Analysis and Visualization of the DX Community with Information Extracted from the Web

F.T. de la Rosa, M.T. Gómez-López, and R.M. Gasca

Departamento de Lenguajes y Sistemas Informáticos, Universidad de Sevilla
{ffrosat, mayte, gasca}@lsi.us.es

Abstract. The aim of the present work is the graphical representation of the structures of a specific knowledge area with data extracted from the Web. Recent Internet development has facilitated access to these new resources and has contributed towards the creation of new disciplines which are aimed at taking full advantage of these resources. The main obstacles to this exploitation are their location and processing. This paper defines a generic architecture which solves the problems of processing these resources by combining techniques of extraction, analysis and visualization of data. Specifically in this work we will automatically explore two of the most important structures which define the DX community: the subjects within the community which generate the greatest interest and its social network. Graphical representations are presented to facilitate the analysis of this community.

1 Introduction

Knowledge extraction or knowledge discovery of the resources available on the Internet is one of the fastest growing areas of research in the scientific community, involving areas such as: Webmining, Cybermetrics, Webometrics, Natural Language Processing (NLP), etc. The aim of this work is to implement architecture to perform this task automatically. The transformation of information into knowledge is understood as a "non-trivial extraction of implicit, previously unknown, and potentially useful information from data" [1]. To this end, the architecture has needed the integration of techniques belonging to different areas, notably: Information Extraction, Information Recovery, Scientometrics, Visualization Techniques and Graph Theory Techniques.

The present work is focused on analyzing two of the most important structures which define a community of researchers: the social network and the thematic areas. In order to carry out this analysis, knowledge of the bibliography of the community is necessary. This information enables both the social network to be analyzed through its co-authorship, as well as the application of the techniques of the co-words and NLP to analyze their thematic areas. As an illustrative example of these techniques, a study applied to the DX community is included in this work. For this reason the web pages with the bibliography of the DX workshops since 1997 to 2003 have been downloaded and semi-automatically annotated with NLP techniques.

Although co-authorship and co-word techniques can seem applicable to only the bibliographic information, they have major potential for those systems which can be modeled as a graph where the nodes are entities (words, web pages, enterprises, etc), and where the edges provide information about the existing relationships between the different entities (joint occurrence, reference, friendship, etc). For example, the co-authorship technique can be used to analyze the relationships produced between members of a discussion forum or between company employees, and the co-word technique allows the analysis of the subjects dealt with in any corpus of documents or field with textual information.

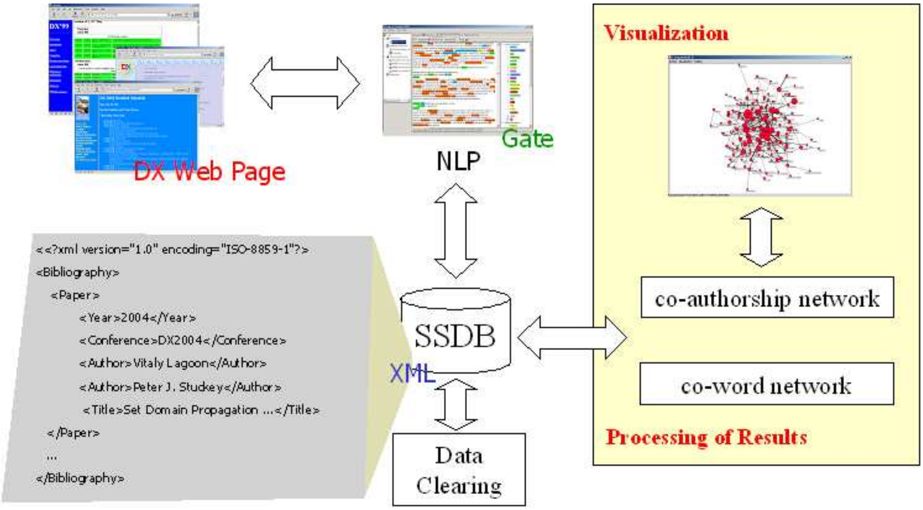
As described in previous work [2], it is necessary to use tools which focus on information from various perspectives in order to create new knowledge from the increasing quantity of information the systems offer. The construction of these perspectives should be guided by our aims or information requirements. For this reason, an architecture is suggested which fulfills our information requirements through the adequate modeling of these networks. For example, for the problem of the analysis of the structures of the DX community, an expert does not have the same information requirements as a new researcher. The expert would be interested in the search for similar authors or emergent lines of research in the same field, whereas a inexperienced researcher would want to discover which subjects are involved and the names of the most influential and relevant authors. Moreover, a company with the intention of obtaining some competitive advantage, would have completely different requirements to others. For example, this company could be interested in looking for experts collaborating on joint projects or in finding out the interests of the researchers of its competitors.

As we have seen, the range of information requirements is very wide and the quality of information under analysis is very large and diffuse, and is sometimes only available from the Internet, as in the case of the communications of the DX community. Development of applications is necessary in order to provide us with automatic selection of desired information.

In order to demonstrate the architecture and techniques used in this work which solve these problems, the article has been divided into the following sections: in Section 2 the developed architecture is described, in Section 3 the process of clearing errors from the extracted data is detailed, and in the Sections 4 and 5 the techniques used to analyze the structures of the DX community are laid out using different visual representations to show its social network and thematic areas. Finally, conclusions are presented in Section 6.

2 Architecture

In previous work [3], an architecture was presented, which integrated all the necessary processes to discover knowledge, from the extraction of information to the processing of analysis and visualization. In Figure 1 the different phases of the architecture can be observed. The architecture has been applied to summarize press articles containing information extracted from different digital newspapers and to analyze the structures of different scientific communities from the bibliographic database DBLP. In both cases the architecture uses a crawler/wrapper to extract the information [4]. One of the most interesting features of the architecture is the decoupling of the process of information extraction from the treatment of the data extracted (analysis and visualization). This has been obtained by making use of a Semistructured Database (SSDB). In this work, the architecture has been enriched by introducing new modules which enable NLP techniques to be used to extract information from heterogeneous web pages, errors in the extracted data to be corrected, and a visual representation of these data to be obtained. For this reason, the architecture has been applied to a more complex problem: the exploration of knowledge areas.



```
<<?xml version="1.0" encoding="ISO-8859-1"?>
<Bibliography>
  <Paper>
        <Year>2004</Year>
        <Conference>DX2004</Conference>
        <Author>Vitaly Lagoon</Author>
        <Author>Peter J. Studkey</Author>
         <Title>Set Domain Propagation ...</Title>
  </Paper>
  ...
</Bibliography>
```

Fig. 1. System Architecture

Following the architecture, web pages of the different DX workshops have been annotated by NLP techniques. The information gathered from the different workshops has been stored in an SSDB, which is recovered information from the authors, articles and titles. The titles have also been annotated with its Noun Phrases (NP) which are groups of words that can function as subjects or objects in sentences.

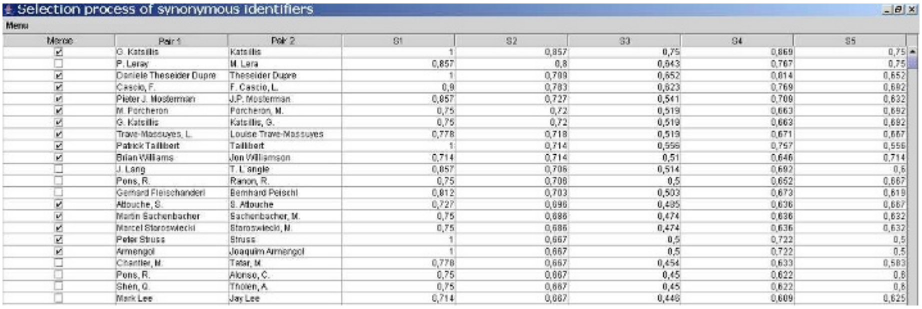
Once the information has been extracted, three processes have to be carried out to obtain the visualization of the information. The first process consists of filtering the authors and the erroneous NP data. For the filtering of these data a data clearing tool is developed which allows the search for and elimination of the errors. In Section 3 it is described in more detail.

Once the SSDB has been filtered of errors, two processes are developed in the phase of processing and analysis, to calculate the co-authorships and co-word networks. Finally, these networks are visually analyzed. In order to carry out the visual analysis task, a tool assists the user in the development and exploration of new perspectives of these networks. The processes of analysis and visualization of the networks will be described in a detailed way in Sections 4 and 5.

3 Data Clearing

After the information extraction of the DX communications, the authors and NP identifiers must be normalized. Some identifiers are synonymous (in that they identified different forms but referred to the same author or NP). For example *“Rafael Ceballos”* is an identifier synonymous to *“R. Cevallos”* and *“model-based diagnostics”* is synonymous to *“Model-based Diagnosis”*. This process of elimination of errors is known as Data Clearing and in [5] a revision of Data Clearing techniques can be found.

A tool has been developed in order to eliminate these errors which calculates several measures of similarity between pairs of identifiers (in this case, strings). The goal of this tool is to enable the user to select synonymous identifiers. For this reason, the tool shows sets of pairs of identifiers, ordered according to their similarity. In the great majority of cases the identifiers with a high similarity are similar identifiers, although not always. From these selections, the tool forms groups of synonymous identifiers, which afterwards are replaced by a single representative identifier. By default this identifier is the most frequent, although the user can change it for another existing or new identifier. This process is totally automatic except for the selection of the pairs.



Merge	Pair 1	Pair 2	S1	S2	S3	S4	S5
☑	G. Katsillis	Katsillis	1	0,857	0,75	0,869	0,75
☐	P. Leray	M. Lera	0,857	0,8	0,643	0,767	0,75
☑	Daniele Theseider Dupre	Theseider Dupre	1	0,789	0,652	0,814	0,652
☑	Cascio, F.	F. Cascio, L.	0,9	0,783	0,623	0,769	0,692
☑	Pieter J. Mosterman	J.P. Mosterman	0,857	0,727	0,541	0,708	0,632
☑	M. Porcheron	Porcheron, M.	0,75	0,72	0,519	0,663	0,692
☑	G. Katsillis	Katsillis, G.	0,75	0,72	0,519	0,663	0,692
☑	Trave-Massuyes, L.	Louise Trave-Massuyes	0,778	0,718	0,519	0,671	0,667
☑	Patrick Taillibert	Taillibert	1	0,714	0,556	0,757	0,556
☑	Brian Williams	Jon Williamson	0,714	0,714	0,51	0,646	0,714
☐	J. Lang	T. L'angle	0,857	0,706	0,514	0,692	0,6
☐	Pons, R.	Ranon, R.	0,75	0,706	0,5	0,652	0,667
☐	Gerhard Fleischanderl	Bernhard Peischl	0,812	0,703	0,503	0,673	0,619
☑	Attouche, S.	S. Attouche	0,727	0,696	0,485	0,636	0,667
☑	Martin Sachenbacher	Sachenbacher, M.	0,75	0,686	0,474	0,636	0,632
☑	Marcel Staroswiecki	Staroswiecki, M.	0,75	0,686	0,474	0,636	0,632
☑	Peter Struss	Struss	1	0,667	0,5	0,722	0,5
☑	Armengol	Joaquim Armengol	1	0,667	0,5	0,722	0,5
☐	Chantler, M.	Tatar, M.	0,778	0,667	0,454	0,633	0,583
☐	Pons, R.	Alonso, C.	0,75	0,667	0,45	0,622	0,6
☐	Shen, Q.	Tholen, A.	0,75	0,667	0,45	0,622	0,6
☐	Mark Lee	Jay Lee	0,714	0,667	0,446	0,609	0,625

Fig. 2. Selection process of synonymous identifiers, ordered by the metrics **s2**

The measures of similarity used in this tool are based on the algorithm of the *Longest Common Subsequence (LCS) of two Sequences*, a technique of dynamic programming and with polynomial complexity. In our case the sequences were the characters of each of the identifiers or NP. The definitions of the metrics used are:

$$s1 = \frac{|lcs(x,y)|}{\min(|x|,|y|)}; \quad s2 = \frac{2*|lcs(x,y)|}{|x|+|y|}$$

$$s3 = \frac{|lcs(x,y)|^2}{|x|*|y|}; \quad s5 = \frac{|lcs(x,y)|}{\max(|x|,|y|)} \tag{1}$$

$$s4 = 0.33*s1 + 0.33*s2 + 0.33*s3$$

In these definitions *lcs(x,y)* is the algorithm which calculates the LCS of the identifiers x and y, and the cardinal represents the length of the identifiers. It is also possible to use other measures of similarity between chains, such as the *Levenshtein distance* or *edit distance* of polynomial cost *O(n*m)* and this measure is less sensitive to the permutation of words in the chains, as for example in the case *“Rafael Ceballos”*and *“Ceballos, R.”*. The *LikeIt distance*, [6], with *O(n+m)* cost, is appropriate to compare strings of large size (*n* and *m* are the lengths of both strings).

Our process has been incorporated into the architecture and its use in different SSDB has been very effective, since the number of pairs of words with a high percentage of similarity is scarce and in only a few minutes a user can decide which pairs of identifiers are synonymous. In Figure 2 the selection process offered by the tool can be observed.

Although the calculation of the measures of similarity for all the pairs of identifiers in the treated SSDB is reasonable, this calculation is considered to be very costly in a very large SSDB. As future works, we suggest using the contextual information that these networks offer us, to improve the efficiency of the algorithm, as well as to automate the process of decision making, by using clustering techniques.

4 Co-authorship Networks

Once data have been cleared, a process is developed to calculate the co-authorship and later the social network is represented. The aim of this technique is the representation of the “invisible colleges” formed in separate established fields of

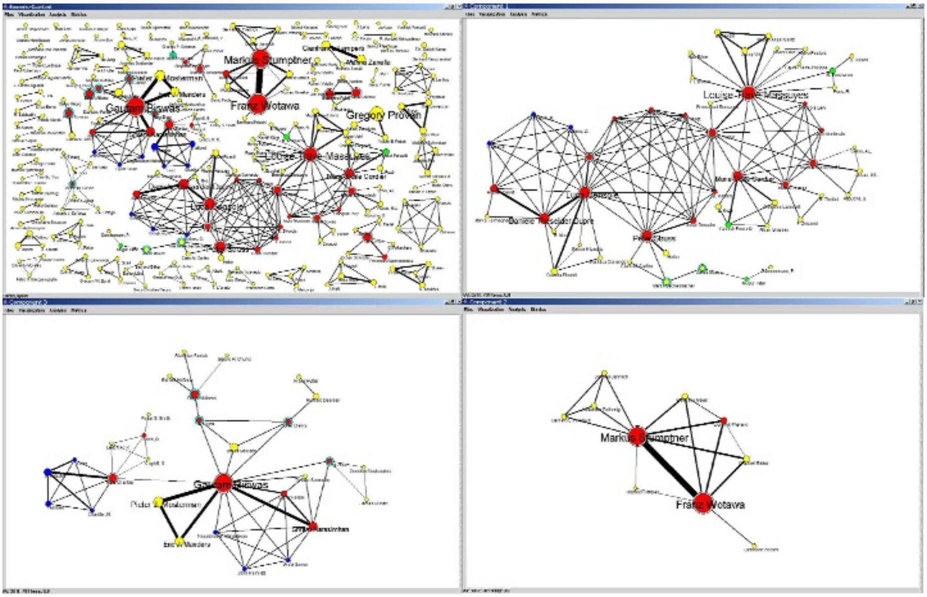


Fig. 3. Co-authorships network of DX community

research. In previous work, [7], the *invisible colleges* are defined as researchers circles influenced by a few researchers of high productivity. There are several metrics to represent the *invisible colleges* through the relationships between the authors: metrics based on *co-citation* [8], metrics based on the *sociometric centrality* and metrics based on the *geodesic distance* [9].

The most appropriate technique to represent the *invisible colleges* from the available data is the *sociogram* [10]. Therefore the impact of the authors and the co-authorship among authors will be used as metrics and are defined as:

$$\text{impact}(w_i) = \#Appearencess\ in\ papers\ titles \qquad (2)$$

$$\text{coocurrences}(w_{i,} w_j) = \#\ Ocurrences\ of\ \ the\ \ words\ \ w_i\ \ and\ \ w_j\ \ in\ papers\ titles \qquad (3)$$

where *ai* and *aj* represent two different authors and # represents the number of publications or collaborations. For a revision of co-authorship visual techniques [11] can be consulted.

With regard to the algorithm for the distribution of the nodes on the screen, the Kamada-Kawai algorithm [12] has been used. The colour of each node indicates the importance of author. This importance is calculated with de PageRank algorithm [13]. Other example of co-authorships networks is presented in the annex.

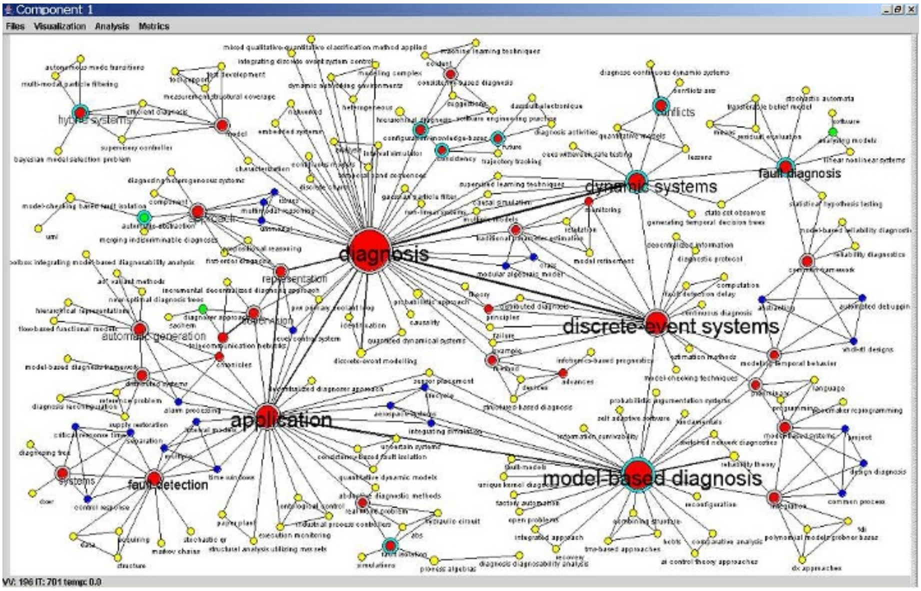


Fig. 4. Co-word network of DX community

The visual analysis of the co-authorship network of the DX community presents of the following features:

1. The network is fragmented into one large, two medium sized and very large number of smaller clusters of authors.
2. The most connected authors in the large cluster are P. Dague, L. Console, L. Trave-Massuyes and P. Struss. This is the kernel of the DX community.
3. The medium sized clusters have G. Biswas, M. Stumptner and F. Wotawa as main researchers.
4. In the smaller clusters, there are several clusters whose size is 5 or 7 authors, however the majority of clusters are composed of one, two or three components.

5 Co-word Network

The aim of the technique of the co-word is the identification of focuses or centers of interest, as well as the relationships that exist between them. These focuses or centers of interest are equivalent to the thematic areas dealt with by the DX community. Our goal is to produce a series of maps from the information obtained with this technique, which deepen the understanding of the thematic structure of the DX community. This technique has been analyzed widely in [14,15,16] and in [17], where we can find the description of a tool, Leximap, which uses this methodology together with techniques of Natural Language Analysis (NLA). The foundations of the technique are based on the construction of a co-word network from a set of documents or their titles and summaries. The co-word network is composed of nodes that represent the terms and of edges whose weights represent the co-occurrences between two words. Unlike the traditional technique, in this work it is assumed that the most frequent terms are the most important, owing to the lack of redundancies in the data available data (we have only the titles of the collected communications). The metrics used are the following ones:

$$\text{impact}(w_i) = \#\textit{Appearencess in papers titles} \tag{4}$$

$$\text{coocurrences}(w_i, w_j) = \#\ \textit{Ocurrences of the words } w_i \textit{and } w_j \textit{in papers titles} \tag{5}$$

where wi and wj are two words or different tokens and # represents the number of wi appearences or the wi and wj joint occurrences. The production process of these thematic maps also differs from the classic technique and consists of the following phases:

- NP annotation with NLP techniques.
- NP data clearing.
- NP normalization. For example, “A Model-Based Diagnosis Framework:” is normalized as “a model-based diagnosis framework”.
- NP stop word elimination. For example, “a model-based diagnosis framework” is normalized as “model-based diagnosis framework”
- Co-word networks construction.

The visual analysis of the co-word network of the DX community presents the following features:

1. The main subject of this community is the diagnosis and an important sub-subject is the model-based diagnosis.
2. The types of the system to be diagnosed are discrete, dynamic and hybrid systems.
3. The most important applications of the work are related with car subsystems, telecommunication networks, aerospace systems, distributed systems and fault detection in industrial systems.
4. The methods used in the DX community are related to artificial techniques where the consistency-based and abductive methods stand out.

6 Conclusion and Future Works

In this paper, two important conclusions can be highlighted. First, the resources available on the Internet can be exploited automatically when an adequate architecture is developed. The architecture presented in this paper holds all the necessary processes to analyze and view the information extracted from the Internet. We name this type of architecture *Internet Knowledge Summarization, Analysis and Visualization (iK-SAV).*

The second conclusions, is the existence of uses for the visual analysis techniques in the development of interactive applications, thereby facilitating the tasks of the final user and the exploration of the research community.

In this case, the DX community has been analyzed and its social network and subject network is represented.

With respect to possible future work, we expect to expand the range of analysis that our tool covers. We intend to implement clustering algorithms which allows the construction of hierarchical maps and which facilitates the exploration of networks. It would also be necessary to improve the application so that the visualization of more complex networks is possible where there are different types of mixed entities. Furthermore, filtering could be carried out by means of the attributes stored in the entities and/or relationships.

Acknowledgements

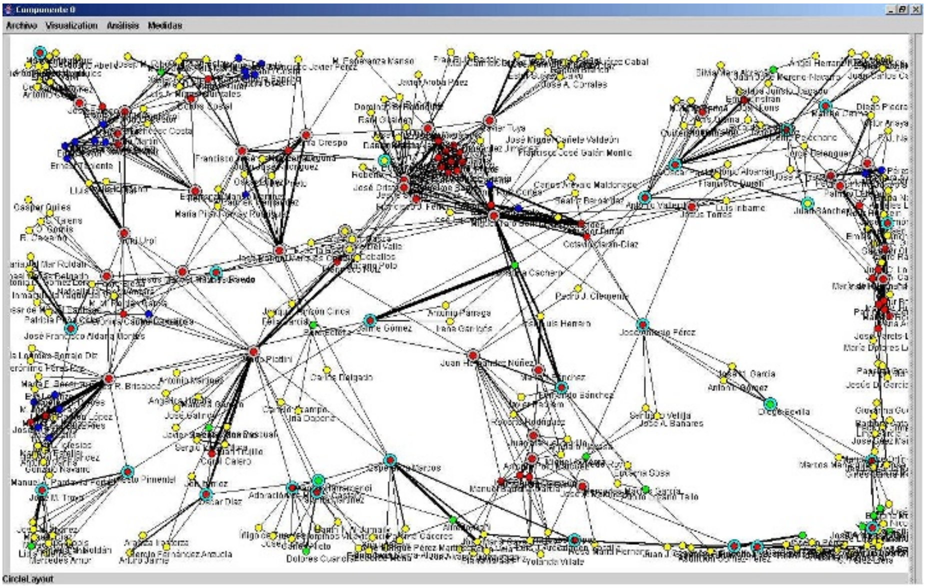
This work has been funded by the Ministerio de Ciencia y Tecnología of Spanish (DPI2003-07146-C02-01) and the European Regional Development Fund (ERDF/ FEDER).

References

1. W. J. Frawley, G. Piatetsky-Shapiro and C.J. Matheus. Knowledge Discovery in Databases: An Overview. In: G. Piatetsky-Shapiro and C.J. Matheus. Knowledge Discovery in databases, pages 1-27, MIT press, 1991
2. Bent Hetzler and Paul Whitney and Lou Martucci and Jim Thomas. Multi-faceted Insight Through Interoperable Visual Information Analysis Paradigms. In: Proceedings {IEEE} Symposium on Information Visualization 1998.

3. F. de la Rosa T., Rafael M. Gasca, Carmelo Del Valle, Rafael Ceballos: Arquitectura de un Crawler para Extraer las Estrcturas y Contenidos de Recursos Electrónicos. JISBD 2002: 259269
4. Thomas Kistler and Hannes Mariais. WebL – A Programming language for the Web. Computer Networks and IDSN Systems (Procceding of the WWW7 Conference). Volume 30. pages 259270. Elsevier.1998.
5. E. Rahm and H.H. Do. Data Clearing: Problems and Current Approaches. IEE Bulletin of the Technical Commitee on Data Enginnering. 23(4), 2000
6. S. R. Bus, And P. N. Yianilos, A bipartite matching approach to approximate string comparison and search, NEC Research Institute Technical Report, 1995.
7. Derek J. de Solla Price. Little Science, Big Science. Columbia Univ. Press, New York, 1963.
8. Henry Small. Co-citation in the scientific literature: a new measure of the relationship between two documents. Journal of the American Society for Information Sciences 24, pp.265-269, Jul-Aug 1973.
9. José Luis Molina, Juan M. Muñoz Justicia y Miquel Domenech. Redes de publicaciones científicas. Un análisis de la estructura de coautorías. Revistas Hispano Americana para el Análisis de Redes Sociales. Vol1. 2002
10. Moreno, J. L. (1934). Who shall survive? New York: Beacon Press.
11. Chen, C. Mapping Scientific Frontiers: The Quest for Knowledge Visualization. Springer. 2003
12. Tomihisa Kamada and Satoru Kawai: An algorithm for drawing general indirect graphs. Information Processing Letters 31(1):7-15, 1989
13. Page, L. and Brin, S. (1999). "The Anatomy of a Large-Scale Hypertextual Web Search Engine".
14. 14. Callon, M., Law, J., and Rip, A. Mapping the dynamics of science and technology: Sociology of science in the real world. London: Macmillan. 1986
15. Callon, M., Courtial, J.P. y Laville, F. Co-Word analysis as a tool for describing the network of interactions between basic and technological research: the case of polymer chemistry. Scientometrics, 1991, vol. 22, nº1, p. 155-205.
16. Coulter, N., Monarch, I. & Konda, S. Software engineering as seen through its research literature: A study in co-word analysis. Journal of the American Society for Information Science, 49(13), 1206-1223. 1998
17. Ira A. Monarch. Information Science and Information Systems: Converging or Diverging? 2000. http://www.slis.ualberta.ca/cais2000/monarch.htm

Annex. Co-Authorships of Journal of Software Engineering and Databases kernel


Componente 0
Archivo
Visualization
Análisis
Medidas
M. Esperanza Manso
Javier Tuya
José Miguel Cañete Valdeón
Francisco José Galán Morillo
Carlos Arévalo Maldonado
Beatriz Bernárdez
Antonio Vallecillo
Jesús Torres
Luis Iribarne
Cotos Marcos-Díaz
Pedro J. Clemente
Juan Hernández Núñez
José Antonio Pérez
Antonio Gómez
Diego Sevilla
Antonio Paniega
Elena García
Jaime Gómez
Carlos Delgado
Mario Piattini
Antonio Martínez
Esperanza Marcos
Adoración de Miguel
César Díaz
Ernesto Pimentel
José M. Troya
Mercedes Amor
Arturo Jaime
Dolores Cuadra
Yolanda Villate
Gaspar Quiles
Francisco José
Lidia Fuentes
Marcos
José Dolores
Jesús D. García
Diego Piedra
CircleLayout