Affinity Analysis between Researchers using Text Mining and Differential Analysis of Graphs

Luís Trigo[1], Pavel Brazdil[1,2]

[1]LIAAD-INESC Tec, [2]FEP, Univ. of Porto

{lptrigo,pbrazdil}@inescporto.pt

Abstract. Finding people with similar skills within a domain may provide an important support for managing research centers. The academic production, albeit in an unstructured format, is easily accessible. Thus, we resort to sources on the web - academic and bibliographic databases - to uncover the affinities among researchers. What interests us most are affinities that are not yet evidenced by co-authorship. Besides, of interest are also other outputs of the method in the form of subgroups and the researchers that play an important role in them.

Keywords: Web Mining, Text mining, Clustering, Social Network Analysis, Differential Analysis of Graphs.

1. Introduction

Researchers seek to discover other researchers with similar interests to follow their work and plan future collaborations. At management level, it enables identifying suitable researchers for a given task, which precedes the implementation of partnerships with other institutions and researchers policies. Another advantage of this analysis is that it goes beyond the formal hierarchical framework within the organization, thereby revealing its unknown connections that can be followed up.

The main scientific contribution is beyond re-using standard techniques of text mining to bibliographic databases, but rather using these techniques to obtain two kinds of graphs, co-authorship and affinity graphs, and exploring a differential analysis with the aim of identifying new useful knowledge.

Our aim is to focus on affinity analysis between certain research centers for various reasons: First, the outcome of the study may be useful to these centers. It may propose that certain collaborations be initiated.

Besides, the outcomes of automatic analysis may be easily verified by some members of these centers. This research could be extended later to cover a larger set of centers.

Regarding the discovery of similarities between researchers, Price et al. (2010) developed a methodology for the Web, called *SubSift*, establishing profiles for researchers on the basis of researchers' publications. Based on these profiles, a typical Information Retrieval task is performed aiming to compare the papers submitted to a scientific conference (playing the role of Query in IR) with different profiles, in order to optimize the task of distributing articles to review.

Rogosky and Goldstone (2002) have pointed out that, in the context of a conceptual network, the meaning of a concept - here "*researcher*" - depends on the relationship with the other concepts in the conceptual framework. Thus we analyze networks of affinities and of particular interest are those that are not covered by the simple co-authorship connections. To uncover these we have to resort to many different techniques, including web mining, text mining, social network analysis, sub-graph discovery and differential analysis of graphs and graph analysis.

The main steps of the method are described in the following.

2. Methodology

This section presents the main steps undertaken to uncover the unknown information regarding affinities. The method involves the following steps:

3. Query user to obtain names of institutions and websites;
4. Web mining to identify researchers' names;
5. Web / Text mining to process researchers' publications;
6. Elaboration of similarity matrix and visualization using graphs;
7. Application of sub-group discovery to the affinity graph;
8. Elaboration of a co-authorship graphs and differential analysis of

graphs;

9. Identification of important nodes (researchers) in the graph.

The details about all these steps are given in the following.

9.1 Query user to obtain names of institutions

So far, our work is in a prototype stage and so we have applied the method to two closely related R&D units, INESC TEC[1] – LIAAD [2] and CRACS [3]. The total number of researchers does not exceed several dozen. Our plan is to apply the method to larger set of units in future, such as the whole INESC or some Faculties of the University of Porto or other Universities.

9.2 Web mining to identify researchers' names

Each research institution has normally a webpage listing their researchers. Lists of researchers can be extracted easily by building an expression in the *XPath* query language to obtain their names from the website.

Regarding tools to extract and process the data, we chose R with its *tm* package for part of text mining, the *XML* package for web mining, as well as *igraph* and *sna* packages for clustering and social network analysis.

9.3 Web / Text mining to process researchers' publications

Each researcher name can be inserted in the search URL for the DBLP [4] which enables direct access to each researcher list of publications. The retrieval of publications can be done automatically, using *XPath* expressions. However, a problem of *named entity identification* arises here. This is because researchers may have several variants of their name. Thus several entries may exist in the bibliographic database for the same researcher, each associated with a particular variant of his/her name. Typically, one of the variants will appear on the institution site. This name may not match the name used in the bibliographic database.

Another problem is that we may have several investigators with the

same name in the bibliographic database. One of the techniques used by Bugla (2009) is the following. To determine whether a given publication of P in some bibliographic database should be attributed to person P' on a given site, a check is made whether both (i.e. P and P') have the same home institution.

Regarding the particular bibliographic database, we have chosen DBLP, because it is an open and comprehensive bibliographic database in the field of computer science. Currently, we are considering to use Authenticus instead, as its design was based on Bugla's work within a project from the University of Porto, and has the advantage that it retrieves publications from several other bibliographic databases (incl. e.g. SCOPUS).

The publications titles are extracted into plain text files, each representing a particular author. The text files are retrieved and preprocessed in the usual manner. We use *BoW* representation, remove numbers, stop-words, punctuation and other spurious elements. After this task, the list of documents is transformed into a document-term vector representation with *tf-idf* weighting (Feldman and Sanger, 2007).

9.4 Elaboration of similarity matrix and visualization using graphs

The vector representation described in the previous step is used to generate the cosine similarity matrix. This matrix can be visualized in the form of a graph and is used as the basis for further processing. Fig. 2 shows an example of an affinity graph, where all links (similarities) below a given threshold have been considered irrelevant and hence removed.

9.5 Application of community discovery to the affinity graph

After transforming the similarity matrix into a graph format, we use the community discovery algorithm called *Walktrap* (Pons and Latapy, 2006). This technique finds densely connected sub-graphs, also defined as communities, through random walks. It assumes that short random walks tend to stay in the same community.

The hierarchical agglomerative approach is based on a measure of distance between vertices (node to node) and an example of the output of clustering can be visualized in the following figure (Fig. 1). An optimal level of modularity of the network, based on the weighted connections between internal and external community is used by the algorithm to identify non-hierarchic communities. In our example below, the method identified three communities, identified as L-ML, L-OR and C on the basis of data gathered in 2011.

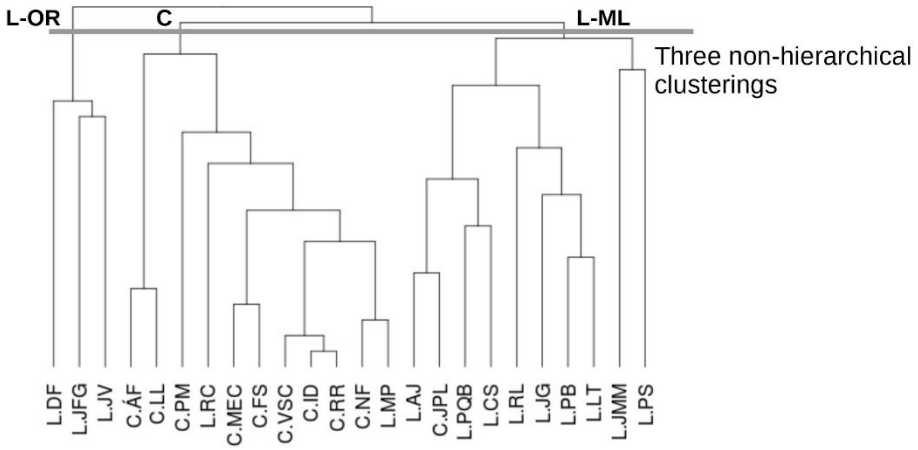

Fig. . Dendrogram generated by the Walktrap clustering algorithm

Different discovered communities can be superimposed on the graph. The result can be seen in Fig. 1, where different communities uncovered by the Walktrap have been identified by ellipses.

The communities discovered correspond well to the organizational structure of the two studied entities. One of the interesting issues to study in the future is - what are the differences between the two organizational structures This differences can suggest that a possible re-organization could be considered by the management in future.

The Researcher Affinity Network graph (Fig. 2) that was generated enables to perform a visual analysis of relationships between researchers. The thickness of the edges represents the similarity weight

between pairs of researchers. Node dimension reflects the number of publications at DBLP for each author.

The same graph shows also several communities that were detected by Walktrap. Further analysis of the community structure is presented in section 2.7.

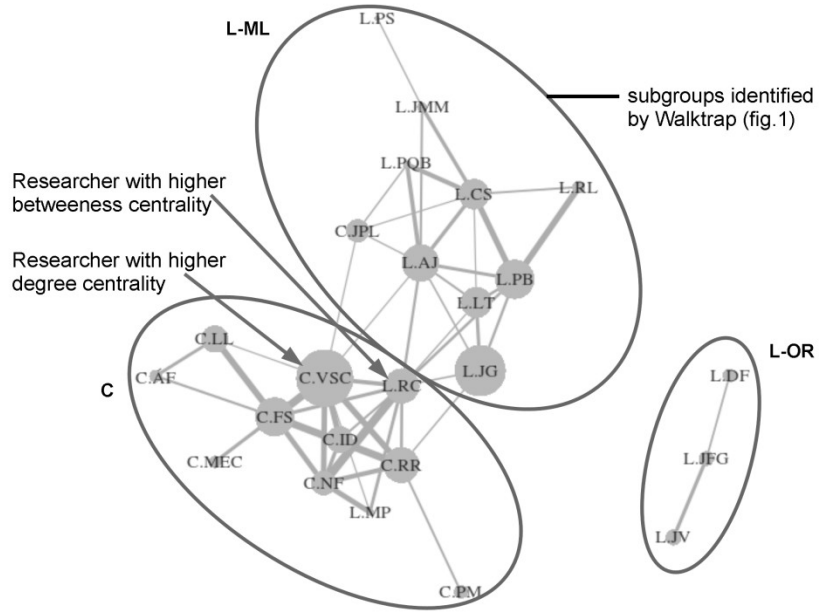

Fig. . Researcher Affinity Network with communities identified by the *Walktrap* algorithm

9.6 Elaboration of a co-authorship graph and differential analysis of graphs

The generation of the co-authorship graph is a relatively simple matter. A link between authors A and B is introduced, if they are co-authors of at least one of the papers. After constructing the affinity graph (G1) and co-authorship graphs (G2), we can proceed to the next step which involves differential analysis. This involves constructing a graph that is basically a difference between G1 and G2. The next figure shows an example.

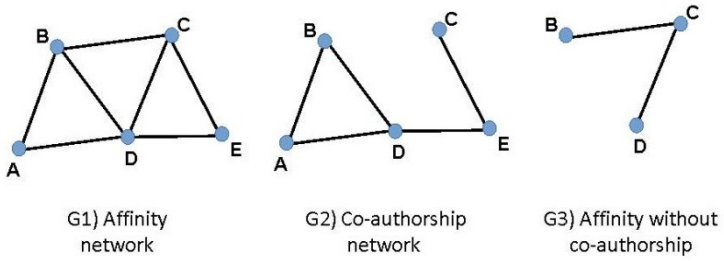

Fig. . Example of differential analysis

9.7 Identification of important nodes (researchers) in the graph.

The affinity network enables to calculate certain measures of importance of the researchers within their community and in the context of different communities. This involves, for instance, *degree centrality* and *betweenness* centrality among others (Wasserman and Faust, 1994).

Some centrality measures can be computed to account for different weights of the connections, as shown in the table below. The degree centrality is based on the number of connections to a vertex. The *betweenness centrality* indicates the number of times a vertex joins two other vertices on the shortest path. The eigenvector centrality shows the importance of vertices that connect to a given vertex.

Table 1 shows an example. The black marked cells indicate that the largest *degree centrality* is located in CRACS, while that the largest *betweenness centrality* belongs to a member of LIAAD (which was clustered with CRACS researchers), as noted in the previous figure visually. It also seems that, as pointed out by the *eigenvector centrality*, the influence within the community from the most central authors in LIAAD is more tenuous than the influence of the most central authors in CRACS.

	L.PB	L.RC	L.AJ	L.JG	C.RR	C.FS	C.VSC
Degree centrality	3.4	4.8	3.1	1.6	4	**5.6**	4.7
Betweenness centrality	41	**179**	130	18	45	88	16
Eigenvector centrality	0.07	0.37	0.06	0.06	0.38	**0.45**	0.44

Table . Centrality measures for some of the most relevant researchers

10. Conclusions and future work

The current work explores Web/Text mining for matching researcher names with their publication titles. This permits to retrieve researchers' publications and process the text files to construct a similarity matrix and a network of affinities.

Further processing leads to quite interesting results in the form of sub-graphs / communities. These can be compared to the formal organization structure. Further work involves differential analysis of graphs on the basis of the affinity and co-authorship graphs. The resulting differential analysis enables to identify pairs of people that could potentially benefit from working together.

In future work, we plan to design an adaptive method capable of retrieving researcher's names from sites with unknown format, or sites that may have altered the format. The method will rely on a fact that at least one researcher's name is known. The HTML/XML source code of the page will be analyzed with the aim of identifying the researcher's name there and elaborating a convenient Xpath expression leading to this name. The command will be adapted so as to be able to retrieve all researchers' names.

We also intend to process the abstracts and consider a substantially higher number of research centers and/or researchers representing some challenges to the process of clustering. To overcome these, we plan to use an incremental / data-streaming approach for this task (Gama et al., 2010).

A validation step needs to be added to the methodology. A brief online survey will be carried out for the most central researchers about who could be their potential collaborator. The outcome will be compared to the results of differential analysis.

An important problem in the text mining phase is that researchers from different domains use different vocabulary/terminology to describe the same things. This problem is difficult to overcome. We will try to use,

as some others did, Wordnet and DBpedia (Leal et al, 2012) to identify synonyms and related words, although this may be harder for some specific domains, which may require specific dictionaries.

Regarding related management needs, we intend to go beyond similarity analysis and study the complementary analysis to uncover potential collaboration between different individuals. The aim of this analysis would be to identify two or more researchers with complementary skills for a given task.

Acknowledgments

This work is partially funded by FCT/MEC through PIDDAC and ERDF/ON2 within project NORTE-07-0124-FEDER-000059 and through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project FCOMP-01-0124-FEDER-037281.

References

11. INESC TEC, http://www.inescporto.pt/
12. LIAAD, http://www.liaad.up.pt/
13. CRACS, http://cracs.fc.up.pt/
14. DBLP, http://www.informatik.uni-trier.de/~ley/db/
15. Bugla, S.: Name identification in scientific publications, Master's thesis, University of Porto (2009)
16. Feldman, R., Sanger, J.: Text Mining Textbook: Advanced Approaches in Analysing Unstructured Data, Cambridge Univ. Press (2007)
17. Gama, J., Rodrigues, P. P., Spinosa, E. J., Ferreira de Carvalho, A. C.: Knowledge discovery from data streams. Boca Raton: Chapman & Hall/CRC (2010)
18. Goldstone, R., Rogosky, B. J.: Using relations within conceptual systems to translate across conceptual systems, Cognition, 84. pp. 295–320 (2002)
19. Leal, J. P., Rodrigues, V., Queirós, R.: Computing semantic relatedness using dbpedia. In OASIcs-OpenAccess Series in Informatics (Vol. 21). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (2012)
20. Pons, P., Latapy, M.: Computing communities in large networks using random walks, Journal of Graph Algorithms and Applications, Vol. 10, no. 2, pp. 191-218 (2006)
21. Price, S., Flach, P. A., Spielgler, S., Bailey, C., Rogers, N.: SubSift web services and workflows for profiling and comparing scientists and their published works, Sixth IEEE International Conference on e–Science (2010)
22. Wasserman, S., Faust, K.:, Social network analysis: methods and applications. Cambridge University Press, New York (1994)