# Database enrichment environment to identify duplicate tuples

Juliano Augusto Carreira
DCCE – IBILCE – UNESP
São José do Rio Preto – SP
*julianocarreira@gmail.com*

Carlos Roberto Valêncio
DCCE – IBILCE – UNESP
São José do Rio Preto – SP
*valencio@ibilce.unesp.br*

Rogéria C. Gratão de Souza
DCCE – IBILCE – UNESP
São José do Rio Preto - SP
*rogeria@ibilce.unesp.br*

**One of the significant problems and inherent to current large databases is the incidence of duplicate tuples. This problem refers to the repetition of records that, in most cases, are represented differently in databases but refer to the same real world entity, which makes the task of identifying those tuples a hard work. Considering that each language has its peculiarities, it is believed that the use of text operations techniques from the area of Information Retrieval can enrich the content of the records for a specific language and thus maximize the amount of identified duplicate tuples and/or improve the confidence level of their classification in relation to current tools. The main contribution of this paper is to provide a language independent environment able to approximate the spelling of the records in a database and thus identify duplicate tuples more efficiently than the isolated application of traditional methods. In addition to only improve database quality this tool can also improve the process of Knowledge Discovery in Databases (KDD).**

*Data Cleansing, Information Retrieval, Duplicate Tuples, Knowledge Discovery in Databases.*

## 1. INTRODUCTION

Given the great technological developments in recent years, the existence of large volumes of data in many areas is something real today. Besides only being concerned with the procedures for storage and retrieval of such data, the data holders are concerned about something more, something that can help them to add value to their businesses and can also assist them in decision making. This need, still unknown to many people, can be provided by a process known as Knowledge Discovery in Databases (KDD) [6].

In the databases in use today, either by a bad design or failures in the process of feeding them, it is very common to find problems related to data inconsistency and also to database structure such as typing errors, values outside the range, null values, duplicate tuples and no primary key definition. In order to address these problems, the KDD process proposes a data cleansing phase within the pre-processing stage, in which the main goal is to ensure the integrity and consistency of data for future steps. One of the most important steps in the KDD process is called data mining [6]. Not performing the data cleansing step can compromise the reliability of the results obtained in the data mining stage and, therefore, undermine the organizations that rely on them for decision making.

The work proposed here operates in the data cleansing phase inside the KDD process and specifically addresses the activities of identifying non-identical duplicate tuples. Realizing that languages are different and each one has unique characteristics, it is believed that building an environment for identifying non-identical duplicate tuples where the factor language is favored, can maximize the amount of duplicates identified and/or increase the accuracy of the results that are now obtained with the use of present techniques. This environment intends to use text operations techniques from the area of Information Retrieval such as thesaurus, spell checkers, stemming, stopwords removal, and others, in order to improve the content of records and approximate them orthographically to further the identification of duplicate records by means of traditional techniques used to identify duplicate tuples.

## 2. RELATED WORK

In recent years, a large number of tools have been proposed to operate in the data cleansing phase. Among them, there are more extensive tools that cover a lot of problems and there are fully specialized tools which handle specific information

such as emails, addresses, names, among others. Given the presented related works regarding non-identical duplicate tuples identification, it can be noted that none of the existing solutions consider the particular characteristics of any language. This is somewhat detrimental to the outcome because each language has features and specific problems that, if improved, can maximize the amount of identified duplicate tuples and/or increase the reliability level of the process.

Febrl [1] is a data cleansing open-source tool developed in Python and basically consists of two main components: the first to deal with data standardization using Hidden Markov Models and the second to address the duplicate tuples identification through a variety of string similarity techniques.

In TAILOR [3], an interactive tool for duplicate tuples identification is proposed In addition to implementing the state of art related to ad-hoc techniques for duplicates identification, TAILOR proposes an approach based on machine learning, specifically three techniques: induction model, cluster model and hybrid model.

WHIRL [2] is a free duplicate tuples identification system for research purposes. It uses strategies from the area of Information Retrieval, more specifically the Cosine similarity algorithm combined with the tf.idf weight schema to identify similarity between two fields in a table.

WinPure Clean & Match 2010 [5] is a proprietary tool that has the mission of cleansing lists of data and also makes the identification of duplicate tuples. The module responsible for the duplicate tuples elimination works in three stages: the first involves identifying the desired tables and columns, the second is to identify which technique should be used in the process; the third involves choosing how duplicate data must be displayed to the user.

## 3. ENRICHMENT ENVIRONMENT

The proposed environment consists basically of two modules: one for duplicate tuples identification and another to perform the enrichment of databases built in the Portuguese language. Although Portuguese is the language chosen for the beginning of the work, the tool is ready to accommodate any other language. Enrichment previously mentioned is through the application of techniques related to the area of Information Retrieval and also by particular adaptations to these techniques in order to obtain better results. The technical specifications required by the enrichment module involve the following concepts [4]:

- **Lexical Analysis:** performs the conversion of a character stream into a stream of words.
- **Stopword Elimination:** eliminates words not considered useful to the meaning of the records in a database.
- **Stemming:** removal of affixes (prefixes + suffixes) from words in order to find their root.
- **Thesaurus:** involves creating a pre-defined synonyms list of words that can be substituted in the records of a database.
- **Spelling standardization:** elimination of common misspellings. In Portuguese, for example, it is very common to find the following exchanges of graphemes: *s* and *ss*, *r* and *rr*, *x* and *ch*, *s* and *ç*, *j* and *g*.
- **Spell checkers:** endeavor to eliminate errors not solved by the function "Spelling Standardization" cited above.

## 4. CONCLUSION

In general, the tool is already functional and presents very interesting results regarding the spelling approximation of records in a database. It's expected to get the feature "spell checker" implemented in order to start formal tests against a real database containing approximately 50,000 records.

## 5. REFERENCES

[1] CHRISTEN, P; CHURCHES, T; HEGLAND, M. (2004) A Parallel Open Source Data Linkage System. In *Proceedings of The 8th Pacific-Asia Conference On Knowledge Discovery And Data Mining (PAKDD '04)*, Sydney, 26-28 May. Springer.
[2] COHEN, W. W.. (2000) Data Integration Using Similarity Joins and a Word-Based Information Representation Language. *ACM Transactions on Information Systems*, p. 288-321.
[3] ELFEKY, M; VERYKIOS, V; ELMAGARMID, A. (2002) TAILOR: A Record Linkage Toolbox. In *Proceedings of the 18th International Conference on Data Engineering (ICDE '02)*, p. 17-28.
[4] MANNING, C. D.; RAGHAVAN, P; SCHÜTZE, H. (2008) *Introduction to Information Retrieval*. Cambridge University Press.
[5] PIATETSKY-SHAPIRO, G. (2010) KDnuggets: Data Mining Community's Top Resource. http://www.kdnuggets.com (25 October 2010).
[6] ZHU, X; DAVIDSON, I. (2007) *Knowledge Discovery and Data Mining: Challenges and Realities*. IGI Global, New York.