ANALYSING AND VISUALIZING POLISH SCIENTIFIC COMMUNITY[1]

Piotr Gawrysiak, Dominik Ryżko

***Abstract**: This paper describes the rationale and partial results of an ongoing experiment aiming to perform analysis and visualization of social and organizational structure of Polish scientific community. Work described in this paper concerns automated extraction of information (related to DSc theses) from Internet based sources and visualization of resulting data set. Primary database that was mined was a scientific information repository "Polish Science ("Nauka Polska" in Polish) maintained by the Information Processing Centre (OPI – "Ośrodek Przetwarzania Informacji" in Polish). The nature of this repository, specifically lack of any data export facility and web scrapping prevention provisions implemented (despite the fact, that database is supported by public funds), complicated data extraction process. However, when finally downloaded, verified and processed the data set proved to be very interesting and valuable, making possible statistical analysis and geospatial visualization. Specifically, this paper describes creation of a software tool able to create geographical maps depicting collaboration between various research institutions in Poland during the process of DSc theses review. It should be regarded as a report on a work in progress. It is a part of larger SYNAT project, being a government funded initiative to create an ICT infrastructure supporting scientific collaboration and exchange of research data in Poland. The results presented herein constitute just a proof of concept for a visualization approach that will be implemented when more detailed data, concerning Polish scientific community, will be collected during other SYNAT activities.*

***Keywords**: bibliometrics, information visualization, social network analysis, web mining*

***ACM Classification Keywords**: J.4 SOCIAL AND BEHAVIORAL SCIENCES*

***Conference topic**: Data Mining, Knowledge Acquisition, Intelligent Web Mining and Applications*

Introduction

Contemporary science is focused on collaboration. The nature of most important research problems - in engineering, but also in physics, biology, bioinformatics, medicine and many other fields - makes solitary research, carried out by an individual researcher or even by a single research center, impractical. One of the reasons is the size of the data and the complexity of the problems that are being tackled by modern science. Another one is the Internet that finally became a useful collaboration tool, linking together research centers across nations and even across continents.

Obviously the science itself was always about sharing results with fellow researchers. The more popular was a given scientific treatise, the more widely cited and commented, the more important it was for subsequent research. However, with the advent of professional bibliometrics in XXth century, the citation count became a driving factor in modern science [Walter, 2003]. On might even risk a statement that nowadays the merits of scientific research are much less important than a number of people knowing about this particular research and

[1] This work is supported by the National Centre for Research and Development (NCBiR) under Grant No. SP/I/1/77065/10 by the strategic scientific research and experimental development program: "Interdisciplinary System for Interactive Scientific and Scientific-Technical Information"

referring to it. This however, being nothing else than a *popularity* of a given idea or a researcher, is related not only to the quality of the research, but also to the quality of marketing (or public relations) strategy employed by the research center or by given scientists. This became especially evident in recent years, with exploding popularity of the Internet, which is used as a primary tool for disseminating scientific information. The teams that are able to use the Web more effectively are also able to promote their research better, which increases the chances of getting high citation count. This could be potentially disadvantageous for disciplines (e.g. in humanities) or teams (e.g. based in less developed countries, where access to computing and communication infrastructure might be problematic) which have not yet fully embraced the Internet as a research tool.

On the other hand, the popularity of the global network as a tool used by global scientific community presents an opportunity for better understanding the collaboration patterns that emerge within this community. Due to the largely open nature of the Internet, but also thanks to the adoption of "open access" [Antelman, 2004] as a one of the most important models of publishing scientific papers, gathering statistical information concerning collaboration patterns between researchers and teams became relatively easy. Of course lack of standardization of protocols and data formats (even for such seemingly simple information as bibliographic references) complicates this process, but it can be – for the first time in history of science – automated.

This is therefore obvious that the growing importance of the Internet presents itself both as an opportunity for a local scientific community and as a threat that needs to be addressed. In recent years most developed countries have launched various initiatives, targeting these two issues. Poland is no exception here. Last year a strategic government project was funded, named SYNAT. It is aiming to improve national and regional (i.e. in the area of Central and Eastern Europe) web based science infrastructure and to map and analyze the state of Polish scientific community. This paper presents information on some initial analysis experiments related to this task.

The paper is structured as follows. Second and third chapters contain an overview of the entire SYNAT project and provide background information concerning the nature of IT systems that the project is implementing. Next chapters describe a web mining experiment, in which data concerning DSc theses has been extracted from the database maintained by the government agency "Information Processing Centre" ("Ośrodek Przetwarzania Informacji" in Polish) and later visualized using geospatial approach. Finally, the last chapter contains concluding remarks and a vision for future work.

Building science infrastructure - the SYNAT project

The research that is the subject of this paper has been carried out within a project, funded by the Polish Ministry of Higher Education and is being implemented by the consortium of Polish universities (see the project website [Synat, 2011] for a complete list). The project is part of a larger effort aiming to improve the state of scientific research in Poland and also in Central and Eastern Europe. One of the main obstacles hindering the growth of scientific research, and perhaps even more importantly, making the distribution of research results more difficult than necessary is lack of information exchange systems pertaining to these results.

In short, the lack of standards and systems supporting storage of scientific oriented information caused large distribution of repositories, storing data mostly in unstructured formats. Even information which is relatively easy to structure, such as bibliographical data, is not stored in an easy to process way. These problems are common both for large universities and research institutes. As a result, the visibility of Polish science in the Internet is very poor and particular centers of research are often not aware of the work conducted by others so intensive scientific collaboration between Polish institutions is relatively rarely implemented in practice. Much more important effect is however a distorted view of Polish science, because the apparent activity of local scientists is much lower than their real efforts and results of their work.

The system to be developed in the SYNAT project is planned to be a heterogeneous repository of data coming from various structured and unstructured sources. Hosting capabilities will be provided to address issues described above. This will be helpful especially with respect to low funded centers of research, which do not have capabilities to set up extensive repositories; additionally it is also required in order to acquire data from external sources. This includes large repositories of scientific papers and information about researchers and projects. In short, the project aims to create a backbone for scientific information storage in Poland. The envisaged system should also be able to retrieve potentially useful unstructured information from the Internet. Blogs, forums, project homepages, science funding schemes etc. constitute a large fraction of available knowledge scattered across the Web. Using such sources means that acquired data can contain missing information, errors or overlaps. The system should be able to: identify duplicates, merge partly overlapping objects, identify object versions, verify completeness of data objects (e.g. bibliography items), identify key words and proper names etc. It is required that the system will be able to perform search for new resources, especially in the areas heavily searched by the end users. The user should also be able to query the system repository but also start an off-line search process in order to discover resources according to specific requirements. Any new findings relevant to a particular user profile should be reported. Once discovered, sources of data have to be monitored in order to track any changes to their contents

Information storage capabilities drafted above constitute, however, only a part of the system's capabilities. Apart from just being able to store results of research in the form of publications and experimental data, the system should also support the research process itself, by providing tools for rapid dissemination of partial research efforts, for discovery of institutions or groups with similar research interests or even supporting some computationally intensive applications on the system infrastructure itself. The system is not being designed as a computational grid, however the distributed nature of storage subsystem means, that it can be also, for some specific use cases, utilized to perform some calculations (such as creating visualizations or statistical and/or knowledge discovery computations).

In a sense, the system's functionality in this context (i.e. not directly related to storage and distribution of bibliographical data) will be similar to the functionality of a social network system. Obviously calling such a system the scientific social network (or even “Facebook for scientists” as it has been unofficially dubbed) might be an overstatement. However, the similarities between Facebook and the system are also visible in a way in which its services are exposed to external parties. The system is structured as a set of modules with well-defined functionality and data types that can be interconnected by external institutions by using public system APIs. In this way it is possible to extend the functionality of the system or rather build other, specialized research support tools basing on the system infrastructure. Such tools could be both of commercial and open nature and possibly in the long run a larger ecosystem of services, supporting the scientific community, could be created.

System architecture overview

The requirements described in the previous chapter indicate an explicit distributed nature of the problems to be addressed. On the data acquisition side, the Internet is a network of loosely connected sources, which can be processed more or less in parallel. On the end user side, each one of them can generate concurrent requests for information. At the same time these parallelisms do not forbid overlap or contradiction. All of the above calls for a highly distributed architecture, with autonomy of its components, yet efficient communication and synchronization of actions between them.

The envisaged approach is based on multi-agent paradigms, which introduce a concept of an intelligent, autonomous and proactive agent. Various agent roles have been identified while analyzing processes to be implemented in the system. Personal agents will be responsible for interaction with end users. They will receive

queries, preprocess them, pass to the knowledge layer and present results returned from the system. User feedback will also be collected here. Personal agents will store history of user queries and maintain a profile of interests to improve results and proactively inform the user about new relevant resources.

The main data acquisition process will be performed by specialized harvesting agents. Their task will be twofold. Firstly, they will perform a continuous search for new relevant resources. Secondly, they will perform special searches for specific queries or groups of queries. The main task of harvesting agents will be to manage a group of web crawlers to perform the physical acquisition of data.

Different users can generate queries, which return partially overlapping results. This means some search tasks should be merged. On the other hand the same results can be delivered from different sources and only one of them should be used. All this means that matchmaking and coordination between demand and supply of data generates some sophisticated problems. This can be mitigated by introduction of brokering agents (also called middle agents) [Klusch2001], whose purpose is to coordinate efficient matching between personal agents and harvesting agents.

Special agents should be dedicated to the process of managing data already incorporated into the system. They will be responsible for finding missing data, inconsistencies, duplicates etc. Finding such situations will result in appropriate action e.g. starting a new discovery process to find new information, deletion of some data, marking for review by administrator etc.

The bottom layer of the system will consist of a group of web crawlers. They will search the Internet for relevant resources and pass the data to appropriate agents responsible for its further processing. The crawlers will use various methods (heuristics, machine learning) to perform focused crawling for new documents based on classified examples.

A focused crawler is a program that traverses the Internet by choosing relevant pages to a predefined topic and neglecting those out of concern. The main purpose of such a program is to harvest more information on the topic that matches the expectation of the user while reducing the number of web pages indexed. A focused crawler has three main components: URL queue (container of unvisited pages), downloader (downloads resources from WWW), classifier (compound model which categorizes the type of information resource, and its domain).

The crawler's classifier includes two modules: extraction module and relevance analysis. The extraction module parses the web page, and identifies the main parts of it. Humans can easily distinguish the main content from navigational text, advertisements, related articles and other text portions. A number of approaches have been introduced to automate this distinction using a combination of heuristic segmentation and features.

In PASSIM we would deploy the solution proposed in [Kohlshutter, 2010]. In this work a combination of two features was used - number of words and link density. This approach leads to simply classification model that achieves high accuracy. Web pages are segmented into atomic text blocks using html tags. Found blocks are then annotated with features and on this basis classified into content or boilerplate. The features are called shallow text ones. They are higher, domain and language independent level (i.e. average word length, average sentence length, absolute number of words). Atomic text blocks are sequence of characters which are separated by one or more html tags, except for “A” tags.The presence of headline tag, paragraph, division text tag are used to split content of web page into set of structural elements. To train and test classifier for various feature combinations we would use well known scientific conference, journal page, home pages of scientists, research institutes. The labeled set is then split into training and a test set (using i.e. 10-fold cross validation) and fed into a classifier mode l(Support Vector Machine).

Relevance analysis uses the significant parts of web resource, which were detected by above described extraction module. It would use intelligent classifier to categorize resource as scientific or not. It is also possible to

do first domain classification, and label document to the field of science (i.e. biology, history, computers). The analysis of topic similarity is the most important stage for a topic-specific web crawling. The relevancy can be determined by various techniques like the cosine similarity between vectors, probabilistic classifiers, or BP neural network. During relevance analysis it may be useful to identify type of resource which was positive categorized. Machine learning classifier is trained with features, which includes i.e. hosting domain, non HTML markup words, URL of page, outgoing links. Such model could be enough relevant to classify resource as home-page, institute page, conference, or blog. Type of resource may be used as a filter during invoking searching process.

We have two types of focused crawlers. First is used in harvesting mode to detect all probably scientific resource, which are further processed by middle layer (specific agents). The second type of crawler is used to find resources relevant to natural language query (NL query). Natural query would be provided by user in similar way as we type queries in Google search engine interface. Next, this query would be analyzed in the context of ontology which describe meta-data of conferences, journals, articles, institutes, researchers. Ontology gives us ability to make user's query much more semantic and structured. This approach achieves advantage over Google like search engines. General purpose searchers could not be ontology-driven, because it is impossible to build ontology of whole world.

Web crawlers have URL queue which contains a list of unvisited web resources. In harvesting mode this queue would be initialized with seed URLs. Seed URLs may be built on data taken from e.g. DBLP [DBLP, 2011] and Citeseer. Those two sources have links to relevant sources, but also meta-data concerning author names, title, publication date. Found URL's within DBLP and Citeseer would build initial URL queue. The mentioned meta-data would be entered to Google Scholar service and returned URL's could enrich seed entries.

The classifier analyzes whether the content of parsed pages is related to topic or not. If the page is relevant, the URLs extracted from it will be added to queue, otherwise will be discarded.

The process described above results in discovery (hopefully) of several classes of resources for various fields of science. Therefore, the data has to be properly classified according to the type of information it represents (e.g. scientific paper, blog, conference homepage etc.). Once we have such classification we can use ontology to decompose the document into appropriate components. For example, if we deal with a scientific paper, we will expect to find title, authors, affiliation, abstract etc. In the case of a conference website, the ontology will tell us what are the roles related to a scientific conference (general chair, organizing chair, program committee member etc.), what is a special session, a paper and so on. Another dimension for classification of resources is the field of science which they belong to. Finally documents – classified and decomposed - will be stored in the system repository. From there they can be accessed by the system users. They will also undergo further processing in order to improve knowledge quality. Duplicates will be eliminated, missing information filled, inconsistencies resolved etc.

More detailed description of the SYNAT project or the infrastructure that will be implemented is out of the scope of this paper. For more information see e.g. [Bembenik, 2011] or [Synat, 2011].

Data extraction from Web resources

Automated or semi-automated extraction of data from web resources, such as roughly outlined in previous chapter, is planned as one of the most important functionalities of SYNAT infrastructure. Because the nature and quality of the data that is currently available is highly inconsistent, various approaches need to be adopted in order to efficiently gather scientific information. From this perspective existing web resources can be roughly classified as follows:

- *structured data sources exposing consistent API* – this category includes scientific information databases (both bibliographic and containing experimental data) that have predetermined structure and that can be accessed not only via web interface (designed for human users) but also via an API, allowing querying the database and downloading contents in a structured format (e.g. in XML or JSON). Examples of such databases include DBLP Computer Science Bibliography [DBLP, 2011] or European Nucleotide Archive [ENA, 2011]. Such databases are obviously the most easy target for web mining and could be even directly linked to other parts of SYNAT infrastructure (a technique which – in the case of bibliographic databases – is most commonly referred to as "*harvesting*" [Sompel, 2004]);
- *structured data sources without external API* – this includes resources, that have an internal controlled structure (e.g. store the data in relational database), but which expose the contents only through web interface. Examples include reference databases such as Polish Science database [OPI, 2011], some open access journals, or information repositories maintained by universities. Such resources are in most cases quite easy to mine, however an initial processing step is usually required where a parser is constructed which is able to extract important structured information from contents of web pages generated by the resource;
- *semi-structured data sources* – this category includes these resources, that are created manually, but where an initial structure has been decided upon and is maintained, usually automatically by some kind of a content management system. Web newspapers and information portals such as CNN are probably most common examples of this resources; Wiki services (especially those using popular server backends such as MediaWiki) are another one – probably much more relevant in the context of scientific information retrieval. Information extraction difficulty is comparable to that described above, provided that some automated tools are available. For the SYNAT project several experimental systems, able to utilize machine learning techniques in order to automatically remove non-essential parts of webpages (such as navigational and static elements) as well as perform appropriate HTML stripping. Mentioned in previous chapter – see e.g. [Kolaczkowski, 2011] and [Kohlschutter, 2010];
- *handcrafted information repositories* – all other manually created data sources, such as personal and institutional web pages, project websites and even – in some cases – full scientific papers posted online. These resources usually require individual approach and in most cases are not susceptible to fully automated harvesting, however most valuable repositories will be imported to internal SYNAT database in later stage of the project.

Second category – the structured data sources without API – was selected as most important subject for initial experiments with web mining in SYNAT project. Most of the scientific databases available in Polish internet fall under this category, contrary to process of attaching external sources exposing well formed API to other IT systems is relatively trivial task, the project team expected some minor difficulties in this case. The aforementioned database Polish Science ("Nauka Polska" in Polish) was selected as one of the first systems that would be mined.

This database constitutes currently probably the most complete information repository regarding Polish scientists. It contains bibliographical records of over 121 000 people, who has been awarded a PhD in Poland, together with information about their professional affiliations, field of study and their professional career (including information about DSc theses and full professor titles awarded). The database is maintained by a government agency Data Processing Centre ("Ośrodek Przetwarzania Informacji" in Polish) [OPI, 2011] and quality of information contained there is generally considered to be high, mostly because of the fact that providing the agency with

relevant information by universities was (until recently) compulsory. The database provides also some information about scientific publications and events (i.e. conferences and symposia) but these parts are very incomplete.

The database is not equipped with any data export or external query provisions (albeit it is possible to formulate some simple queries via a web interface), but the generated HTML is quite consistent so no difficulties in data extraction were expected. However, it quickly turned out, that the database – despite the fact that it has been created and is maintained with public funds – is heavily protected from automated download attempts, blocking access as soon as after just 10 subsequent requests originating from a single IP number. Further investigation showed that one of the reasons is that the agency provides data processing services (such as performing specialized queries over database contents) for a fee, so naturally is not interested in facilitating access for other entities. Because the SYNAT project is also an official government initiative, negotiations concerning data access have been started, but a decision was also made to create a set of tools able to bypass the "data scrapping protection" mechanisms, as such tools might be useful also in future, when dealing with other systems.

The download framework has been implemented in Python and is able to use public anonymous proxy serves in order to "fake" access from the system that is being mined. The framework is able to automatically extract lists of proxies from various web sources (such as public proxy forums) and test their reliability. The accesses to a target database are also appropriately throttled down, with randomization applied both to the delays between consecutive HTTP requests, and to order of records that is being requested.

For test purposes a subset of the entire database was selected, concerning DSc theses defended during 2010. This dataset includes biographical information of authors and reviewers of theses (usually 4 independent reviews are required in the DSc process in Poland) thus allowing to analyze the relationship between various universities. Reviewers are usually invited to the events related to thesis defense by the author's university, so such occasions are usually an occasion for discussion with fellow researchers and in many cases also make possible to start collaboration between various research groups. The extraction process took approximately one week, however during this time the database was offline for two days – later it turned out that such intermittent outages are quite common and might happen even as often as several times a month.

Basic statistical information concerning extracted data is as follows: 783 records describing DSc defended during 2010 has been downloaded. These theses have been reviewed by 3132 university professors working in 1836 institutions, of which only two are outside of Poland. The institutions employing reviewers are located in 76 cities. The quality of records was highly uneven – some records contained full biographical info including list of publications of a given scientists, some only cited the works he or she reviewed or defended, while in some cases only name and affiliation could be extracted. On the contrary, the quality of institutional records was very high, with regard especially to their address information, what turned out to be an important factor for further visualization experiment.

Data visualization

As mentioned above the data set that was extracted contains information about social interactions between Polish scientists. Of course this is not a level of daily communication that is exhibited (and frequently visualized – see e.g. [Butler, 2010]) in social networks such as Facebook. However, due to the reasons mentioned above, it should allow at least to identify the associations between research institutions located in various parts of the country, that under normal circumstances , in day to day work, rarely collaborate together. For this purpose the data set was converted into a graph representation as follows: the bibliographic records of reviewers and authors have been parsed in order to extract information about their home institutions; the institutions records were analyzed in order to extract their addresses and the city portions of the address have been identified. Resulting

graph contains nodes corresponding to the cities, with vertices representing a “collaboration” events i.e. a meeting of two reviewers from two different cities during DSc defense.

Such graph can be obviously simply plotted e.g. via GraphViz and SFDP [Gansner, 2009], however because of the special nature of the data the most interesting way of visualizing it would involve placing it on a map. Of course in order to do it, a geographical coordinates of all the cities present need to be established, or “geocoded” and a specialized library was created that is able to do it for an arbitrary city name, by querying an external geospatial database. Google Maps [Maps, 2011] is probably one of the most popular such databases used currently, however it is not suitable for use in projects such as this due to processing limitations. Instead Geonames database [Geonames, 2011] (which is licensed under Creative Commons license) was used, and proved to be of high enough quality. For the actual drawing the Python Basemap toolkit was used [Basemap, 2011], allowing to incorporate a satellite ground map from the NASA Blue Marble project and draw the graph vertices not as straight lines but as parts of great circles (which will be of course more useful in future, when visualizations of collaboration with institutions in other countries will be performed; for the map of Poland parts of great circle are practically the same as lines). Various experiments with weight and coloring schemes of vertices were performed and finally a relatively simple method was adopted that associates the thickness and saturation of a line with logarithm of number of edges connecting two vertices. Resulting visualizations are presented below.

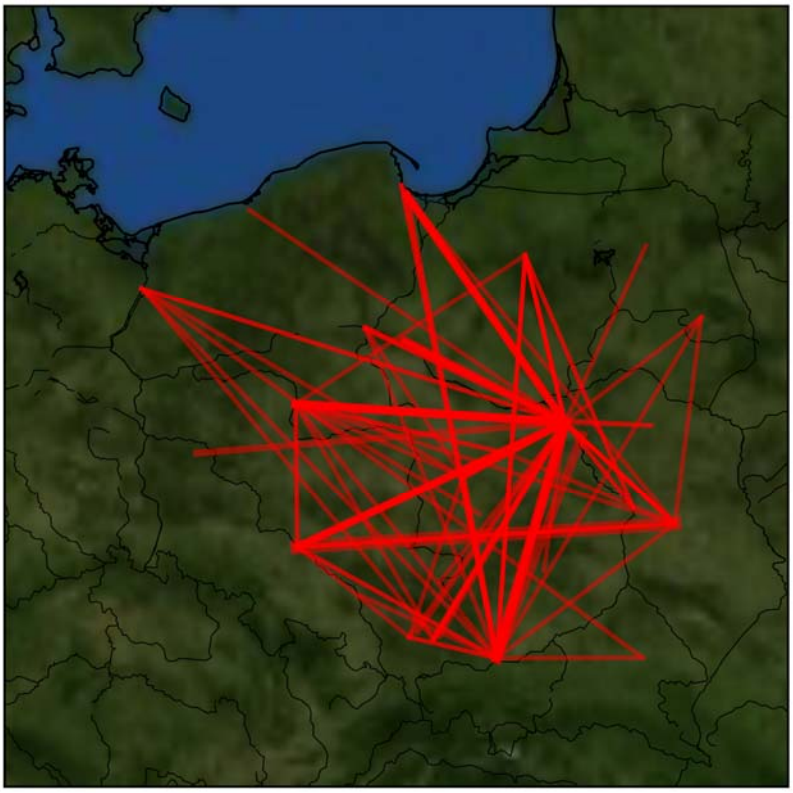

Fig. 1 Collaboration visualization depicting only most connected institutions (at least 3 edges in collaboration graph).

Fig. 2 Alternative approach to representing intensity of collaboration, using only saturation without altering line thickness.

Conclusions and future work

Relatively small temporal scope of the data set (data from only one year i.e. 2010) that was used in this experiment does not allow drawing well founded conclusions about the state of scientific collaboration in Poland and indeed the experiment was thought mainly as a proof of concept and will be expanded in future. However, even within small data set that was a subject of this one can make some interesting observations. For example – it is quite evident that the Warsaw is the most important city as far as scientific research in Poland is concerned, the area most active in research corresponds roughly to central and greater Poland (i.e. the areas that have not been part of Poland before II World War are seem to be not entirely integrated as far as collaboration is concerned).

As it is mentioned above these observations should be taken with a grain of salt. However during next phases of the SYNAT project the entire database maintained by OPI will be downloaded – so this experiment will be repeated, but this time with information spanning several tens of years. Additionally other visualizations will be prepared, using software created during this experiment, but with different data sets (first candidate is obviously an experiment similar to DSc analysis described herein, but concerning PhD theses).

Acknowledgements

This work is supported by the National Centre for Research and Development (NCBiR) under Grant No. SP/I/1/77065/10 by the strategic scientific research and experimental development program: “Interdisciplinary System for Interactive Scientific and Scientific-Technical Information”

Bibliography

[Antelman, 2004] K. Antelman, Do Open-Access Articles Have a Greater Research Impact?, College & Research Libraries vol. 65 no. 5 372-382, 2004

[Basemap, 2011] Basemap library documentation, http://matplotlib.sourceforge.net/basemap/doc/html/, accessed on 30/05/2011

[Bembenik, 2011] R. Bembenik et al. Retrieval and management of scientific information from heterogeneous sources. In: SYNAT Workshop 2011 Proceedings, Studies in Computational Intelligence, Springer Verlag, 2011

[Butler, 2010] P. Butler, Visualizing Friendships, Facebook, USA, 2010, http://www.facebook.com/note.php?note_id=46971639891

[DBLP, 2011] DBLP Bibliography, http://www.informatik.uni-trier.de/~ley/db/, accessed on 30/05/2011

[ENA, 2011] European Nucleotide Archive, http://www.ebi.ac.uk/ena/, accessed on 30/05/2011

[Gansner, 2009] E. Gansner et al., Efficient Node Overlap Removal Using a Proximity Stress Model, Lecture Notes in Computer Science, 2009, Volume 5417/2009

[Geonames, 2011] Geonames Database, http://www.geonames.org, accessed on 30/05/2011

[Klusch, 2011] Klusch M., Sycara K. P. Brokering and matchmaking for coordination of agent societies: A survey. In Coordination of Internet Agents: Models, Technologies, and Applications, pp.197-224. Springer, 2001.

[Kohlschutter, 2010] "Boilerplate Detection using Shallow Text Features" by Christian Kohlschütter et al., presented at WSDM 2010 -- The Third ACM International Conference on Web Search and Data Mining New York City, NY USA, 2010

[Kolaczkowski, 2011] P. Kolaczkowski, P. Gawrysiak, "Extracting Product Descriptions from Polish E-Commerce Websites Using Classification and Clustering". In: 19th International Symposium on Methodologies for Intelligent Systems, Lecture Notes in Computer Science, Springer Verlag, 2011

[Maps, 2011] Google Maps, http://maps.google.com, accessed on 30/05/2011

[OPI, 2011] Nauka Polska, Ośrodek Przetwarzania Informacji, http://nauka-polska.pl/, accessed on 30/05/2011

[Sompel, 2004] H. Sompel et al., Resource Harvesting within the OAI-PMH Framework, D-Lib Magazine, Volume 10 Number 12, 2004

[Synat, 2011] SYNAT Project website. http://www.synat.pl, accessed on 30/05/2011

[Walter, 2003] G. Walter et al., Counting on citations: a flawed way to measure quality, Medical Journal of Australia, Vol. 178, Nr. 6 (2003) , p. 280-281, 2003

Authors' Information

Piotr Gawrysiak *– Deputy Director for Scientific Research, Institute of Computer Science, Warsaw University of Technology, ul. Nowowiejska 15/19, 00-665 Warsaw, Poland; e-mail: P.Gawrysiak@ii.pw.edu.pl*

Major Fields of Scientific Research: natural language processing, text, web and data mining, mobile computing, human computer interaction, information visualization.