EXTENDING THE IST-WORLD DATABASE WITH SERBIAN RESEARCH PUBLICATIONS

Miloš Radovanović[1]*, Jure Ferlež*[2]*, Dunja Mladenič*[2]*, Marko Grobelnik*[2]*, Mirjana Ivanović*[1]

[1]University of Novi Sad, Faculty of Science
Department of Mathematics and Informatics
Trg D. Obradovića 4, 21000 Novi Sad, Serbia
e-mail: {radacha, mira}@im.ns.ac.yu

[2]Department of Knowledge Technologies
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
e-mail: {jure.ferlez, dunja.mladenic, marko.grobelnik}@ijs.si

ABSTRACT

This paper describes an effort of using knowledge technologies to gain insights into research activity, by exploiting publicly available information on research publications. The specificity of this paper is extending the existing IST World database with information on Serbian research publications. We describe the process of information extraction applied in order to fill in the database, based on publication references obtained from the textual repository maintained by the Secretariat for Science of the Serbian Province of Vojvodina. An example shows how we can gain insights into the collaboration and competence of authors.

1 INTRODUCTION

Information on research papers is available in different formats, for instance, a list of authors and the paper title. Electronic forms and large amount of such information naturally leads to development of different methods for enabling analysis of the data. In this paper we deal knowledge technologies for analyzing research publications.

The IST World portal [2] consists of innovative functionalities that help to promote RTD competencies and facilitate and foster the involvement of different research entities in joint RTD activities. It is based on the original idea of Project Intelligence [5]. The portal contains information about RTD actors on the local, national and European level (harvested from existing databases and with Web mining techniques), such as persons, research groups, organizations, projects, and their experience and expertise.

Vojvodina, the northern province of Serbia, is home to many educational and research institutions, most of which operate under the umbrella of the University of Novi Sad, and cover practically every field of science. In 2004, the Provincial Secretariat for Science and Technological Development of Vojvodina started gathering data from researchers employed at institutions within its jurisdiction. Every researcher was asked to fill in a form, provided as an MS Word document, with complete citations of all his/her authored publications, among other data. The gathered information is available in unmodified form on the Web site of the Secretariat: http://apv-nauka.ns.ac.yu.

We show how to obtain a collaboration graph based on this data, expressing coauthorship of papers. Competencies of the authors and organizations are analyzed using competence map based knowledge technologies [4]. The rest of the paper is organized as follows. In Section 2, we describe the data and the extraction process. Section 3 summarizes the functionalities of the IST World portal, while Section 4 illustrates collaboration and competence analysis of extracted data. The last Section concludes, and gives guidelines for possible future work.

2 INCORPORATING PUBLICATION DATA

The process of incorporating publication data consists of data preprocessing and importing into an existing database. This section provides the data description and details on the information extraction process.

2.1 Serbian Data Description

At the time of writing, the collection of documents (with the last update made on July 6, 2006) includes 2,278 researchers from 60 institutions, making the task of manually extracting bibliographical data infeasible for us. We resorted to programming an extractor in Java which, at this time, is able to automatically isolate every researcher's name, affiliation, and list of citations, and save the data in form of [3] compliant XML files, to enable quick import of the data into the existing IST World relational database. Furthermore, the extractor compares citations between different authors, detecting coauthorships between researchers who are included in the collection.

The form to be filled by every Serbian researcher consists of a sequence of tables starting with basic data (name, year of birth etc.), continuing with the tables corresponding to publication types as prescribed by the Serbian Ministry of Science and Environmental Protection. Publication types are labeled by a code of the form R*xx*, where *xx* is a two-digit number. The codes of interest have the first digit in {1,2,5,6,7}, which corresponds to published papers and book chapters, and excludes technical solutions (3) and patents (4). A sample entry is shown in Table 1. We observed that within the tables, the citations were usually entered enclosed in isolated paragraphs or numbered lists.

<table>
<tr><td><b>Spisak rezultata R52 - Rad u časopisu međunarodnog značaja</b>.
Međunarodne časopise i druge navode rangirati (koeficijent R) prema Science Citation *Index-u (Journal Citation Report) odnosno prema kategorizaciji radova, verifikovanih od strane odbora Ministarstva.</td><td><b>Broj</b></td><td><b>10</b></td></tr>
<tr><td colspan="3">1. Bađonski, M., Ivanović, M., and <b>Budimac, Z.</b>, Software Specification using LASS. In <i>Proc. of ASIAN ‘97</i> (Kathmandu, Nepal), Shyamasundar, R. K. and Ueda, K, eds., Lecture Notes in Computer Science vol. 1345, Springer Verlag, Berlin, 1997, pp. 375-376.
2. <b>Budimac, Z.</b> Mašulović, D., Linda as an Abstract Data Type for Concurrent Programming, <i>Novi Sad J. Math</i> 28 (1998) 2, 173-186 (Publisher: Faculty of Science, University of Novi Sad, Novi Sad).
. . .</td></tr>
</table>

Table 1: *Example entry in the form. R52 corresponds to papers published in international journals of category 2.*

2.2 Information Extraction and Import

The present version of the extractor (2.0.2) is able to isolate a total of 101,672 bibliographic units, written in at least five different languages, from current data, and detect 24,262 binary coauthorships, making the total number of citations in the database 77,410. (A paper appearing in *n* researchers' forms can have a maximum *n*–1 detected coauthorships.) The researchers' names and affiliations are extracted from the HTML page on the Web site of the Secretariat in a straightforward fashion, which left the biggest challenge in processing the citation data from MS Word documents.

Parsing. Since the only reliable option for accessing the content of MS Word documents from computer programs at this time is to use the Visual Basic for Applications (VBA) macro language, we found it more convenient to use VBA only to bulk convert all documents to HTML format, and do all actual extraction from HTML. The HTMLParser open source library is used to process the generated HTML files, and isolate the DOM trees of <TABLE> tags corresponding to tables containing the citations of interest, as described in Section 2.1. Further extraction of citations is done using the following scheme: since it was observed that isolated paragraphs and numbered lists in Word documents correspond to <P> and list tags in generated HTML, the citations were "read out" from fixed positions in the DOM trees of <TABLE> tags, taking into account only the two above possibilities.

Somewhat surprisingly, this simple scheme turned out to be rather effective at retrieving strings containing valid citations, although we are unable to give precise figures for precision and recall of detection. After observing the isolated citations, we removed from the collection the 59 forms which were obviously not parsed correctly within this scheme (the citations were either divided up into several, or lumped together). We also removed 62 forms which were filled in using the Cyrillic alphabet, since we elected to leave the conversion of Cyrillic letters for a later date. From the remaining forms, the parser could not correctly process 444 tables (out of a total of 39,688), which roughly corresponds to 17 whole forms. All this amounts to 138 unprocessed forms, putting the upper bound on recall to around 94%. Considering that the average number of detected citations per researcher is 44.63, common sense suggests that true recall should not be an order of magnitude smaller. At the same time, directly observing the recognized citations led us to a conclusion that precision is not worse than about 97%. We consider these estimates of precision and recall to be completely satisfactory for such a simple extraction scheme.

Coauthorship Detection. In order to calculate the similarity of two citations, with the intention to determine a coauthorship relation, the extractor uses an optimized version of the algorithm given in [7]. The algorithm calculates the similarity between two strings by computing the ratio between the number of shared 2-grams (letter pairs) and the number of all 2-grams in both strings, disregarding whitespace, punctuation marks and capitalization. The described ratio is multiplied by two to keep the resulting measure between 0 and 1. The reason for using 2-grams instead of, for instance, whole words, lies in the observed "dirtiness" of manually entered citation data: typographical errors, different or missing information, various citation conventions used (resulting in different ordering of citation information) etc. After parsing a researcher's form and extracting a list of citations, every citation is compared to all citations already in the database which contain the researcher's last name (actually, its first word), retrieved using a maintained index. If the best match of a given citation does not exceed a predetermined similarity threshold (currently set at 0.63 after examining several test cases), the citation is entered as a new one into the database. Otherwise, a coauthorship relation is established, and the entry for the currently processed citation of the researcher is set to refer to the citation already in the database.

As with citation recognition, it is difficult to evaluate the detected coauthorships. Nevertheless, on a sample of the extracted data for colleagues from the Faculty of Science at the University of Novi Sad, we detected no wrongly assigned citation coauthorships, estimating precision close to 100%. As for recall, we can only state that the detected 24,262 coauthorships seems a reasonable number, considering that authors whose forms are not included in the collection are not taken into account, and that the figure exceeds our initial intuitive expectations.

3 THE IST WORLD PORTAL

IST World [1,2] system allows integration of different data sources into a common database. It currently includes data on research publication and RTD projects from several European countries using different languages. In this paper we describe the process of extending the database on an example of incorporating data on Serbian research publications. The IST World portal enables the following functionalities:

Data integration – the main goal is to integrate multiple data sources with similar but different structure into one data structure existing on several levels.

Central data structure – the central data base is organized in a big social network (graph structure) which is organized on several levels: countries, organizations, departments and individuals.

Cross Language Technologies – the data collected in the project is multilingual meaning it is possible to compare the documents written in different languages and be able to identify (despite different languages) the similarity of their contents.

Text mining – enables different analysis of textual data including multi dimensional scaling and latent semantic analysis.

Link analysis – enables community identification and analysis of temporal networks.

Visualization – enables fast graph drawing techniques developed recently at JSI and some text visualization techniques developed in collaboration with Microsoft Research and at SEKT project.

4 SERBIAN RTD ANALYSIS ILUSTRATION

Once the data is imported we can perform any of the analysis supported by the IST World portal. This section describes an example collaboration and competence analysis of Serbian researchers and organizations.

Collaboration. Figure 1a depicts the collaboration diagram of *organizations*, where weighed arcs represent the number of publications mutually coauthored by their affiliates. The diagram is configured to show only the strongest 10% of the arcs, allowing us to get an overview of the most important connections between organizations (at the research level). By far the strongest bond (512 publications) is between the Faculty of Agriculture (*Poljoprivredni fakultet*) and the Institute of Field and Vegetable Crops (*Institut za ratarstvo i povrtarstvo*), which is in tune with Vojvodina being a highly agricultural region with a long tradition of research in this area. The Faculty of Agriculture also has strong ties with the Veterinary Institute (naturally), and also with the Faculty of Science (*Prirodno-matematički fakultet*), a possible result of many graduates of the latter being employed by the Faculty of Agriculture, according to the authors' experience. The Faculty of Science, on the other hand, also collaborates strongly with the Faculty of Technical Sciences (*Fakultet tehničkih nauka*) via its Department of Physics and the Department of Mathematics and Informatics; with the Faculty of Technology (*Tehnološki fakultet*) through its Department of Chemistry; and with the Faculty of Medicine through the Departments of Biology and Chemistry. The most surprising link on the diagram, between Faculties of Medicine and Philosophy, upon closer inspection turned out to be due to an error in the original data: the faculties employ two different researchers with the same first and last name (Slobodan Pavlović), and in the collection they were mistakenly represented by identical forms, resulting in the extractor perfectly matching all 148 publications.

Figure 1b shows the collaboration diagram of *researchers*, where weighed arcs denote the number of coauthored publications. The graph is configured to include only the top 1% of arcs, revealing the cream of the Vojvodinian scientific community. The subgraph of five mutually cooperating researchers is the group of cardiologists and cardio surgeons gathered around Ninoslav Radovanović, the recently retired chief of the Institute of Cardiovascular Diseases in Sremska Kamenica (all researchers are also affiliated with the Faculty of Medicine). The most prominent cooperations also include Ratko Nikolić – Timofej Furman from the Faculty of Agriculture, Spasenija Milanović – Marijana Carić and Jasna Gvozdenović – Vera Lazić from the Faculty of Technology, and the already mentioned Slobodan Pavlović – Slobodan Pavlović, who are actually the same person.

Competence. The competence diagram consisting of researchers form the Faculty of Science is shown in Figure 2. The intention of the diagram is to cluster researchers around colored regions representing competencies, which are labeled by terms extracted from the titles of publications. Unfortunately, the current version of the extractor does not attempt to extract titles from the citation data, and thus whole citations are used instead of titles for producing competence labels. Nevertheless, the introduction of such noise words did lead to an interesting effect: researchers are now being placed around names of their prominent colleagues (Matavulj, Škrinjar, Budimac, Dalmacija...), who now represent another way of labeling competence.

5 CONCLUSIONS AND FUTURE WORK

It is important to note that, with the current state of the extracted data, no collaboration or competence analysis can be considered "the whole truth," since we are unable to give firm guarantees on the recall of extraction. Despite this, the general relationships that *are* observed between organizations and researchers do generally comply with our general picture of Serbian research, suggesting that precision of extraction is adequate.

Currently, the extractor processes only whole citations, with no attempts to isolate the author list, title, journal or conference name, publication date and similar information. Although the task is difficult, when considering the variety of used citation conventions and languages, it may be worthwhile to attempt in future work, because it would allow expressing many other relations beside coauthorship, for instance: being in the same conference/journal issue, same conference stream/journal, or similar conf./ journals [6]. It would also permit the generation of competence maps which are cleansed of noise words appearing in the whole citations. Performing a more comprehensive study in order to tune the similarity threshold and improve precision and recall of citation matching is another, more immediate area for exploration, as well as improving citation recognition by implementing more parsing schemes and Cyrillic letter conversion.

Besides illustrating how knowledge technologies can help gain interesting and useful insights into the relationships between people and organizations, we have also seen that they can help locate and eliminate errors in original data. Such cooperation of extraction and analysis will be beneficial to both phases in the process of further extending the IST-World database.

Acknowledgments

This work was supported by Slovenian Research Agency, Serbian Ministry of Science and Environmental Protection, and EC under IST World (FP6-2004-IST-3-015823).

References

[1] Jörg, B., Jermol M., Uszkoreit H., Grobelnik M., Ferlež J.; Analytic Information Services for the European Research Area. Proc.: eChallenges2006, Barcelona.

[2] Jörg B., Ferlež J., Grabczewski E, Jermol M. IST World: European RTD Information and Service Portal. 8th Int. Conf. on Current Research Information Systems: Enabling Interaction and Quality: Beyond the Hanseatic League (CRIS 2006), Norway, 2006.

[3] CERIF: the Common European Research Information Format. http://cordis.europa.eu/cerif/, 2000

[4] Fortuna B, Mladenič D, Grobelnik M. Visualization of text document corpus. Informatica journal, 29(4):497–502, 2005.

[5] Grobelnik, M., Mladenic, D. Analysis of a database of research projects using text mining and link analysis. In Data mining and decision support : integration and collaboration, Boston; Dordrecht; London: Kluwer Academic Publishers, 157-166, 2003.

[6] Klink, S. et al. Analysing social networks within biblio-graphical data. Proc. of DEXA06, Krakow, Poland, 2006.

[7] White, S. How to strike a match. http://www.devarticles. com/c/a/Development-Cycles/How-to-Strike-a-Match/, 2004

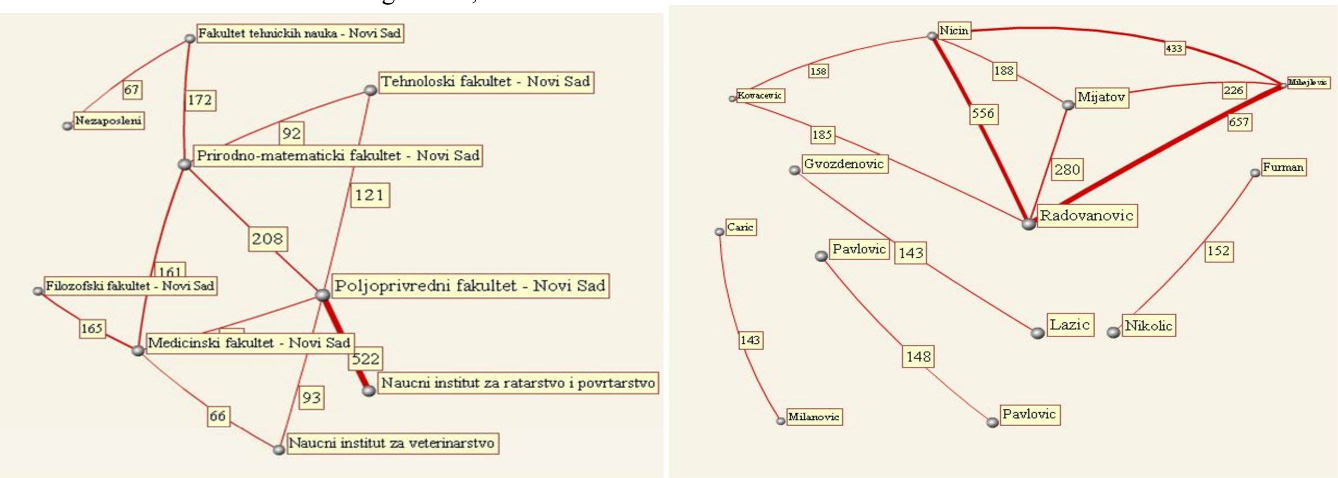
Fakultet tehnickih nauka - Novi Sad
67
172
Nezaposleni
Tehnoloski fakultet - Novi Sad
92
Prirodno-matematicki fakultet - Novi Sad
121
208
161
Filozofski fakultet - Novi Sad
165
Poljoprivredni fakultet - Novi Sad
Medicinski fakultet - Novi Sad
522
Naucni institut za ratarstvo i povrtarstvo
93
66
Naucni institut za veterinarstvo
Nicin
433
158
188
Kovacevic
Mijatov
226
556
657
185
Gvozdenovic
280
Furman
Radovanovic
Caric
Pavlovic
143
152
Lazic
Nikolic
143
148
Milanovic
Pavlovic

Figure 1: *Collaboration diagrams of Serbian organizations – left* (a) *and researchers – right* (b).

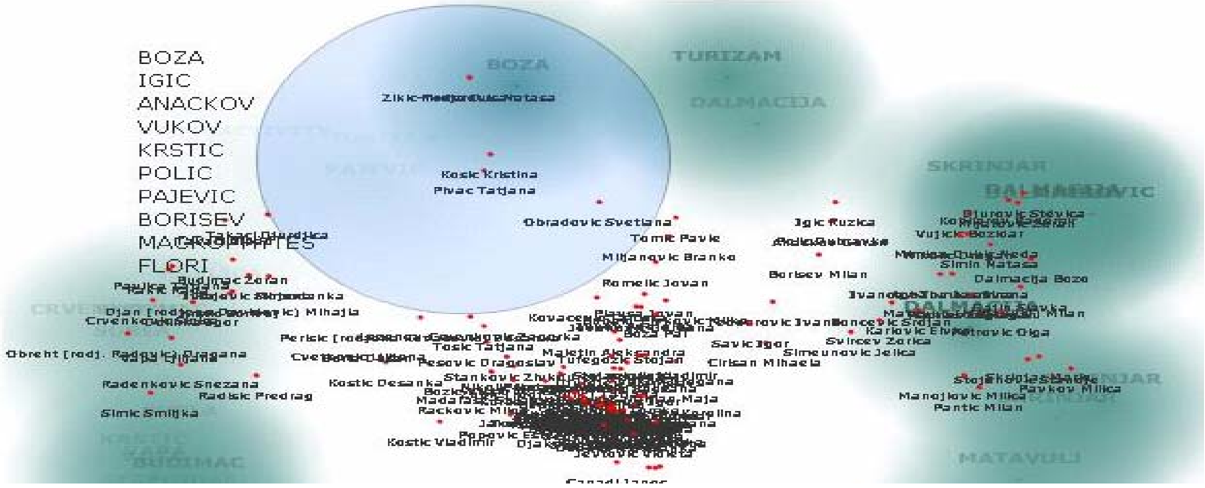

Figure 2: *Portion of a competence diagram of Serbian researches*