Retrieval and Management of Scientific Information from Heterogeneous Sources*

Piotr Gawrysiak, Dominik Ryżko, Przemysław Więch, and Marek Kozłowski

Institute of Computer Science, Warsaw University of Technology,
Nowowiejska 15/19, 00-655 Warsaw, Poland
{p.gawrysiak,d.ryzko,pwiech,m.kozlowski}@ii.pw.edu.pl

Abstract. The paper describes the process of automated retrieval and management of scientific information from various sources including the Internet. Application of semantic methods in different phases of the process is described. The system envisaged in the project is a scientific digital library, with automated retrieval and hosting capabilities. An overall architecture for the system is proposed.

1 Introduction

Rapid advancements in computing and networking technology, that took place during the last two decades, transformed deaply the nature of scientific research worldwide. Nowadays, it is difficult to even imagine conducting a successful research project – both in humanities and in engineering – without exploiting vast knowledge resources provided by the global Internet, and without using the same network to disseminate research results.

The nature of contemporary Internet, used as a research tool, is however drastically different from what was envisioned in the 90-ties. The Internet is just a haphazard collection of non-coordinated knowledge sources. Most valuable repositories are not even centrally controlled. It is sometimes very difficult to evaluate the quality of data contained in non-professional sources, such as some Open Access journals [11]. The situation described above basically means that the concept of Semantic Web [9], promising the coordinated global network of information, failed to materialize. One of the primary reasons for this failure is the difficulty of creating and maintaining useful ontologies, describing all aspects of the world, that would drive exchange of information in the Semantic Web [6], and only such approach would fulfill the needs of a general purpose semangic search engine. The main reason for this is a state of ontology engineering, which is still mostly a manual process, very time-consuming, expensive and error prone. While some automated, or at least semi-automated, ontology building methods that are able to leverage the amount of information present in ever growing repositories of text

* 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".

R. Bembenik et al. (Eds.): Intelligent Tools for Building a Scient. Info. Plat., SCI 390, pp. 37–48.
springerlink.com

data (e.g. obtainable via the Internet) have been created [4], their quality is still vastly inadequate.

In this position paper we argue that building some dedicated ontologies, especially as applied to scientific data and scientific communities, may improve the process of data acquisition. We believe that using contemporary knowledge discovery and natural language processing algorithms and methods, we can achieve much of the goals (as seen from an end user perspective) of the Semantic Web vision, which does not seem to be feasible as applied to the wole Internet. In particular, while it is not feasible to create ontologies to manage whole scientific knowledge, we can create domain ontologies for specific branches of science, which will significantly help in managing knowledge and in gathering new information.

The paper extends [5] and describes the design principles of the PASSIM project. It is a part of a broader strategic initiative of the Polish Ministry of Education and Scientific Research called SYNAT, aimed at creation of universal, open communication and hosting platform for scientific resources. Within PASSIM, led by Warsaw University of Technology, an integrated knowledge base will be created, which will enable acquisition of data from heterogeneous and distributed sources, its safe storage and analysis. On top of hosting capabilities, new services and applications will be built. The project has started in 2010 and will last till the end of 2012. The initial phase of the project has finnished with a analysis of application of Artifcial Intelligence methodologies in the process of data acquisition. The current phase involves development of algorithms supporting this process. Later on selected algorithms will be implemented and tested on real data.

This paper is structured as follows. Section 2 describes challenges targeted by the project and existing research results which can be used to solve them. In Section 3 requirements for the PASSIM project are listed. In Section 4 solutions for system implementation are proposed. Finally, Section 5 summarizes the results.

2 Challenges and Existing Solutions

As mentioned in the introduction, one of the main challenges in PASSIM is automated acquisition of knowledge from various structured and unstructured sources. Among these sources the Internet will play a major role. Despite the overwhelming amount of irrelevant and low quality data, there are several useful resources. This includes researchers' homepages and blogs, homepages of research and open source projects, emerging open access journals, university tutorials, software and hardware documentation, conference and workshop information etc. Finding, evaluating and harvesting such information is a complex task but nevertheless it has to be taken up in order to provide PASSIM users with a wide range of up to date resources regarding science as well as past and ongoing research activities.

Several approaches to harvesting information form the Internet have been proposed in the past. The most popular approach nowadays is the use of general purpose search engines. The improvement in the search quality caused that a vast majority of users say the Internet is a good place to go for getting everyday information [8]. Sites like Google.com, Yahoo.com, Ask.com provide tools for ad-hoc queries based on the keywords and page rankings. This approach, while very helpful on the day to day basis, is

not sufficient to search for large amounts of specialized information. General purpose search engines harvest any type of information regardless of their relevance, which reduces efficiency and quality of the process. Another, even more important drawback for scientists is that they constitute only a tiny fraction of the population generating web traffic and really valuable pages constitute only a fraction of the entire web. Page ranks built by general purpose solutions, suited for general public will not satisfy quality demands of a scientist. One can use Google Scholar, Citeseer or other sites to get more science-oriented search solutions. Although this may work for scientific papers and some other types of resources, still countless potentially valuable resources remain difficult to discover.

Another approach to the problem is web harvesting, based on creating crawlers, which search the Internet for pages related to a predefined subject. This part of information retrieval is done for us if we use search engines. However, if we want to have some influence on the process and impose some constraints on the document selection or the depth of the search, we have to perform the process by ourselves. A special case of web harvesting is focused crawling. This method introduced by Chakrabarti et al. [3] uses some labeled examples of relevant documents, which serve as a starting point in the search for new resources.

The task of retrieving scientific information form the web has already been approached. In [10] it is proposed to use meta-search enhanced focused crawling, which allows to overcome some of the problems of the local search algorithms, which can be trapped in some sub-graph of the Internet.

The main motivation for the work envisaged in the PASSIM project is to create a comprehensive solution for retrieval of scientific information from the heterogeneous resources including the web. This complex task will involve incorporating several techniques and approaches. Search engines can be used to find most popular resources with high ranks, while focused crawling can be responsible for harvesting additional knowledge in the relevant subjects. Additional techniques will have to be used to classify and process discovered resources.

Since many users will use the system simultaneously, a distributed architecture will be required. While this has several benefits regarding system performance, additional measures have to taken in order to avoid overlap in the search process [2]. Various parallel techniques for searching the web have already been proposed [1]. In the PASSIM project multi-agent paradigms will be used, which propose intelligent, autonomous and proactive agents to solve tasks in a distributed environment.

3 Project Requirements

The system that is the subject of this paper would be a direct result of a distributed research project, funded by the Polish Ministry of Higher Education. 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, with most of them created in unstructured formats. Even relatively easy to structure information 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. Obviously as a result 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 effort.

The system to be developed in the PASSIM project is thought to be a heterogeneous repository of data 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. It is also required to acquire data from external sources. This includes large repositories of scientific papers, 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, homepages, projects, conference calls, 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. This process can be greatly improved by relying on existing repositories (e.g. Geonames, LCSH, VIAF). Also idetification of synonyms and homonyms should help with the disambiguation. The same techniques should be applied to iterpret user queries, which is desribed later in the paper.

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.

The data harvesting process will involve a feedback loop. A user will be able to rate the relevance of the resources found. This information will be used to improve the search process as well as classification of documents. Based on the feedback, a user profile should be built, to personalize retrieval and presentation of information for a particular scientist.

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") 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.

4 Envisaged Solution

This section desribes the approach taken in order to reach the goals of the PASSIM project. This includes overall system architecture, information retirieval techniques, knowledge management and presentaion of data to the end users.

4.1 Architecture

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. An example process of user query answering has been shown in Figure 1.

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.

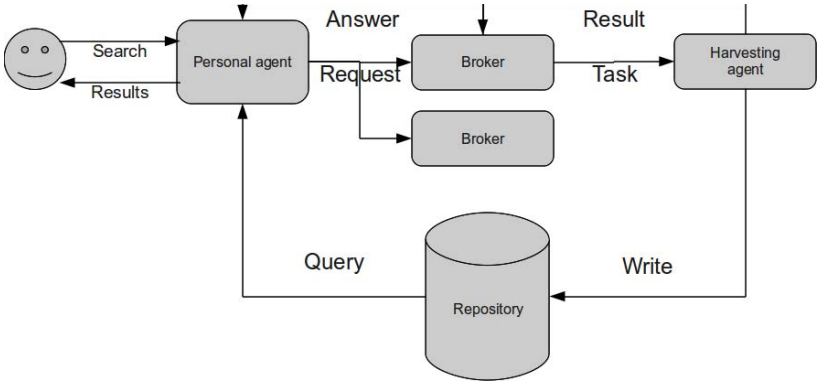

Fig. 1. Query answering process

Different users can generate queries, which return partialy overlaping results. This means some search tasks should be merged. On the other hand the same resutls 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) [7], 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 etc.) to perform focused crawling for new documents based on classified examples. The general architecture of the system has been shown in Figure 2.

While the system will be designed as a set of autonomous agents, we can also look at the system architecture from a data flow perspective. Figure 3 shows how information is passed between functional components of the system. The bottom component in this figure is the data *repository*, which holds all gathered information and provides access for reading, updating and searching its contents. The repository is backed up by an ontology defining the semantics of the contained data. As discussed earlier, we assume several domain specific ontologies maintained by communities interesed in particular subjects as well as a single ontology describing general concepts (e.g. conference, paper, author etc.). The remainder of the components can either be contained in one or more agents or can be viewed as tasks, which operate inside the system.

One of the processing paths deals with crawling Web resources and inserting preprocessed structured data into the repository. The starting points for the crawler can be provided by the results of a search engine, which is queried with selected keywords. This processing path will be described in detail further in the paper.

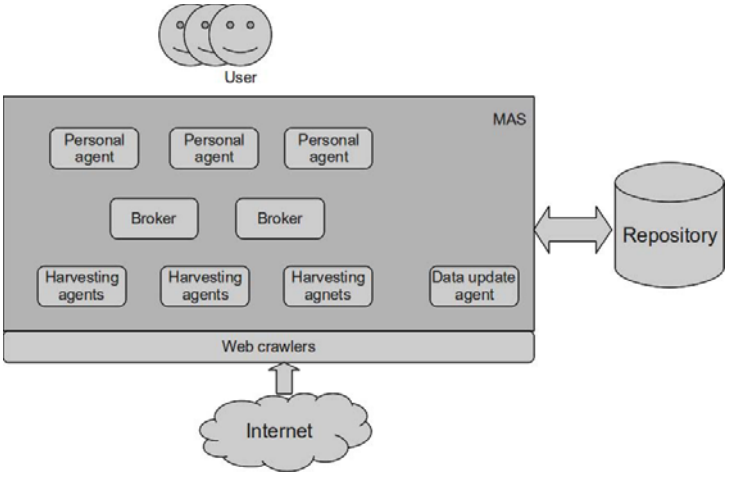

Fig. 2. System architecture

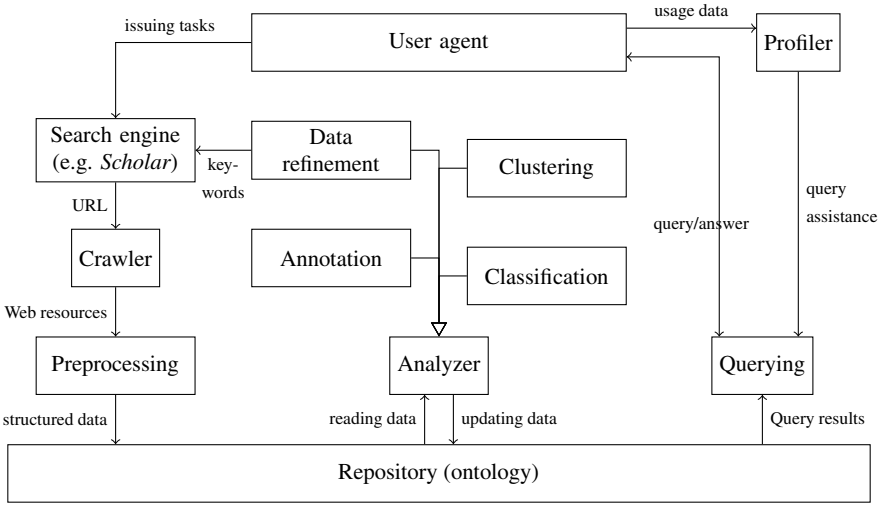

Fig. 3. System data flow

The central part of the diagram presents an architecture for building components, which further process the information that is stored in the data repository. An *analyzer* can execute a processing algorithm such as classification or clustering and based on the results it can add annotations or refine the data, which was input to the data storage. The *data refinement* component is intended to utilize the previously mentioned crawling subsystem to search for additional information about the objects occurring in the data.

The *user agent*, which directly interacts with the user of the system, receives queries and tasks from the user. This agent then queries the data repository for answers or issues tasks to the crawling modules to asynchronously find information relevant for the user. Additionally, the user agent utilizes the *profiler* to collect usage information and personalize the results for the particular user.

4.2 Web Crawling

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 [13]. In this work there were used combination of two features - 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. The found blocks are then annotated with features and on this basis classified into content or boilerplate. The feaures 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 model(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.

In the article [12] authors used Bayes Classifiers and improved TF-IDF algorithm to extract the characteristics of page content and compute relevance to topic. In the paper [14] was described the design and implementation of a university focused crawler that runs on BP network classifier for prediction of the links leading to relevant pages. BP is a three or multi-level topology structure. In [14] a three layer error back-propagation

neural network is used as a network model for predicting the relevance of the crawled page. Before training there is manually built database of keywords, which are closely connected with topics. The fetched pages are sent to extraction module in order to discover significant text parts, which are further matched with database of keywords, and the outcome is a boolean model. The number of inputs of neural network corresponds to the number of keywords in the database. The number of outputs corresponds to the number of topics. Trained BP network classifier achieves precision about 75%, which is the percentage of the Web pages crawled that are relevant to topics.

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 searcher. Next, this query would be analyzed in the context of ontology which describe meta-data of conferences, journals, articles, institutes, researchers. Ontology give us ability to make user's query much more semantic and structured. This approach achieves advantage over google like searcher engines. General purpose searchers could not be ontology-driven, becuase there 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 DBLP, Citeseer. Those two sources have links to relevant sources, but also meta-data concerning author names, title, publication date. Found urls within DBLP, Citeseer would build initial URL queue. The mentioned meta-data would be entered to google scholar and returned urls 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 of NL query-driven searching is as following:

- We begin with an existing manually-created ontology, describing publications, scientist, conferences etc. From hierarchically-structered ontology, using given natural languge query we generate set of alternative queries taking under account concepts and relations in the ontology.
- We use a topic-specific spider(focused crawler) to submit these queries to a variety of Web Search engines and digital libraries. The spider downloads the potentially relevant documents listed on the first page (top-ranked). We also provide options to customize the number of returned results, the formats of returned resources.
- Next crawler applies classifier to filter out documents that match the query but do not belong to scientific world. If in the query is included clear information about domain of interest, focused crawler initially categorizes document to specific domain (i.e. biology, history).

- If the user choose the type of searched resources(only blog, or only home-pages) there is used described above machine learning classifier to filter out irrelevant types.
- Collection of significant resources is processed by information extraction module(IEM). It extracts structured and useful information from the actual text of filtered resources. Simple extraction is used by crawlers to identify outgoing links and add them to crawler's URL query. Much more complex extraction operations (as summarizing) are done in the knowledge system.
- Each relevant resources are automatically described by agents with attributes from ontology. If some information are not presented in the selected document, there is sent query to focused crawlers. In this way we step-by-step fulfill meta-information.

4.3 Information Retrieval

Before becoming available to end users, knowledge has to be discovered and retrieved. As discussed in the Introduction, using general purpose search engines is not an option, while retrieveing specialized information. Therefore, more sophisticated method like focused crawling have to be adopted in order to discover and retrieve valuable data from the Web.

In the search process several classes of resources have to be discovered 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 an ontology to decompose the document into appropriate components. For example, if we deal with a scientific paper, we will expect to find title, authors, affiltiation, 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 classificaion of resources is the field of science which they belong to.

Classifed and decomposed documents 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 knowldge quality. Duplicates will be eliminated, missing information filled, inconsistencies resolved etc.

4.4 Semantic Analysis

When harvesting a new piece of knowledge from the web the system must know its semantics. Unless it is stated explicitly what is a scientific conference, how is it related to papers, sessions, chairman etc., it is not possible to extract automatically any useful information. To allow this task special ontology will be built called knowledge base ontology. It will define most important concepts and their respective relation. This step will be performed manually or semi-automatically. The knowledge base ontology will be used not only to provide semantics to the documents retrieved from the Internet, but also it will allow better interpretation of user queries.

Another group of ontologies will be constituted of domain ontologies. They will contain knowledge about respective fields of science, which will be covered by the project. These ontologies will be build automatically or semi-automatically out of the knowledge gathered in the system.

5 Conclusions

In the paper the concept of scientific knowledge acquisition from the internet in the PASSIM project has been presented. It has been shown shown why special approach is needed here and how semantic technologies play a crucial role in the process. The requirements for the system have been listed and problems to be faced have been outlined. The paper describes also envisaged solution and a general architecture of the system to be developed. The most important technologies selected to achieve the task are multi-agent systems and ontologies.

In the next stages of the project specific algorithms will be selected and implemented. It is important to verify system performance and usability across various branches of science and with large amounts of data. It seams obvious results will vary here. Some branches of science like Computer Science generate large ammount of structured data and valuable knowledge bases, which can be used as starting points (e.g. DBLP). Other especially social sciences will require much more effort to retrieve signifficant knowledge. From the point of view of semantic technologies, the important question to be answered is how complex ontologies will be sufficient to allow retrieval of interesting information.

References

1. Bra, P., Post, R.: Searching for arbitrary information in the www: The fish-search for mosaic. In: Second World Wide Web Conference, WWW2 (1999)
2. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston (1999)
3. Chakrabarti, S., van den Berg, M., Dom, B.: Focused crawling: A new approach to topic-specific web resource discovery. Computer Networks 31(11-16), 1623–1640 (1999)
4. Gawrysiak, P., Rybinski, H., Protaziuk, G.: Text-Onto-Miner - a semi automated ontology building system. In: Proceedings of the 17th International Symposium on Intelligent Systems (2008)
5. Gawrysiak, P., Ryżko, D.: Acquisition of scientific information form the Internet - The PASSIM Project Concept. In: Proceedings of the International Workshop on Semantic Interoperability - IWSI (accepted for publishing, 2011)
6. Gomez-Pérez, A., Corcho, O.: Ontology Specification Languages for the Semantic Web. IEEE Intelligent Systems 17(1), 54–60 (2002)
7. 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, Heidelberg (2001)
8. Manning, C.D., Raghavan, P., Schuetze, H.: An Introduction to Information Retrieval. Cambridge University Press (2008)
9. McIlraith, S.A., Son, C.T., Zeng, H.: Semantic Web Services. IEEE Intelligent Systems 16(2), 46–53 (2001)

10. Qin, J., Zhou, Y., Chau, M.: Building domain-specific web collections for scientific digital libraries: a meta-search enhanced focused crawling method. In: Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries (2004)
11. Suber, P.: Open access overview (2004), http://www.earlham.edu/peters/fos/overview.htm
12. Wenxian, W., Xingshu, C., Yongbin, Z., Haizhou, W., Zongkun, D.: A focused crawler based on Naive Bayes Classifier. In: Third International Symposium on Intelligent Information Technology and Security Informatics (2010)
13. Kohlschutter, C., Fankhauser, P., Nejdl, W.: Boilerplate detection using shallow text features. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining (2010)
14. Hua, J., Bing, H., Ying, L., Dan, Z., YongXing, G.: Design and Implementation of University Focused Crawler Based on BP Network Classifier In: Second International Workshop on Knowledge Discovery and Data Mining (2009)