Thèse de doctorat

UNIVERSITE DE LIMOGES

ECOLE DOCTORALE SCIENCE - TECHNOLOGIE

FACULTE DES SCIENCES & TECHNIQUES

Laboratoire XLIM

Thèse N°

Thèse

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE LIMOGES

Discipline / Spécialité : Informatique

présentée et soutenue par

Anastasios TSOLAKIDIS

le [07 2014]

Systèmes d'aide à l'évaluation à base de visualisation interactive de graphes. Applications à l'évaluation des systèmes et des institutions éducatives

Thèse dirigée par le Professeur Georges MIAOULIS

et le Maître de Conférences HDR, Olivier TERRAZ

JURY :

Président du Jury

M. Djamchid GHAZANFARPOUR, Professeur, *XLIM*, Université de Limoges,

Rapporteurs

M. Yves DUTHEN, Professeur, *IRIT*, Université de Toulouse 1, Rapporteur

Mme Elli PETSA, Professeur, *Department of Surveying Engineering & Geoinformatics*, TEI d'Athènes, Rapporteur

Examinateurs

M. Djamchid GHAZANFARPOUR, Professeur, *XLIM*, Université de Limoges, Examinateur

M. Georgios MIAOULIS, Professeur, *Department of Informatics*, TEI d'Athènes, Examinateur

M. Olivier TERRAZ, Maître de Conférences HDR, *XLIM*, Université de Limoges, Examinateur

Université de Limoges

Acknowledgements

The present document constitutes my research work submitted within the context of my PhD studies at the University of Limoges. The research area is dealing with decision support systems using visual analytic methods in the area of higher educational institutes. I would like to thank everyone who supported this work.

In particular

I would like to thank Dr. Georgios Miaoulis, Professor of the department of Informatics at T.E.I. of Athens, for having supervised and supported my work.

I would like to thank Dr. Olivier Terraz, Professor of the department of Informatics at University of Limoges, for having supervised and supported my work.

I would like to thank Dr. Cleo Sgouropoulou, Associate Professor of the department of Informatics at T.E.I. of Athens, for having supervised and supported my work.

I would like to thank Dr. Ioannis Xydas, Lecturer of the department of Informatics at T.E.I. of Athens, for having supervised and supported my work.

Moreover I would like to thank my wife Anastasia who believed in me and supported me during all this period.

Finally this work is dedicated to my baby-son, wishing him that his innovative ideas will someday change the world.

Résumé

L'objectif de la thèse est d'améliorer les systèmes d'évaluation en utilisant des techniques d'analyse visuelle, des méthodes d'Extraction de Connaissances basés sur des graphiques dynamiques interactifs, analyse de réseaux et d'autres méthodes de représentation visuelle. Donc, nous présentons des visualisations interactives algorithmes afin de fournir des représentations qui aident les analystes à construire des modèles conceptuels précis et instructifs des réseaux de co-auteurs. Sur la base des interfaces visuelles interactives, nous fournissons à l'utilisateur d'extraction de connaissance à partir de la visualisation des données (KDD-V), fournissant ainsi que, l'utilisateur aura une assistance améliorée tout au long de la prise de décision (DM). En ce qui concerne les techniques d'Aide la Décision (AD), nous combinons les résultats de l'analyse des réseaux sociaux qui est appliquée sur le réseau de co-auteurs, avec des algorithmes de classification, afin de prévoir des liens futurs entre les auteurs. En outre, nous reconnaissons les équipes de recherche qui existent entre les auteurs en utilisant des algorithmes de classification et nous mesurons l'efficacité d'entre eux. Notre approche se concentre également sur l'analyse de l'efficacité des structures de collaboration dans le cadre de réseaux scientifiques au sein des institutions universitaires. Le principal domaine d'application des objectifs de recherche ci-dessus est l'élaboration d'un cadre de Gestion Institutionnelle de la Recherche (IREMA). IREMA est basé sur une architecture à quatre niveaux et peut être utilisé pour le développement de services liés à la gestion de la recherche et du développement des activités (R&D) dans les établissements d'enseignement supérieur. Dans notre prototype nous avons réussi l'intégration efficace des techniques de fouille de données, de visualisation et d'interaction homme-ordinateur. En d'autres termes, nous avons développé un Système d'Aide à la Décision qui combine ces domaines différents mais liés en vue de l'obtention de décisions efficaces sur de domaines spécifiques.

Abstract

The aim of the thesis is to enhance evaluation systems by using visual analytics techniques, knowledge discovery methods based on dynamic-interactive graphs, network analysis and visual representation methods. So, we promote interactive visualizations algorithms in order to provide representations that assist analysts in building accurate and informative conceptual models of the co-authoring networks. Based on the interactive visual interfaces, we provide to the user Knowledge Discovery from Data Visualization , thus providing that, the user gets enhanced assistance throughout the decision making (DM) process. Regarding the DM techniques, we combine the results of the social network analysis that are applied on the co-authoring network, with classification algorithms, in order to predict future links among the authors. Moreover, we extract the research teams that exist among the authors using classification algorithms and then, we measure the efficiency among them. Our approach also focuses on analyzing the effectiveness of collaborative structures in the context of scientific networks within academic institutions. The main application area of the above research objectives was the development of a framework for Institutional Research Management (IREMA). The IREMA is based on a four-layered architecture and can be used for the development of services related to the management of research and development (R&D) activities in Higher Education Institutions. In our prototype, we achieve effective integration of techniques from data mining, visualization and human-computer interaction. In other words, we develop a Decision Support System which combines these different but related domains in such a way that will provide effective decisions on specific domains.

Table of Contents

Acknowledgements 2
Résumé 3
Abstract 4
Table of Contents 5
List of Tables 7
List of Figures 8
Abbreviation Table 10
1. Introduction 12
1.1. Overview 12
1.2. Thesis objectives 13
1.3. The proposed methodology 13
1.4. Contribution areas 15
1.5. Thesis organization 16
2. Theoretical Background & State of the Art 17
2.1. Decision Support Systems using Visual Analytics 17
2.1.1. Decision Support Systems 17
2.1.2. Knowledge Discovery using Visual Analytics 18
2.1.3. Data mining 21
2.1.3.1. Classification 21
2.1.3.2. Clustering 23
2.1.3.3. Model-based methods 24
2.1.3.4. Association rules 25
2.1.4. The state of the Art 25
2.2. Visual Analysis by using Graphs 28
2.2.1. Basic Graph Definitions 28
2.2.2. Visual Representations by using Graphs 34
2.2.2.1. Node-Link Diagrams 35
2.2.2.2. Arc Diagrams 36
2.2.2.3. Adjacency Matrix 37
2.2.2.4. Circular Layouts 39
2.2.3. State of the Art 40
2.3. Research & Development Collaboration networks 43
2.3.1. Introduction 43
2.3.2. Co-authorship Networks 43
2.3.3. Efficiency Measure 44
2.3.4. State of the art 46
2.4. Research Information Management Systems for HEIs 48
3. Thesis Contribution to Decision Support Systems for Research Evaluation 51
3.1. Visual Analytics Systems 51
3.1.1. Motivation 51
3.1.2. Visual Analytic Process 52
3.1.3. Ontology for Research & Development Management 56
3.2. Data Visualisation using Graphs 69
3.2.1. Motivation 69
3.2.2. Methodology of ForceAtlas2 69
3.3. Research link Recommendation 74
3.3.1. Motivation 74
3.3.2. Research Link Functions 76
3.3.3. Research Link Algorithm 78

4. Institutional REsearch Management(IREMA) 82
4.1. System Implementation - Evaluation 82
4.1.1. Motivation 82
4.1.2. An overview of IREMA framework 84
4.1.2.1. Data Collection 85
4.1.2.2. Data Preparation 87
4.1.2.3. Data Mining 87
4.1.2.4. Interactive Knowledge Discovery 90
4.1.3. IREMA Use Cases 96
4.1.3.1. Identification of the key researchers and the most prominent research areas 96
4.1.3.2. Research collaborations & Research productivity. 103
5. Evaluation - Discussion & Conclusions 109
5.1. Evaluation 109
5.1.1. Evaluation Form 111
5.2. Discussion 113
5.3. Conclusion 115
References 116

List of Tables

Table 1: Force-directed layout algorithm 43
Table 2: Class Academic Degree 59
Table 3: Class Type 59
Table 4: Class Thesis Degree 59
Table 5: Class Country 60
Table 6: Class Equipment 60
Table 7: Class Event 61
Table 8: Class Conferences 61
Table 9: Class Interview 62
Table 10: Class Performance 63
Table 11: Class Research Area Topic 63
Table 12: Class Prototype Standard 63
Table 13: Class Course 63
Table 14: Class Presentation 64
Table 15: Class Presentation 64
Table 16: Class Patent 64
Table 17: Class Journal 65
Table 18: Class Award 65
Table 19: Class Competition 65
Table 20: Class Workshop 65
Table 21: Description Person 66
Table 22: Description Project 67
Table 23: Description Role 68
Table 24: Matrix of Graph G^1 78
Table 25: Matrix of Graph G^2 79
Table 26 : Paths with length equal to 2 from the node A1 to the other nodes. 79
Table 27: Calculation of scores for Graph G 79
Table 28: Bayesian for Social Science, Engineering and Computer Science 80
Table 29 : ResearchLink Algorithm 81
Table 30 : ResearchLink Functions 81
Table 31: Bayesian of Computer-Social Science 98
Table 32 : Association Rules. 98
Table 33: Definition of the variables included in the efficiency measure 104
Table 34: Efficiency Scores 105
Table 35: Collaboration Clusters 105
Table 36: Importance-Weights of Output Measures 106
Table 37: Efficiency Scores using Projects-Papers as outputs 107
Table 38: Evaluation of the system 112

List of Figures

Figure 1: Visual Analytic Process 19
Figure 2: The layers of Visual Analytic Process 19
Figure 3: The KDD process 20
Figure 4: Classification of Loans 22
Figure 5: Classification Model Construction 22
Figure 6: Classification Model Usage 23
Figure 7: Clustering of Objects 24
Figure 8: The Visual Analytic transformation $F : S \rightarrow I$, ... 26
Figure 9: Graph matrix 29
Figure 10: Co-authoring Network using Centrality Degree 30
Figure 11: Co-authoring Network using Betweeness Degree 31
Figure 12: Co-authoring network using Closeness Degree 32
Figure 13: Co-authoring Network using Louvain Method 34
Figure 14: Force-directed node-link diagrams 35
Figure15 : Arc diagrams 36
Figure 16: An arc covering angle q, with center C. 37
Figure 17: Adjacency Matrix 38
Figure 18: Circular layout 39
Figure 19: A and B are connected by the arc in bold 39
Figure 20: Force-Directed Power 41
Figure 21: Multi-disciplinary research areas 52
Figure 22: IREMA visual analytic process 53
Figure 23: Figure of Visualizations 55
Figure 24 : IREMA Ontology 58
Figure 25: Academic Degree Class 59
Figure 26: Event Class 61
Figure 27: Conference Class 62
Figure 28: Person Class 66
Figure 29: Project Class 67
Figure 30: Role Class 68
Figure 31: Layouts with Fruchterman-Rheingold 70
Figure 32: First stage of Logaritmic layout 72
Figure 33: First stage of Linear Layout 72
Figure 34: Second stage of Logarithmic layout 72
Figure 35: Second stage of Linear Layout 72
Figure 36: Third stage of Logarithmic layout 73
Figure 37: Third stage of Linear Layout 73
Figure 38: Graph with gravity=1 73
Figure 39: Graph with gravity=4 73
Figure 40: The interface for the modification of the graph layout 74
Figure 41: Graph Network 75
Figure 42: IREMA Architecture 84
Figure 43: IREMA Layers 85
Figure 44: Data Collection 86
Figure 45: Co-authoring Graph 91
Figure 46: Parallel coordinator for all the faculty members without any criteria 93
Figure 47: Show all the faculty members with h-index value besides 4-9 93
Figure 48: Map of science 94

Figure 49: Map of science with selected pie 94
Figure 50: Regression Line 95
Figure 51: Identification of the key researchers and the prominent research areas 97
Figure 52: Co-authoring Graph represents the most active researcher 99
Figure 53: Short Path Distance 100
Figure 54: The authors with the most research projects 101
Figure 55: The map of science 102
Figure 56: Research collaborations & Research productivity. 104
Figure 57: Research (Collaboration / Efficiency) 106
Figure 58: Efficient Line of Project-Paper 107
Figure 59: The score of Data Visualisations 113

Abbreviation Table

Higher Education Institutions	HEI
Decision Making	DM
Research Policy Maker	RPM
Institutional Research Management	IREMA
Decision Support Systems	DSS
Knowledge Discovery using Data Visualization	KDD-V
Research and Development	R&D
Human Computer Interaction	HCI
Knowledge Discovery from Databases	KDD
Science Citation Index	SCI
Data envelopment analysis	DEA
Decision Making Unit	DMU
Friend of a Friend	FOAF
Preferential Attachment	PA
Research Information Systems	RIS
Research Portfolio Online Reporting Tool	RePORT
National Institute of Health	NIH
On-Line Analytical Processing	OLAP
Research Management Information Systems	RIMS
Interactive Decision Support Systems	iDSS
Hellenic Quality Assurance Agency	HQAA
Exploratory Data Analysis	EDA
Left Hand Side	LHS
Right Hand Side	RHS
International Journal Articles	JAI
National Journal Articles	JAN
International Conference Papers	CPI
National Conference Papers	CPN
Citation indexes	CI
Book Chapters	BC
Research Project that one of the faculty members has the role of Coordinator	RPC
Research Project that one of the faculty members has the role of partner	RPP
Research Project with external institutes	RPE
Department of Electronic Engineering	EE
Department of Mathematics	MA
Department of Naval architecture	NA
Department of Physics, chemistry & Materials Techology	FY
Department of Land Surveying Technology	TOP
Department of Civil works and Infrastructure Technology	PEU
Department of Energy Technology Engineering	ET
Department of Informatics	CS
Department of Biomedical Engineering Technology	TIO

Evaluation of Visual Data Analysis and Reasoning	VDAR
Evaluation of User Performance	UP
Evaluation of User Experience	UE
Evaluation of Visualization Algorithms	VA

1. Introduction

1.1. Overview

In their effort to secure their place in the modern, knowledge-based society, Higher Education Institutions (HEIs) need to address the new challenges for producing quality outcomes in the context of networked, collaboration-intensive environments; they also strive to ensure social accountability. Within the current global setting, HEIs strive for effectiveness and innovation in research, which is a major factor towards their growth and excellence. Assessment of research outcomes on the other hand plays a key role in the enhancement of university-based research activity. Moreover, research administrators need to be increasingly involved in identifying efficient solutions that will improve the related capturing, analysis, decision making (DM) and strategic planning processes of universities. The more proficient decision makers are in managing their domain's knowledge, the more competitive their organizations can become. The use of technology can support research policy makers (RPM) in making informed, effective recommendations based on quality decision making, and in projecting the effect of these recommendations to the future, according to current and past performance. Towards this direction, the design and implementation of decision support systems requires the combination of multiple models and data mining techniques as well as the exploitation of innovative interactive representation tools that provide the user with enhanced assistance.

In this direction, we develop a framework for Institutional Research Management, a generic architecture for the design and implementation of interactive decision support systems (DSS) that enables Knowledge Discovery and supports Data Visualization using visual analytic processes.

The IREMA framework is based on interactive decision support methods applied within a Higher Education organizational context, where RPMs need to take important decisions for forming institutional research policy. Despite the fact that many techniques can be used to discover knowledge from data in order to formulate useful decisions, some of them are very difficult to utilize [1], since in most cases the end users are not familiar with either statistics or with the underlying data mining methods. These problems can be overcome by providing the user with different choices of interactive representations, which would considerably enhance the search behavior [2]. On this basis, IREMA aspires to offer a backbone for the development of strategic decision tools in the area of HEIs, enabling stakeholders to gain informed insight into the current situation and suggest future directions on research activities. Using interactive visualization, the framework supports visual analytics for the decision making and query execution through semantic extraction and data fusion technologies

1.2. Thesis objectives

The aim of the present thesis is to enhance evaluation systems by using visual analytics techniques, knowledge discovery methods based on dynamic-interactive graphs, network analysis and other visual representation methods. In order to achieve this, we have investigated and integrated several visual analytics techniques and tools, namely: graphs, parallel coordinators, map of science and regression lines; these all have been merged together in order to help analysts browse and explore them, exploring all the facts and information contained therein. A layout algorithm is also proposed, which: i) allows users to make interactive data exploration and ii) is able to display graphs which contain more than 1000 nodes in a way that will be easily understood. Furthermore, the interactive visualizations provide representations that assist analysts in building accurate and informative conceptual models of the co-authoring networks. Based on the interactive visual interfaces, we assist the user through the decision making (DM) process by providing capabilities for Knowledge Discovery from Data Visualization (KDD-V). Regarding the DM techniques, we combine the results of the social network analysis that are applied on the co-authoring network, with link prediction algorithms, in order to predict future links among the authors. We extract using classification algorithms the research teams that can be formed among the authors and then we measure the efficiency among them. Our approach also focuses on analyzing the effectiveness of collaborative structures in the context of scientific networks within academic institutions. The aforementioned research objectives have been materialized in the development of a framework for Institutional Research Management (IREMA). The proposed approach is based on a four-layered architecture (data collection; data preparation; data mining; knowledge discovery) and can be used for the development of services related to the management of research and development (R&D) activities in Higher Education Institutions. Our research, in conclusion aims to integrate data mining, visualization and human-computer interaction techniques by mean of interactive visual representations. In other words, we intend to develop a Decision Support System that will combine these different but related domains in such a way that will provide effective decisions on specific research areas.

1.3. The proposed methodology

The investigation, design and implementation of a DSS based on the Institutional Research Management (IREMA) framework could be said to be the main result of our 3-years research. IREMA tool is a web based system built on Java technologies that supports institutional research management.

In our research approach we used graph analysis with data envelopment analysis (a method for efficiency measurement), which have been combined with data mining techniques as a mean for knowledge extraction. Comparing the R&D outcomes of academic units with the dynamics of the collaboration

patterns extracted from graphs, the developed framework (and the related tool) enables users to evaluate specific criteria and correlate strategic goals with research performance. The aim of our approach is to investigate a framework for the research policy makers that could support them to combine efficiency measurement techniques with data mining methods, in order to take the best decision among a variety of possible alternatives. The process starts by defining the problem and the objectives, proceeds with the visualization of multiple figures-graphs and finishes with the selection of the decision.

The levels of the process are the following:

Data collection and Data cleansing. We use data exported from the Scopus Scientific library in order to get the research publications.

Data transformation. During that process, a graph is created among the academic researchers, based on the co-authoring of research papers. We use five (5) graph measures, namely: Degree centrality, Closeness centrality, Betweenness centrality, Eigenvector and Clustering co-efficient.

Data Mining (DM) is used in order to extract hidden predictive information. The DM method falls into the categories of clustering, classification, and association analysis. Also we propose the Research-Collaboration Algorithm in order to predict future collaborations among the authors.

Knowledge discovery form data visualizations. The proposed framework integrates interactive visual interfaces to support Knowledge Discovery (KD), thus providing the user with enhanced assistance throughout the decision making (DM) process. The proposed framework supports the following visual representation techniques:

Co-authoring Graph, this is created on the basis of the collaboration among faculty members for the publication of a research paper. Furthermore, we have developed a force-directed layout algorithm.

Efficiency Line, which is used to represent the correlation among the indicators.

Parallel Coordinators, as an interactive representation where the user is able to apply a set of criteria (dynamic) depending on his/her objectives.

Map of Science, where each of the research areas is represented in pie charts.

Through the aforementioned process several questions could be evaluated, as the following ones:

- How efficient would be for a Department or Institution to focus on specific areas, and who would be the lead researcher(s) in that case?

- How research collaborations influence R&D productivity?

Based on the evaluation results, the IREMA framework (and the related tool) seems to be satisfactory and helpful.

1.4. Contribution areas

The present thesis focuses on the following research areas:

- Graph layout algorithms and interactive visualization
- Visual analytics and Decision support environment
- Knowledge Discovery
- Evaluation of Research & Development (R&D) activities

In the following section, we will elaborate how we approach each one of the before mentioned research areas.

Graph layout algorithms and interactive visualization

We mainly focused on the graph layout algorithms, as well as on data representation based on dynamic interactive graphs. For the selection of the layout we developed a force-directed algorithm to help users select the most desirable representation. By using the algorithm, the layout is created by applying an energy model on the nodes. The energy model consists of i) the attraction force and ii) the repulsion force.

Visual analytics and decision support environment

In order to enhance visual capabilities of our tool, our research focused on the study of a decision support environment that could be used for the evaluation of the faculty members' research activities of Higher Education Institutes (HEI). The architecture of such a system was also investigated. Additionally, our research focused on aspects related to the Human Computer Interaction (HCI) and visual analytics incorporated in such an environment. Eventually, a web-based system was developed and evaluated by decision and policy makers.

Knowledge Discovery

The process of knowledge discovery from databases (KDD) was enriched by a variety of data mining (DM) methods and the results were displayed using interactive visual representations. The DM methods are used for the clustering, classification, and association analysis of the evaluation data.

Evaluation of Research & Development (R&D) activities

Several studies have attempted to evaluate the R&D activities and quantify the collaborations, giving mainly emphasis to: bibliographic metrics; graph network analysis; qualitative methods and surveys. In our methodology we use graph network analysis in order to evaluate and study the patterns of research collaboration among the faculty members of an institution, using the co-authoring information of research papers as a parameter. By employing the co-authoring networks, we could calculate the metrics of the authors and then make important decisions concerning the performance of the authors. Data Envelopment Analysis (DEA) is used to measure the efficiency of the academic units or the research teams of a HEI.

1.5. Thesis organization

The thesis is organized in five sections. The first section provides a brief overview of the thesis objectives , the proposed methodology and the contributing areas. In the second section we analyze the theoretical background of all the disciplines that have been used in our research. At the third section we present our contribution areas and at the fourth we present the developed prototype. Finally at the fifth section we discuss the findings and propose research directions for future work.

In more detail, the second section explores the relevant research literature about the Research Information Management Systems in HEIs; it also presents the principles behind decision support systems as well as the different approaches in knowledge discovery. It also describes how KDD can be implemented using graphs; the basic graph layout algorithms are also being presented. Finally, collaboration networks are being presented; a discussion is provided on how these can be created based on the R&D activities and how the efficiency among these groups can be measured.

In the third section of the thesis we present our contribution in the areas of visual analytics systems, data visualizations using graphs and link recommendation algorithms.

In the forth section we present in detail our prototype and we provide a case study for the system’s validation, based on three research questions.

In the final section we present the results of our system's evaluation and we conclude with a summary and an outlook on further research.

2. Theoretical Background & State of the Art

Visual analytics is much more than interactive visual representations, as they integrate users' knowledge into data analysis processes. Therefore visual analytic systems support decision making by combing visualizations with human computer interaction and data analysis. In this section, we will discuss about the theoretical background and the state of art of all the different disciplines which are involved in our thesis. Initially, we will discuss about decision support systems by using visual analytics and how we can apply analysis tasks using graphs. Then we will discuss about the different kind of graph representations. In the next subsection, we will see the co-authoring networks as they could be used in order to validate our system. This section concludes with the presentation of the research information system as our prototype aims to offer new potentials in developing enhanced research information systems.

2.1. Decision Support Systems using Visual Analytics

2.1.1. Decision Support Systems

A Decision Support System (DSS) is an information system that analyzes business or organizational data and presents it so that users or experts can take decisions. DSS are designed to support decision-makers at any level in an organization for the financial management and strategic decision-making. A properly designed DSS [3] is an interactive system, which supports decision makers to manipulate quantitative data from a variety of different data sources, in combination with the knowledge of an expert or the related business models in order to support decision making activities. A proposed by Daniel Power taxonomy for a DSS [4] is the following:

A model-driven DSS uses complex models, for example statistical or financial models, to provide decision support. At those systems the decision makers are participating by providing data through the decision process.

A communication-driven DSS is developed to support group of people, which work on a shared task, using network and communication technologies.

A data-driven DSS is developed to support and manipulate time series data by accessing file systems using retrieval tools.

A document-driven DSS provides document retrieval and analysis by using unstructured information as for example catalogs, and corporate historical documents.

A knowledge-driven DSS provides decision making activities using specialized problem-solving applications. These systems are knowledgeable in a particular domain and using data mining techniques try to solve problems.

In our system we will use a Knowledge Driven DSS which is "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data"[5].

2.1.2. Knowledge Discovery using Visual Analytics

In a KDD decision support system, the user initially searches for the most appropriate information, in order to describe and identify the problem [6][7]; then s/he designs the process by developing and analyzing the actions in order to lead in important decisions; and finally, she/he selects among a variety of solutions, the one that best describes the specific problem. This process, as Beynon et al.[9] have shown, may cause problems the most of the times, because the users usually do not know which steps to follow in order to reach a solution; and they often lack the ability to identify when the optimal solution has been reached. A solution to this problem could be the development of an interactive DSS [1][10] which supports human–computer interaction and allows users to identify, explore and solve problems. As Fisher [11] indicates, the success of a DSS depends on its interface for human–computer interaction and not on its capabilities to solve a problem. So the main problem of all those systems is the data transformation to decision where visual analytics seems to be an effective solution.

Visual analytics process combines various related research areas such as visualisation, data mining, data management, data fusion and statistics. As Keim D A et all [12] said: "Visual analytics combines automated analysis techniques with interactive visualisations for an effective understanding, reasoning and decision making on the basis of very large and complex datasets". In figure 1 we can see an overview of the stages of visual analytic process.

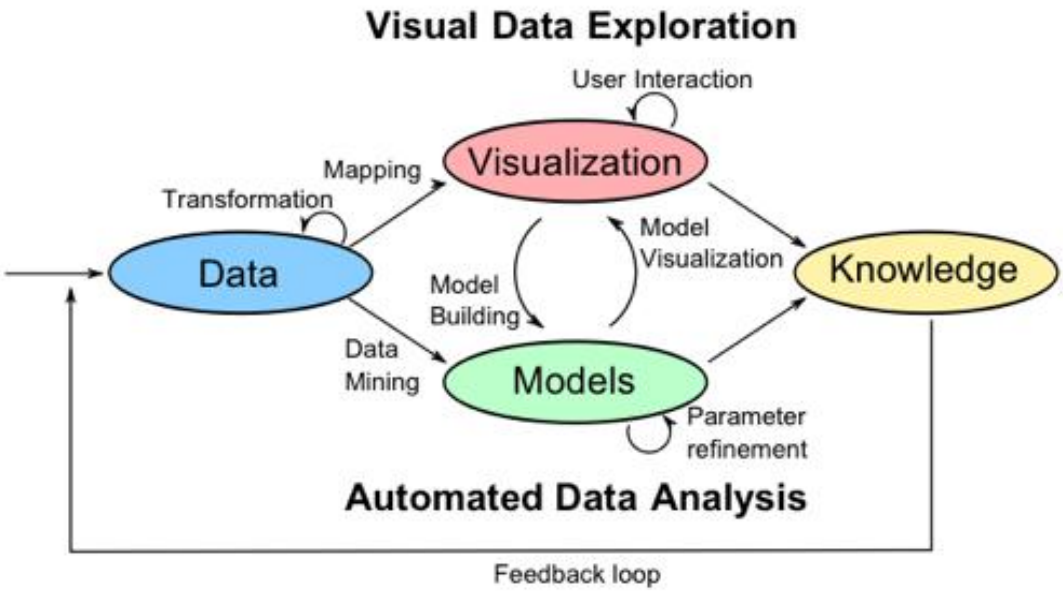


Figure 1: Visual Analytic Process

So the process is based on the interaction among the different stages as the models, the data, the knowledge and the visualization. Also as shown in figure 2, the visualization is the center of the process, as well as the target of all the other processes.

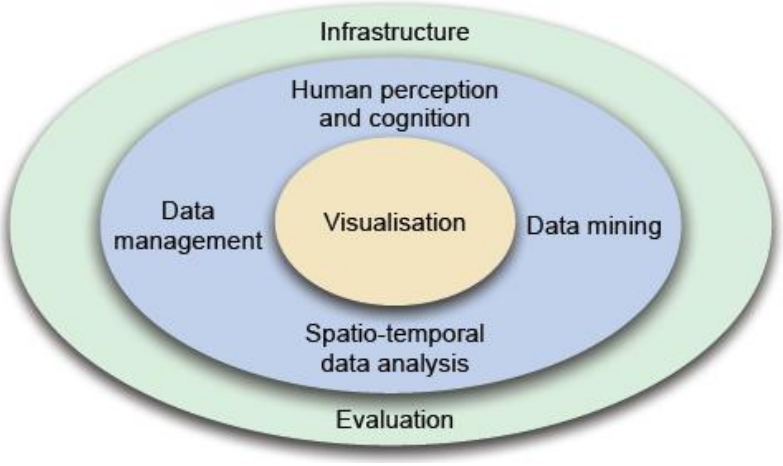


Figure 2: The layers of Visual Analytic Process

Source: [12]

Moreover, in an organization or institute, a decision maker has to deal with rapidly growing volumes of digital data, which should be mined in order to extract knowledge. By using Data Mining techniques, the user is able to extract hidden predictive information from large databases (data warehouses) or to identify hidden patterns, which could not be easily observed. When the process of knowledge discovery enriched by visual analytics, we can get reasoning processes using interactive visualization based on queries through data fusion technologies.

KDD systems use particular statistical procedure or data mining algorithms for evaluating hypotheses. KDD approaches and methods operate best on large data sets with rich data structures. Due to the process of KDD, we construct the appropriate model that will be used for the extraction of the decisions.

The KDD process (figure 3) is interactive and iterative so the user could take several decisions due to the process of KDD. The major steps are the following.

1. Understanding the domain knowledge and identifying the goal of KDD.
2. Selection of the main data set for the purpose of the KDD.
3. Data Cleaning and Data transformation (Data Processing).

   The data transformations may include the reducing of the number of features, discretization of continuous numeric values, the replacement of missing values, etc. Data cleaning helps in the level of confidence in data analysis.
4. Data mining.

   At this step, the user defines the objectives of the KDD process by applying a data-mining algorithm such as clustering, classification, regression, in ways which match with the original goal of the KDD process.
5. Interpretation of the result which should be supported by visual representations.

   The KDD process may contain many interaction and feedback loops between each step in this process.

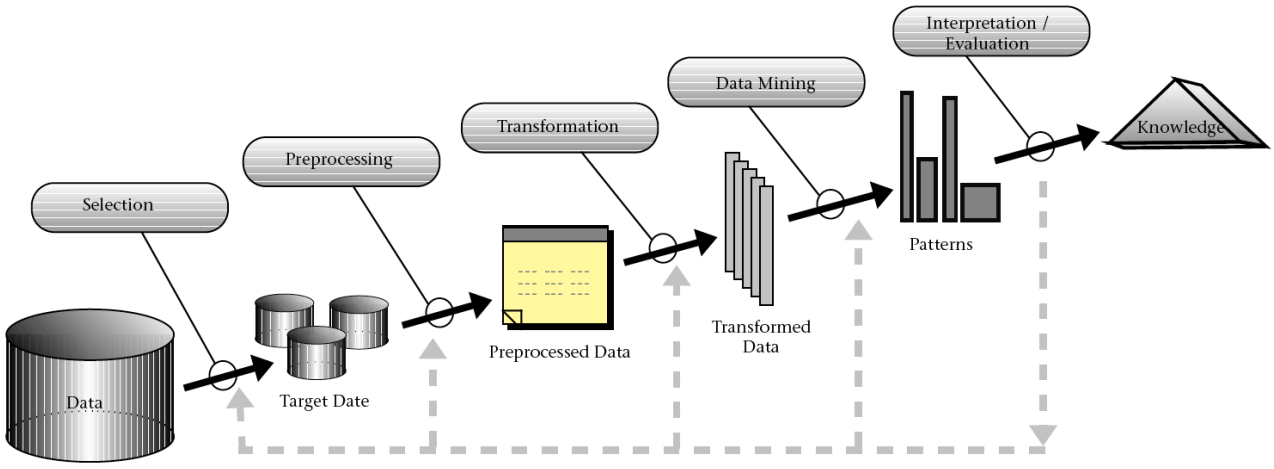


Figure 3: The KDD process

Source: [5]

2.1.3. Data mining

The data mining process is a part of the Knowledge Discovery process in a DSS, which is described as the overall process of discovering useful knowledge from data. The main objectives of data mining process are the prediction, description and the extraction of hidden patterns of information, from large volumes of data. There are various techniques which categorized as i) supervised, where, based on a training set, deterministic or probabilistic algorithms are used to create the classification model which will be used for the classification of new data (decision trees, support vector machines and neural networks, ..) and ii) unsupervised, which attempt to create a model , without previous knowledge for the domain (cluster analysis).

Some of the most well-known algorithms are the following:

1. Classification
2. Clustering
3. Association Rules
4. Sequential Patterns
5. Regression Models
6. Decision Trees

In the next sections we will discuss in details only about the techniques that will be used in our system.

2.1.3.1. Classification

The process of classification is used to predict group membership for data instances by applying supervised learning methods. The classification techniques use a training set where all the instances are assigned on specific classes in order to build the classification model. Then a new entered instance is classified at the corresponding class by using the constructed model. For example, you may wish to use classification to predict whether the weather on a particular day will be “sunny”, “rainy” or “cloudy”. In another example in figure 4 we classify when a bank offer a loan or not in association with the attributes of debt and income. So for a new customer the bank could easily classify him/her in the corresponding group and then decide whether the customer or not is allowed to get a loan.

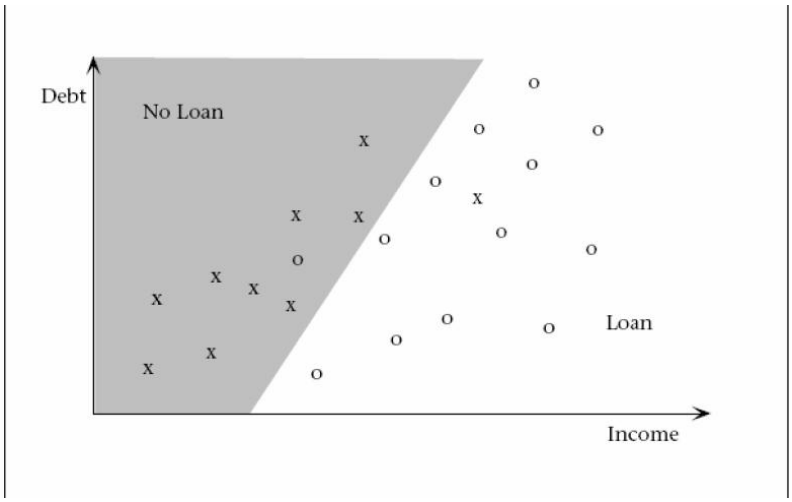
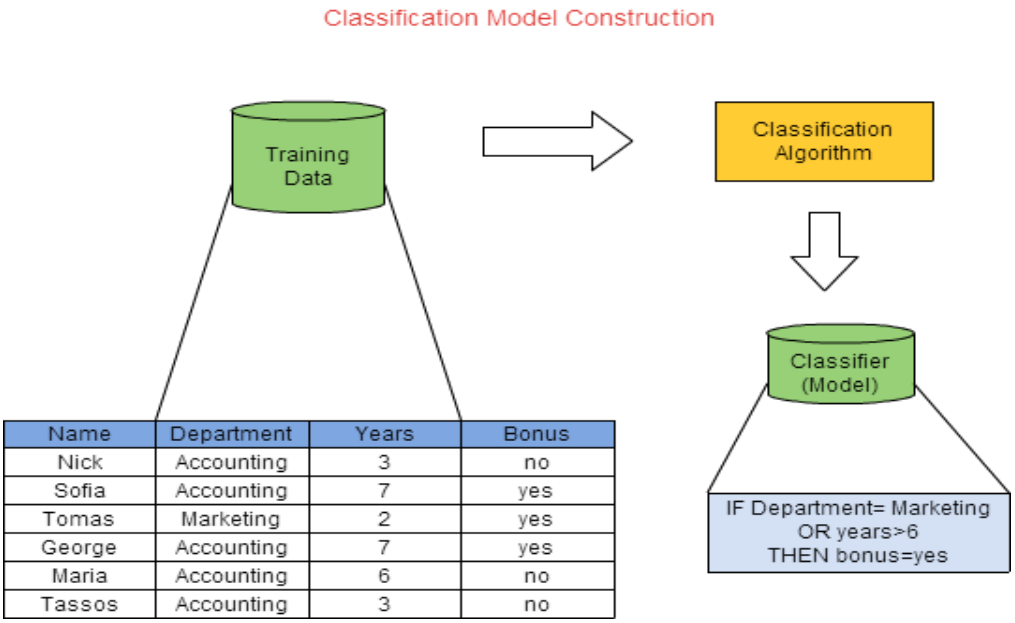


Figure 4: Classification of Loans

Classification is a two-step Process

- Model construction (figure 5). Given a set of data representing examples of a target concept (training data), build a model to "explain" the concept.
- Model usage (figure 6). The classification model is used for classifying future or unknown cases, estimate accuracy of the model.



Name	Department	Years	Bonus
Nick	Accounting	3	no
Sofia	Accounting	7	yes
Tomas	Marketing	2	yes
George	Accounting	7	yes
Maria	Accounting	6	no
Tassos	Accounting	3	no

Figure 5: Classification Model Construction

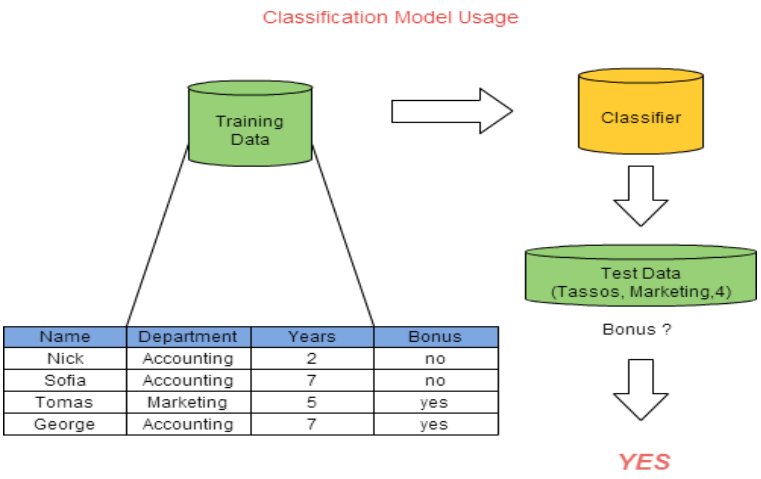


Figure 6: Classification Model Usage

In order to evaluate a classification method we use the criteria of i) Accuracy where we measure the classifier accuracy and the predictor accuracy, ii)the Speed where we measure the time to construct the model (training time) and the time to use the model (classification/prediction time), iii) the Robustness where we measure the handling noise and missing values and iv) the Scalability, where we measure the efficiency in disk-resident databases.

So the main task is the construction of the appropriate classification model that could be used in order to taxonomy the instances into predefined classes. For the classification of the data several techniques exist. The most well known methods are the Decision Trees, the Bayes Networks, the classification of the Nearest Neighbor and the Neural Networks.

2.1.3.2. Clustering

The clustering technique is used to place data elements into related groups (finding similarities between data according to their characteristics) without advanced knowledge of the group definitions. A cluster is a collection of data objects, similar one to another within the same cluster, and dissimilar to the objects in other clusters. For example, in figure 7 we observe 4 clusters. The Cluster analysis method is an unsupervised learning method since there is no predefined class, and in most of the times could be used as the Preprocessing step for other algorithms.

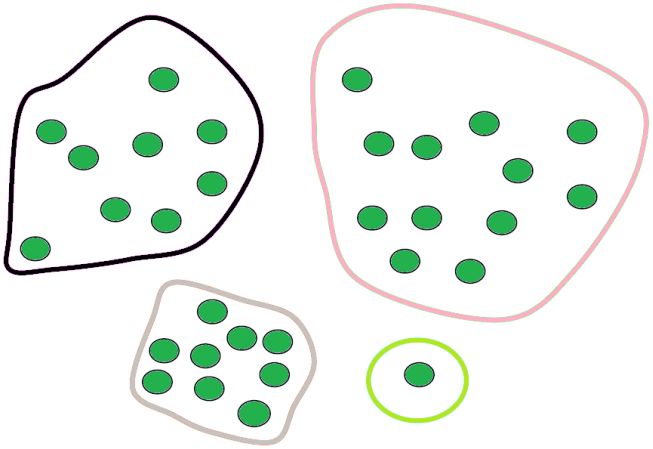


Figure 7: Clustering of Objects

The quality of a clustering result depends on both

- the similarity measure used by the method which is expressed in terms of a distance function, typically metric: d(i, j) and
- its ability to discover some or all of the hidden patterns

The Similarity and Dissimilarity are used to measure the similarity or dissimilarity between two data objects. Some popular methods are the Minkowski, Manhattan and the Euclidean distance.

Two of the major clustering approaches that we will are also going to use in our approach are:

1. Partitioning methods: Construct a partition of a data set containing n objects into a set of k clusters, so to minimize a criterion (e.g., sum of squared distance). The main goal is, for a given k (number of clusters), to find a partition of k that optimizes the chosen partitioning criterion. Typical methods include k-means, k-medoids,..
2. Hierarchical methods: This method is separated into two categories,
   - Agglomerative. Each one of the instances belongs to a cluster. At each step, the closest pair of clusters is merged until only one cluster (or k clusters) left.
   - Divisive. Start with one cluster. At each step, the initial clusters are separated until each cluster contains a point (or there are k clusters).

2.1.3.3. Model-based methods

This method attempts to optimize the fit between the data and some mathematical model. The models that can be used are: i) Statistical, ii) Machine learning and iii) Neural networks.

2.1.3.4. Association rules

The technique of the association rules attracts great attendance because it provides an efficient way to discover frequent patterns, associations and correlations among sets of items or objects in databases, and other information repositories. An association rule is an implication of the form X=>Y, where X and Y are item sets.

In order to evaluate a rule we use the metrics of:

- Support. Fraction of transaction that contain both X and Y.
- Confidence. Measure how often items in Y appear in transactions that contains X.

2.1.4. The state of the Art

In the area of decision making a variety of data mining and statistical methods exist, which produce raw or image based results (charts or pies) without any interaction. In 1977 John W. Tukey [13] introduced the exploratory data analysis, which was the devolution from the confirmatory data analysis to the interactive data analysis.

This first approach was followed by Chen[14] and Spence[15] who enhanced the graphical interface in order to support the knowledge discovery by effective and efficient visualizations. The data mining processes encompassed with interactive representations led to the visual data mining [16]. The integration of both data mining and data visualization create the technique which is called visual analytics. The visual analytics was first discussed in the research and development agenda, Illuminating the Path [17] and it is used in order to describe a multidisciplinary field that combines visualization, human computer interaction, data analysis, data management and statistics.

Our research in the area of visual analytics focuses on the key patterns that should be followed in order to design such a system. The main characteristic of these systems is the collaboration among the human and the computers [18]. Those systems are used in the knowledge discovery and in artificial intelligent for problem solving tasks, in order to solve complex problems by using the knowledge of an expert in combination with the computation capabilities of a computer to solve or execute operations.

In the survey of Bertini and Lalanne [19], three categories of visual analytic systems are being suggested:

1. The systems that express all the results of the computations and data analysis process by using visualizations. So the data visualizations are the primary goal of those systems.

2. The systems where the data analysis is accomplished by data mining and visualizations using interactive interfaces in order to permit the user to modify or to interpret the results. So at those systems the data analysis is supported by interactive interfaces for data mining and data visualization.

3. The systems where the users and the computers are collaborating. At these systems two sub-categories exist:

- The white-box systems where the user and the computer cooperate for the implementation of the model, and
- The black-box ones, where the user gets the model ready and he is only allowed to modify parameters in order to get the corresponding visualizations.

Shneiderman et al.[20] promote the steps of overview, zoom/filter and details for analytical reasoning using visual analytics. In 2006 Keim *et al* [21] extended the previous approach and propose the following visual analytic process which consists of the following steps:

- Analyze,
- Show the important ones,
- Zoom/filter,
- Analyze further,
- Details on demands.

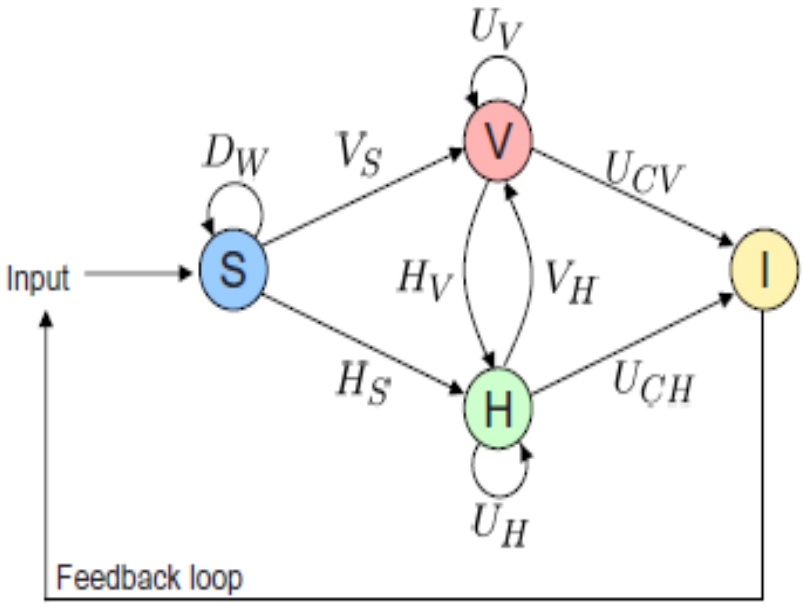


Figure 8: The Visual Analytic transformation $F : S \rightarrow I$,

The visual analytic process is described in figure 8 where the visual analytics process is a transformation $F : S \rightarrow I$, whereas F is a concatenation of functions $f \in \{D_W, V_X, H_Y, U_Z\}$ defined as follows:

D_W describes the basic data pre-processing functionality.

$V_X, X \in \{S, H\}$ symbolizes the visualization functions, which are either functions visualizing data $V_S : S \rightarrow V$ or functions visualizing hypotheses $V_H : H \rightarrow V$.

$H_Y, Y \in \{S, V\}$ represents the hypothesis generation process. Two type of functions exist: these that generate hypotheses from data $H_S : S \rightarrow H$ and functions that generate hypotheses from visualizations $H_V : V \rightarrow H$.

User interactions can either effect only visualizations $U_V : V \rightarrow V$ (i.e., selecting or zooming), or can effect only hypotheses $U_H : H \rightarrow H$ by generating a new hypotheses from given ones. Furthermore, insight can be concluded from visualizations $U_{CV} : V \rightarrow I$ or from hypothesis $U_{CH} : H \rightarrow I$

Also in 2010 Keim *et al.*[22]enhance the visual analytics process, and after their modifications the steps are the following: i) Transformation of the data, ii) Adjustment of visual or data mining methods and iii) Visual data exploration in order to analyze and explore the data.

Bertini et al.[23] proposed the Quality-Metrics-Driven Automation, which automates the numerical/algorithmic data analysis.

Munzner et al.[24] separated the visual analytic process into 4 layers: domain problem characterization, data/operation abstraction design, encoding/interaction technique design, and algorithm design.

Sedlmair et al.[25] proposed a process which consists of nine layers: learn, winnow ,cast, discover, design, implement, deploy, reflect, and write.

In our approach, we have extended the visual analytic process by introducing the use of ontologies, which contain all the knowledge for a specific domain. Moreover, regarding the visual representations we chose to use graph networks as our main visualization tool, in order to support the visual analytic process. Therefore, in the next section we will discuss about visual analysis using graphs.

2.2. Visual Analysis by using Graphs

One of the most serious problems when the user deals with large and complex data, is how those data will be represented. Comprehensive surveys of techniques for data visualization [26] conclude to the result that graph networks are the most representative way for a wide variety of situations and many data sets are most naturally interpreted and depicted as networks. As network visualization seems to be the most representative tool in order to represent complex data sets, several software packages such as: Tulip [27], Graphviz [28], Gephi [29], Pajek[30], and Cytoscape [31], exist. The data exploration using graphs often varies on the development of effective graph layouts. The visualization of large graphs is supported by interactive techniques in such a way, that analytical tasks could be applied. Therefore, in order to support Visual Analytics processes, the representations should follow the terms of [32]:

- Visual representation,
- User interaction, and
- Algorithmic analysis.

In the following subsection we will discuss about some basic graph definitions and then we will see in details the different visual representations by using graphs.

2.2.1. Basic Graph Definitions

In computer science, the area of graph theory is related with the study of graphs and the structures that are created in order to build the associations among two objects (nodes) of the collection. So, a graph (figure 9) consists of nodes and the edges that connect them. A graph could be directed or undirected. A graph network is used in order to describe communication networks, social networks and every other network that there exists a relationship among the users.

The definitions of the Graphs are the following:

• A graph is represented as G = (V, E) where V is the nodes and E, the edges.

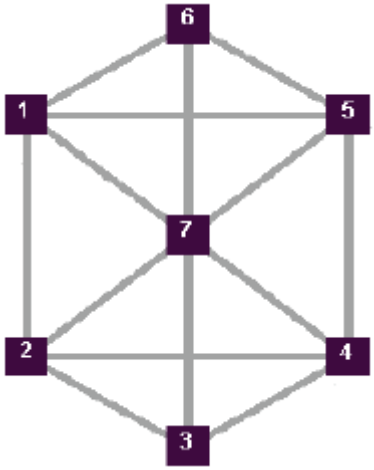


Figure 9: Graph matrix

Based on the connectivity among the nodes and the network topology, a lot of algorithms have been developed in order to measure the importance of a node. Such measures are called graph metrics. Indicatively, we can distinguish:

The **Clustering coefficient** [33] is a measure of the degree to which nodes in a graph tend to cluster together. It shows how well connected are the neighborhood of the node. If the clustering coefficient is 1, then the neighborhood is fully connected, otherwise there are no connections in the neighborhood.

The density of clique-like triangles is measured by calculating the clustering coefficient of the network.

$$C_i = \frac{\text{number of triangles connected to node}_i}{\text{number of triples centered on node}_i}$$

Centrality Degree measures the number of lines incident to a node. Authors with high degree centrality are those who have the most collaboration. By using this measure we could identify the most active researchers (figure 10).

The Centrality Degree is defined as follows:

$$C_D(n_i) = d(n_i),$$

Where $d(n_i)$ is the degree of n_i.

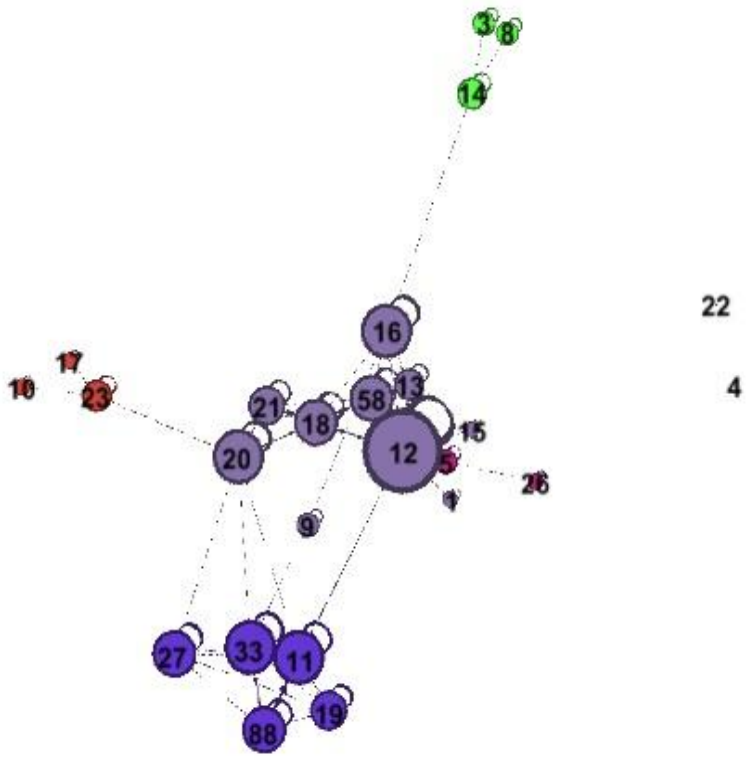


Figure 10: Co-authoring Network using Centrality Degree

Betweenness Degree measures the ability of a node to connect nodes that do not have have any other direct connection (edge) [34]. These nodes are called hubs, because they have the capacity to transfer information from one researcher to another. In Figure 11, we can observe that the author with id=12 has the higher value of betwenness.

The betweenness centrality of a node *u* is given by the expression:

$$g(u) = \sum_{s \neq u \neq t} \frac{\sigma_{st}(u)}{\sigma_{st}},$$

where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(u)$ is the number of those paths that pass through u.

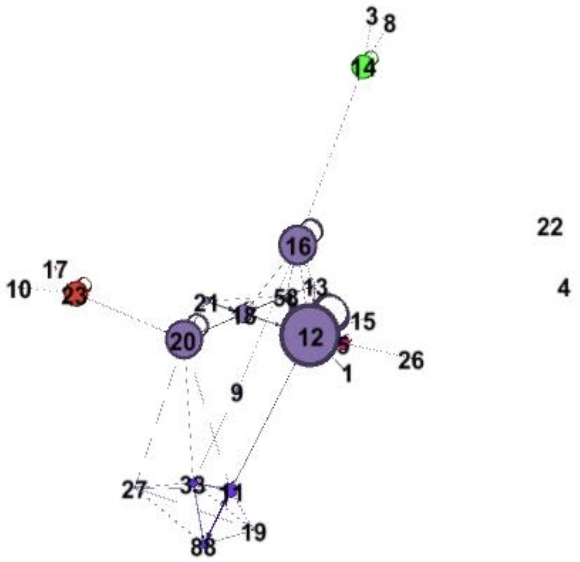


Figure 11: Co-authoring Network using Betweeness Degree

Closeness centralization is based on the total distance between one node and all other nodes. An author is considered with high closeness centrality if he has many, short connections to other authors in the network [35]. For example, in figure 12, authors 18 and 12 have the highest closeness values, meaning that they tend to collaborate easier as they have the shortest paths to the other nodes.

The Closeness Centrality $C_c(n_i)$ is given by the expression:

$$C_C(n_i) = \sum_{i=1}^{n} \frac{1}{d(n_i, n_j)},$$

where $d(n_i,n_j)$ is the distance between two vertices in the network.

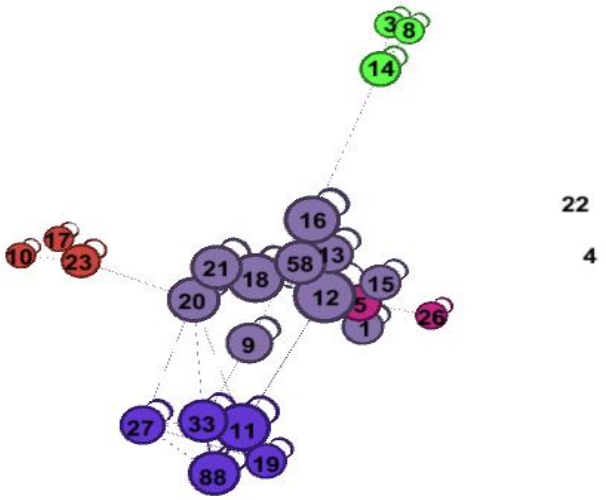


Figure 12: Co-authoring network using Closeness Degree

The **Eigenvector Centrality [**36**]** is a measure of the importance of a node in a network. It assigns relative scores to all the nodes in the network, based on the principle that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes.

For graph G where

$$G := (V, E)$$

with |V| number of vertices and |E| number of edges, let $A=(\alpha_{u,t})$ be the adjacency matrix, i.e. $\alpha_{u,t}=1$ if vertex u is linked to vertex t, and $\alpha_{u,t}=0$ otherwise.

The centrality score of vertex u can be defined as:

$$x_u = \frac{1}{\lambda} \sum_{t \in M(u)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{u,t} x_t$$

where M(u) is a set of the neighbors of u and λ is a constant. With a small rearrangement this can be rewritten in vector notation as the eigenvector equation

Communities Detection algorithms are used to discover clusters of nodes, which has strong connections to each other. The Louvain method is a Multi-Level Aggregation Method for optimizing modularity [33] which could be used for the identification of communities. The method consists of two phases. Initially, it looks for "small" communities by optimizing modularity in the network and then it

builds a new network, the nodes of which represent the communities. These steps are repeated iteratively until a maximum of modularity is attained.

The modularity of a partition is a scalar value [-1, 1] that measures the density of links inside communities as compared to links between communities. In the case of our co-authorship network, having weights on the links such as the number of collaborations between and among the authors, the modularity is defined as:

$$Q = \frac{1}{2m}\sum_{i,j}[A_{ij} - \frac{k_i k_j}{2m}]\delta(c_i, c_j)$$,

where A_{ij} represents the weight of the edge between i and j,

k_i is the sum of the weights of the edges attached to vertex i,

$$k_i = \sum_j A_{ij}$$

c_i is the community to which vertex i is assigned,

$\delta(u, v)$ is 1 if u = v and 0 otherwise, and

$$m = \frac{1}{2}\sum_{ij} A_{ij}$$.

Figure 13 depicts the co-authorship network using the Louvain community detection algorithm which deconstructs the network into six structural communities. The algorithm assigns a membership value to each of these communities (nodes). This value identifies the degree of collaborations, indicated by a unique color. The diameter of the nodes represents the number of publications, where authors with more publications have larger diameters.

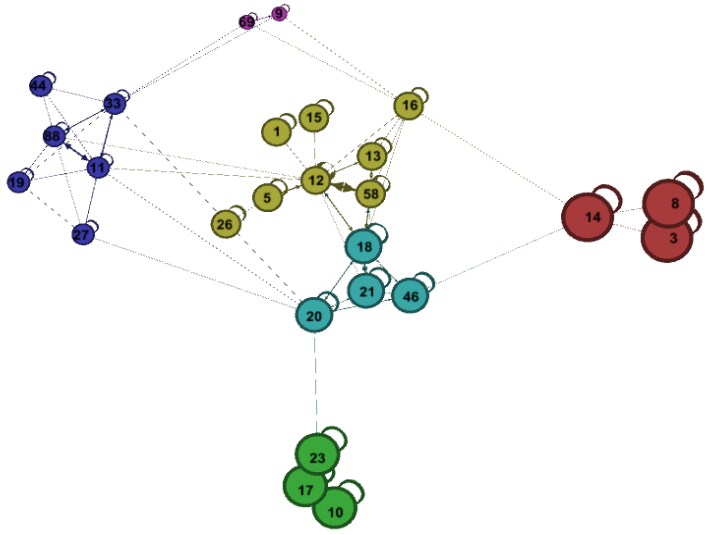


Figure 13: Co-authoring Network using Louvain Method

Educational institutions increasingly need to assess and enhance their activities, in order to provide a balance of tangible and intangible assets, and to measure future capability as well as past performance. Graph analysis can be effectively used on analysing research outputs in universities for the identification of research communities, the most active researchers, the "research hubs" and also the strength of the co-authorship network which reflects the ability for scientific progress. The analysis of a department's research articles provides a detailed insight to the relationships among its faculty members. The structural analysis of the co-authoring network illustrates the research relationship of the scientists, using different graph metrics.

2.2.2. Visual Representations by using Graphs

The most efficient way in order to explore and analyse the graphs, is by using of visualizations. The main parameters that a user should take into consideration in order to design data visualization are:

- The type of visual representations (e.g. matrix or node-link diagrams) is going to use,
- Where on the screen the graph elements should be placed
- The way to map the calculated values on the visual attributes

In the following subsections we will discuss about the layout algorithms, which are separated to those that use matrices in order to represent the elements of the graphs and non-matrix, and morever the way that each one of them represent the elements. More especially we will see the node-link diagrams, arc diagrams, adjacency matrices, and circular layouts.

2.2.2.1. Node-Link Diagrams

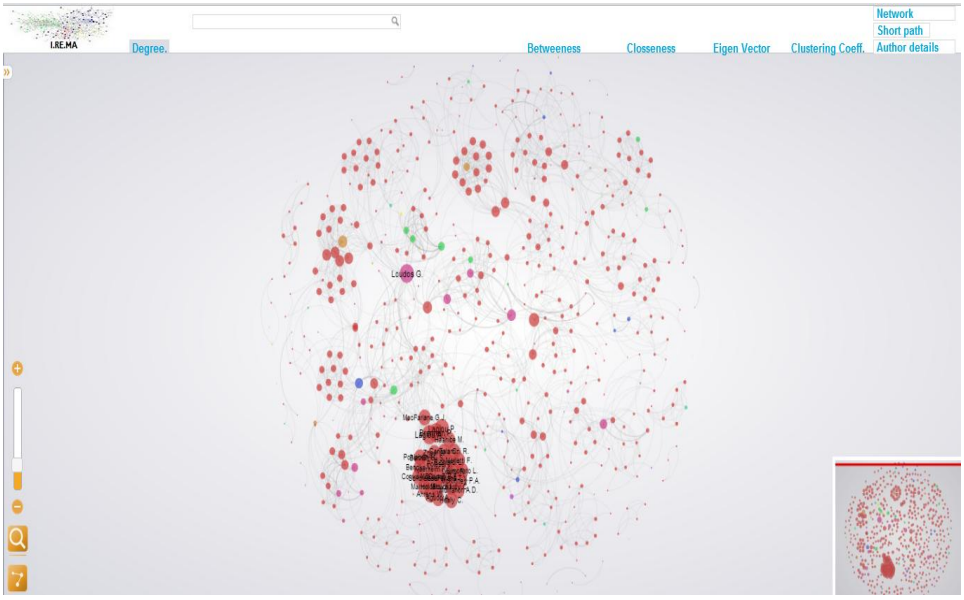


Figure 14: Force-directed node-link diagrams

A node link diagram (figure 14) could be defined as a network or a graph of an ordered pair (N,E) of a set (N) of nodes and a set (E) of edges which can be either directed or undirected. A link among two nodes n_1, n_2 exists only if there exists an edge $\{n_1,n_2\}$. In this case n_1 and n_2 are called neighbors. The degree of a node is the number of neighbors it has. Node-link algorithm is the most common graphical representation where each node is shown as a point and each edge as a line or curve. The main problem that researchers deal with is the computation of the positions among the nodes and the edges in such diagrams. The most reliable algorithms are those which are based on force-directed layout [37] for positioning the nodes.

The main idea of those algorithms is that there exist the two following forces:

(1) a repulsive force between all pairs of nodes, and

(2) a spring force between all pairs of adjacent nodes.

So the points are initialized with random positions, but gradually change their position under the effect of those two forces until the desirable position is reached. The nodes are getting their positions in such way that adjacent nodes are being near each other, but also not so close.

For instance, in figure 14, we can see a force-directed layout, which is generated for the co-authoring data of a Higher Educational Institute. One problem that we can see is that multiple edges can

hide the labels (or any information associated with each node) of some nodes making it unclear when certain edges pass close to a node. Therefore, when using that kind of representation we should take into consideration the length of the labels and how those are overlapped by others.

2.2.2.2. Arc Diagrams

In arc diagrams the nodes are displayed along a straight line where, edges can be drawn as circular arcs (figure 15), yielding an arc diagram. In that kind of representation we have the arcs from the one side and at the left of the nodes, we have free space for the labels. The most serious problem at this representation is the angle of the arcs.

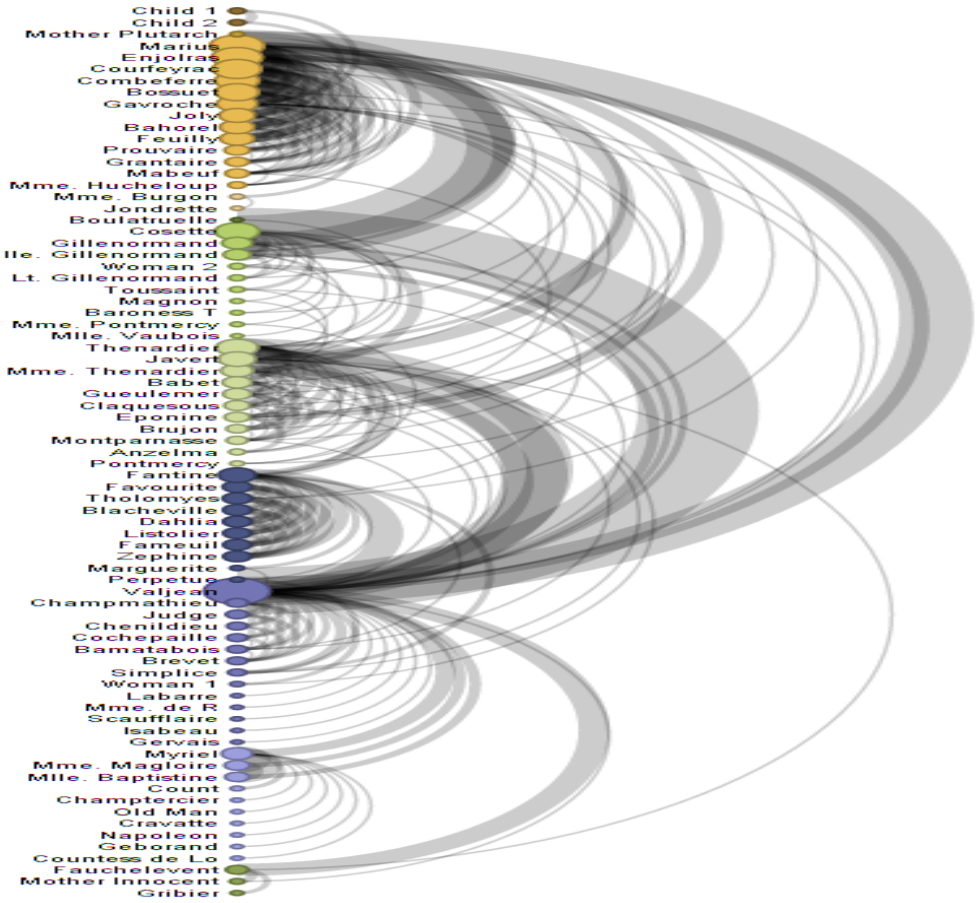


Figure15 : Arc diagrams

In order to implement an arc diagram, two key points exist.

- How to sort the nodes
- How to compute the angle (q) of the arc

Regarding the sorting of the nodes it could be done in various ways as for example alphabetic.

Regarding the computation of the angle, suppose that we have two connecting points.

The point A = (x, y1) and

The point B = (x, y2)

and we have to find the center C of the arc.

In figure 16 we can see an arc connecting A-B and a triangle connecting A, C, at the midpoint of the arc.

So the length of one side of the triangle is $d = |y_1 - y_2|/2$, and we also have $\tan \theta/2 = d/e$, hence $C = (x+e, (y_1+y_2)/2)$ where $e = d/(\tan\theta/2)$.

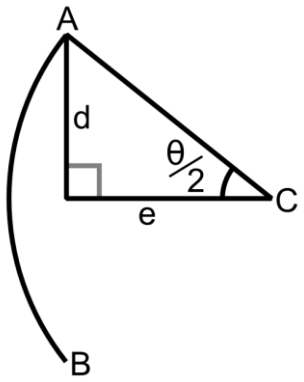


Figure 16: An arc covering angle q, with center C.

2.2.2.3. Adjacency Matrix

An adjacency matrix is a matrix, which describes a graph by representing which vertices are adjacent to which other vertices. If G is a graph of order n, then its adjacency matrix is an n×n square matrix, where each row and column corresponds to a vertex of G. The element a_{ij} of such a matrix specifies the number of edges from vertex i to vertex j or it can be a boolean value indicating if an edge exists between the two nodes. In figure 17 we represent an undirected graph, where the matrix is symmetric

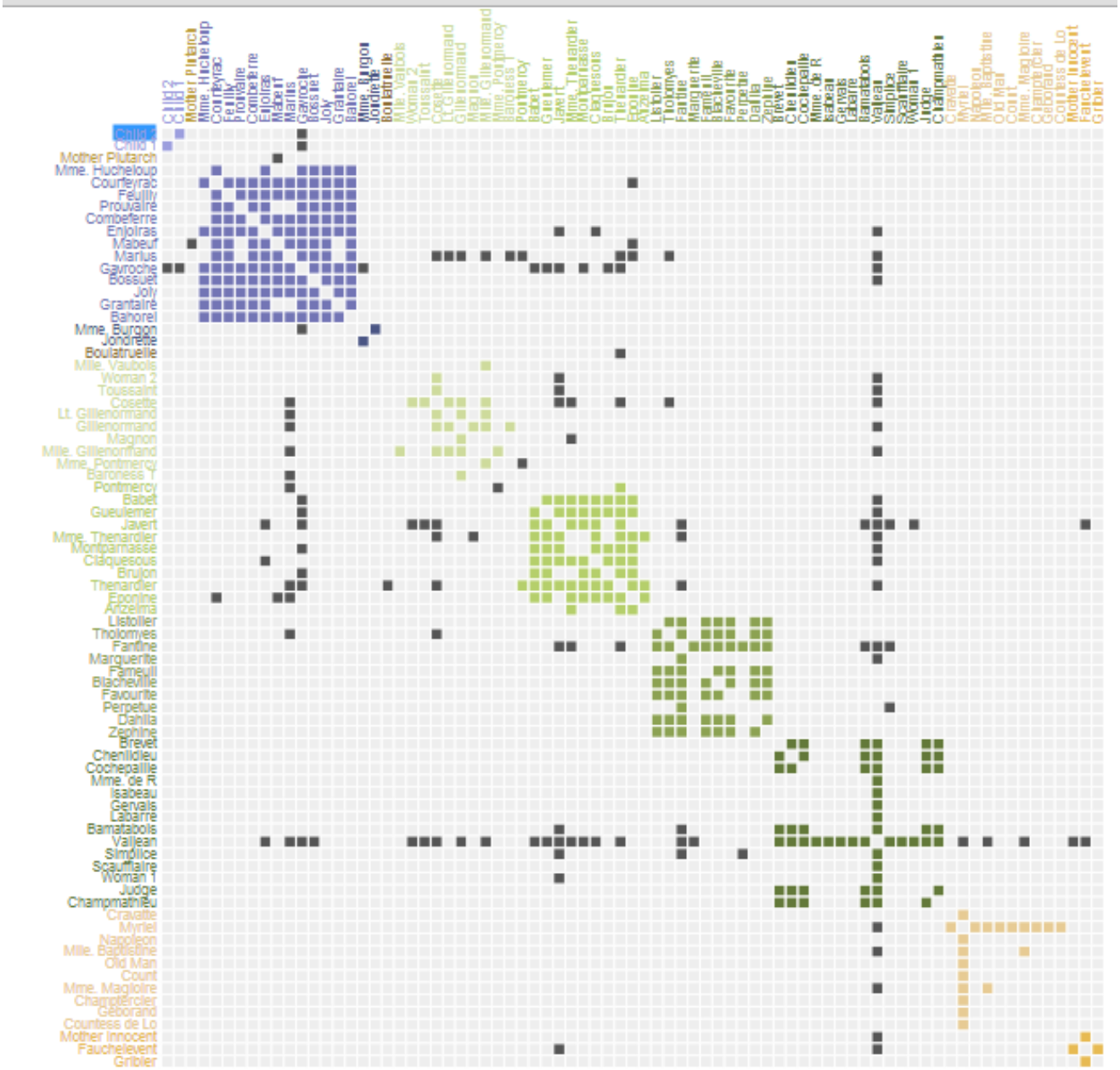


Figure 17: Adjacency Matrix

Source: The Stanford GraphBase.

With that representation it is easier to recognize certain patterns by observing their position into cells. For example cliques (all possible edges connecting the nodes) can be extracted by the matrix. As we can see in figure 17, the blue colored nodes seem to have connections to each other which mean that they have created a clique without any connections with the other nodes. Also, the degree of a node is easy to be recognized by calculating the number of filled cells.

An important disadvantage of using adjacency matrices as pointed out by Henry and Fekete[38], is that the space they require is $O(N^2)$ where N is the number of nodes.

2.2.2.4. Circular Layouts

In figure 18, a circular layout is displayed by positioning nodes on the circumference of a circle. In order to layout the nodes the barycenter heuristic [39] was used. Suppose that *C is* the center of the circular layout and *AC* and *BC* the corresponding points (figure 19). The center *C'* of the arc can be found by finding the intersection between a line through *A* that is perpendicular to *AC*, and a line through *B* that is perpendicular to *BC*.

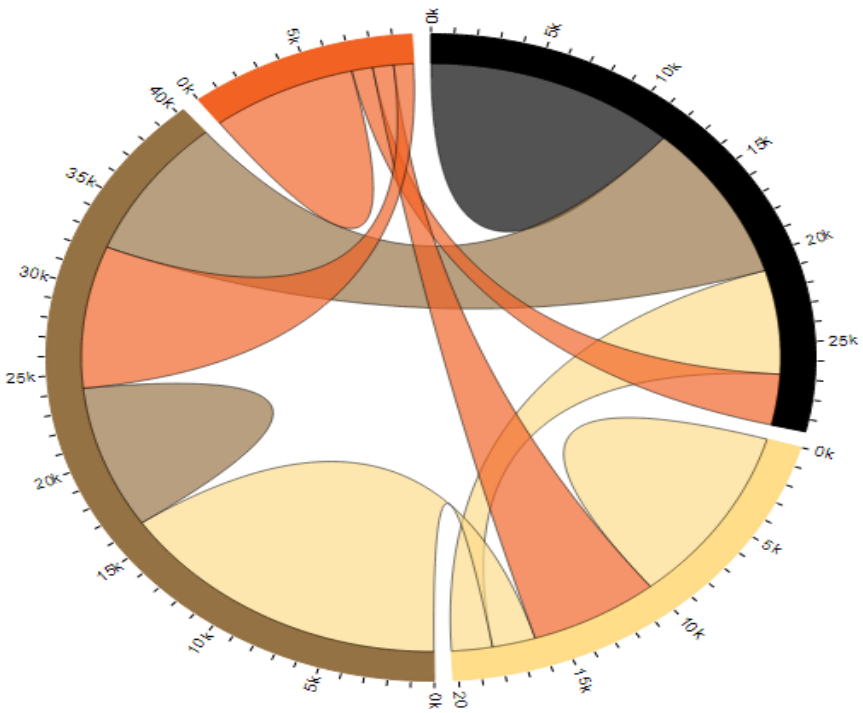


Figure 18: Circular layout

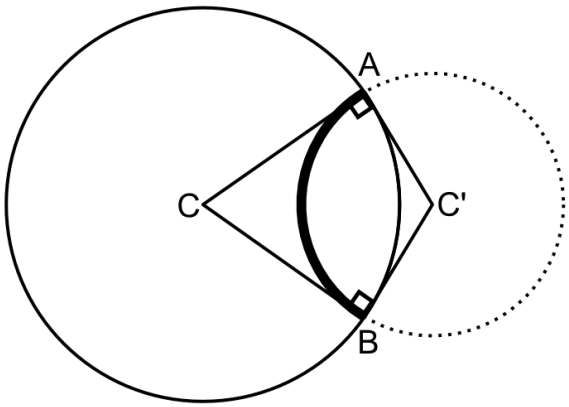


Figure 19: A and B are connected by the arc in bold

In order to compute the position of the neighbors of a node we convert each node to a unit vector in the appropriate direction, we add these unit vectors together, and finally we calculate the angle of the vector sum.

So supposing that we want to categorize the different visualizations, we could say that:

- Node-link diagrams may often be better for showing the topology of the network in a clear and simple manner. Also if the edges are associated with weights then the node-link diagrams are the ideal representation by using the color or the thickness of the edges.
- Matrix-based layouts are more likely to be used to display dense networks.
- Arc diagrams may be used for displaying some extra information about the nodes as they only use a single axis of the diagram.
- Circular layouts may also be used as the arc diagrams with the ability to display larger labels than the arc.

2.2.3. State of the Art

Our research in the visual representation of the graphs is based on the layout algorithms. In order to draw graphs (undirected) the most flexible method for calculating the layout is the Force-directed algorithms also known as spring embedders. At those algorithms forces are assigned among the edges and the nodes of a graph (figure 20). So, spring-like attractive forces are used to attract pairs of endpoints of the graph's edges towards each other, while simultaneously repulsive forces are used to separate all pairs of nodes.

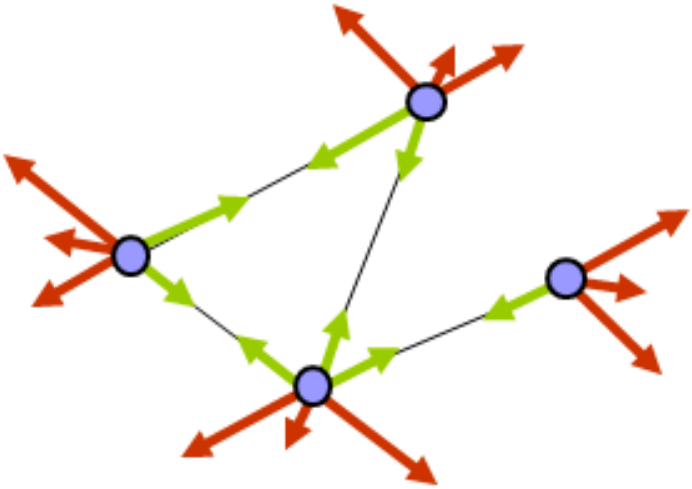


Figure 20: Force-Directed Power

One of the first force-directed graph drawing method is proposed by Tutte [40] based on barycentric representations. The idea behind that algorithm is to solve a system of linear equations for the suitable positions of the vertices, where each node is represented as a convex combination of the positions of its neighbors. Then Eades [41] proposed an algorithm where each edge was replaced with a spring. So the vertices are placed in some initial layout and the spring forces the rings in such a way the system to be in a minimal energy state. Also they make nonadjacent vertices repel each other.

In continuous Fruchterman and Reingold [42] enhance Eades algorithm by introducing a "global temperature" that controls the step of the node movements and the step that the algorithm will terminate. The step width is proportional to the temperature, so if the temperature is hot, the nodes move faster. This temperature is the same for all nodes, and cools down at each iteration. Once the nodes stop moving, the system terminates. Also in other approaches Kamada and Kawai [43] , select to use the graph theoretic distance (All-Pairs-Shortest-Path computation) in order to calculate the forces between the nodes.

The force-directed layout algorithm, as it is presented in the article of Michael J. McGuffin [44] is described in table 1.

Assume that the nodes are stored in an array of nodes, where each element of the array contains a position x, y and the net force force_x, force_y acting on the node. The forces are simulated in a loop that computes the net forces at each time step and updates the positions of the nodes, hopefully until the layout converges to some usefully distributed positions. The inner body of the simulation loop could be implemented like this:

```
L = ... // spring rest length

K_r = ... // repulsive force constant

K_s = ... // spring constant

delta_t = ... // time step

N = nodes.length

// initialize net forces

for i = 0 to N-1

nodes[i].force_x = 0

nodes[i].force_y = 0

// spring force between adjacent pairs

for i1 = 0 to N-1

node1 = nodes[i1]

for j = 0 to node1.neighbors.length-1

i2 = node1.neighbors[j]

node2 = nodes[i2]

if i1 < i2

dx = node2.x - node1.x

dy = node2.y - node1.y

if dx != 0 or dy != 0

distance = sqrt( dx*dx + dy*dy )

force = K_s * ( distance - L )

fx = force * dx / distance

fy = force * dy / distance
```

```
// repulsion between all pairs

for i1 = 0 to N-2

node1 = nodes[i1]

for i2 = i1+1 to N-1

node2 = nodes[i2]

dx = node2.x - node1.x

dy = node2.y - node1.y

if dx != 0 or dy != 0

distanceSquared = dx*dx + dy*dy

distance = sqrt( distanceSquared )

force = K_r / distanceSquared

fx = force * dx / distance

fy = force * dy / distance

node1.force_x = node1.force_x - fx

node1.force_y = node1.force_y - fy

node2.force_x = node2.force_x + fx

node2.force_y = node2.force_y + fy


// update positions

for i = 0 to N-1

node = nodes[i]

dx = delta_t * node.force_x

dy = delta_t * node.force_y
```

node1.force_x = node1.force_x + fx *node1.force_y = node1.force_y + fy* *node2.force_x = node2.force_x - fx* *node2.force_y = node2.force_y - fy*	*displacementSquared = dx*dx + dy*dy* *if (displacementSquared* *MAX_DISPLACEMENT_SQUARED)* *s = sqrt(MAX_DISPLACEMENT_SQUARED* */ displacementSquared)* *dx = dx * s* *dy = dy * s* *node.x = node.x + dx* *node.y = node.y + dy*

Table 1: Force-directed layout algorithm

Those actions will be repeated until the nodes are placed at their final positions.

2.3. Research & Development Collaboration networks

2.3.1. Introduction

Research within Higher Education has, on the one hand, to deal with the increasing and expanding scientific areas and on the other hand, with the diminishing resources. As a direct consequence, we believe that a clear view and capturing of the faculty members' research activity is essential for the creation of an institutional research profile that will in turn enable the promotion of collaboration arrangements as an effective way to enhance scientific performance; it can also increase the quantity and quality of research outcomes (e.g. articles, patents, etc.). In the context of our research we propose that in order to analyse the R&D networks we have to explore the collaboration networks among the authors (co-authoring) as well as to measure the efficiency among the research teams.

2.3.2. Co-authorship Networks

Co-authorship analysis is a useful metric for exploring collaboration patterns in a Higher Education Institute. It is the most common indicator assuming that co-authorship indicates a level of scientific collaboration [45]. As academic scientific collaboration we could define the knowledge sharing and cost rationalization in research infrastructures, in order to achieve the production of new scientific knowledge.

Because of the importance of research collaboration, several studies have attempted to quantify the concept (most concrete examples of that are based on Network analysis and Bibliometric indexes).

Network analysis uses mathematical models and graph theory to analyze graphs, e.g. centrality, distance, diameter, and cluster coefficient [33][34][35][36]. The referenced studies focus either on co-authorship network features or on individual author rankings within the different domains.

Bibliometric studies [46] regarding co-authorship focus mainly on the effects of collaboration to the scientific progress, based on authors as basis of the analysis.

Social network studies have primarily focused on the formation mechanisms of collaboration networks and on the understanding of underlying structures and processes, leading to the observed structures [47]. Relevant research [48] indicates that authors with many collaborators and high scientific prestige gain more connections from new entrants in the network; other work [49] uses co-authoring networks topology in order to observe "the best connected scientist".

Therefore, the evaluation of research activity within a Higher Education academic unit requires measurement along many dimensions and, in many cases, the design and utilization of multidimensional indicators.

In social network analysis, a number of important measures exist in respect to the indication of the importance of a node to the network topology [50]. These measures are used to explore the collaboration patterns based on the co-authoring networks.

A co-authoring network is a set of individuals or groups, each of them having connections with some or all the others. The individuals or groups are called "nodes" and the connections among them are the "edges" representing the collaboration / co-authoship activity. An edge exists between two authors (nodes), if they have at least one co-authored publication. Based on the topology of the network, various metrics have been introduced in order to provide information relevant to the importance of the node. Both nodes and edges can be defined in different ways depending on the research questions of interest.

2.3.3. Efficiency Measure

The literature about universities' efficiency has been largely focusing on the correlation among efficiency and productivity within departments of the same HEI, or among different universities. Madden et al.[51] focused on economic departments in Australian universities, and used as inputs teaching and research personnel, and as outputs graduates and publications. The main aim of their methodology was to

show how government's policy can influence the productivity. In another approach [52], 42 academic units in the USA were examined. They used as inputs staff, financial resources and infrastructures, and as outputs the number of students, FTE enrolments and grant awards. All of those studies examined the comparative efficiency among different universities. Several studies also examined the efficiency among different departments within a university. Pesenti, R. and Ukovich, W [53] derived efficiency scores of the Departments of the University of Trieste, by using human resources and funds as inputs, and teaching, research and fund raising as outputs. In addition, Tommaso Agasisti et all [54], select to use as inputs laboratories, high-qualified human resources and as outputs the yearly number of publications, citations per article, h-index, research funded through regional or national grants, research funded through international grants, and applied research through externally funded orders. They argued that efficiency rankings changed when considering different research-related outputs .Finally, Duk Hee Lee et all [55] examined the impact of collaboration patterns in the R&D performance, using the DEA methodology, using as inputs the full-time researcher and R&D investment expenditure(millions) and as outputs SCI papers, patents, technology licensing income(million).

From all the above we indicate the assumption that in most cases research papers were considered as the main indicator of R&D activities. Based on the officially established Hellenic Quality Framework for research efficiency [56] we considered as a case study to measure the research performance into a Higher Education Institute. In our approach we implement the Data Envelopment Analysis (DEA) (which is generally accepted as a benchmarking technique in order to evaluate the productivity of a unit or individual), by comparing the amount of output(s) produced in comparison to the amount of input(s). Then the performance of a research unit is calculated by comparing its efficiency, with the best observed performance in the data set. In our system we separate the indicators as follows:

Inputs indicators, which are classified under two main groups: human and financial resources.

Outputs indicators, which are grouped into publications, projects and financial indicators (e.g. grants)

DEA is a multi-factor productivity analysis model for measuring the relative efficiencies of a homogenous set of decision making units (DMUs). The efficiency score in the presence of multiple input and output factors is defined as:

$$\text{Efficiency} = \frac{\text{weighted sum of inputs}}{\text{weighted sum of outputs}}$$

$$max \ \frac{\sum_{k=1}^{s} v_k y_{kp}}{\sum_{j=1}^{m} u_j x_{jp}}$$

$$s.t.\ \frac{\sum_{k=1}^{s} v_k y_{kp}}{\sum_{j=1}^{m} u_j x_{jp}} \leq 1 \quad \forall i$$

$$v_k, u_j \geq 0 \quad \forall k, j$$

Where

x_{ji} = amount of input j utilized by DMU i,

v_k = weight given to output k,

u_j = weight given to input j.

k = 1 to s,

j = 1 to m,

i = 1 to n,

y_{ki} = amount of output k produced by DMU i,

s=the number of outputs, m= the number of inputs and i= the number of DMU's.

2.3.4. State of the art

Our research focuses on the co-authoring scientific networks and how to use those so as to extract decisions about the current situation, to predict future links among the authors and to extract a ratio about the efficiency and the collaborations among the scientist of a team.

Co-authoring

Several studies, in the area of co-authoring networks, show that research productivity is influenced not only by the structural advantages of positions within collaboration networks, but also by the relational characteristics of individual nodes (e.g. centrality) that seem to affect research results [55].

In general, analyzing co-authoring networks provides great assistance in evaluating the performance of individuals, groups, or even of the entire network [56]. The methods used as measure the importance of a node by examining the whole network and its participants. Therefore, co-authoring networks could be used for measuring the importance of a node, but they fail to provide the correlation among network characteristics and the efficiency.

According to Reagans and Zuckerman's (2001) [57], the links among the researchers (intensity) and the diversity of collaboration with other researchers affect the research performance.

Rigby and Edler[58] in 2005 indicated that the quality of the research activities is influenced by the collaborations and the length of the co-authoring network, as for example the papers which are initially reviewed by the others in the network and then are submitted.

Padula[59] in 2008 , based on US mobile phone industry, showed that not only the structural but also the relational characteristics of individual nodes affect research results.

Finally, Duk Hee Lee et all[55] in 2012 compared the R&D productivities of public research institutes, and they discovered that the institutes which achieved high productivity are those which have intensive relations with their existing partners.

The main deduction from all of those articles is that the intensity of collaborations among the scientist is highly correlated with their efficiency. In our research, we enhance those results by predicting future links among the scientists based on their research interests and collaborations.

Link Prediction

Another important advantage of social networks is the link predictions applications, which answers the problem of recommending new friends (in a social network) or links more general in graph network. More specific, the network in a certain time t is studied and the links that will be added in another time t' in the future is predicted. In a coauthoring network by applying the same concepts we study the possibility of two scientists to collaborate, or predict who are more likely to collaborate. For example, if we study the network among the scientists in an institute and we observe that two authors do not have any link but have collaborated in the past with common scientists, then these scientists are more likely to collaborate between them than with any others. Usually, in the link prediction algorithms a score (x, y) function exists, which calculates the score between two nodes (x,y) which are expected to establish a collaboration. Links can be predicted either by the study of the topological structure of the network or by studying the whole topology and examining potential paths that could be created [60].

Stanley and Milgram, claimed that all people could be linked using paths with few hops [61]; to prove this, they created an experiment by applying six degrees of separation. Goel et al. [62] executed the same experiment; they found that a 50 % of the human chains (paths) were also executed in 6-7 steps like Milgram [61]. They have defined that the number of steps depends on the selection of the people, and the algorithms used to find the middle steps of the chain.

Tylenda et al. [63] used time and date variables, in their predictions. Zheleva et al. [64], studied the link prediction in social networks using data from people's lives. Other approaches, calculate the similarity measure by examining the nodes and their adjacent nodes; an example of that type is the algorithm Friend of a Friend (FOAF); and another example is the Adamic/Adar, e.t.c. The FOAF [65] algorithm is used in order to predict the likelihood of two nodes trying to establish a connection based on the common neighbors. The Adamic/Adar [66] enriches the previous algorithm by adding more characteristics. In the preferential attachment (PA) [65] algorithm, the likelihood of a new edge is proportional to the current number of neighbors. In [67] a degree which measures the most active and the most influential nodes of a graph is introduced. Another algorithm on the same category is the Random Walk with Restart (RWR) [68], which is based on the Markov model of random walk in a graph. The SimRank [69] algorithm measures the similarity based on the assumption that two objects are similar if they are connecting with similar objects.

2.4. Research Information Management Systems for HEIs

Nowadays, university-based research can undoubtedly be characterized as the primary arena for the production of new knowledge. The impact of Higher Education on the research-innovation eco-system has considerably increased and HEIs have become an important focal point for national policy making[70]. Furthermore, within the European area, universities are considered at the forefront of “Europe’s drive to create a knowledge-based society and economy and improve its competitiveness”[71]. On this basis, universities experience an unprecedented change, having to face crucial challenges regarding institutional strategies and governance, quality assurance, financial and social accountability as well as the increasing impact of globalization, marketization and new technology[72]. Responding to such requirements calls for research intelligence and well-established, technology-enhanced performance management frameworks and tools.

Research Information Systems (RIS) are informational tools “dedicated to provide access to and disseminate research information” [73]. RIS consist of a part of intra or inter-institutional infrastructure for research stakeholders, aiming at supporting a plethora of research processes ranging from the recording and reporting of research activity and its outcomes to the complex analysis of research information for assessment, decision making and planning purposes. For the individual researchers and innovators, RIS can offer opportunities for effective documentation of their research outcomes and profiles, management of their research projects lifecycle, enhancement of researcher networking and collaboration, efficient access to scholarly resources (publications and data), analysis of funding opportunities and trends. At institutional level, RIS constitute a formal log of research, as far as products and collaboration is concerned, and an important tool for evaluation of output-based research, policy making and strategic planning. These

functionality aspects are of crucial importance for research managers and research policy makers as well as for governance bodies and the general public.

A considerable number of systems and tools exist to support the gathering, management and analysis of data, using machine learning and data mining techniques. The Research Portfolio Online Reporting Tool (RePORT)[74], for example, which is developed by the National Institute of Health (NIH), the US medical research agency, captures automatically publications from PubMed [75] and provides access to reports, data and analyses on NIH research activities. In the STAR Metrics [76]approach, the impact of federal science investment on scientific knowledge (using metrics such as publications and citations), social outcomes (e.g. health outcomes measures and environmental impact factors), workforce outcomes (e.g. student mobility and employment), and economic growth (e.g. tracing patents, new company start-ups and other measures), are measured.

Other systems, such as the GridMiner [77], which provide support for data mining and On-Line Analytical Processing (OLAP) have been effectively used for educational purposes. The Development and Mining System [78] developed by the Information Technology and Systems Center at the University of Alabama in Huntsville applies data mining technologies to scientific data; it can also be used as a pattern extraction tool. All of the aforementioned systems focus on the researchers and on the evaluation of research outcomes.

Undoubtedly, there is currently no 'one-fits-all' solution as far as institutional RIS is concerned. The lack of uniformity in orientation and processes drives institutions to develop systems that meet their (perceived to be) specific needs; in addition, suppliers have difficulty in viewing the marketplace from a global perspective. They rather focus on segments of the research environment, building up isolated competencies in one or two specific areas. This results to a cocktail of systems within institutions, some off-the-shelf and many created in-house, which are developed without considering the research management environment in total; these systems lack the capability to integrate with each other, often using conflicting data structures.

Our work explicitly targets specifically at this gap: it addresses the need to develop and deploy integrated RIS, based on flexible, extensible semantic approaches, leveraging widely-accepted interoperability specifications. IREMA constitutes such a framework, with specialized focus on the subgroup of Research Management Information Systems, in the sense that it considers how knowledge from research data can be derived and used to support strategic decision making at institutional and individual researcher levels. More specifically, the functional aspects considered in our approach directly relate to the following research management requirements:

- Aggregate and benchmark research outputs (including publications, funded projects and patents).
- Reveal strengths and weaknesses of individual and institutional research activity.
- Help academics identify strategic priority areas to perform research.
- Help researchers collaborate by capturing and analysing research activity, especially in interdisciplinary areas, within institutions, across departments, and with researchers from other institutions.
- Help institutions collaborate by facilitating and tracking opportunities with industry, national and local government bodies, and with other institutions.

The IREMA solution attempts to combine the DSS domain with the research area of research collaborations evaluation within a higher educational environment, offering thus a new potential for developing enhanced RIS services and providing at the same time added value in research management. Our approach incorporates DSS design principles allowing non-experts to gain insightful research information without any previous knowledge about data mining techniques; special care has been taken for the incorporation of new features that are specific to the area of research collaboration, in order to implement a DSS that provides the most effective solutions.

3. Thesis Contribution to Decision Support Systems for Research Evaluation

The main aim of our work is to design an architecture that will be used for evaluation purposes with the use of graphs. We have introduced the layer of domain knowledge in the visual analytic process, so as to be able to use the proposed architecture in different domains. Also, we studied and developed a graph layout algorithm that could be used to display complex data set in real time. This graph layout algorithm provides an interface to the user from which he/she can modify the graph layout. Finally, we will present a recommendation algorithm, which is used to predict future links among two entities in a graph network.

3.1. Visual Analytics Systems

3.1.1. Motivation

The main aim of visual analytic systems is to provide automatic reasoning without any interaction from the user in the design of the data analysis work flow. The basic idea of visual analytics is to apply data mining techniques, allowing the user to get representations of the results, or to obtain the results of a hypothesis based analysis. These systems make possible the analysis of huge amount of data and as Thomas & Cook [17] defines:

"Visual analytics is a science of analytical reasoning facilitated by interactive visual interfaces."

It is apparent that the main goal of visual analytic systems is to keep the end user away from the data analysis process. The user will get images that display the results of the analysis without participating in the analysis or in the design of the process. The success factor of such systems is based on the visual perception and analysis capabilities of the end user. The results are created in such a form that will be comprehended by users who are familiar with the specific domain, and who can explore the visualization in such way that would lead them in effective decisions. As Fisher [79] indicates, the success of a DSS depends more on its interface for human–computer interaction and less on its capabilities to solve a problem. On the other hand, if the user does not have any previous knowledge of the domain, he will find difficulties about the design of the process or the exploration of the visual interfaces. While the user has the opportunity to use a DSS tool, which supports data mining techniques and provides visual interfaces, because of the lack of knowledge of how to use it, he/she would not be able to interact with it in a proper way, leading thus to ineffective decisions. Moreover, in most of the times the user will get a visual result

that will provide fault answers to specific hypothesis because he/she would not know all the parameters that may exist, and as a result he will get a visual result not corresponding to the real situation.

In order to overcome this problem, we introduce the use of ontology [80] in visual analytic systems which contain relevant domain knowledge, with explicit formal specifications of the terms in the domain and relations among them. So, in order to develop a visual analytic decision support system we have to consider the following:

- The ability to solve complex problems (Data Analysis)
- The visual interface that will display the results and will help the user to explore the data. (Visual Interfaces)
- The human- computer interaction (HCI) and
- The Knowledge domain

In all the aforementioned approaches, the visual analytic systems consisted of the three disciplinary areas of User interface, Data Analysis and Visualization methods as represented in figure 21 (left) [80]. In our approach we also add the Domain Knowledge, which is represented with the dotted line. We can see how the different disciplinary areas collaborate to the decision support systems and how they integrate to visual analytics decision support systems.

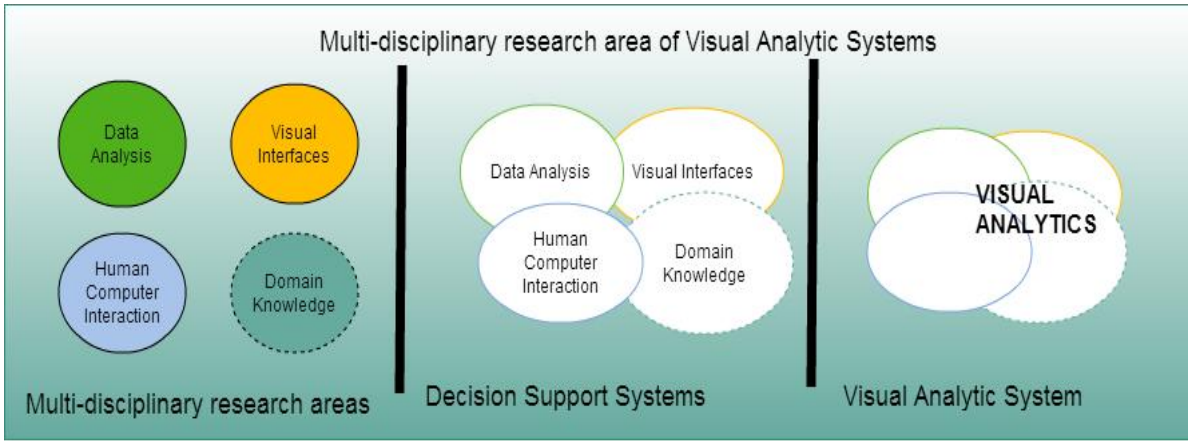


Figure 21: Multi-disciplinary research areas

3.1.2. Visual Analytic Process

In this section, we will discuss about our contribution in the architecture (figure 22) of the visual analytic process. In our proposed process five stages exist:

- The Data (D)
- The Ontology (O)
- The Visualisation interfaces (V)
- The Hypothesis generation (H)
- The Knowledge (K)

The visual analytic process can be seen as a transformation F: D->K, whereas F ∈ {Dx, Oz, V_Y, H_A, K_B }

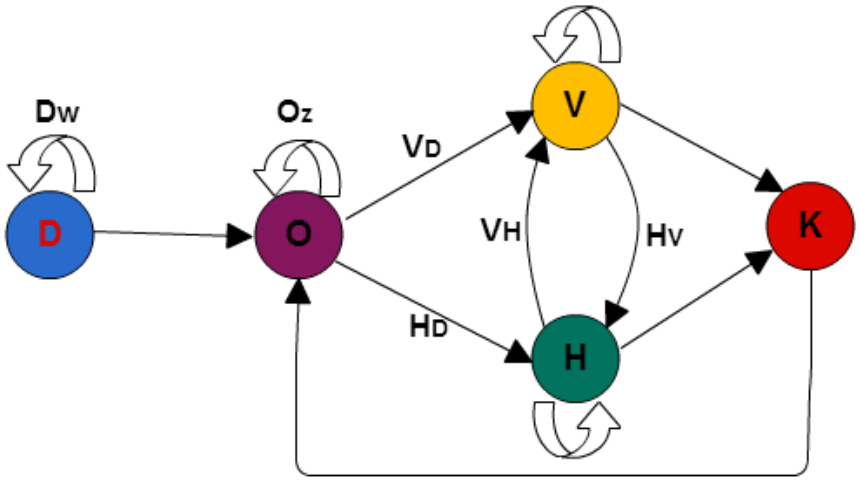


Figure 22: IREMA visual analytic process

The process begins with the input data D and at those data we apply the D_W : D → D where W ∈ {A,C,T}.

The D_A function performs acquisition tasks. The data could be retrieved from a DBMS, through web services, .xls or .csv and .arff files.

The D_c function performs the following tasks:

- Fill in missing values
- Convert the data to the appropriate format (e.g. nominal to numeric, binary to numeric)
- Identify outliers and smooth out noisy data (incomplete data, duplicate records,..). In order to solve such problems we apply both human and computer inspection.

The D_T function includes the data integration and transformation to the appropriate format in order to be aligned to the next stage. The stages of the data processing are defined as follows $D = D_T (D_C (D_A(D)))$. $O_{Z,}$ symbolizes all the modifications that an expert user applies to the ontology. The modifications are not applied to the structure and the classes of the ontology, but to the relations among the classes. V_Y, $Y \in \{D,H\}$ symbolizes the visualization functions. The functions could visualize data $V_D : D \rightarrow V$ or visualize hypotheses $VH : H \rightarrow V$. The visualization process consists of a variety of different representations (figure 23) which can be used to reveal hidden patterns.

H_A , $A \in \{D, V\}$ represents the hypothesis generation process and it could be separated to those that generate hypotheses from data $H_D : D \rightarrow H$ and those that generate hypotheses from visualizations $H_V : V \rightarrow H$. The hypothesis generation process consists of data-mining, social network analysis and statistical methods.

The user interaction process U_A is the most important part of the visual analytic process. The user could affect the results of the visualization images (color, opacity, zoom,..); the user could also change the parameters or the methods for the hypothesis (data mining , or interact using visual tools in a way that new knowledge is generated; finally the user could use the results of a hypothesis in order to generate knowledge. Therefore, the user interaction process is represented by U_B, $B \in \{V,H,KV,KH\}$ where $U_V : V \rightarrow V$, affect only visualizations, $U_H : H \rightarrow H$ affect only hypotheses by generating a new hypotheses from given ones, $U_{KV} : V \rightarrow I$ knowledge can be extracted from visualizations and $U_{KH} : H \rightarrow I$ knowledge can be extracted from hypotheses.

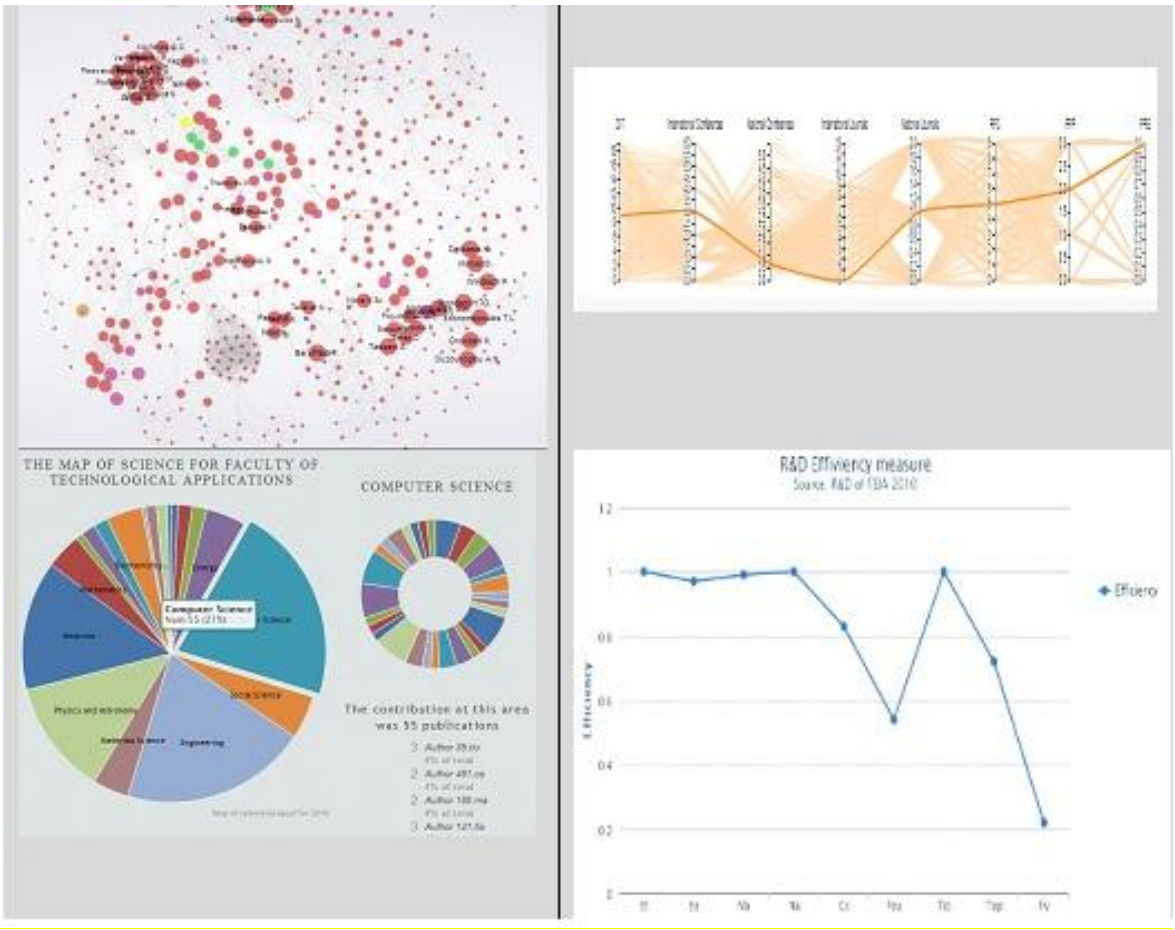


Figure 23: Figure of Visualizations

To sum up, the visual analytic process is based on the knowledge of the user for the specific domain as to be able to get an instant representation of the data or the hypothesis U_{KV} , U_{KH} in order to get decisions for a specific problem or to continue the process by applying a new data analysis function U_V, U_H . The decisions and the results of the process depend on the knowledge that the user could already have for a specific domain. By using the Ontology we enhance the data exploration process by providing semantically meaning to the data. The advantage is that user is not obligated to know in depth the structure of an organization and the possible relations, due to the existence of the ontology which involves such knowledge in a form of statements about the domain of interest. Thus, the user applies reasoning [82] methods to the data in order to get answers for specific queries, without any knowledge about the attributes and the properties among the classes. An ontology could contain rules which have the IF-THEN form and represent and express various complex statements that may exist. By doing so, the end user interacts with the system and gets answers in a visual way, which contains all the specific statements that may exist in an organization.

For example, using the "is-a" relationships among the classes, the ontology uses a hierarchy structure, which means that each member of a child class is also a member of the parent class. Therefore

when a user executes a query for a person that belongs for example in the department of informatics which is "part-of" school of Technological Applications, while the same person is "teacher of" a course and also the person "has-a-ResearchArea", we observe that the query can be expanded to the relationships between the classes or also to exclude them. So using the ontology we could get answers not only for the class "person" but also for the relations that exist among the person and the courses, as well as the department and the research areas.

3.1.3. Ontology for Research & Development Management

Research & Development activities constitute one of the main goals of Higher Education Institutions (HEIs) and a significant quality indicator of their performance and contribution to the society. Academic institutions in Europe and worldwide strive for excellence and invest on the establishment of effective quality assurance systems and processes in order to improve their effectiveness in both directions [83]. Quality assurance is an intriguing and complex process [84] based on methods and tools for capturing past performance and measuring future capability. HEI's evaluation and quality assurance is built around indicators including the quality of teaching, research, services provided, and offered curricula [50][85]. Thus, it is of vital importance to capture, measure and analyze the activities connected with these indicators. Depending on the focus of the evaluation, different dimensions of institutional activity must be considered and analyzed [86].

Networking and collaboration within scientific and academic communities is an essential factor for the promotion of research achievements and contributions [87], the "open discourse" and the increase of the availability of the relevant resources to the community [88]. Science networking applications that are based on Web 2.0 are usually referred to as Research 2.0 or Science 2.0 [89]. To elaborate, nowadays, researchers need to share their research results in wider audiences, to establish relationships within their institution, or at a larger scale.

The IREMA ontology has been developed in order to represent information concerning the research activities within HEIs. It is based on the VIVO extended ontology [80], in order to include collaboration and metrics information. To elaborate, the IREMA ontology addresses the needs of the faculty, the researchers and of both graduate and undergraduate students, creating, storing and exploring academic activities and collaboration possibilities.

Introducing to the IREMA Ontology

The IREMA ontology reuses and extends the VIVO ontology. The IREMA ontology is designed to incorporate all the research aspects of an institution, metrics for the research performance and events

offered within an academic institution, as well as the research co-operations that take place. IREMA is defined by means of a set of data elements, for each of which, the following set of parameters are defined by the IREMA application profile:

Label (Element name) : The element name indicates the name of the specific element in the IREMA ontology.

Properties: The properties of the elements.

SubClass: The class in which the element belongs.

Description: The description of the element

Example: An example of the element

Thus, the basic components that have been used in the IREMA ontology are the following:

- Academic Degree
- Country
- Equipment
- Person
- Project
- Role

These components are important for the conceptualization of the research activities and collaborations within an academic unit. IREMA ontology's main focus is the research interactions and activities of the faculty members. It provides information on the events that happen in an academic department, the authorship and the co-authorship of a research paper. Also it provides relevant information on the research performance, which is useful for the comparison among the scientists.

The IREMA ontology is presented in figure 24.

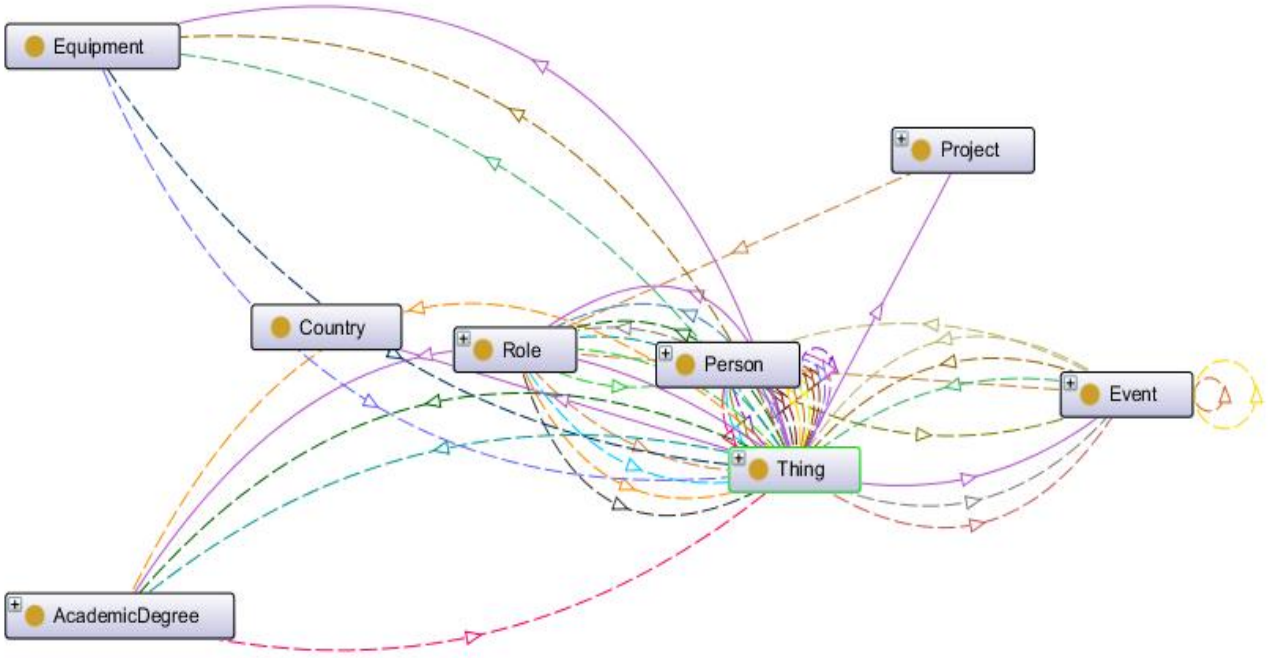


Figure 24 : IREMA Ontology

The ontology consists of the following classes:

The Academic Degree: It describes the academic degrees at any level, both as reported by individuals for employment and as offered by academic degree programs.

Label	**Academic Degree.**
Description	This list may have multiple abbreviations for some degrees.
Example	B.A. Bachelor of Arts.
SubClass	---
Properties	abbreviation only rdfs:Literal, abbreviation min 1

Table 2: Class Academic Degree

Figure 25: Academic Degree Class

The Type: It describes the type of the academic degrees.

Label	**Type**
Description	
Example	phd, mba,msc,ba
SubClass	**Academic Degree**
Properties	

Table 3: Class Type

The Thesis Degree: It describes the academic degree of a thesis.

Label	**Thesis Degree**
Description	Different from general academic degree, thesis degree is achieved through one's completed thesis. Thesis is a document submitted in support of candidature for a degree or professional qualification presenting the author's research and findings (http://en.wikipedia.org/wiki/Thesis_or_dissertation).
Example	Doctor of Philosophy (Ph.D.)
SubClass	**Type**
Properties	

Table 4: Class Thesis Degree

The Country: It describes a country.

Label	**Country**
Description	
Example	Greece, France
SubClass	
Properties	

Table 5: Class Country

The Equipment: It describes a physical object provided for specific purpose, task or occupation.

Label	**Equipment**
Description	A network server is one example. Medical schools and research laboratories can list professional equipment, such as microscopes.
Example	server; Bruker Vector-33 FT-IR
SubClass	
Properties	Free-text Keyword only rdfs:Literal

Table 6: Class Equipment

The Event: It describes something that happens at a given place and time. It consists of the following classes:

Conference, Interview, Performance, Research Area Topic, Prototype Standards, Courses, Presentation, Exhibit, Patent, Journal, Award, Competition, Workshop

Label	**Event**
Description	It describes an event
Example	Conference, Interview, Performance, Research Area Topic, Prototype Standards, Courses, Presentation, Exhibit, Patent, Journal, Award, Competition, Workshop
SubClass	

Properties	contactInformation only rdfs:Literal, realizedRole only Role, hasSubjectArea only owl:Thing , description only rdfs:Literal

Table 7: Class Event

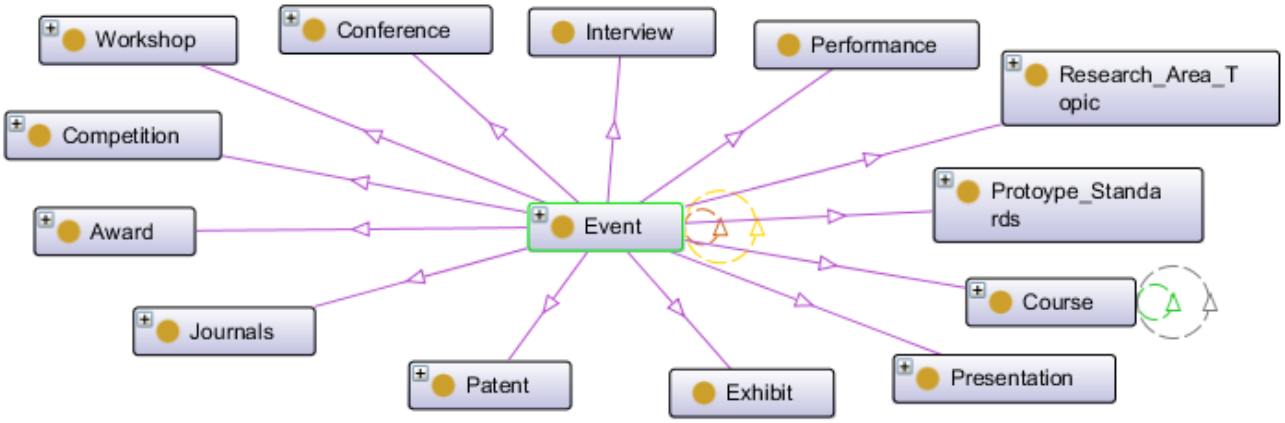


Figure 26: Event Class

- The Conference: It describes a meeting for consultation or discussion.

Label	**Conferences**
Description	A meeting for consultation or discussion
Example	Panhellenic Conference on Informatics 2013
SubClass	Event
Properties	

Table 8: Class Conferences

The conference consists of:

- Research proposals. A proposal for a research grant that has been submitted but not approved; does not represent an existing activity
- Type of Article. A written composition on a specific topic, forming an independent part of a book or other publication such as a newspaper or magazine.
- Document Status. The status of the publication of a document.
- Document Part. A distinct part of a larger document or collected document.

- Metric. The performance metrics for the conference
- ConferencePoster.
- Collected Documents. Work consisting of collections of previously published works
- Book.

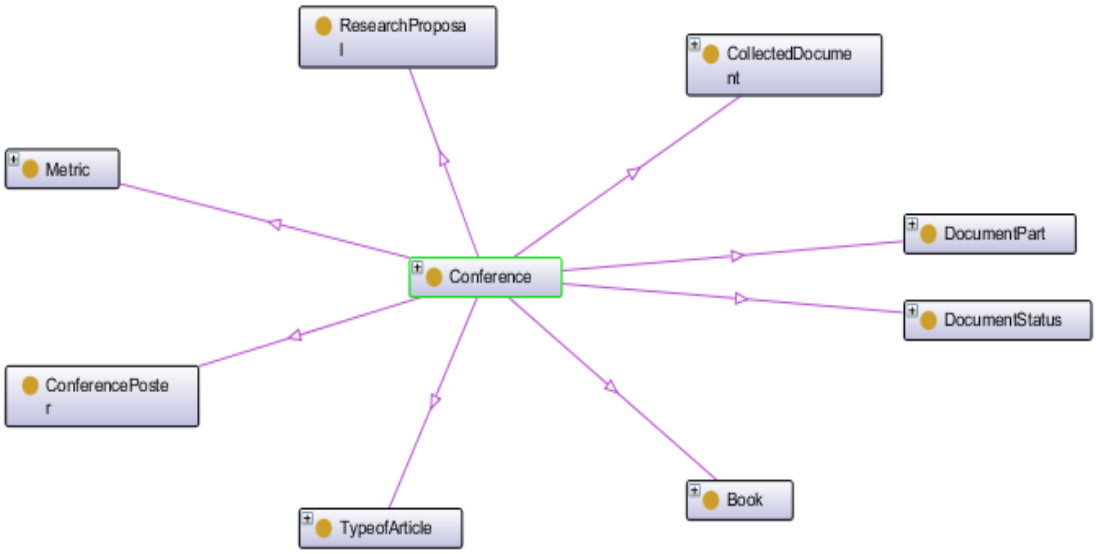


Figure 27: Conference Class

- The Interview: It describes a formalized discussion between two or more people.

Label	**Interview**
Description	A conversation between two or more people where questions are asked by the interviewer to obtain information from the interviewee.
Example	radio or newspaper interview
SubClass	Event
Properties	

Table 9: Class Interview

- The Performance: It describes something carried out, acted or rendered.

Label	**Performance**
Description	Something carried out, acted or rendered.
Example	Painting
SubClass	Event

Properties	

Table 10: Class Performance

- The Research Area Topic: It describes the research areas.

Label	**Research Area Topic**
Description Annotation	It describes the research areas of the Events
Example	Computer Science
SubClass	Event
Properties	

Table 11: Class Research Area Topic

- The Prototype Standards: It describes a prototype standard.

Label	**Prototype Standard**
Description Annotation	It describes the prototype
Example	IEEE
SubClass	Event
Properties	

Table 12: Class Prototype Standard

- The Courses: It describes a course as taught in one time period by one or more instructors, normally but not always for credit.

Label	**Course**
Description Annotation	A course as taught in one time period (such as a semester; although note that a course could consist of only one meeting -teaching session) by one or more instructors, normally but not always for credit.
Example	Advanced issues of databases
SubClass	Event
Properties	courseCredits only rdfs:Literal

Table 13: Class Course

- The Presentation: It describes a presentation

Label	**Presentation**
Description Annotation	Presentation
Example	Encompasses talk, speech, lecture, slide lecture, conference presentation
SubClass	Event
Properties	

Table 14: Class Presentation

- The Presentation: It describes an Exhibition

Label	**Exhibit**
Description Annotation	Exhibit
Example	Encompasses Exhibitions
SubClass	Event
Properties	

Table 15: Class Presentation

- The Patent: It describes a patent

Label	**Patent**
Description Annotation	Patents
Example	DE10161898 A1
SubClass	Event
Properties	

Table 16: Class Patent

- The Journal: It describes a journal

Label	**Journal**
Description Annotation	A group of related documents issued at regular intervals.
Example	Journal of Informatics
SubClass	Event
Properties	

Table 17: Class Journal

- The Award: It describes an award

Label	**Award**
Description Annotation	The award
Example	Nobel prize
SubClass	Event
Properties	

Table 18: Class Award

- The Competition: It describes a Competition

Label	**Competition**
Description	An Innovation Competition
Example	Greece, France
SubClass	Event
Properties	

Table 19: Class Competition

- The Workshop: It describes a seminar, discussion group, etc..

Label	**Workshop**
Description	A seminar, discussion group, or the like, that emphasizes exchange of ideas and the demonstration and application of techniques, skills, etc.
Example	seminar, discussion group
SubClass	Event
Properties	

Table 20: Class Workshop

The Person: It describes all the persons that exist into a higher educational institute, separated into Faculty members, students and Non academic personnel. Also there are Metrics which describe performance indexes for these persons.

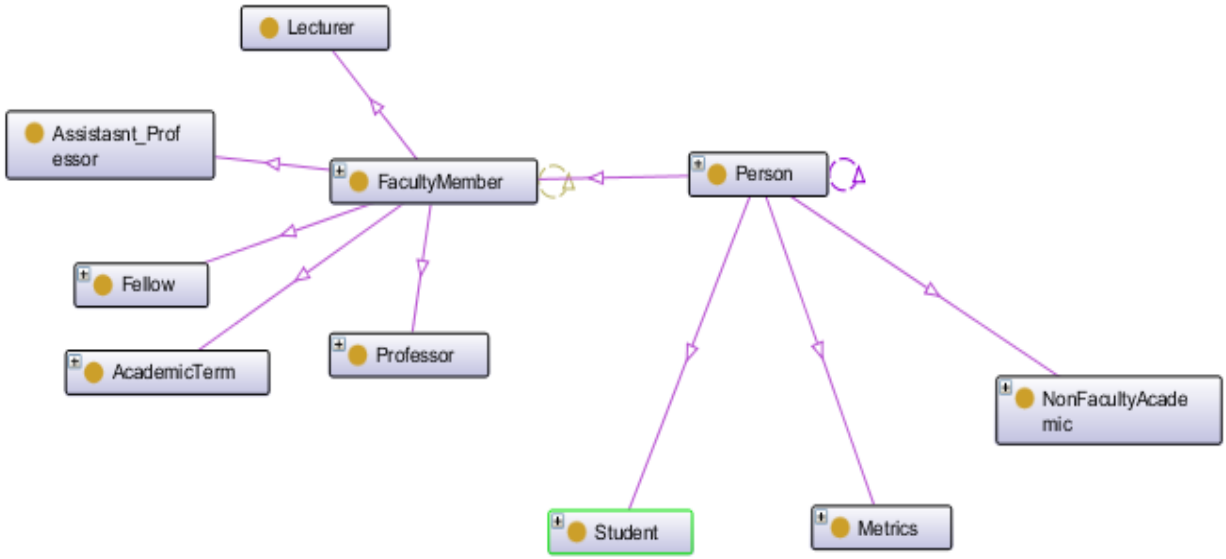


Figure 28: Person Class

Label	**Person**
Description	Used to describe any "agent" related to bibliographic items
Example	-
SubClass	-
Properties	primaryEmail only rdfs:Literal freetextKeyword only rdfs:Literal

Table 21: Description Person

The Project: It defines all the categories of the projects that the faculty members participate. It consists of the

- Industry projects and
- Funded projects that could be either European or national.

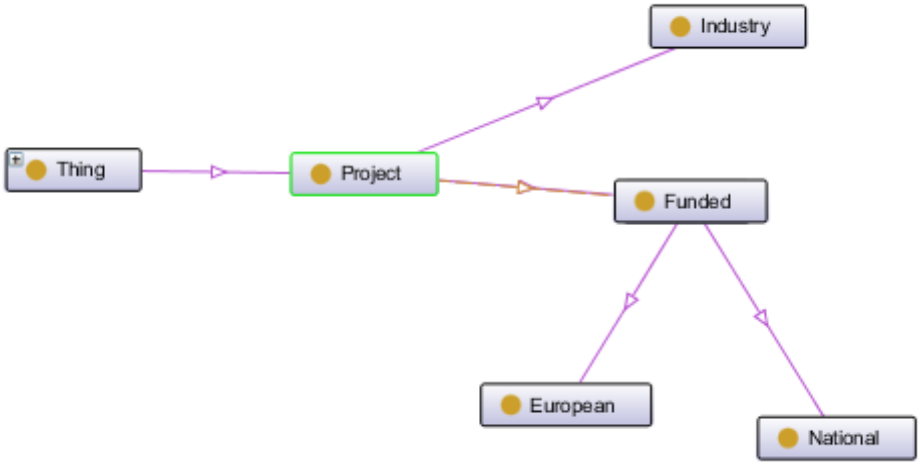


Figure 29: Project Class

Label	**Project**
Description	Used to describe any project.
Example	FP7
SubClass	
Properties	realizedRole only Role description only rdfs:Literal contactInformation only rdfs:Literal

Table 22: Description Project

The Role: It defines all the roles that the faculty members could get. It consists of the

- Editor. An ongoing editorial responsibility for a bibo:Collection, such as a Journal or Series
- Leader. A broad-ranging leader concept, from leading a small temporary committee to head of a large international organization.
- Organizer. A role of organizing
- Outreach provider. An outreach or community service role directed outside a person's primary profession and institution
- Presenter. A role of presenting information
- Reviewer.
- Service Provider. A role of an individual within his or her profession or institution.
- Teacher. A role of serving as an educator.

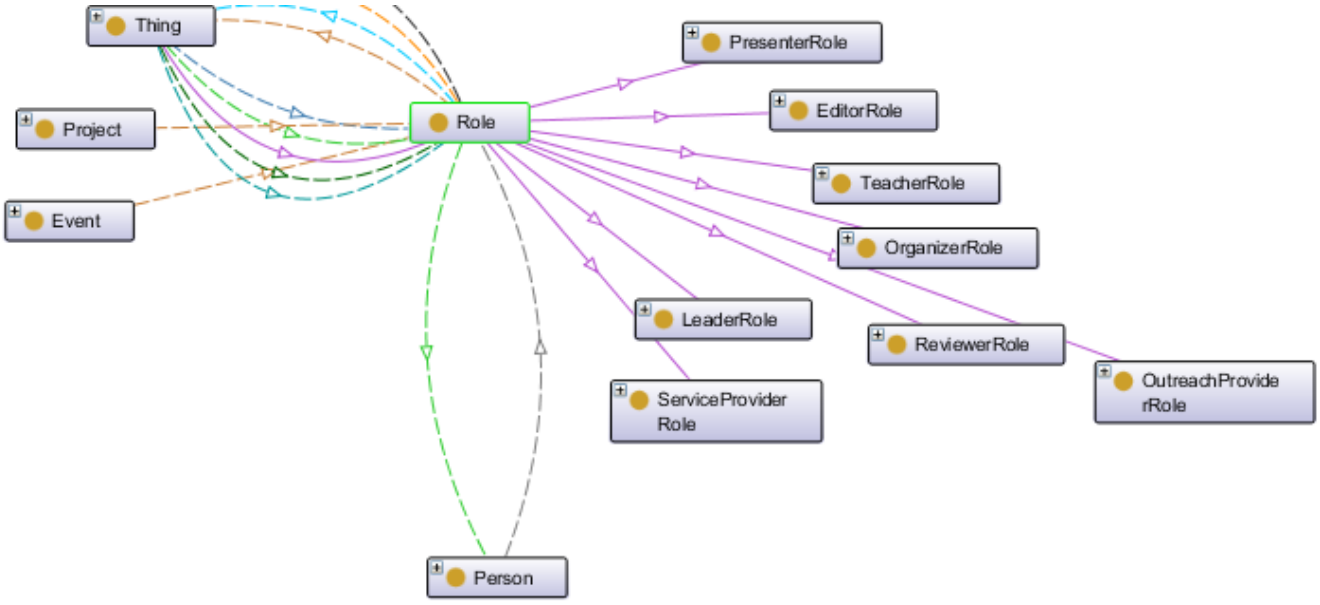


Figure 30: Role Class

Label	**Role**
Description	Used to describe any role
Example	Teacher
SubClass	-
Properties	description only rdfs:Literal

Table 23: Description Role

Finally, note that the class Thing is the root class of ontologies and it is not relevant to our ontology Thus, the definition of the class Thing is not at the scope of this document.

3.2. Data Visualisation using Graphs

3.2.1. Motivation

As the co-authoring networks capture the research activities of the authors including the authors, their papers, the citation number of their papers and the research discipline of them, graph representations seem to be the most suitable to map them since they represent multi dimensional problems. A graph has 4 visual attributes that can be used in order to represent co-authoring values. Those attributes are the topology, the colour, the size and the shape of the nodes.

The layout of the graph can be defined by the ForceAtlas2 algorithm [90], which belongs to the force-directed vector algorithms. ForceAtlas2 is suitable to deal with very small graphs (10 nodes) and fast enough to spatialize 10,000 nodes graphs in few minutes, with the same quality. It is a continuous algorithm that allows the user to manipulate the graph while it is rendering. It is based on a linear model (attraction and repulsion proportional to distance between nodes). In order to shape the graph we can use the Früchterman & Rheingold's [91] layout or Noack's LinLog [92]. Our contribution focuses on the development of a real time interactive tool for modifying the layout, in order to enhance the user interaction as the user will be able to create a more personalized layout based on his/her own needs.

Using the enhanced ForceAtlas2 algorithm the user will be able to:

- Produce a readable spatialization and devise an energy model that could be easily understood by users.
- set different options in such way that users are allowed to tune the shape of their network.

3.2.2. Methodology of ForceAtlas2

In the ForceAtlas2 algorithm there exist forces, which are applied on the nodes attracting and repulsing each other. In the ForceAtlas2 the forces among the nodes are similar with those applied on magnets (repulsion) and springs (attraction). Initially, the nodes are placed in the center of the network and their final position is influenced by the interaction with the other nodes. This way, the final graph layout and the placement of the nodes depend on the forces applied between the nodes. In this algorithm the structural proximities are interpreted into visual mappings, facilitating social network analysis of the communities. In order also to create the layout, an energy model such as Noack's LinLog and Fruchterman and Rheingold is used, where the repulsion forces depend on the degree of nodes. In addition, the network is created step by step where at each step forces are computed, and nodes are displaced until they form the desirable shape. Therefore, the energy model consists of the following forces:

- Attraction and
- Repulsion

In an energy model, the forces are proportional to the distance between the interacting entities. The proportionality between distance and forces can be linear, exponential or logarithmic. For example the Spring model establishes a linear proportionality between the distance and the force and a square proportionality between the distance and the force.

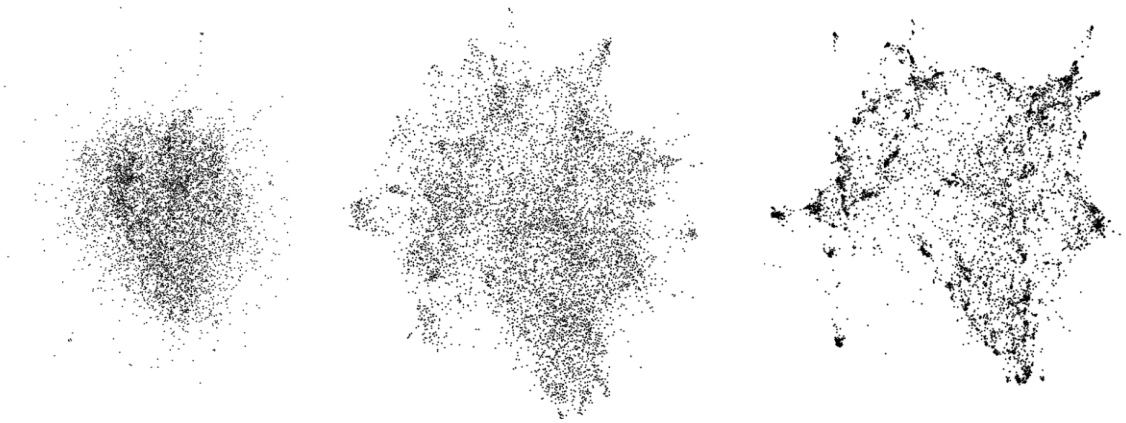


Figure 31: Layouts with Fruchterman-Rheingold

Repulsion force

The repulsion is applied among adjacent nodes in order to bring connected nodes closer to each other. The forces among the nodes depend on the degree centrality. The repulsion forces bring nodes with low degree next to the others with higher degree. In this way the user does not get confused by the overlapping nodes, as the nodes with high degrees will rearranged away from each other. The forces are calculated by the following formula:

$F_r(n1,n2) = k_r((deg(n1))(deg(n2)))/d(n1,n2)$, where k_r is a constant measure which is defined by the user or gets a default value.

By using that formula we do not take into consideration the nodes with degree equal to zero, as we know that the repulsion force will be equal to zero.

Attraction force

When simulating an energy model, apart from the repulsion, we must take into account the attraction forces, in order to create the final distribution of the nodes in the network. The attraction force Fa between two connected nodes n1 and n2 depends linearly on the distance d(n1, n2).

Fa(n1,n2) = d(n1,n2).

As a direct consequence, the closer the nodes are, the more they attract each other with higher forces than the rest of the nodes. The greater the distance is, the more reduced the attraction forces are.

Therefore, attraction and repulsion forces are used in order to create the network and set the nodes to the corresponding positions. Also, regarding the layout of the network it could be modified based on the parameters below, which could be defined real-time by using interactive user interface.

Iteration Step

The algorithm applies attraction and repulsion forces to the nodes, until the graph gets the final shape. Accordingly the user observes the creation of the graph and the iterations among the nodes, setting the final step of the process. Due to the creation of the graph the user observes the nodes without their edges. This functionality is provided by our system because edges might seem a little bit confusing to some users since most of the times they create a network not easily understandable.

LinLog Mode

We provide the user with the option to select between the linear or logarithmic energy models.

Logarithmic layout	Linear Layout
1st Stage	

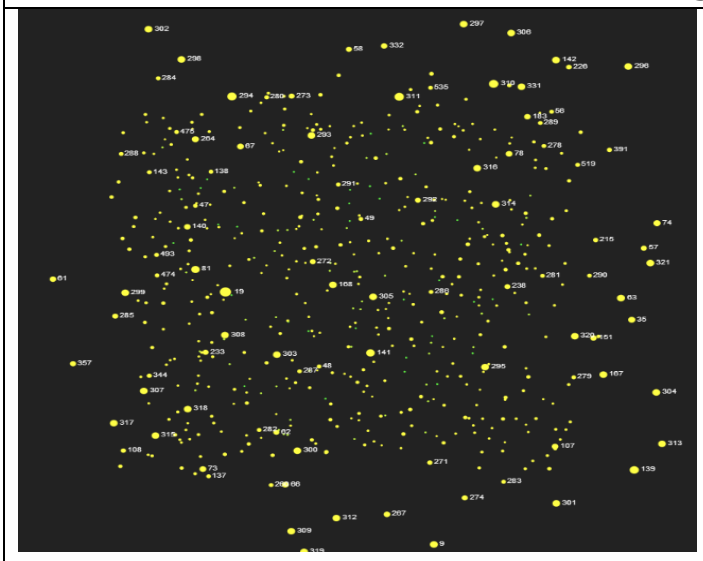
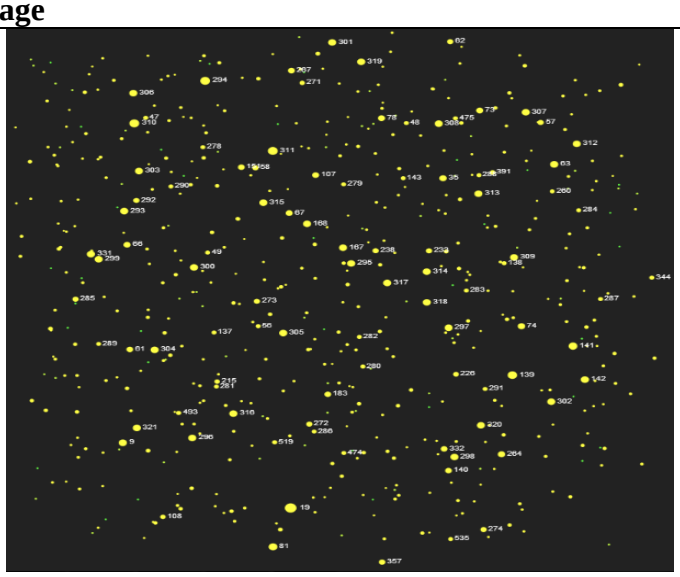

	

Figure 32: First stage of Logaritmic layout	Figure 33: First stage of Linear Layout
2nd Stage	

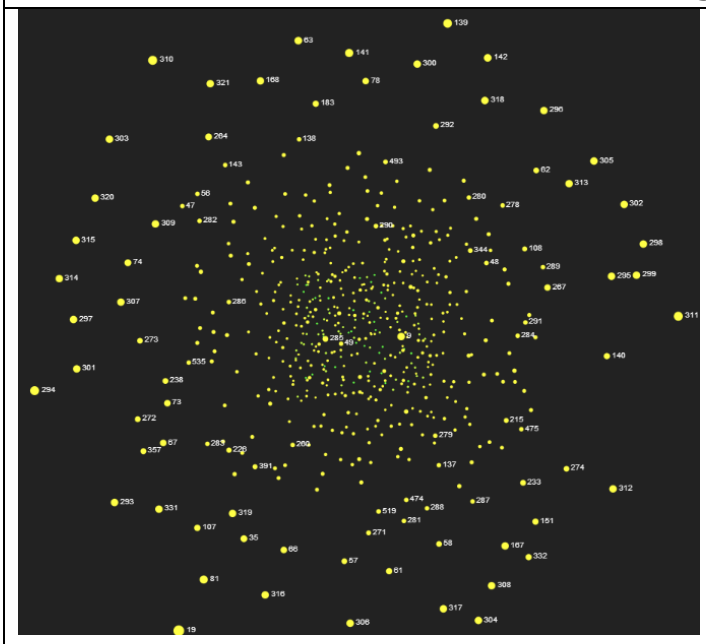
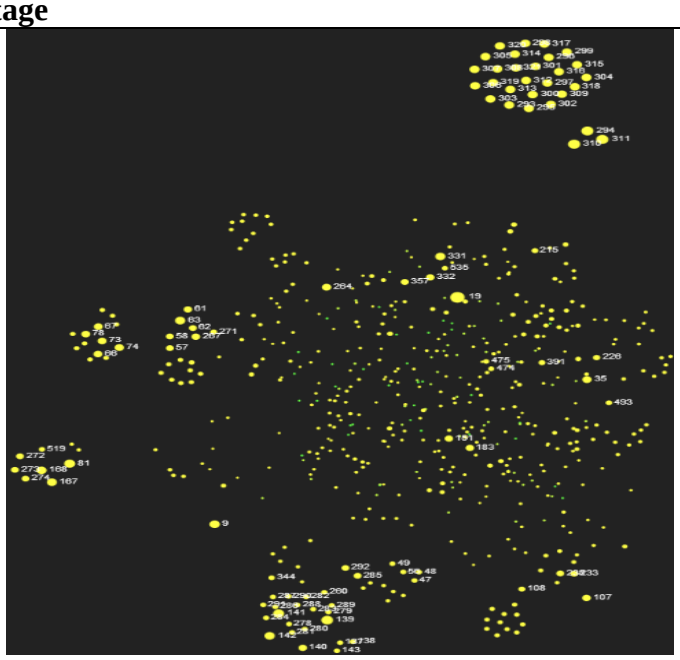

	

Figure 34: Second stage of Logarithmic layout	Figure 35: Second stage of Linear Layout
Final Layout	

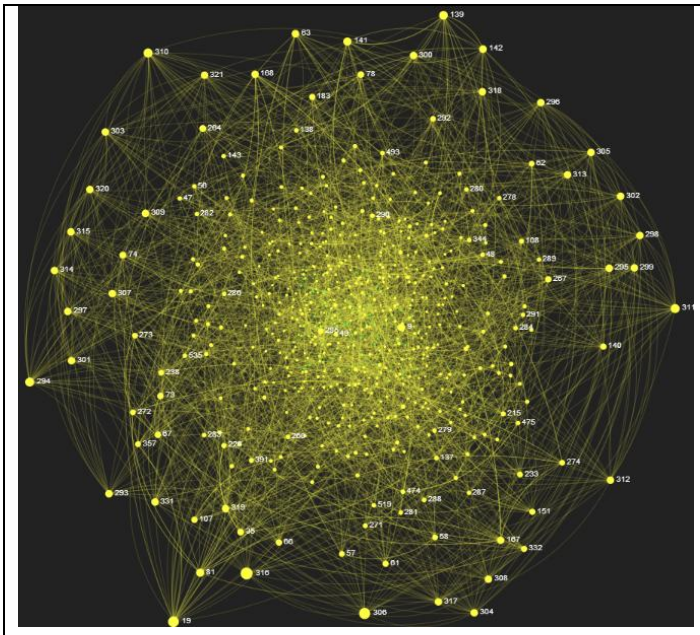


Figure 36: Third stage of Logarithmic layout

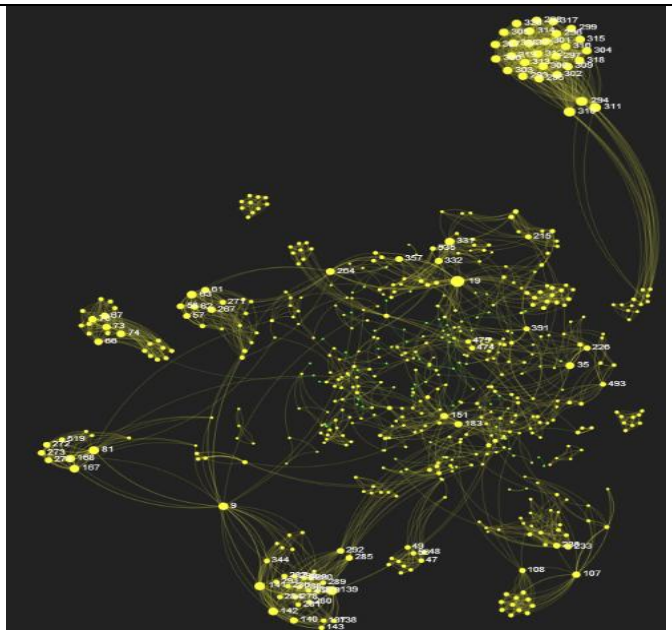


Figure 37: Third stage of Linear Layout

Gravity

The parameter of gravity influences the distribution of the nodes. In figures 32-37, we can see that in the graph with gravity=4 the nodes (disconnected and not) are distributed near to the center of the graph. But in the graph with gravity=1 is easiest to recognize the communities.

Gravity=1

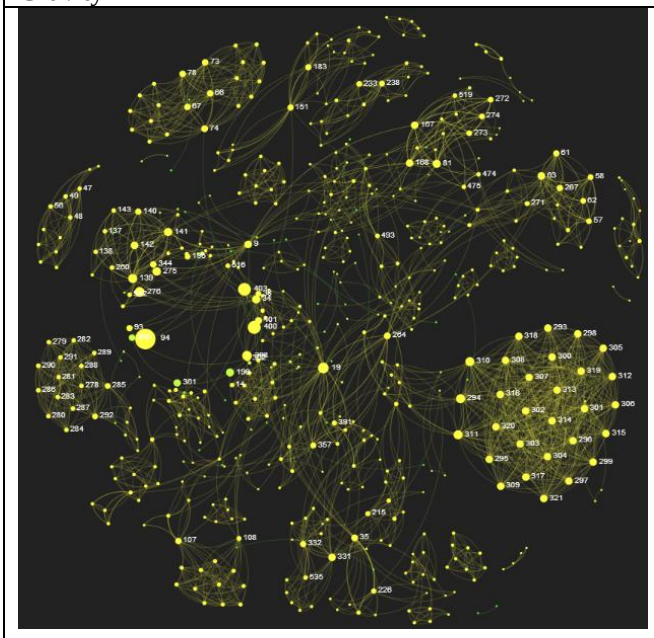


Figure 38: Graph with gravity=1

Gravity=4

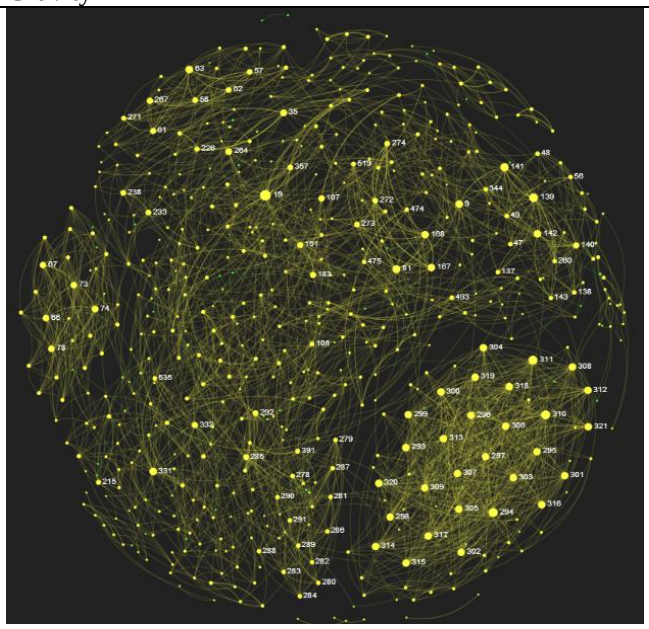


Figure 39: Graph with gravity=4

Edge weight

With this parameter the user could define whether he/she prefers to take into consideration the edge weights or not.

AdjustSizes

With this parameter the user could select whether he/she would like to adjust the sizes of the nodes.

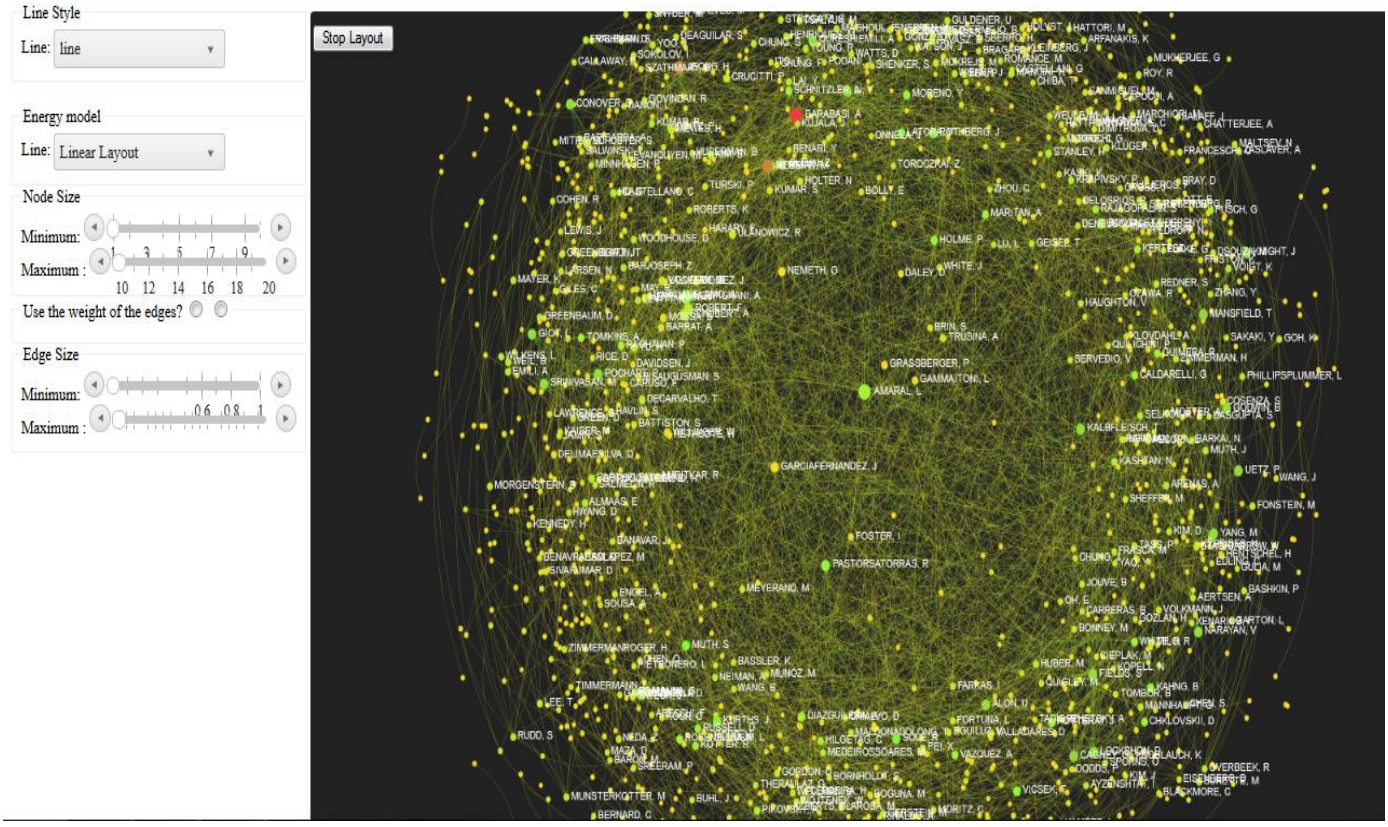


Figure 40: The interface for the modification of the graph layout

3.3. Research link Recommendation

3.3.1. Motivation

Link recommendation algorithms are used in order to predict future links among two nodes on the grounds of a social network. For example, supposing having the network described in figure 41, trying to predict the probabilities, where node A1 is applying a future link with the other nodes that it has not any direct connection yet.

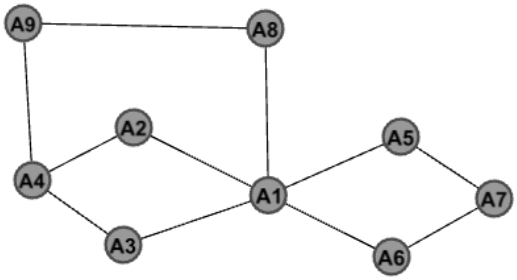


Figure 41: Graph Network

In a more detailed example we see in figure 41, node A1 could connect with A4 (via A3 or A2) or A7 (via A5 or A6) or A9 via A8. By using only the topological structure we could suggest that it is most likely to establish a connection with A4 or A7 than A9, as it has more common nodes. Supposing for example that this network describes a network among friends in facebook, and each one of the nodes has some extra characteristics as for example, the hometown, university graduation, interests, groups, music, e.t.c. If the link prediction in addition calculate those characteristics, then it may suggest that is most likely to establish a connection with node A9 than with the other two, as node A9 and node A1 share the same interest of music and live in the same city, in contrast with A4 where they do not share any common characteristic, despite the fact that they are connected by different nodes. Based on the previous example, the prediction of a future link is influenced not only by the topological structure, but also by the features of the nodes.

Link prediction algorithms are separated in two main categories:

- Those which search the entire network and display all the paths or shortest paths among the nodes and
- Those which have a certain limit in the length of the path (l) and the common neighbors are displayed only for the specified length.

The second approach is considered to be better because of the time and space complexity in contrast with the first one, where the entire network is being searched. On the other hand, the limit in the length of the paths may cause problems in the identification of the ideal neighbor as it may cover a neighbor that is one hoop along the source node and it has more different paths to the source (increase the probabilities) than the nearest one. Based on our previous remarks, the link prediction is a complex task that needs to examine a lot of parameters in order to suggest the most preferable connection in the future. Especially in a Higher Educational Institute, where intensive networks exist among the faculty members, it is much more difficult to predict new links that are out of the "close" network.

Considering all the above, we introduce the research link recommendation algorithm, which takes into account the paths of a specific length among the scientists, their research areas and their performance as members of research teams. At those data, we apply the score() function, in order to find all the alternative paths for a specific number of length and the Bayesian networks to calculate the conditional probabilities among the research areas where they participate.

Our contribution in the link prediction algorithm is based on the assumption that link prediction among researchers is influenced not only by the number of paths among them, but also from the characteristics of their publications.

3.3.2. Research Link Functions

In order to predict the future links among the nodes we propose an algorithm which uses:

- a score() function to measure the ratio of the sum of the "walks" among two nodes to the sum of all the walks, if we suppose that there exist links among all the nodes.
- a bayesian() function, which calculates the conditional probabilities among the research areas.

The score() function.

For a Graph G = (V, E) with nodes (vertices) V and edges E, suppose that we have two nodes v_i and v_j. The score (v_i, v_j) expresses the probability for the nodes to establish a connection. The range of the values is [0,1]. If the value is near to 1 then the nodes are more likely to establish a connection. This way we take into account all the paths, and not only the short path among two nodes, where the short path represents the shorter path among the source and the target node. But this path may not be the ideal as the intermediate nodes may not have any probability to connect based on their research interest. In contrast with other paths that may not be shorter than the "short path", which could contain nodes that are more likely to collaborate to each other. If two nodes are connected with more than one path then the probabilities to establish a connection are increased. For example, if we examine all the paths (with length<=3) for node V1(figure 41) we could see that there exists:

- 1 path with V2 (V1-V3-V4-V2)
- 1 path with V3 (V1-V2-V4-V3)
- 3 paths with V4 (V1-V3-V4, V1-V2-V4, V1-V8-V9-V4)
- none path with V5
- none path with V6

- 2 paths with V7 (V1-V5-V7, V1-V6-V7)
- none path with V8
- 3 paths with V9 (V1-V8-V9, V1-V2-V4-V9, V1-V3-V4-V9)

Therefore, if we check only paths of length 2 then V4 and V7 have the same probabilities to be connected. But if we check paths till length to be equal with 3, then V4 is the predominant. In our approach suggest the following score for measuring the probability to establish a connection among two nodes.

Definition: The probability among two nodes Vx, Vy is measured by the function score:

$$(Vx,Vy)= \sum_{i=2}^{\lambda} \frac{1}{\log(i)} \frac{paths^{i}_{Ux,Uy}}{\prod_{j=2}^{i}(n-j)}$$

where

- n is the number of the nodes
- λ is the length of the path
- $\frac{1}{\log(i)}$ is the attenuation factor
- $paths^{i}_{Ux,Uy}$ is all the paths among Ux, Uy
- $\prod_{j=2}^{i}(n-j)$ is the amount of all the paths , suppose that all the nodes are connected each other.

The bayesian() function.

The Bayesian function is used in order to calculate the conditional probabilities among the examined attributes. The Bayesian network classifier can be defined as

$c(E) = \arg\max_{c \in C} P(c)P(a_1, a_2, ..., a_n \mid C)$, where ($a_1$, a_2… a_i, …, a_n) are the attribute values of an instance E_i.

Variable C is represented as the top node in a Bayesian network. After the execution of the Bayesian function we get the probabilities among the examined research areas, which provide another important measure in the selection of the most prominent researcher.

3.3.3. Research Link Algorithm

The research link algorithm is based on the co-authoring networks and the research areas of their publications. In this section we will describe in details the algorithm using as example the graph described in figure 41. The algorithm consists of the following stages:

- First, we calculate the score (Ux,Uy) among all the authors of the network
- Then we calculate the bayesian probabilities of their research areas
- Finally, we predict the most prominent author

Step1: Initially we calculate the score() among all the nodes of the graph, using as input the nodes A1-A9 of graph G in figure 41. In order to measure the score we create the adjacent matrix (table 24) on the graph G

	A1	A2	A3	A4	A5	A6	A7	A8	A9
A1	0	1	1	0	1	1	0	1	0
A2	1	0	0	1	0	0	0	0	0
A3	1	0	0	1	0	0	0	0	0
A4	0	1	1	0	0	0	0	0	1
A5	1	0	0	0	0	0	1	0	0
A6	1	0	0	0	0	0	1	0	0
A7	0	0	0	0	1	1	0	0	0
A8	1	0	0	0	0	0	0	0	1
A9	0	0	0	1	0	0	0	1	0

Table 24: Matrix of Graph G^1

The matrices can be multiplied by themselves (N times) and the result of that calculation will represent the number of paths of length N from one node to another. Taking into consideration that theorem, if we manipulate a matrix in the Nth power, then the result matrix will show the amount of paths of length N that exist to the graph. Therefore if we want to examine the paths of length 2 we apply the G^2 calculation. Also we apply a cycle detection algorithm in order to remove the paths which include cycles.

	A1	A2	A3	A4	A5	A6	A7	A8	A9
A1	5	0	0	2	0	0	2	0	1
A2	0	2	2	0	1	1	0	1	1
A3	0	2	2	0	1	1	0	1	1
A4	2	0	0	3	0	0	0	1	0
A5	0	1	1	0	2	2	0	1	0
A6	0	1	1	0	2	2	0	1	0
A7	2	0	0	0	0	0	2	0	0
A8	0	1	1	1	1	1	0	2	0
A9	1	1	1	0	0	0	0	0	2

Table 25: Matrix of Graph G^2

As we can see in table 25 all the paths of the graph with length=2 are created. For example the node A1 has more probabilities to establish collaboration with A4 or A7 because of the fact that these are connected by two different paths.

	A1	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**
A1	0	0	0	A1-A2-A4 A1-A3-A4	0	0	A1-A5-A7 A1-A6-A7	0	A1-A8-A9

Table 26 : Paths with length equal to 2 from the node A1 to the other nodes.

After having calculated the paths we can measure the scores for all the nodes. The score for node A1 is calculated based on the fact that it has a two-length path. For example the score for node A1 and A4 is equal to: score $(A1,A4)= 2\,(\text{number of edges})/7\,(\text{number of total edges}) = 0{,}286 \times \frac{1}{\log(2)} = 0{,}086.$ Because there exist two paths of the same length among A1-A4, then the final score of A1, A4 will be score(A1,A4)=0,17 = 0,086*2. The final scores for the nodes are displayed in table 27

	A1	A2	A3	A4	A5	A6	A7	A8	A9
A1	0,000	0,000	0,000	0,172	0,000	0,000	0,172	0,000	0,086
A2	0,000	0,000	0,172	0,000	0,086	0,086	0,000	0,086	0,086
A3	0,000	0,172	0,000	0,000	0,086	0,086	0,000	0,086	0,086
A4	0,172	0,000	0,000	0,000	0,000	0,000	0,000	0,086	0,000
A5	0,000	0,086	0,086	0,000	0,000	0,172	0,000	0,086	0,000
A6	0,000	0,086	0,086	0,000	0,172	0,000	0,000	0,086	0,000
A7	0,172	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
A8	0,000	0,086	0,086	0,086	0,086	0,086	0,000	0,000	0,000
A9	0,086	0,086	0,086	0,000	0,000	0,000	0,000	0,000	0,000

Table 27: Calculation of scores for Graph G

Apart from the prediction of the future links we considered as useful to extract forecasting reports about future research activities based on the estimation of present research outcomes so we call the

Bayesian() function, where nodes represent attributes, and edges represent attribute dependencies. For example in our system the user can select the research areas and then he/she can observe the conditional probabilities for each node (research area). After the execution of the bayesian() function the user can identify the conditional probabilities among the research areas of the authors(nodes). For example node A4 has probabilities to establish a link with A1 and A8. The research areas of these nodes interests are: social science, computer science and engineering. The conditional probabilities are the following:

Social Science			
Engineering	Computer Science		
0	0	0.166	0.833
0	1	0.75	0.25
1	0	0.722	0.277
1	1	0.658	0.342

Table 28: Bayesian for Social Science, Engineering and Computer Science

The **ResearchLink** algorithm has the following steps:

1. The user selects the source target (Ux)
2. The score among Ux and the other nodes is calculated.
3. The bayesian probabilities for the research areas are calculated.
4. The final scores are displayed.

The algorithm is defined as follow:

```
Algorithm ResearchLink (A,n,ℓ,V<id,R,flag>)
Input
A: adjacent matrix of Graph,
n: the degree of the Graph
ℓ: the length of the path,
V<id,R,flag>: id is author's id, R is the research areas and flag takes values 0/1 if
              the author has or not publication at an area,
Output
score(i, j): the final score among the nodes i,j to establish a connection

Main Program
for m = 2 to ℓ
        Compute Paths(A,m)
        Compute Baysian(V<id,R,flag>)
end for m
End Main Program


Function Paths()
 for i = 1 to n {
        for j = 1 to n {
```

```
        d = 1
                 for k = 2 to m{
                 d= d * (n - k)
                 }
                if
                {
                (cycle_detection(path(i,j))
                                                    1      paths^i_{i,j}
                paths(i,j)= paths(i,j)+             ------ ------
                                                    log(m)   d
                }
         }
 }
 return paths(i, j)
End Function
```

Table 29 : ResearchLink Algorithm

```
Function Bayesian(V<id,R,flag>)
size=count(id)
for i = 1 to size {
execute bayesian (V<id,R,flag>)
 }
 return score(V<R,i>) //where R the research areas and i the probabilities.
End Function

Function cycle_detection(Path p){
for (Node v1:p.getVectrices)
  v1 = startNode;
        if (v1 != v1.next());
        return true;
        }else return false;
}
```

Table 30 : ResearchLink Functions

4. Institutional REsearch Management(IREMA)

In this section we present in detail the development of the Institutional REsearch MAnagement framework, a system built to evaluate the research and development activities among the faculty members of an institute. The framework is validated based on hypothesis queries using the interactive visual interfaces.

4.1. System Implementation - Evaluation

4.1.1. Motivation

An implementation of a system called IREMA that implements the aforementioned concepts has been completed for the needs of the Visual Analytics project. The idea was to create a generic architecture for the design and implementation of interactive decision support systems that enables research Knowledge Discovery and supports Data Visualization (KDD-V). A tool is presented that provides to the user a variety of different ways to explore data in order to extract knowledge. The functionalities of that tool allow the user to perform analysis, to explore the data through interactive visualizations and also to provide additional details depending on the nature of the query.

The development of the IREMA started with the specification of the requirements [93]. Upon the completion of this stage, we were able to design the system taking into consideration all the current workflows and by creating associations that would enable the extraction of new patterns and that would enable recommendations for future activities. The analysis consists of two steps:

Step 1: The initial step, before the implementation, is to understand the "problems" and set the aims of "What system are we looking to create?" which could be described from the questions below.

What is the context of use of visualizations?

It will be used for institutional research management.

In which data analysis processes should the visualization tool be used?

It will be used to analyze the co-authoring network among the faculty members of an institute, to explore the characteristics of their collaboration, to measure the efficiency among the departments and finally to make the data available to the user in such way that will be used to find hidden patterns among research collaborations between the faculty members.

Who will use it and what are the user characteristics?

It will be used by experts on the specific domain, but it will be also used by non experts without any previous knowledge on data mining and related concepts.

What data will be used?

The data input will be research papers and research projects related data.

Which kind of visualizations will be familiar to the user?

They will be familiar with pie and charts.

What challenges and usage barriers can we see for a visualization tool?

Using the data visualization module the users will create for all the data appropriate visual representations without any personal intervention. A limitation of such a representation is the readability of the results since the users are provided with an image that was created automatically without their participation in the calculations.

Step 2: The next step is the overview of the different layers of the system and the definition of the methodologies that we are going to apply. These layers can be distinguished as:

Data Overview: The data is displayed using several visualization representations depending on the user's selection. The available data visualizations are graph networks, the map of science, the parallel coordinators and the regression or efficiency lines.

Analysis. The tool includes a variety of methods to analyze the data, which could be used alone or complementary to each other, so as the results of one could be input for the other and vice versa. The methods that the user could use are: social network analysis, bayesian networks, tree classification, k-means clustering, Data Envelopment Analysis and association rules. The user has the ability to define the parameters or the weights of the parameters.

User interaction. All of the results are displayed using visual representations. The user can explore the data, zoom in-out and change the parameters of the displayed visualization (color, opacity). The user could set weights on the parameters of data mining; he/she could also interact with all the available visualization methods until he/she gets the desired result.

Hypothesis Validation. The tool makes possible to validate hypothesis based queries using the interactive visual interfaces.

Data source selection and integration. The data to be analysed can be collected from several data sources as web services, xls, csv or RDBMS.

4.1.2. An overview of IREMA framework

The IREMA framework (figure 42) is a web based system built on Java technologies that supports institutional research management. The system enables multiple users (end - advanced users, developers) to work on a common shared ontology [80].

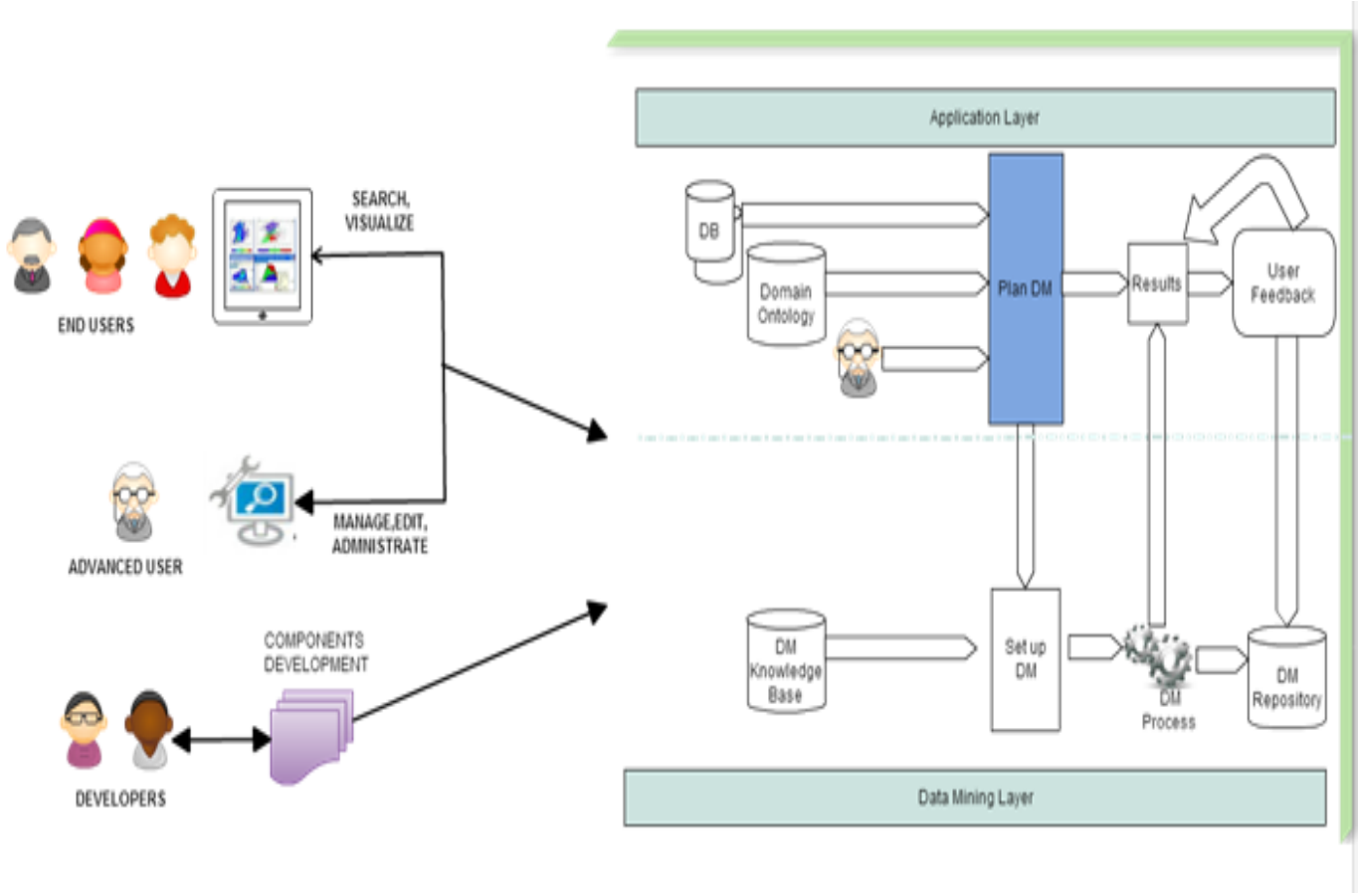


Figure 42: IREMA Architecture

The advanced users (RPMs) are responsible for: a) setting the main parameters (criteria) of the decision making process, b) modifying the set of the evaluation criteria (by adding or deleting criteria) and c) setting the criteria weights or the type of the corresponding data mining process, and the associated parameters. Developers on the other hand, have full access to all features of the multi-criteria evaluation

process; they can add data mining processes, or develop custom functions based on advanced users needs and also expand the ontology. Except from advanced users and developers, the DSS includes a user-friendly interface that facilitates the preparation of several reports in graphical and tabular format.

The following subsections describe the capabilities that the framework provides and the inner architecture through which our system provides decision support mechanisms. In our implementation, we have selected SNA in combination with data envelopment analysis, as a method for efficiency measurement and we combine it with data mining techniques as a means for knowledge extraction. Comparing the R&D outcomes of academic units with the dynamics of the collaboration patterns extracted from graphs, the developed system enables RPM to evaluate specific criteria and correlate strategic goals with research performance. In figure 43 we can see the layers of our decision support system and in the bottom the methods of the visual analytic process (described in figure 22). The IREMA consists of the Data Collection, Data preparation, Data mining and Interactive Knowledge Discovery that will be described in details in the next subsections.

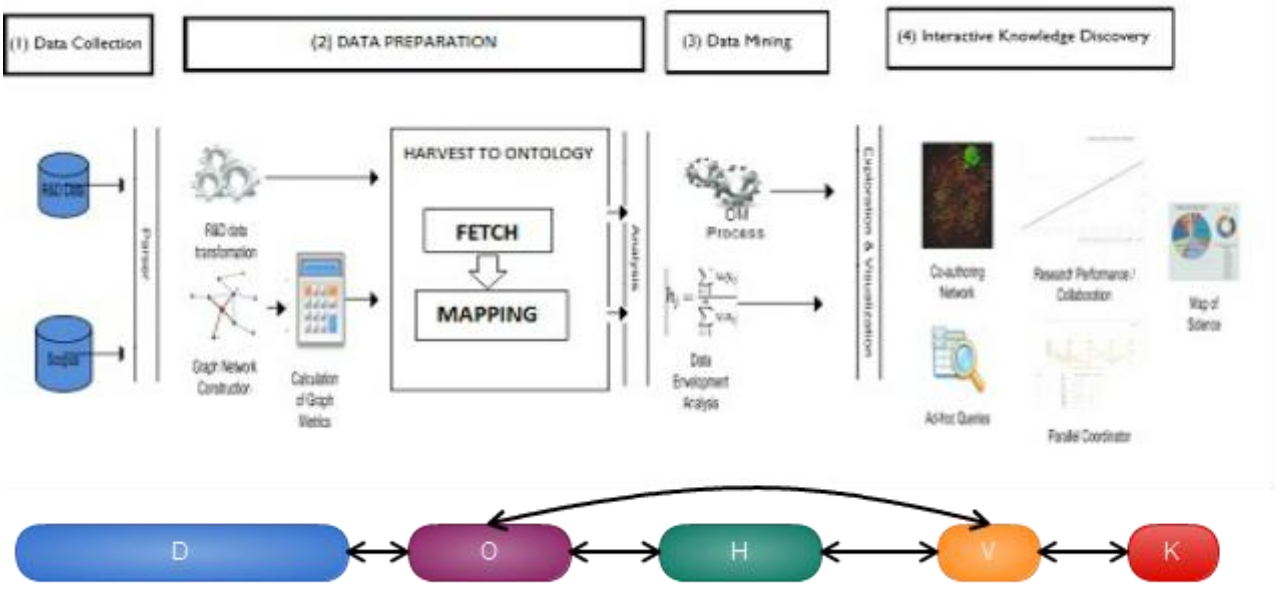


Figure 43: IREMA Layers

4.1.2.1. Data Collection

Data collection is the initial operation of our system. The data that are being collected are separated into two categories: co-authoring and R&D data. These data are collected either online by using a web Service, either by uploading a .csv file, or finally by connecting to RDBMS. In order to describe the

functionality of data processing we present the work flow that the user should follow in order to store the coauthoring data from the Scopus bibliographical database. The steps in this stage are described on the UML activity diagram on figure 44. Initially, the user imports a csv file from SCOPUS which includes the authors, the papers and the citation number; next, a list which contains the faculty members by department is loaded, which enables to filter out those who came from other institutes or departments. Then, the research areas for each one of the faculty members are inserted in the system. Regarding the R&D data we collect data based on the criteria set by the European Association for Quality Assurance in Higher Education – the Hellenic Quality Assurance Agency.

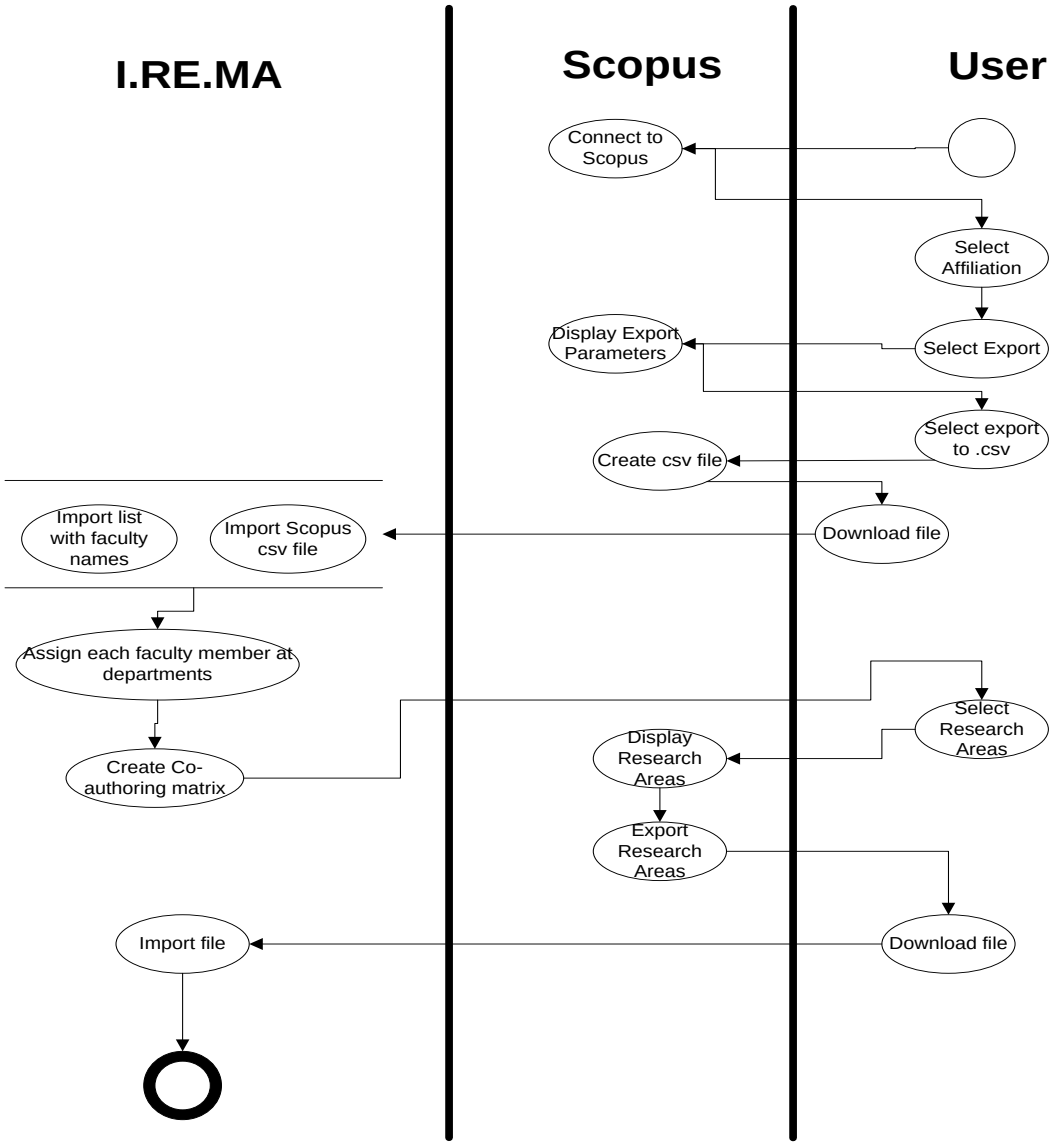


Figure 44: Data Collection

4.1.2.2. Data Preparation

The system is designed to represent the co-authoring networks, where nodes represent the authors and the edges the collaboration (co-authoring activity) among them. In order to build this, for all the faculty members we construct the co-authoring matrix and we calculate the graph metrics for each one of them, based on the network topology. In order to enable semantically enhanced relationships in the composed graph we make use of an appropriate ontology. An ontology in general is a structure <D, W, R >, where D is a domain and W is a set of maximal states of affairs of such domain (also called possible worlds). For instance, D may be a set of blocks on a table and W can be the set of all possible spatial arrangements of these blocks. Similarly R is a set of relevant relations on D. The domain space in our case is a set of research-related concepts for a Higher Education Institution. In our case we use the ontology described in section 6.3, which has been reconstructed and adapted so as to be aligned with the profile of an academic unit.

In order to map the evaluation data to the ontology we have implemented a process that enables the user to assign entities of the database to the terms of the ontology. In more detail the mapping is performed as a two-step process: First the data need to be fetched from the data sources where they are kept either as raw data, either in relational form. Then, through the mapping process the elements of the database are translated to the particular domain ontology. The transformation between the relational schema to the XML ontology is performed using the D2R [94] which is a tool that enables mappings of relational structures to OWL/RDFS ontologies.

4.1.2.3. Data Mining

Data mining is a technique for the extraction of hidden predictive information from large databases and it is considered as a sub-process, within the overall KDD process. Our system provides support for a variety of clustering and classification methods; more specifically we provide different choices to the user, who is able to select from k-means clustering, a-priori association rule mining, tree classification and Bayesian networks. We provide in the following a brief explanation of each different operation:

- **The k-means clustering:**

K-Means is an algorithm for creating clusters (k) of n objects in which each object belongs to the cluster with the nearest mean. The main difficulty in this approach is the selection of the number of clusters which is not straightforward. It still remains to the user either to select by intuition either by repetitions until an optimal k has been found. We have selected to define the optimal number of clusters by estimating

it as the one that minimizes the merging cost for these clusters. An algorithm that helps estimate the merging cost and within cluster distance is Ward's algorithm [95].

Ward's method considers the distance between two clusters A and B by calculating the sum of squares when we merge them. The merging cost D of the two clusters is given by

$$D(A,B)= \sum_{i\in A\cup B}\| \vec{x_i} - \overrightarrow{m_{A\cup B}} \|^2 - \sum_{i\in A}\| \vec{x_i} - \overrightarrow{m_A} \|^2 - \sum_{i\in B}\| \vec{x_i} - \overrightarrow{m_B} \|^2 = \frac{n_A n_B}{n_A + n_B} \| \overrightarrow{m_A} - \overrightarrow{m_B} \|^2$$

where $\overrightarrow{m_j}$,is the center of cluster j, nj is the number of points in it.

Ward's method is both greedy, and constrained by previous choices as to which clusters to form. Therefore, the sum-of-squares for a given number k of clusters is usually larger than the minimum for that k, and even larger than what k-means will achieve.

After having estimated the optimal number of clusters, the steps that we follow according to the k-means algorithm [96] in order to determinate the clusters of data are the following:

- Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
- Assign each object to the group that has the closest centroid.
- When all objects have been assigned, recalculate the positions of the K centroids.
- Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

- **<u>Performing rule-based mining on evaluation data.</u>**

We enable the user also to extract associations or casual structures from the data using rule mining techniques. For this purpose we have implemented the Apriori association rule mining technique [97], which is useful for discovering interesting relationships hidden in large data sets. An association rule is an implication expression of the form $X \rightarrow Y$, where X and Y are disjoint item-sets, i.e., $X \cap Y = \emptyset$. In order to measure the strength of an association rule the terms of support and confidence are used. Support determines how often a rule is applicable to a given data set, while confidence determines how frequently items in Y appear in transactions that contain X.

The formal definitions are:

Support: $$c(X \rightarrow Y) = \frac{\sigma(\mathrm{X} \cup Y)}{N}$$

Confidence: $$c(X \rightarrow Y) = \frac{\sigma(\mathrm{X} \cup Y)}{\sigma(\mathrm{X})}$$

Support is an indication of how frequently the items appear in the data-set. Confidence indicates the number of times the if/then statements have been found to be true. In our case the association rules enable the user to analyze present research activities in specific research areas and also to forecast future expectations.

- **Efficiency Measure**

Research publications are used to shape the co-authoring network of contributing to institutional research faculty members. Apart from revealing the existing collaboration patterns, co-authoring networks can also be used as a measure of research performance. In our approach, and in order to measure the performance of R&D activities, we use criteria specified by the Hellenic Quality Assurance Agency (HQAA). For quantifying the R&D productivity we use DEA and calculate the ratio of output(s) to input(s). The performance of a Decision Making Unit (DMU) is calculated by comparing its efficiency, with the best observed performance in the data set.

- **Social Network Analysis**

In the previous stage (data processing), we have calculated the graph metrics, such as betweeness, closeness, eccentricity, clustering co-efficient and eigenvector. Those metrics measure the importance of the nodes and also we calculate the link analysis algorithms (PageRank, HITS), with the purpose of "measuring" the relative importance of specific nodes within a specific set.

- **Research link Recommendation**

The research link recommendation method provides the most likely future links among the nodes. The algorithm was described in details in subsection "Research link Recommendation".

4.1.2.4. Interactive Knowledge Discovery

The proposed framework integrates interactive visual interfaces to support Knowledge Discovery (KD), thus providing the user with enhanced assistance throughout the decision making (DM) process. The IREMA supports the following visual representation techniques:

- Co-authoring Graph, which is created on the basis of the collaboration among faculty members for the publication of a research paper
- Parallel Coordinators, as an interactive representation where the user is able to apply a set of criteria (dynamic) depending on his objectives
- Efficiency Line, which is used to represent the correlation among indicators
- Map of Science, where each of the research areas is represented through pie charts

- **Co-authoring Graph**

A co-authoring (or co-authorship) network is the network created on the basis of the collaboration among the faculty members for the publication of a research paper. Figure 45 illustrates such a co-authoring network, where nodes represent the authors and edges represent the collaboration (co-authoring activity) among them. Each of the edges has been assigned with a value denoting the number of citations for the paper. The diameter of the nodes and their color depends on the number of publications the authors have submitted. In our implementation the user may select a node and view only the authors that have links to the selected author. The appearance of the graph is based on the criteria (graph metrics) that the user has selected. Different criteria provide different graph networks, thus resulting in different nodes' colour and size.

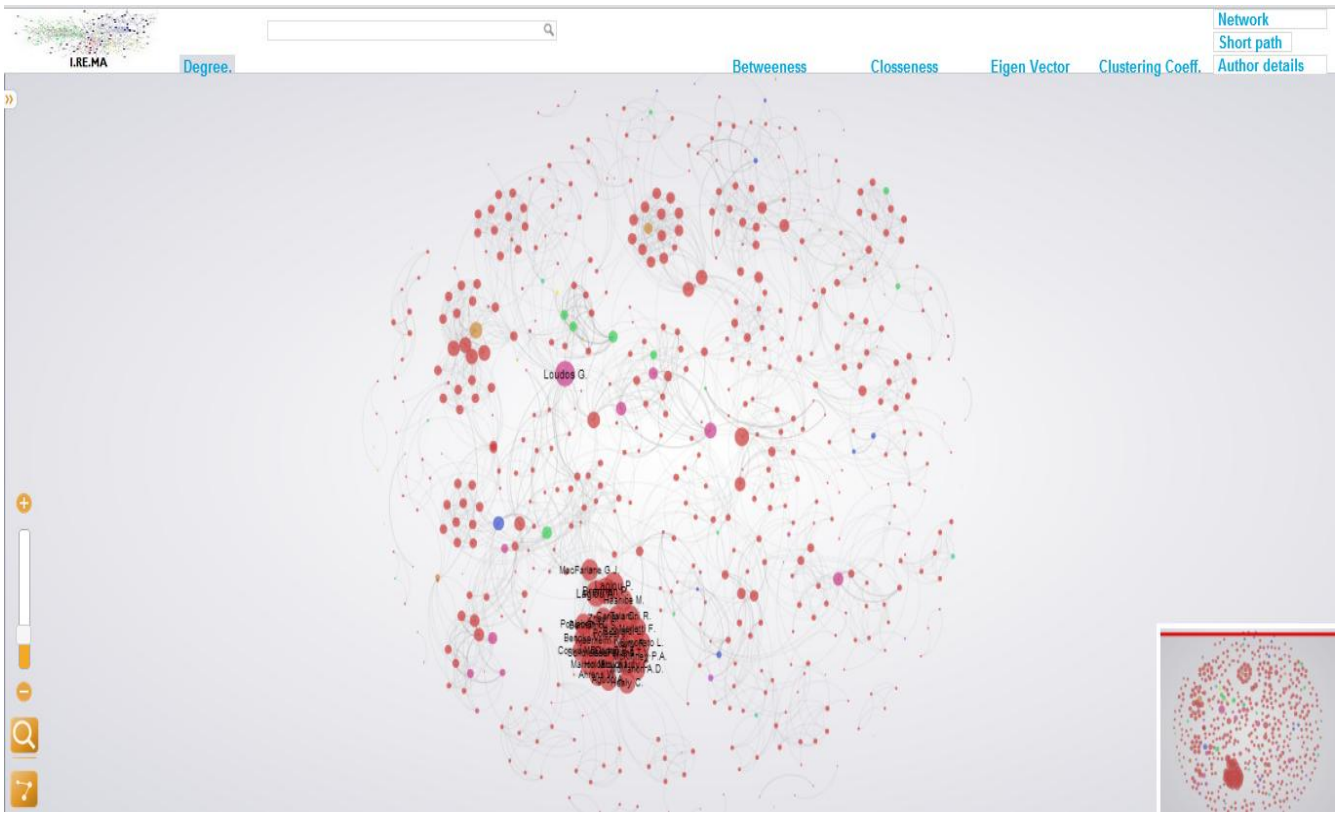


Figure 45: Co-authoring Graph

- **Parallel Coordinators visualisation**

Parallel Coordinate plots have been shown to be useful in Exploratory Data Analysis, especially when they are implemented in interactive software. Exploratory Data Analysis (EDA) refers to methods and procedures for exploring the data space to learn about a data set. Interactive graphics are excellent for EDA. They are designed for exploring rather than presenting information.

Parallel coordinates display multi-dimensional data in a two-dimensional space: One for axes, and the other for data lines. The parallel coordinator system was developed using the d3js, JavaScript library for manipulating data using HTML, SVG and CSS. In our system, JavaScript functions pulled data from the server using asynchronous HTTP Requests. Also we use SVG elements which display rendering positions which consist of three sets of drawing primitives: Axes, selected lines, and non-selected lines. The axes data structure encodes the display location of each axis and the minimum/maximum values on that axis. Finally, brushing of data using a mouse-controlled ”highlight” line is used for selection. The user may interactively set the efficiency criteria for the corresponding data and get visual comparative results for the faculty members that fall within a given value range. The parallel coordinator visualisation method provides the extraction of that knowledge.

One of the most common research evaluation criteria is to provide the most qualified researcher among the faculty members. Using the parallel coordinator visualization method the user is able to apply a set of criteria depending on his objectives. Figure 46 illustrates the parallel coordinator visualization which uses, as case study, the faculty members of a Department of Informatics of a Greek Higher Education Institution. The parallel lines represent criteria such as for example the h-index, the position held (Professor, Assistant Professor,..), the projects, the number of publications and the research field of the faculty members. Using this representation the user has the ability to set filters manually and select only those data with values within specific, desired ranges. Hence, in figure 46 the user may view all the 25 members of the department, which are considered as authors of research papers. Figure 47 shows the application results on the basis of the h- index criterion and, specifically, the authors with h-index among 4 and 9.

Using parallel coordinators, the user could set priorities by selecting the range of values for the examined fields. Therefore, the user defines the priorities by observing the overall research performance of all the faculty members. We thought that it is of high importance for a decision maker to take decisions in that way, because at any time he could change the ranges of the values or the order of the filters so as to get different results and different decisions.

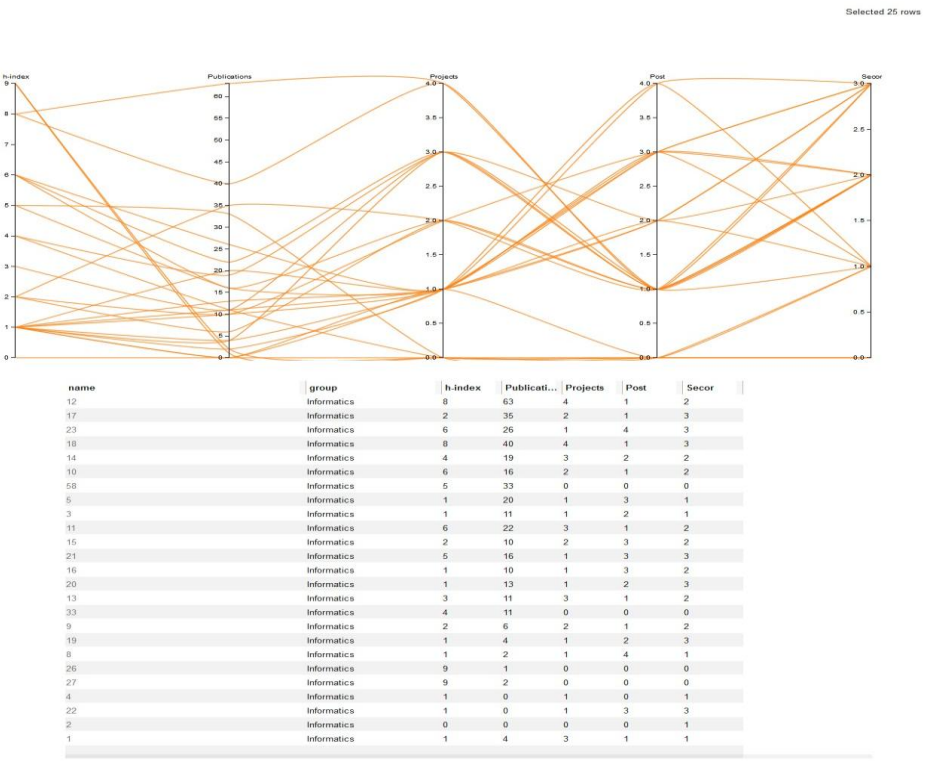
The types of queries that our system supports are:

Linear

e.g. “Find all the authors with their publications”

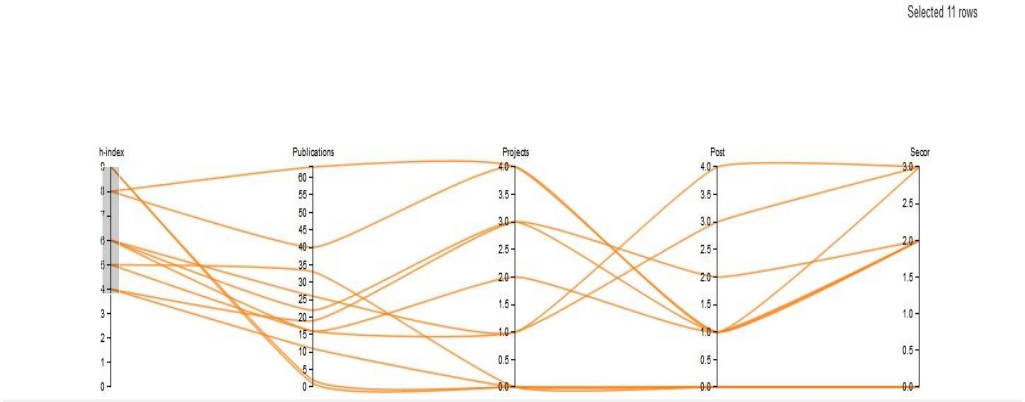
Filtering

e.g. “Find all the authors who have more than 10 publications”



name	group	h-index	Publicati...	Projects	Post	Secor
12	Informatics	8	63	4	1	2
17	Informatics	2	35	2	1	3
23	Informatics	6	26	1	4	3
18	Informatics	8	40	4	1	3
14	Informatics	4	19	3	2	2
10	Informatics	6	16	2	1	2
58	Informatics	5	33	0	0	0
5	Informatics	1	20	1	3	1
3	Informatics	1	11	1	2	1
11	Informatics	6	22	3	1	2
16	Informatics	2	10	2	3	2
21	Informatics	5	16	1	3	3
16	Informatics	1	10	1	3	2
20	Informatics	1	13	1	2	3
13	Informatics	3	11	3	1	2
33	Informatics	4	11	0	0	0
9	Informatics	2	6	2	1	2
19	Informatics	1	4	1	2	3
8	Informatics	1	2	1	4	1
26	Informatics	9	1	0	0	0
27	Informatics	9	2	0	0	0
4	Informatics	1	0	1	0	1
22	Informatics	1	0	1	3	3
2	Informatics	0	0	0	0	1
1	Informatics	1	4	3	1	1

Figure 46: Parallel coordinator for all the faculty members without any criteria



name	group	h-index	Publicati...	Projects	Post	Secor
12	Informatics	8	63	4	1	2
23	Informatics	6	26	1	4	3
18	Informatics	8	40	4	1	3
14	Informatics	4	19	3	2	2
10	Informatics	6	16	2	1	2
58	Informatics	5	33	0	0	0
11	Informatics	6	22	3	1	2
21	Informatics	5	16	1	3	3
33	Informatics	4	11	0	0	0

Figure 47: Show all the faculty members with h-index value besides 4-9

- **The map of science**

In order to represent the map of science we have developed an interactive pie representation (figure 48). The pie represents all the research areas that faculty members have published their research papers in.

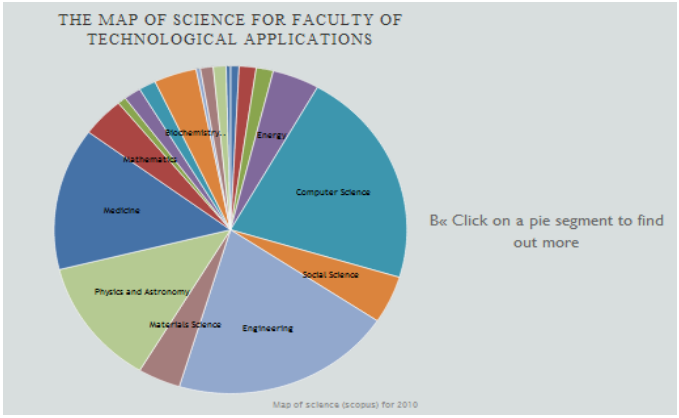


Figure 48: Map of science

When the user clicks at a certain area, a second pie opens at the right side of the screen which informs about the authors that participate in this certain area, the number of papers, the authors' id and the department to which they belong.

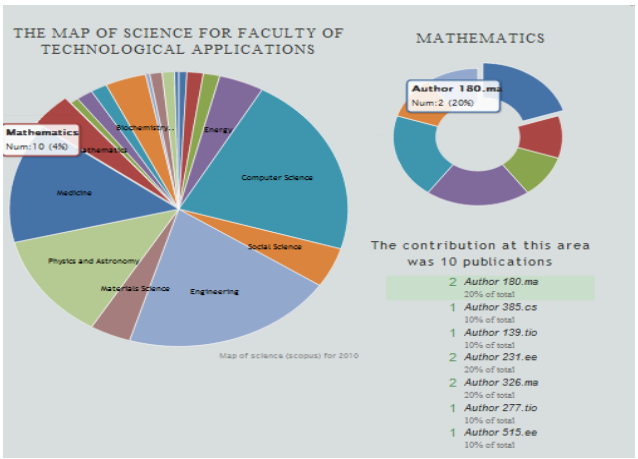


Figure 49: Map of science with selected pie

- **The regression line.**

The regression is selected in order to represent the correlation among selected indicators,

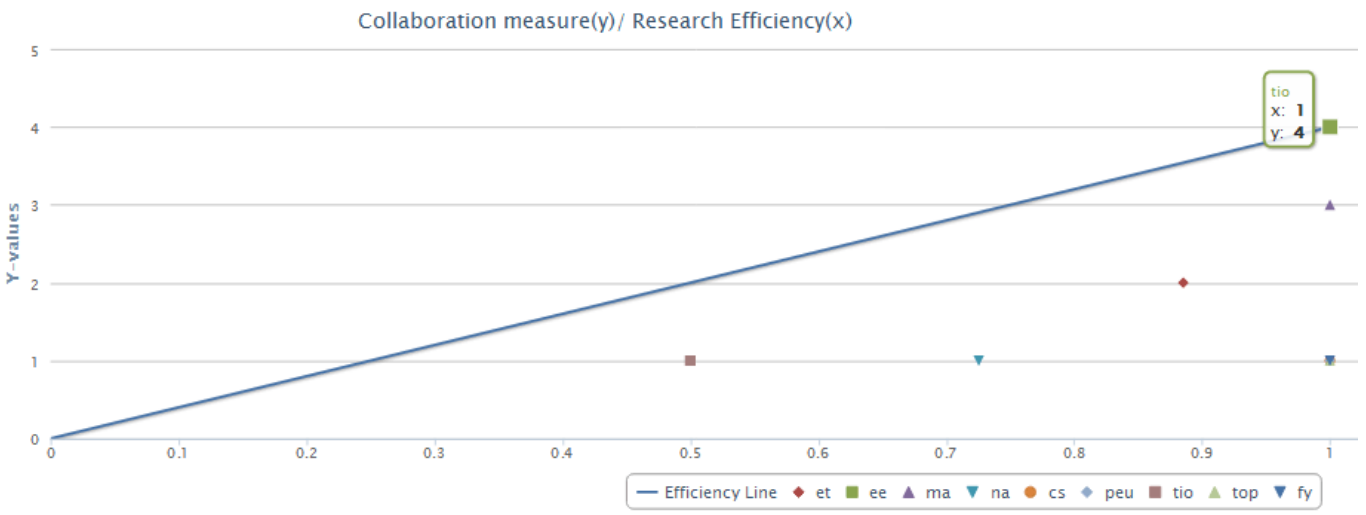


Figure 50: Regression Line

Based on the principles of the visual analytics process, we conclude that the developed representations have the following characteristics:

- They are interactive
- The results of the visualizations could be adjusted according to user preferences (colour, opacity, zoom,..)
- They can be used to extract knowledge.

4.1.3. IREMA Use Cases

The IREMA aims to provide feedback to questions like:

- How efficient would be for a Department or Institution to focus on specific areas, and who would be the lead researchers in that case?
- How research collaborations influence R&D productivity?
- With whom is more likely a scientist to collaborate with?

The user can explore the visual representations of the IREMA framework showing the values on specific parameters and metrics. From the layouts the user can select measurements of interest, perform data mining methods and get hypothesis based data visualizations. Then the user can choose from the resulting visualizations for further analysis. In the next sub sections we will discuss in details the process in order to get answers about the first and second question, as the third has already discussed in the section Research link Recommendation.

4.1.3.1. Identification of the key researchers and the most prominent research areas

In order to be able to identify the most promising research areas and the lead researchers for each area, the user has the opportunity to select from a number of different techniques; these different selection choices enable the user to estimate the probabilities to direct research to specific areas and also to observe the correlation among the criteria set for research assessment. The visual analytic process is described in figure 51, which consist of the following stages:

- Data processing (Data Acquisition, Data cleaning, Data transformation).
- Ontology (IREMA Ontology).
- Hypothesis generation (Bayesian Networks application, Association rules, Social Network Analysis and Short Path).
- Visualization processes (Map of science, graph and parallel coordinators).
- Knowledge generation (Identification of the key researchers and the prominent research areas).

Since we have already discussed about the stages of data processing and ontology at previous sections, we will now describe the process for the hypothesis (data mining) layer. Initially the user selects the research areas; then, using the Bayesian Network module the system calculates the probabilities for specific authors to publish a paper in selected areas (the Bayesian network applies prior knowledge to predict the possibility of future activities). In table 30, indicative results are described. For example, we see

that for a specific author having a publication in computer science (1), the probability to also have a publication in Social Science (1) is 0.61. This was extracted based on prior knowledge the system has and using the Bayesian formulas to calculate the probabilities.

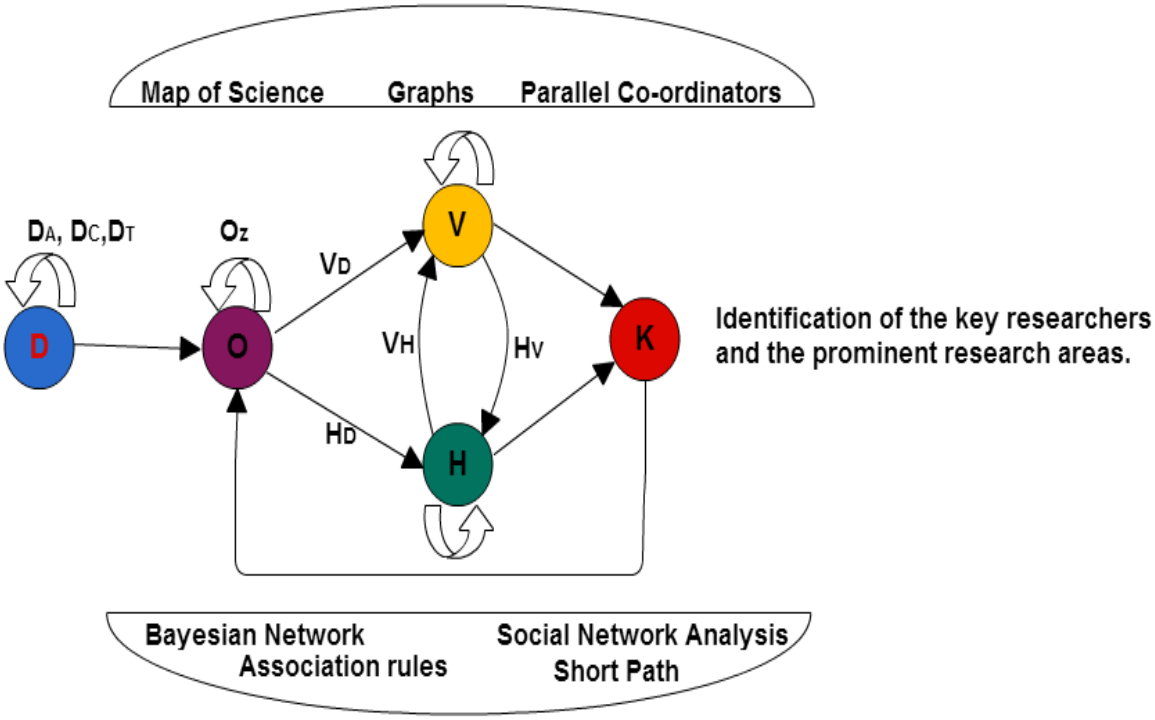


Figure 51: Identification of the key researchers and the prominent research areas

Alternatively the user can extract association rules by applying the Apriori algorithm; for example the user is able to generate rules based on the data described on table 32. The values of those attributes are discretized into 3 counterparts with minimum support 0.8 and confidence level 0.9. In table 31, we can see some of the calculated rules concerning the number of times both LHS (Left Hand Side) and RHS(Right Hand Side) appear (in brackets the confidence level). For example, the first rule indicates that those who have participated as coordinators in research projects for about 2 or 3 times (value 2-3), have also published at least 3 to 5 articles in international journals ([3-5]). The confidence of the rule is 1 and that rule appears 60 times in both LHS and RHS. The explanation behind this is that usually faculty members that participate as coordinators to research projects are likely to publish more journal articles than the others. Also, the second rule indicates a similar situation: if the faculty members have not participated in research projects, they are unlikely to have publications in international journals. Therefore, journal publications in our case seem highly correlated with the participation in research projects and research teams in general.

Computer Science	Social Science	Probabilities
0	1	0.5
1	1	0.61
0	0	0.5
1	0	0.39

Table 31: Bayesian of Computer-Social Science

1. RPC='[2-3)' 60 ==> JAI='(3-5]' 60 conf:(1)

2. RPC='(0)' RPP='(0)' 56==> JAI '(0)' conf:(0.98)

.....

10. RPC='(1-2)' RPE='(1-2]' 57 ==> CPI='(2-5]' JAI='(0-2)' 56 conf:(0.93)

Table 32 : Association Rules.

Apart from making use of the Bayesian Network or association rules, the user has the option to select the most suitable method in order to discover key-researchers among the others. This can be done through four different visual representations:

- The co-authoring graph.
- The interactive clustering using the parallel coordinators.
- The short path exploration.
- The map of science.

We describe in the following subsections, these different choices and their applicability.

Co-authoring Graph

In our case 178 articles were analyzed: 144 have as an author only one faculty member of the department, while the rest 107 articles comprise of 63 dual-authorship papers, 43 triple-authorship papers and 1 four-author paper. For our experiment we constructed a binary 577*577 matrix where each author constitutes a node. Hence, the resulting co-authorship network contains 577 nodes (authors) connected by 2154 collaboration ties.

It is of major importance for the institutional policy to identify the most important researchers among the faculty members who could play a key role in an attempt to enhance scientific performance as well as to increase the research outcomes (e.g. papers, patents, etc.). For this purpose, we have developed an interactive visual interface which illustrates such a co-authoring network, where nodes represent the authors and edges represent the collaboration (co-authoring activity) among them. The appearance of the graph is affected by the criteria (degree centralities, closeness) that the user has set. Different criteria provide different graph networks, thus resulting in different nodes' size.

The color of the nodes is different according to the department a specific user belongs to. Therefore, by using this representation the user could get answers to specific questions like for example:

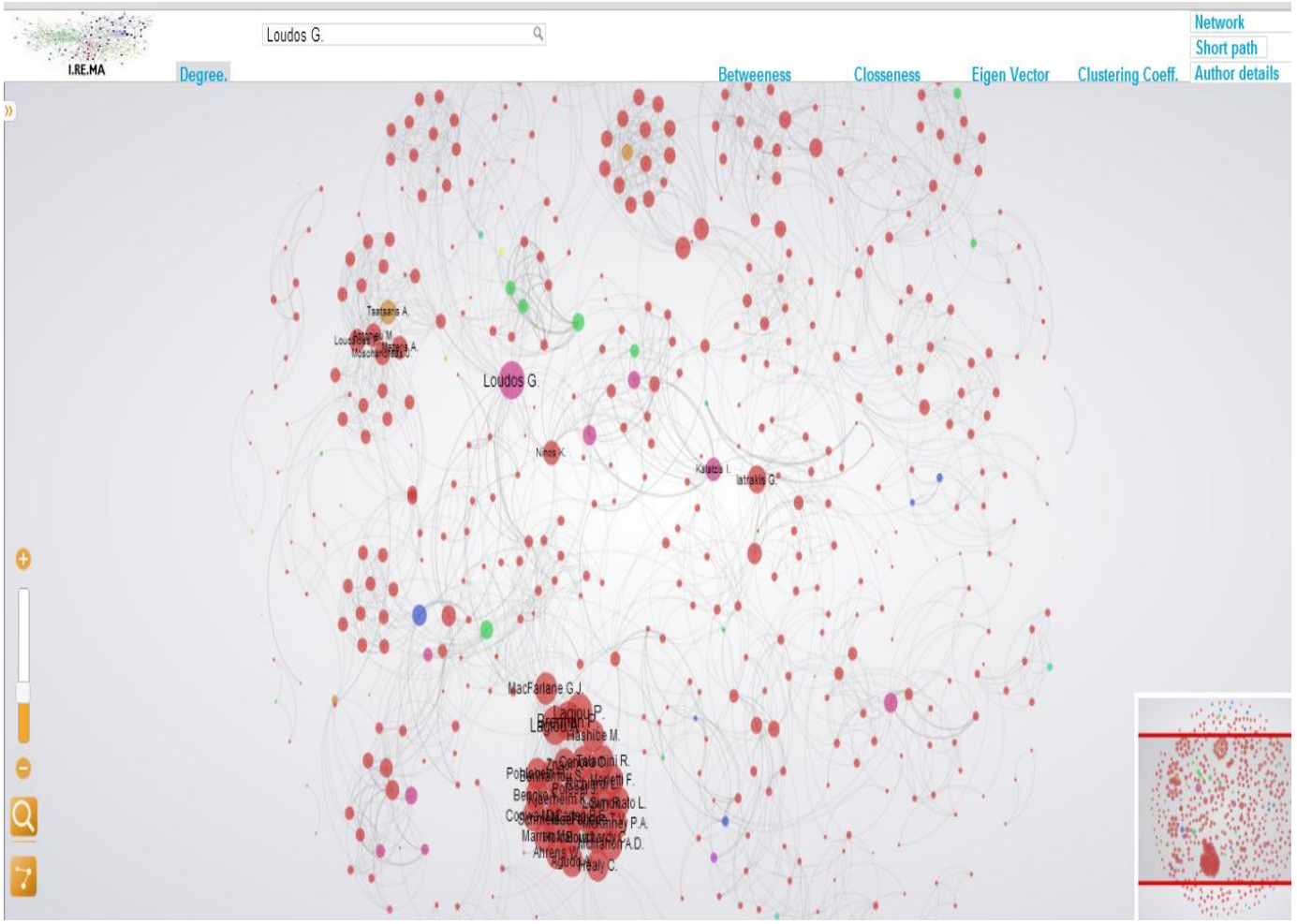


Figure 52: Co-authoring Graph represents the most active researcher

Who are the most active researchers (figure 52). The most active author is the one that has published the most research papers. In our example we select to use the degree centrality measure, which measures the number of lines that are incident to a node. So authors with the biggest diameter than the others are those who have more edges than the others which could imply that they have the higher number of collaborations.

The researchers who act as "research hubs". Research hubs are called those author-nodes who have the capability to transfer information from one author-node to another. By using the betweenness degree, we can measure the ability of a node to connect nodes that do not have any direct connection (edge). In order to get the appropriate representation, the user only needs to click the corresponding link (Degree of Betweeness, Closeness, EigenVector and Clustering Coefficient). The user also can select a node and observe details for the selected author (co-authors, the department, metrics and CIT) and the corresponding network.

The Short path exploration

In addition to the research link recommendation, we have developed an application for the calculation of the shortest path; this application aims in exploring potential collaboration patterns, using as indicators the path distance between two authors at the co-authoring network as well as the research areas of their publications. In graph theory, the shortest path problem is the problem of finding a path between two vertices (or nodes) in a graph such that the sum of the weights of its (constituent) edges is minimized. In our application, we try to find the minimum path between 2 authors. The user initially specifies the source and the target nodes-authors and then he/she is able to view (figure 53) common research areas, if they exist; in addition, the length and the edges of the path are displayed. The user can also view the intermediate steps (the authors) of the path and their research areas.

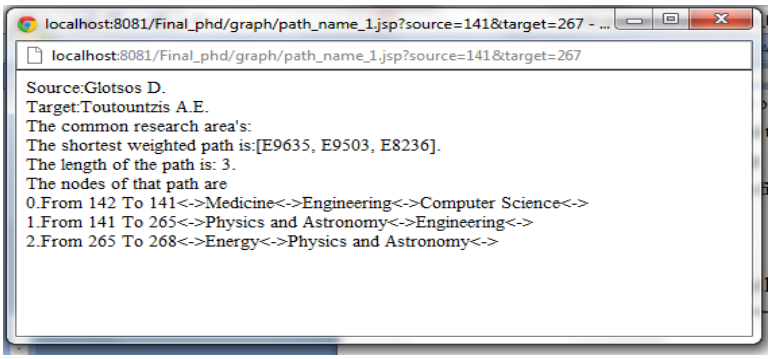


Figure 53: Short Path Distance

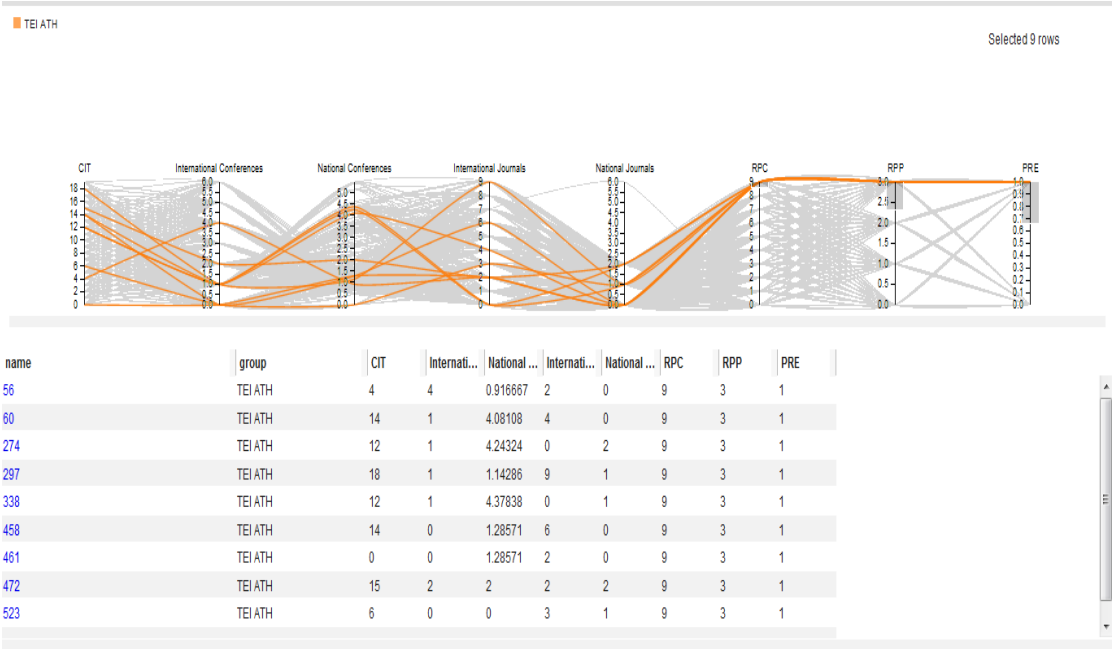
For a hypothetical scenario, if we would like to examine the possibility to establish collaboration between two authors, we should calculate the distance of the path between those authors and their research areas. The existence and accordingly the length of the path provide a degree of support for potential collaborations or against it.

Interactive Clustering

The interacting clustering process is based on the parallel coordinator visualization method. In this method the parallel lines (figure 54) represent all the faculty members, while the axes represent the criteria used for the research assessment (table 33). By using the parallel coordinator visualization method, the user is able to apply a set of criteria depending on his objectives. The criteria can be defined interactively by setting ranges on the axes. In our example the user initially sets the criteria and subsequently he/she gets the corresponding figure, in order to answer questions, such as:

“Who has coordinated the most research projects? (figure 54)” or

“Who has co-authored the most journals?”



name	group	CIT	Internati...	National ...	Internati...	National ...	RPC	RPP	PRE
56	TEI ATH	4	4	0.916667	2	0	9	3	1
60	TEI ATH	14	1	4.08108	4	0	9	3	1
274	TEI ATH	12	1	4.24324	0	2	9	3	1
297	TEI ATH	18	1	1.14286	9	1	9	3	1
338	TEI ATH	12	1	4.37838	0	1	9	3	1
458	TEI ATH	14	0	1.28571	6	0	9	3	1
461	TEI ATH	0	0	1.28571	2	0	9	3	1
472	TEI ATH	15	2	2	2	2	9	3	1
523	TEI ATH	6	0	0	3	1	9	3	1

Figure 54: The authors with the most research projects

The map of science

Another useful option in our system is the so-called map of science. This feature enables to extract useful information according to the faculty a specific user belongs to, or according to the research areas. The map of science represents all the research areas that the faculty members have published their research papers. When the user clicks on an area a second pie opens at the right side of the screen, which informs about researchers' activities in this certain area, or about the number of papers, the authors' id and the department to which they belong to (figure 55). For example, the author with id 180 has co-authored 2 papers, belongs to the department of mathematics and his work in this represents 20 % of the total produced work. Using this representation, the users can get a general view about the overall contribution in certain research areas and have also the ability to observe the most active authors.

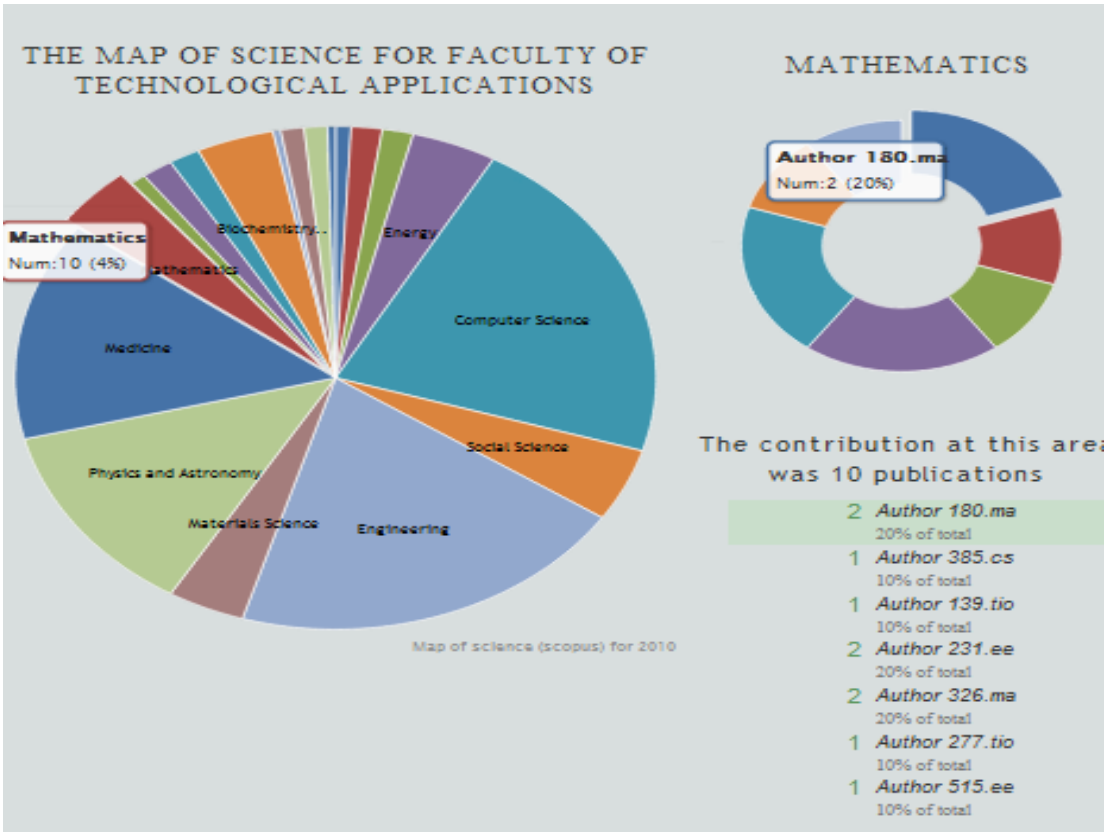


Figure 55: The map of science

Discussion

In order to answer and further explore the question "how efficient for a department would be to focus on specific areas" or "who could be the lead researcher", the user in our example executes the following actions:

- He/she selects specific research areas for example the areas of Computer Science and Social Science; he/she then observes the conditional probabilities, using the past performance in order to forecast the previous activity to future activities.
- The user is able to extract association rules by setting different criteria; for example how likely is for people who participate in projects to publish more research papers than others that don't. Using the Bayesian networks the user is able to get information about the correlation between the criteria set. In our case, it is proved that the most promising research areas (using Bayesian networks) are those in which faculty members have ongoing research projects. So the current and the future research activities are highly correlated with the research projects.
- In addition, the question "who could be the lead researcher?" can be answered by using a variety of visual representation or short path queries as.

- The co-authoring graph through which the user could observe: a) The most active researcher (Degree centrality), b) the authors that have the capacity to transfer information from one researcher to another (Betweeness centrality), c) the authors with the higher average distance to all the other nodes in a network. (closeness centrality), d) the most important node in the network (EigenVector Centrality) and e) how close are the authors and its neighbors to creating a group, or within the given context a "clique" (clustering coefficient)
- The parallel coordinators where the user could set filters and observe the authors that have values at the specified ranges.
- The short path queries where the user sets the source and the target author and observes if there exists a path among the selected authors and if so, then he could view: a) the length of the path, if the length is greater than 1, the authors that participate in the path and the research areas of each one of them and b) the common research areas - if those exist - beside the source and the target author.
- Finally, the user could use the map of science in order to have an overall view about the research publications and the most active authors at each one of them.

4.1.3.2. Research collaborations & Research productivity.

Regarding to the Research & Development activities [98] of the faculty members of Technological Educational Institute of Athens and due to the significant differences that may exist among the different scientific areas, we take into consideration only those activities that have been extracted from the data in respect to the faculty members of Technological Applications. The data set has been separated into input and output variables about research and technology transfer. The advantage of this dataset is that the external conditions for all the faculties are the same. The indicators that we select for the R&D efficiency measure are described in table 32. The visual analytic process for this question is described in figure 56. The stages of the process are:

- Data processing (Data Acquisition, Data cleaning, Data transformation).
- Ontology integration (I.RE.MA Ontology).
- Hypothesis generation (K-means clustering, Social Network Analysis and Data Envelopment Analysis).
- Visualization processes (Regression line).
- Knowledge generation. (Correlation among Research collaborations & Research productivity).

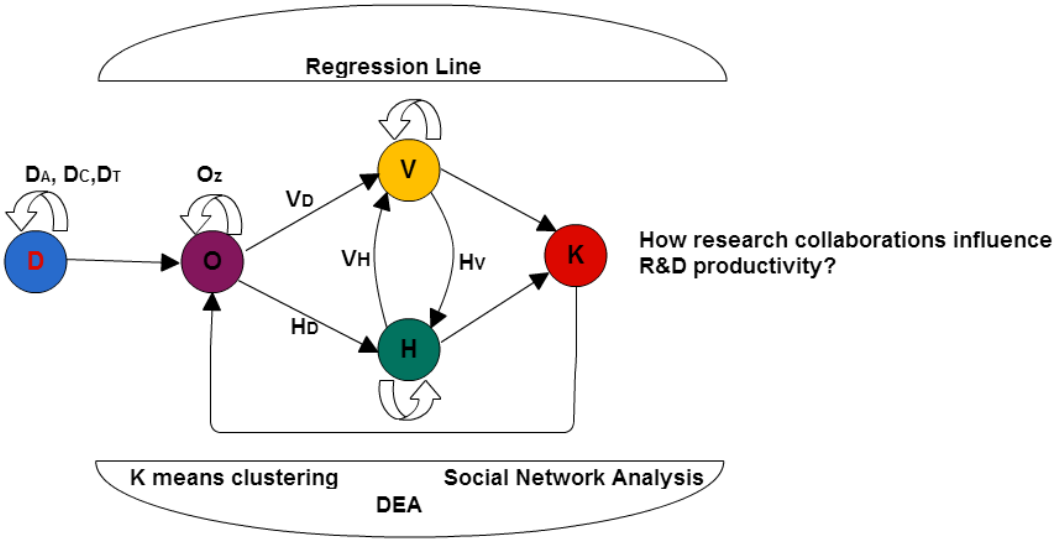


Figure 56: Research collaborations & Research productivity.

Index	Variable	Description	Unit of measurement	Type
1	Staff	Faculty members	# of Persons	Input
2	JAI	International Journal Articles	# of Journals	Output
3	JAN	National Journal Articles		Output
4	CPI	International Conference Papers	# of Conference Papers	Output
5	CPN	National Conference Papers		Output
6	CIT	Citation indexes	# of Citations	Output
7	BC	Book Chapters	# of Book Chapters	Output
8	RPC	Research Project that one of the faculty members has the role of Coordinator	# of Projects	Output
9	RPP	Research Project that one of the faculty members has the role of partner	# of Projects	Output
10	RPE	Research Project with external institutes (RPE).	# of Projects	Output

Table 33: Definition of the variables included in the efficiency measure

In order to measure the research efficiency among the departments of the faculty of Technological Applications we implement the DEA methodology. After the execution of the DEA we get the corresponding score as those described in table 34.

Dept	JAI + JAN	CPI + CPN	BC	Cit	RPC	RPP	RPE	Staff	Eff. score
Energy Technology (Et)	5	3	0	84	2	0	0	15	1
Electronics (Ee)	18	26	2	319	1	1	3	29	0,97
Mathematics (Ma)	8	7	4	12	0	0	1	5	0,99
Naval architecture (Na)	4	7	0	41	0	0	9	14	1
Computer Science (Cs)	16	27	5	91	0	4	3	24	0,83
Civil works and Infrastructure Technology (Peu)	1	1	1	1	0	0	4	13	0,54
Medical Instruments Technology (Tio)	21	41	1	170	1	4	2	15	1
Land Surveying Technology (Top)	3	6	2	9	1	0	2	17	0,72
Physics (Fy)	0	3	0	0	0	0	0	5	0,22

Table 34: Efficiency Scores

Regarding to the collaboration measure among the faculty members of the institute, we assume that the graph metrics are the most suitable representation measures [80]; therefore, we execute the k-means based process in order to find the most suitable grouping for the different departments. The attributes of the instances for clustering are the graph metrics which provide a degree of collaboration. After the clustering of the departments, the expert classifies the clusters using values from 1 to n (n= number of clusters) in order to classify them in a range starting from the worst to the best observed cluster.

CLUS #4 (Very Good=3)	CLUS #3 (Good=2)	CLUS #2 (Bad=1)	CLUS #1 (Excellent=4)
Ee	Cs	Et,MA, Na,Fy,Top,Peu	Tio

Table 35: Collaboration Clusters

In figure 57, the user can see the efficiency scores for the departments and he/she can observe the correlation among collaboration and research performance. For example we observe that there is an instance ("tio") with the highest degree of collaborations as well as the highest value of research performance. In a similar way the user can observe if and how the collaboration among the authors influences the research efficiency of the department.

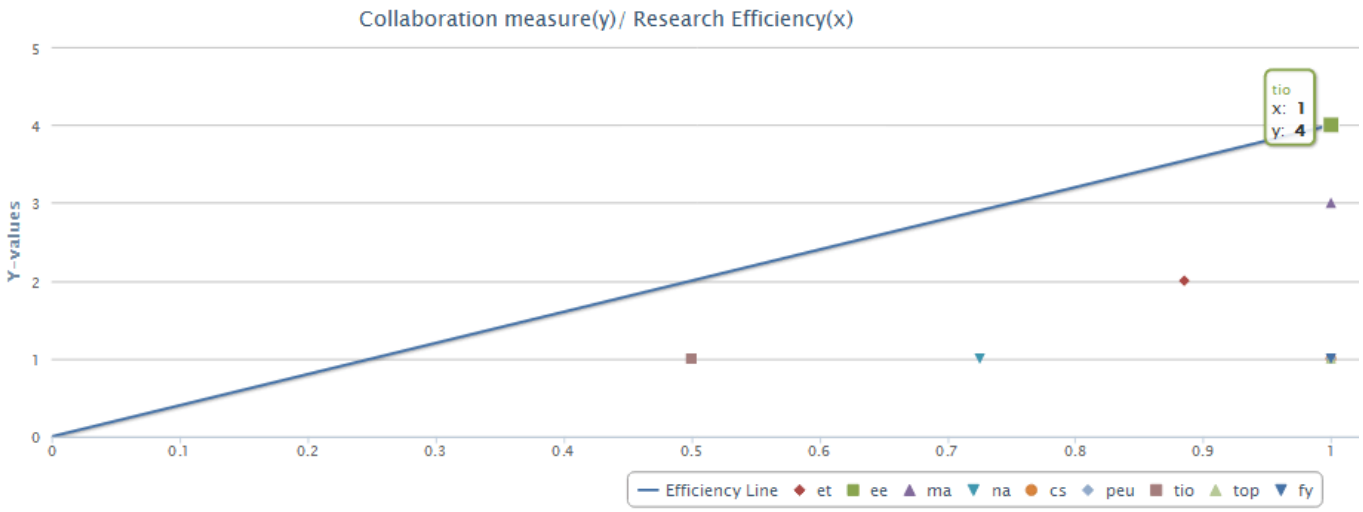


Figure 57: Research (Collaboration / Efficiency)

Another classification we have performed was to classify the output measures in respect to project outcomes and research publications, based on the answers of a group of experts. The classification and the weights obtained for each of the output parameters are shown in table 35. The objective of this approach is to maximize their output using the CCR (input oriented) DEA model.

Output Measure	Weight	Group
JAI	0,15	Research publications
JAN	0,05	
CPI	0.10	
CPN	0.05	
CIT	0,1	
RPC	0.35	Research Projects
RPP	0.15	
RPE	0.05	
SUM:	**1**	

Table 36: Importance-Weights of Output Measures

	OUTPUT		INPUT	Efficiency	
DMU Name	Research Projects/ Staff	Research publication/ Staff	Staff	Value	Efficient
Et	0,7	9,05	15	0,6666666	
Ee	0,65	35,5	29	0,7880716	
Ma	0,05	2,3	5	0,2961373	
Na	0,45	5,05	14	0,4591836	
Cs	0,75	12,8	24	0,4464285	
Peu	0,2	0,35	13	0,2197802	
Tio	1,05	23,3	15	1	Yes
Top	0,45	1,65	17	0,3781511	
Fy	0	0,45	5	0,0579399	

Table 37: Efficiency Scores using Projects-Papers as outputs

After the execution of the DEA, we get the efficiency scores as those described in table 36. In figure 58 we can observe the efficiency line which connects the efficient departments. Using that representation we can see how close the departments are to be efficient (near to efficiency line) and also how the efficiency line separates the departments to those which are "project-oriented" and those that are "publications-oriented". For example the 'ee' department could be classified as more "publications-oriented" because it is located between the efficiency line and the x'x axis and the ‘na’ department as "project-oriented", but both of them are below the efficiency line threshold. Also 'ee' is closer to the efficiency line which means that it indicates higher efficiency value than 'na'.

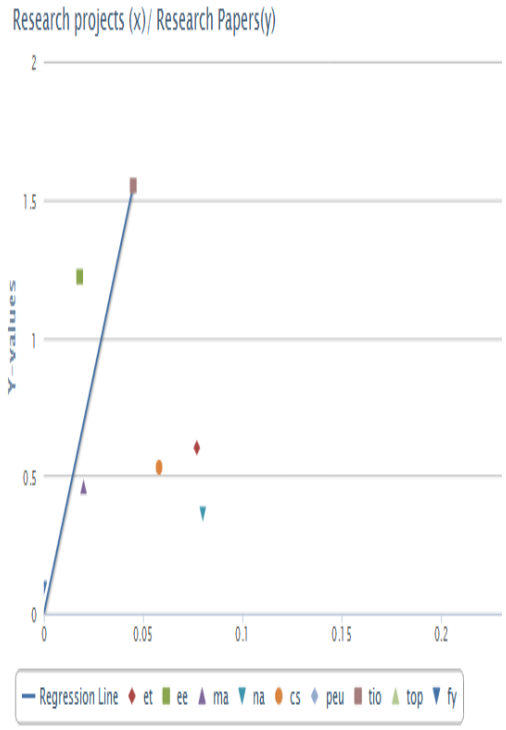


Figure 58: Efficient Line of Project-Paper

Discussion

Based on the co-authorship data we calculate the graph metrics and then we employ a data envelopment analysis, comparing the R&D productivities of each one of the departments in order to examine the impact of collaborating patterns on the R&D performance. The results indicate that departments of higher productivity, retained intensive relations among the faculty members. Therefore, using the proposed methodology we could observe the correlation between concepts like research productivity and research collaboration; moreover, in our example the findings are could implicitly create constructive suggestions on how to set an institutional policy on collaborations among the faculty members in order to enhance the R&D results.

5. Evaluation - Discussion & Conclusions

In this section we will discuss in detail the evaluation process of the I.RE.MA framework. We will present the evaluation results, based on an evaluation form distributed to a group of experts. Then we will discuss about the results of our methodology and finally we will conclude and suggest the future plan.

5.1. Evaluation

The evaluation of DSS using visual analytics is complex since it is primarily designed to provide visual interactive interfaces, data analysis and reasoning. Its evaluation [93] could be performed on a dual basis: based on users who are able to understand concepts related to data analysis and based on these who are able simply to evaluate the provided visualizations.

For those users with basic understanding of data analysis concepts we apply the method that is known as Evaluation of Visual Data Analysis and Reasoning (VDAR)

For those which simply are able to evaluate the visualization characteristics of the application, we evaluate based on the following criteria:

- Evaluation of User Performance (UP)
- Evaluation of User Experience (UE)
- Evaluation of Visualization Algorithms (VA)

VDAR:

The evaluation of Visual Data Analysis and Reasoning is based on the assessment of the visualization tools and its ability to support data analysis and reasoning using data visualizations. The metrics that we collect are:

- the number of insights or new knowledge that is being created
- evaluation of the quality of such results.

In order to execute the data analysis and reasoning we evaluate each one of the following stages in the process of visual analytics:

- Data exploration

  How well according to the user's opinion searching, filtering, reading and extracting information capabilities are supported.

- Knowledge discovery

  We ask users to record how many different capabilities for data visualizations for knowledge discovery are provided

- Hypothesis generation

  We ask users to evaluate how well the system supports generation of hypotheses using data mining methods enriched by interactive visualizations.

- Decision making

  Evaluation of the degree to which the system supported the decision making process by encompassing workflows which lead to decisions. For example if we received accurate answers about the most active researcher.

Evaluation of overall user experience:

The evaluation of the overall user experience using the system was based on the measure of user satisfaction while using the data visualization tool. The main goal was to understand to what extent the visualization supports the initial design purpose, and to make the appropriate changes or modifications in order to be aligned with the initial purpose. For this reason we distributed and evaluation form to the users, in order to get answers to the following questions:

- How “easy to use” is the system?
- What features do you consider as useful?
- Are the reasoning tasks satisfactory?
- Is the tool understandable and how easily can it be learned?

Evaluating Visualization Algorithms (VA)

The evaluations in the visual analytics algorithms are based on the output that they generate and how these results provide answers to the problems. A visualization algorithm should optimize data to a

given visualization goal. The proper selection of a visualization algorithm should provide answers to the following questions:

- Which visualization provides the best answer to a problem?
- Assess the speed required in order to solve a problem.
- How does the algorithm scale to different data sizes and complexities?
- What is the flexibility to modify the user interface.

The evaluation form used to evaluate the IREMA prototype is provided in the next section.

5.1.1. Evaluation Form

In order to evaluate the prototype, we distributed an evaluation form to a group of faculty members which had published papers in the examined period of time. These researchers had no prior knowledge about the recommendation approaches mentioned in this research. They were chosen to explore the decisions that they could get using our system and to evaluate the recommendation results on a 5-point scale ranging from strong satisfaction to strong dissatisfaction (1, very unsatisfied; 2.5, average; 5, very satisfied.). They were also asked to give their satisfaction score based on the degree to which they get efficient answers to the following questions:

Q1." How research collaborations influence R&D productivity?”

Q2. “How efficient would you consider a decision to focus on specific areas and who can be the key researchers? Overall, for the aforementioned system we delivered questionnaires designed according to certain specifications as those described by Ong C.-S., et al [99]. We have separated the questionnaire into the following sections:

- E. Ease of use
- U. Usefulness
- I. Information quality
- S. Satisfaction on Questions.

The results we received from the evaluation of our system are described in table 38.

	Item Description	Av.Degree(0-5)
E1	My interaction with the I.RE.MA is clear and understandable.	3.9
E2	Learning to use the I.RE.MA is easy.	3.6
E3	It is easy for me to become skillful at using the I.RE.MA.	4.2
E4	I find it easy to use the I.RE.MA to do what I want it to do.	3.8
E5	I find the I.RE.MA easy to use.	3.8
E6	The speed in order to solve a problem	4.2
E7	Flexibility to modify the user interface	4.8
U1	Using the I.RE.MA would enhance my effectiveness on the management of research activities.	4.8
U2	Using the I.RE.MA would improve the management of research activities.	4.7
U3	Using the I.RE.MA would get awareness about the research collaborations.	4.6
U4	Using the I.RE.MA would get awareness about the research performance.	4.7
U5	Using the I.RE.MA would get awareness about the correlation among research efficiency and research collaborations.	4.8
U6	Using the I.RE.MA would make it easier to find the key researchers.	5
I1	Information provided in the I.RE.MA is easy to understand.	2.8
I2	Information provided in the I.RE.MA is relevant.	5
I3	Information provided by the I.RE.MA is complete.	3.8
I4	Information provided in the I.RE.MA is personalized.	3.5
S1	Satisfaction on Q1	4.2
S2	Satisfaction on Q2	4.5

Table 38: Evaluation of the system

From table 37, we can see that the average satisfaction score for the proposed approach (E,U,I) is 4.2, and the average satisfaction score of the Q1 and Q2 approach is 4.2 and 4.5 respectively. The high average satisfaction score may suggest that the proposed approach generated satisfactory decisions and the results are considered as positive. Finally regarding to the question “Which visualization provides the best

answer to a problem" the results are displayed in the next representation (figure 59) and we can see that the graph representation achieves the highest score among the others.

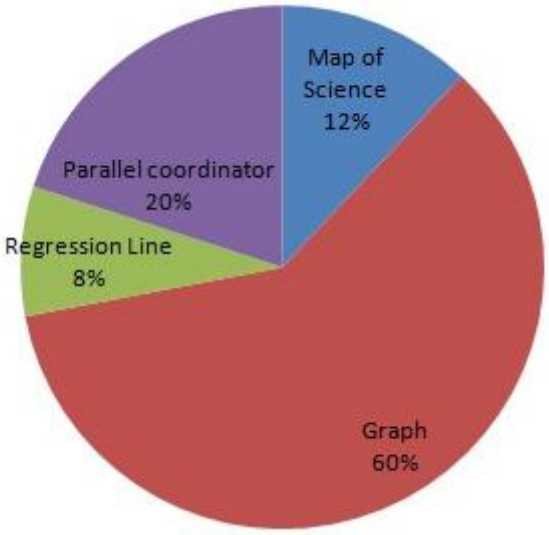


Figure 59: The score of Data Visualisations

5.2. Discussion

In this work, we presented a DSS that used Knowledge Discovery techniques based on visualization methods (KDD-V) in order to support research policy makers to shape correct decisions without being decision making experts or familiar with those techniques.

To the best of our knowledge, this is the first attempt to develop a KDD-based on visualizations using interactive features together with data at the domain of research management in Higher Education Institutes. The aim of this study is to provide a methodology that will help the research policy maker with the ability to combine efficiency measure techniques with data mining methods in a way that they will be able to select the best decision among a variety of possible alternatives. The initial step in the methodology is to define the problem and the objectives; the second step is to apply through the visual interface a categorization of the data and apply data-mining techniques on the data. Finally the policy maker of an institution can proceed with the selection of appropriate recommendations. In the related literature, many models focus either on the efficiency measure of R&D activities **[52]**,**[53]** in collaboration analysis using the co-authoring networks **[49]**,**[48]** or use a combination of these techniques **[55]**, but none of these models is adapted to develop a KDD-V based DSS.

As a proof of concept implementation we presented a system that provides to the user visual analytic capabilities in addition to its ability to support automatic reasoning without any interaction from the user. Our system is able to apply data mining techniques and to acquire various visual representations of these

results; it is also able to validate a specific hypothesis. Through the visual interface the user is able to acquire a report based on the analysis of huge amount of data without further interaction. A strong characteristic of our system is that it does not require from the user to be aware of data analysis concepts. It provides an easy interface to perform complex analysis tasks, such as those based on Bayesian networks or those based on the k-means algorithm; it also guides the user through a process that enables to identify the optimal number of clusters which at the same time abstracts the selection criteria and additional internal details -such as the Ward algorithm which is applied in the background- from the user. The results are presented in such a form that can be comprehended by users who are familiar with the specific domain but not knowledgeable of the processes at the background, and who can explore the visualization in a way that leads them to effective decisions.

We have also introduced semantics in our system by introducing an ontology, namely the IREMA ontology which has been developed to represent information concerning the research activities within HEIs. This ontology addresses the needs of the faculty, the researchers and of both graduate and undergraduate students, creating, storing and exploring academic activities and collaboration possibilities.

The IREMA framework is also able to provide with a degree of probability the possibility for two nodes (which are representing researchers) to collaborate in the future, by applying a link recommendation algorithm on a social research network. Associations can also be extracted based on data structures using rule mining techniques. Another also characteristic of our system is the ability to build the co-authoring graph which represents the researchers-authors and also links together the authors which have cooperated in the past.

Another feature which is also implemented in our system provides the ability to the user to display multi-dimensional data in a two-dimensional space. For the variables of interest such as performance results for faculty members, it allows the user to set the results for faculty members. The user of the system is also able to explore visual representations by adjusting values on specific parameters and metrics. From the layouts the user can select measures of interest, he/she can also perform data mining techniques or is also able to get hypothesis-based data visualizations.

The most important research contribution of this work is the development of a KDD-V based DSS which implements data mining techniques using data visualizations. Our approach focuses on the concept that if a DSS proves difficult to use then it may be rejected by users; our system was designed with this in mind taking into consideration to hide the theoretical concepts from the end user but also to provide an easy to use tool that facilitates Human Computer Interaction. Based on the evaluation results, the IREMA achieves this purpose up to a satisfactory level.

5.3. Conclusion

The increasing demands that research policy makers are facing in order to improve the quality of research management processes, outline the vital importance of a DSS tool that satisfies those demands. Such a DSS tool should make possible the exploration of the research activities and the discovery of useful knowledge. Since our main aim is to achieve a high level of interaction between the user and the system, we have developed a variety of data visualization methods towards that direction. So as to validate our system, we examine and explore some indicative questions about the R&D activities of the Technological Educational Institute of Athens and we provide to the RPMs the most appropriate solution from a set of feasible alternatives.

The IREMA was developed in the basis of four modules as those described in section 3. Regarding the support in the decision making process, we have implemented the following data mining and efficiency techniques:

- Bayesian Networks,
- SNA,
- K-means Clustering,
- Apriori rules associations and
- DEA.

The results illustrate that the combination of data mining with different visualizations can facilitate effective decisions. The future research on the IREMA includes two major directions. The first direction is to enhance the data processing module by implementing a web service which will transform data from scientific databases (as Scopus, dblp,..) to our ontology; the second direction is to enhance the DM process by attracting DM experts to define the appropriate methodologies for facilitating the decision making, the planning processes, and the management of R&D activities in Higher Education Institutions. The future direction of our work focuses on the improvement of the HCI concepts that our system supports and also on how decisions by a policy maker could be supported more effectively in order lead to efficient solutions.

References

[1] Mounir Ben Ayed, Hela Ltifi, Christophe Kolski, Adel M. Alimi. "A user-centered approach for the design and implementation of KDD-based DSS: A case study in the healthcare domain". Decision Support Systems Volume 50, Issue 1, December 2010, Pages 64–78

[2] Moreno AA, Tadepali R . "Assessing Academic Department Efficiency at a Public University". Managerial and Decision Economics 2002;23:385-397.

[3] Arnott, D. and G. Pervan, "A critical analysis of decision support systems research", Journal of Information Technology, 20, 2, 2005, 67-87.

[4] Power, D. J. (1996). What is a DSS? The On-Line Executive Journal for Data-Intensive Decision Support 1(3).

[5] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, " From Data Mining to Knowledge Discovery in Databases". pp. 1-35. AAAI/MIT Press, 1996.

[6] Klein M., Methlie L.B., "Expert Systems: A Decision Support Approach with Applications in Management and Finance, "Addison-Wesley, 1990

[7] H. Ltifi, M. Ben Ayed, C. Kolski, A.M. Alimi, HCI-enriched approach for DSS development: the UP/U approach, in: 14th IEEE Symposium on Computers and Communications, ISCC 2009, July 5–8, Sousse, Tunisia, (2009) pp. 895–900.

[8] Russell, S., Gangopadhyay, A., & Yoon, V. (2008). "Assisting decision making in the event-driven enterprise using wavelets". Decision Support Systems, 46(1), 14–28.

[9] Beynon M., Rasmequan S., Russ S., "A new paradigm for computer-based decision support", Decision Support Systems 33 (2002) 127–142.

[10] Zorrilla M., García-Saiz D. , " A service oriented architecture to provide data mining services for non-expert data miners ", Decision Support Systems Volume 55, Issue 1, April 2013, Pages 399-411.

[11] Fischer G., Human–computer interaction software: lessons learned, challenges ahead, IEEE Software 6 (1) (1989) 44–52.

[12] Keim D A, Kohlhammer J, Ellis G, Mansmann F. Mastering the Information Age: Solving Problems with Visual Analytics. Florian Mansmann, 2010.

[13] J.W. Tukey. Exploratory Data Analysis. Addison-Wesley, Reading MA,1977.

[14] C. Chen. Information Visualization - Beyond the Horizon. Springer, 2004.

[15] R. Spence. Information Visualization - Design for Interaction. Pearson Education Limited, 2nd edition, 2007.

[16] D. A. Keim. Visual exploration of large data sets. Communications of the ACM (CACM), 44(8):38–44, 2001.

[17] J. Thomas and K. Cook, editors. Illuminating the Path: Research and Development Agenda for Visual Analytics. IEEE Press, 2005.

[18] B. Shneiderman and C. Plaisant. Designing the user interface: strategies for e_ective human-computer interaction. Addison-Wesley, 4th edition, 2004.

[19] E. Bertini and D. Lalanne. Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery. SIGKDD Explorations, 11(2):9–18, May 2010.

[20] Shneiderman B. The eyes have it: A task by data type taxonomy for information visualizations. In Proc. the IEEE Symposium on Visual Languages, Sept. 1996, pp.336-343.

[21] Keim D A, Mansmann F, Schneidewind J, Ziegler H. Challenges in visual data analysis. In Proc. the IEEE Conference on Information Visualization, Oct. 2006, pp.9-16.

[22] Keim D A, Kohlhammer J, Ellis G, Mansmann F. Mastering the Information Age: Solving Problems with Visual Analytics. Florian Mansmann, 2010.

[23] Bertini E, Tatu A, Keim D A. Quality metrics in high dimensional data visualization: An overview and systematization. IEEE Transactions on Visualization and Computer Graphics, 2011, 17(12): 2203-2212.

[24] Munzner T. A nested model for visualization design and validation. IEEE Transactions on Visualization and Computer Graphics, 2009, 15(6): 921-928.

[25] Sedlmair M, Meyer M, Munzner T. Design study methodology: Re-ections from the trenches and the stacks. IEEE Transactions on Visualization and Computer Graphics, 2012, 18(12): 2431-2440.

[26] T. von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. J. van Wijk, J.-D. Fekete, and D. W. Fellner. Visual analysis of large graphs. In EuroGraphics: State of the Art Report, 2010.

[27] Tulip.http://www.tulip-software.org/

[28] Graphviz. http://www.graphviz.org/

[29] Gephi. http://gephi.org/

[30] Pajek. http://pajek.imfm.si/

[31] Cytoscape. http://www.cytoscape.org/

[32] Landesberger T, Kuijper A, Schreck T, Kohlhammer J, van Wijk J J, Fekete J D, Fellner D W. Visual analysis of large graphs: State-of-the-art and future research challenges. Computer Graphics Forum, 2011, 30(6): 1719-1749.

[33] Watts, D.J., Strogatz, S.H. "Collective dynamics of 'small-world' networks", Nature 393:440-442

[34] Freeman, L. C. (1977) "A set of measures of centrality based on betweenness". Sociometry 40, 35-41.

[35] Dangalchev Ch., Residual Closeness in Networks, Phisica A 365, 556 (2006).

[36] Bonacich P. " Some unique properties of eigenvector centrality". Social Networks 2007, 29(4):555-564.

[37] Thomas M. J. Fruchterman and Edward M. Reingold. Graph drawing by force-directed placement. Software: Practice and Experience, 21(11):1129–1164, 1991.

[38] Nathalie Henry and Jean-Daniel Fekete. MatLink: Enhanced matrix visualization for analyzing social networks. In Proceedings of IFIP TC13 International Conference on Human-Computer Interaction (INTERACT), pages 288–302, 2007.

[39] Erkki M¨akinen and Harri Siirtola. The barycenter heuristic and the reorderable matrix. Informatica, 29(3):357–363, 2005.

[40] W. T. Tutte. How to draw a graph. Proc. London Math. Society, 13(52):743–768, 1963.

[41] P. Eades. A heuristic for graph drawing. Congressus Numerantium, 42:149–160, 1984.
[42] T. Fruchterman and E. Reingold. Graph drawing by force-directed placement. Softw. – Pract. Exp., 21(11):1129–1164, 1991.
[43] T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. Inform. Process. Lett., 31:7–15, 1989.
[44] Michael J. McGuffin, Simple algorithms for network visualization: A tutorial," Tsinghua Science and Technology, vol. 17, no. 4, pp. 383{398, 2012.
[45] Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. J Informetr. 2011 January 1; 5(1): 187 - 20.
[46] Glänzel and Schubert, "Analysing scientific networks through co-authorship", Handbook of quantitative science and technology research (2004), pp. 257–276
[47] Milojevic S. "Modes of collaboration in modern science: Beyond power laws and preferential attachment", Journal of the American Society for Information Science and Technology, 61 (7) (2010), pp. 1410–1423
[48] Moody J." The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999", American Sociological Review, 69 (2) (2004), p. 213
[49] Newman M.E.J. "Coauthorship networks and patterns of scientific collaboration", Proceedings of the National Academy of Sciences of the United States of America, 101 (Suppl. 1) (2004), p. 5200
[50] Tsolakidis A., Sgouropoulou C. , Xydas I., Terraz O. , Miaoulis G.," Academic Research Policy-Making and Evaluation Using Graph Visualisation". Panhellenic Conference on Informatics 2011: 28-32
[51] Madden G, Savage S, Kemp S."'Measuring Public Sector Efficiency: a Study of Economics Departments at Australian Universities". Education Economics1997;5(2):153-167.
[52] Moreno AA, Tadepali R . "Assessing Academic Department Efficiency at a Public University". Managerial and Decision Economics 2002;23:385-397.
[53] Pesenti, R. and Ukovich, W. 1996." Evaluating academic activities using DEA", Italy: Internal Report, DEEI, Università di Trieste.
[54] Tommaso Agasisti, Giuseppe Catalano, Paolo Landoni and Roberto Verganti." Evaluating the performance of academic departments: an analysis of research-related output efficiency", Research Evaluation (2012) 21(1): 2-14.
[55] Duk Hee Lee ,Il Won Seo, Ho Chull Choe and Hee Dae Kim. "Collaboration network patterns and research performance: the case of Korean public research institution". Scientometrics Volume 91, Number 3 (2012), 925-942,
[56] Abbasi, J. Altmann, "A social network system for analyzing publication activities of researchers Symposium on collective intelligence", COLLIN 2010, Advances in Intelligent and Soft Computing, Springer, Hagen, Germany (2010).
[57] Reagans, R., & Zuckerman, E. W. (2001). Networks, diversity, and productivity: The social capital of corporate R&D teams. Organization Science, 12(4), 502–517.
[58] Rigby, J., & Edler, J. (2005). Peering inside research networks: Some observations on the effect of the intensity of collaboration on the variability of research quality. Research Policy, 34(6), 784–794.
[59] Padula, G. (2008). Enhancing the innovation performance of firms by balancing cohesiveness and bridging ties. Long Range Planning, 41(4), 395–419
[60] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In Proceedings 12th International Conference on Information & Knowledge Management (CIKM'2003), pages 556--559, New Orleans, LO, 2003.
[61] S. Milgram. The small world problem. Psychology Today, 22:61--67, 1967.
[62] S. Goel, R. Muhamad, and D. Watts. Social search in 'small-world' experiments. In Proceedings 18th International World Wide Web Conference (WWW'2009), pages 701--701, Madrid, Spain, 2009.
[63] T. Tylenda, R. Angelova, and S. Bedathur. Towards time-aware link prediction in evolving social networks. In Proceedings 3rd Workshop on Social Network Mining & Analysis (SNA-KDD'2009), pages 9:1--9:10.Paris, France, 2009.
[64] E. Zheleva, L. Getoor, J. Golbeck, and Kuter U. Using friendship ties and family circles for link prediction. In Proceedings 2nd Workshop on Social Network Mining & Analysis (SNA-KDD'2008), pages 97--113, Las Vegas,2008.
[65] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In Proceedings 12th International Conference on Information & Knowledge Management (CIKM'2003), pages 556--559, New Orleans,LO, 2003.
[66] L. Adamic and E. Adar. How to search a social network. Social Networks, 27(3):187--203, 2005.
[67] L. Katz. A new status index derived from sociometric analysis. Psychometrika, 18(1):39--43, 1953.
[68] J. Pan, H. Yang, C. Faloutsos, and P. Duygulu. Automatic multimedia cross-modal correlation discovery. In Proceedings 10th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'2004), pages 653--658, Seattle, WA, 2004.
[69] G. Jeh and J. Widom. SimRank: a measure of structural-context similarity. In Proceedings 8th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'2002), pages 538--543,Edmonton, Canada, 2002.
[70] European Commission, Directorate-General for Research , Assessing Europe's University-Based Research - Expert Group on Assessment of University-Based Research, Luxembourg: Publications Office of the European Union (2010), ISBN 978-92-79-14225-3
[71] Council of the European Union (2007) Council Resolution on modernising universities for Europe's competitiveness in a global knowledge economy, 16096/1/07 REV 1. Retrieved November 2013, from http://www.consilium.europa.eu/ueDocs/cms_Data/docs/pressData/en/intm/97237.pdf
[72] Taylor J., "Managing the Unmanageable: The Management of Research in Research-intensive Universities", Higher Education Management and Policy – Volume 18, no. 2, pp. 9-33,ISSN 1682-3451, OECD 2006
[73] EuroCRIS. Current Research Information Systems
[74] National Institutes of Health. Research portfolio online reporting tool. Available from: URL:http://projectreporter.nih.gov/reporter.cfm. Accessed November 30, 2009.
[75] PubMed , http://www.ncbi.nlm.nih.gov/pubmed
[76] Lane, J., & Bertuzzi, S. "The STAR METRICS project: current and future uses for S&E workforce data.." National Science Foundation; National Institutes of Health.
[77] Brezany P. et al., "Gridminer: An Infrastructure for Data Mining on Computational Grids," Conf. Advanced Computing, Grid Applications, and eResearch, Australian Partnership for Advanced Computing (APAC), 2003, www.apac.edu.au/APAC03.

[78] Rushing J., Ramachandran R., Nair U., Graves S., Welch R., Lin H.," ADaM: a data mining toolkit for scientists and engineers" , Computers & Geosciences 31 (2005) 607–618.

[79] Fischer G., Human–computer interaction software: lessons learned, challenges ahead, IEEE Software 6 (1) (1989) 44–52.

[80] Tsolakidis A, Sgouropoulou C, Papageorgiou E, Terraz O., Miaoulis G., " Using Visual Representation for Decision Support in Institutional Research Evaluation". Intelligent Computer Graphics 2012, 41-57.

[81] Aigner. W, Miksch, S., Müller, W., Schumann, H. & Tominsk, C. 2008. Visual methods for analyzing time-oriented data. Transactions on Visualization and Computer Graphics, Vol. 14, No. 1, pp. 47–60.

[82] Pellet. http://clarkparsia.com/pellet

[83] Costreie, S., Ianole, R. and Dinescu, R.. 2009. An Evaluation of the Quality (Assurance) Evaluation – Case Study: The University of Bucharest. Quality Assurance Review, Volume l, No. 2.

[84] Kandil, M.S., Hassan, A. E., Asem, A. S. and Ibrahim M. E.. Prototype of Web2-based system for Quality Assurance Evaluation Process in Higher education Institutions. International Journal of Electrical & Computer Sciences IJECS-IJENS, Volume 10, No 2.

[85] Hellenic Quality Assurance Agency, http://www.hqaa.gr/

[86] Lueger, M., Vettori, O.. 2008. “Flexibilising” standards? The role of quality standards within a participative quality culture. Implementing and using quality assurance: strategy and practice a selection of papers from the 2nd european quality assurance forum. European University Association. pp 11-16

[87] Haynes, L. 2008. Mentoring and networking: How to make it work. NATURE IMMUNOLOGY.

[88] Kalb, H.. 2011. Social networking services as a facilitator for scientists’ sharing activities. ECIS 2011 Proceedings.

[89] Waldrop, M.M.. 2008. Science 2.0: Great New Tool, or Great Risk?. Scientific American.

[90] M. Jacomy, S. Heymann, T. Venturini, M. Bastian, Forceatlas2, a graph layout algorithm for handy network visualization, Tech. rep., Gephi Consortium (2011).

[91] T. M. J. Fruchterman and E. M. Reingold, “Graph drawing by force-directed placement,” Softw: Pract. Exper., vol. 21, no. 11, pp. 1129–1164, Nov. 1991.

[92] A. Noack, “Energy models for graph clustering,” J. Graph Algorithms Appl., vol. 11, no. 2, pp. 453–480, 2007.

[93] Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, Sheelagh Carpendale, " Empirical Studies in Information Visualization: Seven Scenarios" Visualization and Computer Graphics, IEEE Transactions on (Volume:18 , Issue: 9)

[94] Bizer C., " D2R MAP – A Database to RDF Mapping Language" , in 12th International WorldWide Web Conference (WWW 2003)

[95] Ward, J. H., Jr. (1963), "Hierarchical Grouping to Optimize an Objective Function", Journal of the American Statistical Association, 58, 236–244.

[96] Wu J., “Cluster Analysis and K-means Clustering: An Introduction,” in Advances in K-means Clustering, Springer Berlin Heidelberg, 2012, pp. 1–16.

[97] Hegland, M. "Algorithms for association rules.". Advanced Lectures on Machine Learning. LNCS (LNAI), vol. 2600, pp. 226–234. Springer, Heidelberg (2003)

[98] Tsolakidis A., Sgouropoulou C. , E Papageorgiou, Terraz O. , Miaoulis G.," Institutional Research Management using an Integrated Information System", Procedia-Social and Behavioral Sciences 73, 518-525

[99] Ong C.-S., Day M.-Y., Hsu W.-L." The measurement of user satisfaction with question answering systems" Information and Management, 46 (7) , pp. 397-403.