Approaching Analysis of EU IST Projects Database

Marko Grobelnik
Jožef Stefan Institute, Ljubljana, Slovenia
Marko.Grobelnik@ijs.si

Dunja Mladenić
Jožef Stefan Institute, Ljubljana, Slovenia
Dunja.Mladenic@ijs.si

Abstract: *We present the first results of the analysis of publicly available EU IST project descriptions. The database was automatically built from the publicly available information on the Web and organized to enable further analysis. We have used Text Mining methods to group the projects according to their content and the institutions participating in the projects. Two systems were developed, the first enabling grouping and visualization of the projects and the second enabling searching over the projects and partners. One of the advantages of the proposed methodology is that it is scalable and can be easily applied on large collections of project descriptions. By changing the filters needed for automatic creation of the database from the Web site, the approach could be used on other Web sites giving similar type of information as EC site with funded project descriptions. Our intention here is mainly to show potential of Text Mining and stimulate further development of application oriented Text Mining systems. Our customers here are EC officers and analysts in charge of EU funding of projects. The first feedback we got from them on two presentations we had was very positive, contributing many ideas for potential problems and improvements of the first version of the system.*

Keywords: *Text Mining Application, Information Extraction from the Web, EU Project Description Analysis*

1. INTRODUCTION

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 [1]. 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*. We have adopted that phases for our Text Mining efforts.

Our intention here is mainly to show potential of Text Mining and stimulate further development of application oriented Text Mining systems. In the process of addressing Text Mining and Natural Language related [4] practical problem, we have identified European Commission as a potential end-user. We have concluded business and data understanding involving also discussion with the representative of the end-user. In our case the end-user representative is the project officer of SolEuNet [5] project inside which the described Text Mining prototype solution development is taking place. We have defined two problems: grouping IST projects based on their content and finding connections between the partners in the projects. After meeting with the end-user we have started work on understanding the problem and defining the prototype goal. We managed to get all the data for FP5 IST projects by crawling the EC Web site. It is also possible to extend our work on other EU non-IST projects, if this turns to be interesting for the end-user.

2. IST DATABASE DESCRIPTION

The data was obtained from the database of 5FP EU projects at <http://www.cordis.lu/> [2]. We have about 9000 5FP project descriptions that were downloaded as HTML files. Out of that HTML files the relevant information was extracted: project name, acronym, start and end dates, textual project description, a list of partner institutions. The next step was cleaning the data, mainly because of inconsistencies (mainly in referencing the institutions). For our preliminary analysis we have used only descriptions for IST projects, resulting in about 1700 project descriptions. Each project is represented with text description of the project content and a list of the institutions acting as partners in the project.

3. VISUALIZATION OF PROJECT DESCRIPTIONS

The usual approach when dealing with text for visualization is first to transform the text data into some form of high dimensional data and in the second step to carry out some kind of dimensionality reduction down to two or three dimensions that allows to graphically visualize the data. There are several (but not too many) approaches and techniques offering different insights into the text data like: showing similarity structure of documents in the corpora (e.g. WebSOM, ThemeScape), showing time line or topic development through time in the corpora (e.g. ThemeRiver), showing frequent words and phrases relationships between them (Pajek), etc.

One of the most important issues when dealing with visualization techniques is scalability of the approach to enable processing of very large amounts of the data. We propose two text visualization techniques both based on clustering [7] of the textual project descriptions and working in linear time and space complexity [3].

The first procedure is a combination of the K-Means clustering procedure and a technique for nice graph drawing. The idea is first to build certain number of document clusters (with K-Means procedure), which are in the second step transformed into the graph structure where more similar clusters are connected and bound more tightly. The third step performs one sort of multidimensional scaling procedure by aesthetically drawing of the graph. Each node in the graph represents the set of similar documents represented by the most relevant and distinguishing keywords denoting the topic of the documents. In the graph based visualization project descriptions were clustered in K (in example in Figure 1 we set K=20) groups based on the document similarity defined as cosine similarity between the bag-of-word representation [6] of the text documents. On the top of the clusters a graph was formed, where each node represents a group of similar documents based on their content.

The second procedure performs hierarchical K-Means clustering producing a hierarchy of document clusters. In the next step the hierarchy is drawn as a two-dimensional area splits reflecting the hierarchy splits. Like in the first approach, each cluster (group of documents) in the hierarchy is represented by the set of the most relevant keywords from the project descriptions.

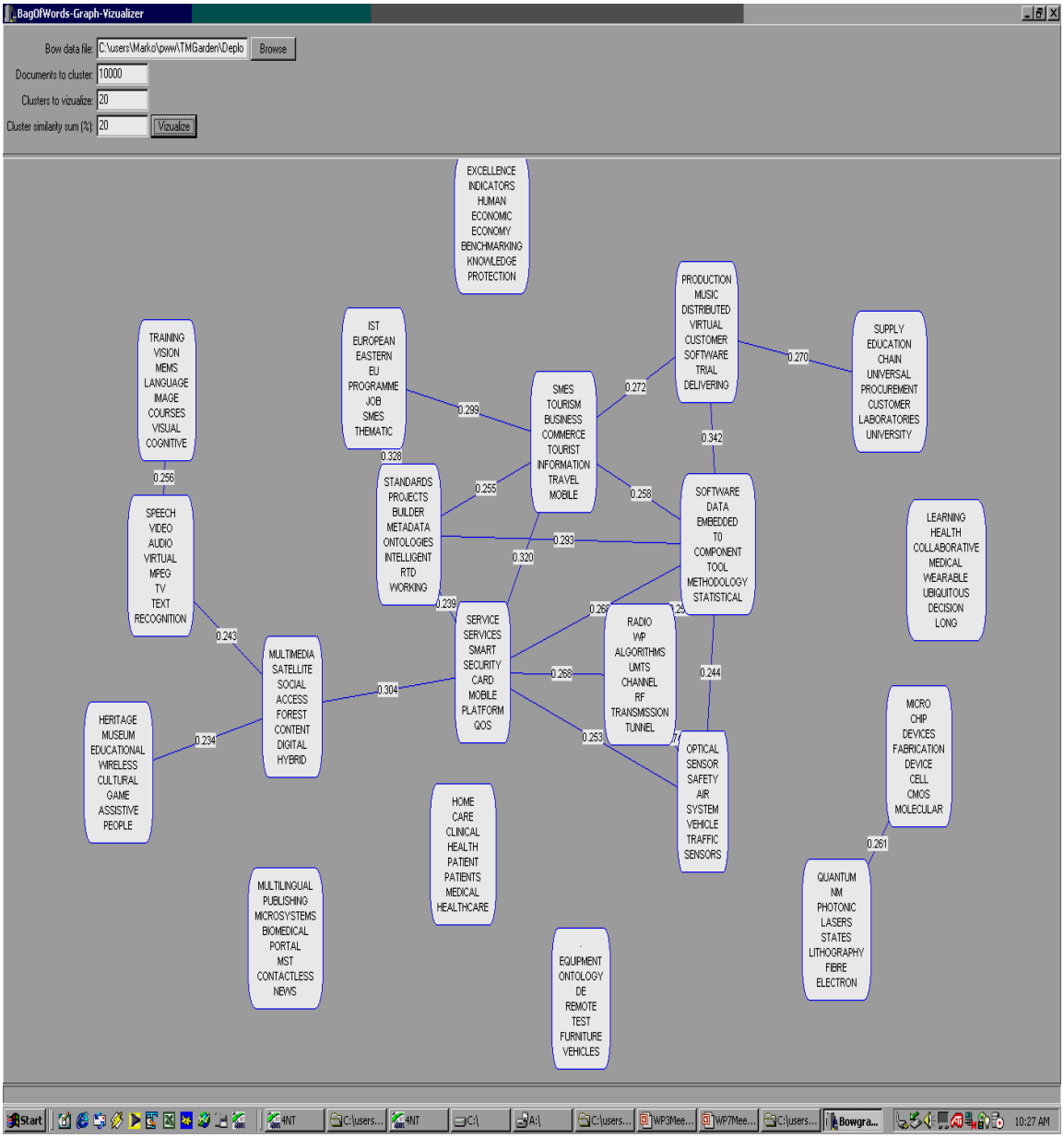

Figure 1: Graph based visualization of EU IST 5th FP projects using 20 groups, labeling them with the most characteristic words and connecting the most similar groups.

4. IDENTIFYING THEMATIC CONSORTIA OF INSTITUTIONS

We have identified an interesting problem of identifying a set of the most relevant institutions for a certain thematic area. As the input the user provides a set of keywords describing a topic. The output of the system is a list of institutions sorted by their relevance for the given set of keywords. The generated set of institutions could be understood as a proposal of consortium for the given thematic area. As an example we give the top 20 institutions forming a possible Data Mining consortium for the future projects (see Figure 2). These top 20 institutions were obtained using the following set of

"data-mining" related keywords: "knowledge discovery text mining classification machine learning data mining data analysis personalization decision support".

Rank. (Relevance) Institution name - [list of relevant projects]

1. (1.564) GMD FORSCHUNGSZENTRUM INFORMATIONSTECHNIK - [SOL-EU-NET, SPIN!, XML-KM, CYCLADES, VESPER, COGITO]
2. (1.404) UNIVERSITAET DORTMUND - [MINING MART, KDNET, DREAM, CYCLADES, APPOL II]
3. (1.059) DIALOGIS SOFTWARE SERVICES - [MINING MART, SOL-EU-NET, SPIN!]
4. (0.782) EUROPEAN COMMISSION JOINT RESEARCH CENTRE - [MINEO, KDNET, LINK3D, ETB, CTOSE, NOSE II]
5. (0.758) UNIVERSITA DEGLI STUDI DI BARI - [SPIN!, KDNET, LINK3D, COGITO]
6. (0.745) FRAUNHOFER GESELLSCHAFT ZUR FOERDERUNG DER ANGEWANDTEN FORSCHUNG - [KDNET, CERENA, VOSTER]
7. (0.744) UNIVERSITA DEL PIEMONTE ORIENTALE AMEDEO AVOGADRO - [MINING MART, KDNET]
8. (0.744) SCHWEIZERISCHE LEBENSVERSICHERUNGS UND RENTENANSTALT SWISS LIFE - [MINING MART, KDNET]
9. (0.744) PEROT SYSTEMS NEDERLAND - [MINING MART, KDNET]
10. (0.649) BUREAU DE RECHERCHES GEOLOGIQUES MINIERES - [MINEO]
11. (0.636) KATHOLIEKE UNIVERSITEIT LEUVEN - [SOL-EU-NET, KDNET, VIBES, UP-ARIADNE]
12. (0.622) INSTITUT NATIONAL DES SCIENCES APPLIQUEES DE LYON - [CINQ]
13. (0.612) UNIVERSITY BRISTOL - [SOL-EU-NET, KDNET]
14. **(0.612) INSTITUT JOZEF STEFAN - [SOL-EU-NET, KDNET]**
15. (0.612) CZECH TECHNICAL UNIVERSITY PRAGUE - [SOL-EU-NET, KDNET]
16. (0.586) PIXELPARK - [KDNET, CERENA]
17. (0.579) ENGINEERING - [BIOSIM]
18. (0.557) UNIVERSITY LEEDS - [SPIN!, LIQUID, SOQUET]
19. (0.516) TEKNILLINEN KORKEAKOULU - [KDNET, NOMAD, MYGROCER, ALMA]
20. (0.500) FUNDACIO IMIM - [LIQUID, LINK3D]

Figure 2: Possible Data Mining consortium for the future projects.

We have developed prototype software for consortia searching that for the given set of keywords calculates:

- Set of the most relevant institutions
- Set of the most relevant projects

First the set of 100 most relevant projects is found using cosine similarity measure between the issues query and the text description of the project content. Then for each of that selected projects the similarity score is used to assign weight to all the partners in that project. If the same partner (institution) appears in several projects, the weights are added. Finally, all the institutions are sorted according to their weight and presented to the user.
In the direction of deployment, our prototype currently works as Windows GUI application and we plan to make a web service with the same functionality.

5. FURTHER WORK

In addition to these 1700 project descriptions of IST EU projects, there are other EU project description that we have collected. In further work we can consider extending our analysis to include these description following by preparation of the data for all 9000 FP projects. The prototype itself can benefit from improvements either giving better support for user interaction or producing results that fit better to a particular user needs. An interesting further direction is towards identifying communities (cliques) of institutions working in related areas based on the project partnership network. We also see potential in generating interactive visualizations for on-line browsing of project and partner relationships.

Further work can be also seen in the direction of identifying potential new partners for proposed projects, identifying consortia from the text on web home pages of the companies (not only from the CORDIS project database), identifying new trends from the web data (for companies or for topics) and monitoring web activity of projects, companies.

ACKNOWLEDGMENT

The work reported here was in part supported by EU IST project Sol-Eu-Net, IST-11495, and by the Slovenian Ministry of Education, Science and Sport.

REFERENCES

[1] Chapman P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. CRISP-DM 1.0: Step-by-step data mining guide, 2000.

[2] CORDIS: Public Web site of 5FP EU projects by European Commission <http://www.cordis.lu/>

[3] Grobelnik M., Mladenić D. Efficient Visualization of Large Text Corpora, The 7th TELRI Seminar "Information in Corpora", 2002.

[4] Manning, C.D., Schutze, H. Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge, MA, 2001.

[5] Mladenić D. EU project: Data mining and decision support for business competitiveness: 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, pp. 172–173, 2000.

[6] Mladenić D. Text-learning and related intelligent agents: a survey. *IEEE Inteligent. Systems and. their Appications.*, V:14, pp. 44-54, 1999.

[7] Steinbach, M., Karypis, G., and Kumar, V. A comparison of document clustering techniques. In Proceedings of KDD Workshop on Text Mining, 2000.