Exploiting Text Mining in Publishing and Education

Marko Grobelnik MARKO.GROBELNIK@IJS.SI
Dunja Mladenic DUNJA.MLADENIC@IJS.SI
J.Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
Mitja Jermol MITJA.JERMOL@DZS.SI
DZS d.d., Mali Trg 6, 1000 Ljubljana, Slovenia

Abstract

The paper describes an experience on using Text mining methods for enhancing a customer application in the area of publishing and education. We described the whole process of identifying customers with potentially interesting problems and narrowing down to the one of them followed by the five text mining phases (adopted from CRISP data mining methodology). From the application and research point of view, the project involved in particular: (1) taxonomy/ontology building from a plain set of documents, (2) searching the document database, and (3) addressing (non-English) language specific issues.

1. Introduction

Educational materials as well as other materials published in the last years on paper usually exist also in electronic form. Many publishers also provide their materials on electronic media exclusively or in addition to the published paper version. We can notice that technology advances influenced the publishing industry in many aspects. The most fundamental fact is the ongoing transformation process from traditional publishing to the educational service business. With the help of Information and Communication Technology (ICT), advanced methods and tools the availability of different kind of published materials in electronic and complex multimedia form is raising rapidly. In our opinion, educational branch from the side of business as well as society can profit a lot from the new technologies. Computer Aided publishing, digital archives, multimedia, on-line publishing, teleworking they all strongly influence on traditional thinking and working. The right and cleaver use of ICT can help publishers to become more flexible, reliable and to have quality. Although introducing ICT means a lot of problems, the use of it is a must. We focus here on technology advances for analysing usually large amount of data in text format. The methods we base our experience on are so called *Text Mining* methods and can be viewed as an extension of Data Mining to text data. Text Mining is relatively new interdisciplinary area involving *Machine Learning and Data Mining* - contributing data analysis within different knowledge representations and potentially having large amount of data, *Statistics* - contributing data analysis in general, *Information Retrieval* - contributing text manipulation and handling mechanisms and *Natural Language Processing* – contributing mechanisms for analysing natural language. Some of the typical aspects of Text Mining research involve development of models for reasoning about new text documents based on words, phrases, linguistic and grammatical properties of text and; extracting information and knowledge from large amounts of text documents.

Different application areas have been identified as having potential for Text Mining. Some of the typical Text Mining applications are text categorization and keyword assignment where we model the existing set of documents in order to automatically assign topic categories or keywords to new documents. User profiling is another area, where the profiles of user interests are build based on the text of documents the user is handling (or accessing from the Web), potentially also using interaction of the users with similar documents and connections between the documents themselves (e.g., considering hyperlink on the Web pages). Providing help for building ontologies of documents, performing intelligent text search, text segmentation, topic tracking are some of the other application areas for the research and development of Text Mining methods.

2. Approaching the real world

Extending research and development results to *the real world applications* has been found as a difficult task by many researchers. By real world we mean business, industry, public sector and even individuals in their every day life. Usually the real world can profit from the applications of research results but the level of awareness of the potentials of any new research achievements is rather low. From the research point of view transfer of technology is demanding, requires time and capabilities

beyond research itself. From the business point of view making profit out from the research results is difficult and risky, requires considerable investigation of time to get the research results to the right form, training of marketing people. Nevertheless research and development including technology transfer is considered important and thus supported by governments through different channels. European Commission forms different programs to support research, development and cooperation between academic and commercial institutions. One of them is IST program inside which a part of the described work was accomplished in Sol-Eu-Net RTD project [Mladenic 2001] inside the subproject on practical applications of Text and Web Mining.

In order to find suitable problems and interested customers, we made contact with about a dozen different customers (here also referred to as end-users) and with about half of them we had more than one meeting. Out of that pool we identified four that contributed to our list of promising end-user problems. We had several meetings with each of them. Following is the list of the four potential customers and a brief description of their problem we were considering:

- **DZS - Slovenian publishing house. They are interested in different support to search on text databases and also in automatic document categorization.** The collection they have consists of text document giving educational materials for different areas and different levels of primary, secondary and high school education. Materials are prepared with contractual authors, mostly distinguished authorities from the specific field and then edited by either in-house editors or in cooperation with experts for pedagogy, methodology and didactic. As different authors have different possibilities and preferences when working on computers, material are potentially provided in different formats and transformed only as needed for the "classical paper-publishing" procedure. There was an ongoing project at DZS on developing uniformly formatted database of educational material based on uniform ontologies with the future goal of offering access on electronic media (CDs and Web) as well as on printed materials as well as on printed materials and to automate the process for the selection and categorization of educational material
- **IBMI - institute for biomedical informatics expressed interest in any help we could provide for hierarchical document categorization.** There work includes also building and updating a database of published medical paper following an international structure of medical keywords Medical Subject Headings – MeSH. Each paper or abstract is equipped with a set of keywords, depending on it content. The process is time consuming requiring a high level of expertise in mapping the content of the papers with the given structure of keywords. They have a couple of domain experts and higher a number of medical students to help. However, one of the domain experts complained at one of our joint meetings that many times the students lack of familiarity with the prescribed keyword structure causes additional work for the experts. They were very interested in our research experience of automatic document categorization based on the given hierarchy (ontology) of keywords and documents. For our research work we have used large, publicly available hierarchy of Web documents – Yahoo! [Mladenic 1998], [Mladenic Grobelnik 1999].
- **GZS – Chamber of commerce and Industry of Slovenia. They were interested in user profiling and Web access analysis.** One of their important activities is providing information via well-structured and maintained Web site <http://www.gzs.si/eng/index.htm>. They have a number of users subscribed to their databases and different levels and also different information they have about their subscribers. Some of the information they have is publicly available while more detailed information can be obtained only via subscription. They were interested in improving their Web sites and we found that User profiling and Web access analysis are interesting directions to start our cooperation on. We had a couple of meetings but it seems that the priorities in their organization have changed in the meantime and we postponed further meetings for some time later.
- **GV – Business newspapers. They were interested in Text Mining in general including User profiling and Web access analysis.** We had preliminary meeting discussing possibilities but we did not have a chance to get involved more with this end-user, mainly due to the time constraints and the good progress with the first two end-user.

In the next phase we proceeded by selecting the two most promising problems/end-users that also have shown great enthusiasm in results of data mining on their text data. The first is the biggest Slovenian publishing house of educational materials interested in improving access to their materials that are offered in electronic form (DZS) and to develop methods and tools to support editors and authors in the process of searching, selecting and refining web based materials. The second is institute for biomedical informatics that among others also builds a database of medical papers equipping them with some additional information describing their content (IBMI).

From both end-users we obtained a sample data and made some preliminary inspection in order to see if and how we would proceed with the more precise definition of a target problem and development of a prototype solution. Based on those preliminary studies and taking into account our

time constraints, we decided to proceed with the first end-user that is the publishing house of educational materials.

3. Identifying text mining problems in publishing and education

In the context of publishing and education we have identified several problem that can potentially profited from text mining. Here we briefly describe each of them to illustrate how the methods themselves can be put in the context of applications.

Editorial support is the most obvious problem, where text mining can be used as support for editors when gathering and preparing the materials as well as for the users when accessing them. There are several aspects of editor support we identified as interesting in cooperation with one of the biggest Slovenian publishing house of educational materials. We can envision the system where the input is complex, content based query and the output is a list of the most relevant documents regarding to the query within specified document set. The method to be used is content indexing of documents with adapted scoring mechanisms for meta search engines (on the Web).

Text Categorization can be applied when we are given a set of predefined categories potentially organized in some ontology like for instance the Yahoo! ontology of Web documents. The task here is to assign one or more content categories to a new document. Related to this is also task of assigning one or more keywords to a new document. On that kind of problem setting we already have successful experiment with Yahoo! ontology of Web documents [Mladenic 1998], [Mladenic Grobelnik 1999].

Building Ontologies is usually approached manually and can be supported by the text mining methods. As input the system takes possibly large set of documents and outputs a hierarchy (or ontology) of documents and document sets grouped by similarity. The method to be used here is document clustering.

User Profiling can be based on a trace of documents the user read by browsing a closed set of textual materials. The output of the system is a model of user's interests and the most common browsing patterns. This model can be used for personalized selection of information in various contexts.

Group work support is another potentially interesting problem, where the goal is to associate people (students, professors) with similar interests and to achieve some synergetic effects in e.g. education and editorial work. Input to the system would be a model of user interests and texts the user wrote (reports, articles, newsgroups messages). The output is recommendation for groups of people with potentially similar interests.

We were glad to confirm our initial hypothesis that there is a great potential of text mining in publishing and education. We can observe that the common characteristics of the techniques needed for addressing the described problems is identification of contents and structure within free text

4. Phases in the solution development

Data mining processes consist of a number of phases that are not necessarily executed in a linear manner. For instance, the results of one phase may reviele details related to some of the previous phasees and thus require more effort on a phase that was already considered as completed. There are different definisiont of the pahses of Data mining process. In our work on real-world problems involving end-users, we have adopted the CRISP-DM methodology — CRoss Industry Standard Process for Data Mining [Chapman et al.2000]. The methodology has been developed by a consortium of industrial data mining companies as an attempt to standardise the process of data mining. In CRISP-DM, six interrelated phases are used to describe the data mining process: *business understanding*, *data understanding*, *data preparation*, *modelling*, *evaluation*, and *deployment*.

4.1 Business and data understanding

Our work on text mining for publishing house of educational materials also approximately followed these five phases. The first two phases of *business and data understanding* resulted in focusing on two text mining problems: different support to search on their text databases and automatic document categorization. It should be pointed out that we had to deal with non-English texts that practically made other potentially available partial solutions inappropriate. This fact was in addition to our deep understanding of the text mining mechanisms probably one of the more important reasons for getting the business case. Namely, from the business point of view it is often not clear how much can the application really profit from incorporation of the recent research achievements, especially considering the additional efforts needed for its development and cooperation with the researchers.

4.2 Data preparation

After the definition of the target problems, the end-user provided all the relevant datasets from the three education subject areas: history, physics, and biology. The whole dataset contained several tens of thousands of documents in HTML and XML format. Closer inspection of the data reviled differences between the subsets of the data that were mainly caused by the involvement of different editors and in-house priorities during the previous work in their project on collecting the data. In the next phase of *data preparation* we agreed that some additional tagging was needed. For instance, adding meta-information to XML document about the paragraph content. The team of

editors at the end-user site annotated a subset of the materials with the additionally needed information. There were also some other differences in the coding of the materials that were handled automatically at our site. For instance, taking care of different coding of special characters used in the addressed natural language (Slovenian).

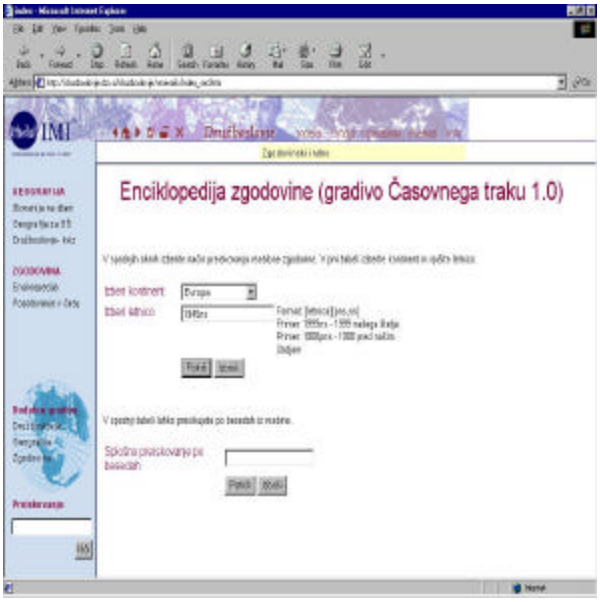

Figure 1 Customized history encyclopedia search engine.

4.3 Modeling and evaluation

In our prototype development we have addressed two problems identified as interesting for the customer: (1) support to search on end-user in-house text databases and (2) support to ontology construction from in-house `.XML` documents. Each of them is described in the following subsections. Since the used modeling techniques are well established in the research community, the evaluation part was mainly concentrated on correctness of the results and was performed in the loop with the end-user in the last phase of finalizing the prototype.

Support to search on end-user in-house text databases was partially, but not sufficiently, covered by the existing search engines. The problems were mainly because of the natural language specifics and the additional functionality required by the end-user that was not offered by the general search engines. We have developed a customized search engine supporting the native language in which the provided educational materials are written (Slovenian). Through the communication with the end user, the following important characteristics of the systems were identified. The system should support extended querying language that in addition to usual queries also enables range queries (referring to the era the document is describing) and has capabilities of being extended also for querying problem specific keywords. The system should be suitable for searching electronic documents such as books or sets of several thousands of documents. It is also important that it can be used as a CGI script either on the Web or Wap and also as standalone (in our case Microsoft ActiveX) component integrated in the information system of the end-user. All these requirements were met, meaning also that from the technical point of view efforts were invested in using the recent technology and that was actually as an important part of the prototype development. We show example page in the final end-user system for accessing the developed customize search engine via text interface Figure 1 and via graphical interface Figure 2.

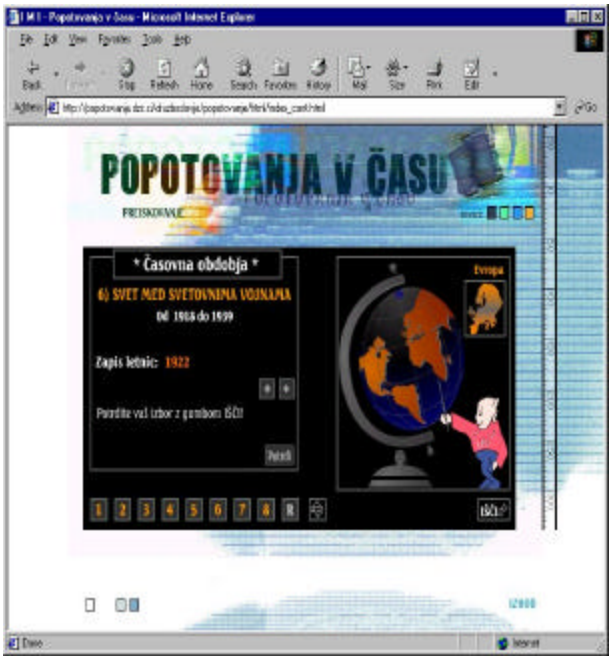

Figure 2 Customized graphic-supported search for history-geography knowledge base.

Support to ontology construction of the end-user in-house text database was not covered by the available systems, involved more research efforts from our side, but was a bit lower on the end-user's priority list. This problem was addressed by automatically building a hierarchy (taxonomy/ontology) of documents providing an easy way for the end-user to organize a large collection of documents. In order to handle the large amount of text data and build a hierarchical structure, a library was developed implementing efficient hierarchical clustering approach. The standard K-means clustering was used as a base for the top-down hierarchical K-means clustering for text data. The clustering approach was used in combination with the given taxonomy of topics and keywords describing the preferred way of structuring the documents. The approach was similar to the co-training type of algorithms where two independent sets of attributes mutually reinforce each other to converge to a

better model – in this case hierarchy of documents. The other important, non-standard part of the system was handling natural language specifics for Slovenian language. In particular, three main issues were: (1) stop word selection, (2) lemmatization (normalization) of the words and more trivial, (3) dealing with several standards for Slovenian character set. First two issues were solved using the dictionary of word pairs in the form (normalized word, inflected word) each being annotated with the word type. The list was collected in one of the previous projects. Using this list with the selection of certain word types, we created stop word list used in the further work. Lemmatization was done using the above word pair list by a simple lookup into the table of inflected words each being accompanied with its normalized version. Figure 3 shows a part of generated ontology as used through the search engine.

Figure 3 Part of the generated ontology as used behind the customized search engine (the topmost part shows the three hits related to the issued query.

5. Deployment

The resulted prototypes are included in one of the main projects of multimedia division of DZS publishing house, the IMI project (Just-for-You Education). IMI is putting into practice the vision of the DZS Educational Publishing Division, which is based on comprehensive support for education in the information society, the humanisation of education and respect for the interest and needs of each individual.

Besides the complex publishing system, IMI project resulted in four independent (but interconnected) Web educational portals (see Figure 3) with well structured, ontology based dynamic contents, supported by communication tools and services mentioned in section 2. More than 15.000 atoms of knowledge (app. 18.000 pages) were prepared, several methods and courses to support self-learning and self-testing were implemented for Civic education (druzboslovje.dzs.si), Biology (biologija.dzs.si, Physics (fizika.dzs.si) and Pedagogy (pedagogika.dzs.si). Portals were already sold to over 70 schools all over the country and so targeting more than 35 thousands of individual users. The publishing house expressed their strong believe that the included prototype we have provided improved the quality of their product and potentially also brought financial benefit for their company. When dealing with educational materials and particularly with distance learning then what makes the service successful, competitive and interesting to the users are additional methods and tools. With the help of developed prototypes DZS educational services become the most successful in Slovenia and so making DZS one of the core partners in supporting Information Society in Slovenia. The exact details for direct revenue of the prototype is hard to give, because of the tight connection of the developed prototypes to their comprehensive new product, but indirect influences to the DZS position on Web education business market are very positive.

Figure 4 Publishing house IMI Web portals that incorporate text mining prototypes.

6. Summary and the lessons learned

The described cooperation between J.Stefan Institute and the publishing house DZS gave us two kinds of experience: (1) motivation for new technical solutions and (2) specifics on cooperation with the people working in the publishing house.

First experience is meeting very unclean text data edited by several editors, existing in several formats (XML, HTML, ordinary text). Since the requirements were to build several solutions in the areas of data search engine and construction of hierarchical index, we built as many customized filters, as there were data formats. This solution enabled us to have text in the same format that made the system more complex and more expensive to maintain.

The second experience is about requirements for customized search engine. Usual requirements include simple query language with basic logical operators (AND, OR, NOT). In our case the publishing house required special type of query operators since the system was used for the special type of data (history). As a result we added a mechanism for inexpensive inclusion of new complex query operators at the expense of additional space.

Next, an interesting technical solution was found for the problem of hierarchical index construction, where the hierarchy was given in advance with no additional descriptions. We made simple approximate join between given hierarchy and documents in the database using search engine, which produced very useful results.

For the second problem when making the hierarchy out of the set of documents, we used relatively novel approach using 2-means clustering, dividing the set of documents into two groups, and further dividing each of the group into new subgroups etc. since some intuitive the stopping condition was met. This procedure constructed very interesting hierarchy of structured subjects, which can serve as a good starting point for human editors.

As the last experience we have to note specifics when dealing with the people from the publishing house. Their primary concern is the contents in the text documents while the technical issues are always in the second plan although the technical support in the company seems at the very good level. Usual practice is to start with certain standard (like XML) but soon, because of the time constraints or some similar reasons, editors in the publishing house stop respecting agreements on the standard to fulfill current obligations which make their solution hard to maintain and more expensive on the longer run.

For the conclusion we could say that text mining seems very interesting and promising for the content based industries like publishing and media houses or companies dealing with large amounts of documents. There are two major aspects the publisher found as interesting and promising area for text mining. First is that methods and tools based on data and text mining add value to on-line services. The quality and success of Learning through ICT is strongly dependent on appropriate methods and tools. Personalisation of learning, ontology based dynamic catalogs of content, extensive search and find methods as well as group work are already becoming distance learning system necessities. Second founding is that with the help of data and text mining tools several activities in the publishing process (on-line, CD and paper) could be optimised and automatised. Automatic categorization makes in the world of information overload the basic publishing process of "knowledge refinery" realistic and manageable. Extensive searchers help editors to overcome the "click-and-miss" stresses in daily work. So in general the filling from the publisher about these kind of technology was enthusiastic, positive and very promising.

But before they came to that conclusions the experience from our site was that the solutions provided by text mining technology seemed rather new for this type of customers where the most common practice is two level manual work (editing and design) with not enough experience in on-line contents. This led us to long initial phase of persuasion of the customer about the benefits of new approaches work the business. For that reason we think the most convenient way to work with customers is first to find or educate somebody within the customer company with some general conceptual knowledge about data and text mining methods.

Acknowledgement

The reported work was partially supported by the EU Fifth RTD Framework Programme project Sol-Eu-Net (IST-1999-11495, www.SolEuNet.ijs) and by the Slovenian Ministry of Education, Science and Sport.

References

[Chapman et al 2000] Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler) *(2000) CRISP-DM 1.0: Step-by-step data mining guide.*

*[*Mladenic 1998] Mladenic, D. (1998) *Turning Yahoo into an Automatic Web-Page Classifier*, Proceedings of the 13th European Conference on Aritficial Intelligence ECAI'98 *(pp. 473-474)*.

[Mladenic Grobelnik 1999] Mladenic, D. and Grobelnik, M. (1999). *Feature selection for unbalanced class distribution and Naive Bayes*, Proceedings of the 16th International Conference on Machine Learning ICML-99. Morgan Kaufmann Publishers, San Francisco, CA, pp.258-267.

[Mladenic 2001] Mladenic, D. (2001) *EU project: data mining and decision support for business competativeness : a European virtual enterprise (Sol-Eu-Net).* In: D'ATRI, A., SOLVBERG, A., WILLCOCKS, L. (eds.). OES-SEO 2001: open enterprise solutions : systems, experiences and organizations: Rome, 14-15 September 2001. Roma: LUISS, 2001, pp. 172-173.