Expert Systems with Applications 38 (2011) 8835–8849

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A clustering study of a 7000 EU document inventory using MDS and SOM [☆]

Patrick A. De Mazière [*], Marc M. Van Hulle

Laboratorium voor Neuro-en Psychofysiologie, K.U.Leuven, Leuven, Belgium

ARTICLE INFO

Keywords:
Text mining
Data warehousing
High dimensional data
Visualisation methods
High performance computing (HPC)
Parallel algorithms

ABSTRACT

In this article, we discuss a number of methods and tools to cluster a 7000 document inventory in order to evaluate the impact of EU funded research in social sciences and humanities on EU policies. The inventory, which is not publicly available, but provided to us by the European Union (EU) in the framework of an EU project, could be divided into three main categories: research documents, influential policy documents, and policy documents. To represent the results in a way that non-experts could make use of it, we explored and compared two visualisation techniques, multi-dimensional scaling (MDS) and the self-organising map (SOM), and one of the latter's derivatives, the U-matrix. Contrary to most other approaches, which perform text analyses only on document titles and abstracts, we performed a full text analysis on more than 300,000 pages in total. Due to the inability of many software suites to handle text mining problems of this size, we developed our own analysis platform. We show that the combination of a U-matrix and an MDS map, which is rarely performed in the domain of text mining, reveals information that would go unnoticed otherwise. Furthermore, we show that the combination of a database, to store the data and the (intermediate) results, and a webserver, to visualise the results, offers a powerful platform to analyse the data and share the results with all participants/collaborators involved in a data- and computation intensive EU-project, thereby guaranteeing both data- and result consistency.

1. Introduction

Text mining is a quite mature research field (see *e.g.* Erhardt, Schneider, & Blaschke, 2006; Yang, Akers, Klose, & Yang, 2008 and references therein) with many successful applications in a variety of domains such as biomedicine (Collier & Takeuchi, 2004; Duch, Matykiewicz, & Pestian, 2008; Erhardt et al., 2006; Lourenço et al., 2009), document classification (Isa, Kallimani, & Lee, 2009; SanJuan & Ibekwe-SanJuan, 2006) and document generation (Yang, 2009; Yang & Lee, 2005), ontology mapping (Tsoi, Patel, Zhao, & Zheng, 2009), computer intrusion/fraud detection (Adeva & Atxa, 2007; Holton, 2009), etc. Text mining by itself is a multidisciplinary field, and it involves many subtasks such as semantic analysis, clustering, categorisation, etc.

Almost all text mining analyses can be split into three major stages: a data preprocessing and/or formatting stage, an extensive information extraction-, transformation-, and selection stage, and, finally, the actual analysis stage and accordingly the visualisation. In the preprocessing stage, the documents are formatted in such a way that they become *computer interpretable*: binary and non-binary formats like Microsoft Word or Hypertext Markup language (HTML) are converted into flat text (ASCII), with or without preservation of the document structure.

In the second stage, one extracts information from the documents, reduces the existing redundancy, and finally converts the document in a numerical usable format (Manning & Shütze, 1999; Salton & McGill, 1983). This second stage is usually concerned with the semantic/lexical analysis methods to discern informative words (or word groups) from non-informative ones, *e.g.*, part-of-speech taggers (POS, Toutanova & Manning, 2000; Toutanova, Klein, Manning, & Singer, 2003), natural language parsers (NLP, Klein & Manning, 2002), or named-entity-recognition (NER, Finkel, Grenager, & Manning, 2005) methods. POS methods name every word according to its type: nouns, verbs, plural proper nouns, etc., while NER methods concentrate on sequences of words in a text which are the names of things, persons, institutions, company names, etc. Natural language parsers group (sequences of) words together that act e.g., as subject, adverb, or direct object

[☆] This work is performed in the framework of the SSH project, which is funded by DG Budget Framework Service Contract BUDG06/PO/01/Lot3 (EU project officer: Dr. Nikolaos Kastrinos; project coordinator: Dr. Geert Steurs; project partners: IDEA Consult, Ecorys Nederland BV, NIFU STEP, and Akilian NV). We thank Dr. Vincent Duchêne, Dr. Geert Steurs, and Nathalie Pasquier of IDEA Consult, Brussels and Dr. Nikolaos Kastrinos for their valuable contributions to this article. Furthermore, we thank Prof. Ron Wehrens (Radboud University Nijmegen) for his help to use his kohonen R- package as optimal as possible, and the K.U.Leuven for the use of their High Performance Computing infrastructure during many weeks, and Wim Obbels (K.U.Leuven, ICTS/HPC) for the given support in establishing the required communication channels.

[*] Corresponding author.

E-mail addresses: patrick.demaziere@med.kuleuven.be (P.A. De Mazière), marc.vanhulle@med.kuleuven.be (M.M. Van Hulle).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.01.094

clause in the given sentence. Sometimes, different extraction methods are combined within the same tool, and many of them are part of larger machine learning and/or text processing suites like WEKA (Witten & Frank, 2005), GATE (Cunningham, Maynard, Bontcheva, & Tablan, 2002) or Rapid-I (formerly known as YALE).[1] For an overview of (commercial) text mining suites we refer to the literature, e.g., Yang et al. (2008). From the semantically/lexically analysed sentences, one selects usually only the most informative types of words, e.g., only the extracted nouns (Yang & Lee, 2005), and converts them into a numerical usable format by applying the vector space model (Salton & McGill, 1983). This method encodes documents as vectors, in which each vector component corresponds to a different term.[2] The vector components have as values the frequencies the corresponding terms occur in a given document.

Finally, one eliminates words that have the same meaning but a different spelling by applying techniques like, e.g., stemmers (Lovins, 1968; Porter, 1980) that reduce a conjugated verb to its stem, case simplifiers that make words in upper, lower or mixed case appear in the same case, synonym lists or co-word/co-occurrence analysis techniques (Erhardt et al., 2006), or metrics based on the inverse document frequency (IDF). The IDF expresses how many documents contain a given term. At the end of the second stage, one retains a collection of *document vectors* that are maximally informative with a minimum amount of redundancy.

In this article, we report on an EU project, the aim of which was to analyse, classify, and visualise EU funded research in social sciences and humanities in EU framework programmes FP5 and FP6. This project, called the SSH project for short, was aimed at the evaluation of the contributions of research to the development of EU policies (see Section 2). In this article, we focus on discourse links, which group documents together when they are semantically similar. Consequently, they cover the same topic/domain and/or refer to the same group of other items (documents). For the visualisation of the document clustering, we focus on the self-organising maps (SOM), a very commonly applied method, also in text mining (Isa et al., 2009; Vesanto, 1999; Yang, 2009; Yang & Lee, 2005), and on the multi-dimensional scaling (MDS) technique, which is rather rarely applied for text mining (Chen, Tzeng, & Ding, 2008; Duch et al., 2008), but offers great advantages when dealing with huge data sets.

Since the total document inventory consists of 7038 unstructured documents, all together over 300,000 pages, we had to be very selective within the pool of the existing methods and software-suites to be able to obtain results within an acceptable timespan. Due to the size of the inventory, we had to perform any linkage analysis in an unsupervised way: we have category-labels for each document of the inventory (see Fig. 1), but these refer to the origin of the document rather than its contents, albeit that both are related. Therefore, supervised methods like naive Bayesian classifiers (see e.g., Isa et al., 2009) or exploratory techniques that rely on the document structure, are not applicable. In addition, one of the most frequently occurring problems with the mentioned (commercial) software-suites are their limited processing capabilities: they can handle only a limited number of documents and/or pages or do perform weakly. The size of our inventory makes also that some interesting methods like co-word/co-occurrence analyses (Erhardt et al., 2006) are too time- and memory consuming. Consequently, we opted to develop our own analysis platform. This platform is described in more detail in Appendix B.

This article is structured as follows: we first briefly discuss the data set, and then the methodology and how it is applied: data preprocessing, information retrieval and extraction, dimensionality reduction of the document vector space, and linkage analysis and visualisation of the results. Next, we present and discuss the obtained results. Finally, we end this article with a conclusion and some perspectives for further research.

2. The document inventory

We start with a 7038 document inventory of ±6 GiB gathered by IDEA Consult, Brussels with the aid of the EU, in the framework of this EU-funded SSH-project. It contains documents divided across three main categories (Fig. 1): research documents (54% of all documents), influential policy documents (28%), and EU policy documents (18%). The EU policy documents are predominantly 'Communications' issued by the European Commission; EU influential policy documents are analytical, policy-supporting documents produced within or outside the EU institutions; the research documents are delivered by the research consortia that were engaged in EU-funded research and summarise their research contributions to these projects. The latter category contains, for example, the scientific reports that need to be delivered at the end of an EU-funded project. In our case, all documents are within the social sciences and humanities domain. We refer to Deliverable 1 (2009) of this project for more details about the document inventory.

Albeit that we determined for each document the category to which it belongs, this was not the final goal of this SSH project. Instead, the EU wanted to know whether there was a convergence between the three main document categories or between any of its subcategories. The overall objective was to evaluate the contributions of research supported through the socio-economic key action of FP5, and the priorities 7 and 8 of FP6 to the development of EU policies. More in particular, it was aimed at:

- identifying the positive influences from the SSH research programme to European policy in a selected number of policy domains;
- comparing these influences with the stated objectives and ex-ante expectations;
- drawing lessons on how to improve the positive influences of SSH research on European policies in these specific policy domains.

All documents of the inventory are stored in a database server as depicted in Fig. 2. This database server contains also all intermediate and final results as calculated by the workstation or the K.U.Leuven High Performance Computing (HPC) infrastructure. Moreover, to speed up the visualisation of the results, which are computation intensive as well, given the size of the document inventory, we decided to store the resulting images in the database server too. Consequently, all data *and* results can be consulted, generated, or viewed in one place, the database server, making that both the data *and* results are kept consistent to all processing units and/or users. This is an important advantage with respect to other projects where data is moved around by e-mail, ftp, or other, which is a possible cause to data inconsistencies or losses.

3. Methodology: data conversion, analysis and visualisation of the results

We consider three major stages when performing a document clustering: a raw preprocessing stage to make documents computer readable, an (extensive) information extraction, transformation, and selection stage, and finally the actual linkage analysis, which leads to the easily interpretable (graphical) results. An

[1] Table 3 in Appendix A gives the websites with information, source and/or binary code of the (Linux) programs mentioned throughout this manuscript.

[2] It should be mentioned here that we will use '"word" and '"term" interchangeably wherever appropriate.