AN EVALUATION PROTOCOL FOR TEXT MINING TOOLS : ALCESTE, SAS TEXT MINER, SPAD-CRM AND TEMIS TEXT MINING SOLUTIONS TESTING

Yasmina Quatrain, Sylvaine Nugier, Anne Peradotto
ELECTRICITE DE FRANCE Research & Development
1, avenue du Général de Gaulle 92141 Clamart Cedex – France
{yasmina.quatrain, sylvaine.nugier, anne.peradotto}@edf.fr

Extract

Within the context of the opening of the electricity market, EDF needs to be able to analyse large volumes of text data to enable the company to have a better knowledge of its customers. With this in mind, several text mining tools intended for analysing this very diverse information in large quantities have been evaluated using three different corpora. It appeared essential to create a table to enable easy comparison of the software. Inspired by existing expertise in data mining tools, this was carried out while being careful not to favour statistical over linguistic results. This table has ten subjects varying from the editing company to the fields of application passing through data access and lexical table analysis. In addition to the carrying out of the evaluation and its results on four market tools, this article retraces the method for creating the test table, the choice of the tools evaluated and the criteria retained. Moreover, this experience supports the use of a detailed protocol permitting indispensable functions to be identified and evaluated according to the objectives and the profile of the software user and the nature of the corpus to be analysed.

INTRODUCTION

The growing volume of text data coming from the Internet and customer contacts (by mails, the transcription of telephone messages, claims letters, enquiries, etc.) provides a quantity of information which cannot be exploited manually and which is at present little used at EDF. Nevertheless, this data is essential in order to have good customer knowledge and to improve customer management particularly in the present context of market opening.

The process of extracting this information from a large volume of unstructured text data is called text mining. It starts with a primordial stage of data preparation which can range from structuring to enriching passing through filtering. The methods then used to extract the relevant information are principally clustering and classification.

It is in this context that we have been confronted with a choice of software processing text data (text mining), for different types of studies. By text mining tools, we mean here tools permitting the analysis of non-structured data (textual in our case) associated with structured data, such as data related to customer consumption or housing type. In order to help with the evaluation and comparison of these tools, a test table needed to be created.

This table thus had two principal objectives:

- Permitting the comparative evaluation of software analysing textual data;

- An aid to decision making when buying such a software which must be suited to the users' needs.

The method was deliberately pragmatic, seeking first similar experience in the data mining and natural language processing fields.

An evaluation of data mining software (CXP, 2001) gave a good insight into the subjects which should figure in our table. It did not treat the part relating to the text processing functions which are non existent in this field. We found very few references to the creation of evaluation protocols or the evaluation of this type of tool (Brugidou et al., 2000). The articles dealing with this subject concern very specific software, such as evaluation for automatic summary (Barthel et al., 2002) or syntax analysers (Aït Mokhtar et al., 2003). The European project TECHNOLANGUE, co-financed by the French Ministry for Research and New Technologies, the French Ministry for Industry and the Ministry for Culture and Communication includes an evaluation programme divided into natural language processing fields, in which text mining is not given (such as EVALDA-ARCADE II for the evaluation of the alignment of multi-lingual documents or EVALDA-CESTA for automatic translation systems).

The table permitting the evaluation of our tools was created using the above-mentioned references and evolved during the tests on the three document collections' types selected (open enquiry questions, comment fields in a customer contact database, discussion forums).

We finally selected a set of criteria organised into three streams:

- commercial (price, services, documentation, etc.) ;

- technical (architecture, volume limits, etc.) ;

- functional (possible processing, user friendliness, usefulness of results, etc.).

This article contains the method, the summarised criteria of the test table[1] and the results of the four tools evaluated. A more long-term objective is the creation of a real text mining tool test protocol.

DRAWING UP THE TEST TABLE

METHOD

During the creation of the test table, we sought to avoid favouring a statistical over an automatic language processing solution, the first being more sensitive to the means available for text analysis, the second seeing to find data analysis or modelling methods. Moreover, we did not wish to list more or less exhaustively the functions available in all the tools without concentrating more particularly on the methods applicable to the text.

A language should also be used which is understandable by the two communities. Moreover, unlike the data mining tools which offer wider and wider ranges of methods, it seems important to refocus the studies on the essential properties (capacity to clusterize, capacity to classify) expected from a text mining tool where the results can be interpreted and are relevant.

CHOOSING THE SOFTWARE TO BE EVALUATED

The software tested does not form an exhaustive list of the market offer, nevertheless we voluntarily chose very different software products representing a wide technical range of solutions (in the methodologies proposed) and a wide operational cover (in the possible fields).

The choice of software took place in two steps: SAS being the EDF reference statistical tool, it appeared obvious that the new SAS/Text Miner module should be tested. The same was true for Image/Alceste, a tool used for several years at EDF R&D for text analysis. The two other tools were then chosen to complete the range with the objective of covering together a maximum of functions offered by these so-called text mining tools. We thus selected the TEMIS Insight Discoverer series presented as a real text mining solution and a statistical software permitting text analysis, SPAD/CRM of the company DECISIA.

These four software packages therefore give a very good overview of the existing text mining tools for the following reasons:

- The “commercial” orientations are different.
Alceste is a dedicated text data analysis software used to process speech without references. SAS Text Miner is a text mining solution integrated into the data mining software series. SPAD/CRM is also positioned on the data mining market. TEMIS Insight Discoverer is an exclusively text mining tool.

- A wide range of methods and functions is adopted.
We can observe major differences in the purely text processing (presence or not of a more or less efficient linguistic tool), in the construction and reduction of lexical tables (factorial or other analyses) and finally in the proposed methods of analysis.

- The degree of maturity is different.
Alceste and SPAD/CRM (formerly called SPAD-T) are software tested in the analysis of text data, the SAS Text Miner solution is on the other hand very recent (the version of SAS which we are testing is the first commercial version) as is the TEMIS Insight Discoverer solution.

RUNNING THE TEST

It is important that the text mining tools are all tested according to the same principle in order to guarantee that the results can be compared and are relatively durable over time. The tests have been carried out by a single machine and the same person within a professional degree data mining training course[2].

For each corpus, the functions of each software were tested and the various points of the test table progressively filled in.

The document collection chosen for the tests is a non representative sample of all the test data which could be analysed by text mining methods. However, they were selected for their variable nature and because they correspond to the various types of documents that we have to analyse, such as:

- “QO”: replies to an open question in an EDF satisfaction enquiry;

- “comments”: comment fields extracted from an EDF database; this field is filled in by the employee following telephone conversations with customers;

- “forums”: Lincoln discussion forums on the Internet.

The type of data (natural language or retranscribed) varies according to the type of text: the “QO” and “forum” texts are in natural language compared to that of the “comments” where the reasons for the customer call have been retranscribed in abbreviations by an operator.

The volume is also very variable from one type of text to another: the “comments” text consists of 100,000 call reasons, “forums” contains 400 actions and “QO” 2,000 replies to an open question concerning satisfaction.

RESULTS AND INTERPRETATION

TESTING THE FOUR TOOLS

The evaluations carried out reveal a certain number of differences and similarities between the various tools whether in their use, the method, the linguistic functions or the statistical functions. A ranking has been given to

[1] The test table and detailed results can be provided upon request and at the posters session of the LREC 2004 conference.

[2] End of study training course in collaboration with the company Lincoln, 92774 Boulogne-Billancourt Cedex – France.

each of them based on 10 macro-criteria summarising those given in detail in our test table (see figure 1 below):

- The company (durability of the editor, country of origin, etc.)

- the product (architecture, product costs, learning curve)

- data access (volume, pre-formatting)

- the linguistic tools (presence and quality of the linguistic tool)

- automation (necessity for manual action and quality of tool provided)

- dimension reduction (quality and diversity of methods to transform the lexical table)

- the clustering methods (diversity of methods to file documents or reveal the subjects studied)

- the classification methods (variety of methods to create automatic classification models)

- reading results (presentation and legibility of results or help in interpretation)

- and finally the report (presence and quality of the study report automatically produced by the software)

INTERPRETATION

The four products are very different.

For Alceste, the data pre-processing is automatic and efficient (data highly enriched by linguistic tools). The thematic classes obtained are homogenous and their characterisation with exogenous variables, if these exist, is efficient and useful. The analysis report, automatically generated, is also an advantage. The two main problems are the limited volume and the absence of modelling methods (no filing method, for example). Alceste is thus a very efficient tool for rapidly detecting the subjects of a text type but cannot be seen as a text mining tool.

On the other hand, SAS Text Miner has a great number of modelisation methods and the volume is not limited except by the machine's specifications. However, for the linguistic part, it is necessary to wait for the next version to be able to really test its possibilities (erroneous results, apparently not tested for French). The help for interpreting results is also very disappointing (no characterisation of the classes obtained, no classification according to the relevance of documents of one class, display interface not very ergonomic, etc.). This text mining tool is clearly intended for experienced "data miners" who already know the SAS Enterprise Miner data mining product in which the text mining module is included.

The TEMIS Insight Discoverer software series is a text mining tool which permits an efficient text analysis and which can process a large volume of data. However, it does not have statistical functions permitting, for example, the classes obtained by text clustering to be characterised with illustrative, non textual variables. In order to use the product to its full potential, it is for the moment necessary to know a programming language to be able to manipulate the data.

The SPAD/CRM is between Alceste and Text Miner. This product has data analysis tools permitting a refined exploration of the body of text and modelling tools permitting the creation of filing models over large volumes of data. However, it has no linguistic tool and even if the "word" filter interface is quite user-friendly, the data preparation phase before analysis is long and fastidious for large corpora.

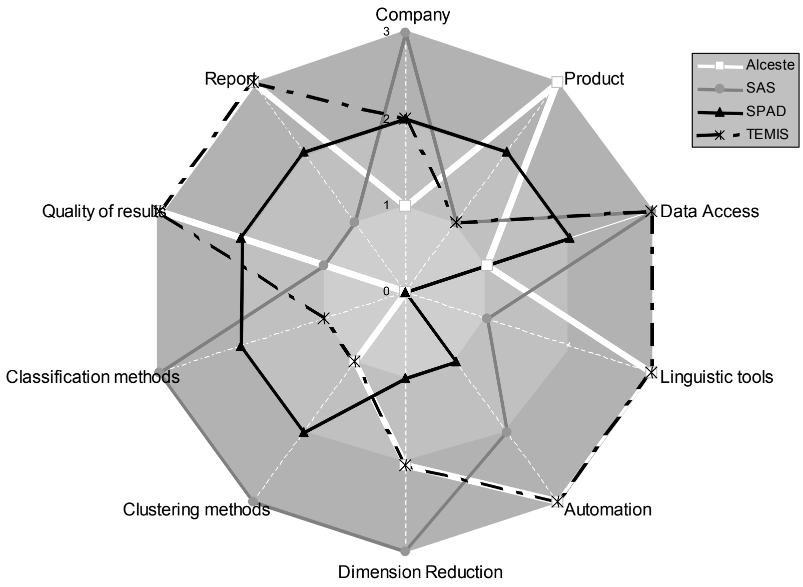

Figure 1 : The range of the four Text Mining Tools

CONCLUSION

The main objective of this study, that is to recommend a company software, has evolved towards a comparative evaluation of the software in the sense where the diversity of text to be treated, as well as the objectives, are such that the tools are more complementary than competing.

As far as EDF's specific needs and texts are concerned, Alceste remains the preferred tool for sorting and exploring the replies to the open questions of satisfaction enquiries, with a small volume and presenting the illustrative explanatory variables. The TEMIS Insight Discoverer software series is well adapted to the processing of comment fields for our customer contact data base where the volume exceeds 100,000 documents and in which is found a very specific language (technical vocabulary, use of abbreviations, etc.).

This evaluation comforts the idea of using a detailed test protocol in order to identify a set of essential functions both depending on the objectives in using a text mining tool, the corpus to be analysed by also the user profile (linguist, statistician, data miner). Moreover, at the end of this, it seems that the tools with the initial vocation being to analyse text remain the most efficient.

The test table presented is intended to help this evolution towards a true text mining tool test protocol, on the one hand by testing other tool types (tools dedicated to information intelligence, for example) or other types of text (multilingual) and on the other hand by choosing a panel of users with variable levels of knowledge.

REFERENCES

- Aït Mokhtar S., Hagège C. et Sàndor A.(2003). Problèmes d'intersubjectivité dans l'évaluation des analyseurs syntaxiques. Actes de TALN 2003. www.sciences.univ-nantes.fr/irin/taln2003/articles/eval1.pdf
- Barthel M.P., Khouas L., Sanford E. et Couillault A. (2002). Evaluation automatique pour résumé automatique. ATALA study days on automatic text summaries: solutions and perspectives. http://www.atala.org/je/021214/Barthel.pdf
- Brugidou M., Escoffier C., Folch H., Lahlou S., Le Roux D., Morin-Andréani P. et Piat G. (2000). Les facteurs de choix et d'utilisation de logiciels d'Analyse de Données Textuelles. In : JADT 2000 (5èmes Journées Internationales d'Analyse Statistique des Données Textuelles).
- CXP. PackExperts (2001) "Business Intelligence : outils de Data Mining", société CXP International. 19-21 rue du rocher, 75008 PARIS.
- Nugier S. Garrouste D., Peradotto A. et Quatrain Y. (2003). Grille de test de logiciels de text mining. EDF internal note. Available upon request.