How to deal with heterogeneous data?

Mathieu Roche
UMR TETIS (Cirad, Irstea, AgroParisTech) – France
mathieu.roche@cirad.fr
LIRMM (CNRS, University of Montpellier) – France
mathieu.roche@lirmm.fr
Web: www.textmining.biz

1 Introduction

The Big Data issue is traditionally characterized in terms of 3 V, i.e. volume, variety, and velocity. This paper focuses on the variety criterion, which is a challenging issue.

Data heterogeneity and content

In the context of Web 2.0, text content is often heterogenous (i.e. lexical heterogeneity). For instance, some words may be shortened or lengthened with the use of specific graphics (e.g. emoticons) or hashtags. Specific processing is necessary in this context. For instance, with an opinion classification task based on the message `SimBig is an aaaaattractive conference!`, the results are generally improved by removing repeated characters (i.e. `a`). But information on the sentiment intensity identified by the character elongation is lost with this normalization. This example highlights the difficulty of dealing with heterogeneous textual data content.

The following sub-section describes the heterogeneity according the document types (e.g. images and texts).

Heterogeneity and document types

Impressive amounts of high spatial resolution satellite data are currently available. This raises the issue of fast and effective satellite image analysis as costly human involvement is still required. Meanwhile, large amounts of textual data are available via the Web and many research communities are interested in the issue of knowledge extraction, including spatial information. In this context, image-text matching improves information retrieval and image annotation techniques (Forestier et al., 2012). This provides users with a more global data context that may be useful for experts involved in land-use planning (Alatrista Salas et al., 2014).

2 Text-mining method for matching heterogenous data

A generic approach to address the heterogeneity issue consists of extracting relevant features in documents. In our work, we focus on 3 types of features: thematic, spatial, and temporal features. These are extracted in textual documents using natural language processing (NLP) techniques based on linguistic and statistic information (Manning and Schütze, 1999):

- The extraction of **thematic information** is based on the recognition of relevant terms in texts. For instance, terminology extraction techniques enable extraction of single-word terms (e.g. `irrigation`) or phrases (e.g. `rice crops`). The most efficient state-of-the-art term recognition systems are based on both statistical and linguistic information (Lossio-Ventura et al., 2015).

- Extracting **spatial information** from documents is still challenging. In our work, we use patterns to detect these specific named entities. Moreover, a hybrid method enables disambiguation of spatial entities and organizations. This method combines symbolic approaches (i.e. patterns) and machine learning techniques (Tahrat et al., 2013).

- In order to extract **temporal expressions** in texts, we use rule-based systems like HeidelTime (Strötgen and Gertz, 2010). This multilingual system extracts temporal expressions from documents and normalizes them. HeidelTime applies different normalization strategies depending on the text types, e.g. news, narrative, or scientific documents.

These different methods are partially used in the projects summarized in the following section. More precisely, Section 3 presents two projects that investigate heterogeneous data in agricultural the domain.[1]

3 Applications in the agricultural domain

3.1 Animal disease surveillance

New and emerging infectious diseases are an increasing threat to countries. Many of these diseases are related to globalization, travel and international trade. Disease outbreaks are conventionally reported through an organized multilevel health infrastructure, which can lead to delays from the time cases are first detected, their laboratory confirmation and finally public communication. In collaboration with the CMAEE[2] lab, our project proposes a new method in the epidemic intelligence domain that is designed to discover knowledge in heterogenous web documents dealing with animal disease outbreaks. The proposed method consists of four stages: data acquisition, information retrieval (i.e. identification of relevant documents), information extraction (i.e. extraction of symptoms, locations, dates, diseases, affected animals, etc.), and evaluation by different epidemiology experts (Arsevska et al., 2014).

3.2 Information extraction from experimental data

Our joint work with the IATE[3] lab and AgroParisTech[4] deals with knowledge engineering issues regarding the extraction of experimental data from scientific papers to be subsequently reused in decision support systems. Experimental data can be represented by n-ary relations, which link a studied topic (e.g. food packaging, transformation process) with its features (e.g. oxygen permeability in packaging, biomass grinding). This knowledge is capitalized in an ontological and terminological resource (OTR). Part of this work consists of recognizing specialized terms (e.g. units of measures) that have many lexical variations in scientific documents in order to enrich an OTR (Berrahou et al., 2013).

4 Conclusion

Heterogenous data processing enables us to address several text-mining issues. Note that we integrated the knowledge of experts in the core of research applications summarized in Section 3. In future work, we plan to investigate other techniques dealing with heterogeneous data, such as visual analytics approaches (Keim et al., 2008).

References

H. Alatrista Salas, E. Kergosien, M. Roche, and M. Teisseire. 2014. ANIMITEX project: Image analysis based on textual information. In *Proc. of Symposium on Information Management and Big Data (SimBig), Vol-1318, CEUR*, pages 49–52.

E. Arsevska, M. Roche, R. Lancelot, P. Hendrikx, and B. Dufour. 2014. Exploiting textual source information for epidemio-surveillance. In *Proc. of Metadata and Semantics Research - 8th Research Conference (MTSR) - Communications in Computer and Information Science, Volume 478*, pages 359–361.

S.L. Berrahou, P. Buche, J. Dibie-Barthelemy, and M. Roche. 2013. How to extract unit of measure in scientific documents? In *Proc. of International Conference on Knowledge Discovery and Information Retrieval (KDIR), Text Mining Session*, pages 249–256.

G. Forestier, A. Puissant, C. Wemmert, and P. Gançarski. 2012. Knowledge-based region labeling for remote sensing image interpretation. *Computers, Environment and Urban Systems*, 36(5):470–480.

D.A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler. 2008. Visual analytics: Scope and challenges. In *Visual Data Mining*, pages 76–90. Springer-Verlag.

J.A. Lossio-Ventura, C. Jonquet, M. Roche, and M. Teisseire. 2015. Biomedical term extraction: overview and a new methodology. *Information Retrieval Journal (IRJ) - special issue "Medical Information Retrieval", to appear.*

C.D. Manning and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

J. Strötgen and M. Gertz. 2010. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proc. of International Workshop on Semantic Evaluation*, pages 321–324.

S. Tahrat, E. Kergosien, S. Bringay, M. Roche, and M. Teisseire. 2013. Text2geo: from textual data to geospatial information. In *Proc. of International Conference on Web Intelligence, Mining and Semantics (WIMS).*

[1] http://www.textmining.biz/agroNLP.html

[2] Joint research unit (JRU) regarding the control of exotic and emerging animal diseases – http://umr-cmaee.cirad.fr

[3] JRU in the area of agro-polymers and emerging technologies – http://umr-iate.cirad.fr

[4] http://www.agroparistech.fr