Acquisition of Scientific Information from the Internet: The PASSIM Project Concept

Piotr Gawrysiak and Dominik Ryżko

Institute of Computer Science, Warsaw University of Technology
Nowowiejska 15/19, 00-655 Warsaw, Poland

Abstract. The paper describes the concept of automated acquisition of scientific information from the Internet. The work is part of PASSIM project a strategic initiative of the Polish Ministry of Education and Scientific Research. Different methods like web mining, data mining and other techniques of Artificial Intelligence will be applied in order to harvest, extract, classify and store science oriented information form the Web.

1 Introduction

Rapid advancements in computing and networking technology that took place during the last two decades transformed the nature of scientific research. 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 quality of data contained in non-professional source, such as some Open Access journals [9]. The situation described above basically means that the concept of Semantic Web [7], 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, that would drive exchange of information in the Semantic Web [5]. 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 data (e.g. obtainable via the Internet) have been created [4], their quality is still vastly inadequate.

In this position paper we argue that the Semantic Web strategy, especially as applied to scientific data and scientific communities, simply does not make sense. However we believe that using contemporary knowledge discovery and natural language processing algorithms and methods, we can achieve much of goals (as seen from an end user perspective) of the Semantic Web vision.

In this paper we describe the design principles of the PASSIM project. PASSIM is

Gawrysiak P. and Ryżko D..
Acquisition of Scientific Information from the Internet: The PASSIM Project Concept.
DOI: 10.5220/0003527008200087
In *Proceedings of the International Workshop on Semantic Interoperability* (IWSI-2011), pages 82-87
ISBN: 978-989-8425-43-0

a strategic initiative of the Polish Ministry of Education and Scientific Research aiming to create regional ICT infrastructure supporting storing, processing and sharing of scientific research data and results.

This paper is structured as follows. Chapter 2 describes challenges targeted by the project and existing research results which can be used to solve them. In Chapter 3 requirements for the PASSIM project are listed. In Chapter 4 solutions for system implementation are proposed. Finally, Chapter 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 search engines. The improvement in search quality caused that a vast majority of users say the Internet is a good place to go for getting everyday information [6]. 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 [8] 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 several 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 to be developed in the PASSIM project is thought to be a heterogeneous repository of data from various structured and unstructured sources. This means, that acquired data can contain missing information, errors or overlaps. In order to address these issues, methods for data cleansing will have to be introduced. To this end NLP (Natural Language Processing), text mining, data mining and other methods will be applied. As a result, the system should be able to: identify duplicates, merge partly overlaping objects, identify object versions, verify completeness of data objects (e.g. bibliography items), identify key words and proper names etc.

Before cleansing, the data has to be discovered and harvested. 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 information etc.).

Once the appropriate class is identified, its structure has to be decomposed. For example the system should know how a scientific paper is structured (title, author, affiliation, abstract etc.), what are the roles related to a scientific conference (general chair, organizing chair, program committee member etc.) and so on.

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 start an off-line search process in order to discover resources according to specific requirements. 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 relevance of the resources found. This information will be used to improve the search process as well as classification of documents.

4 Envisaged Solution

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 high level architecture of the system has been shown in Figure 1 below.

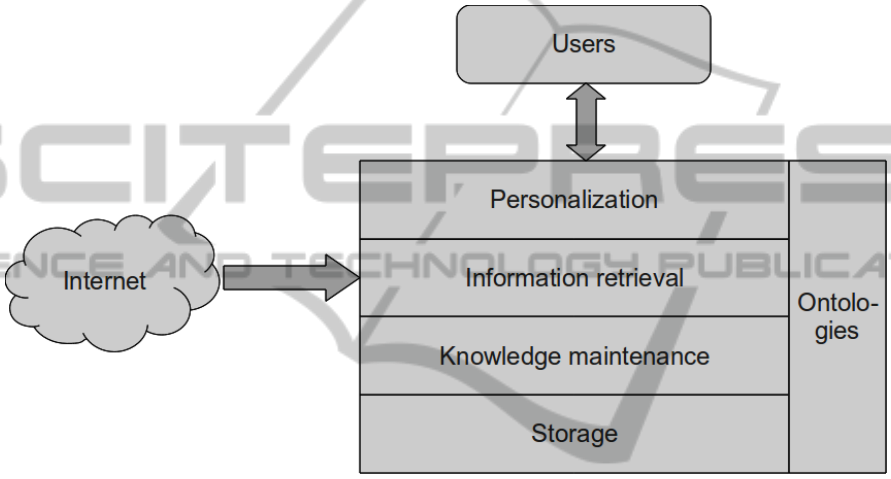

Fig. 1. System architecture.

The envisaged approach is based on multi-agent paradigms, which introduce a concept of an intelligent, autonomous and proactive agent. Various agent roles will be designed and developed. 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.

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 heuristics to perform focused crawling for new documents based on classified examples.

An important part of the knowledge acquisition process will be the classification of documents. Each document after being discovered and preprocessed needs to be properly labeled. Such classification is non-trivial and can be done along various dimensions. One aspect is what kind of information has been found e.g. scientific paper, science funding scheme etc. Another dimension is the field of science which is being referred in the particular document. To address all these aspects, a multi-step classification will be performed.

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 ontologies will be built. They will define most important terms and their respective relation. This step will be performed manually or semi-automatically.

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. 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. 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 vol.31 n.11-16 pp.1623 1640 (1999)
4. Gawrysiak P., Rybinski H., Protaziuk G. Text-Onto-Miner - a semi automated ontology building system Proceedings of the 17th International Symposium on Intelligent Systems (2008)
5. Gomez-Prez A., Corcho O. Ontology Specification Languages for the Semantic Web IEEE Intelligent Systems v.17 n.1 pp.54-60 (2002)

6. Manning C. D., Raghavan P., Schuetze H. An Introduction to Information Retrieval Cambridge University Press (2008)
7. McIlraith S. A., Son C. T., Zeng H. Semantic Web Services IEEE Intelligent Systems vol.16 n.2 pp.46-53 (2001)
8. Qin J., Zhou Y., Chau M. Building domain-specific web collections for scientific digital libraries: a meta-search enhanced focused crawling method Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries (2004)
9. Suber P. Open access overview http://www.earlham.edu/ peters/fos/overview.htm (2004)