Novi Sad J. Math.
Vol. 37, No. 2, 2007, 161-180

MINING AND VISUALIZING SCIENTIFIC PUBLICATION DATA FROM VOJVODINA

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

Abstract. This paper describes the process of information extraction and analysis of a document collection representing publication references of the majority of researchers in the Serbian province of Vojvodina. After an introduction to techniques for text mining and visualization, a review of the IST World system is given, which provides online services for search, navigation and visualization of bibliographic data collected from multiple sources. The process adopted for information extraction consists of reference recognition and coauthorship detection. Evaluation on a representative subset of the data demonstrates good performance of the extraction as measured by precision and recall. An example analysis by a domain expert using IST-World system shows how insights can be gained into the collaboration and competence of authors using the visualization functionalities of IST World.

AMS Mathematics Subject Classification (2000): 68U15, 68U35, 91C15, 91C30

Key words and phrases: Text mining, information extraction, bibliography processing, link analysis, text visualization

1. Introduction

Most European countries collect and store their research information in national research and technology development (RTD) repositories. This information is often spread across several regional or local repositories that are realized with proprietary encoding and structure. It is very difficult to get additional information value out of multiple collections of RTD information, spread over several individual sources.

The IST World portal [11, 12] 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 [9], applying text mining and link analysis techniques to analyze European research space based on the RTD projects. The IST World portal contains information about RTD actors on the local, national and European level (harvested from several existing databases and with Web mining

[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

techniques), such as persons, research groups, organizations, projects, and their experience and expertise. By integrating information from various sources into the integrated database that is based on the CERIF standard, IST World offers possibilities to discover existing and potential competency and collaboration networks.

On the other hand, 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 references of all his/her authored publications, among other data. The gathered information is available in an unmodified form on the Web site of the Secretariat: `http://apv-nauka.ns.ac.yu`. Notable properties of the data at this stage are incompleteness (unfortunately, many researchers did not submit their forms) and diversity of approaches to giving references, permitted by information being entered in free text format.

This paper demonstrates how to extract references and coauthorship relations from the collection of documents describing Vojvodinian researchers, extending our previous work [20] with a formal experimental evaluation of the accuracy of extraction and presenting several improvements initiated by the evaluation. Next, it is shown how to obtain a collaboration graph expressing coauthorship of papers between researchers and organizations by employing the functionalities of the IST World portal. Furthermore, competencies of the authors and organizations are analyzed using competence map based knowledge technologies [7].

The rest of the paper is organized as follows. In Section 2, essential text mining and visualization techniques are presented, while Section 3 summarizes the functionalities of the IST World portal. Section 4 describes the process of information extraction from the researcher document collection and illustrates the possibilities for analysis using visualization of collaboration and competence. The last section presents the conclusions and guidelines for possible future work.

2. Text Mining and Visualization

This section will review the techniques for mining and visualizing textual data relevant to the work described in the paper. After presenting methods for representing documents (Section 2.1) and calculating their similarity (Section 2.2), visualization techniques based on dimensionality reduction are described in Section 2.3.

2.1. Document Representation

Visualization of document collection as applied in this work does not require the full richness of the content of natural language texts. We transform documents to a simpler and more manageable form. The most widely used approach is the bag-of-words (BOW) representation, where word order is completely discarded from a document.

Let W be the *dictionary* – the set of all terms (words) that occur at least once in a collection of documents D. The BOW representation of a document d_n is a vector of weights $(w_{1n},\ldots,w_{|\mathrm{W}|n})$. In the simplest case, the weight $w_{in} \in \{0, 1\}$ denotes the presence or absence of a particular term in a document. More commonly, w_{in} represents the frequency of the ith word in the nth document. Normalization can be employed to scale the term frequencies to values between 0 and 1, accounting for differences in the lengths of documents. The natural logarithm function can also be applied to term frequencies, replacing the weights with $\log(1 + w_{in})$. The *inverse document frequency* transform is defined as $\log(|D|/\mathrm{docfreq}(\mathrm{D}, i))$, where $\mathrm{docfreq}(\mathrm{D}, i)$ is the number of documents from D the ith word occurs in. It can be used by itself, or multiplied with term frequency to yield the popular TFIDF representation.

Besides words, *n-grams* may also be used as terms in the vector document representation. But, two very different notions have been referred to as "n-grams" in the literature. The first are *phrases*, as sequences of n words; this meaning was adopted by the Statistical Natural Language Processing community [18]. The other notion are n-grams as sequences of *characters*.

N-grams as phrases can be viewed as a generalization of words, for 1-grams *are* words, so 2-grams up to 5-grams are usually used to *enrich* the BOW representation, rather than on their own. The main problem is sheer magnitude – the number of n-grams grows exponentially with n – therefore many strategies for efficient generation of a useful set of n-grams have been developed. One such algorithm, presented by Mladenić [19], iterates over n, generating all possible n-grams from known $(n - 1)$-grams, immediately discarding all which appear too infrequently in the document set.

N-grams as sequences of characters are used *instead* of words in the BOW representation, which means that only is the word order lost, but words themselves are not preserved. Nevertheless, character n-grams proved very useful in situations with grammatical and typographical errors in documents, in handling highly inflected languages, and are also an effective way to achieve language independence [2, 17].

2.2. Distance Measures

For our document visualization we would like to have a visual representation that reflects the similarity of documents, placing similar documents closer to each other in the visual representation. This is achieved by an appropriate distance measure that is applied on documents. Since documents in the BOW representation are vectors, vector metrics may be used to express the distance (and, reciprocally, similarity) between two documents. We will review the commonly used L_k, cosine and Tanimoto vector distance/similarity measures. Comprehensive overviews of metrics are given by Chapman [4] and Kohonen [16], while Cohen et al. [5] provide a thorough experimental comparison of string distance metrics for name-matching.

$\mathrm{L_k}$ distance. For two document vectors d_n and d_m, the L_k distance (also known as the Minkowski metric) is defined as the k-norm of their difference:

$$\mathrm{L}_k(d_n, d_m) = \|d_n - d_m\|_k = \left(\sum_{i=1}^{|\mathrm{W}|} (w_{in} - w_{im})^k \right)^{1/k} .$$

When $k = 1$, the metric is also referred to as Manhattan distance, while L_2 is the well-known Euclidean distance.

It is a known fact that the L_k distance tends to deform with high numbers of dimensions [1], with "high" starting from as low as 20. In textual domains, where dimensions are often counted in tens or hundreds of thousands, this drawback may be very pronounced. Therefore, the cosine measure is usually employed, which has proven to be particularly suited for application on text.

Cosine similarity. For two document vectors d_n and d_m, the cosine similarity is defined as

$$\mathrm{CS}(d_n, d_m) = \frac{\langle d_n, d_m \rangle}{\|d_n\| \cdot \|d_m\|} = \frac{\sum_{i=1}^{|\mathrm{W}|} w_{in} w_{im}}{\sqrt{\sum_{i=1}^{|\mathrm{W}|} w_{in}^2} \cdot \sqrt{\sum_{i=1}^{|\mathrm{W}|} w_{im}^2}} .$$

The measure expresses the cosine of the angle between the two vectors in their $|\mathrm{W}|$-dimensional space. A smaller angle means greater similarity between two document vectors and, presumably, greater similarity between the semantics of their contents.

Tanimoto similarity. The Tanimoto similarity (also referred to as Jaccard similarity) between vectors d_n and d_m may be defined as

$$\mathrm{TS}(d_n, d_m) = \frac{\langle d_n, d_m \rangle}{\|d_n\|^2 + \|d_m\|^2 - \langle d_n, d_m \rangle} = \frac{\sum_{i=1}^{|\mathrm{W}|} w_{in} w_{im}}{\sum_{i=1}^{|\mathrm{W}|} w_{in}^2 + \sum_{i=1}^{|W|} w_{im}^2 - \sum_{i=1}^{|\mathrm{W}|} w_{in} w_{im}} ,$$

which is reminiscent of the cosine measure. But, its origins in the comparison of *sets* allow a different definition. Let N and M be the sets of words present in documents d_n and d_m. The Tanimoto metric then represents the ratio between the number of words shared by d_n and d_m and the number of all words appearing in both documents:

$$\mathrm{TS}(\mathrm{N}, \mathrm{M}) = \frac{|\mathrm{N} \cap \mathrm{M}|}{|\mathrm{N} \cup \mathrm{M}|} = \frac{|\mathrm{N} \cap \mathrm{M}|}{|\mathrm{N}| + |\mathrm{M}| - |\mathrm{N} \cap \mathrm{M}|} .$$

This definition is applicable on *multisets* as well.

The Tanimoto distance is often used in situations where a common dictionary is not available, or not necessary for solving the problem at hand. In the BOW setting, this would mean that the dictionary is formed online only from the two documents being compared.

2.3. Visualization Techniques Based on Dimensionality Reduction

Text visualization is useful in situations where insight needs to be gained into the structure and underlying patterns in large document collections. Several approaches and techniques are available, e.g. showing the similarity structure of documents in the collection, showing time line of topic development through time, and showing frequent words and phrases relationships between them [8]. This section focuses on techniques that utilize the similarity structure of documents, and presents visualization possibilities offered dimensionality reduction techniques, i.e. latent semantic indexing coupled with multidimensional scaling.

By using term-document matrix manipulations, *latent semantic indexing* (LSI) techniques transform document vectors to a lower dimensional space, while preserving (and making explicit) the correlations between terms, referred to as "latent semantics." Singular value decomposition (SVD) is the concrete linear algebra technique that is employed by LSI for matrix transformation. The resulting k new features (terms) are then used to represent original documents in the new space.

For the purposes of document visualization setting $k = 2$ when applying LSI gives poor results as all documents are described using only the two main LSI concepts which is usually not sufficient [7]. The Document Atlas system [7] proposes an alternative, employing multidimensional scaling (MDS) to reduce the k-dimensional space with $k > 2$ obtained by LSI down to two dimensions. MDS achieves this by placing the documents in two dimensions, at the same time minimizing an energy function, for instance

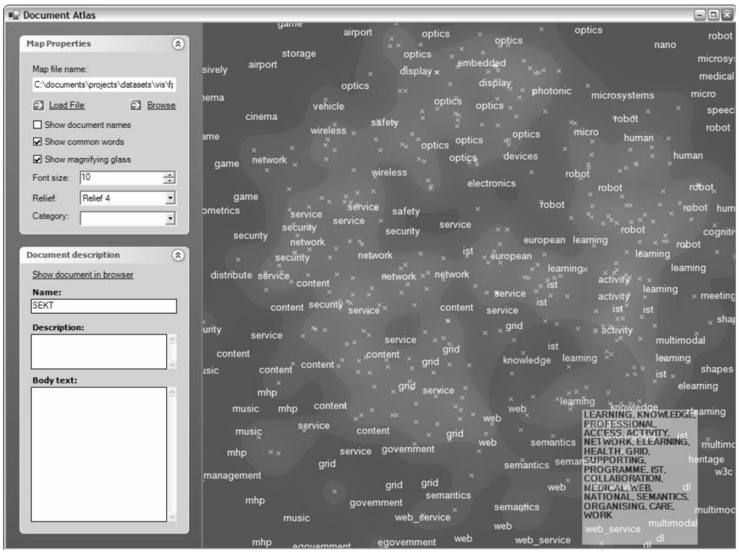
$$\mathrm{E} = \sum_{m \neq n} (\mathrm{L}_2(d_m, d_n) - \mathrm{L}_2(x_m, x_n))^2,$$

where d_m and d_n are documents in the "semantic space" obtained by LSI, and x_m and x_n are points in the two-dimensional space.

Figure 1 shows the Document Atlas visualization of a collection of documents describing the 6th framework European IST projects [6]. The landscape is generated using the density of points, where lighter areas denote bigger density, and hence "height." Individual documents are labeled by crosses, while most common words are placed on the map at randomly chosen points. The commonality of a word for a given point on the map is calculated by averaging TFIDF vectors of documents which appear within a circle of a certain radius originating at the point. The system offers a more detailed view of common words that can be obtained by using the mouse to move the dark circle to a desired area on the map.

3. The IST World Portal

The IST World system [11, 12] 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.



Document Atlas

Map Properties

Map file name: C:\documents\projects\datasets\vis\f

Load File Browse

□ Show document names

☑ Show common words

☑ Show magnifying glass

Font size: 10

Relief: Relief 4

Category:

Document description

Show document in browser

Name: SEKT

Description:

Body text:

Figure 1: Document Atlas visualization of European IST projects from the 6th framework

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 that 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 multidimensional scaling and latent semantic analysis

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

Visualization – enables fast graph drawing and text visualization techniques.

3.1. IST World Portal Services

The following IST World portal services built on top of a CERIF-based data repository are available at the time of writing this paper: (1) complex search and navigation functionalities to retrieve relevant information from the data repository, (2) automated analytical methods and visualization techniques to present the results. The analytic processes depend on the predefined sets of information. First, in a selection step the entities (organizations, experts, publications, projects) or subsets of the entities that represent the target of interest are to be specified by making use of available search and navigation functionalities. Second, in an analysis step one of the analytical methods is applied upon the retrieved set of entities. The results are presented by advanced visualization techniques.

Search and navigation. Full text search is currently employed for all of the IST World entities, simple and advanced query templates are available for organizations, projects, experts and publications. A category-based navigation interface is provided, based on the Science part of the dmoz taxonomy. Furthermore, the "Partner finder" service allows searching for experts and organizations whose fields of expertise, as derived from the data in the repository, match best a given set of keywords.

Automated analysis and visualization techniques. A variety of tools are available in the IST World portal to analyze and visualize search and navigation results. Automated methods provide insight into current, past and partly to the future research activities based on subset selections from the IST World information repository. The following functionalities are implemented:

- community identification by using collaboration graphs to visualize the cooperation relationships between entities selected by a query,
- expertise identification via competence diagrams built from selected entities,
- partner finder, which allows searching for experts and organizations whose fields of expertise, as derived from the data in the repository, match best a given set of keywords,
- trend identification, utilizing time information within the repository to indicate the trends in the work of the selected entities.

Portal enhancements. Some of the services presented are only available for registered users or members. In order to organize and manage the different views and access rights, the IST World portal will employ techniques for social trust networking. As IST World intends to take advantage of a warm welcome, a multilingual user interface in several languages will be made available.

3.2. IST World Repository

The services of the IST World portal are offered on top of the IST World repository and thus depend on data input. Data is provided in specified formats [10], by the community via Web forms and from the Web by automated crawling. The main sources of data are public databases.

Conceptual baseline. During the initial phase of the project the portal uses only the relational model, implemented in a conventional RDBMS. Currently, the conceptual model (or ontology) is not operationally involved in the portal and is used only as a design guide and a base for developing a proper expertise modeling schema. Later, the ontology will be used in addition to a semantic repository for properly integrating the RDBMS data [14].

Technological baseline. The Common European Research Information Format (CERIF) [3] was formerly developed under the coordination of the European Commission to harmonize national Current Research Information Systems (CRISes) within Europe. CERIF is an open set of guidelines prepared to deal with research information systems.

The IST World implementation uses a relevant subset of CERIF entities and their relationships and follows the current CERIF practices in extending the data model. Database creation was facilitated by the use of SQL scripts provided by the CERIF task group and extended with additional entities and relationships.

To meet the necessary storage requirements for the repository and for the offered portal functionalities, extensions to the CERIF model were necessary to cover the following requirements:

- to support the display of information on trends and the prediction of the state of European and national research activities,
- to support computer aided social networking,
- to allow user authentication,
- to allow source identification,
- to store the content of publication documents and not only their metadata.

The trends detection and prediction functionalities require that the extended CERIF data model stores additional data and meta-data on scientific publications and information. Moreover, the functionality to support the provision of computer aided social networking enables users to search and collaborate with existing social networks and requires addressing the issues of privacy, trust and the interests of network members. All mentioned issues were input to the CERIF extensions for the IST World CERIF-based data model.

4. Information Extraction and Analysis

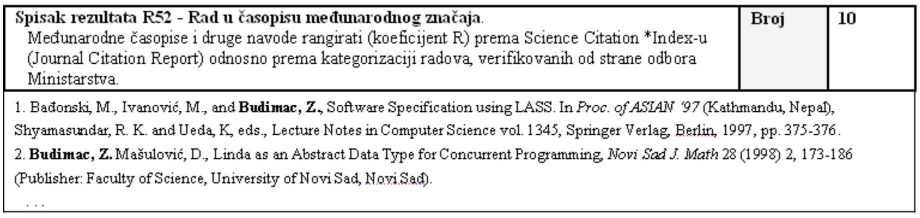
This section presents the steps taken to extract and analyze information about references contained in a collection of semi-structured documents providing a research bibliography. After describing the data in Section 4.1, the process of reference recognition and coauthorship detection is presented in Section 4.2, together with the evaluation of their performance on a representative subset of the data. Section 4.3 demonstrates the possibilities for visualization and analysis using collaboration and competence charts from IST World.

4.1. 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. Despite the large number of entries considering the size of the Vojvodinian region, the collection is still in an early stage of development. Many researchers have not yet submitted their data and some information in the collection appears to be out of date, which should be remedied in part by the planned future updates of the database.

The number of existing entries in the collection still made the task of manually extracting bibliographical data infeasible. 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 references, and save the data in the form of CERIF compliant XML files, to enable quick import of the data into the existing IST World relational database. Furthermore, the extractor compares references 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 Rxx, 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 references were usually entered enclosed in isolated paragraphs or numbered lists. The collection includes references written in more than five natural languages, the most prominent being Serbian, English, Hungarian, Slovak, and Romanian.



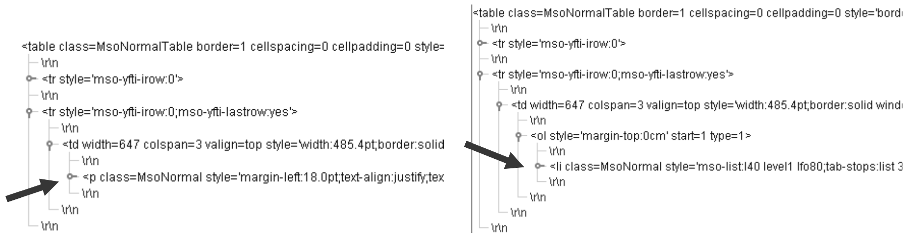
Spisak rezultata R52 - Rad u časopisu međunarodnog značaja. Međunarodne časopise i druge navode rangirati (koeficijent R) prema Science Citation *Index-u (Journal Citation Report) odnosno prema kategorizaciji radova, verifikovanih od strane odbora Ministarstva.	**Broj**	**10**
1. Bađonski, M., Ivanović,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

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

4.2. Information Extraction

Version 2.0.2 of the extractor is able to isolate a total of 101,672 bibliographic units from current data, and detect 24,262 duplicate references (which correspond to coauthorships – a paper appearing in n researchers' forms can have a maximum of $n - 1$ detected duplicates). This makes the total number of references in the database 77,410. 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 reference data from MS Word documents.

Reference recognition. From the limited number of options for accessing the content of MS Word documents from outside programs, we found it most convenient to bulk convert all documents to HTML format via a Word macro, 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 references of interest, as described in Section 4.1. Further extraction of references 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 references were "read out" from fixed positions in the DOM trees of <TABLE> tags, taking into account the two above possibilities. The DOM trees with the indicated positions are illustrated in Figure 2.



```
<table class=MsoNormalTable border=1 cellspacing=0 cellpadding=0 style=
  \r\n
  <tr style='mso-yfti-irow:0'>
  \r\n
  <tr style='mso-yfti-irow:0;mso-yfti-lastrow:yes'>
    \r\n
    <td width=647 colspan=3 valign=top style='width:485.4pt;border:solid
      \r\n
      <p class=MsoNormal style='margin-left:18.0pt;text-align:justify;tex
      \r\n
    \r\n
  \r\n
```

(a)

```
<table class=MsoNormalTable border=1 cellspacing=0 cellpadding=0 style='bord
  \r\n
  <tr style='mso-yfti-irow:0'>
  \r\n
  <tr style='mso-yfti-irow:0;mso-yfti-lastrow:yes'>
    \r\n
    <td width=647 colspan=3 valign=top style='width:485.4pt;border:solid wind
      \r\n
      <ol style='margin-top:0cm' start=1 type=1>
        \r\n
        <li class=MsoNormal style='mso-list:l40 level1 lfo80;tab-stops:list 3
        \r\n
      \r\n
    \r\n
  \r\n
```

(b)

Figure 2: Possible positions of references in the HTML DOM tree

Somewhat surprisingly, this simple scheme turned out to be rather effective at retrieving strings containing valid references. After observing the isolated references, we removed from the collection the 59 forms which were obviously not parsed correctly within this scheme (the references were either divided up into several, or lumped together). The forms which were filled in using the Cyrillic alphabet were also removed (62 in total), since it was decided 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%.

In order to evaluate more precisely the success of reference recognition, we examined the results of extraction on 42 forms submitted by researchers from the Department of Mathematics and Informatics. We counted the *true positives* (TP), which in this context is the number of extracted strings that are truly bibliographic references, the *false positives* (FP) – the number of strings which are only partial references or not references at all, and the real number of references in the submitted forms. The classical measures from information retrieval, *precision* and *recall*, may now be expressed as TP/(TP + FP) and TP/Real, respectively.

	TP	**FP**	**Real**	**Precision**	**Recall**
True estimate	1082	23	1732	97.92%	62.47%
9 forms removed	1033	21	1158	98.01%	89.21%

Table 2: Evaluation of reference recognition for extractor v2.0.2

TP	**FP**	**Real**	**Precision**	**Recall**
1580	5	1732	99.68%	91.22%

Table 3: Evaluation of reference recognition for extractor v2.0.3

The results of the evaluation, summarized in Table 2, revealed a recall of only 62.47%. It was immediately evident that this was due to some forms being filled in an unexpected manner – references have sometimes been written in the header rows of the tables instead of the provided second row. After removing 9 such forms, recall jumped to 89.21%, meaning that a simple adjustment of the parsing scheme should considerably raise recall. On the other hand, precision was determined to be 97.92%, exceeding our initial estimate of 97% [20]. Also, it was observed that some tables were being selected for reference extraction when they should not have been, that way damaging precision.

Based on the above observations, in extractor v2.0.3 we introduced several enhancements:

- When references are enclosed in <P> tags, if the content of the <P> tag does not begin with an ordinal number it is concatenated with the previous reference. This fix ensured that references are no longer divided;
- The first row of tables is now being searched for references;
- Selection of tables was made more accurate.

Table 3 summarizes the evaluation of the new version of the extractor on identical data. It can be seen that both precision and recall are considerably improved. The reason for recall not being closer to 100% lies in specific parsing issues within certain forms and tables. We decided not to address these details in order to leave the evaluation unbiased: introducing fixes that solve problems specific to the chosen evaluation sample would have led to "overfitting" and producing overly optimistic estimates of precision and recall. The enhancements that *were* introduced into the extractor are general, in the sense of pertaining to all forms, not just the chosen evaluation dataset.

In summary, version 2.0.3 of the extractor isolates 110,394 references (about 9,000 more than v2.0.2) and detects 30,822 duplicates (about 6,500 more), making the total number of references in the database 79,572. Detection of duplicates is discussed next.

Coauthorship Detection. In order to calculate the similarity of two references, with the intention to determine a coauthorship relation, the extractor uses an optimized version of the algorithm described in [21], which calculates

TP	FP	Real	Real (no lang.)	Precision	Recall	Recall (no lang.)
583	26	625	612	95.73%	93.28%	95.26%

Table 4: Evaluation of coauthorship detection for extractor v2.0.3

the value of the Tanimoto similarity metric over the space of character 2-grams. The algorithm computes 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 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 "noise" in manually entered reference data: typographical errors, different or missing information, various referencing conventions used (resulting in different ordering of reference information), etc. After parsing a researcher's form and extracting a list of references, every reference is compared to all references 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 reference does not exceed a predetermined similarity threshold (set at 0.63 after examining several test cases), the reference is entered as a new one into the database. Otherwise, a coauthorship relation is established, and the entry for the currently processed reference of the researcher is set to point to the reference already in the database.

Evaluation of coauthorship detection conducted on the same data as the evaluation of reference recognition is summarized in Table 4. Numbers for true positives, false positives and the real count actually represent *authorship relations* of researchers to multi-authored papers – papers with two detected authors are counted twice, with three authors three times, etc. Precision is 95.73%, lowered from the perfect score by wrong assignments of one author to a two-author paper, which happened among authors with same last names and similar scientific interests. Undetected coauthorships arose mainly for three reasons: (1) one author was supplying much less information within a reference than another, that way lowering the calculated string similarity, (2) an author changed her name, or (3) authors wrote references in different languages. Information which could help solve case (2) was usually not available within the forms. Since situation (3) requires a sophisticated solution, recall was calculated separately for the two cases when different language references are considered equal and not equal in reality, resulting in recall values of 93.28% and 95.26%, respectively.

4.3. Example Analysis

Once the data is imported, any kind of analysis supported by the IST World portal can be performed. This section describes an example of collaboration and competence analysis of Vojvodinian researchers and organizations.

Collaboration. Figure 3 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 (778 publications) is between the Faculty of Agriculture (*Poljoprivredni fakultet*) and the Institute of Field and Vegetable Crops (*Naučni 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. Not surprisingly, the Faculty of Agriculture also has strong ties with the Veterinary Institute (98 publications), but also with the Faculty of Science (*Prirodno-matematički fakultet*, 486 publications), which is a result of cooperation with the Department of Biology of the Faculty of Science, and also a result of many graduates of the Faculty of Science being employed by the Faculty of Agriculture. 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 (396 publications); with the Faculty of Technology (*Tehnološki fakultet*) through its Department of Chemistry (93 publications); and with the Faculty of Medicine through the Departments of Biology and Chemistry (324 publications). 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.

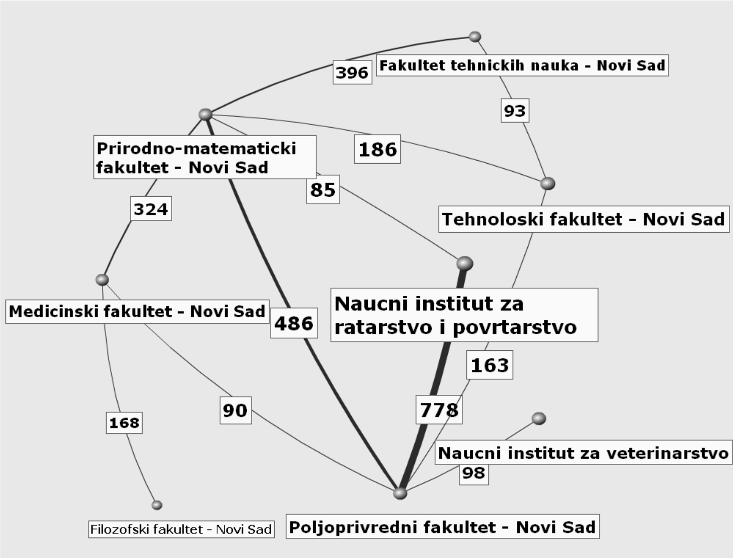

396
Fakultet tehnickih nauka - Novi Sad
93
Prirodno-matematicki fakultet - Novi Sad
186
85
324
Tehnoloski fakultet - Novi Sad
Medicinski fakultet - Novi Sad
486
Naucni institut za ratarstvo i povrtarstvo
163
168
90
778
Naucni institut za veterinarstvo
98
Filozofski fakultet - Novi Sad
Poljoprivredni fakultet - Novi Sad

Figure 3: Collaboration diagram of Vojvodinian organizations (top 10%)

Figure 4 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 majority of publications from this group of researchers are in abstract form, which explains the seemingly impossible numbers of joint papers. Most prominent cooperations also include Ratko Nikolić – Timofej Furman from the Faculty of Agriculture (151 joint publications), Spasenija Milanović – Marijana Carić (143 joint publications) and Jasna Gvozdenović – Vera Lazić (142 joint publications) from the Faculty of Technology, and the already mentioned Slobodan Pavlović – Slobodan Pavlović.

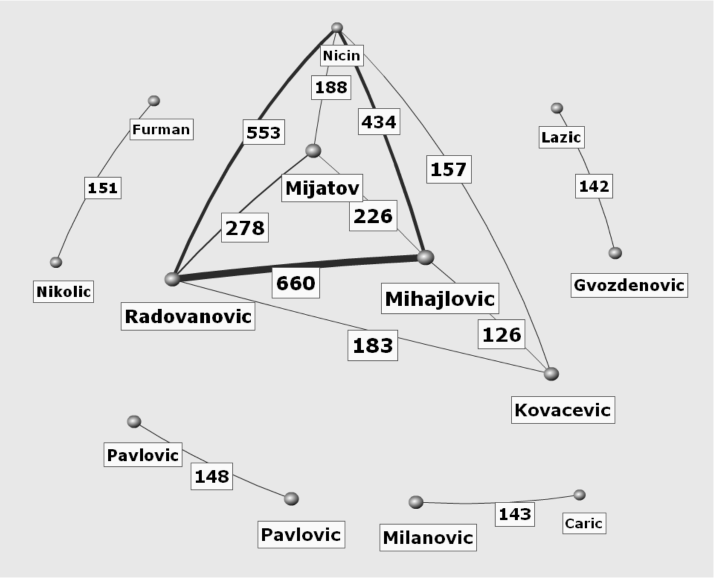


Figure 4: Collaboration diagram of Vojvodinian researchers (top 1%)

When observing collaboration between researchers, we can restrict the view to only a subset of researchers, e.g. researchers from the selected organization. For instance, the collaboration of all researchers from the Department of Mathematics and Informatics (DMI) of the Faculty of Science, University of Novi Sad, is depicted in Fig. 5. The strongest cooperation is exhibited between professors Zoran Budimac and Mirjana Ivanović, and professors Miloš Racković and Dušan Surla, who represent the "backbones" of the two Informatics chairs at the Department – the Chair of Computer Science, and the Chair of Information Systems, respectively. All other members of both chairs collaborate directly with at least one member of the backbone, except for prof. Dragan Mašulović from the CS chair and Đorđe Herceg from the IS chair, while inter-chair cooperation is expressed by two joint papers of Zoran Budimac and Dušan Surla.

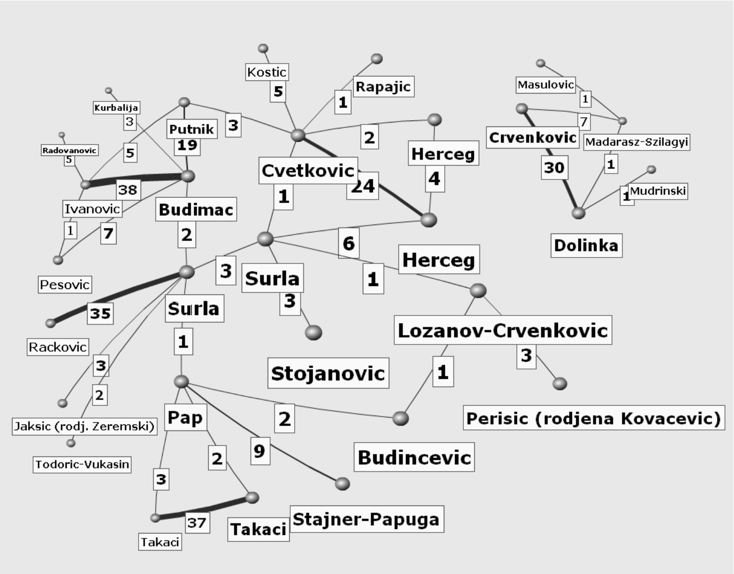


Figure 5: Collaboration diagram of researchers from the Department of Mathematics and Informatics

Unfortunately, members of Mathematics chairs, who form the majority of the Department, show more sparsely than in the diagram than they could. The Chair of Numerical Mathematics is represented by prof. Ljiljana Cvetković, prof. Dragoslav Herceg, prof. Katarina Surla and Vladimir Kostić; the Chair of General Algebra and Theoretical Computer Science by profs. Siniša Crvenković, Igor Dolinka and Rozália Madarász-Szilágyi, and assistant Nebojša Mudrinski; the Chair of Applied Analysis by profs. Endre Pap, Ivana Štajner-Papuga, Đurđica and Arpad Takači; the Chair of Mathematical Analysis, Probability and Diff. Equations by profs. Mirko Budinčević, Dušanka Perišić and Zagorka Lozanov-Crvenković; the Chair of Applied Algebra by profs. Branimir Šešelja and Andreja Tepavčević; and the Chair of Functional Analysis, Geometry and Topology by prof. Mirjana Stojanović. Extraction errors (i.e. lower recall) are only partly to blame for the fact that many members of the Department are missing from the diagram – a number of senior researchers have not submitted their data to the Secretariat for Science, which resulted not only in their exclusion from the diagram, but also in the exclusion of many young researchers who have a low number of publications coauthored only by their mentors.

The top 15% of collaboration between affiliates of the Faculty of Science is exhibited in Fig. 6. Completeness of data for the Department of Biology and the Department of Chemistry clearly shows its advantage – the tightly interconnected subgraphs of researchers are easily recognizable in the diagram, with inter-departmental cooperation expressed by many joint publications of Božo Dalmacija with Olga Petrović and Slavka Gajin. Separate components of the graph which represent members from the Department of Physics and the Department of Biology are also identifiable, corresponding to research within separate subfields of the two sciences.

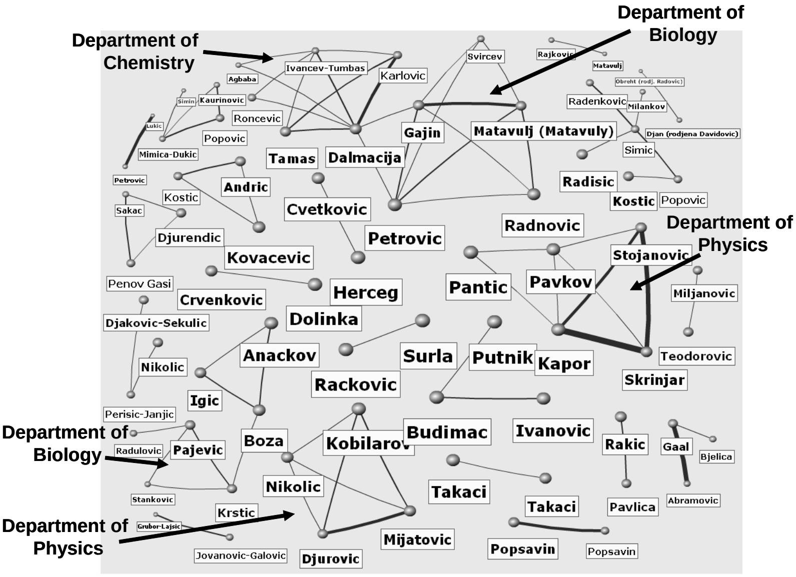
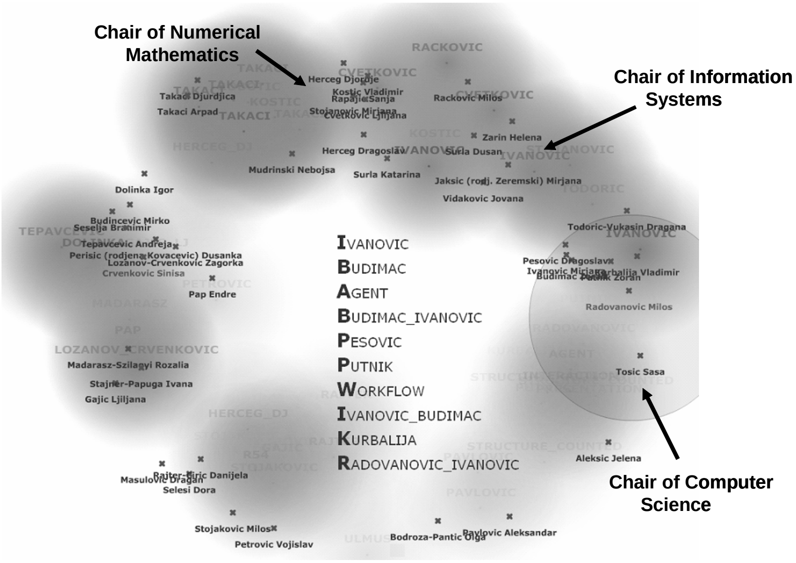


Figure 6: Collaboration diagram of researchers from the Faculty of Science (top 15%)

Competence. The competence diagram was constructed for researchers from the Department of Mathematics and Informatics of the Faculty of Science, as shown in Fig. 7. The intention of the diagram is to show how the researchers group around highlighted 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 reference data, and thus whole references 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 (Dolinka, Pap, Ivanović. . .), who now represent another way of labeling competence.

Despite the noise present in the data, the two Informatics chairs of the Department and the Chair of Numerical Mathematics were properly depicted by three clusters of researchers, and separated from each other and the rest of the Department. The only inconsistency with the real-world situation was the inclusion of Helena Zarin (a numerical mathematician) to the cluster representing Information Systems. We argue that for the proper clustering of members of the two chairs the inclusion of proper names in publication titles actually *assisted* the clustering by bringing actual coauthors closer together. In our


Chair of Numerical Mathematics

Chair of Information Systems

Chair of Computer Science

Figure 7: The competence diagram of researchers from the Department of Mathematics and Informatics

situation, where the chairs are formed by small groups of 5–10 people closely interconnected by coauthorship relations, this effect of the noise words is welcome. However, if chairs were larger, with several distinct groups of researchers who write papers together, but with all chair members actually doing research in a similar field (a reasonable assumption), the noise words would draw the different clusters coauthors further apart than they should (because their paper "titles" do not contain names from other groups of coauthors).

Although some researchers from the majority of the Mathematics part of the Department were correctly placed in proximity with each other in Fig. 7 (e.g. Ivana Štajner-Papuga and Ljiljana Gajić; Endre Pap, Dušanka Perišić and Zagorka Lozanov-Crvenković; Đurđica Takači and Arpad Takači), the overall clustering does not reflect the organizational structure of the mathematics chairs of the Department, nor the research interests of their members. Since the data for the two Informatics chairs and the Chair of Numerical Mathematics (which were clustered almost perfectly) was much more complete, we may assume that a future update of the data will enable a better clustering of the Department as a whole. This is corroborated by the fact that this same diagram, when generated using data output by v2.0.2 of the extractor, did not include a distinguishable cluster for the Chair of Numerical Mathematics. Extractor v2.0.3 was able to process the forms of the numerical mathematicians better than the previous version, providing enough data to enable a more successful clustering.

The competence diagram consisting of researchers form the whole Faculty of Science is shown in Fig. 8. We identified main areas of research in the diagram, but it can be noticed that many researchers were placed around the center of the picture, not close to any area. This is a consequence of the requirement for the clustering algorithm to work in the online Web setting of IST World – since the size of the Faculty of Science data is rather large, the algorithm was not given enough time to run and converge to an adequate clustering of researchers. Optimization of the clustering algorithm in order to make it capable of coping with larger datasets is an area for future work.

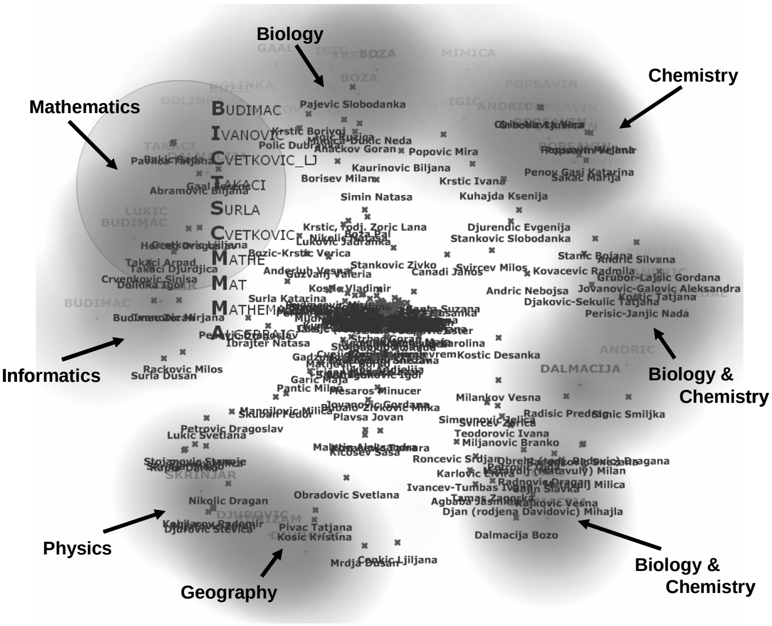


Figure 8: The competence diagram of researchers from the Faculty of Science

5. Conclusions and Future Work

It is important to note that, with the current state of the extracted data regarding missing forms and less than perfect precision and recall of extraction, no collaboration or competence analysis can be considered “the truth and the whole truth.” However, the observed relationships between organizations and researchers do comply with our own view of research in Vojvodina, suggesting that the values of precision and recall of extraction obtained on the evaluation dataset are good overall estimates.

Currently, the extractor processes only whole references, 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 referencing conventions and languages, it may be worthwhile to attempt this in future work, because it would allow expressing many other relations besides coauthorship, for instance: being in the same conference/journal issue, same conference stream/journal, or similar conferences/journals [15]. It would also permit the generation of competence maps which are cleansed of noise words appearing in the whole references. Another area for exploration is a more comprehensive study for tuning the similarity threshold, and investigating different similarity measures in different spaces, not only in the n-gram space, in order to improve the precision and recall of coauthorship detection.

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 can only be beneficial to both of the phases, as parts of the struggle to manage and make useful the vast amounts of bibliographic data available from many sources and in many forms.

Acknowledgments

This work was supported by Slovenian Research Agency, Serbian Ministry of Science, EC under IST World (FP6-2004-IST-3-015823), and NeOn Lifecycle Support for Networked Ontologies (IST-4-027595-IP).

References

[1] Aggarwal, C. C., Hinneburg, A., Keim, D. A., On the surprising behavior of distance metrics in high dimensional spaces. In: Proceedings of ICDT'01, 8th International Conference on Database Theory, volume 1973 of Lecture Notes in Computer Science, London, UK: Springer-Verlag 2001.

[2] Cavnar, W. B., Trenkle, J. M. N-gram-based text categorization. In: Proceedings of SDAIR'94, 3rd Annual Symposium on Document Analysis and Information Retrieval, pages 161–175, Las Vegas, USA 1994.

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

[4] Chapman, S., String similarity metrics for information integration.
`http://www.dcs.shef.ac.uk/~sam/stringmetrics.html` (2007).

[5] Cohen, W. W., Ravikumar, P. Fienberg, S. E. A comparison of string distance metrics for name-matching tasks. In: Proceedings of IJCAI'03 Workshop on Information Integration on the Web (IIWeb'03), Acapulco, Mexico 2003.

[6] CORDIS FP6: Home Page `http://cordis.europa.eu/fp6/` (2005).

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

[8] Grobelnik, M. Mladenic, D., Efficient visualization of large text corpora. In: Proceedings of the 7th TELRI Seminar, Dubrovnik, Croatia 2002.

[9] 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, pp. 157–166, Boston, Dordrecht, London: Kluwer Academic Publishers 2003.

[10] Jörg, B., Public IST World Deliverable – 3.1 Data import/export specification as XML Schemata 2005.

[11] Jörg, B., Jermol M., Uszkoreit, H., Grobelnik, M., Ferlež, J., Analytic Information Services for the European Research Area. In: Proceedings of eChallenges e-2006 Conference, Barcelona 2006.

[12] Jörg, B., Ferlež, J., Grabczewski, E., Jermol, M., IST World: European RTD Information and Service Portal. In: Proceedings of CRIS'06, 8th International Conference on Current Research Information Systems: Enabling Interaction and Quality: Beyond the Hanseatic League, Norway 2006.

[13] Kaski, S., Dimensionality reduction by random mapping: Fast similarity computation for clustering. In: Proceedings of IJCNN'98, International Joint Conference on Neural Networks, Piscataway, NJ (1998), 413–418.

[14] Kiryakov, A., Grabczewski, E., Ferlež, J., Uszkoreit, H., Jörg, B., Public IST World Deliverable 1.1 – Definition of the Central Data Structure (2005).

[15] Klink, S. et al., Analysing social networks within bibliographical data. In:Proceedings of DEXA'06, 17th International Conference on Database and Expert Systems Applications, volume 4080 of Lecture Notes in Computer Science, pp. 489–498, Krakow, Poland: Springer-Verlag 2006.

[16] Kohonen, T., Self-Organizing Maps. Third edition, Springer-Verlag 2001.

[17] Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., Watkins C., Instance-based learning algorithms. Journal of Machine Learning Research 2 (2002), 419–444.

[18] Manning, C. D., Schütze, H., Foundations of Statistical Natural Language Processing. MIT Press 1999.

[19] Mladenić, D., Machine Learning on non-homogenous, distributed text data. PhD thesis, University of Ljubljana, Slovenia 1998.

[20] Radovanović, M., Ferlež, J., Mladenič, D., Grobelnik, M., Ivanović, M. Extending the IST-World database with Serbian research publications. In: Proceedings of IS 2006, 9th International Multiconference on Information Society, pp. 251–254, Ljubljana, Slovenia 2006.

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

Received by the editors July 16, 2007