space limitation which we can solve by using link list.

5. REFERENCES

[1] “Winkler, William E. "Overview of record linkage and current research directions."US Bureau of the Census. 2006.,” Tech. Rep. RR2006/02, 2006.

[2] Vladu, Adrian, and Cosmin Negruşeri. "Suffix arrays–a programming contest approach." (2005).

[3] Gog, Simon, Alistair Moffat, J. Culpepper, Andrew Turpin, and Anthony Wirth. "Large-scale pattern search using reduced-space on-disk suffix arrays." IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 8, AUGUST 2014

[4] Christen, Peter. "A survey of indexing techniques for scalable record linkage and deduplication." Knowledge and Data Engineering, IEEE Transactions on 24.9 (2012): 1537-1555.

[5] P. Christen, “A comparison of personal name matching: Techniques and practical issues,” in Workshop on Mining Complex Data, held at IEEE ICDM’06, Hong Kong, 2006.

[6] Christen, P., Churches, T., & Hegland, M. (2004). Febrl–a parallel open source data linkage system. In Advances in knowledge discovery and data mining (pp. 638-647). Springer Berlin Heidelberg

[7] Clark, D. E. (2004). Practical introduction to record linkage for injury research. Injury Prevention, 10(3), 186-191.

[8] Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Eng. Bull., 23(4), 3-13.

[9] Churches, Tim, et al. "Preparation of name and address data for record linkage using hidden Markov models." BMC Medical Informatics and Decision Making 2.1 (2002): 9.

[10] Christen, Peter, and Karl Goiser. "Quality and complexity measures for data linkage and deduplication." Quality Measures in Data Mining. Springer Berlin Heidelberg, 2007. 127-151. [11] L. Gu and R. Baxter, “Decision models for record linkage,” in Selected Papers from AusDM, Springer LNCS 3755, 2006

[11] Su, Weifeng, Jiying Wang, and Frederick H. Lochovsky. "Record matching over query results from multiple web databases." Knowledge and Data Engineering, IEEE Transactions on 22.4 (2010): 578-589.

[12] Dey, Debabrata, Vijay S. Mookerjee, and Dengpan Liu. "Efficient techniques for online record linkage." Knowledge and Data Engineering, IEEE Transactions on 23.3 (2011): 373-387..

[13] Bernecker, Thomas, et al. "Scalable probabilistic similarity ranking in uncertain databases." Knowledge and Data Engineering, IEEE Transactions on 22.9 (2010): 1234-1246

[14] Bilenko, Mikhail, Beena Kamath, and Raymond J. Mooney."Adaptive blocking: Learning to scale up record linkage." Data Mining, 2006. ICDM'06. Sixth International Conference on. IEEE, 2006.