UNIVERZA V LJUBLJANI

FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Jure Ferlež

Metode analize podatkov o raziskovalni dejavnosti na primeru aplikacije IST World

MAGISTRSKO DELO

Ljubljana, 2007

UNIVERZA V LJUBLJANI

FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Jure Ferlež

Metode analize podatkov o raziskovalni dejavnosti na primeru aplikacije IST World

MAGISTRSKO DELO

Mentor: akad. prof. dr. Ivan Bratko
Somentor: doc. dr. Dunja Mladenič

Ljubljana, 2007

Metode analize podatkov o raziskovalni dejavnosti na primeru aplikacije IST World

POVZETEK

V magistrskem delu izdelamo in uporabimo (1) metode strojnega učenja za integracijo podatkov iz več podatkovnih virov in (2) metode rudarjenja v podatkih o raziskovalni dejavnosti v podporo iskanju partnerjev v kontekstu procesa prenosa znanja. Najprej opišemo spletni informacijski sistem IST World, ki je dosegljiv na spletnem naslovu http://www.ist-world.org. Ta predstavlja okolje, v katerem delujejo opisane metode integracije in analize podatkov o raziskovalni dejavnosti. Nato opišemo aplicirane metode strojnega učenja, ki jih uporabimo za podporo integraciji podatkov, ki izvirajo iz različnih podatkovnih virov. Za reševanje tega problema v kontekstu sistema IST World smo razvili metodologijo integracije podatkov, ki temelji na sodobnih metodah analize podatkov, kot so rudarjenje v besedilih, obrnjeno indeksiranje besed z dokumenti in strojno učenje. Uspešnost razvite metodologije smo empirično potrdili s poskusom integracije raziskovalnih podatkov iz baze evropskih projektov CORDIS. V drugem delu magistrske naloge predstavimo uporabljene metode analize podatkov o raziskovalni dejavnosti za podporo procesa iskanja partnerjev. Cilj analize je avtomatsko identificirati in slediti intenzivnosti tematike dela ter vzorcev sodelovanja znanstvenih akterjev. V tem delu opišemo metode rudarjenja v besedilih in grafih sodelovanja, ki omogočajo razpoznavanje kompetenc, konzorcijev, razvoja kompetenc in razvoja konzorcijev. Uspešnost algoritmov analize prikažemo na poskusih, ki potrdijo, da se rezultati ujemajo s človeško intuicijo.

Ključne besede: integracija podatkov, povezovanje zapisov, iskanje podvojenih zapisov, rudarjenje v besedilih, urejevalna razdalja, aktivno učenje, strojno učenje, metoda podpornih vektorjev, iskanje partnerjev, vizualizacija podatkov, kompetenca, konzorcij, rudarjenje v podatkih, singularni razcep, večdimenzionalno lestvičenje in gručenje.

UNIVERSITY OF LJUBLJANA

FACULTY OF COMPUTER AND INFORMATION SCIENCE

Jure Ferlež

Methods for Analysis of Research Related Data in the IST-World Application

M.SC. THESIS

Supervisor: Acad. Prof. Dr. Ivan Bratko
Co-Supervisor: Doc. Dr. Dunja Mladenič

Ljubljana, 2007

Methods for Analysis of Research Related Data in the IST-World Application

ABSTRACT

In this master thesis we implement and apply (1) machine learning algorithms to support the integration of data coming from different data sources and (2) data mining algorithms for analysis of research related data to support the partner search process in the knowledge transfer scenario. We begin by giving an overview of the IST World portal, which is an online information system we developed for supporting partner search in the knowledge transfer process. The portal, accessible at http://www.ist-world.org is the environment in which the described algorithms for data integration and data analysis are put to use. We then describe the applied machine learning methods for integrating research related data, which originates from several data sources, into a single integrated dataset. We developed an integration approach based on state of the art data analysis methods such as text mining, inverted indexing and active learning for solving the record linkage problem in the scope of the IST World system. The approach was empirically evaluated with an experiment in integration of research related data from the European CORDIS database of research projects. The second part of the thesis is centered on research related data analysis for the purpose of supporting the partner search process. The goal of the developed and applied data mining algorithms is to automatically identify and track the topics of work and collaboration communities of the analyzed research actors. We describe the used text and graph mining algorithms enabling identification of competences, consortia, competences development and consortia development. We conclude by illustrating the effectiveness of these algorithms in several experiments and by showing that the results agree with human intuition.

Keywords: data integration, record linkage, duplicate detection, machine learning, text mining, string kernel, edit distance, active learning, support vector machine, partner search, data mining, data visualisation, competence, consortium, latent semantic indexing, singular value decomposition, multidimensional scaling and clustering.

Acknowledgments

I would like to thank my mentor Acad. Prof. Dr. Ivan Bratko at the Faculty of Computer and Information Science, for accepting my extraordinary path towards M.Sc. degree and for very valuable advice on perfecting this thesis. I am grateful to my co-mentor Doc. Dr. Dunja Mladenič at Jožef Stefan Institute for her numerous advise on writing scientific papers including this master thesis. I would also like to thank Marko Grobelnik and Mitja Jermol for giving me the opportunity to work on the IST World project. Thank you.

The IST World portal would not have existed without tremendous efforts of Brigitte Jörg, Sebastjan Mislej, Boštjan Pajntar, Peter Ljubič and Drago Trbežnik. I am privileged to be a part of such a great team. I would also like to thank Prof. Dr. Hans Uszkoreit for successfully supervising the IST World project and for taking part in my defending of this thesis.

I am grateful to Janez Brank, Blaž Fortuna, Miha Grčar and Blaž Novak for helping me understand and use the machine learning and data mining algorithms described in this thesis.

The DocumentAtlasBG algorithm was developed by joining forces with Boštjan Pajntar.

Last but not least, I am grateful to my family and friends, who have never stopped supporting me during the long five years of my graduate study.

Jure Ferlež

Darji

Table of Contents

1 Introduction

We start this thesis with the motivation for our work on using several machine learning and data mining methods to facilitate integration and analysis of research related data. We continue this chapter by highlighting our contributions in terms of the implemented information system, used machine learning algorithms and extended data mining algorithms. We end this chapter with an introduction to the structure of this thesis.

1.1 Motivation

We use the term *partner search* to describe the process of searching for actors in the academic sector (researchers or research organizations) and the industrial (e.g. commercial companies) sector, which have the potential for effective transfer of scientific knowledge into a particular line of industry. This describes searching for academic partners with the scientific state of the art that could be used by the known industrial partners. It includes searching for industrial partners, which could use the knowledge of the known academic partners. The match is better if the potential partners already have experience in mutual collaboration. That is, they have already cooperated or at least shared a common partner in the past attempts of knowledge transfer. Therefore, when considering partners for knowledge transfer the following questions concerning the topic of work and the social network of a potential partner are usually asked: Which are the main topics of an individual actor's work in the absolute context and in the context of the considered partnership? What are the past collaborations of an academic or industrial organization in the absolute context and in the context of the considered partnership? How did the topics of work and collaboration patterns change with time in the absolute context and in the context of the considered partnership?

Automated tools in the form of databases storing research related data are developed to help answer these questions. Examples include structured databases containing information about the funded academic projects and structured dictionaries of commercial companies looking for commercial opportunities. We use the term *research related data* to describe the data usually provided by such information systems. *Research related data* are the records of individual actor’s description of work and collaboration. The stored records of actors’ work usually include textual information on the topic of work and collaboration information on the partners performing the work. In case of the academic actors (researchers or research organizations) the records are usually publications published in academic journals or scientific conference proceedings. Another type of records of academic activity is the research project description in which the academic actors participated. Assessing the records of activity of the non academic actors (e.g. commercial companies) is more difficult as usually these do not publish detailed information on their business projects, topics of their research work and collaboration. Usually the reason lies in preserving the secrets of the trade, which are the main assets of the high technological industrial actors. Instead, these actors usually provide short description of their expertise in the form of keywords and product descriptions. The description of collaboration is provided in a form of referential lists of customers and partners.

As the size of research related content stored within such information systems grows very rapidly, it quickly becomes too big for the traditional search techniques and traditional ways of organizing the information. This causes new data analysis techniques to emerge, which

help to organize the information contained in large repositories of research related data. An example is the co-citation index (Small 1999) measure of similarity and importance of research publications. Another example is the full text search (Zobel et al 1998), which allows searching for actors according to the textual information stored in the databases. The latest promising approach for organizing and presenting research related data includes combining machine learning and data mining methods to enable easier research data integration and better research data analysis.

When acquiring information about potential knowledge transfer partners, we have to use several data sources. In order to aggregate the data inside a single information system, the data must first be transformed into a common database schema and then searched for possible redundancies in records describing the same real world entities – a process also called record linkage (Winkler 2006). Therefore a method for efficient linkage of records coming from different data sources has to be developed.

The problem of solving the issue of record linkage is the first problem we address in this thesis. We show how we dealt with the issue of record linkage inside the developed IST World information system. We show that it is possible to perform this task with the use of the machine learning methods such as active learning (Tong et al 2002) and the SVM classifier (Vapnik 1995).

Acquiring information about the potential knowledge transfer partners includes (1) identifying the main topics of partners' conducted research, (2) analyzing the temporal evolution of the identified research topics, (3) identifying the main collaboration patterns between potential partners and (4) analyzing the temporal development of the identified collaboration patterns. Therefore, data mining methods could be used to enable extraction of this information from the research related data.

The problem of identifying and tracking through time the topics of actors' work and patterns of their collaboration based on the available records of work and collaboration is the second problem we address in this thesis. We show how to use the data mining methods to extract this information from the research related data by searching for low dimensional data embedding, using unsupervised clustering and enabling interactive data visualization.

1.2 Contributions

First contribution of this thesis is the design and evaluation of the data integration approach for solving the record linkage problem in the IST World data repository. The approach relies on the established (Winkler 2006) method for record linkage and adapts it in a novel way with text mining and machine learning methods for integration of research related data.

Second contribution of this thesis is the design, implementation and intuitive evaluation of the DocumentAtlasBG algorithm, which is an extension of the DocumentAtlas (Fortuna et al 2005) algorithm used for text corpus visualization. Extension in DocumentAtlasBG algorithm allows identification and interactive visualization of the main topics identified in the text corpus.

Next contribution is the extension of the temporal clustering method developed by (Caruana et al 2005) with a hierarchical clustering algorithm. This enables interactive visualization and analysis of the temporal development of topics of work.

Another contribution is the setup and intuitive evaluation of the text mining methods used on graph data in a form of vertices and edges. We show that graph representation in a form of a matrix enables us to reuse the text mining methods to identify main connection patterns in a graph.

Final contribution of this thesis is the implementation of an information system called IST World for supporting the partner search in a novel way by providing information on research actors' topics of work and collaboration patterns. The information system is developed in a form of a data mining application allowing its users to select the data for analysis and the analysis method separately. It provides the users with an environment for interaction with the analysis results.

1.3 Organization of the Thesis

Chapter 2 provides an overview of the developed IST World (Jörg et al 2006) portal, an online information system for supporting the partner search process by providing information on research actors' topic of work and collaboration. The portal is the environment in which the machine learning and data mining algorithms for data integration and analysis are implemented and used.

In Chapter 3 we describe the methods used for integration of research related data coming from many data sources into a single integrated dataset. We begin by describing the known problem of record linkage between the two datasets (Winkler 2006). We continue by presenting the particular record linkage problem present in the database of the IST World information system. Next, we present the integration approach designed to solve the record linkage problem in the scope of the IST World system. We conclude with an empirical evaluation of the approach.

Chapter 4 describes the data analysis methods used to identify topics of research work, collaboration patterns and how these change with time. We begin by formally presenting the analytical goals of identifying competence, consortia, competence development and consortia development. We then describe the data mining methods, which enable us to reach these goals. We conclude with a series of experiments which show that the results of the data mining methods agree with human intuition.

Chapter 5 concludes the thesis with a summary, concluding remarks and ideas for future work.

2 IST World Portal

In this chapter we provide an overview of the IST World portal (Jörg et al 2006), an online information system we developed for supporting the partner search process. The portal is the environment in which the machine learning algorithms for data integration and data analysis are used. The portal is available online on the World Wide Web at http://www.ist-world.org and is currently used by more than 2000 users per average working day. Figure 1 displays the portal's start page. To assist the partner search process the portal is setup in the form of a data mining application specifically customized to enable discovery of knowledge about researchers, research organizations, research projects and research papers on a local, national and European level.

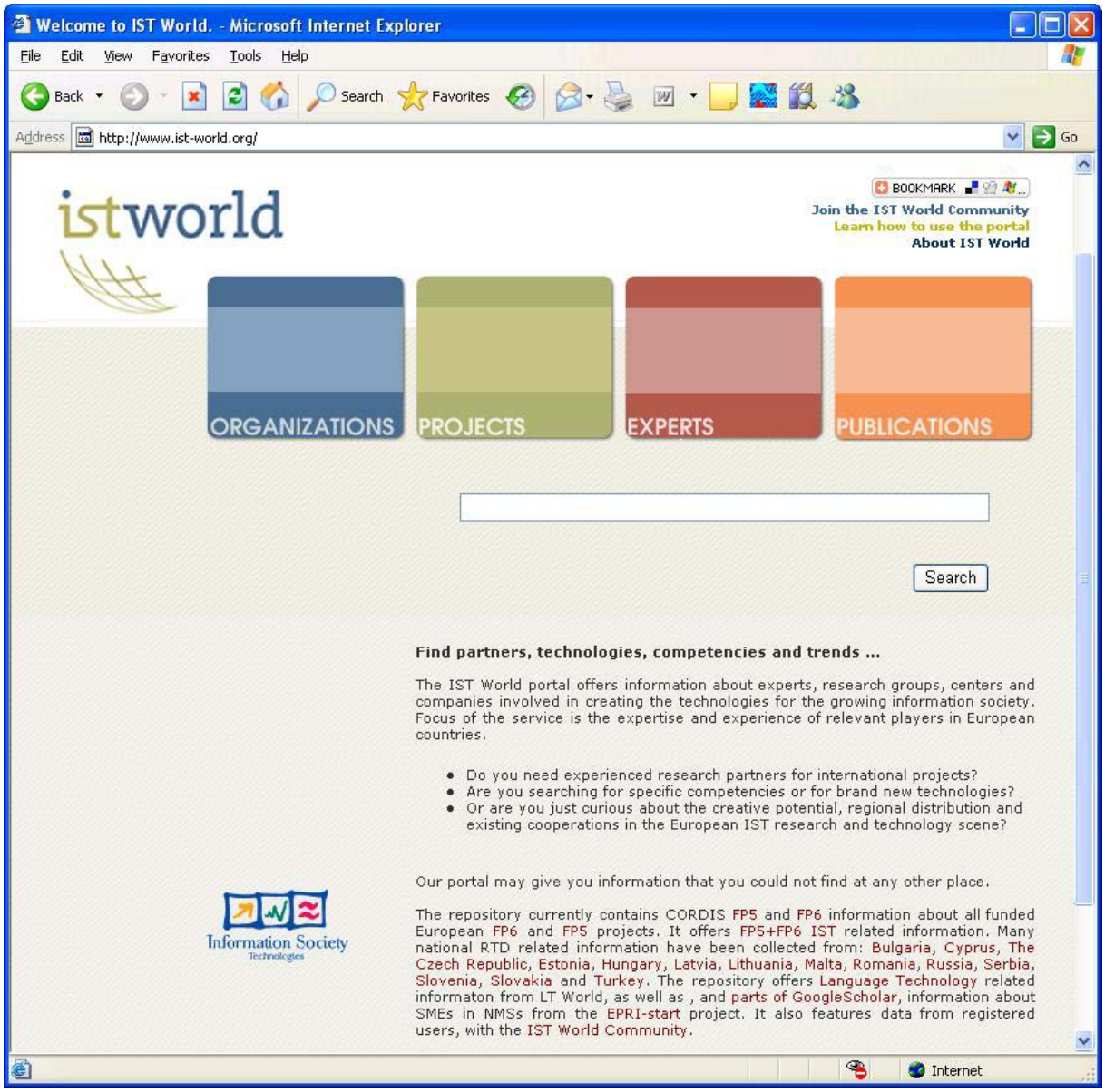


Figure 1: The start page of the IST World Portal.

The rest of this section is continued by describing systems related in the function to the IST World portal. We continue by highlighting the functionality of the IST World portal and the data stored in the IST World database. We conclude with a quick analysis of the portal usage by the web visitors.

2.1 Related Systems

Let us start the review of the IST World portal by giving an overview on some similar information systems, which have been setup in the past to support the partner search in the knowledge transfer scenario. We conclude this section with comparison of these systems with the IST World.

A system called Project Intelligence (Grobelnik and Mladenic 2003) enables analysis of collaboration of the research institutions collaborating in the context of projects financed by the European Union. It facilitates automated project keywords extraction and automated summary of work description. The system however only covers data describing European Funded Research. CISTRANA (Lambert 2006) is a system which provides an insight into integrated data about the state funded research in the European new member states. The amount of presented information is however small due to manual data integration. A system called Google Scholar (Jacsó 2005) offers access to the world's biggest collection of scientific publications and corresponding authors. The system successfully integrates data on scientific research originating from all over the world. Its functionality is however rather limited only offering basic search and browse of the articles and authors. It does not provide systematic analysis of topics of work or communities of collaboration of these publications and researchers.

The most important difference between the IST World information system described in this thesis and the systems mentioned above is the amount of integrated data and the incorporated state of the art analysis techniques. The IST World information system differs from the Project Intelligence and CISTRANA systems by the bigger number of integrated data sources and available automated analysis techniques. It distinguishes from the Google Scholar system by better analysis techniques for mining smaller dataset of research related data.

2.2 IST World Functionality

The functionality of the IST World portal (Jörg et al 2006) aims at providing the user with assistance on discovering topic of work and collaboration patterns knowledge on researchers and research organizations. To meet these goals the portal is setup in the form of a data mining application. Figure 2 displays the user interface of the online analytical application. The typical scenario of the application usage consists of three steps:

1. **search or navigation** to select relevant records from the database (the very left part of the UI in Figure 2),
2. **analysis tool selection** to automatically analyze the selected records (the very right part of the UI in Figure 2).
3. **results observation** to analyze the results of automated data analysis (the centre of the UI in Figure 2)

In the selection step, the records (organizations, experts, publications, projects) representing the target of interest are specified by the user with the help of available search and navigation functionalities. Second, in the analysis step the user selects the appropriate analysis method, which is applied to the retrieved set of entities. In the third step, the analysis results are presented to the user.

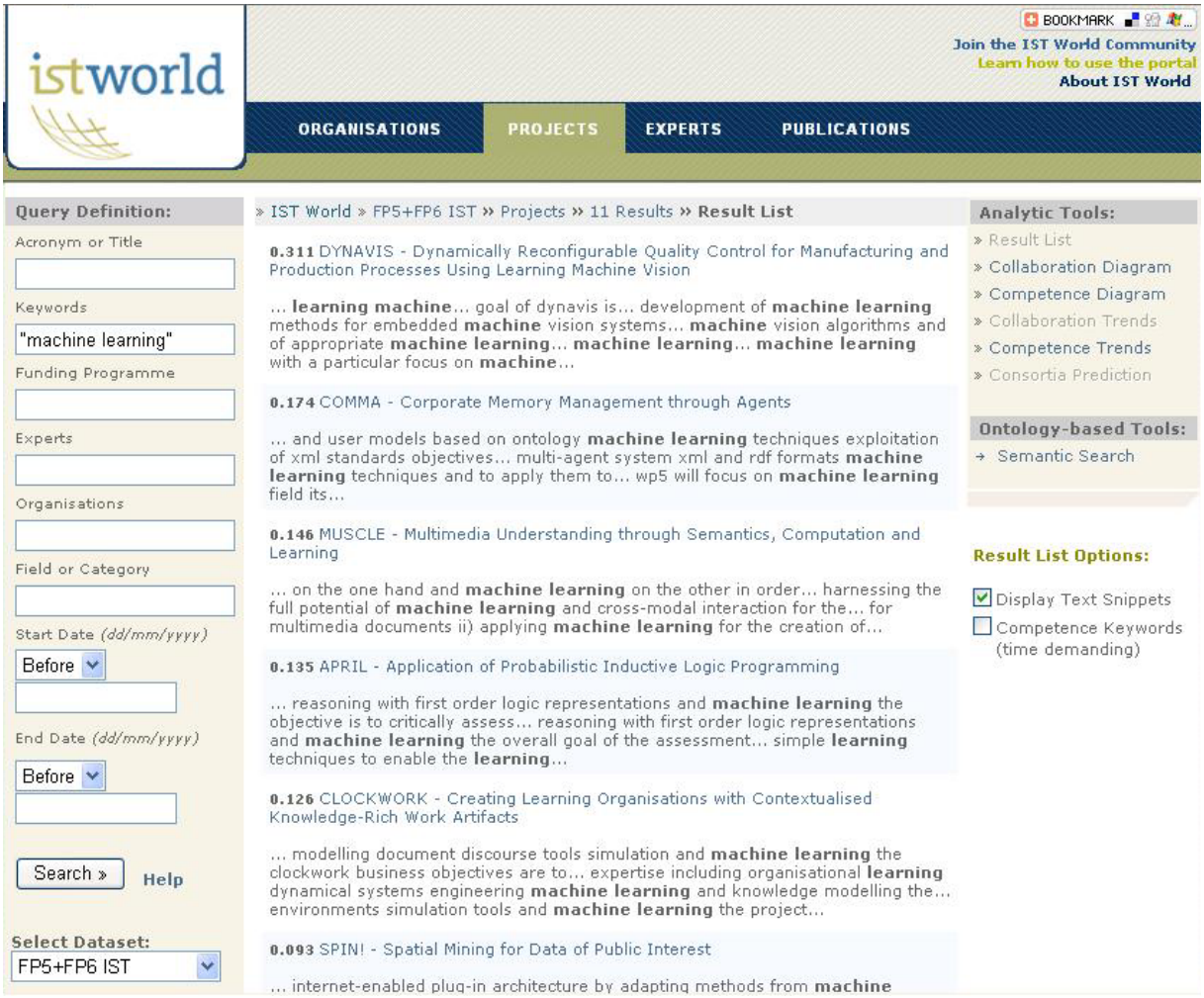


Figure 2: Example user interface of the IST World portal, when searching for projects on "machine learning". The user selects the data for analysis on the left hand side *(Query Definition)*, he or she selects the appropriate analysis method on the right hand side (*Analytical Tools)* and then observes the results in the middle *(Result List)*.

2.2.1 Search or Navigation

The goal of the search or navigation step is to select which records from all the data on people, organizations, projects or publications in the IST World database are to be analyzed. The portal supports various ways of selecting the data:

- **Complex Search**: The portal allows selecting the records via textual keywords, which should appear in the names or topic description of the records retrieved from the database. E.g. a user can search for all the projects which have "machine learning" mentioned in the textual description of the project and which started after year 2002.
- **Record Relatedness**: The portal allows selecting the records which are administratively related to another record. E.g. a user can select all the departments which are related to the Faculty of Computer Science of the University of Ljubljana.

- **Record Collaborations**: The portal allows selecting the people or organizations which are collaborating most with an individual person or organization. E.g. a user can select to analyze the records of organizations which collaborate most with the Jožef Stefan Institute according to common projects or published publications.
- **Record Category Assignment**: The portal allows selecting the records which match a certain category. Majority of the data held in the IST World database was automatically classified (Grobelnik and Mladenic 2005A) into DMOZ categories. E.g. the user can select those publications which were automatically classified in to the category: "TOP/SCIENCE/MATH/".

A limitation of data selection step is that the user is only able to select at most 200 records for data analysis in order to keep low the time demanding complexity in the analysis step. When a certain selection step results in more data records only those records are selected which are most relevant according to the used data selection method. For example, when keyword search is used, only the top 200 hits are used as they correspond most strongly to the search query.

2.2.2 Analysis Tool Selection

The goal of the analysis tool selection step is to select the analysis method and corresponding parameters for analysis of the records selected in the search and navigation step. Currently IST World portal supports six different analysis methods as observed on the very right hand side in Figure 2. The methods are centered on text and graph analysis for the purposes of revealing the topics of work and most important collaboration communities of the actors represented by the selected set of records. The analysis methods are:

- **Textual Competence Description**: This analysis method automatically calculates the textual keywords best describing the topic of work of the selected entities. The user is able to read the keywords with every associated entity and at the same time observe which keywords are the most suitable for the description of work of the entire selected set of entities. E.g. Figure 3 displays the calculated keywords of topic of work of all the organizations participating in the ALVIS EU project. This analysis method is not described in detail in the scope of this thesis.

- **Collaboration Graph Visualization**: This analysis method allows complex graph visualization by automatically calculating the best position of the graph vertexes according to the weight and number of the graph edges. E.g. Figure 4 displays a project collaboration graph of the organizations participating in the ALVIS EU project. The user is able to set a few parameters for better results of the analysis like the type of node ranking, the type of collaboration to take into account and the time allowed for the analysis. This analysis method is also not described in detail in the scope of this thesis.

- **Competence Diagram**: The competence diagram visualization allows analysis of the textual description of the selected records for the purpose of visualizing the similarity between the different descriptions of the selected records and for revealing the main topics found in the descriptions. E.g. Figure 5 shows a competence diagram of organizations involved in the ALVIS EU project. The Competence Visualization technique description is one of the main topics of this thesis and is described in details in the Data Analysis chapter.

- **Consortia Diagram**: The consortia diagram visualization technique automatically analyzes the collaboration graph of the selected set of records in order to reveal the most important collaboration communities and the correspondence between the individual records with the identified communities. E.g. Figure 6 shows the Collaboration Patterns Visualization of the organizations involved in the ALVIS EU project based on the collaboration on common projects. The user can set the parameters describing which collaboration records to include in the analysis. The Collaboration Patterns Visualization technique description is also one of the main topics of this thesis and is described in detail in the Data Analysis chapter.

- **Competence Development Diagram**: This analysis technique combines temporal and textual analysis for the purpose of visualizing the temporal development of the automatically identified topics in the textual descriptions of the selected records. E.g. Figure 7 displays the development of the topics of work identified from the textual description of activity of the organizations involved with the ALVIS EU project. The user can set the parameters describing which textual records to include in the analysis. The Competence Trend Visualization technique description is also one of the main topics of this thesis and is explained in the Data Analysis chapter.

- **Consortia Development Diagram**: This technique combines temporal and graph analysis to present the development of importance of communities with time according to cooperation on the common projects. E.g. Figure 8 shows the temporal development of importance of the identified communities of the organizations collaborating in the ALVIS EU project. The user can again set the parameters describing which collaboration records to include in the analysis. The Collaboration Patterns Trends Visualization description is the last of the four data analysis cornerstones of this thesis and is described in details in the Data Analysis chapter.

2.2.3 Results Analysis

The goal of the Results Observation step is to allow the user an interactive observation of the results of the selected automated data analysis tool. Most of the data analysis tools are visualizations enriched with interactivity which allows interactive inspection of the analyzed data. The user is usually able to interact and explore the visualization by moving or dragging the mouse cursor around the visualization. What is more the user is usually able to select the record which he or she finds most interesting for further analysis. For example, in the Collaboration Graph Visualization, presented in Figure 4, the user is able to move the presented graph or the individual vertexes around the visualization area. In the Competence Visualization, presented in Figure 5, the user is able to observe the topic of work of an individual actor by moving the mouse pointer over the dot representing the record of the selected actor. In the Competence Trend Visualization, presented in Figure 7, the user is able, by clicking on the individual trend patterns, to explore and drill down on the individual trends of identified topics of work. The Collaboration Patterns Visualization and the Collaboration Patterns Trends Visualization techniques, presented in Figures 6 and 8, are similar in terms of the provided interactivity. Instead of focusing on topics of work these techniques allow interactivity which enables the user to explore the collaboration patterns of a certain institution. Every visualization technique allows the user to select a particular record for detailed analysis of all the data associated with it. Figure 9 displays the detail record of the ALVIS EU project. We can see that the user is able to select by using hyperlinks the next analysis data and analysis method directly from this detail page

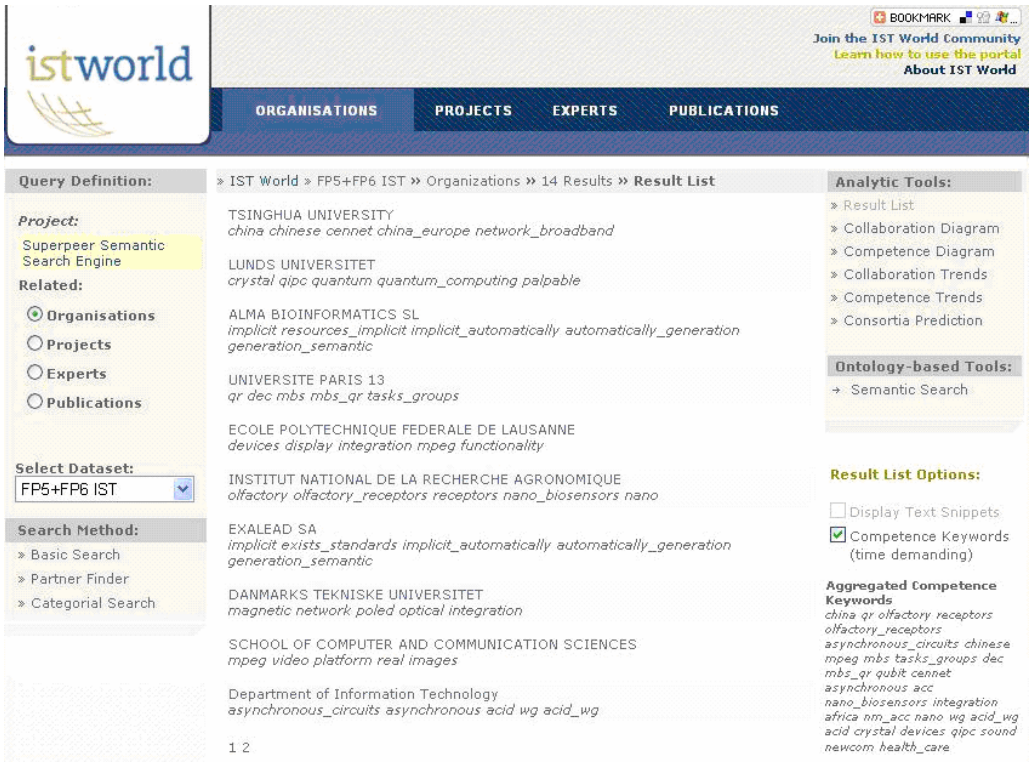


Figure 3: Textual Competence Description analysis option is selected. Calculated competence keywords of all the organizations participating in the ALVIS EU project are displays. Most important keywords overall the search results are presented in the lower right hand side.

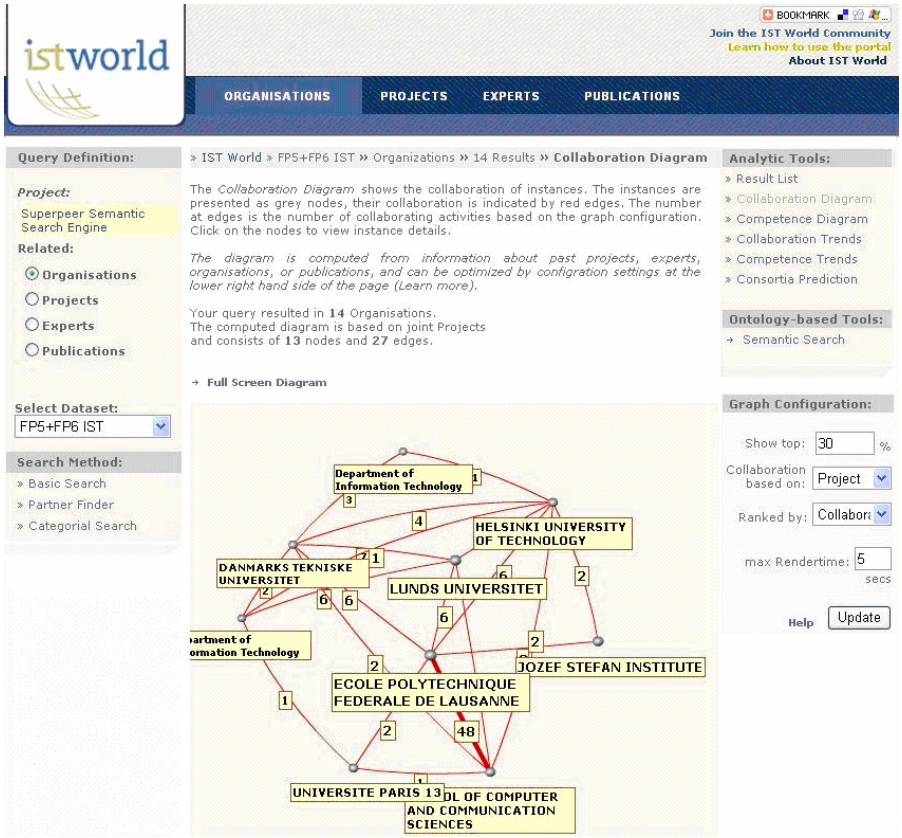


Figure 4: Collaboration Graph Visualization analysis option is selected. Visualization of the collaboration graph of all the institutions participating in the ALVIS EU project according to common projects is visible.

Figure 5: Competence Diagram analysis option is selected. Competence visualization of the organizations involved with the ALVIS EU project reveals the organizations most important topics of work is visible.

Figure 6: Consortia Diagram analysis option is selected. Consortia visualization of the organizations involved with the ALVIS EU project based on common projects reveals the communities and the connections between them.

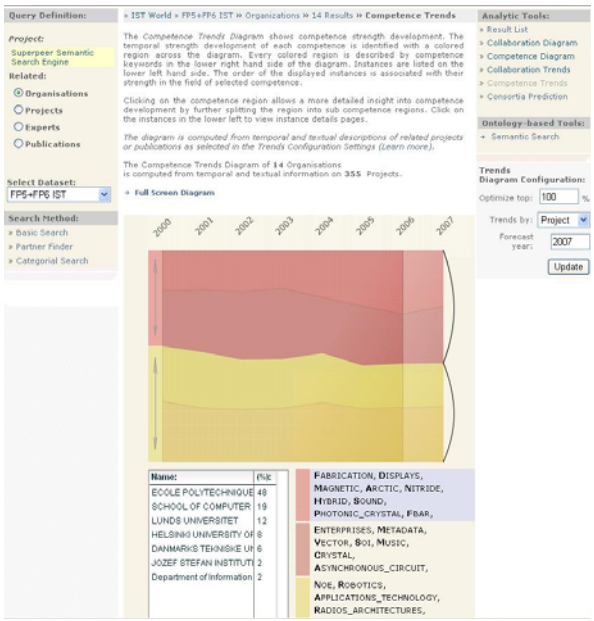


Figure 7: Competence Development analysis option is selected. Competence trend visualization of the organizations involved with the ALVIS EU project reveals the temporal development of the identified topics of work is visible.

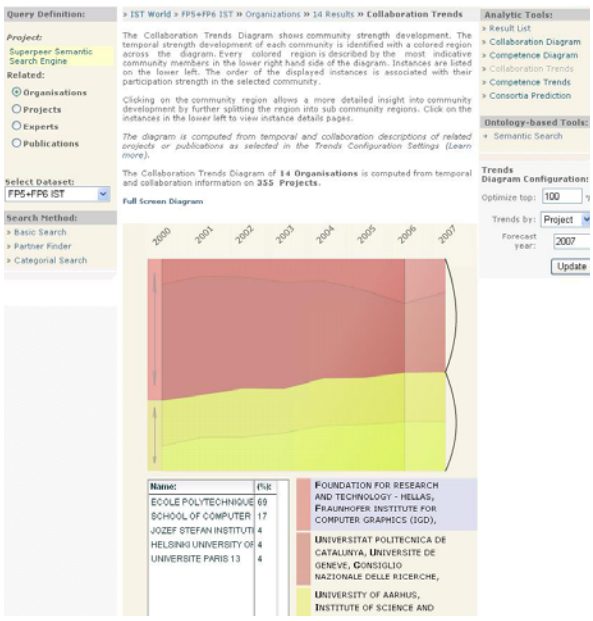


Figure 8: Consortia Development analysis option is selected. Collaboration patterns trends visualization of the organizations involved with the ALVIS EU project based on the temporal information in the shared projects reveals the temporal development of the identified communities.

Superpeer Semantic Search Engine

Acronym: ALVIS
Start: 1/1/2004
End: 12/31/2006

Project Status History

unknown	1/1/2004

Funding Programmes

CORDIS Specific Targeted Research Project	2260000

Abstract

The project will conduct research in the design, use and interoperability of topic-specific search engines with the goal of developing an open source prototype of a distributed, semantic-based search engine. Existing search engines provide poor foundation for semantic web operations, and are becoming vital access paths to the Internet. Our approach is not the traditional Semantic Web approach with coded or semi-automatically extracted metadata, but rather an engine that can build on content through automatic analysis of free text. Linguistic processing will be inside the search engine and a probabilistic document model will provide a principled evaluation of relevance to complement existing standard authority scores. This will facilitate semantic retrieval and incorporate pre-existing domain ontologies using facilities for import and maintenance. The distributed design will be based on exposing search objects as resources, and on using implicit and automatically generated semantics (not ontologies) to distribute queries and merge results. Because semantic expressivity and interoperability are competing goals, developing a system that is both distributed and semantic-based is the key challenge: research will involve both the statistical and linguistic format of semantic internals, and determining the extent to which the semantic internals are exposed at the interface. The consortium assembles a team of leading international researchers from the areas of peer-to-peer computing (P2P), information extraction and search, and probabilistic modeling together with some of Europe's leading SMEs. The combination of design goals, the distributed operation and open source development have been chosen to support incremental growth, third-party involvement, low barrier to entry, and next generation knowledge services so it provides a foundation for European SMEs and be accepted internationally.

Organizations Analyze, Collaboration (Trends, Prediction), Competence (Trends)

ALMA BIOINFORMATICS SL
DANMARKS TEKNISKE UNIVERSITET
Department of Information Technology
Department of Information Technology
ECOLE POLYTECHNIQUE FEDERALE DE LAUSANNE
EXALEAD SA
HELSINKI UNIVERSITY OF TECHNOLOGY
INDEX DATA APS
INSTITUT NATIONAL DE LA RECHERCHE AGRONOMIQUE
JOZEF STEFAN INSTITUTE

1 2

Figure 9: The detail record of the ALVIS EU project. The links (*Analyze, Collaboration, Trends, Prediction, etc*), before the organizations list, allow quick selection of the organizations in the list together with selection of the analysis method to be performed on the records in this list.

2.3 IST World Database

The data in the IST World repository must be as rich as possible in order to provide best information to the IST World users. This is why the data in the portal database is imported from several types of public data sources. The database includes data from various institutionalized databases like project databases, national research databases, regional business indexes and global publication citation databases. The separate data sources currently contributing to data contained inside the IST World database are:

- **SICRIS**: The IST World database contains data from the Slovenian CRIS World Wide Web service called SICRIS available at http://sicris.izum.si (Seljak 2001). All together the database contains information on Slovenian researchers (11930), research organisations (1794) and research projects (4780).
- **CORDIS**: The IST World database contains data from the CORDIS database (Thévignot 2000) containing research and development projects funded by the European Union. The data contained in the CORDIS database are available as a World Wide Web service called CORDIS at http://cordis.europa.eu/. The database contains information on all the projects in the context of the Fifth and Sixth European Framework Programme for research and technological development together with a list of organisations collaborating on each of the projects. All together the database contains information on 24965 organisations and 5094 projects from the FP6 programme and 35010 organisations and 15343 projects from the FP5 programme.
- **GOOGLE Scholar**: The IST World database contains data from the world's largest citation index service called Google Scholar (Jacsó 2005). The service provides information on a large collection of academic researchers and their publications. The data in this index is available as World Wide Web service available at http://scholar.google.com/. All together, the database contains information on 847706 researchers and their 1138782 publications.
- **LT World database:** Information on 2109 organizations, 763 projects and 2750 researchers involved with language technologies.
- **Bulgarian CRIS database:** Information on 794 organizations, 73 projects, 10940 researchers and 19023 publications.
- **Estonian RIS database:** Information on 114 organizations, 2167 projects, 5062 researchers and 32712 publications.
- **Serbian database**: Information on 60 organisations, 2278 persons and 77410 publications participating in Serbian national research programmes.
- **Singleimage database**: Information on 15870 organisations, 3219 projects and 16514 persons participating in IST FP5 and FP6 activities.
- **Huncris CRIS dataset:** Information on 1245 organisations, 1297 projects and 2425 experts.
- **Latvian CRIS dataset:** Information on 106 organisations, 830 projects and 701 researchers.
- **Turkish SME database:** Information on 285 Turkish small and medium sized technological organisations is stored in the IST World database.
- **Czech database:** Information on 16 Czech organisations and 23 projects.
- **Cyprus database:** Information on 29 Cyprus organisations.
- **Romanian database:** Information on 169 organisations, 68 projects, and 87 experts.

- **SINNIN database**: Information on 136 organisations and 709 projects participating in Russian technical projects.

The data originating from different sources should be merged together to create one single dataset in which every real life entity is only represented once. The problem of identifying the record pairs which actually correspond to only one real world entity is called record linkage (Winkler 2006). The description of solution to the record linkage problem in the scope of data inside the IST World database is the first of the two parts of this master thesis. It is in detail described in the Data Integration chapter.

2.4 IST World Portal usage

We continue the overview of the IST World portal by evaluating its success in terms of the frequency of the usage of the portal which is freely available on the World Wide Web. The usage frequency is assessed with the analysis of the usage log files. The presented report in Figure 10 shows the usage of the portal from the 1st of April until the 15th of July 2007.

Figure 10: Overview Report on Visitors of the IST World portal from April 1st to July 15th 2007.

In order to get a basic impression of the portal usage we examine the Overview Report displayed in Figure 10. The top diagram in Figure 10 shows the number of daily visitors for the selected time period. We can observe aggregated statistics like the number of visitors during this time (Visits), the number of different people visiting the site (Absolute Unique Visitors), the count of all pages that were viewed by these visitors (Pageviews), the average number of served pages per visit (Average Pageviews), the average amount of time spent on the site by an average visitor (time on site), the percentage of visitors that only viewed one page and then left (Bounce Rate) and the percentage of visits by new visitors in the number of all visits (New Visits). The little trends diagrams associated with each of these statistics show the statistics' trend development.

Figure 10 shows that the number of visitors of the portal increased substantially from April 1st to July 15th. What is more, we can observe that the number of visitors per day depends on the working day of the week. Usually there are fewer visitors during the weekend days. These aggregated statistic shows that during the specified time range, the portal was visited by approximately 86 000 people who created approximately 100 000 visits. Each visit was approximately 3 minutes long and included a usage of about 3 pages. The worrying fact is that the bounce rate metric is quite high, signaling that only one third of the visitors viewed more than one page. The trend associated with this metric is rather stable and suggests that improvements could be done in this area.

3 Data Integration

In this chapter we describe the methods used for integration of data on different research activities (e.g., research papers, research projects, research collaboration, etc.) coming from many data sources into a single integrated dataset. We begin by describing the known problem of record linkage between two datasets and the known problem of duplicate detection in a single dataset (Winkler 2006). We first illustrate the problem using examples and then put forward the problem's formal definition. We continue by presenting the particular record linkage problem present in the database of the IST World portal. Next, we present the IST World Integration approach designed to solve the record linkage and duplicate detection problem in the scope of IST World. We conclude with an empirical evaluation of the approach on the duplicate detection task in one of the research related datasets.

3.1 Problem Definition

"Record linkage is the means of combining information from a variety of computerized files. It is also referred to as data cleaning or object identification." (Winkler 2006) The most basic application is identifying duplicates within a dataset or identifying duplicates across two datasets. If a single large dataset is considered, then the record linkage or matching procedures may be intended to identify duplicates.

(Winkler 2006) has also shown that computerized record linkage procedures can significantly reduce the resources needed for identifying duplicates in comparison with methods that are primarily manual. (Newcombe et al 1975) have demonstrated the purely computerized duplicate detection in high quality person lists can often identify duplicates at greater level of accuracy than duplicate detection that involves a combination of computerized procedures and reviews by highly trained clerks. The reason is that the computerized procedures can make use of the overall information from large parts of a list. For instance, the purely computerized procedure can make use of the relative rarity of various names and combinations of information in identifying duplicates. The relative rarity is computed as the files are being matched. (Winkler 2006) observed that the automated frequency-based (or value-specific) procedures could account for the relative rarity of a name such as 'Martinez' in cities such as Minneapolis, Minnesota in the US in comparison with the relatively high frequency of "Martinez' in Los Angeles, California.

We illustrate some of the issues of record linkage with an example. When integrating data from the CORDIS (Thévignot 2000) and the SICRIS (Seljak 2001) data source, we run into the problem of both sources containing information on research projects, the former on the level of EU and the latter on a national level. A real world organization like Jožef Stefan Institute in Slovenia can therefore appear in both data sets. The two records should therefore be merged into one single record including data from both datasets. Figure 11 presents the two records from two different data sources side by side.

Figure 11: A record about Jožef Stefan Institute, Ljubljana, Slovenia from the CORDIS data source (*left*) and another one from the SICRIS data set (*right*)

Another possible record linkage problem is that of a dirty data source, which does not include a unique list of the contained entities. Again the example is the CORDIS database (Thévignot 2000), which keeps a unique list of project but does not include a unique list of organizations collaborating in these projects. This causes many different spelling and word orderings in the names of organizations in the organization list. This again results in duplicate records about the same organization. In this setting the problem is referred to as a duplicate detection problem (Winkler 2006). Figure 12 displays an example of the records of 2 organizations coming from the same CORDIS data source actually representing one single real life organization Jožef Stefan Institute in Ljubljana, Slovenia.

Figure 12: Two duplicate records about one and only Jožef Stefan Institute, Ljubljana. The two instances use different spelling in the first word of the names of the institute: Josef Stefan Institute (*left*) and Jozef Stefan Institute (*right*).

Let us now formally define the problem as described in (Winkler 2006):

PROBLEM: Record Linkage between record sets A and B.

GIVEN: two sets of records A and B,
CLASSIFY: every pair (a,b) $\in$ A × B into M set of true matches or U set of true non matches.

The records from the two datasets A and B are matched. The idea is to classify pairs of records (a,b) from the product space A × B into M, the set of true matches or U, the set of true non matches.

In a substantial number of situations, the datasets are too big to test every pair (a,b) from A × B whether it represents a match or not. Newcombe (1962, 1988) showed how to reduce the number of pairs considered by only testing pairs which agree on a characteristic such as surname or date-of-birth. Such reduction in the number of pairs is called blocking (Winkler 2006). After the reduced list of pairs is brought together more advanced, computational intensive methods are used for comparing them. This step is usually called matching (Winkler 2006).

In the following sections we describe the concrete record linkage problem present in the IST World database and evaluate our approach to perform the blocking and matching steps with the appliance of machine learning and text mining methods.

3.2 IST World Data Integration Task

The integration task within the IST World portal is necessary because the database behind the portal contains data from different sources which should all be merged together to create one single dataset in which every real life entity is only represented once. The number of different data sources providing data to the IST World database is described in details in the chapter on IST World Portal. As many data sources describe the same real world entity this results in two or more IST World records about one entity. What is more, in the same IST World portal we need to overcome the problem of duplicate detection inside a single dataset. For example, the CORDIS (Thévignot 2000) dataset, as publicly provided, already contains duplicate records of the same entity. Overall, any two records describing the same real world entity should be combined into one record in the IST World portal.

The data in the IST World portal database is imported from several types of public data sources like institutionalized databases, regional resources and project databases, national research databases, regional business indexes, global citation databases, the IST World Community Portal and web crawls (Grabczewski et al. 2005). Our goal was to establish an effective IST World Integration Approach for combining the data coming from all these data sources into one integrated data set.

In order to get a basic idea about the data integration problem and possible solutions in the scope of the IST World database we conveyed a survey of the typical duplicate detection and record linkage problem patterns present. Survey (Ferlez et al 2007) on a random sample from the IST World dataset showed that most obvious errors have been identified within duplicate organizations contained inside the CORDIS database (Thévignot 2000), especially because of different name variants and multiple entries of organizations names. The results of the survey suggest that the majority of the differences actually come in the form of additional special characters or character modifications in the names of the entities. The most important patterns in differences between duplicates were identified as follows (in descending order of importance):

- Capitalization / Lowercase Letters
- Blanks, extra Spaces
- Hyphens
- Quotes
- Coma in Different Places
- Article in Name
- Full stop in Name
- Incomplete Names
- English Translation
- Word Order
- Language Specific Characters (encoding Jorg instead of Jörg)
- Special Characters (wrong encoding of '&' and '?')
- Differences in Addresses

The mentioned survey also showed that only name and location information of the organizations held in the IST World portal can be used to identify the matching pairs of organizations. The rest of the information related with the pair of organization records like records of work and collaboration usually differ from one duplicate organization to another.

3.3 IST World Data Integration Approach

We developed an IST World Data Integration Approach to solve the record linkage and duplicate detection problem in the IST World database. The approach was constructed by combining the traditional record linkage methods with the machine learning algorithms according to the specifics of the data contained in the IST World database.

A standard setting of methods for solving the record linkage problem is described in (Winkler 2006). Search for redundant records is usually performed in two steps. In the first step a heuristic is used to identify the potential redundant pairs of records. This step is called blocking. Many approaches to blocking have already been developed such as sliding window method (Hernandez et al 1995), based on feature selection (Newcombe 1998) or pre-clustering of records (McCallum et al 2005). The second step of the standard record linkage solution is called matching. The goal of which is to decide on every potential pair of records whether they describe the same real world entity. Many matching approaches have been described in (Winkler 2006).

The IST World Data Integration Approach features a blocking method exploiting an inverted index of all the words (Zobel et al 1998) used in the examined records. The matching step was performed by introducing the algorithms for automatic classifier induction, and text mining feature construction.

We created a five step approach for tackling the IST World data integration task. We describe it using the notation established in the Problem Definition section of this chapter. In the first step, we effectively reduce the magnitude of the problem by effectively reducing the considered space of all pairs (a,b) of records from the cross product $A \times B$ of datasets A and B (**Blocking Step**). In the second step, we create a function for mapping a potential pair (a,b) into a useful pair comparison space vector Γ (**Feature Generation Step**). In the third step, we use the machine learning approach called active learning to support the manual creation of the

M and U sets of labeled pairs (**Active Learning Step***)*. Next we use a machine learning algorithm to automatically construct a decision rule like (2) in a comparison vector space Γ (**Classifier Induction Step***)*. Last step of the approach uses the induced classifier on the still unlabeled pairs to automatically assign them into sets M and U (**Classifier Application Step**). In the following sub sections we review every step in more detail.

3.3.1 Blocking

In order to overcome the problem of combinatorial explosion of pairs we first need to reduce the initial space of record pairs A × B. We try to exploit the full text indexing and full text querying capabilities of the relational database to reduce the number of potential pairs of redundant records.

Survey (Ferlez et al 2007), mentioned in the previous section, showed that the most frequent pattern in the cause of the duplicates problem is a different spelling in one of the words in the names or different word ordering in the names of the pair. The duplicate records will therefore most probably have at least one word in common or very similar. This is why we looked for a *blocking* procedure which would efficiently compare every record with all the records and reduce or block those pairs which do not share at least one common word.

Full text indexing and querying (Zobel et al 1998) is the approach which can be used to effectively reduce the number of potential pairs of records in the scope of the IST World. Full text indexing is a technique for enabling efficient searching for the records of the table by any word contained in these records. This is achieved by building a so called inverted index. An inverted index is an index structure storing a mapping from words to their locations in a document or a set of documents, allowing full text search. The full text indexing allows us to match records with the same words or same inflection of the words regardless of their order or case. Full text search capabilities of the relational database include searching with special character independence, case independence, word order independence and capability of inflection matching. However it does not allow us to search for any records that may have been completely mistyped and therefore do not share at least one common word with the query. This technique is thus most suitable for those records with correct spelling in at least one of the common words, which is the case in the IST World data integration task. Therefore the blocking procedure based on full text search on names can find the most probable duplicate records candidates to every record in the context of data inside IST World.

We created a blocking procedure using a full text query with inflections to query the database for names of each entity in a dataset A. For each record we produced a short list of length n of most similar records. For the duplicate detection problem we established a list of $n << |A|$ of similar name matches according to the query results for every records in A. This produces a list of potential pairs of size $s \leq n*|A|$ for the duplicate detection problem, which is a good result in comparison to the size $s = |A|*|A|$ of the non blocked list of all pairs in A × A. For the record linkage problem we established a list of $n << |B|$ of similar name matches according to the query results for every record in A and $n << |A|$ of similar name matches according to the query results for every record in B. This approach allowed to produce a short list of potential pairs of size $s \leq n*(|A| + |B|)$ for the record linkage problem, which is much better than the size $s = |A|*|B|$ of the non blocked list of all pairs in A × B. Table 1 displays an example of a short list of 10 most similar organizations to Jožef Stefan Institute for a duplicate detection example, when we use n=10. We can conclude from the entries in this table that the

developed blocking procedure successfully assembles most probable matching pairs for the Jožef Stefan Institute.

1	JOZEF STEFAN INSTITUTE
2	ENVIRONMENTAL SCIENCES / JOZEF STEFAN INSTITUTE
3	TECHNOLOGY TRANSFER OFFICE/ JOZEF STEFAN INSTITUTE
4	INSTITUT "JOZEF STEFAN"
5	STEFAN BATORY FOUNDATION
6	STEFAN TISCHER LANDSCHAFTSARCHITEKT
7	INSTITUT JOSEF STEFAN
8	UNIVERSITY OF PAVOL JOZEF SAFARIK
9	FORESTRY FACULTY, UNIVERSITY STEFAN CEL MARE SUCEAVA
10	INSTITUTE OF MICROBIOLOGY OF THE GERMAN ARMED FORCES (BUNDESWEHR INSTITUTE OF MICROBIOLOGY)

Table 1: A short list of 10 most similar organizations to Jožef Stefan Institute according to full text search on the name

3.3.2 Feature Generation

In the Feature Generation step we develop a function γ: $A \times B \rightarrow \Gamma$ for mapping a potential pair of records (a,b) into a finite vector space Γ. Every $\gamma(a,b) \in \Gamma$ is a vector describing a comparison of records a and b from different aspects. Every particular aspect of comparison result corresponds to a particular component of the vector. The result of the particular comparison affects the value of this component. As we use efficient machine learning algorithms for analysis of these vectors we would like our comparison pattern to capture as many different aspects of similarity between records a and b as possible. In other words we want our model to have a rich hypothesis language which will enable it to better recognize the difference in patterns of comparison vectors of record pairs in the two sets M and U.

We chose to represent a comparison vector $\gamma(a,b)$ between two records a and b as a finite length vector with every component of the vector quantitatively representing another aspect of similarity between the two records. This makes the vector space Γ a space with a defined standard scalar product enabling comparison of different comparison vectors $\gamma(a,b)$ and $\gamma(c,d)$ from the two sets of pairs. This will become useful in the Active Learning and Classifier Induction Step as it enables use of the Support Vector Machine algorithm with linear kernels (Vapnik 1995).

The mentioned survey of the IST World data (Ferlez et al. 2007) also showed that only name and location information of the organizations held in the IST World portal can be used to identify the matching pairs of organizations in the set of organization pairs from $A \times B$. In detail the survey showed that organizations should typically be matched if they share almost the same spelling. The difference in spelling of names mostly comes in the form of mixed order of words, addition of special words like LTD, GMBH or DOO, inclusion of special characters and case of letters distinction. An example of the two records with some of these differences was already displayed in Figure 12. In order to capture these patterns of distinction between matches we had to develop comparison techniques on the names of organizations which would allow to grade as more similar those records which only differ from each other in one of the mentioned ways.

In order to capture these types of similarities between name and location information of the two records (a,b) we developed four comparison approaches. First, we presented both names and location information string with the vector of words using the well known technique called bag of words. This presentation of words in a document neglects their position inside the document and only captures the frequency of appearance of every word in a document. This allowed us to capture the similarity of the two records in the usage of the special words like LTD, GMBH or DOO. Next, we compared the order of the words and the order of the characters with the technique called string kernels (Lodhi et al 2002) to cope with the mixed word order or inclusion of special characters. Next, in order to cope with spelling mistakes in the words and alternative word ordering we used the edit distance measure (Levenshtein 1966). To combine all the mentioned approaches we normalized the features of each of the approaches and then combined them into one feature vector describing similarity of the pair from different comparison aspects. Each of the feature generation techniques and the feature combination are in detail described bellow.

Bag of Words Comparison

We developed a set of comparison features to capture matching of records with special words like LTD and GMBH in their names. This is why we used the well established bag-of-words representation, which relies on the weights associated with the words, for the representation of the words appearing in names and location fields of the records. The weights of the words are calculated by so called tf-idf weighting (Salton et al 1983).

The tf–idf weight (term frequency–inverse document frequency) is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a document collection. The importance increases proportionally to the number of times a word appears in the document but is offset by the number of documents containing the word.

First set of vector components describing pair comparison patterns is there fore represented by a component for every word in the name field and every word of the location field that appear in the candidate pair. This implies that if a word appears in both records' fields than its weight is bigger, if it appears in only one of the records its weight is smaller, and if it does not appear in any of the records then its weight is set to zero. Table 2 represents feature vector assignments for the comparison of a pair of records with names "Jozef Stefan Institute" and "Institut Jozef Stefan".

Feature Word	Jozef	Stefan	Institute	Institut
Feature Weight	2	2	1	1

Table 2: Bag of Words Comparison vector of a pair of names: "Jozef Stefan Institute" and "Institut Jozef Stefan"

String Kernel Comparison

To enable comparison of records with different character ordering, missing characters or use of special characters in the record fields we created a second set of features based on the order of the characters in words of the record pair. These features capture the similarity of organizations in terms of the difference in the order of the individual characters between the

two records in the corresponding fields, e.g. organization name fields. This type of comparison is called a string kernel on the characters.

Let Σ be a finite alphabet. A string is a finite sequence of characters from Σ, including empty sequence. For strings s and t we denote with |s| the length of string $s = s_1 \ldots s_{|s|}$, with st the string obtained by concatenating the strings s and t and with s[i : j] substring $s_i \ldots s_j$. We say that u is a substring of s if there exists indices $\mathbf{i} = (i_1, \ldots, i_{|u|})$ with $1 \leq i_1 < \ldots < i_{|u|} \leq |s|$ such that u = s[**i**]. The length l(**i**) of the subsequence in s is $i_{|u|}-i_1+1$. Feature mapping Φ for string s is given by defining Φ_u for each $u \in \Sigma^n$ as

$$\Phi_u(s) = \sum_{i:u=s[i]} \lambda^{l(i)} \tag{1}$$

for some $\lambda \leq 1$. These features measure the number of occurrences of subsequences in the string s weighting them according to their lengths. Hence, the inner product of the feature vectors for two strings s and t give a sum over all common subsequences weighted according to their frequency of occurrence and lengths

$$K_n(s,t) = \sum_{u \in \Sigma^n} \Phi_u(s)\Phi_u(t) = \sum_{u \in \Sigma^n} \sum_{i:u=s[i]} \sum_{j:u=s[j]} \lambda^{l(i)+l(j)} \tag{2}$$

Computation of features using this definition for n > 4 is very expensive, hence the explicit use of such would be impossible. It however turns out that the kernel function (2) can be calculated very efficiently using a recursive definition. Please see (Fortuna 2004) for more information.

"The main idea of string kernels is to compare documents not by words, but by the sub strings they contain. These sub strings do not need to be contiguous, but they receive different weighting according to the degree of contiguity. For example: sub string ‘c-a-r’ is present both in word ‘card’ and ‘custard’ but with different weighting. Weight depends on the length of sub string and decay factor λ. In previous example sub string ‘car’ would receive weight 4 as part of ‘card’ and 7 as part of ‘custard’." (Lodhi et al 2002). Every string kernel can be tuned using two mentioned parameters. First is the decay factor λ, signaling the importance of gaps in the ordering of the characters in the two strings. Second parameter is the length n of sub strings to be examined when comparing two strings. Actually string kernel can be computed for many lengths at once. Resulting scores can then be weighted and summed together. This is why a parameter vector representing linear combination of importance of a kernel on a given sub string length is used to combine many kernels into one. We can combine kernel parameters to get sub string order comparison in many different aspects.

As string kernels are dependant on the length of the compared strings it is a good idea to normalize them between 0 and 1. We remove any bias introduced by different lengths of the compared strings by defining kernelized cosine distance measure. We can compute kernelized cosine distance K_c by replacing the inner product with the kernelized inner product in the cosine distance definition as done in equation 3:

$$K_c(s,t) = \frac{K(s,t)}{\sqrt{K(s,s)K(t,t)}} \tag{3}$$

We have used 6 different values of λ parameters and 25 different linear combinations of combining kernels according to the length parameter to compare names and location fields of the pair of organization records. This results in 150 different ways of comparing the examined textual fields. This allowed us to capture difference in misspellings and occurrence of special characters in the two compared fields. If one of the fields only distinguishes from the other by one special character or simple misspelling the kernel measure of similarity will still grade the similarity between the two fields as very similar. Second set of record comparison vector components is there fore represented by components corresponding to string kernel comparison based on different combinations of kernel parameters. Table 3 displays feature vector assignments for the comparison of pair of records with the names "Jozef Stefan Institute" and "Institut Jozef Stefan".

Kernel Parameters	n=11	n=10	n=9	n=8	n=7	n=6	n=5	n=4	n=3	N=2
Feature Weight	0,005	0,009	0,014	0,023	0,036	0,053	0,077	0,109	0,148	0,194

Table 3: String Kernel Comparison vector of characters in a pair of names: "Jozef Stefan Institute" and "Institut Jozef Stefan". First 10 features are displayed. The displayed features use a static parameter value λ=0.8 with a varying length n of the considered substrings. The similarity increases with smaller lengths of the considered substrings.

To enable comparison of records with wrong word ordering or missing words in their fields we created a third set of features based on the order of the words in the fields of the record pair. This type of comparison is called a string kernel on the words. We again use string kernels to capture the difference in the ordering. This time we first pre-process the strings to represent them by order of the words that appear in them. The string "Institut Jozef Stefan" would this way get represented by the sequence of three numbers: "1 2 3". The string "Jozef Stefan Institute" would then be represented by the sequence of numbers: "2 3 4". In this setting number 1 represents word "Institut", number 2 corresponds to "Jozef", 3 to "Stefan" and 4 to "Institute". The two sequences of numbers "1 2 3" and "2 3 4" can now be compared with the same string kernel technique as it was previously used on the characters.

We have again used 6 different values of λ parameters and 25 different linear combinations of combining kernels according to the length parameter to compare words in names and location fields of the pair of organization records. This results in 150 additional ways of comparing the examined textual fields. It allowed us to capture difference in wrong word ordering or missing words in the two compared fields. If one of the fields is only distinguished from the other by an addition of an extra word or word ordering of the part of the string, then the kernel measure of similarity would still grade the similarity between the two fields as very similar. Table 4 displays feature vector assignments for the record with the names "Jozef Stefan Institute" and "Institut Jozef Stefan".

Kernel Parameters	n=11	n=10	n=9	n=8	n=7	n=6	n=5	n=4	n=3	n=2
Feature Weight	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,064

Table 4: String Kernel Comparison vector of the words in pair of names: "Jozef Stefan Institute" and "Institut Jozef Stefan". First 10 features are displayed. The displayed features use a static parameter value λ=0.8 with a varying length n of the considered sequences. Due to only comparing sequences of words of length 3, which only have 2 words in common, the resulting scores are zero everywhere except when comparing according to sub sequences of length 2.

Edit Distance Comparison

We used the edit distance (Levenshtein 1966) comparison measure to capture the notion of typical typing mistakes in the IST World data - typos. The edit distance between the two strings of the characters is the number of operations required to transform one of them into the other. This way we compare the entities in the database with proximity based on usual human typing mistakes. According to the survey on the data to be integrated (Ferlez et al 2007) this is the right measure we need to capture the differences between true matches of pairs which are a consequence of the manual human input of the data.

In information theory and computer science, the edit distance between two strings of characters is the number of operations required to transform one of them into the other. There are several different algorithms to define or calculate this metric. In our approach we used two most commonly used metrics: the Levenshtein Distance (Levenshtein 1966) and the Damerau-Levenshtein (Damerau 1964) distance.

Levenshtein distance is a string metric which is one way to measure edit distance. The measure of distance between two strings is given by the minimum number of operations needed to transform one string into the other, where an operation is an insertion, deletion, or substitution of a single character. It is named after Vladimir Levenshtein, who considered this distance in 1966. It is useful in applications that need to determine how similar two strings are, such as spell checkers.

Damerau-Levenshtein distance is alternative string metric for measuring edit distance between two strings. Like Levenshtein distance, it finds the difference between two strings by giving the minimum number of operations needed to transform one string into the other where an operation is an insertion, deletion, or substitution of a single character. Unlike Levenshtein distance, it counts transposition as a single edit operation, rather than two. Damerau-Levenshtein distance is therefore equal to the minimal number of insertions, deletions, substitutions and transpositions needed to transform one string into the other. Damerau in his seminal paper (Damerau 1964) not only distinguished these four edit operations but also stated that they correspond to more than 80% of all human misspellings.

To remove any bias introduced by different lengths of strings we normalize the feature vectors by transforming the measures to give values between 0 and 1. We do that by taking into an account that the maximum edit distance equals the length of the larger of the two compared strings. This way we can transform the edit distance measure into edit similarity measure which grows with similarity. Details of the normalization are described by equation 4:

$$S = 1 - \frac{D(s,t)}{\max(|s|,|t|)} \tag{4}$$

We have used the transformed Levenshtein and Damerau Levenshtein Distance to compare organization names and location fields in the pair of potential records of organizations. We first used the measures on the characters in these fields. As said this allows us to compare pairs of strings with regards to the possible human typing mistakes. For example, when comparing records with the names "Jozef Stefan Institute" and "Jozef Stefan Lnstitute" then

the two measures yield the same result of 0,559330761432648 for both the Levenshtein distance and the Damerau Levenshtein distance.

The same measures could be used to compare strings not character wise but word wise. This type of comparison allows us to capture similarity in the compared strings according to missing or added words, and changes in word ordering. For example, when comparing records with the names "Jozef Stefan Institute" and "Jozef Stefan Lnstitute" word wise, then both of the two measures again yield the same result of 0,432607352733612 for both the Levenshtein distance and the Damerau Levenshtein distance.

Feature Combination

We combined the comparison features vectors extracted with the use of the mentioned comparison techniques by reweighing of every set of features to effectively create a function $\gamma: A \times B \rightarrow \Gamma$, which is used for mapping a potential pair of records into a useful comparison vector space Γ. Every feature set originating in different comparison technique was uniformly reweighed to set the norm of the feature vector of the corresponding set to one. After combining the feature vectors of the sets into one vector we again rescaled the features to acquire a normalized feature vector with the norm 1. This unified feature vector contains all the generated features describing the pair of records from many different aspects of similarity.

3.3.3 Active Learning

In the active learning step of the IST World Integration Approach we show how to use the machine learning approach called active learning (Tong et al 2002) to support the manual creation of the set of true matches M and the set of true non matches U mentioned in the record linkage problem definition. That is, in order to prepare the training set of potential pairs for automatic learning of the decision function, a set of potential pairs needs to be evaluated and labeled manually into sets M and U. These sets will later be used by a machine learning algorithm to automatically find a good decision function on the potential pairs of records.

We expect many of the potential pairs of records to be true non matches and only very few of them to be true matches. This is why it is hard to produce balanced learning sets M and U in terms of the amount of information which can be used for learning the decision function. For example if the examples in M only cover one particular type of true matches then it will be very hard to generalize on this information to cover other types of true matches. Another similar example is if the set of true non matches only covers one type of non matches then we can easily end up with the same problem of generalizing from the learning data in U to the rest of the potential pairs.

First possible approach to gather the records for manual labelling is to randomly select a sample of pairs. Because of the imbalance between the two sets we would need to acquire a very large sample to gather sufficiently enough learning data in set of true matches M. Therefore a very large sample of pairs would have to be manually labelled only to capture enough information about the set M.

A better approach is to use a machine learning algorithm for deciding which potential pairs to label into sets M and U. In this approach we exploit the sampling behaviour of the active learning algorithm (Tong et al 2002) to effectively fill sets M and U with labelled pairs based

on the features generated in the Feature Generation Step. The active learning paradigm enables a human to assist the learning algorithm by providing the label on the examples, which the algorithm asks for. One possible strategy of the learning algorithm to most effectively use this source of information is to present to human those examples, for which resulting answers would probably result in the biggest information gain for the learning algorithm. In other words, one strategy of the algorithm could be sampling from the list of all potential pairs, maximizing the expected information gain for the task of inducing a good decision function in the vector space Γ. For example, when the Support Vector Machine (SVM) (Vapnik 1995) is used as the underlying learning algorithm, the algorithm might be tuned to select those examples for inquiry which lie closest to the separating hyper plane in the feature space, in the expectation that these are the examples which might have the most impact on efficiently truncating the version space (Mitchell 1982) of the induced decision function. Thus, the machine learning paradigm Active Learning might enable us to fill the sets M and U with those examples that carry the most information with regards to the examined data and the selected decision problem.

We selected to use Support vector machine (Vapnik 1995) algorithm as the underlying learning algorithm to effectively fill the sets M and U. Selection of the Support Vector Machine as the underlying Active Learning algorithm results in optimal sampling of the list of pairs for maximal information gain about the given classifier induction problem because the same algorithm will also be used for decision function induction.

Support vector machine is a machine learning algorithm which maps input vectors to a higher dimensional space where a maximal separating hyper plane is constructed to achieve best classifier generalization error on the test data. Two parallel hyper planes are constructed on each side of the hyper plane that separates the data. The separating hyper plane is the hyper plane that maximizes the distance between the two parallel hyper planes. An assumption is made that the larger the margin or distance between these parallel hyper planes the better the generalization error of the classifier will be. SVM algorithm uses a parameter called C to signal the importance of the misclassifications in the learning data.

Figure 13: Active Learning Application for labeling pairs of potential matching organizations as true or false matches. The user can answer the presenting question with "yes" (the organization records represent one real life organization), "no" (the organization records represent two different real life organizations), or "I do not know" in case the user is himself not able to answer the question with sufficient confidence. The "Rethink the situation" reruns the learning algorithm to take the new labeling into account and produce an updated set of questions.

We created a tool with a web interface visible in Figure 13 which allows the user to interact with the active learning algorithm. The tool allows the user to answer the questions produced by the algorithm and thus label pairs into sets M and U. What is more the user is able to observe the opinion and opinion confidence of the algorithm on the given candidate pair. We set the SVM parameter C to 1, which is a standard default setting usually producing satisfactory results (Brank et al 2003).

3.3.4 Classifier Induction Step

The goal of the Classifier Induction Step is to automatically construct a decision rule, which solves the record linkage problem as defined in the beginning of this chapter. We try to achieve this goal by using a machine learning algorithm to automatically generalize from vectors $\gamma(a,b) \in \Gamma$ of the examples (a,b) in the M and U sets. The generalized knowledge in the form of a decision function should then be used to label the rest of potential pairs as matches or non-matches. In detail, the induced decision function uses information on comparison features $\gamma(a,b) \in \Gamma$ that were generated in the Feature Generation Step to separate the vector space Γ into two subspaces corresponding to probable matches and probable non matches. Every new example of pair of records (a,b) is then classified according to the subspace which contains the corresponding comparison pattern $\gamma(a,b)$.

The traditional record linkage problem as described in Section 3.1 can be solved by a standard supervised machine learning setting for learning a binary classification model. Using the same notation the following definition captures the record linkage problem in a machine learning setting.

PROBLEM DEFINITION: Learn a classification model for distinguishing between duplicate (class m) and non duplicate pairs of records (class u)

GIVEN:

- two datasets of records A and B
- a test set of record pairs from $A \times B$ with missing class assignment
- a learning set of pairs from $A \times B$ contained in the sets M and U with the corresponding class assignment m and u
- a vector space Γ with function γ to map from $A \times B$ into Γ: $\gamma(a,b) \in \Gamma$, where $a \in A$ and $b \in B$

FIND:

- a classification model T on the records from $A \times B$ with attributes defined in Γ, so that based on the class assignments in the training set, it will classify the examples in the test set with
- Performance requirement: defined as precision (Grossman et al 1998) on the test set to be over 99% while retaining as high recall (Grossman et al 1998) values as possible.

We have used an SVM two-class classification model with linear kernel (Vapnik 1995) to generalize from the information in the training set for automatic labeling of the pairs in the test set. This model was shown (Brank et al 2003) to work reasonably well when the attributes of the examples include bag of words features with added other artificially constructed features like topological feature PageRank (Brin 1998) and others. We therefore expect it to

also perform well on our generated features, which capture the different aspects of comparison of potential pairs.

As already mentioned the Support vector machine (Vapnik 1995) is a machine learning algorithm which maps input vectors to a higher dimensional space where a maximal separating hyper plane is constructed to achieve best classifier generalization error on the test data. Two parallel hyper planes are constructed on each side of the hyper plane that separates the data. The separating hyper plane is the hyper plane that maximizes the distance between the two parallel hyper planes. An assumption is made that the larger the margin or distance between these parallel hyper planes the smaller the error of the classifier will be on the unseen examples. SVM algorithm uses a parameter called C to signal the importance of the misclassifications in the learning data. The mapping of the input vectors to a higher dimensional space is called a kernel trick (Aizerman et al 1964). The kernel trick is a method for converting a linear classifier algorithm into a non-linear one by using a non-linear function to map the original observations into a higher-dimensional space; this makes a linear classification in the new space equivalent to non-linear classification in the original space. In our approach we will be using a linear kernel signaling that a comparison of two comparison vectors s and t translates into computation of a standard inner product between the two vectors: <s,t>.

Our goal is to induce a classifier meeting the prescribed expected precision of 99% on the training examples while still retaining as good recall values as possible. The SVM algorithm tries to maximize the margin between the positive and negative examples to produce best classifier in balancing the trade-of between precision and recall (Grossman et al 1998). The classifier balances with equal importance between the incorrectly classified positive and incorrectly classified negative examples. We therefore need to correct the decision making process of the SVM classifier by moving the decision boundary away from the hyper plane, which separates the positive and negative examples, towards the positive examples. This forces the classifier to only classify those examples as positive which are far away from the separating hyper plane. How far away should they be is determined by the threshold parameter, let us call it P.

All together this means we have to set two parameters C and P in order to learn a classifier for distinguishing between duplicates and non duplicates pairs of records. We used an AUC classifier estimation score and a standard cross validation method to set both parameters.

One way to measure the goodness of a classification algorithm is the AUC (Bradley 1997) score. This measure can be interpreted as the probability that when we randomly pick one positive and one negative example, the classifier will assign a higher score to the positive example than to the negative. The closer the AUC score of a classifier is to 1 the better the classifier.

Cross-validation (Kohavi 1995), sometimes called rotation estimation, is the statistical practice of partitioning a sample of data into subsets such that the testing is performed on a single subset, while the other subset(s) are used in training. Cross-validation is important in guarding against testing hypotheses suggested by the data, the so called "Type III error", especially where further samples are hazardous, costly or impossible to collect. There exist many types of cross validation, according to different ways of partitioning the data which is available. We have used the so called K-fold cross validation. In K-fold cross-validation, the

original sample is partitioned into K sub samples. Of the K sub samples, a single sub sample is retained as test data for testing the model, and the remaining K − 1 sub samples are used as training data. The cross-validation process is then repeated K times, with each of the K sub samples used exactly once as the validation data. The K results from the folds can then be averaged to produce a single estimation. We have used 10 as the value for K. Cross validation is typically used to grade the goodness of a classification algorithm in a conservative way. In the case of K-fold cross validation the algorithm learns from the K-1 sets of data and then tests the induced hypothesis on the one remaining fold of data.

We have performed a search for the best C parameter by trying different versions of the learning algorithm with regards to this parameter. We started with the default value of 1. We then performed iterative steps of increasing or decreasing the parameter and observed the resulting values of AUC obtained with cross validation. The parameter value which resulted in the highest AUC score was selected as the best setting for the final classifier induction.

The threshold parameter P was estimated by averaging the threshold value across the 10 folds. Inside every test fold the threshold was set to the value which effected in 99% precision. The overall threshold was then computed as the average of all the thresholds set in the individual folds.

The Classifier Induction thus works this way: First we use the cross validation to estimate the best C parameter for the SVM Classifier. Then we use the cross validation again to search for the best threshold value P signaling how far away from the hyper plane should the examples be in order to be classified as positive to achieve the expected false positive rate.

3.3.5 Classifier Application Step

The Classifier Application step is used to automatically assign the potential pairs into sets M and U by using the induced classifier. We simply run the classifier on every feature vector $\gamma(a,b) \in \Gamma$ of the unlabeled potential pairs (a,b) from A × B and compare the resulting score against the selected threshold. If the score is bigger than the threshold the pair (a,b) is assigned to the M set of matching pairs, otherwise it is assigned to the non matching set U.

3.4 Approach Evaluation

We experimentally evaluated our approach by applying it to a duplicate detection problem. First we describe our evaluation methodology and then we report on the resulting evaluation scores of our approach.

3.4.1 Evaluation Methodology

Evaluation of the IST World Integration Approach is performed by measuring the percentage of the incorrectly classified examples in the randomly selected subset of the automatically labeled candidate pairs. The classifier goodness is further evaluated by observing the ROC curve on the test sample. The goodness is summed by providing the AUC scores of the classifier on the sampled set of pairs. In this subsection we first describe how we acquired the test set and then describe the mentioned evaluation measures.

Test Set Acquisition

Our test set was randomly selected from the reduced set of potential pairs (a,b) from A × B, which were labeled automatically by the tested classifier. We assigned the correct labels to the classifiers by using manual evaluation of every pair of sampled records (a,b). As the manual resolving of the duplicate detection task is slow and tedious we decided to have a test set of moderate size of 1000 pairs. In order to overcome the problem of very small share of positive matches in the set of all pairs from A X B, we decided to sample from the reduced set of pairs which is a result of the Blocking Step of the Integration Approach. We have reused the tool developed for Active Learning for manual labeling of the candidate pairs by answering the questions posed by the learning algorithm. This time the questions were asked by the algorithm only to provide the human evaluators with good environment for fast data labeling.

Classifier Evaluation Techniques

We have used several standard techniques of evaluating the goodness of the IST World Integration Approach: Precision, Recall, Specificity, ROC Analysis and AUC statistics. The goodness of the approach signals the capability of the algorithm to correctly recognize the duplicate pairs of records to be linked. The task of the approach is to classify the pairs of records as either matching or not matching. This corresponds to solving a binary classification problem, for which standard evaluation measures have already been established. We describe the used measures bellow.

To measure the performance of a binary classification model like ours, the concepts precision, recall or sensitivity and specificity (Grossman et al 1998) are often used. Precision is defined as the percent of positive correctly classified examples: (true positives) / (true positives + false positives). It can be seen as the probability that the example is manually evaluated as positive when automatically classified as positive. Sensitivity or recall measures the proportion of examples that were correctly declared as positive by the automatic classifier (true positives) of all the examples that were manually evaluated as positive (positives); that is (true positives) / (positives). It can be seen as the probability that the example is classified as positive when a manual evaluation would reveal it is in fact positive. The higher the sensitivity, the fewer positive examples go undetected. Specificity is the proportion of examples that were classified negative of all the negative examples tested; that is (true negatives) / (true negatives + false positives). As with sensitivity, it can be looked at as the probability that the example is classified negative given that the example is in fact negative. The higher the specificity, the fewer negative examples are labeled as positive by the automatic classifier.

The relationship between sensitivity and specificity, as well as the performance of the classifier, can be visualized and studied using the ROC curve (Spackman 1989). ROC analysis provides tools to analyze the pattern of sensitivity and specificity trade-offs of the particular classification algorithm. It allows to select possibly optimal models and to discard suboptimal ones independently from the cost context or the class distribution. This evaluation technique is thus very appropriate for the binary classification problem at hand which features a strong imbalance in frequency distribution of the two classes.

To draw an ROC curve, the true positive rate (TPR) and false positive rate (FPR) measures are needed. TPR determines a classifier performance on classifying positive instances

correctly among all positive instances available during the test. FPR, on the other hand, measures how many negative instances were classified as positive among all the negative instances available during the test. An ROC space is defined by FPR and TPR as x and y axes respectively, which depicts relative trade-offs between true positive (benefits) and false positive (costs). Since TPR is equivalent with sensitivity and FPR is equal to 1 - specificity, the ROC graph is sometimes called the sensitivity vs. (1 - specificity) plot. Each prediction result or one instance of a confusion matrix represents one point in the ROC space. The best possible prediction method would yield a point in the upper left corner or coordinate (0,1) of the ROC space, representing 100% sensitivity (all true positives are found) and 100% specificity (no false positives are found). The (0,1) point is also called a perfect classification. A completely random guess would give a point along a diagonal line (the so-called line of no-discrimination) from the left bottom to the top right corners. An example of a ROC curve, borrowed from the Wikipedia, is depicted in Figure 14.

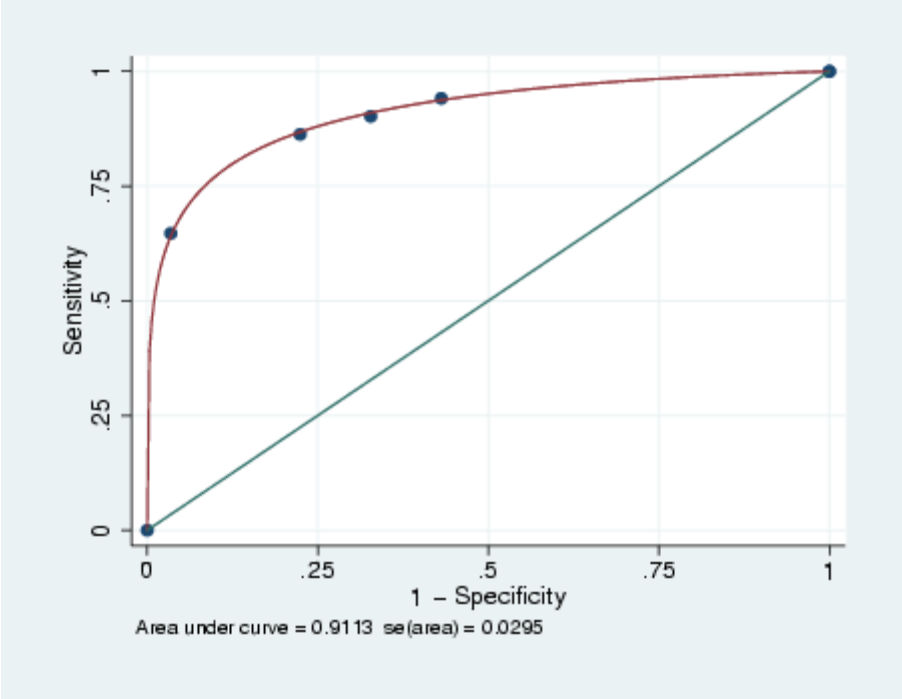


Figure 14: An example of a ROC Curve shows the selection of the sensitivity and specificity trade-of by selecting a classification threshold.

The ROC can also be used to generate a summary statistic of a goodness of a classification algorithm. The machine learning community most often uses the ROC AUC statistic. As the AUC name suggests this statistics measures the Area Under the (ROC) Curve. This measure can be interpreted as the probability that when we randomly pick one positive and one negative example, the classifier will assign a higher score to the positive example than to the negative. For example the AUC statistics of the ROC curve depicted in Figure 14 is 0.9113.

3.4.2 Duplicate Detection Experiment in CORDIS FP6 Dataset

We have used the IST World Integration Process described in the previous section for cleaning the data from the publicly available CORDIS FP6 (Thévignot 2000) data source. We evaluated the success of the cleaning using the evaluation measures described in the previous section.

The CORDIS FP6 data source contains information on organizations and projects funded by the European Union in the scope of 6th Framework Programme. The data source contains a unique list of the 5094 funded projects. The list of organization collaborating in them is not unique across the dataset, but is maintained with every individual project. All together 24965 different organisations are mentioned. As the data about the collaborating organizations is entered manually per every project this results in inconsistency in the naming and spelling of the individual organizations. Therefore the problem of duplicate records of the same organization emerges, e.g. Jozef Stefan Institute, Institut Jozef Stefan, Lnstitute Jozef Stefan are all mentioned in the data source as different organizations instead as a single organization Jožef Stefan Institute. Figure 15 displays an example of two project web pages, with overlapping organizations participating in them. In this example it is easy to spot the duplicate record of Jozef Stefan Institute (Institut Jozef Stefan). These two duplicate records combine the difference in the spelling of organization names and its word ordering.

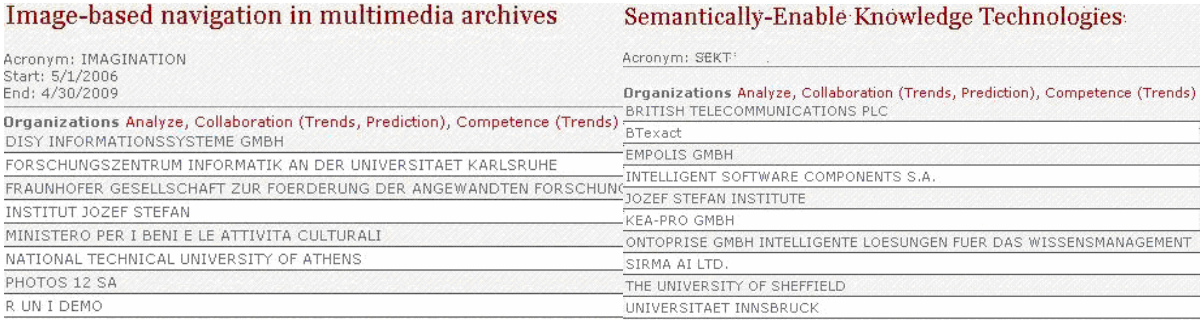


Figure 15: An example of two records of two projects from CORDIS database with overlapping participating organizations. Jožef Stefan Institute appears in both lists, however in two different spelling variants: INSTITUT JOZEF STEFAN (*left*) and JOZEF STEFAN INSTITUTE (*right*).

Our task was to resolve the duplicate record problem of such organizations in the CORDIS data source. These records need to be cleaned by merging them into a single record. We used and evaluated the IST World Integration Approach on this task.

First, **Blocking Step** was used to get the potential pairs of records pointing to the same real world organizations. For every one of 24965 organizations we produce a list of at most 10 most similar organizations according to the full text querying mechanisms on their names. After performing full text search in the full text index of the organization names a list of around 220,000 pairs out of 462,000,000 of all possible pairs was created. This list featured the pairs of organizations with highest similarity in their name.

Feature Generation Step was used to construct numeric features on every potential pair by taking into an account different similarity aspects of name and country of origin of the organizations in the pair. All together a sparse set of features was constructed for every pair out of an overall set of around 20 000 features.

The **Active Learning Step** was then used to assist the human in creating relevant labeled training dataset for learning a decision function on all the rest of the potential pairs of records. We successfully used the Active Learning Application to answer around 2000 questions posed by the algorithm which produced 700 positive and around 1300 negative data labels.

In the **Classifier Induction Step** the labeled pairs were used to learn a model of the sought

decision function. We have used the SVM algorithm with linear kernel as the model for our decision function. The 10-fold cross-validation was first performed on training data to find the best model induction parameter C. After investigating the goodness of classifiers based on the AUC measure when changing the parameter C, we decided to fix the parameter at the value 15. This signaled the classifier to try harder to avoid the miss classifications in the training data. Goodness of such classifier calculated in 10 fold cross validation measured by the average AUC was 0.91. As the precision was set to be 99 percent, we used another 10 fold cross validation to calculate the average classifier threshold to meet this requirement. The average threshold was found to be 1.58 and corresponded to 38 percent average recall on the training data.

In the **Classification Step** we applied our induced classifier to all the still unlabeled pairs. All the pairs that were classified further than threshold away from the induced separating hyper plane in the comparison space were marked as positive pairs and all the other pairs as negative pairs. In this way, we automatically identified around 5,000 pairs out of 220,000 possible pairs as the records which are describing the same real world entities.

We evaluated our model by random sampling and hand evaluating additional 1000 pairs of records. The comparison of manual and SVM assigned labels yielded 30 TP, 1 FP, 35 FN and 934 TN pairs. The observed precision was found to be 97% and recall 46%.

Figure 16 displays the ROC curve of the classifier on the evaluation sample. The model achieved an AUC score of 0.96.

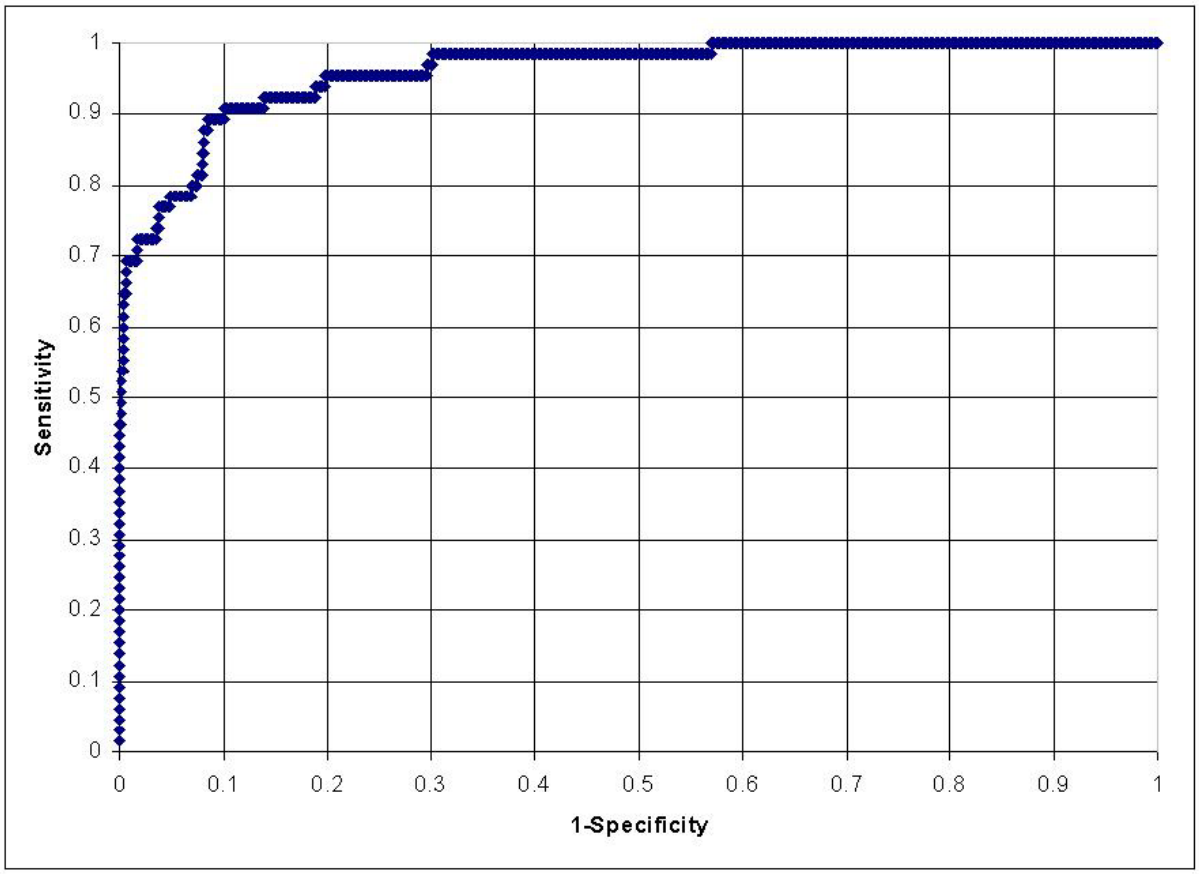


Figure 16: The ROC curve of the classifier on the CORDIS FP6 evaluation sample. AUC of this curve is 0.96.

The experiment showed that the developed integration approach works well. Our model performed slightly worse than expected in terms of the targeted precision rate of 99 % - it achieved a precision score of 96%. The model slightly over performed in terms of the expected recall on the test data, as measured on the training data (38%), achieving the recall of 46%. The difference between the targeted and acquired precision and recall probably originates in the difference of distributions between the learning and test data. The data in the learning set is centered on the separating SVM hyper plane because of the nature of learning set acquisition via active learning. The precision and recall estimation on learning data with cross validation is thus heavily biased by the examples which lie close to the hyper plane. The test data on the other hand is randomly sampled from the unlabeled dataset, meaning that more data points lie further away from the hyper plane. According to the higher recall there are more positive examples lying further away from the hyper plane than anticipated from the learning data.

The ROC curve in Figure 16 of the classifier on the test data looks smooth. The turning point of the curve lies approximately at specificity of 0.91. However due to targeted high precision the turning point of the curve could not be used as threshold for classifier induction. The ROC curve reveals another pattern as it features two turning points first at 0.7 sensitivity level and another one at 0.9 sensitivity level. This might indicate that there exists a special class of false positives which could be prevented by further analysis. The AUC score of 0.96 reveals that when a positive and a negative example are compared the classifier almost always correctly orders the positive as the one more probably a match than the negative one.

The conducted experiment proved that the developed IST World Data Integration approach works well for the task of solving the duplicate detection task in one of the data sets contained in the IST World database. It facilitated semi automatic identification of 4000 duplicate pairs with very high accuracy and reasonable recall.

The approach will be used in the future to solve the record linkage problem in the IST World database. The Feature Generation step can be adapted to incorporate any assumptions on the patterns in matching records. These similarity assumptions can simply be coded using the corresponding Feature Generation Step. This allows using the same approach on any record linkage or duplicating detection problem regardless of the assumptions on the similarity of the records.

The described integration methodology could be further developed by employing a transductive instead of inductive learning function. As all the test data is already know in advance it could be used to generate even better model for classification. Transductive SVM machine introduced by (Vapnik 1999) could be used for these purposes. These would also enable to better take into account the imbalance in class distributions. It would however create a dilemma whether the transductive setting should also be used in the Active Learning step of the approach. The Feature Generation step could be improved using the collaboration information as a way of comparing the two records. It could also be redesigned for every particular data integration problem at hand. Transfer learning methods might enable reuse of one integration model on several instances of integration problems.

4 Data Analysis

In this chapter we describe the data analysis methods used to identify topics of research work, collaboration patterns and how these topics and patterns change with time. We begin by formally describing the analysis goals of identifying competence, consortia, competence development and consortia development. We continue with description of the data mining methods used for reaching these goals. We conclude with a series of experiments which show that the results of the data mining methods agree with human intuition.

4.1 Analysis Goals

The data analysis in the scope of the IST World portal is centered on information provision for supporting the partner search process. The IST World system tries to assist its users by automatically providing information on actor's topic of work and main communities of collaboration.

When considering partners for knowledge transfer the following questions concerning topic of work and the social network of a potential partner are usually asked: Which are the main topics of an individual actor's work? Which are the main past collaborations of an actor? How did the actor's topics of work and collaboration patterns change with time? What is more, the actor's work description and collaboration patterns description depend heavily on the context of the inquiry. Therefore the context of the partnership in which the individual actor is considered becomes important when acquiring information on the actor's topic of work and its collaborators.

The goal of the data analysis in the IST World portal is to answer the above questions automatically. In the following sub sections we formally define the analysis goals of automatically identifying **competence**, **consortia**, **competence and consortia importance development** of an actor. We then describe the text analysis and graph analysis methods which can be used for analyzing the research data. Next we present the use of these algorithms for automatic extraction of **competence**, **consortia**, **competence development** and **consortia development**. We illustrate the effectiveness of the algorithms on several examples and show that the results agree with human intuition.

4.1.1 Competence

Which are the main topics of an individual actor's work? Usually the answer depends on an implicitly or explicitly defined context of other actors. We illustrate this idea by an example of analyzing the topics of research work at Jožef Stefan Institute (JSI). If the context of the analysis is defined to consist of all the Slovenian academic institutions, then we say that the institute is mainly active in the fields of physics, chemistry, robotics and informatics. When we compare the JSI with all the Slovenian organizations which are active in the informatics field, then the topic of works are: natural sciences, robotics, data mining, intelligent systems and network security. Notice how the chemistry and physics are not so important in this context so they are replaced by single "natural sciences" keyword. On the other hand the informatics field is expanded into details which describe what areas in particular is the JSI most interested as a research actor in the informatics field. Therefore, it is essential to define the analysis context before performing the analysis of the actor's topic of work.

To capture the importance of context in the topic of work analysis we introduce a term **competence**. We define competence as follows: *Competence of a particular actor is the most general topic of work, which distinguishes the actor from the majority of other actors in the context of the analysis. Every topic of work is identified by keywords which best describe the competence*. In short, competence of an actor is the topic of work in the context of other actors. Reusing the previous example, we say that the competences of the JSI in the context of the Slovenian academic institution are physics, chemistry, robotics and informatics. Its competences in the narrower context of the Slovenian organizations active in the informatics field are: natural sciences, robotics, data mining, intelligent systems and network security.

We will analyze the textual data records which describe actor's activity to automatically identify the competence of an academic or industrial actor. Our goal is to automatically extract keywords, which best describe the competence as it is described above, from the textual description found in these records. Thus, the selected words should be as general as possible in their semantics but still able to distinguish the actors between them selves. For instance, the competence of the Jožef Stefan Institute in the context of the institutions involved with informatics can be highlighted with words extracted from this organization's records of work like: earth, artificial, intelligence, model, security and communication.

Let us now formally define the problem:

PROBLEM: **Competence Identification.**
GIVEN: A set of textual documents, corresponding to aggregated textual description of actors' records of work,
IDENTIFY: for every actor, most general topics of actor's work, where every such topic is associated more with a particular actor than with the majority of other actors in the analysis scope. Topics must be identified with extracted keywords, which most generally describe the topic.

In the second section of this chapter we show how to effectively use a series of dimensionality reduction algorithms on textual data for solving this task.

4.1.2 Consortium

Which are the main past collaborations of an actor? Again the answer depends on an implicitly or explicitly defined context of other actors. Usually, we would like to identify the main collaborations of a particular actor in the context of a collaboration graph, defined by other actors in the analysis scope. Let us illustrate this by an example of analyzing the collaboration of the Jožef Stefan Institute (JSI). When the context of analysis is defined by collaboration graph of all the Slovenian academic institutions, then we say that the institute is mainly active in the community led by the JSI. This includes collaborators like the Faculty of Mathematics and Physics, Department of Experimental Particle Physics and Solid State Physics Department. Another two communities the institute is involved in are led by the University Clinical Centre and the National Forestry Institute. On the other hand, when we analyze collaboration of the JSI in the smaller context of collaborating organizations participating in project named Experimental Particle Physics, then the important communities are found to be the one led by the JSI and the Faculty of Mathematics and Physics, and a disjunctive one led by the JSI and the Department of Experimental Particle Physics. Notice

how the community of Faculty of Mathematics and Physics, Department of Experimental Particle Physics and Solid State Physics Department, as we observed it when using larger analysis context, is expanded into two disjunctive communities, when observed in a smaller context.

To capture the importance of context in the analysis of collaborating communities we introduce a term **consortium**. It is defined as follows: *Consortium of a particular actor is the biggest intensely collaborating community of actors, with which the analyzed actor collaborates more than majority of other actors in the context of the analysis. Each consortium must be identified with a few actors, which collaborate most inside one consortium.* In short consortium of an actor is the collaborating community in which the actor is more involved than other actors in the context of analysis. Reusing the example from the previous section, we say that the consortia of the JSI in the collaboration context of Slovenian academic institutions consist of a community led by the JSI, Faculty of Mathematics and Physics, Department of Experimental Particle Physics and Solid State Physics Department. Two additional consortia include the one led by the University Clinical Centre and the one led by the National Forestry Institute. If we narrow the context of analysis to organizations which collaborate on a project named Experimental Particle Physics, then the consortia of JSI can be found to consist of the community led by the JSI and the Faculty of Mathematics and Physics, and a disjunctive one led by the JSI and the Department of Experimental Particle Physics.

To automatically identify the consortia of an actor we analyze the records which describe this actor's collaboration. Our goal is to extract those names of actors in analysis context, which are most suitable for description associated with the analyzed actor. For instance, one collaborating consortium of the JSI in the context of participants in a project named Experimental Particle Physics might be identified with the organization named Faculty of Mathematics and Physics.

Let us now formally define the problem of consortium identification:

PROBLEM: **Consortium Identification.**
GIVEN: A set of actors and a collaboration graph between them
IDENTIFY: for every actor, the biggest communities of collaboration, where with every such community a particular actor collaborates more than the majority of other actors. Every community must be described with a few most collaborating actors in the community.

In the algorithm description part of this section we show how to effectively use a series of graph to matrix transformations and dimensionality reduction algorithms for solving this task.

4.1.3 Competence and Consortium Development

After identifying competence and consortia of the examined actor we would also like to answer the following question: What is the temporal development of the identified patterns? Is the competence of an actor in the specified field only recent or has the actor been involved with the field for a longer time? The same question may arise concerning the identified collaboration consortium. Is the identified collaboration pattern an artifact of a dying cooperation or is its intensity increasing?

Let us now define the concepts of **competence development** and **consortium development**.

Actor's competence development is a trend of a share of a certain topic of work in the overall work performed by the actor in the context of the work of other actors. In short competence development measures relative trends in importance of a certain competence. For example, competence in the field of information technologies of the Jožef Stefan Institute, when analyzed in the context of all Slovenian academic institutions might be increasing.

Actor's consortium development is a trend of a share of a certain collaborating community in the actor's overall collaboration in the context of the collaboration of the other actors. In short consortium development measures relative trends in collaboration with a certain consortium. An example of consortium development could describe the declining trend of collaboration between the Jožef Stefan Institute and the Faculty of Computer Science in the scope of collaboration between all European institutions.

Let us first formally define the competence development identification problem:

PROBLEM: **Competence Development.**
GIVEN: A set of actors which constitute the context of the analysis, a textual description of each actors' work and a time label associated with every textual description,
IDENTIFY: the competences of each actor and the temporal development of their relative importance

Following is the Consortium Development identification problem definition:

PROBLEM: **Consortium Development.**
GIVEN: A set of actors which constitute the context of the analysis, records of each actors' collaboration and a time label associated with every collaboration record
IDENTIFY: the consortia of each actor and the temporal development of their relative importance

In the next section we show how to use a hierarchical k-means (Kanungo et al 2002) clustering algorithm to present the development of the identified competences and consortia through time.

4.2 Analysis Methods

This section will review the data analysis techniques for visualizing textual and graph data, which will be used with the purpose of identifying competence, consortia, competence development and consortium development of the analyzed actors. After presenting methods for representing textual documents, a selection of text visualization techniques is described, based on the dimensionality reduction and clustering techniques. This enables competence and competence development analysis. Next we describe how graphs can be represented in a similar way as textual data. This representation enables us to reuse the existing text analysis methods also for the analysis of graph data to enable consortium and consortium development analysis.

In the past many methods have been suggested for analysis of the research related data or data with similar properties. Large text corpora analysis (Rebholz-Schuhmann et al 2005) and big social network temporal developments have been studied (Leskovec et al 2005). Spectral data

analysis (Golub et al 1970), unsupervised clustering (Grobelnik et al 2005B) and latent semantic indexing combined with multidimensional scaling (Fortuna et al 2005) were used for effective visualization of textual data. What is more, a topic of research work can be identified with searching for the most important words in a corpus (Brank et al 2002). Unsupervised clustering of citation data can produce diagrams showing the change in the strength of the research topics over the years (Caruana et al 2005). Machine learning methods (Leskovec et al 2004) were researched for an automated text document summarization. Methods for text visualization can also be used in the domain of large graph visualization (Mladenič et al 2003, Pajntar 2006).

In this thesis we describe the used data mining methods, which rely on algorithms for text visualization described in (Fortuna et al 2005), extended in a novel way with a method for main patterns identification and visualization. Next we describe the methods for visualization of clustering, similar to those described in (Caruana et al 2005), but extended with recursive clustering and interactive functionality. Overall the methods described in this thesis differ from the methods mentioned in the previous paragraph by enabling more interactive data analysis and by adaptation for their use on graph data.

4.2.1 Text Representation

Analysis of topics in textual document collection does not require the full richness of the content of natural language texts (Wallach 2006). 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 representation.

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 document d_n is a vector of weights $(w_{1n},\ldots,w_{|W|n})$. In the simplest case, the weights $w_{in} \in \{0, 1\}$ denote the presence or absence of a particular term in a document. More commonly, w_{in} represents the frequency of the word i in the document n. 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| / docfreq(D, i)), where docfreq(D, i) is the number of documents from D in which the word i occurs in. It can be used by itself, or be multiplied with term frequency as used in the popular TFIDF measure of word importance in the BOW setting.

Besides words, n-grams may also be used as terms in the vector document representation. These are phrases represented as sequences of n words (Manning and Schütze 1999). 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 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 (Mladenič 1998), iterates over n, generating all possible n-grams from known (n–1)-grams, immediately discarding all which appear too infrequently in the document set.

For analysis of topics of work in research related data, we decided to represent the textual records using the TFIDF measure of importance of words in the BOW presentation. We extended the vocabulary of the BOW with the most frequent 2-grams.

4.2.2 Text Visualization

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 (Grobelnik and Mladenic 2002), e.g. showing the similarity structure of documents in the collection, showing the time line of topic development through time, showing frequent words and phrases relationships between them, etc. This section focuses on techniques utilizing the analysis of patterns in the BOW presentation of the documents. It continues with presentation of visualization possibilities offered by latent semantic indexing (Golub et al 1970), multidimensional scaling (Borg et al 1997) and document clustering (Kanungo et al 2002).

Latent semantic indexing and Multidimensional Scaling

To visualize a whole document collection, it is preferable to employ some kind of projection method which maps the documents from their high-dimensional vector space of terms to two, possibly three dimensions, at the same time preserving their distance relations. Two approaches to such document projection and visualization are the subject of the remainder of this section.

"A well known and used approach for extracting latent semantics (or topics) from text documents is Latent Semantic Indexing (LSI) (Landauer et al 1998). In this approach we first construct term-document matrix A from a given corpus of text documents. This is a matrix with vectors of documents from a given corpus as columns. The term-document matrix A is then decomposed using singular value decomposition (SVD) (Golub et al 1970), so that $A = USV^T$; here matrices U and V are orthogonal and S is a diagonal matrix with ordered singular values on the diagonal. Columns of matrix U form an orthogonal basis of a subspace in the BOW space where vectors with higher singular values carry more information -- this follows from the basic theorem about SVD, which tells that by setting all but the largest k singular values to 0 we get the best approximation for matrix A with matrix of rank k). Vectors that form the basis can be also viewed as concepts and the space spanned by these vectors is called the *Semantic Space*. Each concept is a vector in the BOW space, so the elements of this vector are weights assigned to the words coming from our documents. The words with the highest positive or negative values form a set of words, which is found most suitable to describe the corresponding concept."(Fortuna 2005) The resulting k concepts are then used to represent original documents in the new *Semantic Space*. The k number of concepts forming the new space is the only parameter of this method.

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 (Fortuna et al. 2005). The Document Atlas system (Fortuna et al. 2005) proposes an alternative, employing Multidimensional Scaling (MDS) (Borg et al 1997) to reduce the k-dimensional *Semantic Space,* with k > 2, obtained by LSI down to two dimensions. MDS algorithm achieves this by placing the documents into two dimensions by minimizing some energy function, for instance:

$$E = \sum_{m \neq n} \sqrt{L_2(d_m, d_n) - L_2(X_m, X_n)} \qquad (5)$$

In (5) d_m and d_n are documents in the "semantic space" obtained by LSI algorithm, and x_m and x_n are projection points of the same documents in the two-dimensional space. Function $L_2(x,y)$ measures Euclidian distance between vectors x and y. We use the gradient descent algorithm to find the best x_m and x_n for every m and n in order to minimize the energy function (5).

Figure 17 shows the DocumentAtlas visualization of a collection of documents describing the 6th framework programme of European IST projects (Thévignot 2000). 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 in the point. The system offers a more detailed view of the common words, which can be obtained by using a mouse to move the dark circle to a desired area on the map.

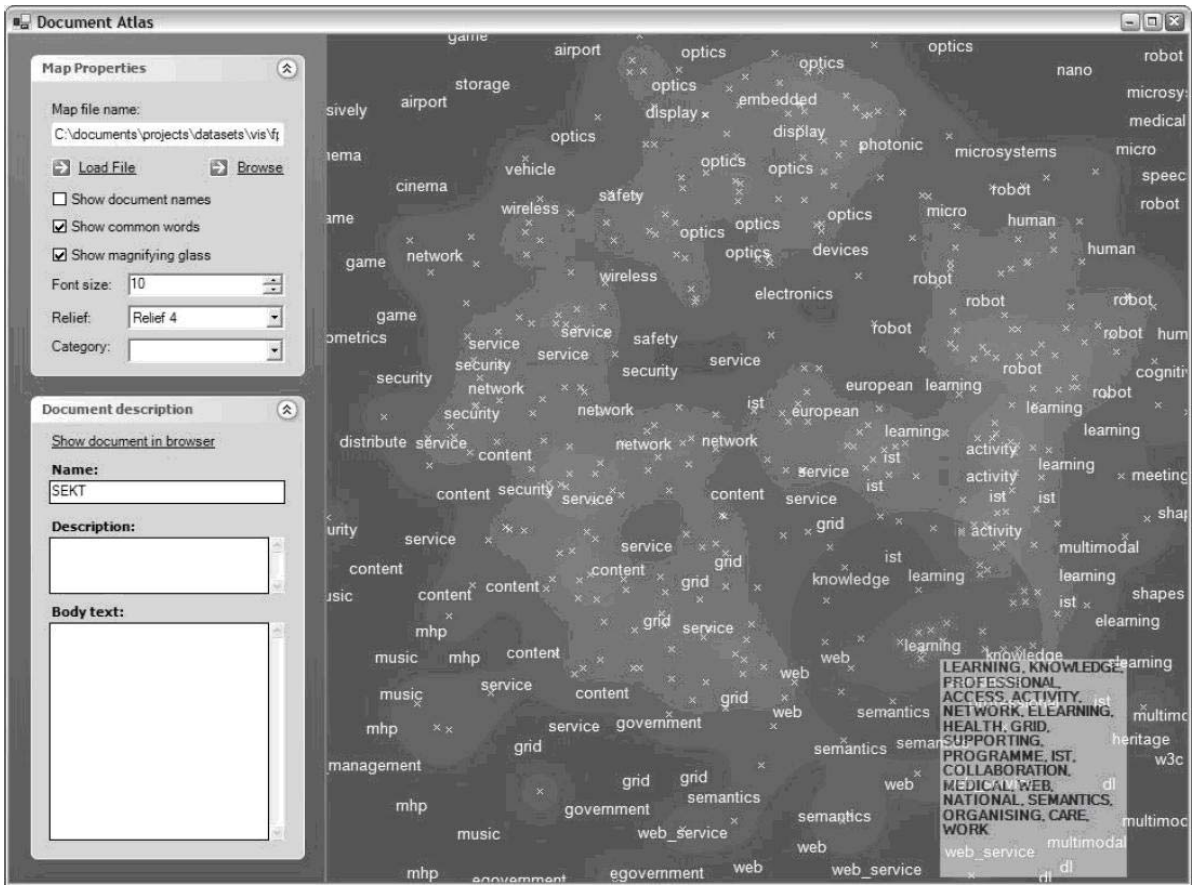


Figure 17: Document Atlas visualization of European IST projects from the 6th framework.

DocumentAtlasBG

We have extended the DocumentAtlas algorithm to allow identification and visualization of the main textual topics in the analyzed corpus of documents. The algorithm is named ***DocumentAtlasBG.***

In (Fortuna et al 2005) the authors mention that the most important components of the singular vectors in SVD computation correspond to main topics found in the documents. For

instance, when analyzing the second singular value vector the often observation is that the largest positive components describe one set of related topics while the set of negative components describes another set of topics unconnected to the first set.

If such topics are identified it is beneficial to include them into the text visualization. Therefore we developed an innovative method of visualizing documents together with the most important topics in the document collection. Let us first formally define the problem:

PROBLEM: Visualize textual documents together with the most important topics.
GIVEN: a set of textual documents represented with the BOW feature vectors,
FIND: a set of most common patterns (topics), their association with each of the documents' BOW vectors and visualize them together with the documents in a 2D space.

We have developed an algorithm for solving this problem named ***DocumentAtlasBG***. The name indicates that the background of the visualization represents the identified topics in the document corpus. The DocumentAtlasBG uses the standard SVD and MDS algorithms exactly as described in the DocumentAtlas algorithm. It however employs an additional twist. Not only are the documents subject to Multidimensional Scaling, but also a list of artificial documents called **topics** is added to the visualization. The **topic** documents are constructed from the singular value vectors which are a bye product of the SVD dimensionality reduction step. Every singular value vector corresponds to two topic documents. The positive components of the vector correspond to the feature vector of the first artificial topic document, and the negative components of the vector correspond to the feature vectors of the second artificial topic document. This way we acquired twice as many topic documents as there are dimensions after the SVD step of dimensionality reduction.

Every artificial topic vector corresponds to the most important patterns captured by the given lower dimensional embedding. The main property of the SVD decomposition is that the process assures minimal information loss when projecting the data into lower dimensional space (Golub et al 1970). This can only be achieved when the new lower dimensional space captures the most important patterns in the analyzed feature vectors. The lower dimensional space is defined by the singular vectors of the decomposition. It is these vectors which carry the information on the most important patterns in the feature vector set. When we artificially create the two vectors out of a single singular vector, we create two documents which correspond to the most important patterns found in the data.

The inner product between the artificial and the analyzed documents corresponds to the importance of the artificial pattern in the examined document. This follows from the fact that the components of any of the compared features vectors cannot be smaller than zero. When the inner product is high this corresponds to the inclusion of the pattern in a document. This allows measuring the importance of every topic with every document.

The artificial topic documents and the original documents can be visualized in 2D with the use of the classical MDS step. As both are described with the feature vectors, we are allowed to capture the distance between them the same way as with the set of original documents. Once the distance matrix is computed the appliance of the MDS step is straightforward to produce a 2D embedding of documents and the identified topics.

The distance or importance between the topics and the documents allows including another type of interactivity into the computed diagram. When the user selects a certain document, the topics which are associated with this document are highlighted. This produces a visualization which is very effective in providing information on which are the main topics in the analyzed documents. Moreover the interactivity can also serve as the guide to similar documents in terms of common topics. This can be achieved by inverting the interactivity visualization by selecting a specific topic in the diagram and highlighting the corresponding documents. An example of the output of the DocumentAtlasBG analysis is displayed in Figure 18.

Competence Analysis with the DocumentAtlasBG algorithm

The DocumentAtlasBG algorithm can be used to tackle the problem of identifying competence of actors as we defined it in the first section of this chapter. This is because the DocumentAtlasBG algorithm enables efficient topic of document identification in the scope of all the other analyzed documents as described in the previous paragraphs. This is exactly the solution to the problem of identifying topics of work in contextualized setting as we defined it in the Competence Identification problem.

The visualization technique first tries to identify the main patterns in the textual description of all actors in the context and then calculate the strength of every actor with each of these patterns. The parameter of the LSI method is set to at most 20 and can be less if there are less documents to be analyzed. At last the patterns and the entities are transformed into a two dimensional space by preserving their similarities and relationships as much as possible. Details of the algorithm are described in the previous section.

DocumentAtlasBG algorithm allows interactive competence identification by means of positioning the actor in the visualization and by displaying appropriate keywords describing the identified competence in the selected area of visualization. The 2D visualization diagram like the one depicted in Figure 18 shows documents or in our case actors on the background of competence of all the actors in the analysis context. The topics of work are shown as dark clouds. Actors are shown as crosses. The position of each actor is associated with the most important topics. This signals the actor's major competence. The visualization is interactive and allows identifying every actor’s competence by highlighting corresponding topic clouds, when the user drags the mouse cursor over the actor. E.g. in Figure 18, the most characteristic computed research topics of Luc de Raedt in the context of the people involved with "machine learning" are Probabilistic_Logic, Inductive_Queries, Logic Learning etc. This signals that Luc de Raedt's competence in this context is identified as Probabilistic Logic Programming, Inductive Queries and Logic Learning.

The textual content, related with the actors, which is the input to the DocumentAtlasBG algorithm, can usually be extracted from various types of records describing the work. In case of the academic actors the records are usually publications published in academic journals or scientific conference proceedings. Another type of records of academic activity is also research project descriptions in which the academic actors participated. The textual content of a scientific paper or a project description can be used for the purpose of competence identification. Assessing the records of activity of the non academic partners is more difficult as usually they do not publish detailed information on the business projects and the topics of their research work. Usually the reason lies in preserving the secrets of the trade, which are the main assets of the high technological industrial actors. Instead these companies usually

provide short description of their expertise in the form of keywords. Therefore these keywords can be analyzed as well to automatically identify industrial actor's competence.

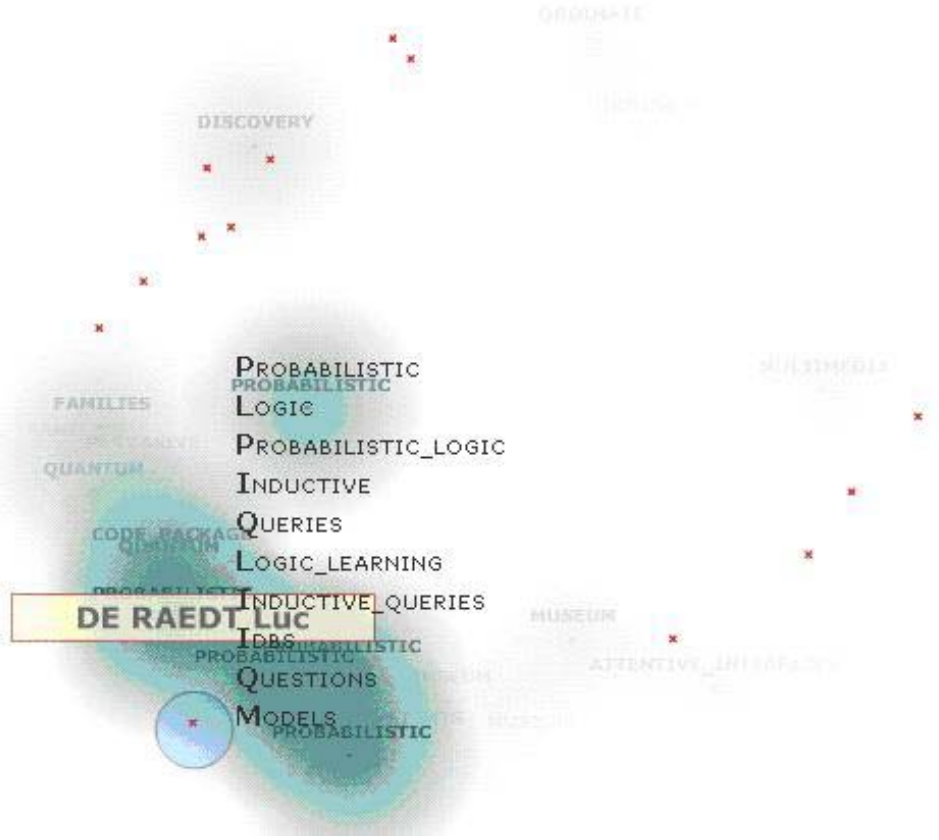


Figure 18: Competence identification with DocumentAtlasBG visualization in Competence Diagram.

Currently only one value of the parameter of the DocumentAtlasBG algorithm is used when running the algorithm in the scope of the IST World portal, due to required simplicity of use of this analysis technique. This parameter sets the number of *Semantic Space* dimensions at 20 in the SVD computation. This parameter can be smaller, if the number of the analyzed documents is less then 20.

Clustering

The basic k-means clustering algorithm (Kanungo et al 2002) is one of the oldest and simplest clustering algorithms to be applied to text, which still produces good results because of its fast execution and apparent suitability to textual data distributions. It involves randomly choosing k points to be the centroids of clusters, and grouping documents around centroids based on proximity with regards to a certain distance/similarity measure. Then, centroids are iteratively recomputed for each cluster, and documents regrouped until there is sufficiently little change in centroid positions. This algorithm depends heavily on the choice of k, and the initial positioning of centroids, therefore the whole k-means algorithm can be run several times with different choices of k and/or initial centroids, and final cluster assignments determined on the basis of the independent clustering results.

Once clusters have been established they can be automatically labeled with the terms best describing them. These can be determined simply by taking terms from the cluster centroid that has the highest weight. As the labels are selected in the context of other clusters and actually describe the topics of the clusters, this algorithm enables us to identify competence as defined in the beginning of this chapter in the first section of this chapter.

A hierarchy of clusters may be obtained by repeated application of k-means, or by using some other inherently hierarchical clustering algorithm. It is also possible to integrate the expert knowledge of a human analyst into the process of cluster derivation, by allowing a semi-automatic interactive approach to cluster generation and visualization, like for instance in the OntoGen tool for semi-automatic data-driven ontology construction (Fortuna et al 2006).

What is more, when temporal information is associated with every document classified into a hierarchy of clusters, then the visualization showing the relevance of a cluster in a certain time period can be very beneficial for the temporal topic analysis as this is shown in (Caruana et al 2005). An example of such visualization is displayed in Figure 19. The idea is very simple. In every time period the cluster's presence corresponds to the share of its documents which belong to the time period. If none of the contained documents is present in the selected time period then the cluster itself is not present in it. On the other hand if some of the documents in a clustering are present then the share of the documents with comparison to other cluster documents is visualized. This enables us to identify the trends of importance of the clusters.

Competence Development Analysis with Clustering

Our task is to identify the competence development of the selected actors as it is defined in the Analysis Goals section of this chapter. That is we seek to identify the development with time of the importance of the topics distinguishing between the records of work of the actors in the context of the analysis.

We achieve this goal by visualizing clusters of the related research documents over time. We have used the mentioned clustering visualization method (Caruana et al 2005) which shows temporal cluster development. For clustering we used the hierarchical k-means algorithm (Kanungo et al 2002) minimizing intra-cluster variance until convergence, which hierarchically split all the documents into a binary clustering tree. This corresponds to using k=2 value for the k parameter of the k-means clustering. This setting is fixed due to required simplicity of the usage of the tools inside the IST World portal. The visualization algorithm then calculates the strength of every cluster in the hierarchy per every time period. The time slots are distributed uniformly over the time span of the analyzed document corpus.

The acquired Competence Development Diagram, e.g. the one displayed in Figure 19, shows research competence strength development through time. The temporal strength development of each research competence is identified with a shaded region across the diagram. Every colored region is described with keywords in the lower right hand side of the diagram. The association between the clusters and the researchers is indicated in the lower left hand side. The order of the displayed researchers is associated with their strength in the field of the selected research topic. Clicking on the research topic region allows a more detailed insight into topic development by further splitting the topic region into sub topic regions.

The textual description information of the analyzed actors is usually time stamped. It can be extracted from various types of records describing the work of the actors. In case of the academic actors these records are usually the publications published in academic journals or scientific conference proceedings. They contain textual information (title, abstract), collaboration (authors, affiliated organizations) and temporal information (date of publication). Every record about an academic research project includes the information on the project's content in the form of a textual description, collaborating participants, and duration of the project. The textual data with temporal stamps can be used for tracking the development of competence.

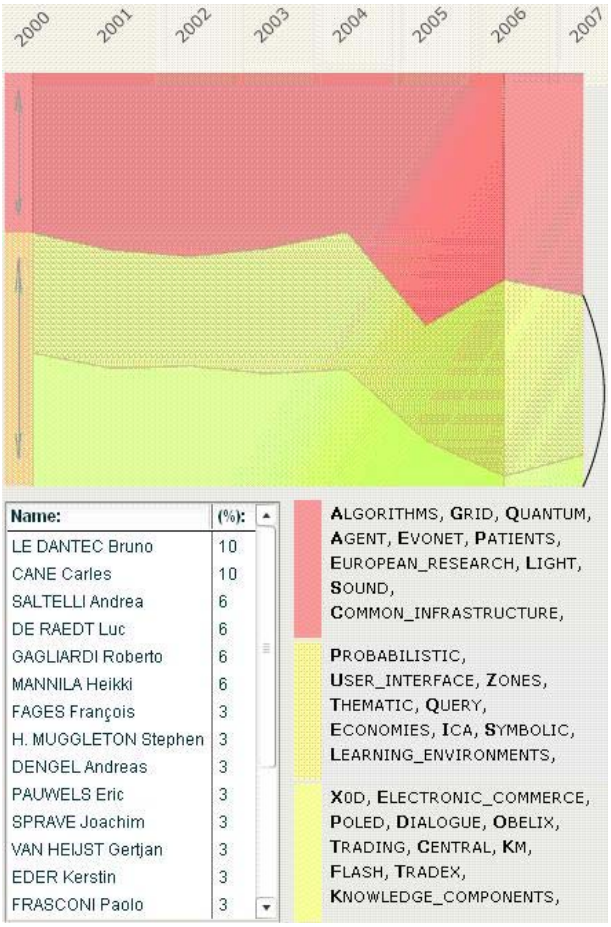


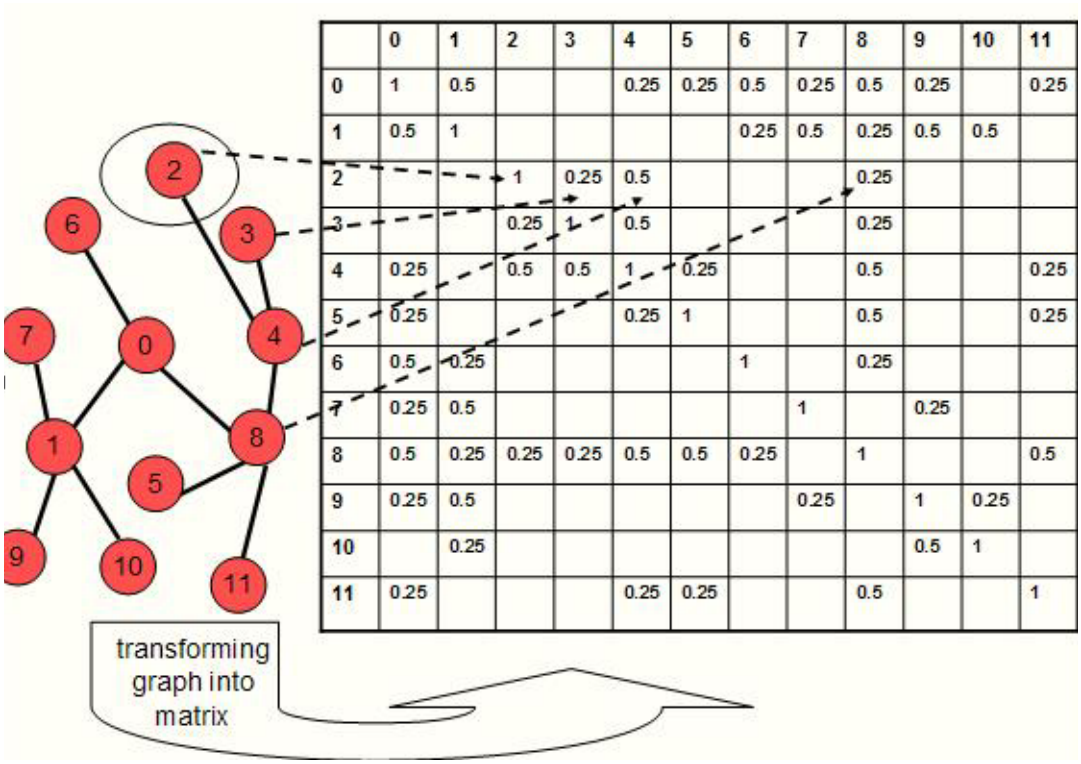
Figure 19: Temporal development of competence is visualized with document clustering algorithm in Competence Development Diagram.

4.2.3 Graph Representation

Representing a graph in the form in which its vertexes correspond to feature vectors enables us to apply some of the existing text mining methods for analysis of the graph structures. When considering a particular graph, the task is to describe each vertex in the graph with a feature vector.

For the purpose of graph representation one can adopt the techniques proposed in (Mladenic and Grobelnik 2005). Graph on *N* vertexes is represented as an *NxN* matrix. The matrix is constructed so that the *X*th row gives information about vertex *X* and has nonzero components for the columns representing vertexes from the neighbourhood of vertex *X*. The neighbourhood of a vertex is defined by its (restricted) domain. The domain of a vertex is the

set of vertexes that are path-connected to the vertex. More generally, a restricted domain of a vertex is a set of vertexes that are path-connected to the vertex at a maximum distance of d_{max} steps. The Xth row thus has a nonzero value in the *X*th column because vertex *X* has zero distance to itself) as well as nonzero values in all the other columns that represent vertexes from the (restricted) domain of vertex *X*. A value in the matrix represents the importance of the vertex represented by the column for the description of the vertex represented by the row. The authors propose to compute the values as 1/pow(2,d), where d is the path distance between the two vertexes represented by the row and column. Figure 20, borrowed from (Mladenic and Grobelnik 2005), illustrates this graph representation technique on a simple example.



	0	1	2	3	4	5	6	7	8	9	10	11
0	1	0.5			0.25	0.25	0.5	0.25	0.5	0.25		0.25
1	0.5	1					0.25	0.5	0.25	0.5	0.5	
2			1	0.25	0.5				0.25			
3			0.25	1	0.5				0.25			
4	0.25		0.5	0.5	1	0.25			0.5			0.25
5	0.25				0.25	1			0.5			0.25
6	0.5	0.25					1		0.25			
7	0.25	0.5						1		0.25		
8	0.5	0.25	0.25	0.25	0.5	0.5	0.25		1			0.5
9	0.25	0.5						0.25		1	0.25	
10		0.25								0.5	1	
11	0.25				0.25	0.25			0.5			1

Figure 20: Illustration of the transformation process. The rows of the matrix represent the instances (vertexes) and the columns represent the neighborhood with weights relative to the distance from the vertex in the corresponding row. Here we have set the maximal distance to dmax = 2. Notice that the diagonal elements all have weight 1 (showing that each vertex is in its own neighborhood). The dashed lines point out the neighboring vertexes and the corresponding weights for the vertex labeled as 2. This vertex has four non-zero elements in its sparse vector representation: 1, 0.25, 0.5, and 0.25. These elements correspond to the four neighboring vertexes (labeled in the graph as 2, 3, 4, and 8).

The graph in Figure 20 is a very simple one and does not include cases in which two vertexes are connected with more than one path. This also means that the graph does not contain any cycles. Furthermore, all the edges have the same weight. Of course, this may not be always the case. Therefore, in order to correctly represent general graphs with a matrix, we need to decide how to take multiple paths and cycles into account and how to include edge weights into the matrix.

In order to resolve the issue of multiple paths and cycles we simply measure the distance according to the shortest path (also called the geodesic distance between two vertexes).

Consequently, if there is a cycle along the path between two vertexes, we always use the shortest way around the cycle between the two vertexes.

Last but not least, edge weights need to be included into the matrix. One way is to use the weights only for setting the d_{max} threshold used in the representation of (Mladenic and Grobelnik 2005). This results in removing all the edges from the matrix that have weights below this threshold. This way the information on the weights gets incorporated into the matrix.

Another option is to use the ScentTrails algorithm (Olston et al 2003). The idea is to metaphorically "waft" scent of a specific vertex in the direction of its out links (links with higher weights conduct more scent than links with lower weights). The scent is then iteratively spread throughout the graph. After that one can observe how much of the scent reached each of the other vertexes. The amount of scent that reached a target vertex denotes the importance of the target vertex for the description of the source vertex. To compute the spreading of the scent, we first create an adjacency matrix A. The matrix element t_{ij} denotes the weight of the edge from vertex i to vertex j, if such edge exists, or is set to zero if edge does not exist. Next, the rows of matrix A are normalized so that they sum to 1. Scent conduit matrix $\mathbf{S}^{(t)}$ is then computed in every iteration of the algorithm. The scent conduit matrix element $s_{ij}^{(t)}$ denotes the amount of scent that reached vertex i from vertex j within t iterations. The scent conduit matrix computation is done as described in equations (6):

$$S_0 = I,$$
$$S_t = I + (1-\beta)*zdiag(A_t * S_{t-1}), \quad (6)$$

where zdiag(X) sets the diagonal entries of the matrix X to zero (to prevent the scent being directed back to its origin), and β is the parameter signaling the fraction of the scent intensity that is lost during the propagation through each link. Formal derivation and further explanation can be found in (Grčar et al 2007).

In the work described in this thesis we represented all the graphs with N edges with a matrix using a ScentTrails algorithm running for N steps and by using the parameter value $\beta = 0.5$.

4.2.4 Graph Visualization

For visualizing graphs many types of algorithms exist, for a survey see (Pajntar 2006). Most common for simple undirected graphs are algorithms that use an analogy from physics. These algorithms tend to focus on visualizing graphs by carefully placing the vertexes on a 2D surface to minimize the predefined energy function dependant on the positioning of the vertexes and the edges between them. On the other hand our task is to identify the most important connected components in a graph, which is similar to the goal of finding the most important topics in the textual records. This is why we will try to employ the same visualizing techniques as for analyzing text also for the analysis of graphs. In the previous section we described how to successfully represent graph vertexes with feature vectors. Now we explain how to reuse the described Clustering, SVD and MDS algorithms as they were described in the Text Visualization Section on graph matrixes obtained with techniques presented in the previous section.

Graph SVD and vertex MDS

As the graph is represented with feature vectors we can treat each vertex in a similar way as a textual document, which is also represented with a feature vector. Graph visualization is then obtained by applying Latent Semantic Indexing in a combination with Multi Dimensional Scaling to acquire a 2D representation of a graph. The effect is a two dimensional embedding of the graph vertexes. Every vertex is visualized close to the vertexes which share the similar feature vector pattern, indicating that the vertexes share a similar connectivity in the original graph.

When applying the DocumentAtlasBG algorithm to such graph representation we acquire a 2D embedding of not only the vertexes of the graph but also the main patterns in vertexes description which are represented as artificial vertexes. Every artificial vertex is described by the most important components of the singular value vector describing the dimensionality reduction step. Every artificial vector therefore represents the most important patterns of the connected vertexes. Therefore the acquired 2D visualization shows the vertexes of the graph with the most important communities in the graph. The interactivity of the diagram allows to view the communities in which every vertex participates and to view the vertexes which are part of any identified community. Figure 21 shows an example of using the DocumentAtlasBG algorithm on a graph with such feature vector representation.

Consortium Analysis with DocumentAtlasBG algorithm

The DocumentAtlasBG algorithm for analyzing collaboration network by SVD and MDS methods also corresponds to solving the problem of identifying consortium of the actors as defined in the Analysis Goals section of this chapter. Our task is to extract the consortium of the selected actor described by records of the actor's collaborators. That is, we seek to identify the most important communities of collaboration of the selected researcher or research organization in the selected context of other researchers or research organizations.

We achieve this goal by reemploying the 2D visualization technique DocumentAtlasBG already used for analysis of research competence. The DocumentAtlasBG algorithm enables efficient consortium of actor identification by identifying the patterns in collaboration network and the association between the patterns and the vertexes in the graph. The visualization technique tries to first identify the main patterns in the collaboration graph and then calculate the strength of every actor in these collaboration patterns. At last the patterns and the entities are transformed into a two dimensional space by preserving their similarities as much as possible. The algorithm is very similar to the one used for Analysis of Competence and was described in detail in the previous paragraphs. The developed method thus allows efficient collaboration community identification.

Currently only one value of the parameter of the DocumentAtlasBG algorithm is used when running this algorithm in the scope of the IST World portal, due to required simplicity of use of this analysis technique. This parameter sets the number of *Semantic Space* dimensions at 20 in the SVD computation. This parameter can be smaller, if the number of vertexes in the analyzed graph is less then 20.

The Consortia Diagram, e.g. the one in Figure 21, shows researchers on the background of most related consortia based on the past collaboration network. The consortia are identified as

background clouds. Researchers are visible as small crosses. The position of researchers is associated with their most related communities and other researchers selected into the context of analysis. Moving the mouse highlights the specific related communities in the range of the dynamic blue circle, as indicated with the communities' most representative researchers. For example, in Figure 21, the characteristic computed community of Luc de Raedt in the context of the people involved with machine learning consists of the following people: Luc De Raedt, Heikki Manilla, Fronçois Fages, etc.

The collaboration information of the analyzed actors, which is the input to the graph visualization algorithm, can usually be extracted from various types of records describing the work. In case of the academic actors these records are usually again the publications published in academic journals or scientific conference proceedings. Every published scientific paper includes information on the authors of the publication and their affiliated organizations. Every record on an academic research project includes the information on the collaborating participants. The co-authorship and partnership information in the projects and publications can be used to extract input collaboration graph of the studied academic actors. This approach does however not work very well with the industrial actors who usually do not publish the information on collaboration with their partners or clients. In the context of the knowledge transfer it is more important to capture the existence of their social network in terms of cooperation with the academic partners. We can therefore reuse the existing records of academic activity to capture at least the academic part of the social network of the industrial actors.

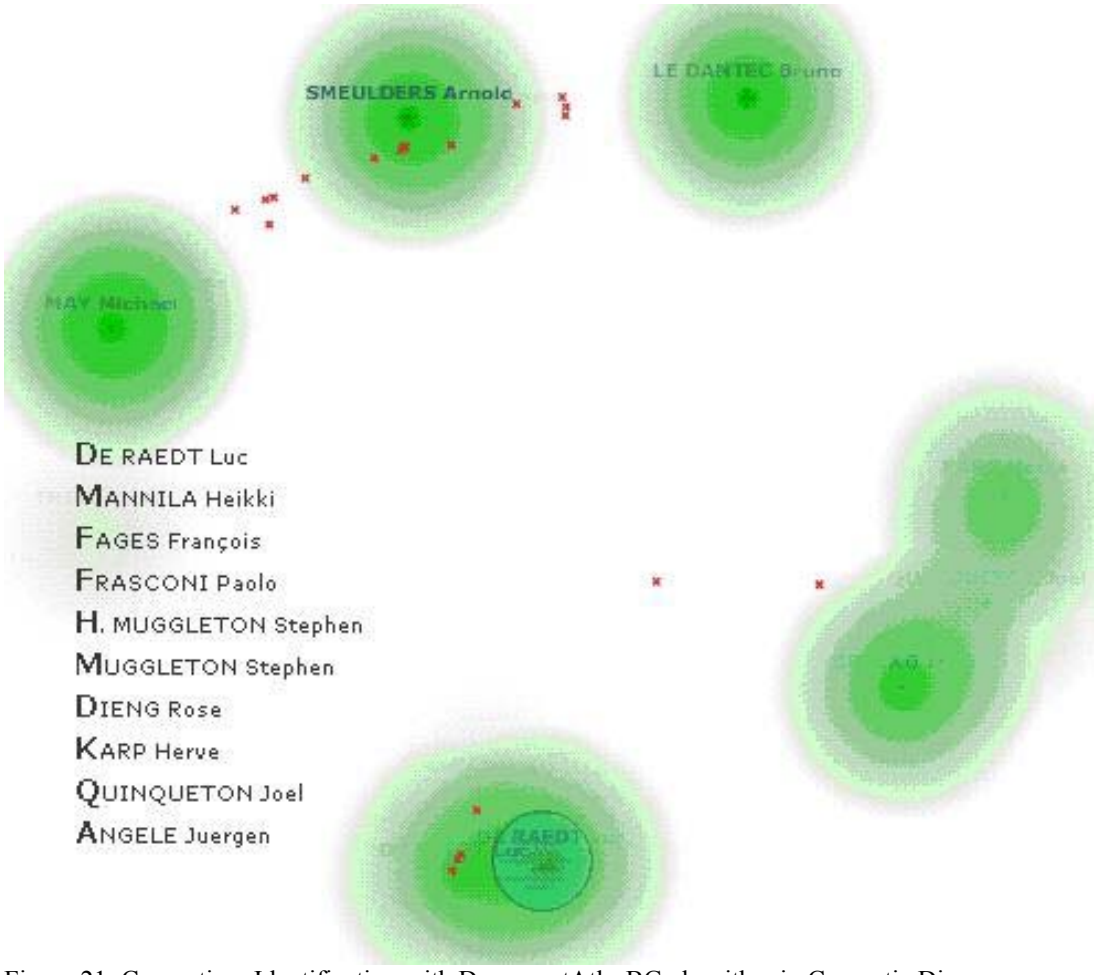


Figure 21: Consortium Identification with DocumentAtlasBG algorithm in Consortia Diagram.

Graph partitioning

As the graph is represented with feature vectors we can treat each vertex as a textual document. If we apply classical k-means clustering (Kanungo et al 2002) already described in the previous sections on such a feature vector set, we in effect partition the underlying graph into K partitions. The partition is based on the description of vertexes in the form of the feature vectors. This results in high connectivity features between vertexes inside a partition but not between the partitions. The partitioning can be done repeatedly to acquire hierarchical partitioning of the graph. When the vertexes include temporal information it is easy to visualize the partitioning using the already mentioned method for visualization of clusters with assigned temporal information (Caruana et al 2005).

Graph partitioning for Consortia Development Analysis

Our task is to identify the consortia importance development. That is we seek to identify the development with time of the importance of the consortia distinguishing between the records of collaboration of the actors in the context of the analysis.

We can reuse a method for analysis of Competence Temporal Development to perform the interactive community trends identification. The algorithm is the same with the exception of step one which has to deal with different kind of document representation that is used instead of bag-of-words: Every document is now described by the entities collaborating over the document like authors of the paper or organizations involved in a project. Preprocessing is applied to acquire a matrix of documents described in terms of the collaborators instead of words. This operation effectively expanded the uni-partite collaboration graph consisting of nodes representing collaborative items (e.g. papers) into a bi-partite graph in which one set of nodes represents the collaborative items (e.g. papers) while the other set of nodes represents the collaborating actors (e.g. paper authors). All the rest of the algorithm's steps are the same. This way we obtained a 2D visualization of temporal development of automatically computed consortia like the one displayed in Figure 22. It features identified trends in collaborating consortia of the researchers in the field of machine learning.

The Consortia Development Diagram, e.g. the one displayed in Figure 22, shows consortia strength development through time. The temporal strength development of each consortium is identified with a colored region across the diagram. Every colored region is described by community's most important members in the lower right hand side of the diagram. The association between the community clusters and the researchers is indicated in the lower left hand side. Clicking on the community region allows a more detailed insight into community development by further splitting the region into sub community regions.

The collaboration information on the analyzed actors is usually time stamped. It can be extracted from various types of records describing the collaboration of the actors. In case of the academic actors these records are usually the publications published in academic journals or scientific conference proceedings. Every record about an academic research project includes the information duration of the project and its participants and partner organizations. The co-authorship and partnership information in the projects and publications can be used to identify the consortia and the development in importance of the communities of collaboration of the studied academic actors.

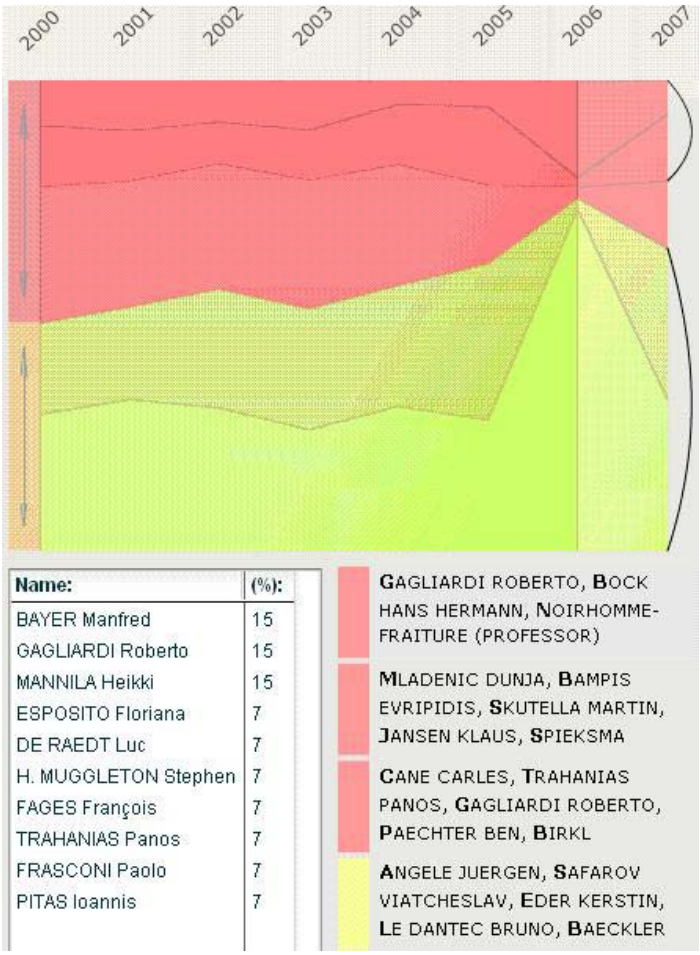


Figure 22: Temporal Consortia Visualization with graph partitioning algorithm in Consortia Development Diagram.

4.3 Analysis Experiments

In the previous section we reviewed the data mining algorithms for visualizing text and graph data. In this section we show how we experimented with these algorithms for the tasks of identifying competence, competence development, consortia and consortia development. With every analysis goal we describe the corresponding experiment methodology, used data and the acquired results.

4.3.1 Analysis of Competence

First experiment was used to intuitively evaluate the competence analysis of the faculties which are administratively part of the University of Ljubljana and intuitively compare the extracted competence words with the expected competence given the faculty name. For instance, the competence of the Faculty of Law in the context of University of Ljubljana as a collection of faculties is most probably best described by keywords which are centered on the general concept of law. Another example is the competence of the Faculty of Pharmacy. Its competence keywords should in general describe the notion of pharmacy. In this setting we expect from our automatic competence analysis method to extract those words from the textual description of the faculty activity which semantically match the meaning of the faculty name.

All together we analyzed a list consisting of 20 faculties which are administratively part of the University of Ljubljana. The activity of every faculty is described by the textual keywords associated with every faculty and the textual content associated with the projects, in which the faculties participate as recorded in the SICRIS database (Seljak 2001). The names of the analyzed faculties are: Faculty of Mechanical Engineering, Faculty for Social Work, Faculty of Electrical Engineering, Veterinary Faculty, Faculty of Sport, Faculty of Administration, Faculty of Theology, Faculty of Economics, Faculty of Pharmacy, Bio-technical Faculty, Faculty of Mathematics and Physics, Faculty of Education, Faculty of Natural Sciences and Engineering, Faculty of Social Sciences, Faculty of law, Faculty of Medicine, Faculty of Chemistry and Chemical Technology, Faculty of Civil and Geodetic Engineering, Faculty of Maritime Studies and Transport and Faculty of Computer and Information Science.

The textual data, which we analyze, stems from the SICRIS database (Seljak 2001). E.g. the activity records of the Faculty of Computer Science include textual data like "Engineering sciences and technologies", "Computer science and informatics", "Intelligent systems - software", "Systems engineering", "computer technology", "Artificial intelligence", "Education", "research" and "education". Additionally, some of the projects in which the Faculty of Computer Science collaborates are for example described with the following text: "Algorithms for control of scanning probe microscopes using quartz tuning forks", "Automatic construction of 3-D geometric models and object recognition", "Clinical paths data mining with soft computing", "Computer Vision", etc.

Figure 23 shows the visualization result of the automatic competence analysis using the DocumentAtlasBG algorithm. The diagram shows the plot of the 20 faculties of the University of Ljubljana, depicted as small crosses on the background of identified competence keywords depicted as clouds. The list of the extracted competence keywords includes Slovene and English words.

Keywords in Figure 23 describe the most important competence which can be identified in this group of 20 organizations. The position of every organization is associated with the positions of the most important competence words. This enables us to identify the main competence of the individual organizations. The diagram reveals approximately five clusters of faculties according to their position in the diagram. The clusters identified from the Competence Diagram are:

- Pharmaceutical Cluster: Faculty of Pharmacy, Veterinary Faculty, Biotechnological Faculty, Faculty of Medicine, (Faculty of Chemistry and Chemical Engineering)
- Construction Cluster: Faculty of Natural Sciences and Engineering, Faculty of Civil and Geodetic Engineering, Faculty of Mathematics and Physics, (Faculty of Computer and Information Science, Faculty of Electrical Engineering, Faculty of Mechanical Engineering)
- Sports Cluster: Faculty of Sports, Faculty of Maritime Studies and Transport
- Social Work Cluster: Faculty for Social Work, Faculty of Education
- "Law" and "Administration" Cluster: Faculty of Law, Faculty of Social Sciences, Faculty of Arts, Faculty of Administration, Faculty of Economics

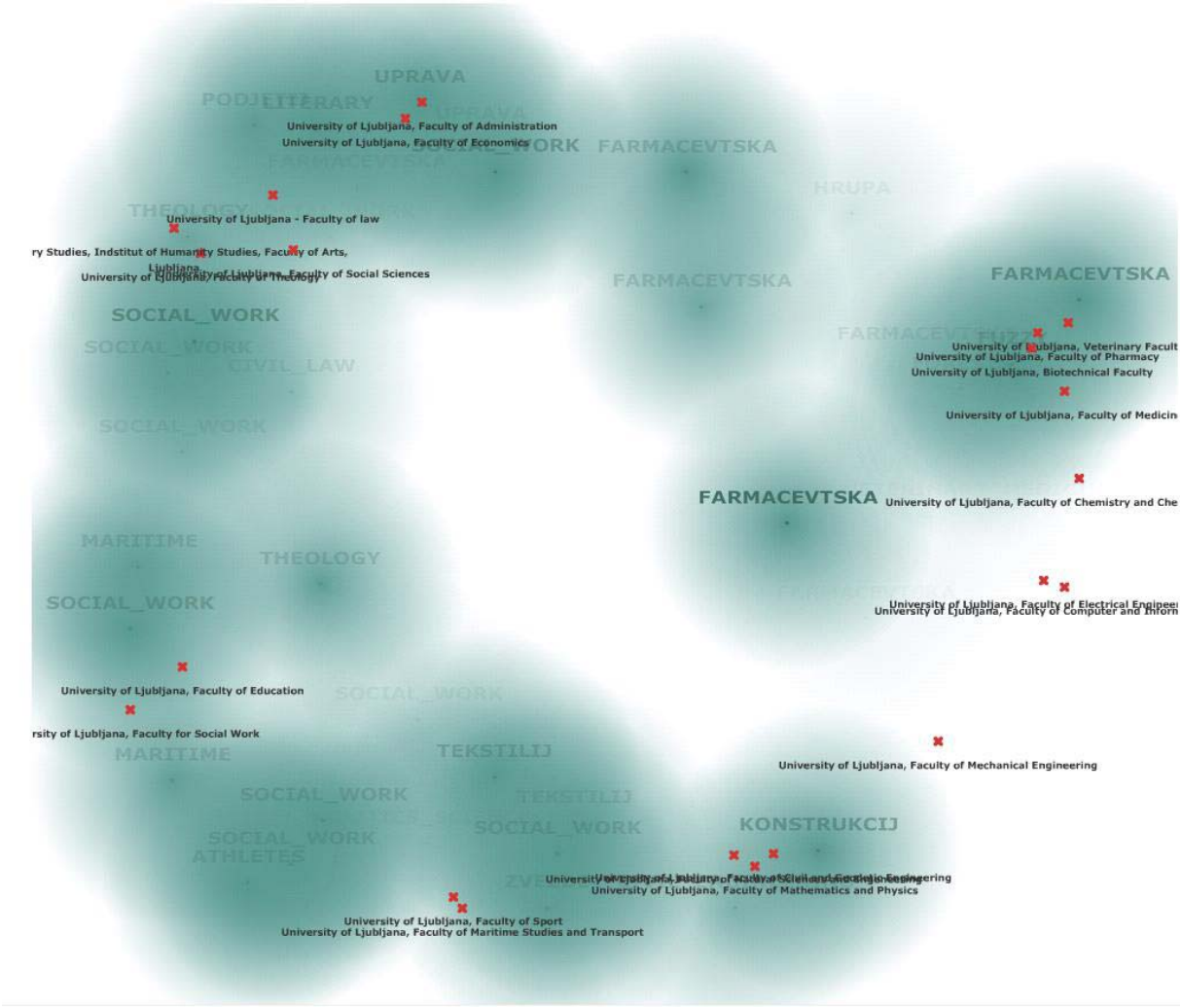


Figure 23: Competence Diagram of the members of the University of Ljubljana.

We can conclude from the intuitive evaluation of cluster members that the Competence Diagram in Figure 23 already gives a correct rough estimation of the competencies of the individual institutions.

The interactivity of the diagram, as seen in Figure 24, allows even more detailed analysis of the competencies of individual faculties by focusing the mouse pointer over the red points representing individual organizations.

Figure 24 presents the highlighted competence of the Faculty of Pharmacy. The identified competence words of this faculty according to this context mostly include just one word: FARMACEVTSKA (*PHARMACEUTICAL*) which is correct given the name of the faculty. What is more, the additional competence keywords are displayed to the user describing the competence in even greater detail. These are: FARMACEVTSKA (*PHARMACEUTICAL*), (UČ)INKOVINA (*DRUG*), *PHARMACY*, *PHARMACEUTICAL*, *DRUG*, *MATHEMATICAL_PHARMACY* and *PHARMACY_REPRESENTATIVE*.

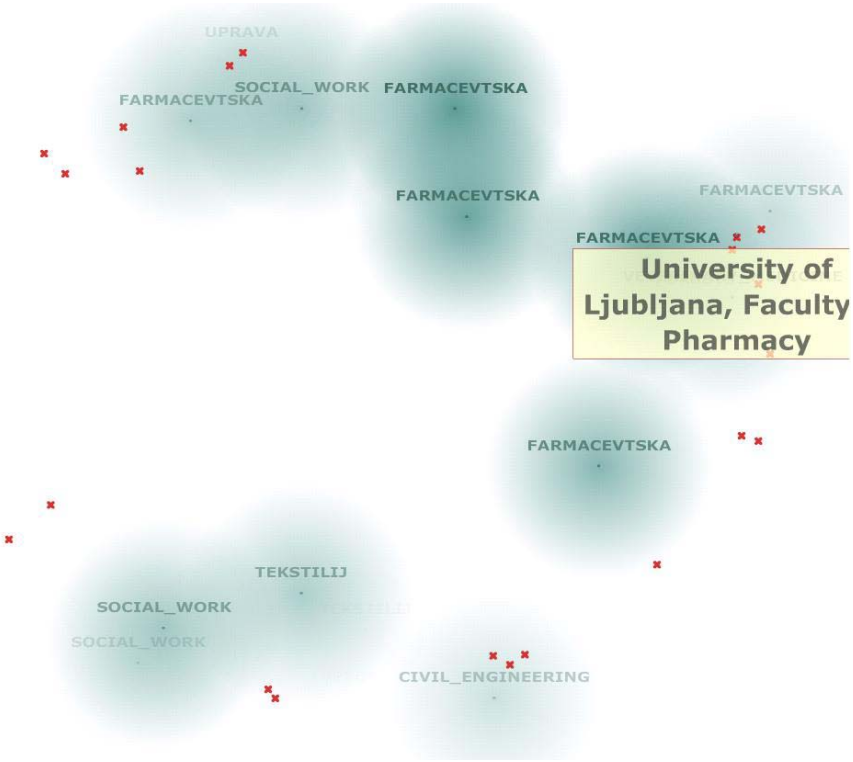


Figure 24: The Competence Diagram of the Faculty of Pharmacy in the context of 20 Faculties of the University of Ljubljana

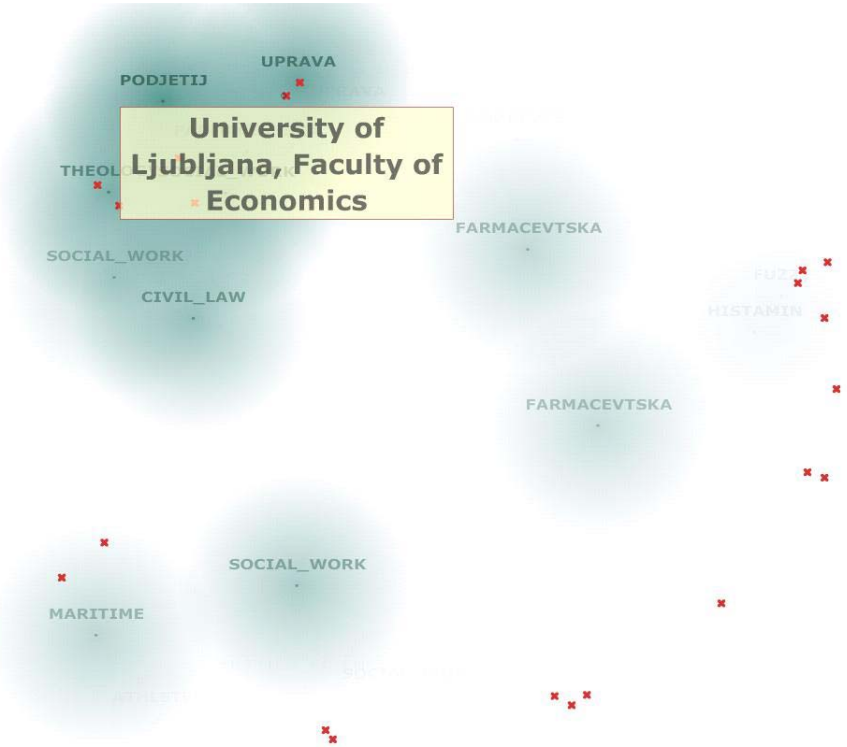


Figure 25: The Competence Diagram of the Faculty of Economics in the context of 20 Faculties of the University of Ljubljana.

Another example of competence detail estimation is the analysis of the Faculty of Economics. Figure 25 depicts the highlighted competence of this organization. The most important

competence words are: UPRAVA (*MANAGEMENT*), PODJETIJ (*COMPANIES*), *THEOLOGY* and *CIVIL LAW*. Apart from *THEOLOGY* all the keywords are consistent with our expectations.

To further investigate identified competence of the 20 faculties we show the extracted competence keywords for every one of them in Table 5. We can observe that more than half of the extracted keywords correspond accordingly to the name of the associated faculties in 14 out of 20 organizations. These faculties are marked with light gray background color in Table 5. The 6 faculties with less than half of the correctly assigned competences are marked with dark gray background color in Table 5. These are: Faculty of Social Sciences, Faculty of Education, Faculty of Mathematics and Physics, Faculty of Electrical Engineering, Faculty of Computer and Information Science and Faculty of Chemistry and Chemical Technology.

Organization Name	Extracted Keywords
Bio-technical Faculty	**FORESTS**, **SOCIAL WORK**, **NOISE**
Faculty of Administration	UPRAVA (**MANAGEMENT**), **SOCIAL WORK**
Faculty of Chemistry and Chemical Technology	**FORESTS**, FARMACEVTSKA (**PHARMACEUTICAL**), TEKSTILIJ (**TEXTILE**)
Faculty of Civil and Geodetic Engineering	**CIVIL ENGINEERING**, ZVEZDE (**STARS**), TEKSTILIJ (**TEXTILE**)
Faculty of Computer and Information Science	**FUZZY**, KONSTRUKCIJ (**CONSTRUCTIONS**), FARMACEVTSKA (**PHARMACEUTICAL**)
Faculty of Economics	UPRAVA (**MANAGEMENT**), PODJETIJ (**COMPANIES**), **THEOLOGY**, **CIVIL LAW**,
Faculty of Education	**SOCIAL WORK**, FARMACEVTSKA (**PHARMACEUTICAL**), **SOCIAL WORK**, UPRAVA (**MANAGEMENT**)
Faculty of Electrical Engineering	**FUZZY**, KONSTRUKCIJ (**CONSTRUCTIONS**), **FORESTS**
Faculty of Law	UPRAVA (**MANAGEMENT**), **CIVIL LAW**, **SOCIAL WORK**, **MARITIME**
Faculty of Maritime Studies and Transport	**MARITIME**, TEKSTILIJ (**TEXTILE**)
Faculty of Mathematics and Physics	KONSTRUKCIJ (**CONSTRUCTIONS**), ZVEZDE (**STARS**), FARMACEVTSKA (**PHARMACEUTICAL**),
Faculty of Mechanical Engineering	HRUPA (**NOISE**), KONSTRUKCIJ (**CONSTRUCTIONS**), FARMACEVTSKA (**PHARMACEUTICAL**)
Faculty of Medicine	**HISTAMINE**, **SOCIAL WORK**, FARMACEVTSKA (**PHARMACEUTICAL**)
Faculty of Natural Sciences and Engineering	KONSTRUKCIJ (**CONSTRUCTIONS**), TEKSTILIJ (**TEXTILE**), **FORESTS**
Faculty of Pharmacy	FARMACEVTSKA (**PHARMACY**)
Faculty of Social Sciences	**THEOLOGY**, FARMACEVTSKA (**PHARMACEUTICAL**), **POLITICS SCIENCE**
Faculty of Social Work	**SOCIAL WORK**, **THEOLOGY**, **MARITIME**
Faculty of Sports	**ATHLETES**, **SOCIAL WORK**,

	PHARMACEUTICAL
Faculty of Theology	**THEOLOGY**, **LITERARY**, **SOCIAL WORK**, **MARITIME**
Veterinary Faculty	VETERINARSKA MEDICINA (**VETERINARY MEDICINE**), FARMACEVTSKA (**PHARMACEUTICAL**)

Table 5: The extracted competence keywords for every of the 20 faculties in the scope of the analysis. Light gray background color depicts correctly extracted keywords. Dark gray indicates incorrectly extracted keywords.

We can conclude that the DocumentAtlasBG algorithm as we described it in the previous sections works well for this particular example of analyzing the competence of the Faculties of the University of Ljubljana. Most likely the inappropriate parameter setting of the used analysis methods prevent it to do even better.

4.3.2 Analysis of Competence Development

In the second experiment we evaluated the algorithm for identification of temporal trends in competence development. The main idea of the experiment was to extract the competence development of the Acad. Prof. Dr. Ivan Bratko for the last 30 years of his academic career and evaluate the findings by reviewing them with the professor himself. The review process is a simple one and consists of collecting intuitive remarks of the Prof. Dr. Ivan Bratko on the correctness of the produced interactive visualization.

The experiment was conducted on 329 publications, which were extracted from the Google Scholar Database (Jacsó 2005) as described in the IST World Portal chapter. Every publication record consists of data on the publication title, publication authors and the publication date. For instance, one of the publications of the Prof. Dr. Ivan Bratko was published in 1986, titled "Detection of positional patterns in chess". The authors are I. Bratko, P. Tancig and I. Tancig. Another publication was published in 1994; the title is "Reconstructing Human Skill with Machine Learning"; its authors are I. Bratko, and T. Urbančič. The date labels of the publications vary from 1978 to 2006.

Figure 26 depicts the visualization of the results of the competence development analysis using the temporal clustering algorithm described in the previous section. The results show the competencies of the respected professor and how they developed through the 30 years of his scientific work. The hierarchical clustering algorithm split all the publications into two groups. The first group denotes the competence of the Prof. Dr. Ivan Bratko described by keywords like QUALITATIVE MODELS, DYNAMIC SYSTEMS, CHESS, INTERACTION, RECONSTRUCTING, AI, PREDICTION, etc. The other competence is described by keywords like PROGRAMMING, ARTIFICIAL, INTELLIGENCE, DISCOVERING, DIAGNOSTIC, NOISY DOMAIN, ANALYSIS, CONCEPTS, PROLOG, etc. The trend of importance of each of the two competencies shows that the QUALITATIVE MODELS competence was prevailing when the professor was starting his research around 1980. Later, around 1985 the ARTIFICIAL INTELLIGENCE PROGRAMMING competence grew stronger and reached its peak in 1987. In the following years the two major competences have become balanced.

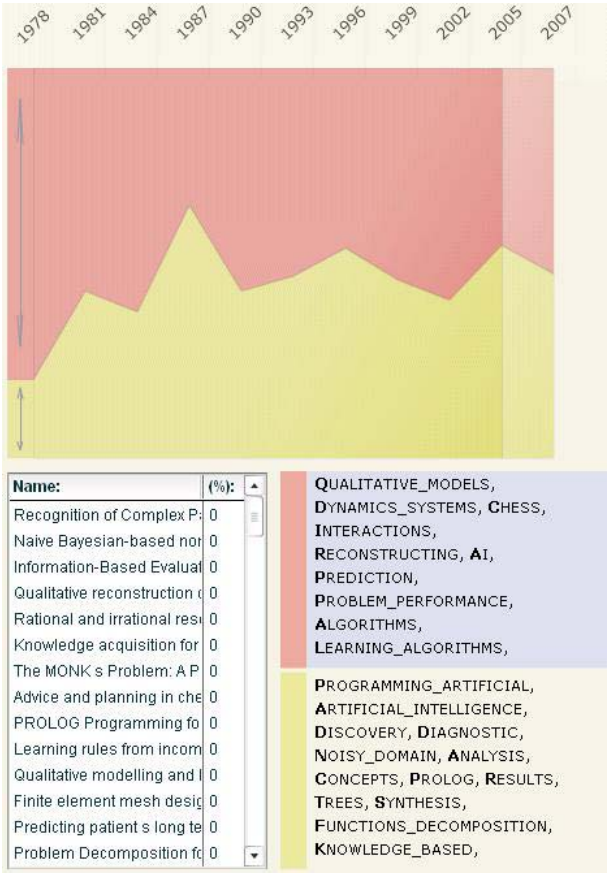


Figure 26: Competence Development Diagram of the 329 publications of the Acad. Prof. Dr. Ivan Bratko shows the development of the author's competence in the absolute context.

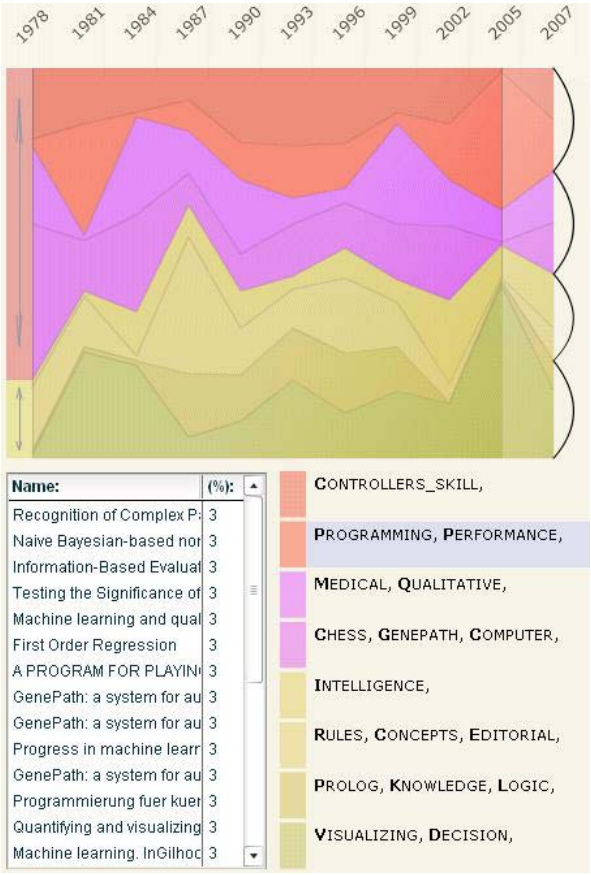


Figure 27: Detailed Competence Development Diagram of the 329 publications of the Acad. Prof. Dr. Ivan Bratko enables to explore development of sub competences inside every competence.

We can use the hierarchical structure of the identified competence patterns in the Competence Development Diagram to review the competence evolution in more detail. The interactive functionalities of the Competence Development Visualization enable us to explore the two identified competences in more detail as this is depicted in Figure 27. We can identify 8 distinct competences and their relative importance trends during the years. The CONTROLLERS_SKILL competence is more or less stable through the 30 years of professor's research. The PROGRAMMING and PERFORMANCE competence is not that stable with big jumps in importance at the beginning of the professor's career and his current work. MEDICAL QUALITATIVE competence is also more or less stable through the years with mirroring behavior to the just described PROGRAMMING PERFORMANCE competence. The CHESS, GENEPATH, COMPUTER competence is however slowly diminishing starting from early years, when it was one of the most important. The INTELLIGENCE competence is slowly gaining in importance but is still relatively insignificant. The competence described with RULES, CONCEPTS, EDITORIAL has been quite strong through out and has only recently ceased. The PROLOG, KNOWLEDGE, LOGIC competence on the other hand was significant only in the years between 1984 and 2000. At last the VISUALISATION, DECISION competence seems to be one of the most important competences. Its importance is quite stable through out the years and exhibits strong peaks before 1984 and after 2002.

Prof. Bratko's comments on the results depicted in Figure 26 were in general positive. He first confirmed that the extracted keywords correctly identify his major two topics of work. Out of 23 extracted keywords, describing the two main competences, only 6 were labeled as questionable or inappropriate. In the first competence cluster the inappropriate keywords were: INTERACTIONS and RECONSTRUCTING. In the second cluster the inappropriate keywords were: ANALYSIS, RESULTS and SYTHESIS. The questionable keyword in the first competence cluster was the PROBLEM_PERFORMANCE. In the second cluster the keywords PROGRAMMING_ARTIFICIAL was deemed as questionable extracted keyword when considered independent from the rest of the extracted keywords. All the rest of the 23 extracted keywords were recognized as correct. What is more Prof. Dr. Bratko also recognized the depicted trends of development of the two clusters in Figure 26 as correct and in correspondence with his impression of the development of his topic of work. He noted that the big jump in the importance of the second competence group around the year 1987 corresponds very well with the start of his PROLOG related research.

We can conclude that the clustering algorithm which produced the Competence Development Diagram in Figure 26 performed rather well for the task of extracting competence development trends of Prof. Dr. Bratko from his publication records.

4.3.3 Analysis of Consortia

The main idea of the third experiment was to automatically identify the main collaboration consortia of every faculty of the University of Ljubljana using the ScentTrails and DocumentAtlasBG algorithms and then intuitively compare the automatically extracted patterns with the manually extracted patterns from the visualization of the complicated collaboration network between the faculties. We expect that the automatically identified patterns would signal the main ways of inter faculty collaboration. For instance we expect that if two faculties collaborate more intensively within each other than with other faculties in the scope of the analysis, then this should be reflected as a distinct consortium of the two organizations.

The data that specifies the context of the analysis is the same as with the analysis of competence. This time we focus on the collaboration graph between the faculties. The graph consists of nodes which represent the faculties. All together the nodes consist of 20 faculties which are administratively part of the University of Ljubljana. An edge in the graph represents collaboration in a common project, as listed in SICIRS database (Seljak 2001), between the two vertexes - faculties. Every edge is weighted by the number of common projects between the two faculties. For instance, the edge between the Faculty of Medicine and the Faculty of Pharmacy has a weight value 5, because there exist 5 common projects between the faculties. Figures 28 and 29 display the collaboration network between the faculties in the analysis scope.

The manual analysis of the network in Figure 28 reveals that there are at least two very strong distinct consortia, which are formed around the edges with the biggest weights. Figure 29 shows the part of network from Figure 28 with the names of the organizations with the strongest collaboration links. The first collaboration cluster is formed around the Faculty of Medicine and the Veterinary Faculty (10 common projects). The second cluster is formed around the faculties of Natural sciences and Engineering and the Faculty of Mechanical Engineering (9 common projects).

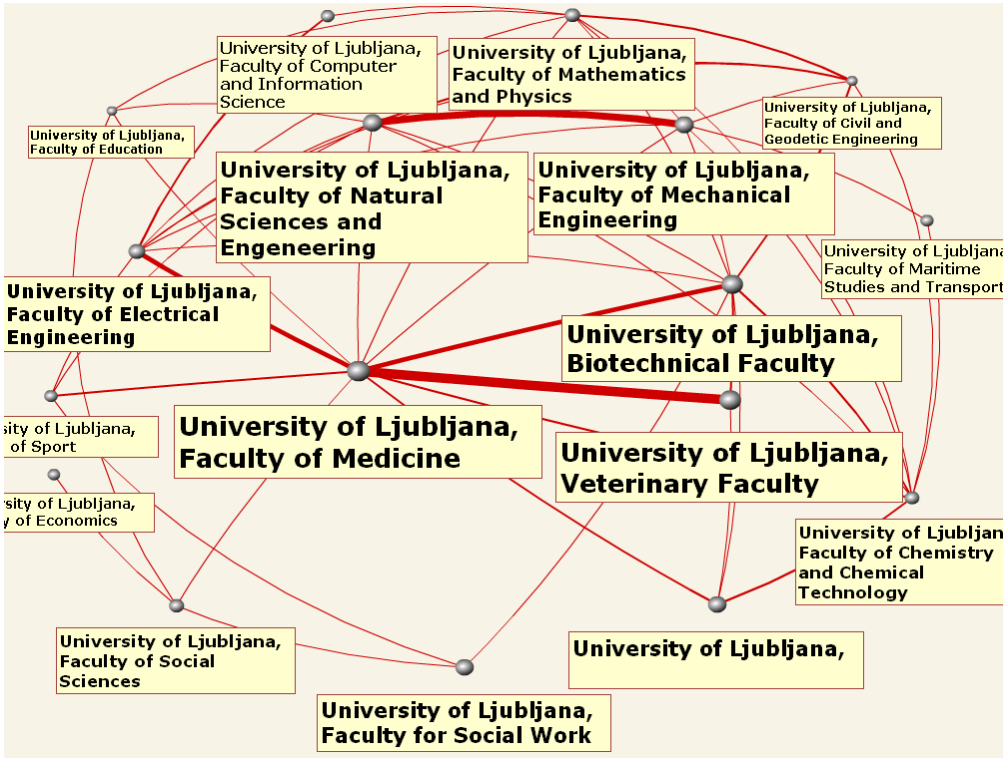


Figure 28: Collaboration Network between the 20 faculties of the University of Ljubljana

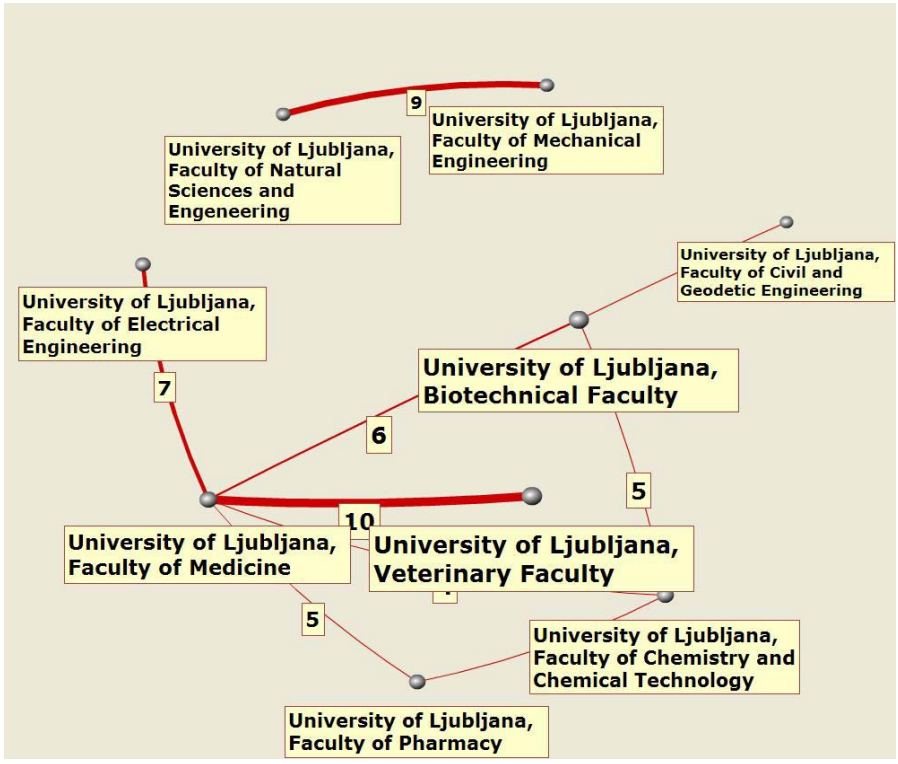


Figure 29: Collaboration Network of the faculties with the most intensive inter collaboration.

When running the DocumentAtlasBG algorithm described in the previous section on the graph in Figure 28 we get the visualization depicted in Figure 30. The diagram presents the faculties as small crosses and the identified consortia as background clouds. The transparency level of the consortium reveals its relative importance. The more the consortium is rendered transparent the more insignificant is the identified consortia. The diagram reveals that the

most important consortia identified within the collaboration network of 20 organizations is the one centered on the Faculty of Medicine, which is placed at the top right hand side of the diagram. The rest of the identified consortia include the one around the Faculty of Mechanical Engineering and the Faculty of the Economics, which are rendered as consortia clouds in the middle of the diagram. The rest of the consortia in the diagram are centered on the Veterinary Faculty and the Faculty of Arts.

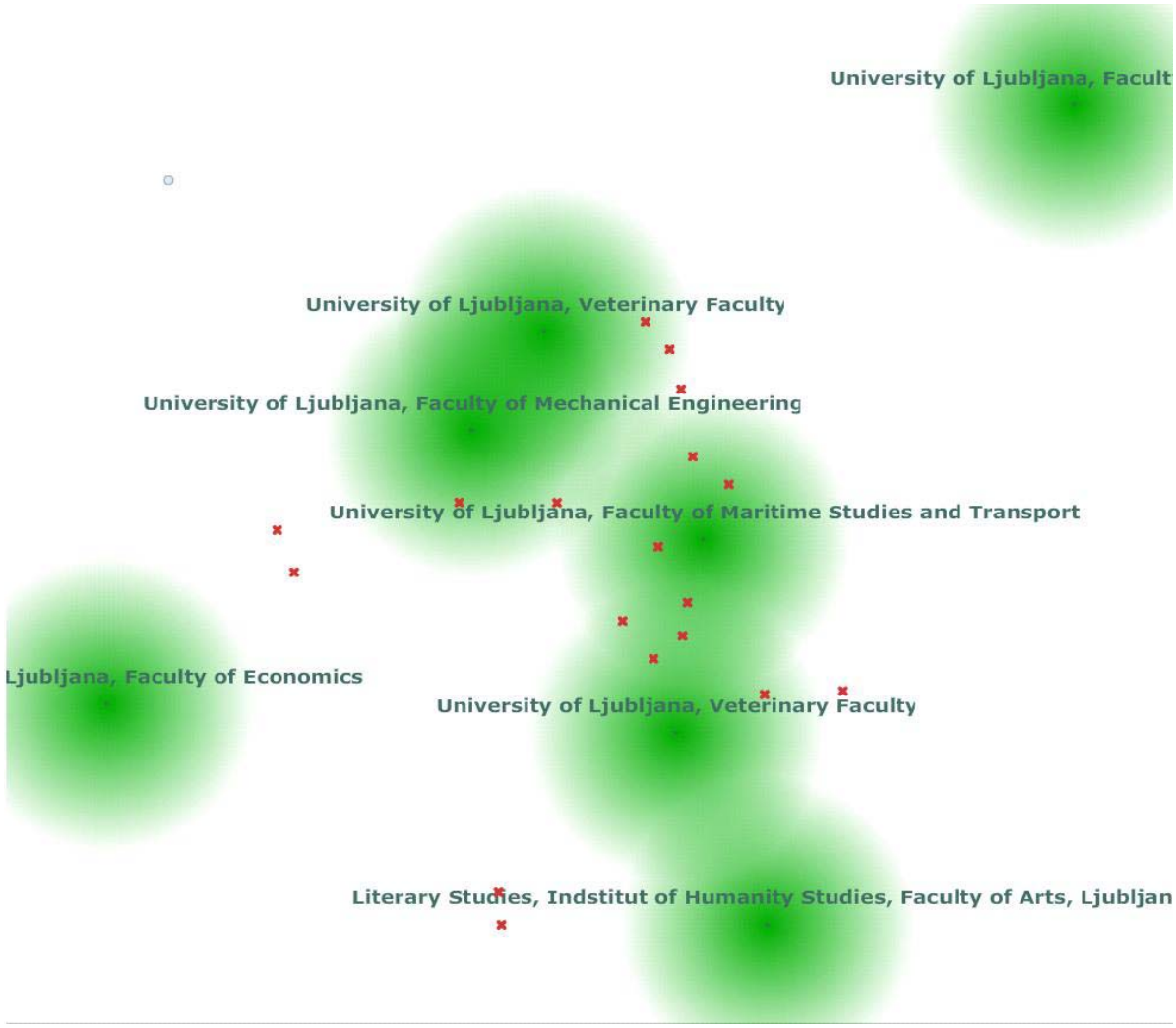


Figure 30: Consortia Diagram of the collaboration network between the faculties of the University of Ljubljana.

The consortia are described by a single organization name. We interact with the diagram to get a more information on which organizations constitute the consortia in the focus. Figure 31 displays detailed information about the top right hand side consortium around the Faculty of Medicine. The user is able to observe that the Bio-technical Faculty and the Faculty of Chemistry and Chemical Technology are important parts of the consortia centered on the Faculty of Medicine. Figure 32 presents the details of the consortia centered on the Faculty of Mechanical Engineering. We observe that the Faculty of Natural Sciences and Engineering and the Faculty of Economics are important parts of this consortia.

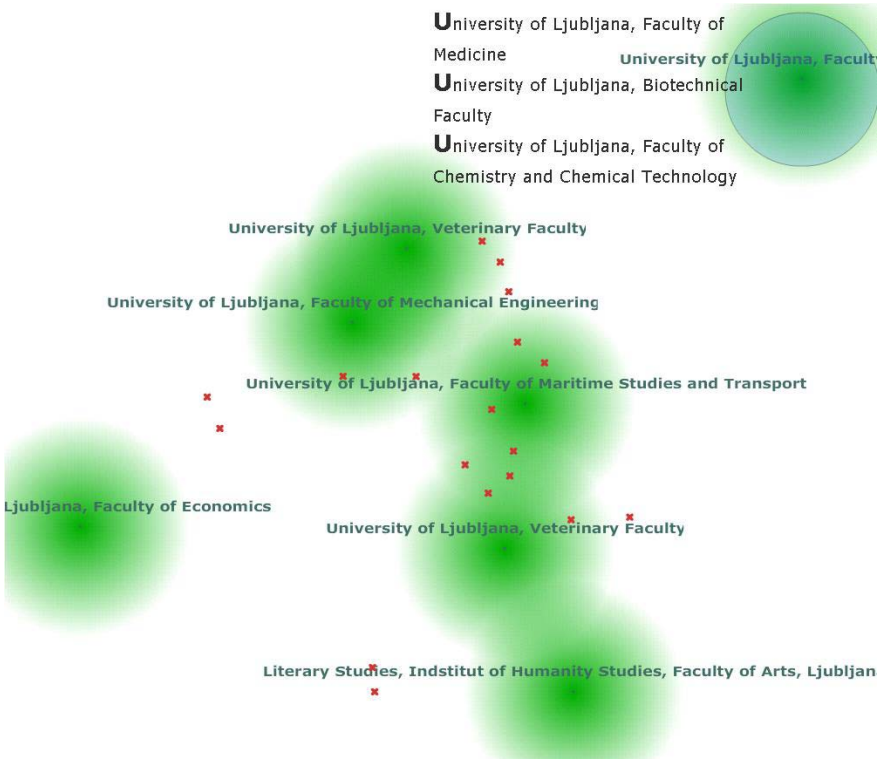


Figure 31: Consortia diagram reveals the strongest consortia centered on the Faculty of Medicine in the upper right hand corner of the diagram.

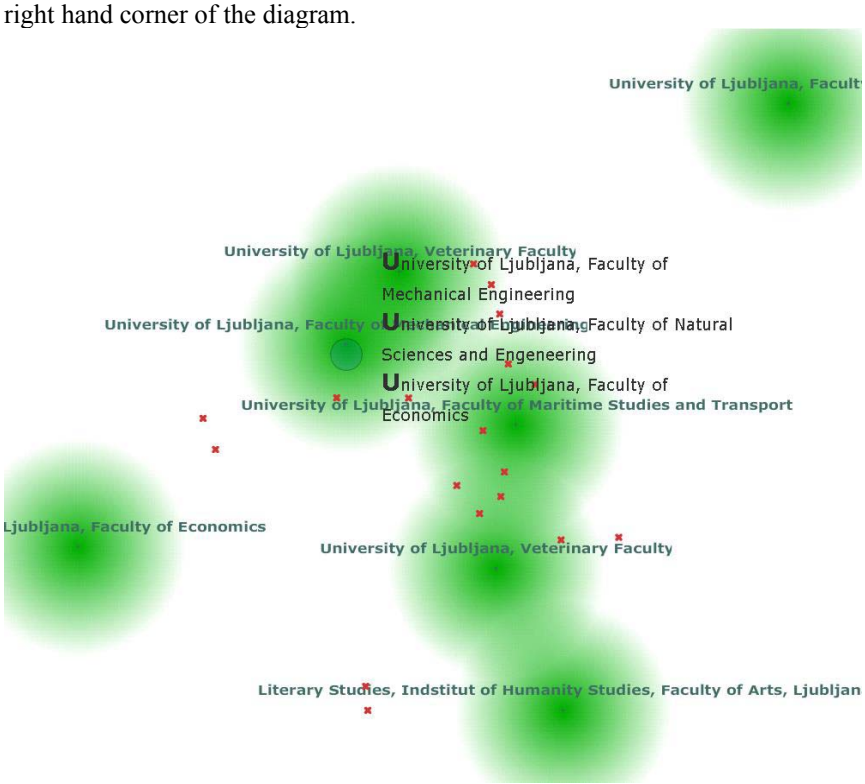


Figure 32: Consortia diagram reveals the consortia centered on the Faculty Mechanical Engineering.

The general findings of the analysis as these can be observed in Figures 30, 31 and 32 correspond reasonably well to our manual pattern identification. The automatically identified consortium centered on the Faculty of Medicine corresponds to our observation of the centrality of the Faculty of Medicine in the network in Figure 29. The automatically identified consortium around the faculty of Mechanical Engineering also corresponds very well to the

second strongest component in Figure 29. The identified consortium centered around the Veterinary Faculty corresponds to the fact that the Veterinary Faculty only collaborates with very few other faculties except the Faculty of Medicine, with which the collaboration is very strong. More detailed investigation of the diagram in Figure 28 explains the consortia of the Faculty of Economics and the Faculty of Arts found in Figure 30. Actually all the rest of the nodes of the graph are connected to at least two other nodes except the nodes of the Faculty of Economics and the Faculty of Arts. These consortia thus describe the collaboration which is not captured by the more general consortia like the one centered on the Faculty of Medicine or the Faculty Mechanical Engineering.

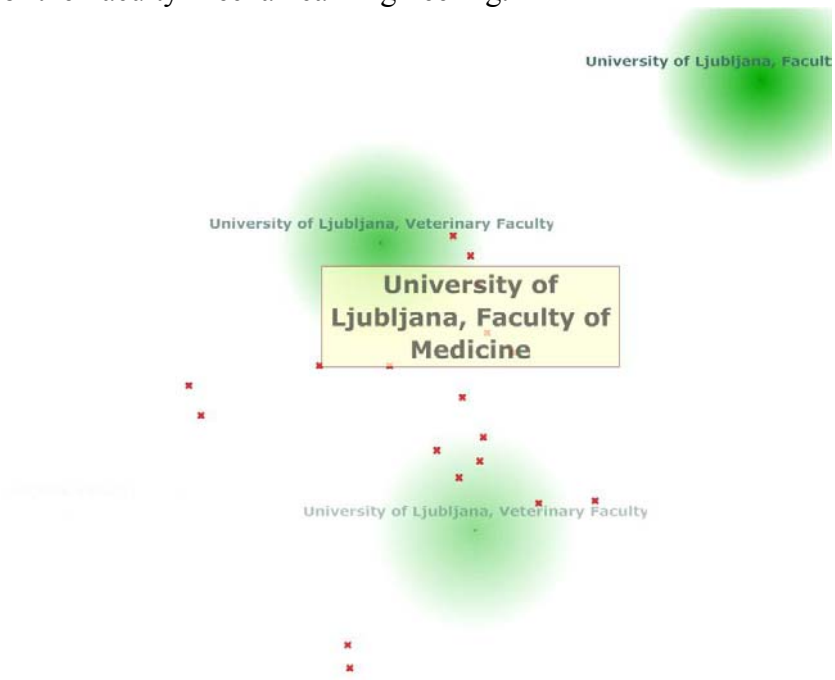


Figure 33: Consortia Diagram shows the most important consortia of the Faculty of Medicine.

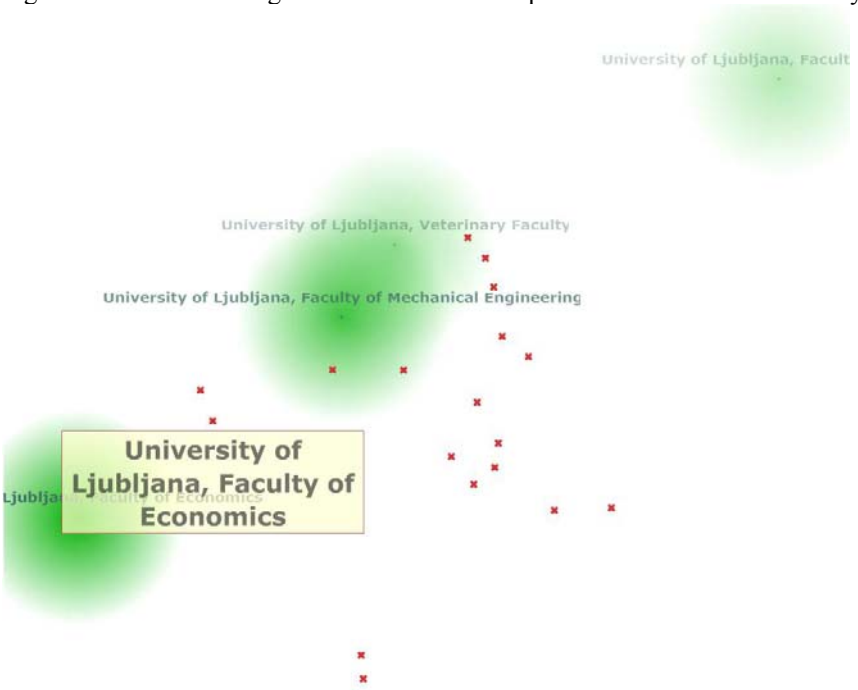


Figure 34: Consortia diagram shows the most important consortia of the Faculty of Economics.

The interactivness of the Consortia Diagram allows also identifying the collaboration patterns of every individual organization present in the analysis context by revealing the association between the organizations and the identified consortia. The diagram in Figure 33 reveals the most important consortia of the Faculty of Medicine. According to the diagram the Faculty of Medicine is most active in the consortia named after the same faculty and the consortia around the Veterinary Faculty. Faculty of Economics on the other hand is actively collaborating in very different consortia. It is most active in the consortia named after Faculty of Economics and the consortia around the Faculty of Mechanical Engineering as this can be observed from the diagram in Figure 34.

We have collected the most associated consortia for every one of the 20 organizations presented in Figure 14 and assembled them in Table 6. Every consortium is described by means of its most important organization.

We can see that most of the extracted leading organizations correspond to the names of the most important faculties in the collaboration network. The extracted consortia of the 12 faculties with correctly extracted consortia are marked with light gray in Table 6. The 7 faculties with imprecise consortia are marked with dark gray in Table 6.

Organization Name	Associated Consortia Names
Bio-technical Faculty	Faculty of **Medicine, Veterinary** Faculty
Faculty of Administration	**(no collaboration)**
Faculty of Economics	Faculty of **Economics**, Faculty of **Mechanical Engineering**
Faculty of Civil and Geodetic Engineering	Faculty of **Medicine,** Faculty of **Arts**
Faculty of Computer and Information Science	Faculty of **Medicine**
Faculty of Chemistry and Chemical Technology	Faculty of **Medicine**
Faculty of Education	Faculty of **Economics**, **Veterinary** Faculty, Faculty of **Arts**
Faculty of Electrical Engineering	Faculty of **Medicine**
Faculty of Law	**(no collaboration)**
Faculty of Maritime Studies and Transport	Faculty of **Mechanical Engineering**, Faculty of **Medicine,** Faculty of **Maritime Study**
Faculty of Mathematics and Physics	Faculty of **Medicine,** Faculty of **Arts**
Faculty of Mechanical Engineering	Faculty of **Mechanical Engineering**, Faculty of **Medicine,** Faculty of **Arts**
Faculty of Medicine	Faculty of **Medicine, Veterinary** Faculty
Faculty of Natural Sciences and Engineering	Faculty of **Mechanical Engineering**, Faculty of **Medicine,** Faculty of **Arts**
Faculty of Pharmacy	**Veterinary** Faculty, Faculty of **Medicine**
Faculty of Social Sciences	Faculty of **Economics**, Faculty of **Mechanical Engineering**
Faculty of Social Work	Faculty of **Economics**, Faculty of **Mechanical Engineering**, Faculty of **Medicine**,
Faculty of Sports	Faculty of **Medicine, Veterinary** Faculty

Faculty of Theology	**(no collaboration)**
Veterinary Faculty	**Veterinary** Faculty, Faculty of **Medicine,**

Table 6: The identified consortia leader names for every of the 20 faculties in the scope of the analysis. Light gray background color indicates correctly extracted consortia information. Dark gray color indicates in majority incorrectly extracted consortia information.

We can conclude that the ScentTrails algorithm coupled with DocumentAtlas algorithm as we described it in the previous sections works reasonably well for this particular example of analyzing the collaboration of the Faculties of the University of Ljubljana. It still however produces some unexpected results and must there fore be further improved.

4.3.4 Analysis of Consortia Development

In the fourth and final experiment we evaluated the algorithm for identification of temporal trends in consortium importance development. The main idea of the experiment was to extract the consortium development by analyzing the collaboration network of people around Acad. Prof. Dr. Ivan Bratko. Prof. Dr. Ivan Bratko is one of the long time centers of collaboration in the artificial intelligence field of research in Slovenia. He is therefore one of the most credible people for interpreting the ten years of development of collaboration between the Slovenian researchers in this field. The review process is a simple one and mostly consists of collecting intuitive remarks of the Prof. Dr. Ivan Bratko on the correctness of the produced interactive visualization showing identified consortia and their development in importance through time.

The experiment was conducted on collaboration links of the 48 people, which are listed in the SICRIS database (Seljak 2001) as collaborators of Prof. Dr. Ivan Bratko. Every person record consists of data on all the projects this person was involved with. Each project record of a person includes the project start and end date information. All together the subject of our analysis was the 48 persons connected with a network of 244 projects. For instance, one of the projects of the Acad. Prof. Dr. Ivan Bratko started in 1999, ended in 2003, was titled "Artificial Intelligence" and had the following collaborators: Bratko Ivan, Bevk Matjaž, Demšar Janez, Juvan Peter, Kononenko Igor, Kukar Matjaž, Robnik Šikonja Marko, Sadikov Aleksander, Šuc Dorian, Vladušič Daniel and Zupan Blaž. Another of his projects started in 1999 and ended in 2003; its title is "Intelligent Systems"; and the collaborator of this project are: Bežek Andraž, Bohanec Marko, Bojadžiev Damjan, Brank Janez, Bratko Ivan, Demšar Damjan, Drobnič Matija, Filipič Bogdan, Gams Matjaž, Križman Viljem, Luštrek Mitja, Mladenić Dunja, Pivk Aleksander, Rajkovič Vladislav, Šef Tomaž, Špegel Marjan, Urbančič Tanja, Volovšek Miha, Zupan Blaž and Žnidaršič Martin. The date labels of the analyzed project vary from 1995 to 2006.

Figure 35 depicts the visualization of the results of the consortia importance development analysis using the temporal clustering algorithm described in the previous section. The results show the consortia identified in the network of collaboration between the colleagues of Prof. Dr. Bratko. The plot also shows how the importance of individual consortia developed over the 11 presented years. The hierarchical clustering algorithm split all the projects into two groups. The first group denotes the consortia of people centered around: Lavrač Nada, Erjavec Tomaž, Džeroski Saso etc. The other identified consortia can be described as centered around: Bogdan Filipic, Drobnič Matija, Pivk Aleksander, Arnež Zoran Marij, etc. The trend of importance of each of these two consortia shows that the Lavrač Nada consortium was prevailing around 1995. Later, starting at around 1997 the consortia around Bogdan Filipič and Drobnič Matija started growing stronger and prevailed most around 2003. Notice that

one person can be a member in several consortia (e.g., Mladenić Dunja is present in both consortia).

We can further exploit the hierarchical structure of the identified consortia patterns in the Consortia Development Diagram. The interactive functionalities of the visualization enable us to explore the two identified patterns in more detail, as this is depicted in Figure 36. We can identify 8 distinct consortia and their relative importance trends during the years. The Dunja Mladenic and Aleksander Pivk consortia only had importance around 1996 and 2004. Otherwise, it was quite unimportant. The Lavrač Nada and Cestnik Bojan consortia started as very important around 1996, but then diminished rapidly to nonexistence in 2002. The Erjavec Tomaž and Dunja Mladenic cluster is however low in importance, however still mostly gaining weight from 1998 to 2005. The consortium of Lavrač Nada and Džeroski Sašo is behaving in a similar way achieving even bigger importance in 2003. The consortium around Arnež Zoran Marij only started around 1998 and was then gaining in high importance and stability since then. The consortia around Filipič Bogdan and again Arnež Zoran Marij was however short lived and only lasted between the years 1997 and 2001. Around 2001 another important consortia got started. It is centered on Demšar Janez and Kukar Matjaž. The most stable of all is the consortia around Filipič Bogdan and Pivk Aleksander. Its importance is high and has only decreased slightly in the years from 1999 to 2002.

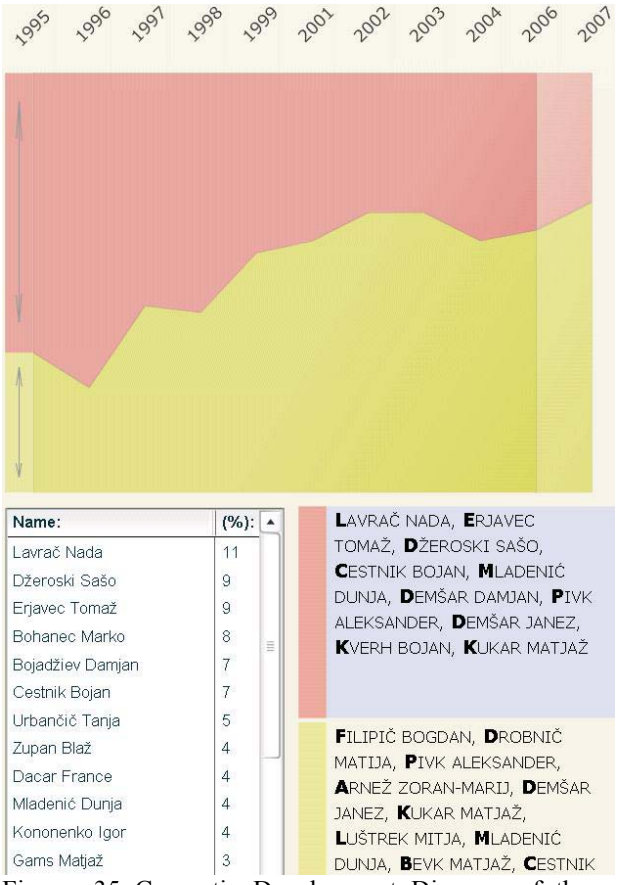


Figure: 35 Consortia Development Diagram of the collaboration network of Acad. Prof. Dr. Bratko's collaborators.

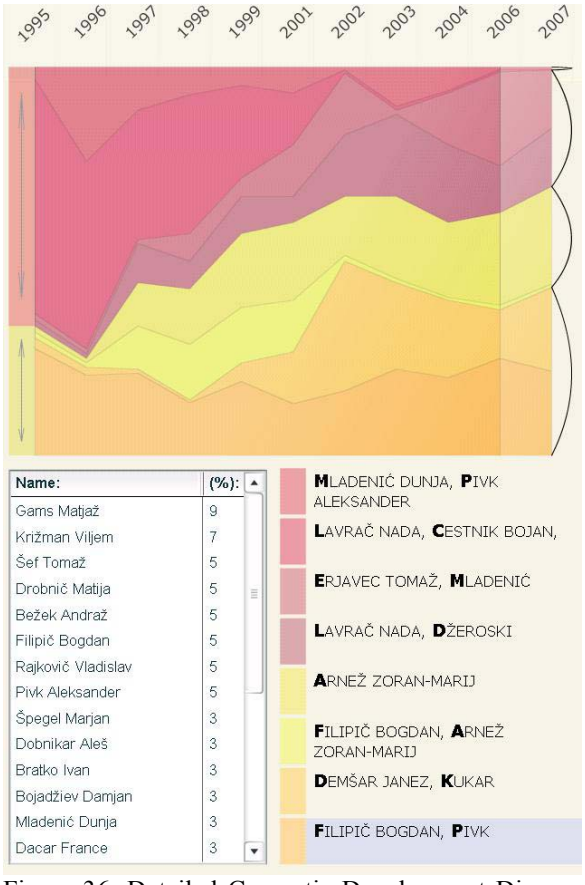


Figure 36: Detailed Consortia Development Diagram of the collaboration network of Acad. Prof. Dr. Bratko's collaborators.

Prof. Dr. Bratko's comments on the results depicted in Figure 35 included mixed impressions. He first realized that the selected names of his collaborators mostly correctly identify two

main consortia of his collaboration. Out of 20 selected names, describing the two main consortia, only 3 were labeled as questionable or inappropriate. In the first consortia cluster the inappropriate selected name is KVERH BOJAN. In the second cluster the inappropriate name was: ARNEŽ ZORAN-MARIJ. In the second cluster the name MLADENIC DUNJA was deemed as questionable. All the rest of the 20 extracted names were recognized as correct. On the other hand Prof. Dr. Bratko also missed some names of the more important people in the diagram. These are GAMS MATJAZ, ZUPAN BLAZ and ŠUC DORIAN. With regards to the depicted trends of consortia importance Prof. Dr. Bratko also recognized the depicted trends of development of the two clusters in Figure 35 as correct and with correspondence with his impression of the collaboration within the two clusters.

We can therefore conclude that the graph partitioning algorithm, which produced the Consortia Development Diagram in Figure 35, performed rather well for the task of extracting consortia trends of people collaborating with Prof. Dr. Bratko. However the mentioned problems of incorrectly selected and missing people in the description of the two collaborative clusters suggests that there is still room for improvement of the method. On the other hand it is quite possible that the mentioned problems actually originate from the incomplete input data to the analysis. For instance, the input collaboration data only includes information on collaboration of people via common projects and not via national programme groups. This may explain the non appearance of Prof. Dr. Blaz Zupan and Doc. Dr. Dorian Šuc.

5 Conclusion

In this master thesis we presented the data analysis methods for research data integration and research data analysis. We first presented machine learning methods used for semi automatic solving of the record linkage problem. Next we describe how the developed data mining methods allow discovering topics of work in textual data and collaboration patterns in collaboration networks.

The data integration and analysis methods were used in the scope of the developed IST World portal, accessible at http://www.ist-world.org. The portal attracts approximately 2000 users per day. The goal of the portal is to support the partner search process in the knowledge transfer scenario. The portal aims to support the search for best match between the industry and the academic actors by providing information on actors' past work and past collaboration. This is achieved by aggregating and integrating data on actors involved with research and development and by providing summarized information on key topics of work and most important collaborations of the selected actors. The data analysis methods were applied to facilitate the IST World portal's functionality. We applied the supervised machine learning methods on the task of integrating data describing research and development. We used the dimensionality reduction techniques in text and graphs to identify and track the research actor's competence and collaboration consortia.

The integration task within the IST World portal is necessary because the database behind the portal aggregates data from different sources. As many data sources describe the same real world entity this results in two or more records about one entity. The same problem occurs when a single data set contains duplicate records of the same entity. The problem of searching for such duplicate records is called record linkage (Winkler 2006). The task at hand was to merge all the different duplicate records together to create one single record. In the acquired integrated dataset every real life entity should only be represented once. Therefore, the ultimate goal of the IST World data integration is to combine any two records describing the same real world entity into one record inside the IST World portal.

The task of data integration in IST World was represented as a supervised machine learning problem. The standard record linkage methodology (Winkler 2006) prevents combinatorial explosion and gives theoretical grounds for decision making on whether two records are duplicates. First, we showed how to effectively reduce the magnitude of the problem by effectively reducing the considered space of all pairs of records. We exploited the full text indexing and full text querying (Zobel et al 1998) to reduce the number of potential pairs of redundant records. Next, we showed how to effectively create a function for mapping a potential pair into a useful comparison vector space. Comparison vectors were generated in an innovative way by using text mining methods like bag of words document presentation and comparison, string kernels (Lodhi et al 2002) for character and word order comparison, and edit distance (Levenshtein 1966) for capturing the notion of human typing mistakes. We then showed how to use the machine learning approach called active learning to support the manual creation of the positively and negatively labeled pairs of records. We selected to use the Support Vector machine (Vapnik 1995) algorithm as the underlying learning algorithm. Next we showed how to use the same algorithm to automatically construct a decision rule in the comparison vector space of the labeled pairs of records. The decision rule was then used to automatically label the unlabeled potential pairs as duplicates or non duplicates.

The presented data integration approach was evaluated by experimentally integrating approximately 25,000 records of organizations participating in European research projects (Thévignot 2000). We used an independent test data sample for evaluation of the integration approach. The evaluation showed that the developed integration approach works well. Our model performed slightly worse in terms of the targeted precision of 99 % - it achieved a precision score of 96%. It however slightly over performed in terms of recall (expected recall obtained on training data was 38%) by achieving recall rate of 46%. The methodology has been proven successful enough to be used in the future to integrate or clean data coming from more than 20 data sources. We believe that the data integration methods, developed and presented in this thesis, are innovative and useful.

The problem of automatic identifying and tracking the research actor's competence and collaboration consortia was tackled next by using data mining techniques. The goal was to answer the following questions automatically: Which are the main topics of an individual actor's work? Which are the main past collaborations of an actor? How did the actor's topics of work and collaboration patterns change with time? We formally defined the analysis goals as problems of automatically identifying the competence, consortia, competence development and consortia development of an actor. We described the text and graph analysis used for the task of analyzing research data. We relied on methods for text and graph representation with feature vectors for representing these modalities in a form of matrixes. Pattern identification was enabled by using dimensionality reduction with Latent Semantic Indexing (Landauer et al 1998) and Multi Dimensional Scaling (Borg et al 1997). We used k-means clustering (Kanungo et al 2002) methods for acquiring interactive temporal visualizations. These methods have already been used as described in (Fortuna et al 2005) and (Caruana et al 2005). We extended these approaches by providing additional interactive functionality to the acquired visualizations and by enabling explicit identified pattern visualization.

We developed a DocumentAtlasBG algorithm which extends the existing DocumentAtlas (Fortuna et al 2005) approach for visualizing textual data. The key extension is the identification of the main patterns in the input data and the addition of the entities representing these patterns to the resulting visualization. This allows the user to explore the association between the analyzed data and the major patterns found in the input data. The identified patterns are the most important singular vectors (Golub et al 1970) of the analyzed input matrix according to the obtained singular values. We showed that the identified patterns correspond to topics of research work when the input data to the algorithm is a corpus of textual documents representing research actors' work. We showed that the identified patterns represent the most important communities of collaboration when the input data is a matrix, which represents a network of collaboration.

We extended the temporal development visualization technique described in (Caruana et al 2005) by using hierarchical clustering and interactive features in the clustering visualization diagram. The method allows the user to interactively observe the trends of development of cluster importance in the hierarchical chain of clusters. When this visualization technique is used on textual data, the diagram can be used to show the development of topics of work through time. When used on the collaboration graph, the method can be used to partition the graph and show the development of importance of communities in the graph.

We presented the use of these data mining algorithms on four examples and showed that the results agree with human intuition. First experiment was used to intuitively evaluate the competence analysis of the faculties which are administratively part of the University of Ljubljana and intuitively compare the extracted competence words with the expected competence given the faculty name. The experiment results were satisfactory as the analysis showed that most of the extracted keywords agree with our intuition. We were satisfied with extracted competence description for 14 out of 20 faculties. In the second experiment we evaluated the algorithm for identification of temporal trends in competence development. The main idea of the experiment was to extract the competence development of the Acad. Prof. Dr. Ivan Bratko for the last 30 years of his academic career and evaluate the findings by reviewing them with the professor himself. Prof. Dr. Bratko's comments on the results were in general positive. Out of 23 extracted keywords, describing the two main competences, only 6 were labeled as questionable or inappropriate. What is more, Prof. Dr. Bratko also recognized the depicted trends of competence development as correct and in correspondence with his impression of the development of his topic of work. The main idea of the third experiment was to automatically identify the main collaboration consortia of every faculty of the University of Ljubljana using the Consortia Diagram and then intuitively compare the automatically extracted patterns with the manually extracted patterns from the visualization of the complicated collaboration network between the faculties. The results were in general positive. We were satisfied with extracted consortia description for 12 out of 18 collaborating faculties. The experiment therefore showed that the consortia analysis algorithm works well for this particular intra university collaboration network. However it also produced some unexpected results and must therefore be further improved. In the fourth and final experiment we evaluated the algorithm for identification of temporal trends in consortium importance development. The main idea of the experiment was to extract the consortium development by analyzing the collaboration network of people around Acad. Prof. Dr. Ivan Bratko. Prof. Dr. Ivan Bratko is one of the long time centers of collaboration in the artificial intelligence field of research in Slovenia. He is therefore one of the most credible people for interpreting the ten years of development of collaboration between the Slovenian researchers in this field. Prof. Dr. Bratko's comments on the automatically identified consortia development included mixed impressions. The selected names of his collaborators correctly identify two main consortia of his collaboration. On the other hand Prof. Dr. Bratko expressed that some names of the more important people were missing in the diagram. With regard to the identified trends of consortia importance, Prof. Dr. Bratko recognized their development as correct and in correspondence with his impression of the collaboration within the two clusters. However, the mentioned problems suggest that there is either still room for improvement of the algorithm or that there are latent problems with the input collaboration data to the analysis.

In general the four conducted experiments increased our belief in the usefulness of the developed and implemented data mining algorithms.

Possible future work includes improvement of the data integration and data analysis methods together with conducting more empirical evaluations. We first plan to more rigorously evaluate the described data analysis methods. We can involve the users of the IST World portal in the methods assessments. Another way of improving the data analysis is to try to automatically tune the parameters of the used methods. Future steps include extensive use of the data integration methodology to integrate more than 20 different datasets in IST World repository into a single dataset. The Feature Generation step of the integration approach could be improved using the collaboration information as a way of comparing the two records. It

could also be redesigned for every particular data integration problem at hand. Active Learning step and the Classifier Induction step of the approach might be further improved by utilizing transductive machine learning algorithms. Transfer learning methods might enable reuse of one integration model on several instances of integration problems. The embedding IST World portal could be further improved by adding introspection and drill down data selection methods.

6 Razširjeni povzetek v slovenskem jeziku

6.1 Uvod

V magistrskem delu opišemo uporabo metod strojnega učenja za nalogi integracije in analize podatkov o raziskovalni dejavnosti. Metode so implementirane v okviru informacijskega sistema IST World, ki omogoča iskanje partnerjev v kontekstu procesa prenosa znanja iz akademskega v industrijski sektor.

Ob iskanju partnerjev za prenos znanja se sprašujemo predvsem vprašanja o preteklem delu in sodelovanju potencialnih partnerjev: Kaj so glavna področja dela akterjev v kontekstu predlagane skupine in v absolutnem kontekstu? Kakšni so glavni vzorci sodelovanja akterjev v kontekstu predlagane skupine in v absolutnem kontekstu? Kako se pomembnost različnih področij dela in vzorcev sodelovanja spreminja skozi čas, glede na predlagano skupino in v absolutnem okviru?

Za lažji odgovor na ta vprašanja so se pojavili informacijski portali, ki zbirajo podatke o potencialnih partnerjih. Omogočajo vpogled v podrobnosti dela in sodelovanja raziskovalcev, kot so npr. informacije o financiranih znanstvenih projektih in uslugah ter izdelkih, ki jih nudijo industrijski akterji.

Z uporabo metod strojnega učenja in rudarjenja v podatkih lahko take informacijske sisteme močno izboljšamo v smeri povečevanja količine vsebovanih podatkov ter z omogočanjem avtomatske analize le teh. Metode strojnega učenja omogočajo učinkovito zlivanje podatkov iz dveh podatkovnih zbirk v eno podatkovno zbirko. Podatke v obliki zapisov o delu in sodelovanju partnerjev lahko analiziramo z metodami strojnega učenja, kar omogoča vizualizacijo glavnih tematik dela, glavnih vzorcev sodelovanja ter spreminjanja tematike dela in sodelovanja skozi čas.

Glavni prispevek tega magistrskega dela je uporaba nekaterih algoritmov strojnega učenja in rudarjenja v besedilih za podporo (1) integracije podatkov, ki prihajajo iz različnih virov in (2) uporaba ter in razširitev algoritmov za rudarjenje v podatkih za pridobivanje informacij o področjih dela in vzorcih sodelovanja raziskovalnih akterjev. Algoritmi delujejo v kontekstu razvitega spletnega informacijskega sistema, ki se imenuje IST World portal.

6.2 Portal IST WORLD

V tem podpoglavju bomo opisali informacijski sistem IST World (Jörg et al 2006). IST World je spletni informacijski sistem, za podporo procesa iskanja partnerja za prenos znanja. Portal je dostopen na spletnem naslovu: http://www.ist-world.org, kjer ga dnevno obišče približno 2000 uporabnikov. Ker je portal namenjen podpori iskanja partnerjev je njegova funkcionalnost izražena v obliki aplikacije za rudarjenje v podatkih, specializirani za odkrivanje znanja o raziskovalcih, raziskovalnih organizacijah, projektih in publikacijah. V okolju tega sistema delujejo v tem magistrskem delu opisani algoritmi strojnega učenja in rudarjenja v podatkih za namen integracije in analize podatkov.

6.2.1 Funkcionalnost portala IST World

Funkcionalnost portala IST World (Jörg et al 2006) je namenjena iskanju glavnih tematik raziskovalnega dela in glavnih vzorcev sodelovanja. Zato je portal implementiran v obliki aplikacije za rudarjenje v podatkih. Tipični scenarij uporabe spletnega portala je sestavljen iz treh korakov:

1. **iskanje in navigacija** za izbiro relevantnih zapisov v podatkovni zbirki. Portal omogoča štiri različne načine izbire zapisov za analizo. Ti so kompleksno iskanje, sorodnost zapisov, sodelovanje zapisov in kategorizacija zapisov
2. **izbira orodja analize** za avtomatsko analizo izbranih zapisov. Implementirane metode analizirajo predvsem tekstovne podatke in podatke o sodelovanju izbranih entitiet. Uporabnik si lahko izbere enega od šest različnih načinov analize skupaj s pripadajočimi parametri. Metode analize so: tekstovni prikaz kompetenc, prikaz grafa sodelovanja, vizualni prikaz kompetenc, prikaz trendov kompetenc, prikaz vzorcev sodelovanja in prikaz trendov vzorcev sodelovanja
3. **pregled rezultatov** za analizo rezultatov avtomatske analize izbranih zapisov. V tem koraku lahko uporabnik uporabi interaktivne lastnosti prikazov rezultatov različnih metod analize za poglobljeno analizo prikazanih rezultatov.

6.2.2 Zbirka podatkov v portalu IST World

Da bi uporabnikom portala IST World omogočili kar najbolj točno informacijo o tematikah raziskovalnega dela in vzorcih sodelovanja raziskovalnih akterjev morajo podatki, iz katerih s pomočjo metod analize pridobivamo informacije, biti kar se da bogati in točni. To pomeni, da želimo v zbirki portala IST World združiti podatke iz različnih podatkovnih virov. Podatki portala IST World so zato zbrani iz različnih podatkovnih virov kot so: institucionalne zbirke podatkov, nacionalne zbirke raziskovalnih podatkov, regijski poslovni imeniki in globalni indeksi publikacij. Trenutno zbirka podatkov vsebuje podatke iz naslednjih podatkovnih zbirk:

- Slovenska zbirka **SICRIS** (Seljak 2001) vsebuje informacijo o 11930 slovenskih raziskovalcih, 1794 slovenskih raziskovalnih organizacijah in 4780 slovenskih raziskovalnih projektih.
- Evropska zbirka **CORDIS** (Thévignot 2000) vsebuje podatke o približno 70 000 organizacijah in 20 000 projektih financiranh s strani Evropse Unije.
- Del zbirke podatkov **GOOGLE Scholar** (Jacsó 2005) vsebuje podatke o 847706 znanstvenikih in 1138782 njihovih publikacij.
- Poleg tega vsebuje še podatke iz drugih podatkovnih virov: LT World database, Bulgarian CRIS database, Estonian RIS database, Serbian database, Singleimage database, Huncris CRIS dataset, Latvian CRIS dataset, Turkish SME database, Czech database, Cyprus database, Romanian database, Slovak database, SINNIN database.

Podatki, ki izvirajo iz različnih podatkovnih virov, morajo biti združeni v enotno podatkovno zbirko, v kateri mora biti vsaka entiteta, ki obstaja v resničnem življenju, predstavljena samo enkrat.

6.3 Integracija Podatkov

V temu poglavju opišemo metode, ki smo jih uporabili za integracijo podatkov pridobljenih iz različnih podatkovnih virov v en samo zbirko podatkov. Opis začnemo s foramalno teorijo povezovanja zapisov (*record linkage*) (Winkler 2006). Nato predstavimo metodologijo, ki nam omogoča reševanje tega problema v zbirki podatkov portala IST World. Končamo pa z empiričnim ekperimentom, ki potrdi delovanje predstavljene metodologije za integriranje podatkov.

6.3.1 Definicija problema

Problem bomo formalno predstavili po (Winkler 2006). Vsak par zapisov (a,b) iz množice kartezičnega produkta dveh množic podatkov A × B je preiskušen na ujemanje. Cilj je klasificirati tak par zapisov (a,b) iz množice kartezičnega produkta A × B v množico pravih parov M in množico ne pravih parov U.

6.3.2 Metodologija integracije podatkov v zbirki podatkov IST World

Razvili smo metodologijo integracije podatkov v zbirki podatkov IST World portala. Metodologija nadgrajuje tradicionalne metode integracije podatkov z modernimi algoritmi strojnega učenja, ki omogočajo hitro in učinkovito integracijo podatkov.

Tradicionalni način reševanja problema integracije podatkov oz. problema povezovanja dvojnih zapisov je opisan v (Winkler 2006). Iskanje dvojnih zapisov se izvrši v dveh korakih. V prvem koraku se iz iskanja izločijo tisti pari zapisov, ki glede na določeno hevristiko zelo verjetno niso pravi pari. Ta korak se imenuje bločenje. Drugi korak postopka iskanja pravih parov se imenuje parjenje. V tem koraku je potrebno za vsak preostali potencialni par zapisov sprejeti odločitev, ali je par predstavlja pravi ali ne pravi par.

V razviti IST World metodologiji integracije podatkov smo v prvem koraku uporabili hevristiko, da so pravi pari tisti, ki vsebujejo vsaj podobne besede. Zato smo s tehniko obrnjenega indeksa besed v zapisih (Zobel et al 1998) poiskali le tiste pare, ki vsebujejo vsaj približno enake besede. V drugem koraku smo uporabili algoritme strojnega učenja za avtomatsko indukcijo klasifikatorjev in algoritme za odkrivanje znanja v besedilih.

Bolj natančno je razvita metodologija integracije podatkov v portalu IST World sestavljena iz petih korakov. V prvem koraku **bločenja** uspešno omejimo problem na majhno podmnožico parov iz A × B z uporabo obrnjenega indeksa besed. V drugem koraku **generiranja atributov** realiziramo funkcijo za opis vsakega para (a,b) v vektorskem prostoru primerjav Γ. Funkcija opisa γ temelji na metodah iskanja zakonitosti v besedilih, kot so primerjave besedil s pomočjo vreče besed (*bag of words*), jeder nizov (*string kernel*) in urejevalne razdalje (*edit distance*). V tretjem koraku **aktivnega učenja** uporabimo metodo podpornih vektorjev (*support vector* machine) za podporo ročnega klasificiranja nekaterih parov zapisov v množico pravih parov M in množico nepravih parov U. V četrtem koraku uporabimo metode strojnega učenja, kot so metoda podpornih vektorjev in prečno preverjanje, za avtomatsko **indukcijo klasifikacijskega pravila** (2) glede na pare v množicah M in U. V zadnjem koraku **uporabe avtomatsko induciranega pravila** to uporabimo na preostalih parih in jih tako avtomatsko klasificiramo v množici M in U.

6.3.3 Ocena metodologije

Razvito metodologijo integracije podatkov lahko ocenimo s preiskusom na nalogah odkrivanja dvojnih zapisov v množici podatkov. Uspešnost metodologije lahko ocenimo z merjenjem klasifikacijske točnosti, priklica (Grossman et al 1998), oblike krivulje ROC (Spackman 1989) in površine (AUC) pod krivuljo ROC (Spackman 1989) na naključnem vzorcu avtomatsko klasificiranih potencialnih parov, ki jo pridobimo z naključnim izbiranjem parov iz množice parov (a,b) v že bločeni podmnožici množice A × B.

Metodologijo integracije podatkov smo ocenili na nalogi iskanja dvojnih zapisov v zbirki podatkov CORDIS FP6 (Thévignot 2000). Ta zbirka vsebuje podatke o raziskovalnih projektih in raziskovalnih ustanovah, ki so financirani s strani Evropske Unije v okviru 6. Okvirnega programa. Po eni strani ta zbirka vsebuje seznam 5094 raziskovalnih projektov. Po drugi strani pa ne vsebuje enotnega seznama organizacij, ki na teh projektih sodelujejo, ampak le seznam za vsak projekt posebej. To pomeni, da je zaradi ročnega vnosa podatkov prišlo do različnih zapisov enih in istih sodelujočih organizacij. Pojavljajo se na primer različni zapisi organizacije Inštitut Jožef Stefan kot so: Jozef Stefan Institute, Institut Jozef Stefan, Lnstitute Jozef Stefan. Naša naloga je identificirati take dvojnike organizacij z uporabo razvite metodologije integracije podatkov v IST Worldu.

Najprej smo z uporabo **bločenja** zmanjšali število potencialnih parov 24965 organizacij, ki so omenjene v tej zbirki podatkov. Za vsako od teh smo zgradili seznam 10 najbolj verjetnih dvojnih zapisov. Tako smo pridobili seznam dolg približno 220,000 potencialnih parov. S korakom bločenja smo torej uspešno omejili množico vseh 462,000,000 potencialnih parov organizacij. Z **avtomatskim generiranjem atributov** smo vsak par opisali s približno 20,000 atributi, ki opisujejo različne načine ujemanja para. V koraku **aktivnega učenja** smo z izrabo interakcije med učnim algoritmom SVM in uporabnikom uspešno našli in označili približno 700 pravih in 1300 nepravih parov organizacij. Označene pare smo uporabili za **induciranje klasifikacijskega pravila** nad atributnim prostorom, kar nam omogoča avtomatsko klasifikacije še neoznačenih parov, zapisanih v tem atributnem prostoru. S pomočjo prečnega preverjanja smo določili parameter uporabljenega algoritma C = 15 in parameter praga p=1.58. Take nastavitve so na učnih podatkih s prečnim preverjanjem pomenile uspešnost AUC = 0.91. V zadnjem koraku **uporabe induciranega pravila** smo pravilo pognali na vseh preostalih 218,000 potencialnih parih. Na ta način smo avtomatsko odkrili še 4300 dodatnih pravih parov zapisov.

Uspešnost metodologije smo ocenili z ročnim pregledom 1000 naključno izbranih parov iz množice 218,000 neoznačenih bločenih parov. Ugotovili smo, da je model pravilno ocenil 30 pravih parov in 934 ne pravih parov. Model je nepravilno klasificiral 1 nepravi par in 35 pravih parov. Klasifikacijska natančnost (*precission*) modela na tej naključni množici je torej 97%. Priklic (*recall*) pa je 46%. Področje pod krivuljo ROC (AUC) je veliko 0.96.

Eksperiment je torej dokazal, da je razvita metodologija integracije podatkov uspešna pri reševanju naloge za avtomatsko iskanje dvojnikov znotraj ene zbirke podatkov. Metodologija je omogočila pol avtomatsko identifikacijo 5000 parov pravih dvojnih zapisov. V prihodnosti bomo metodologijo uporabili za iskanje dvojnikov med vsemi pari množic podatkov, ki so vsebovane v zbirki podatkov IST World. Metodologija je namreč robustna na vzorce, ki se pojavljajo v dvojnikih, saj lahko korak generiranja atributov po potrebi prilagodimo lastnostim množic podatkov, ki jih bomo integrirali. Metodologijo bi lahko še dodatno izboljšali z uporabo transduktivnih metod strojnega učenja (Vapnik 1999). V korakih

aktivnega učenja in induciranja modelov bi tako poleg označenih primerov lahko upoštevali tudi vzorce v še neoznačenih primerih. To bi omogočilo še bolj uravnoteženo vzorčenje v koraku aktivnega učenja in induciranje še bolj natančnih klasifikatorjev v koraku induciranja modela.

6.4 Analiza raziskovalnih podatkov

Analiza raziskovalnih podatkov v kontekstu portala IST World je osredotočena na iskanje informacij, ki lahko pripomorejo k bolj uspešnemu procesu iskanja partnerjev. Z uporabo portala je mogoče dobiti informacije o tem, kaj so glavna področja dela raziskovalnih akterjev in kakšni so vzorci njihovega sodelovanja. V nadaljevanju zato natančno definiramo pojme kot so **kompetenca**, **konzorcij** ter **časovni razvoj kompetenc in konzorcijev**. Zatem opišemo algoritme za avtomatsko identifikacijo kompetenc, konzorcijev in njihovega časovnega razvoja. Na koncu s praktičnimi eksperimenti pokažemo, da algoritmi glede na intuitivno presojo vračajo pravilne rezultate.

6.4.1 Cilji analize

Kompetenco definiramo takole: *Kompetenca nekega akterja je s ključnimi besedami opisano področje dela, ki akterja razlikuje od ostalih akterjev v kontekstu analize.* Problem avtomtskega iskanja kompetenc torej lahko definiramo tako:

PROBLEM: **Iskanje kompetenc.**
IZ: tekstovnih dokumentov, ki vsebujejo tekstovne opise dela akterjev
POIŠČI: tiste besede, ki najbolj splošno razlikujejo področja dela med posameznimi akterji, ki so vključeni v analizo.

Konzorij definiramo takole: *Konzorcij nekega akterja je opis skupine akterjev, s katerimi dotični akter sodeluje bolj kot z ostalimi akterji. Opis konzorija sestavljajo tisti člani skupine, ki v njej najbolj sodelujejo.* Problem avtomatskega iskanja konzorcijev torej lahko definiramo tako:

PROBLEM: **Iskanje konzorcijev.**
IZ: uteženega grafa sodelovanja akterjev, ki so del konteksta analize
POIŠČI: imena tistih organizacij, ki najbolj splošno razlikujejo med posameznimi akterji glede na sodelovanje med akterji, ki so vključeni v analizo.

Časovni razvoj kompetenc definiramo tako: *Časovni razvoj kompentenc je trend deleža določene tematike dela v skupnem delu akterja, skozi določeno časovno obdobje, v kontekstu dela ostalih akterjev zajetih v analizo.* Problem avtomatskega iskanja časovnega razvoja kompetenc torej definiramo tako:

PROBLEM: **Razpoznavanje časovnega razvoja kompetenc.**
IZ: časovno označenih tekstovnih podatkov o aktivnostih akterjev
POIŠČI: ključne besede, ki opisujejo kompetence vsakega avtorja in trend časovnega razvoja te kompetence

Časovni razvoj konzorcijev definiramo tako: *Časovni razvoj konzorcijev je trend deleža določene skupine sodelujočih v celotnem sodelovanju nekega akterja v kontekstu vseh*

akterjev, ki so vključeni v analizo. Problem avtomatskega iskanja časovnega razvoja konzorcijev lahko torej definiramo tako:

PROBLEM: **Razpoznavanje časovnega razvoja konzorcijev.**
IZ: uteženega grafa sodelovanja akterjev s časovno označenimi vozlišči
POIŠČI: imena akterjev, ki opisujejo konzorcije sodelovanja posameznega avtorja, in trend časovnega razvoja tega konzorcija.

6.4.2 Metode analize

Metode analize, ki smo jih uporabili za avtomatsko razpoznavanje kompetenc in konzorcijev, temeljijo na metodah vizualizacije besedil (Fortuna et al 2005), razširjenih z metodo razpoznavanja glavnih vzorcev v besedilih. Metode za analizo časovnega razvoja komptenc in konzorcijev temeljijo na metodah za gručenje podatkov (Kanungo et al 2002) v skupine ter vizualizacije časovnega razvoja skupin (Caruana et al 2005) z dodano funkcionalnostjo, ki omogoča hierarhično iskanje skupin in interaktivnost. Te metode so, z namenom iskanja kompetenc in njihovega časovnega razvoja, uporabljene nad tekstovnimi podatki. Za namen iskanja konzorcijev in časovnega razvoja le teh pa so uporabljene nad podatki, ki opisujejo grafe sodelovanja.

Analiza kompetenc z algoritmom DocumentAtlasBG

Za analizo kompetenc smo razvili algoritem za vizualizacijo besedil skupaj s prepoznanimi vzorci v teh besedilih. Algoritem se imenuje ***DocumentAtlasBG.***

Algoritem poiskuša zmanjšati dimenzionalnost analiziranih vektorjev z zaporedjem dveh metod SVD (Golub et al 1970) in MDS (Borg et al 1997). Pri tem v drugem koraku ne upošteva le analiziranih dokumentov ampak tudi seznam umetnih dokumentov, ki označujejo tematike dela. Ti dokumenti so umetno skonstruirani iz singularnih vektorjev, ki so stranski produkt metode za zmanjšanje dimenzionalnosti SVD. Vsak singularni vektor tako pretvorimo v dva nova tematska dokumenta. Pozitivne vrednosti singularnega vektorja ustrezajo besedam, ki jih najdemo v prvem tematskem dokumentu. Negativne vrednosti singularnega vektorja pa ustrezajo besedam v drugem tematskem dokumentu. Na ta način smo pridobili dvakrat toliko tematskih dokumentov kolikor singularnih vektorjev nam je uspelo izračunati. V koraku MDS se vsi vhodni dokumenti, skupaj s pridobljenimi tematskimi dokumenti, projecirani v 2D prostor na način, ki najbolje opisuje njihovo podobnost. Tako smo pridobili vizualizacijo tekstovnih podatkov skupaj z razpoznanimi tematikami v teh podaktih.

Analiza časovnega razvoja kompetenc z gručenjem tekstovnih dokumentov

Za analizo časovnega razvoja kompetenc smo uporabili metode gručenja besedil z metodo k-means (Kanungo et al 2002) in prikaza časovnega razvoja gruč skozi čas (Caruana et al 2005).

Z metodo gručenja k-means (Kanungo et al 2002) najprej iz seznama dokumentov pridobimo hierarhično urejene skupine dokumentov. Razmerja pomembnosti skupin, glede na število vsebovanih dokumentov pa nato izrišemo za vsako časovno obdobje posebej (Caruana et al 2005). Časovna obdobja so enakomerno razporejena glede na najstarejšo in najnovejšo časovno oznako v analiziranih dokumentih.

Analiza konzorcijev z algoritmi DocumentAtlasBG

Za analizo konzorcijev v grafih sodelovanja smo ponovno uporabili algoritem **DocumentAtlasBG**, ki je bil razvit za vizualizacijo besedil.

Graf sodelovanja smo najprej zapisali v obliki vektorskega zapisa, ki opisuje vsako vozlišče z atributi, ki določajo stopnjo povezave z ostalimi vozlišči. Za ta namen smo uporabili algoritem ScentTrail (Olston et al 2003), ki vektorje, ki opisujejo sosednost v uteženem grafu, določijo s simuliranjem širjenja plina po povezavah grafa. Prepustnost povezav za plin je pri tem povezana z njihovimi utežmi v grafu sodelovanja. Na ta način pridobimo atributni opis vozlišč grafa, ki ga lahko naprej analiziramo tudi z metodami za analizo matričnih podatkov.

Uporaba algoritma **DocumentAtlasBG** za analizo z matriko predstavljenega uteženega grafa ustreza tudi reševanju problema iskanja konzorcijev kot smo to opisali zgoraj. Algoritem namreč omogoča razpoznavanje vzorcev v matriki sodelovanja in povezavo med temi vzorci ter vozlišči v grafu. Razpoznani vzorci izvirajo iz singularnih vektorjev matrike, ki opisujejo ravno glavne smeri sodelovanja med vozlišči. Zato vsak vzorec določa svoj konzorcij sodelovanja. Ko so vzorci prepoznani, jih algoritem skupaj z vozlišči izriše v 2D diagramu.

Analiza časovnega razvoja konzorcijev z gručenjem vozlišč grafa

Za analizo časovnega razvoja konzorcijev smo ponovno uporabili metode gručenja z metodo k-means (Kanungo et al 2002) in prikaza časovnega razvoja gruč skozi čas (Caruana et al 2005).

Ta problem smo reševali z uporabo gručenja na vozliščih grafa dvo partitnega grafa, kjer ena particija predstavlja zapise dela raziskovalnih akterjev, npr. publikacij ali projektov. Druga particija vozlišč pa predstavlja akterje , ki so avtorji aktivnosti. Povezava med publikacijo in njenim avtorjem npr. ustreza povezavi med vozliščem v particiji akterjev in pariticiji zapisov aktivnosti teh akterjev. Tak graf sodelovanja je v vektorskem opisu predstavljen z matriko, v kateri so vektorji opisa zapisov aktivnosti predstavljeni z akterji, ki na teh zapisih sodelujejo. Vektorji torej predstavljajo razmerja in podobnosti sodelovanja akterjev na posameznih zapisih. Vsak vektor je dodatno označen s časovno oznako, ki določa časovni okvir zapisa aktivnost predstavljene z vektorjem.

Z metodo gručenja k-means (Kanungo et al 2002) najprej iz seznama vektorjev pridobimo hierarhično urejene skupine vektorjev. Razmerja pomembnosti skupin, glede na število vsebovanih dokumentov pa nato narišemo za vsako časovno obdobje posebej (Caruana et al 2005). Časovna obdobja so enakomerno razporejena glede na najstarejšo in najnovejšo časovno oznako v analiziranih dokumentih.

6.4.3 Poskusi analize podatkov o raziskovalni dejavnosti

V prejšnjih sekcijah smo opisali algoritme za razpoznavanje kompetenc in konzorcijev. V tem podpoglavju pa opišemo nekaj poskusov analize raziskovalnih podatkov z opisanimi metodami.

Analiza kompetenc

V prvem poskusu poskušamo ugotoviti, kaj so glavne kompetence fakultet članic Univerze v Ljubljani. Ideja poskusa temelji na intuiciji, da ime fakultete dokaj dobro opisuje kompetence fakultete v kontekstu vseh fakultet članic dotične univerze. Kompetence poskušamo avtomatsko razpoznati z analizo tekstovnega opisa delovanja fakultet v okviru projektov, na katerih so sodelovale glede na podatke v zbirki SICRIS (Seljak 2001).

Slika 37 prikazuje rezultate avtomatske analize kompetenc. Diagram prikazuje 20 fakultet Univerze v Ljubljani v obliki rdečih križcev na ozadju razpoznanih kompetenc v obliki zelenih ključnih besed. Seznam predstavljenih ključnih besed vsebuje besede kot so: FARMACEVTSKA, SOCIAL WORK, UPRAVA, THEOLOGY, TEKSTILIJ, KONSTRUKCIJ, etc. Te besede predstavljajo najpomembnejše kompetence, ki smo jih našli v tej skupini dvajsetih organizacij. Položaj vsake organizacije na sliki je povezan s položajom kompetenc, ki so za organizacije pomembni. To omogoča ugotavljanje kompetenc vsake organizacije posebej. Na sliki 42 lahko glede na njihov položaj opazimo pet glavnih skupin fakultet. Glavne skupine so:

- Farmacevtska skupina: Fakulteta za Farmacijo. Veterinarska Fakulteta, Biotehnična fakulteta, Medicinska Fakulteta, (Fakulteta za kemijo in kemijsko tehnologijo)
- Konstrukcijska skupina: Fakulteta za naravoslovje, Fakulteta za gradbeništvo in geodezijo, Fakulteta za Matematiko in Fiziko, (Fakulteta za računalništvo in fiziko, Fakulteta za elektrotehniko, Fakulteta za strojništvo)
- Športna skupina: Fakultea za šport, Fakulteta za pomorstvo in promet
- Skupina za socialno delo: Fakulteta za socialno delo, Pedagoška fakulteta
- Pravniška in administrativna skupina: Pravna Fakulteta, Fakulteta za družbene vede, Filozofska Fakulteta, Fakulteta za upravo, Ekonomska Fakulteta

Na podlagi teh rezultatov lahko intuitivno sklepamo, da kompetenčni diagram na sliki 37 v grobem pravilno izraža kompetence posameznih fakultet.

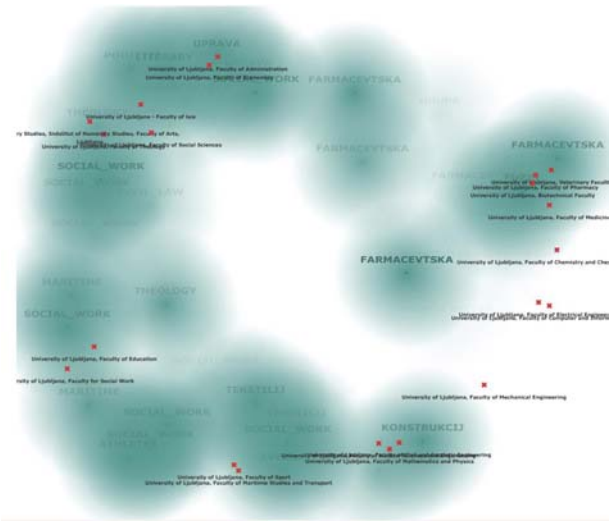
Analiza časovnega razvoja kompetenc

V drugem poskusu poskušamo oceniti algoritem za avtomatsko identifikacijo trendov v časovnem razvoju kompetenc. Cilj poskusa je razpoznavanje časovnega razvoja kompetenc akad. Prof. dr. Ivana Bratka s pomočjo opisanih algoritmov. Rezultate poskusa smo ovrednotili s pomočjo samega profesorja, ki je podal njegovo mnenje o pravilnosti rezultatov. Algoritem za časovno analizo razvoja kompetenc smo uporabili na besedilih v naslovih 329 publikacij prof. dr. Bratka v zbirki podatkov Google Scholar (Jacsó 2005).

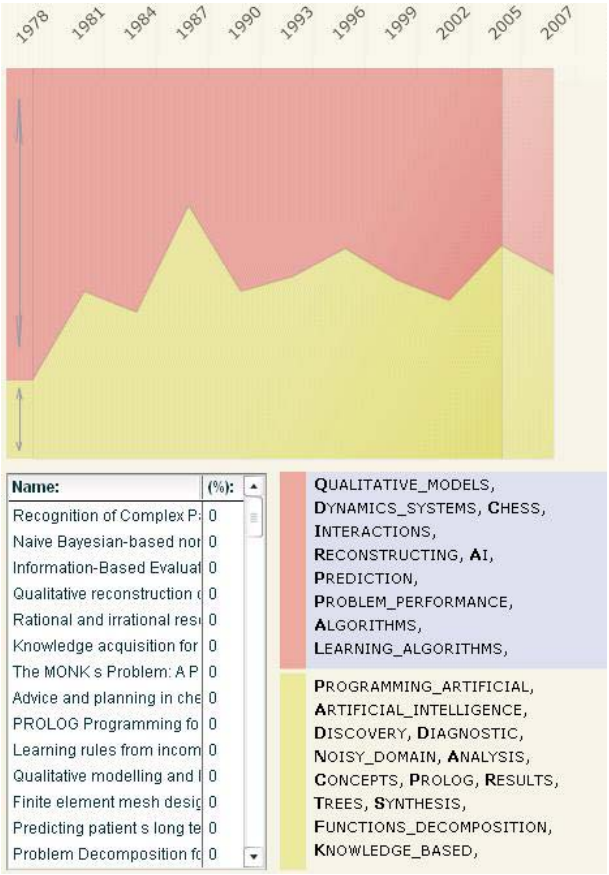
Slika 38 prikazuje vizializacijo razvoja kompetenc pridobljeno z uporabo algoritmov, ki smo jih opisali v podpoglavju Metode analize časovnega razvoja kompetenc. Slika prikazuje glavni področji kompetenc profesorja in razvoj le teh v tridesetletnem časovnem obdobju njegovega znanstvenega dela.

Komentarji prof. dr. Bratka k rezultatom na sliki 38 so bili v splošnem pozitivni. Najprej je potrdil, da izbrane ključne besede pravilno določajo njegovi glavni dve tematiki dela. Od triindvajset izbranih besed jih le šest ni popolnoma ustrezalo njegovi presoji. Poleg tega je prof. dr. Bratko ocenil tudi, da so prikazani trendi razvoja njegovih kompetenc pravilni in ustrezajo razvoju tematik njegovega dela. Primer pravilno razpoznanega trenda je velik skok v

obdobju okoli leta 1987 v pomembnosti druge skupine kompetenc, saj ta ustreza začetkom bolj intenzivnih raziskav na področju programskega jezika PROLOG.



Slika 37: Kompetenčni diagram fakultet članic Univerze v Ljubljani.

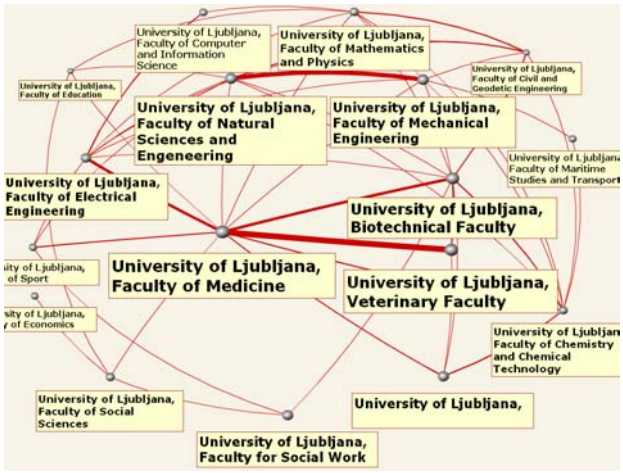


Slika 38: Diagram razvoja kompetenc na podlagi 329 publikacij acad. prof. dr. Ivana Bratka prikazuje časovni razvoj pomembnosti posameznih tematik njegovega dela.

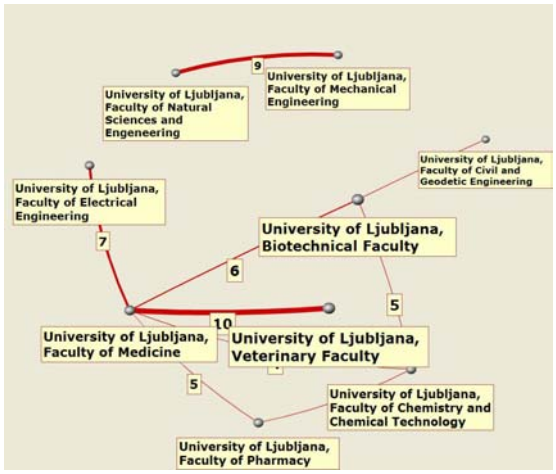
Analiza konzorcijev

V tretjem poskusu smo analizirali vzorce sodelovanja na skupnih projektih vseh fakultet članic Univerze v Ljubljani. Cilj poskusa je avtomatsko razpoznati konzorcije v tej mreži sodelovanja in jih intuitivno primerjati z vzorci v grafu sodelovanja, ki jih lahko razpoznamo z ročno analizo le tega. Ta poskus smo izvedli na podatkih o sodelovanju dvajsetih fakultet članic Univerze v Ljubljani na slovenskih nacionalnih raziskovalnih projektih (Seljak 2001). Mreža sodelovanja je prikazana na slikah 39 in 40. Vsaka točka predstavlja fakulteto, vsaka utežena povezava predstavlja sodelovanje med fakultetami na skupnih projektih.

Ko algoritem za analizo konzorcijev uporabimo na mreži sodelovanja na sliki 39 pridobimo vizualizacijo na sliki 41. Na sliki so organizacije vidne v obliki rdečih križcev, razpoznani konzorciji pa v obliki zelenih oblakov v ozadju. Vsak konzorcij je označen z vodilno organizacijo konzorcija. Prosojnost oblaka označuje pomebnost konzorcija. Diagram razkrije, da je najpomembnejši razpoznani konzorcij v analizirani mreži sodelovanja tisti okoli Medicinske Fakultete (v desnem zgornjem kotu). Ostali pomembni identificirani konzorciji so še okoli Fakultete za Strojništvo in Ekonomske Fakultete. Konzorciji okoli Veterinarske Fakultete in Filozofske Fakultete pa so manj pomembni.



Slika 39: Mreža sodelovanja dvajsetih fakultet članic Univerze v Ljubljani.



Slika 40: Mreža sodelovanja fakultet, ki bolj intenzivno sodelujejo.

Z uporabo interaktivnosti diagrama lahko uporabnik za vsako organizacijo razbere konzorcije, v katerih organizacija najpogosteje sodeluje. Na ta način smo primerjali pripadajoče konzorcije posameznih fakultet s konzorciji, ki smo jih sami razbrali iz slik 39 in 40. Ugotovili smo, da večina izbranih konzorcijev ustreza intuitivnim vzorcem na teh slikah. Od dvajsetih fakultet so bili dvanajst fakultetam pravilno določeni najpomembnejši konzorciji. Konzorciji šestih fakultet pa niso popolnoma ustrezali našim opažanjem. V enem primeru so bili identificirani konzorciji napačni.

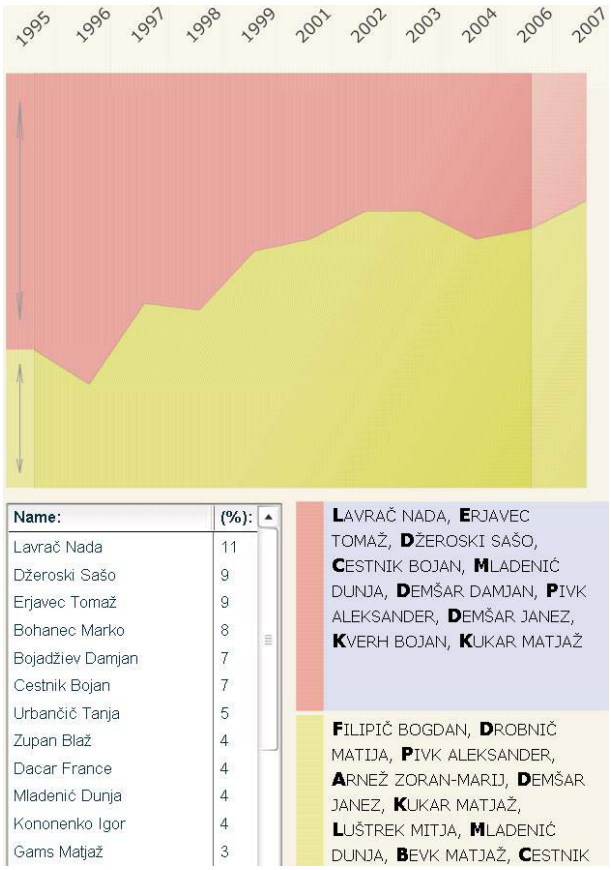
Analiza časovnega razvoja konzorcijev

V četrtem in zadnjem poskusu smo preiskusili algoritem za analizo časovnega razvoja konzorcijev. Cilj poskusa je analizirati razvoj konzorcijev na mreži sodelovanja ljudi, s katerimi sodeluje ali je sodeloval akad. prof. dr. Bratko. Rezultate poskusa smo ocenili s pomočjo prof. dr. Bratka samega, ki je podal njegovo videnje ustreznosti rezultatov analize. V analizo smo vključili 48 ljudi, ki so v zbiriki podatkov SICRIS (Seljak 2001) zapisani kot sodelavci prof. dr. Bratka.

Slika 42 prikazuje vizualizacijo rezultatov analize časovnega razvoja konzorcijev, s pomočjo algoritmov opisanih v podpoglavju Metode analize. Slika prikazuje dve skupini sodelavcev in 11 let razvoja intenzivnosti njihovega sodelovanja. Iz slike je vidno, da je okoli leta 1995 prevladovala skupina zbrana okoli Nade Lavrač, zatem, okoli leta 1997 pa je začela prevladovati skupina zbrana okoli Bogdana Filipiča in Drobniča Matije.

Prof. dr. Bratko je rezultate na sliki 42 komentiral z mešanimi občutki. Najprej je ugotovil, da je algoritem pravilno izbral imena ljudi, ki so najbolj pomembni za opis dveh pomembnih skupin sodelovanja. Izmed dvajsetih prikazanih imen so bila le tri označena kot vprašljiva za prikaz. Po drugi strani pa je je prof. dr. Bratko v prikazanih seznamih pogrešil imena ljudi, ki so po njegovem menju ravno tako pomembni za opis skupin. Glede prikazanih trendov razvoja pa je bilo ugotovljeno, da ustrezajo njegovemu videnju razvoja sodelovanja v prikazanih skupinah raziskovalcev.

Slika 41: Diagram konzorcijev v mreži sodelovanja fakultet članic Univerze v Ljubljani.



Slika 42:Diagram časovnega razvoja konzorcijev sodelavcev akad. prof. dr. Bratka

6.5 Sklep

V magistrskem delu smo predstavili metode strojnega učenja za podporo integracije in analize podatkov o raziskovalni dejavnosti. Najprej smo predstavili, kako lahko metode strojnega učenja uporabimo za pol avtomatsko iskanje podvojenih zapisov. Zatem smo predstavili metode rudarjenja v podatkih, ki nam omogočajo razpoznavanje tematik dela in vzorcev sodelovanja znanstvenih akterjev.

Metode integracije in analize so bile predstavljene in preiskušene v kontekstu spletnega portala IST World, ki je dosegljiv na spletnem naslovu: http://www.ist-world.org. Cilj portala je podpora procesu iskanja partnerjev v scenariju prenosa znanja.

V prihodnosti bomo poskušali še bolj empirično preiskusiti opisane metode rudarjenja v raziskovalnih podatkih. Metode rudarjenja v raziskovalni dejavnosti bi lahko izboljšali s pomočjo še bolj fleksibilnih metod izbire podatkov za analizo, ki bi omogočale operacije rudarjenja kot so vpogled in povečava. Dalje bi lahko metode analize podatkov izboljšali z uporabo algoritmov za avtomatsko izbiranje parametrov metod ali z vključitvijo uporabnika v izbiro teh parametrov. Metodo integracije podatkov bi lahko izboljšali z uporabo transduktivnih tehnik strojnega učenja, ki bi poleg učne množice pravih in ne pravih parov upoštevala tudi lastnosti že vidnih neoznačenih parov.

References

M. Aizerman, E. Braverman, and L. Rozonoer (1964). Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control 25: 821-837, 1964.

I. Borg, and P. Groenen (1997). Modern Multidimensional Scaling: theory and applications. Springer-Verlag New York, 1997

AP. Bradley (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms - Pattern Recognition, 1997

J. Brank and J. Leskovec (2003). Download Estimation on KDD cup 2003, in KDD Cup 2003, eds., Johannes Gehrke, Paul Ginsparg, and Jon Kleinberg, 2003.

J. Brank, M. Grobelnik, N. Milič-Frayling, D. Mladenić (2002) Interaction of feature selection methods and linear classification models. In: Proceedings of the ICML-2002 Workshop on Text Learning, [Nineteenth International Conference on Machine Learning, 8-12 July 2002.

S. Brin, L. Page (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 1998

B., Caruana R., Gehrke J., Joachims T. (2005). Identifying Temporal Patterns and Key Players in Document Collections. Proceedings of the IEEE ICDM Workshop on Temporal Data Mining: Algorithms, Theory and Applications (TDM-05), 165–174, 2005.

F.J. Damerau (1964). A technique for computer detection and correction of spelling errors. Communications of the ACM, 1964.

J. Ferlez, B. Jörg, (2007). D8.1 Component for automated content acquisition and database integration, public IST World Deliverable 2005, available online at: http://ist-world.dfki.de/internal/Deliverables/wp8/deliverable8_1/

B. Fortuna (2004) String Kernels Proceedings of the 7th International multi-conference Information Society IS-2004, Ljubljana, Slovenia

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

B Fortuna, D. Mladinic and M. Grobelnik (2006). System for Semi-Automatic Ontology Construction. In Proc. of the 3rd European Semantic Web Symposium. http://www.eswc2006.org/demo-papers/FD18-Fortuna.pdf [last seen 14/09/2007]

G. H. Golub and C. Reinsch (1970). Singular value decomposition and least squares solutions, Numerische Mathematic, Volume 14, Number 5, April 1970

E. Grabczewski, B. Jörg, (2005). D3.2 Base set of Data, public IST World Deliverable 2005, available online at: http://ist-world.dfki.de/internal/Deliverables/wp3/deliverable3_

M. Grcar, D. Mladenic, M. Grobelnik, B. Fortuna, J. Brank, (2007). EU-IST Project IST-2004-026460 TAO D2.2 Ontology learning implementation, http://gate.ac.uk/tao/live/resources/publicdeliverables/d2-2.pdf [last seen 14/09/2007] 2007

M. Grobelnik and D. Mladenic (2003). Analysis of a database of research projects using text mining and link analysis. In Mladenic, D., Lavrac, N., Bohanec, M. and Moyle, S. (eds), Data Mining and Decision Support Integration and Collaboration, Kluwer, Dordrecht, 2003, pp. 157-166.

M. Grobelnik and D. Mladenic (2005A). Simple classification into large topic ontology of Web documents. In Proceedings: 27th International Conference on Information Technology Interfaces (ITI 2005), 20-24 June, Cavtat, Croatia.

M. Grobelnik and D. Mladenic, (2005B). Contexter - a system for visualization of large collections of novel stories. V: 29th Annual Conference of the German Classification Society, March 9-11, 2005, Magdeburg. From data and information analysis to knowledge engineering : program and abstracts. Magdeburg: Otto-von-Guericke-University, page 269, 2005

D. Grossman and O. Frieder (1998). Information Retrieval: Algorithms and Heuristics. 1998.

M. Hernandez, and S. Stolfo (1995). The Merge-Purge Problem for Large Databases. *Proceedings of ACM SIGMOD 1995*, 127-138.

P. Jacsó (2005). Google Scholar: the pros and the cons. Online Information Review, Volume 29, Issue 2, Page 208-214. Apr 2005

B. Jörg, J. Ferlež, E. Grabczewski, M. Jermol (2006). IST World: European RTD Information and Service Portal. 8th International Conference on Current Research Information Systems: Enabling Interaction and Quality: Beyond the Hanseatic League (CRIS 2006), Bergen, Norway, 11-13 May 2006

T. Kanungo, D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu (2002). An efficient k-means clustering algorithm: Analysis and implementation, IEEE Trans. Pattern Analysis and Machine Intelligence, 24 (2002), 881-892.

R. Kohavi, (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2 (12): 1137–1143. (Morgan Kaufmann, San Mateo)

C.S. Lambert (2006). Coordinating IST research and development across Europe: the CISTRANA outlook. Proc. 8th International Conference on Current Research Information Systems (CRIS 2006), Bergen, Norway, 11-13 May 2006

T. Landauer, P. W. Foltz and D. Laham (1998). Introduction to Latent Semantic Analysis". Discourse Processes 25: 259-284, 1998.

J. Leskovec, M. Grobelnik, and N. Milic-Frayling (2004). Learning Sub-structures of Document Semantic Graphs for Document Summarization , Conference or Workshop Item August 2004

J. Leskovec, J. Kleinberg, C. Faloutsos (2005). Graphs over time: densification laws, shrinking diameters and possible explanations, Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. Pages: 177-187, 2005

V. I. Levenshtein (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10 (1966):707–710.

H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, C. Watkins, (2002) Text classification using string kernels, The Journal of Machine Learning Research, Volume 2, Pages 419-444, March 2002

A. McCallum, K. Bellare and F. Pereira (2005), A Conditional Random Field for Discriminativelytrained Finite-state String Edit Distance, *UAI* 2005.

T. Mitchell (1982) . Generalization as search. Artificial Intelligence, 28:203–226, 1982..

D. Mladenic and M. Grobelnik, (2003). Text and Web Mining. In Mladenic, D., Lavrac, N., Bohanec, M. and Moyle, S. (eds), Data Mining and Decision Support Integration and Collaboration, Kluwer, Dordrecht, 2003.

D. Mladenic, and M. Grobelnik (2005). Visualizing very large graphs using clustering neighborhoods. In: Morik, K., Boulicaut, J-F, Siebes, A. (eds.). Local pattern detection: international seminar : Dagstuhl Castle, Germany, April 12-16, 2004 : revised selected papers, (Lecture notes in computer science, Lecture notes in artificial intelligence, 3539), (State-of-the-art survey). Berlin; Heidelberg; New York: Springer, cop. 2005, pp. 89-97.

H.B. Newcombe and J. M. Kennedy (1962) Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information. Communications of the Association for Computing Machinery, 5, 563-567, 1962.

H. B. Newcombe, (1988), Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business, Oxford: Oxford University Press. 1988.

H. B. Newcombe and M. E. Smith, (1975). Methods for Computer Linkage of Hospital Admission- Separation Records into Cumulative Health Histories. Methods of Information in Medicine, 14 (3), 118- 125, 1975.

B. Novak, D. Mladenić, M. Grobelnik, (2005) Text classification with active learning. 29th Annual Conference of the German Classification Society, 2005, Magdeburg, Germany.

C. Olston and H. E. Chi (2003) ScentTrails: Integrating browsing and searching on the Web. In ACM Transactions on Computer-Human Interaction (TOCHI), Volume 10, Issue 3, Pages: 177–197, ACM Press, New York, NY, USA. 2003

Pajntar B. (2006), Overview of algorithms for graph drawing, In proc. of Slovenian KDD Conference 2006. Oct. 2006

D. Rebholz-Schuhmann, H. Kirsch, F. Couto (2005) Facts from Text—Is Text Mining Ready to Deliver? PLoS Biol 3(2):

G. Salton and M. J. McGill (1983). Introduction to modern information retrieval. McGraw-Hill, 1983.

M. Seljak, (2001). IZUM-ova ponudba e-virov. U COBISS & SICRIS 2001, 13. in 14. 11. 2001. Maribor ([presentation at open debate])

H. Small (1999), Visualizing science by citation mapping, Journal of the American Society for Information Science, 1999

K. A. Spackman, (1989). "Signal detection theory: Valuable tools for evaluating inductive learning". Proceedings of the Sixth International Workshop on Machine Learning: 160–163, San Mateo, CA: Morgan Kaufman.

C. Thévignot (2000). The redesigned CORDISweb service contributes to the Commission's eEurope Initiative, Conference on European Research Information Systems CRIS2000, Helsinki 25-27 May 2000

S. Tong, D. Koller (2002). Support vector machine active learning with applications to text classification; The Journal of Machine Learning Research, Volume 2, Pages: 45-66. March 2002.

V. Vapnik (1995). The Nature of Statistical Learning Theory. Springer-Verlag, 1995.

V. Vapnik (1999). The Nature of Statistical Learning Theory. Springer-Verlag, 1999. ISBN 0-387-98780-0

H. Wallach. (2006). Topic modeling: beyond bagof-words. In Proceedings of the International Conferenceon Machine Learning (ICML).

E. Winkler (2006). Overview of Record Linkage and Current Research Directions. RESEARCH REPORT SERIES (Statistics #2006-2)

J. Zobel, A. Moffat and K. Ramamohanarao (1998). Inverted files versus signature files for text indexing. ACM Transactions on Database Systems (TODS), Volume 23, Issue December 1998, Pages: 453 - 490., 1998

Izjava o avtorstvu

Spodaj podpisani Jure Ferlež izjavljam, da sem avtor magistrske naloge z naslovom *Metode analize podatkov o raziskovalni dejavnosti na primeru aplikacije IST World*, oziroma angleškim naslovom *Methods for Analysis of Research Related Data in the IST-World Application*. Magistrsko nalogo sem izdelal samostojno pod mentorstvom akad. prof. dr. Ivana Bratka in somentorstvom doc. dr. Dunje Mladenič. Izkazano pomoč drugih sodelavcev sem v celoti navedel v zahvali.

Jure Ferlež, univ. dipl. ing. rač. in inf.