COMPETITIVE INTELLIGENCE BASED ON SOCIAL NETWORKS FOR DECISION MAKING

Fco Fernando de la Rosa Troyano[1], María Teresa Gómez López[1]and Rafael Martínez Gasca[1]

[1] Departamento de Lenguajes y Sistemas Informáticos. Dpto. Lenguajes y sistemas informaticos. University of Seville, Spain
{ffrosat, maytegomez, gasca}@us.es

Abstract. In previous works a framework has been presented to extract from internet the scientific community interested in a specific topic. The process uses search engines query results and e-mails address co-occurrences to obtain the invisible colleges and subtopics of a community. This work presents the use of this technique to implement several competitive intelligence tasks to help in decision making in a research area. In order to show an illustrative purpose, this technique is applied to analyze the social network of participants in several Veille Stratégique Scientifique & Technologique editions.

Keywords: Social networks extraction, competitive intelligence, search engine mining, emails address mining.

1 Introduction

One of the consequences caused by internet expansion is the exponential growth of public information. This information is stored in a set of interlinked heterogeneous sources. Search engines play a key role in internet searching information, but the general approaches to analyze the information systems cannot integrate different sources. The analysis of the stored information in a systematic and automatic way can help to decision making in competitive intelligence. In this context, the co-occurrences analysis becomes imperative to implement intelligence competitive tasks.

In previous works a process to extract automatically the scientific community interested in a specific topic was presented. The process carries out a systematic search engines queries. These queries are specially designed to extract the goal network through a posterior analysis of the search engines requests. Analysing this extracted information is also possible to determine the community subtopics of interest. This paper discusses the use of this technique to implement several competitive intelligence tasks. For example: searching for experts, gathering of information, organizations collaborations analysis, countries collaborations analysis or products impacts analysis.

The paper is divided into the following sections. Section 2 presents the related works and compares them with our proposal. Section 3 presents the Social networks topic-driven extraction algorithm. Section 4 analyzes some Competitive Intelligence Tasks. In order to illustrate the purpose, the defined process is applied to extract the social network of participants in the Veille Stratégique Scientifique & Technologique (VSST). Finally, conclusions and future work are presented.

2 Related Works

Previous works have defined social networks extraction processes for different information web sources: search engines [8][10][12], chats [14], DBLP [5], FOAF archives [12][13], SourceForge [1], mailing lists [2], etc. The review of this paper is restricted to approaches using search engines to extract social networks. One of the systems that uses search engines to extract social networks is REFERRAL WEB [8] which was designed to use Altavista engine, the obtained network is focused on a specific person (egocentric network). For this reason, it only needs to know the name of the ego. By mean of an entity recognition system extracts a list of related people. To measure the significance of the relationship between X and Y (the variable Y contains any of the persons related to X) uses the Jaccard coefficient [7]. The above process may be repeated recursively.

Recently two systems POLYPHONET [10][11] and FLINK [12][13] were launched, these systems obtain the social network using a list of the members of a determined community. The basic algorithm of both systems must do a query for each pair X and Y to create the affinity matrix (X and Y are two different names of the list). Both systems use a threshold to determine when relationships are meaningful. These systems have several disadvantages, for example the calculated affinities are difficult to understand. The terms co-occurrence on the indexed pages may be due to several factors: co-authorship, participation in the same event (i.e., program committee), referenced in the same paper, etc. By using personal names in the query could add ambiguity to the results, since the probability that the name refers to more than one person is high. Sometimes several names reference to the same person (i.e., 'Rafael Martínez Gasca' or 'R.M. Gasca'). Another disadvantage is the high cost of extracting the social network, since in the worst case for each pair of members a query must be performed. This is a major problem given that licenses to use search engines may have limitations. For example, Google does not allow more than 1000 queries per day (to calculate a network of 500 actors would need 125 days). The scalable algorithm implementations [11] have reduced the numbers of necessary queries to complete a whole network (according to the author, for 503 actors 19,852 queries are needed, 20 days).

The reviewed works use lists of names to extract social networks. In contrast there is the possibility of replacing the lists of names by e-mails address. In [9] is described

an algorithm to extracting social relationships of a specific internet domain using e-mails address. This algorithm is used for social engineering attacks in security of computer systems area and has a high cost of extracting the social network, for each pair of e-mail address a query must be performed. This paper proposes a social network extraction algorithm based on e-mails to improve the complexity of reviewed algorithms. The query model that is used makes it more robust to ambiguity, and allows a clear interpretation of the relationships extracted. It does not use learning process. Unlike previous work [5], the proposed algorithm does not need an initial list of e-mails and it uses heuristics for driving the construction of a social network by topic and by importance of the network members.

The practice cases presented in this paper have been developed with TREDAR tool. This tool permits to define business processes in an interactive way. The specific analyzed case is a competitive intelligence process [19] that permits recollect and extract data from internet about VSST community and perform different types of analyses for the decision making to solve different queries. This tool also allows us to spread and visualize the obtained data to decision-making support. In this way, all the necessary stages to provide a vigilance service and competitive intelligence are integrated through web technology.

3 Social networks topic-driven extraction algorithm

In this section the algorithm to social networks topic-driven extractions formalize in [6] is described. The algorithm is divided into three steps:

- To seed e-mails address extraction: To start the extraction process is necessary to have a set of e-mails. Different processes for extracting e-mail addresses from web can be defined. The e-mails address selection process used by de algorithm is described below:
  - To perform queries with the '<topic>' and visit the web documents returned by the search engine
  - To extract the e-mails from the <a> tags of the html pages and
  - For each e-mail address, to check through the query '<email>' '<topic>' the association degree of e-mail to the topic. The process considers a high association degree if the number of documents returned by the previous query exceeds a threshold.
- To expand the e-mail address: The algorithm expands the social network with new relationships extracted from search engine queries result. For each mail address the algorithm do:
  - To create a query using the schema: '<username>' '<domain>' filetype:pdf. For example, the username of the e-mail address ffrosat@ us.es is ffrosat and the domain is us.es, hence the query is: 'ffrosat' 'us.es' filetype:pdf and extracts the result contexts.

  - The contexts were analyzed using regular expressions, to analyze only the contexts in which verify the e-mail appearance. Each new e-mail that appears in the context is added to the social network as a new node and the relationships are added. The nodes and the relationships are associated with a counter which will represent their importance in the network. If any e-mail address appears again at some context, the counter will increase by one unit, also the relationships.
- Social networks topic-driven: This principal process is iterative and stores the e-mails address in a priority queue.
  - Initially, the queue is initialized with the e-mails address seeds.
  - In each iteration an e-mails address is extracted from the queue top, and it is checked the association degree of the e-mail address to the topic. If the association degree exceeds a threshold, then it is expanded with the neighbors of the e-mails address.
  - The iteration is repeated until the queue is empty and other strategies can be implemented, for example limiting the number of web pages visited or the number of emails-address of the social network.
  - After completing the expansion, the new nodes are added to the queue.

Although it is possible that a person can have more than one e-mail address (for example a personal and a institutional e-mail address), it is also true that an e-mail address typically identify a person. It could avoid the ambiguity problems of other methodologies, but not the variety problems. Rearranging the queue several driven strategies can be implemented: maximizing the association degree of e-mail to the topic or other social networks analysis measures such as the degree, pagerank or the betweenness.

4 Competitive Intelligence Tasks

In order to illustrate the **social networks topic-driven extraction** technique, the 'VSST' and 'Interlligence competitive' topic has been used. The goal is the extraction of the social network of the community of the several Veille Stratégique Scientifique & Technologique (VSST) editions. The 327 seeds were extracted from different VSST web pages editions. After running the extraction algorithm, the social network had 2.107 nodes and 7.102 edges. Of these nodes, 432 nodes exceeded the minimum threshold, only these nodes were expanded.

Table 1. Numbers of documents and queries associated to each topic

Query	#Queries	#ndocs>0
<login> <domain> filetype :pdf	932	
<email> 'VSST'	932	377
<email> 'competitive intelligence'	932	367
<email> 'text mining'	932	435
TOTAL	3728 (4 days)	

The social network extracted can be used to find experts in certain topics, [15] uses the number of pages for develop experts ranking. In our case we can use the query <email> "<topic>" to estimate de number of documents associated and develop a topic ranking of experts, as it is shown in Table 1 and Table 2. There are alternatives to the impact measure such as Mindshare. The advantage that we have proposed in the experts finding task is the own of the social network, this allows using the ARS measures to classify the experts, for example: degree, authority, centrality, betweens, etc [16].

Table 2. Numbers of documents associated to each topic

VSST <email> "VSST"		COMPETITIVE INTELLIGENCE <email> "competitive intelligence"		TEXTMINING <email> "text mining"		MIN(VSST+IC,DEGREE)	
Bernard Dousset	86	David Doose	200	M. Boughanem	118	Bernard Dousset	0,525
Y. Bertacchini	41	Eric Andonoff	200	Josiane Mothe	104	Luc Quoniam	0,44
Odile Thiery	34	Luc Quoniam	125	Alessandro Zanasi	50	Bertacchini	0,32
Jacques Ducloy	26	Bernard Dousset	105	Luc Quoniam	43	Marisela Rodriguez-Salvador	0,295
Humbert Lesca	24	Bertacchini	95	Pascal Poncelet	31	Henri Dou	0,25
Amos David	22	Humbert Lesca	59	Stanley Loh	26	Yara Rezende	0,2
Henri Dou	19	Marisela Rodriguez-Salvador	59	Key-Sun Choi	25	Eric Andonoff	0,2
Luc Quoniam	10	Dou Henri	50	Jian-Yun NIE	25	Odile Thiery	0,19
Humbert Lesca	10	Yara Rezende	43	Alessandro Zanasi	24	Carlos Merino	0,16

An appropriate selection of email addresses can be used to model an area of interest and optimize the gatherer of information. For example, using the query <email> filetype:pdf to download pdf documents of experts in the area. This gathering of information can be used to refine the expert ranking. In order to analyze the global topics impact of the social network, it is possible analyzing the impact of products. In [5] this approach is used to analyze two ARS tools, Pajek and Ucinet. It concluded that the global impact of Ucinet on the social network was a 20% larger than Pajek. Probably the difference of impact is due to its usability. The mayor problem of this approach is the words ambiguity.

By using Figure 1 it is possible to analyze the cooperation relations between the countries of the VSST participants. In this case underlines the strategic position of France and his historical connecting with the Maghreb countries. And using Figure 2 is possible to analyze the cooperation between different organizations through their

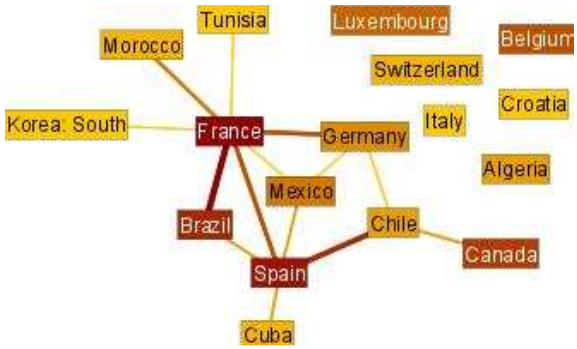

Fig. 1. VSST countries collaborations

domains. For example, analyze the domains we can discover new competitors (IALE, CDE, ISCOPE, IMCSLINE, etc), new clients (LAPOSTE, CEA, EADS, etc) or new investigation centers (IRIT, INIST, etc). And analyzing the relationships we can appreciate the central position of diverse universities in the network. Also using the IP domains is also possible to allocate the organizations [5].

Figure 3 represents in a lexical network the VSST community topics of interest. In order to build this map, two tasks have been executed: (1) the key words of the paper published in VSST events in 2007 have been extracted (2) the concurrency network has been calculated [18]. Using this map, it is possible to determine the interesting centers of the researcher that participate in the events. The follow topics of interest have been extracted in this example: text-mining, natural language processing, creativity and innovation, information retrieval and filtering, co-word analysis, business-intelligence.

The results for the VSST case study are online available in http://www.lsi.us.es/~ffrosat/index.php/Ffrosat/MapaListaVSST2007Es.

5 Conclusions and future work

In this paper some techniques have been combined helping in the making-decision process of an research organization. The topics that have been used are : extraction of information from the web, analysis of social networks and co-occurrences.

The queries that can be solved with these techniques are :

- Which are the experts in an specific area?
- Which are the most influence groups ?
- How the organizations are structured?
- Which are the most influential countries ?
- Which are the goal network?

This work provides a framework for making-decision processes to analyse the distributed data in different and heterogeneous sources, non centered in any database. Figure 4 shows in a visual way the main ideas developed in this work.

As future work we propose an automatic classification of the domains using the available information in the researh center webs. Also it will be interesting to develop techniques to extract information about the entities to be represented in the maps (names, departments, telephone numbers…)

Acknowledgments. This work has been partially funded by Junta de Andalucía (P08-TIC-04095).

References

1. K. Crowston and J. Howison. (2005) The Social Structure of Free and Open Source Software Development, Vol. 10, No. 2, First Monday.
2. A. Culotta, R. Bekkerman, and A. McCallum. (2004) 'Extracting social networks and contact information from e-mail and the web', Proceedings of the Conference on Email and Spam.
3. F. de la Rosa T., S. Pozo and R. M. Gasca. (2005) 'Análisis y visualización de comunidades científicas con información Extraída de la web', IEEE Latin America Transactions, Vol. 3, No. 1, ISSN: 1548-0992.
4. F. de la Rosa T, R. M. Gasca, L. González y F. Velasco (2005). Análisis de Redes Sociales mediante Diagramas Estratégicos y Diagramas Estructurales. Redes: Revista Hispana para el Análisis de Redes Sociales. ISNN: 1579-0185. Vol. 8.
5. F. de la Rosa T., F. and Gasca, R.M. (2007) 'Sistemas de inteligencia web basados en redes sociales', Revista Hispana para el Análisis de Redes Sociales, Vol. 12, ISSN: 1579-0185.
6. F. de la Rosa T. and R. M. Gasca (2008) Automatic extraction of social networks by topics of interest. IJCAT , Vol. 33 , Nr. 4, p. 292-299.
7. P. Jaccard (1901) 'Étude comparative de la distribution florale dans une portion des Alpes et des Jura', Bull Soc Vaudoise Sci Nat, Vol. 37, pp.547–579.
8. H. Kautz, B. Selman and M. Shah (1997) 'The hidden web', AI Magazine, Vol. 18, No. 2, pp.27–35.
9. J. Long (2005) The Google Hacker's Guide, Retrieved, ISBN 1-59749-176-4.
10. Y. Matsuo, H. Tomobe, K. Hasida, and M. Ishizuka. (2003) 'Mining social network of conference participants from the web', Proceedings of the International Conference on Web Intelligence, pp.190–194.
11. Y. Matsuo, J. Mori, M. Hamasaki, H. Takeda, T. Nishimura, K. Hashida and M. Ishizuka. (2006) 'Polyphonet: an advanced social network extraction system, Proceedings of WWW2006.

12. P. Mika. (2004) 'Bootstrapping the FOAF-Web: an experiment in social network mining', Proc. 1st Workshop Friend of a Friend, Social Networking and the Semantic Web.
13. P. Mika. (2005) 'Flink: Semantic web technology for the extraction and analysis of social networks', Journal of Web Semantics, Vol. 3, No. 2.
14. P. Mutton. (2004) Inferring and Visualising Social Networks on Internet Relay Chat, InfoVis, Austin, TX, pp.35–43.
15. Da Silva, A, Manniana, B., Quoniam, L. and Rostaing, H. 'Searching for Experts on the Internet' Competitive Intelligence Rewiew, USA, v. 11, n. 4, 2000.
16. Scott, J. P. (2000) 'Social Network Analysis: A Handbook. Second edition' Sage Publications.
17. X. Canaleta, P. Ros, A. Vallejo, D. Vernet and A. Zaballos. "A system to extract social networks based on the processing of information obtained from Internet", Proceedings of the 11th International Conference of the Catalan Association for Artificial Intelligence(CCIA 2008), IOS Press, ISBN: 978-1-58603-925-7, 2008
18. N. Coulter, I. Monarch and S. Konda. (1998). "Software engineering as seen through its research literature: A study in co-word analysis". Journal of the American Society for Information Science, 49(13), pp. 1206-1223.
19. P. Escorsa P.y R. Maspons (2001), De la Vigilancia Tecnológica a la Inteligencia Competitiva. Madrid. Prentice Hall.
20. V.A. Bucheli y F. González (2007) Herramienta informática para vigilancia VIGTECH-. Avances en Sistemas e Informática, v. 4 n.1, pp 117-126, Junio 2007, Facultad de Minas, Universidad Nacional de Colombia, Medellín

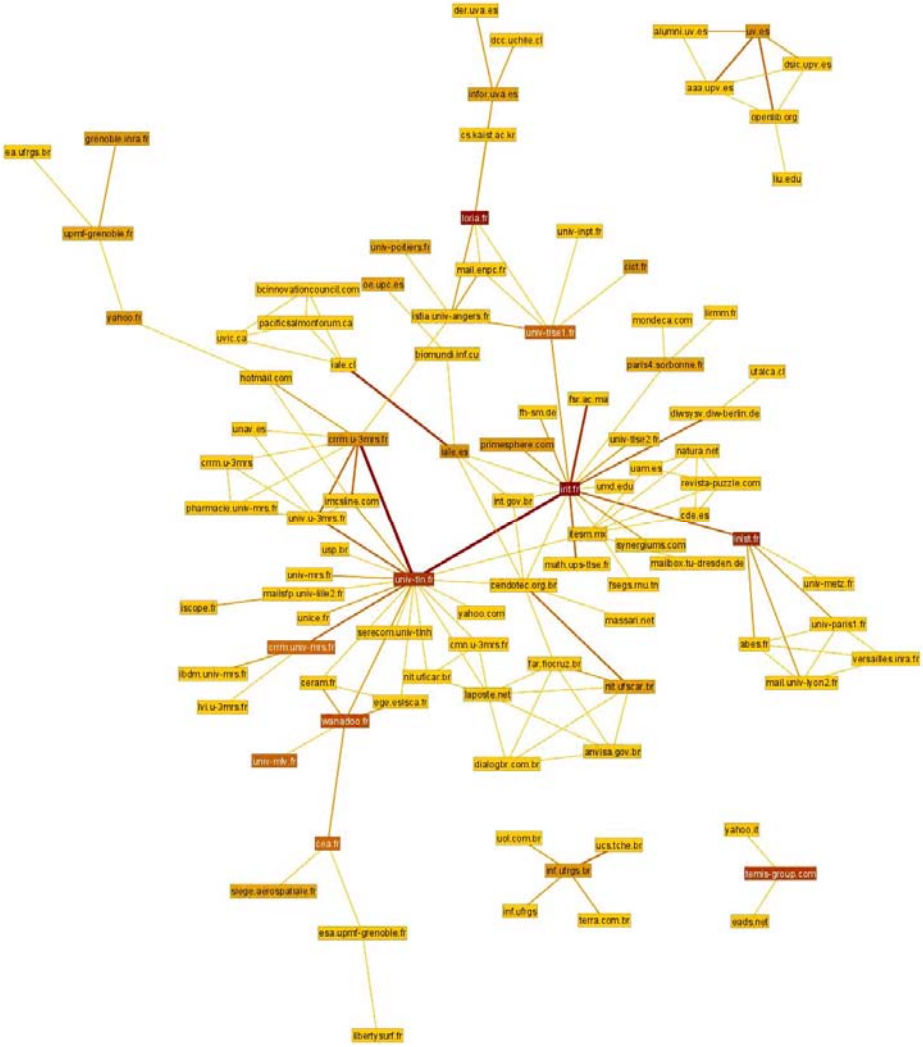

Fig. 2. VSST organizations collaborations

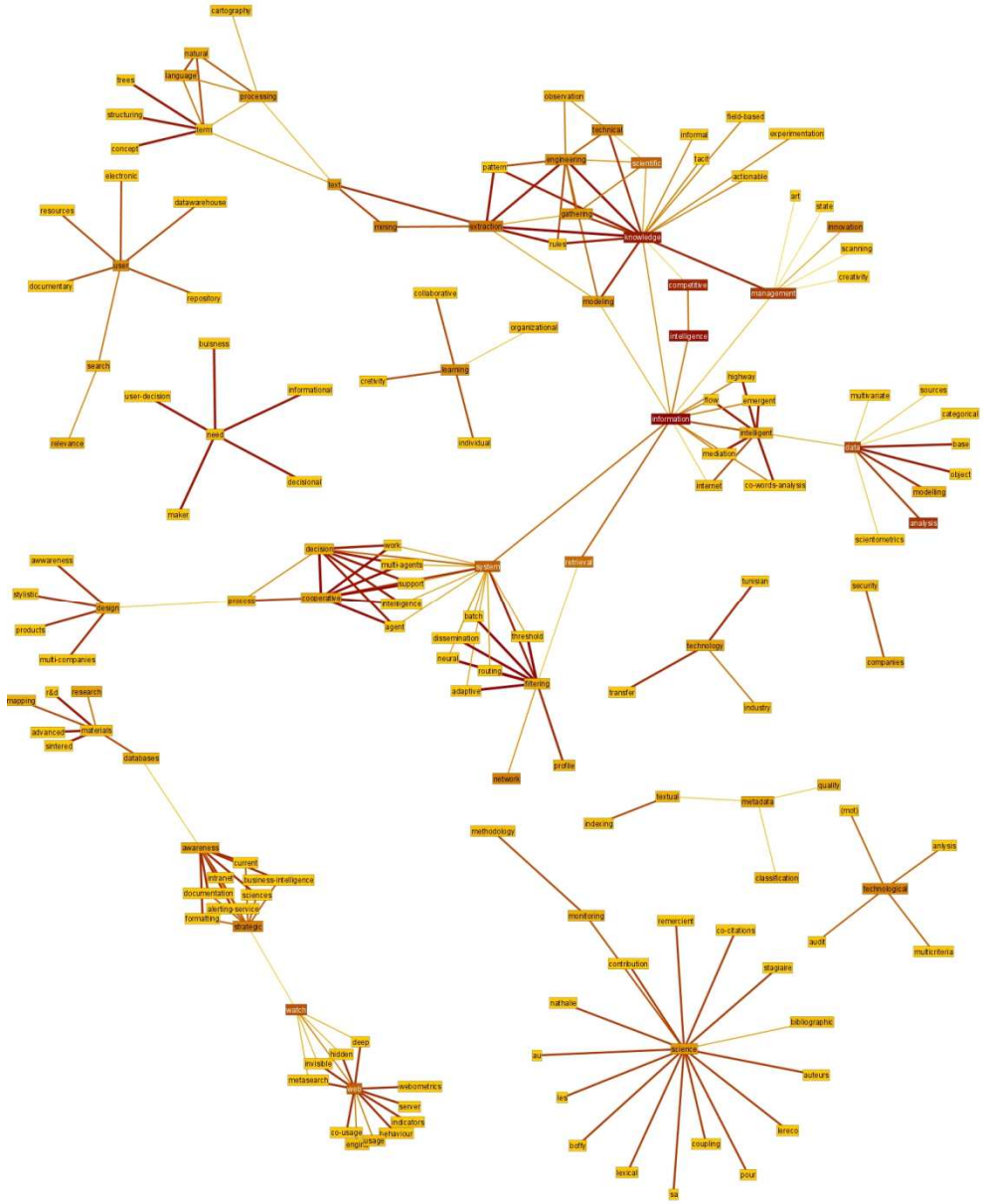

Fig. 3. VSST topics of interest

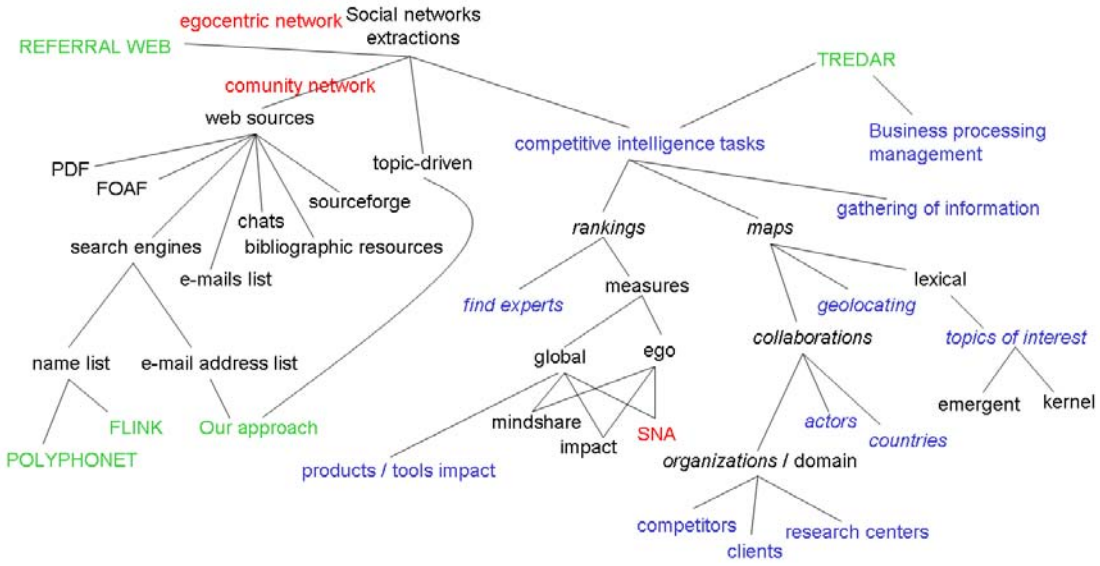

Fig. 4. Work mind map