The challenge to incorporate the Socials Sciences with Computer Science

Pilar Vanessa Hidalgo León
Universidad Andina del Cusco
phidalgo@uandina.edu.pe

Abstract

The research shows a collection of the tools that today are complemented the qualitative investigations in the Social Sciences through the Sciences of the computer, in addition, it shows some examples and theories used in processing qualitative information and to transform the quantitative studies. The construction of a corpus of approximately 300 records and extraction of entities through the tools under evaluation (using Polyglot, NER and Sckecth). These entities defined in English by Actors, Sentiments, motivations, etc. As a result they have the rules of association of the actors with the most frequent sentiments and the topics they discuss. In conclusion, it is possible to articulate in a multidisciplinary way the qualitative studies with the Computer Science to give greater reliability and interpretation of the results.

1 Introduction

In order to use the advantages of informatics for the social sciences, the large amount of information (more than 300 testimonies in natural language) makes the task of analysis and interpretation difficult. Concept Analysis and emotion-based analysis of feelings will be possible using association rules with traditional algorithms such as Apriori and FPGrowth to find relationships between the places mentioned by the subjects mentioned by the people who gave their testimony about experiences, motivations, Feelings during immigration.

2 Methods and materials

Three corpus are builded, each one belongs each testimonial in english and French in Natural Language extracted from letters and testimonials: Lucie Bacon, 2015 (4 testimonies in French and 18 in English), "Variations sur l ',"" one of the three testimonies in English and French in Natural " migreire épisode of Kreuz. Car-tographies mises en scène ", Mediapart; Cyril Roussel, 2013 (43 testimonies in French), "Les migrations des membres du Komala entre l'Iran et le Kurdistan d'Irak", Atlas du Kurdi-stan d'Irak; Nelly Robin, 2013 (99 in french), «Mineurs en mobilité entre l'Afrique sub-saharienne et l'UE», Program Migrinter - Unesco.

In the first stage of the research the thesaurus of terms in French was constructed and the main concepts were written in HTML format: 3 levels (eg Actors -> Authorities -> Police Authorities), 5 main concepts: Actors (famille, amis, Organizations, etc.), Resources (food and lodging, transport), Controls (military campamento) (War, opportunities) and Feelings (problème, rencontre, très). The extraction of space entities was done in the French and English corpus, obtaining: 3 levels (Continent -> Countries -> Cities), 3 type locations (spatial entities, toponomies, adverbs of place). In terms of space entities, it was found: in English: 87 Attributes, 2 Continents, 28Countries, 57 cities and French: 86 Attributes, 2 Continents E: 21 Countries, 63 cities. Various tools were used for the ex-traction: Polyglot, Unitex, SketchEngine, as well as several data or geodetic repositories of Geonames, TreeTagger and NER. The extraction of countries and adverbs of place with better results was through Unitex. Unfortunately the French-language dictionary is too large for UNITEX, Polyglot and TreeTagger were used for extracting spatial entities as shown in Table 1.

That's <ADV>near</ADV> to the Kunar <Topo>Province</Topo>	UNITEX-Dictionary by own
Iran;Kabul;Serbia;Sofia;Istanbul;Dimitrograd;Kunar	Polyglot-TreeTagger

Table 1: Results from Testimonial 0: **** *recit_01 *sex_Masculin *pays_Afghanistan *celibataire *prenom_Ahmad *terrain1

In each case four expedients containing the space entities mentioned in the testimonies were obtained. To apply the association rule algorithms, the data sets were constructed in a format acceptable to Weka, the CSV format and the headers must match the established thesaurus (entities thesaurus and spatial concepts). The experiments were performed in the sample Mineurs (99 instances). The characteristics of the Mineurs data set in French are as follows:

- Title: MineursFrench
- Format: Horizontal
- Number of instances: 99
- Number of Attributes in French: 232 main concepts (nominal 3 levels) and 69 spatial entities (nominal level 1).

3 Results

By doing the emotion-based analysis of feelings, there are 38 (feeling) selected attributes of 232 attributes of the data set are taken. According to Shapiro (1991), for an association rule to be considered interesting, it must fulfill the following principle: *Confidence <Lift*. Under this principle, the Lift measure was changed to 1. The experiment was performed comparing the level 2 and 3 attributes in the case of spatial entities; (Ei, difficulté), Motivations (ei, voyage) and members of the family (ei, père). The rules of association that comply with the principle, which are more interesting in the sample of minors, are shown below:

Main Concepts	**Spatial Entities**
Motivations	Cities
Gao=si avenir=si => Bamako=si projet=si Gao=si projet=si => Maghnia=si avenir=si	

Table 2: Association rules in Minors ,(99 instans)

These sets of common articles (Table 2) could be used to generate sets of common concepts related to a generalized spatial entity. However, Apriori in a small set of data performs small procedures, and in a long sample or sets of frequent itemset varies the MinSupport will be larger than the sets that will be obtained, making it less interesting.

4 Conclusions

It is frequent that during the analysis and syntactic of the messages, the dictionary changes some words and removes of context the message. The manual review should unfortunately be included before performing data processing and algorithm application.

The dictionaries or gazetteer of localities depend on the ambiguity, since there are many names of localities in distant places of the world that have the same name.

In this paper, we compare and investigate some techniques that make possible the extraction of a large amount of data on the web that can support the qualitative studies in the social sciences.

One of the objectives achieved is the preprocessing of a dataset managed by a data mining tool. The analysis of the concepts was done by analyzing them through association rules and finding hidden relationships between them and spatial entities and through measures that provided interesting rules shown in the results of the QoDSS project.

Acknowledgements

This article was written thanks to the sup-port of the TETIS Group (Télédétection Irstea-Cirad-AgroParisTech) and the QDoSSi project, Quali-té des Données multi-Sources.

References

Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In Acm sigmod record (Vol. 22, No. 2, pp. 207-216). ACM.

Zaki, M. J., Meira Jr, W., & Meira, W. (2014). Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press.

Mishra, A. K., Pani, S. K., & Ratha, B. K. Association rule mining with Apriori and FPGrowth using Weka.

Elgaml, E. M., Ibrahim, D. M., & Sallam, E. A. (2015). Improved FP-growth Algorithm with Multiple Minimum Supports Using Maximum Constraints. Analysis, 6160, 10001351.

Mehay, A., Singh, D. K., & Sharma, D. N. (2013). AnalyzeMarket Basket Data using FP-growth and Apriori Algorithm international Journal on Recent and Innovation Trends in Computing and Communication, 16, 693-696.