A REVIEW ON DATA NORMALIZATION OF DUPLICATE RECORDS FROM MULTIPLE SOURCES

NEKKANTI MOUNIKA Student, M.Tech (CSE), VIKAS GROUP OF INSTITUTIONS, A.P., India.

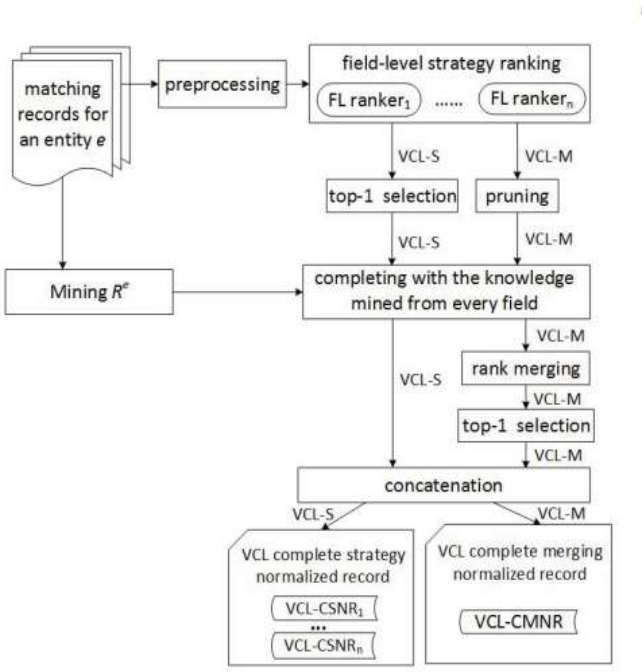
Mr. B.PHANI KRISHNA Assistant Professor, Dept. of Computer Science & Engineering, VIKAS GROUP OF INSTITUTIONS, A.P., India.

Abstract — In this paper, The bulk data is generated in the world wide web. Based on the user search parameter the data is collected from various sources. The usefulness of data increases when it is linked and fused with other data from numerous (Web) sources. The promise of Big Data hinges upon addressing several big data integration challenges, such as record linkage at scale, real-time data fusion, and integrating Deep Web. Although much work has been conducted on these problems, there is limited work on creating a uniform, standard record from a group of records corresponding to the same real-world entity. We refer to this task as record normalization. Such a record representation, coined normalized record, is important for both front-end and back-end applications. In this paper, we formalize the record normalization problem, present in-depth analysis of normalization granularity levels (e.g., record, field, and value-component) and of normalization forms (e.g., typical versus complete). We propose a comprehensive framework for computing the normalized record. The proposed framework includes a suit of record normalization methods, from naive ones, which use only the information gathered from records themselves, to complex strategies, which globally mine a group of duplicate records before selecting a value for an attribute of a normalized record.

INTRODUCTION

The relevant data collection is done from various warehouses like Google, Bing Shopping. Google Scholar is an important mining domain. Web data integration is an important component of many applications collecting data from Web databases, such as Web data warehousing (e.g., Google and Bing Shopping; Google Scholar), data aggregation (e.g., product and service reviews), and metasearching [3]. Integration systems at Web scale need to automatically match records from different sources that refer to the same real-world entity [4], [5], [6], find the true matching records among them and turn this set of records into a standard record for the consumption of users or other applications. There is a large body of work on the record

matching problem [7] and the truth discovery problem[8]. The record matching problem is also referred to as duplicate record detection [9], record linkage [10], object identification [11], entity resolution [12], or deduplication [13] and the truth discovery problem is also called as truth finding [14] or fact finding [15] a key problem in data fusion [16], [17]. In this paper, we assume that the tasks of record matching and truth discovery have been performed and that the groups of true matching records have thus been identified. Our goal is to generate a uniform, standard record for each group of true matching records for end-user consumption. We call the generated record the normalized record. We call the problem of computing the normalized record for a group of matching records the Record Normalization Problem (RNP), and it is the focus of this work. RNP is another specific interesting problem in data fusion Record normalization is important in many application domains. For example, in the research publication domain, although the integrator website, such as Citeseer or Google Scholar, contains records gathered from a variety of sources using automated extraction techniques, it must display a normalized record to users. Otherwise, it is unclear what can be presented to users: (i) present the entire group of matching records or (ii) simply present some random record from the group, to just name a couple of ad-hoc approaches. Either of these choices can lead to a frustrating experience for a user, because in (i) the user needs to sort/browse through a potentially large number of duplicate records, and in (ii) we run the risk of presenting a record with missing or incorrect pieces of data.



LITERATURE SURVEY

Accessing the web: From search to integration

AUTHORS: K. C.-C. Chang and J. Cho

We have witnessed the rapid growth of the Web-- It has not only "broadened" but also "deepened": While the "surface Web" has expanded from the 1999 estimate of 800 million to the recent 19.2 billion pages reported by Yahoo index, an equally or even more significant amount of information is hidden on the "deep Web," behind query forms, recently estimated at over 1.2 million, of online databases. Accessing the information on the Web thus requires not only search to locate pages of interests, from the surface Web, but also integration to aggregate data from alternative or complementary sources, from the deep Web.

Although the opportunities are unprecedented, the challenges are also immense: On the one hand, for the surface Web, while search seems to have evolved into a standard technology, its maturity and pervasiveness have also invited the attack of spam and the demand of personalization. On the other hand, for the deep Web, while the proliferation of structured sources has promised unlimited possibilities for more precise and aggregated access, it has also presented new challenges for realizing large scale and dynamic information integration. These issues are in essence related to data management, in a large scale, and thus present novel problems and interesting opportunities for our research community. This tutorial will discuss the new access scenarios and research problems in Web information access: from search of the surface Web to integration of the deep Web.

Query-time record linkage and fusion over web databases

AUTHORS: E. K. Rezig, E. C. Dragut, M. Ouzzani, and A. K. Elmagarmid

Data-intensive Web applications usually require integrating data from Web sources at query time. The sources may refer to the same real-world entity in different ways and some may even provide outdated or erroneous data. An important task is to recognize and merge the records that refer to the same real world entity at query time. Most existing duplicate detection and fusion techniques work in the off-line setting and do not meet the online constraint. There are at least two aspects that differentiate online duplicate detection and fusion from its offline counterpart. (i) The latter assumes that the entire data is available, while the former cannot make such an assumption. (ii) Several query submissions may be required to compute the "ideal" representation of an entity in the online setting. This paper presents a general framework for the online setting based on an iterative record-based caching technique. A set of frequently requested records is deduplicated off-line and cached for future reference. Newly arriving records in response to a query are deduplicated jointly with the records in the cache, presented to the user and appended to the cache. Experiments with real and synthetic data show the benefit of our solution over traditional record linkage techniques applied to an online setting.

PROPOSED METHOD

Integration systems at Web scale need to automatically match records from different sources that refer to the same real-world entity find the true matching records among them and turn this set of records into a standard record for the consumption of users or other applications. There is a large body of work on the record matching problem and the truth discovery problem. The record matching problem is also referred to as duplicate record detection, record linkage, object identification, entity resolution, or deduplication and the truth discovery problem is also called as truth finding or fact finding - a key problem in data fusion.

Drawbacks of existing framework:

- Truth Discovery Problem.
- Record Matching Problem.

PROPOSED SYSTEM:

- We propose three levels of granularities for record normalization along with methods to construct normalized records according to them.
- We propose a comprehensive framework for systematic construction of normalized records. Our framework is flexible and allows new strategies to be added with ease. To our knowledge, this is the first piece of work to propose such a detailed framework.
- We propose and compare a range of normalization strategies, from frequency, length, centroid and feature-based to more complex ones that utilize result merging models from information retrieval, such as (weighted) Borda.
- We introduce a number of heuristic rules to mine desirable value components from a field. We use them to construct the normalized value for the field.
- We perform empirical studies on publication records. The experimental results show that the proposed weighted-Borda-based approach significantly outperforms the baseline approaches.

Favorable circumstances of proposed framework:

- High Accuracy.
- Best Performance.
- We analyzed the record and field level normalization in the typical normalization.
- In the complete normalization, we focused on field values and proposed algorithms for acronym expansion and value component mining to produce much improved normalized field values.

IMPLEMENTATION

System Model

In this first module, we develop two entities: User and Secure-Cloud Service Provider.

User: The user is an entity that wants to outsource data storage to the S-CSP and access the data later. In a storage system supporting deduplication, the user only uploads unique data but does not upload any duplicate data to save the upload bandwidth. Furthermore, the fault tolerance is required by users in the system to provide higher reliability.

S-CSP: The S-CSP is an entity that provides the outsourcing data storage service for the users. In the deduplication system, when users own and store the same content, the S-CSP will only store a single copy of these files and retain only unique data. A deduplication technique, on the other hand, can reduce the storage cost at the server side and save the upload bandwidth at the user side. For fault tolerance and confidentiality of data storage, we consider a quorum of S-CSPs, each being an

independent entity. The user data is distributed across multiple S-CSPs.

Data Deduplication

Data Deduplication involves finding and removing of duplicate datas without considering its fidelity.

Here the goal is to store more datas with less bandwidth.

Files are uploaded to the CSP and only the Dataowners can view and download it.

The Security requirements is also achieved by Secret Sharing Scheme.

Secret Sharing Scheme uses two algorithms, share and recover.

Datas are uploaded both file and block level and the finding duplication is also in the same process.

This is made possible by finding duplicate chunks and maintaining a single copy of chunks.

File Level Deduplication Systems

To support efficient duplicate check, tags for each file will be computed and are sent to S-CSPs.

To upload a file *F*, the user interacts with S-CSPs to perform the deduplication.

More precisely, the user firstly computes and sends the file tag ϕF= TagGen(F) to S-CSPs for the file duplicate check.

If a duplicate is found the user computes and sends it to a server via a secure channel.

Otherwise if no duplicate is found the process continues, i.e secret sharing scheme runs and the user will upload a file to CSP.

To download a file the user will use the secret shares and download it from the SCSP's .

This approach provides fault tolerance and allows the user to remain accessible even if *any* limited subsets of storage servers fail.

Block Level Deduplication Systems

In this module we will show to achieve fine grained block-level distributed deduplication systems.

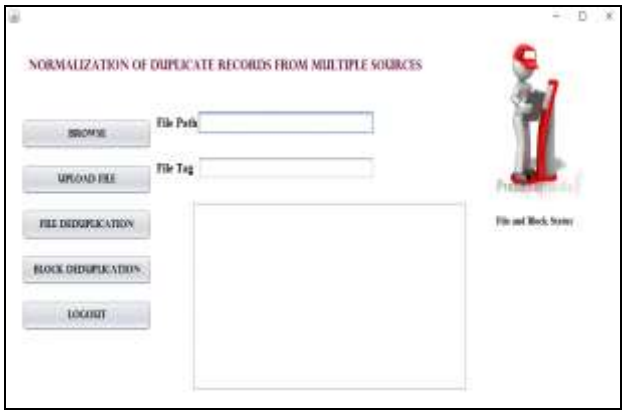
In a block-level deduplication system, the user also needs to firstly perform the file-level deduplication before uploading his file.

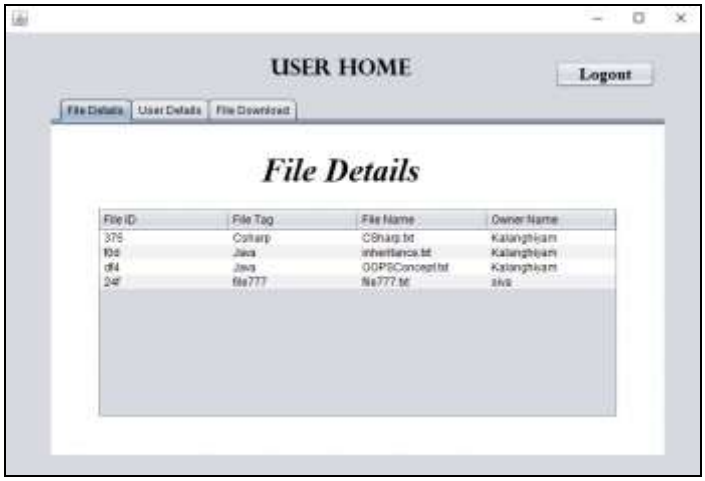
If no duplicate is found, the user divides this file into blocks and performs block-level deduplication.

The System setup is similar to the file level deduplication except the parameter changes.

To download a block the user gets the secret shares and download the blocks from CSP.

SAMPLE RESULTS





CONCLUSION

In this paper, we studied the problem of record normalization over a set of matching records that refer to the same real-world entity. We presented three levels of normalization granularities (record-level, field-level and value component level) and two forms of normalization (typical normalization and complete normalization). For each form of normalization, we proposed a computational framework that includes both single-strategy and multi-strategy approaches. We proposed four single-strategy approaches: frequency, length, centroid, and feature-based to select the normalized record or the normalized field value. For multi strategy approach, we used result merging models inspired from meta searching to combine the results from a number of single strategies. We analyzed the record and field level normalization in the typical normalization. In the complete normalization, we focused on field values and proposed algorithms for acronym expansion and value component mining to produce much improved normalized field values. We implemented a prototype and tested it on a real-world dataset. The experimental results demonstrate the feasibility and effectiveness of our approach.

REFERENCES

[1] K. C.-C. Chang and J. Cho, “Accessing the web: From search to integration,” in *SIGMOD*, 2006, pp. 804–805.

[2] M. J. Cafarella, A. Halevy, D. Z. Wang, E. Wu, and Y. Zhang, “Webtables: Exploring the power of tables on the web,” *PVLDB*, vol. 1, no. 1, pp. 538–549, 2008.

[3] W. Meng and C. Yu, *Advanced Metasearch Engine Technology*. Morgan & Claypool Publishers, 2010.

[4] A. Gruenheid, X. L. Dong, and D. Srivastava, “Incremental record linkage,” *PVLDB*, vol. 7, no. 9, pp. 697–708, May 2014.

[5] E. K. Rezig, E. C. Dragut, M. Ouzzani, and A. K. Elmagarmid, “Query-time record linkage and fusion over web databases,” in *ICDE*, 2015, pp. 42–53.

[6] W. Su, J. Wang, and F. Lochovsky, "Record matching over query results from multiple web databases," *TKDE*, vol. 22, no. 4, 2010.

[7] H. K¨opcke and E. Rahm, "Frameworks for entity matching: A comparison," *DKE*, vol. 69, no. 2, pp. 197–210, 2010.

[8] X. Yin, J. Han, and S. Y. Philip, "Truth discovery with multiple conflicting information providers on the web," *ICDE*, 2008.

[9] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate record detection: A survey," *TKDE*, vol. 19, no. 1, pp. 1–16, 2007.

[10] P. Christen, "A survey of indexing techniques for scalable record linkage and deduplication," *TKDE*, vol. 24, no. 9, 2012.

[11] S. Tejada, C. A. Knoblock, and S. Minton, "Learning object identification rules for information integration," *Inf. Sys.*, vol. 26, no. 8, pp. 607–633, 2001.

[12] L. Shu, A. Chen, M. Xiong, and W. Meng, "Efficient spectral neighborhood blocking for entity resolution," in *ICDE*, 2011.

[13] Y. Jiang, C. Lin, W. Meng, C. Yu, A. M. Cohen, and N. R. Smalheiser, "Rule-based deduplication of article records from bibliographic databases," *Database*, vol. 2014, 2014.