Digital Healthcare Empowering Europeans
R. Cornet et al. (Eds.)

doi:10.3233/978-1-61499-512-8-227

A new approach for cleansing geographical dataset using Levenshtein distance, prior knowledge and contextual information

Adrien UGON[a,1], Thomas NICOLAS[a], Marion RICHARD[a] , Patrick GUERIN[b], Pascal CHANSARD[b], Christophe DEMOOR[b] and Laurent TOUBIANA[a, b]
[a] *INSERM, U1142, LIMICS, F-75006, Paris, France;*
Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1142, LIMICS, F-75006, Paris, France;
Université Paris 13, Sorbonne Paris Cité, LIMICS, (UMR_S 1142), F-93430, Villetaneuse, France.
[b] *IRSAN, 37 rue des Mathurins, F-75008, Paris, France*

Abstract. Epidemiological studies are necessary to take public health decisions. Their relevance depends on the quality of data. Doctors in continuous care collect a big amount of data that can be used for epidemiological purpose, but spatial data may be dirty; based on city names, the localization is imprecise, even more if it is misspelled. The only way to identify a city without ambiguity is to use its identifier, which can be retrieved by cleansing geographical textual data. In France, cities are organized in administrative zones called departments and some city names are shared by several cities in several departments. The clear identification of the department and the city name allows to deduce the city unique identifier and to make some spatial analysis such as epidemiological studies. In this paper, we propose a method to cleanse such data, using several steps. After having standardized the text to cleanse, we use the Levenshtein distance to generate a first set of propositions. Finally, the propositions are filtered, by removing the less likely candidates, so that it remains only one, which becomes the chosen city. Tested on a dataset of 9818 entries, we obtained 89.1% of concordance, whereas the standard Levenshtein distance obtained 70.5%. This demonstrates that our method has better results.

Keywords. Levenshtein distance, geographical database, context-based analysis

Introduction

IRSAN is a result of the partnership between INSERM (French National Institute of Health and Medical Research) and SOS-Médecins-France (SMF), the main private network of associations for out-of-hours healthcare covering 60% of French population. Since 2006, an average of 7,000 interventions are aggregated at national level every day.

Efficient data mining requires a good quality of data. In the context of epidemiological studies, one of required fields is the location. When given as free text, a satisfying identifier needs to be retrieved. The difficulty comes from different reasons:

[1] Corresponding author adrien.ugon@inserm.fr

(a) wrong spelling (b) truncated city name (c) use of abbreviations (d) ambiguity due to homonyms (e) presence of noisy additional information (f) entry of a district name and not a city. A bad treatment of this data may lead to misunderstandings and wrong results.

The INSEE (National Institute of Statistics and Economic Studies) defines a "geographical code" for each city, which is different from the "zip code" that can be shared by several neighboring cities. Both include the identifier of the "department".

To solve this problem, we propose a knowledge-based system using Levenshtein distance to cleanse data using contextual information. The process of cleansing data is composed of 3 steps: (a) standardization of each entry (b) determination of the area of work of each association (c) determination of the geographical code for each entry.

1. Methods

1.1. Description of data

Our dataset is composed of every different pair of values for 2 fields from original database: the identifier of the called association and a free text field containing the name of the city where the patient is from. It contains 10126 records.

All were expertized by 3 different people. For each entry, the "geographical codes" and the name of all cities compatible with the entry were asked to be given. In 177 cases, it was impossible to decide any real city name; In 131 cases, the city was located abroad.

1.2. First step: standardization of the city name entered by the doctor

Standardization consists in (a) capitalizing (b) removing accent from characters (c) substituting common abbreviations (d) removing noisy substrings.

Steps (c) and (d) use regular expressions to make sense.

1.3. Second step: automatic learning of the departments associated to an association

When specified, the zip code allows easily to deduce the department associated to a city. Otherwise, a perfect match is searched with all names of cities in France. In case of a single perfect match, we assume that it is the right city. The department is then deduced.

We dispose also now from the count of cities directly detected as belonging to it. This is an estimation of the likelihood of each department.

1.4. Third step: cleansing of city names using Levenshtein distance

The Levenshtein distance, also known as edit distance, is used to measure the lexicographic distance between two strings(1). It is defined as the minimal number of edits operations to transform one string to the other. We consider as edit operations in this algorithm insertions, deletions and replacements with a penalty cost of 1.

The decision needs 2 steps: (a) the standardized entered city name is compared using the Levenshtein distance to all city names of the departments found previously.

This gives a list of possible cities, all at the minimal distance of the entry. (b) we choose the most likely city name by using (a) the Levenshtein distance to not truncated previously selected city names, (b) the similarity between first letters of the selected city names and the standardized entry; it is admitted that first letters are more significant (2) (c) the likelihood of each department found in previous steps, (d) arbitrarily, the first city name.

2. Results

We tested our method on our database composed of 9818 remaining decidable cities. On the complete dataset, we obtained 89.1% of concordance with experts, which is a significant improvement of the reference simple "Levenshtein" method (70.5%).

Sources of disagreement with experts are: a) Use of the name of a hamlet or a sub-part of the city b) the right answer was found in the first steps but then filtered c) Cities located in departments not identifying as belonging to working area of an association

3. Discussion

String to string correction of erroneous texts are generally solved by edit distances like the Levenshtein distance (1) or by the Longest Common Subsequence algorithm (3). Some of them also use context information (4,5).

Most works of city names cleansing just try to find city names with high similarity, and do not try to identify clearly a city. Homonyms can thus not be distinguished.

4. Conclusion

Epidemiological studies need proper data. To cleanse geographical fields, we propose a method based on the Levenshtein distance, using prior knowledge and contextual information. Tested on a dataset of 9818 cities, concordance with experts was 89.1%.

For future work, we will add post-processing when the edit-distance is too high.

References

1. Levenshtein V. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Sov Phys Dokl. 1966;10:707‑10.
2. Lim S. Cleansing Noisy City Names in Spatial Data Mining. 2010 International Conference on Information Science and Applications (ICISA). 2010. p. 1‑8.
3. Goodrich MT, Tamassia R. Algorithm Design: Foundations, Analysis and Internet Examples. 2nd éd. New York, NY, USA: John Wiley & Sons, Inc.; 2009.
4. Ruch P, Baud R, Geissbühler A. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. Artif Intell Med. oct 2003;29(1-2):169‑84.
5. Siklósi B, Novák A, Prószéky G. Context-aware correction of spelling errors in Hungarian medical documents. Comput Speech Lang.