Ontology-Based Deep Web Search For E-Science[1]

Tatyana I. Ivanova[2]

Abstract: **This paper makes a brief exploration of Deep Web search technologies and proposes a new semantic ontology-based approach for personalized searching scientific publications in digital libraries, books in web catalogs of scientific-content books, and other scientific data in web databases. Our main aim is to investigate main deep web search tools and digital libraries and in the base of them develop a conceptual model of personalized searching tool for scientists.**

Keywords: Deep Web Search, Personalized Search, Searching Web databases, Ontology-based Search, Searching digital libraries

I. INTRODUCTION

User queries on the Web can be classified into three types according to user's intention [3]: informational query (The intent is to acquire some information assumed to be present on one or more web pages), navigational query (The immediate intent is to reach a particular site) and transactional query [2] (The intent is to perform some web-mediated activity, as downloading or purchasing). General search engines usually don't recognize the user intent and disregarding the result type, return the mixed result list. Sometimes, it is difficult to make a strict classification of user queries according to his intent. For example, searching the digital library for scientific information is informational query, but downloading the chosen paper is transactional operation.

Web search engines can't index most of the possible pages, that can be returned by the dynamic web sites, or data, stored in Web databases (so called Deep Web[6]) and it is difficult to find such information if (the location of) source site is unknown. Google scholar for example is very useful for searching free scientific publications, but it has indexed only a little part of all of the scientific papers, published in the Web. It is very important for scientists to be able to find easy all the new research papers, related to his subject, purchase new issues of valuable books or download needed software. Another drawback of Web search engines is that during searching or ranking results they do not take into account personal user preferences or interests. Federated search tools help users to identify the databases that are best suited to the subjects they are searching. It allows users to search across multiple resources: subscription databases, library catalogs, and other types web databases.

[1] The research, presented in this publication is funded by the Internal Research Project 102ни013-10 Research and Development sector at Technical University of Sofia for 2010

[2]Tatyana I. is from the Technical University of Sofia, Bulgaria , E-mail: tiv72@abv.bg

In this paper a new semantic ontology-based approach for personalized searching scientific publications in digital libraries and web catalogs of scientific-content books is proposed. Our aim is to develop a conceptual model and as a future research, a tool for personalized searching of scientific publications in digital libraries and scientific books in web catalogs (for purchasing). As such resources are stored in full text databases or web catalogs, and are intended for users with specific research interests, we have to made research about and develop a specialized personalized Deep Web search tool. It will be used as part of the virtual scientific laboratory to facilitate the search for scientific publications, books, or specific information in Internet databases relating to scientific research.

The paper is organized as follows: Section 2 discusses earlier research in Deep Web searching; Section 3 proposes a new semantic ontology-based approach for searching scientific objects; Section 4 discuss the expected problems, strengths and drawbacks of proposed approach and it's further realization; Section 5 concludes the article.

II. DEEP WEB SEARCH STATE OF THE ART

Most of Search engines rely on programs known as crawlers (or spiders) that gather information by following the trails of hyperlinks that tie the Web together. Traditional search engines [5] such as Google, or Yahoo can be searched, retrieved and accessed only sources that have been indexed by the search engine's crawler technology. That approach works well for the pages that make up the surface Web, but for online databases that are set up to respond to grand amount of typed queries it is practically impossible to index all possible responses. The large volumes of documents that compose the Deep Web are not open to traditional Internet search engines because of limitations in crawler technology.

There are two main approaches for Deep Web search: searching previously harvested metadata (in search engine indexes, as in surface web), and federated search. Deep Web indexing methods are very different from those in surface web, as all indexing process is based on automatically querying and retrieving data behind web database search interfaces. Search engines, indexing deep web content (as Google, or Yahoo) use specific deep web crawlers. They detect the index features by issuing probe queries against the search and build a sample of the queried database by issuing a set of queries. Next, they select the most frequent words in the documents in samples to crawl the database, assuming they also have a high frequency in the actual database/index. Interface.

Federated search makes deep web documents in databases searchable by sending queries directly to native search interfaces of these databases. Additionally, federated search provides a singular search interface to numerous underlying

deep web data sources. Federated search is the technology of simultaneously searching multiple content sources from one search form and aggregating the results into a single results page. This reduces the burden on the search patron by not requiring knowledge of each individual search interface or even knowledge of the existence of the individual data sources being searched. Federated search process consists of four phases: 1. transforming a query and broadcasting it to a group of disparate databases or other web resources, with the appropriate syntax; 2. merging the results collected from the databases; 3. presenting them in a unified format with minimal duplication; 4. providing a means, performed either automatically or by the portal user, to sort or cluster the merged result set. Federated search is a type metasearch. We can build our own metasearch engines for federated search, using database or other federated search engine interfaces.

For our purposes, it is important to build our metasearch personalized tool for searching scientific digital libraries, e-commerce book catalogs and some type specific scientific databases.

A digital library is a library in which collections are stored in digital formats and accessible by computers. Many academic and government organizations provide libraries, some of which are actively involved in building institutional repositories of the institution's books, papers, theses, and other works which can be digitized. Many of these repositories are made available to the general public with few restrictions, in accordance with the goals of open access, in contrast to the publication of research in commercial journals, where the publishers often limit access rights.

Digital libraries frequently use the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) to expose their metadata to other digital libraries, and search engines like Google Scholar, Yahoo! and Scirus. The OAI-PMH compliant digital repositories in the world may be find on [15]. Metadata according to this protocol are represented in XML format. The main drawback of this representation is the lack of explicit formal semantic. As digital libraries are Deep Web recourses (mainly textual databases), there are three general strategies for searching digital libraries: Searching, using the library search interface, Federated search and Searching previously harvested metadata.

Almost every one digital library proposes internal searching tools. It provides full access to all stored in the library resources and make full use of specific library metadata. The main disadvantage of this approach is the need for the user to know in which library to search, as there are many different metadata standards [1]. There are a lot of digital libraries and choosing the best one for concrete search is a problem.

For building effective federated search engines, the knowledge of internal architecture and metadata standard of used libraries is needed. For example, DSpace architecture has three layers and two APIs : Storage layer to store digital objects and their metadata in databases and file systems; Business logic layer for key operations such as searching and browsing services; and Application layer for users to access DL system through its user interface; Networked Digital Library of Theses and Dissertations (NDLTD) is based on Federated Architecture, with MARIAN as a mediation middleware; CiteSeer uses a Service-Oriented Architecture (SOA), Open Digital Libraries uses Component-based DL architecture. Various architectures and metadata standards are the main sourse of problems in federated search engine building.

Federated search engines searching in digital libraries perform typical vertical search, as most of libraries contain resourses, related to one or few domains. For example, CiteSeer.IST[9] search engine (and digital library) search information within scientific literature, Scopus finds academic information among science, technology, medicine, and social science categories, GoPubMed searches for biomedical articles in PubMed, PubFocus searches Medline and PubMed for information on articles, authors, and publishing trends, Scitation search over one million scientific papers from journals, conferences, magazines, and other sources, Scirus moves beyond journal articles and also includes searches among such resources as scientists' webpages, courseware, patents, and more, Sesat is an open sourced Search Middleware with federation capabilities and a built-in search portal framework, CompletePlanet uses a query based engine to index 70,000+ deep Web databases and surface Web sites, WorldWideScience is composed of more than 40 information sources, several of which are federated search portals themselves. One such portal is Science.gov which itself federates more than 30 information sources representing most of the Federal government articles. This approach of cascaded federated search enables large number of information sources to be searched via a single query. For effective searching user have to have some knowledge about digital library search engines, mainly which libraries they search, papers, related to which domains store corresponding libraries and what metadata is important in searching. List of important academic databases and search engines can be found in [4].

Big search engines as Google, Yahoo, or Bing index a nearly every web site (web developers take care of this by complying with search engine optimization rules), and one may rely on them for finding emerging digital libraries before choosing the best tool for search them. Strength of general purpose search engines, having deep web searching capabilities is that they (for example Google Scholar) can offer many of freely available in the internet scientific publications in nearly every domain.

There are three main ways to search web catalogs: direct usage of building search engines, using general purpose or deep web (e-commerce) meta search engines, or making own federated search engine to search in many (directly chosen from the user) catalogs simultaneously.

For efficient direct usage of building search engines user can be previously informed about type and coverage of catalog content and corresponding search engine capabilities (accuracy, relevancy of returned results; misspelling correction capabilities; ability of searching and sorting according to different criteria, as price, brand, availability; ability in finding related words and common synonyms for terms; helping in query formulation).

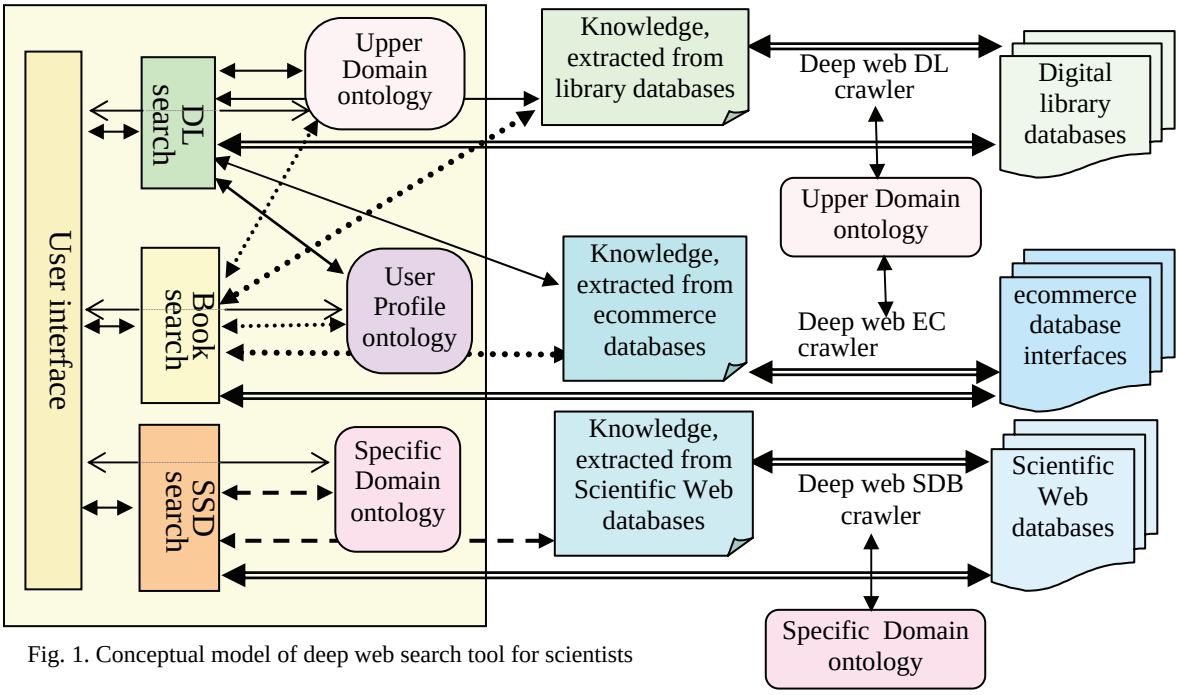

Fig. 1. Conceptual model of deep web search tool for scientists

Ecommerce Meta Search Engines (EMSEs) [8] provide a unified access to multiple ESEs allowing users to search and compare products from multiple sites with ease. A number of EMSEs have been created for some application areas on the Web. For example, addall.com searches books from multiple book sites.

Search engine technology had to grow up dramatically during the last few years [7], in order to keep up with the rapidly growing amount of information available on the web. Despite of all the continuous improvement, usually search engines return thousands results, but it is difficult to find those that are needed for the user, or they are not among the returned results. We believe that for easy and successful finding of needed information on the web, the search has to be personalized focused and semantic-based. User has to be advised or supported in full text unambiguous query formulation and choosing the best for his concrete purpose search engines. In the second chapter we propose a conceptual model of semantic ontology-based tool for scientists. It is intended for personalized searching scientific publications in digital libraries, books in web catalogs of scientific-content books, and other scientific data in web databases.

III. Semantic Ontology-Based Approach For Searching Scientific Objects

Recent research in both Data Integration, Semantic Web, or EScience, witness increasing needs for semantically driven data access, and in particular for the so called Ontology Based Data Access (OBDA). The aim of OBDA is to use ontology, i.e. a formal conceptualization of the application domain, to mediate access to data. Ontologies provide a semantic access to domain specific application data and the expression of constraints allow overcoming incompleteness that may be present in the actual data. Our idea is to model the user profile of the scientist and his scientific domain semantically using ontologies to achieve more flexibility in choosing the right search engine, decrease query ambiguity and in such a way increase the precision and recall in searching scientific publications, books or another type of scientific data. We also may rerank the returned (from one or several similar search engines) results according to particular user profile. Conceptual schema of the proposed deep web search tool for scientists is shown on figure 1.

The main differences between our tool and other Deep Web tools are:

– It is specialized for scientific papers, books and scientific data;

– It is intended to search three main Deep Web recourse types and recourse type is explicitly specified in the sending query.

The tool will offer three type services: Searching specific digital libraries for reading or purchasing scientific papers, searching book catalogs for purchasing scientific books, and searching scientific databases for integrated circuits scientific data.

We propose rich user interface, allowing selection of query intention (informational, for reading papers, or transactional, for book purchasing), selection of preferred libraries or repositories, as well as searching, dependent from chosen data or metadata. As the tool use semantic metadata (domain and profile ontologies), we will experiment some query expansion techniques [12], as well as automatic library selection or returned results reranking according to particular user profile.

Searching specific digital libraries everything uses various metadata, related to keywords, bibliographic and other specific to concrete library metadata. We will manually explore widely used technical and scientific digital libraries (for example [16], [17], [18]) and supply user with a tool for

automatic library selection according to the query and his profile information. User also may manually choose a library before sending a query. As a whole, stored in digital libraries paper metadata are not sufficient for efficient searching. We propose on the fly annotation of selected papers before it ranking and recommending to the users. For such annotation, domain ontologies, scientific experiment ontology (EXPO [14]) and paper structure ontology may be used. Annotation will be discussed in another paper. We also may use a Deep Web crawler for extracting metainformation from digital libraries or finding automatically ones, which we don't explore manually.

For searching book catalogs for purchase scientific books we will experiment several Ecommerce Meta Search Engines. Their drawback is that they query many databases (not only these that are preferable for scientists) and syntactic search approach may cause appearing the best for concrete user results very backward in the result list, or even disappearing. That is why we plan using specialized Deep web Electronic catalog semantic crawler to extract specific metadata from electronic catalogs or find emerging catalogs (figure 1).

For searching Web databases, containing scientific data, related to our electronic circuit testing domain, we firstly will exporting database schema as ontology, representing semantic of our data and then will develop and test semantic scientific data search tool, which realize Deep Web crawling for similar databases, metadata extraction and searching. Query-based sampling [10] can be used to discover the content available at a remote database server. Database translation as ontology will discuss in another paper. We expect that access rights to analogous Web databases nay be serious problem for deep web crawling and data extraction.

IV. CONCLUSION

More than a half of Web data are hidden from the surface-web search engines in databases of financial information, shopping catalogs, medical and other research in digital libraries. It is of great importance for scientific research to have easy and continuous access to the latest developments in the scientific area (presented in publications, books, and other scientific resources, usually stored in web databases).

As a result of rapidly growing number of scientific publications and books in electronic catalogs, the search precision and accuracy are becoming more and more important. One of the main trends for improving search quality is increasing the recourse metadata quality by using collaborative or semantic web technologies for metadata extraction, representation, and usage. Another important trend is digital library and web catalog standardization, exporting recourse metadata in mashine-processable format, development of the more and more effective deep web search engines.

In this paper after an analytical survey of deep web tools and approaches, we propose a conceptual model of specialized personalized Deep Web search tool for scientific information, stored as publications in digital libraries or specific databases. It uses ontology-based semantic search approach to improve search quality. It relies on rich collection of metadata, extracted from repositories, or by using methods of direct automatic otology-based annotation of textual resources to propose a flexible user friendly search interface and user query disambiguation capabilities. After analyzing the query and taking into account user profile, domain ontology and explicitly selected from the user options, the tool may reformulate the query and take a decision to which search engine (s) forward it. The tool will be implemented and tested as part of our research project. We plan to experiment dynamic selection of search strategy among several variants: direct forwarding the query to one or more scientific database or e-commerce search engines, manage user feedback and store processed information in user profile ontology for future usage in strategy-selection process.

REFERENCES

[1] Links to metadata standards http://archive.ifla.org/II/metadata.htm

[2] Y. Liz, R Krishnamurthy, S. Vaithyanathan, "Getting Work Done on the Web: Supporting Transactional Queries", http://www.almaden.ibm.com/cs/projects/avatar/sigir06.pdf, 2006

[3] A. Broder, " A taxonomy of web search" , IBM Research http://www.sigir.org/forum/F2002/broder.pdf, 2002

[4] Wikipedia's List of academic databases and search engines, http://en.wikipedia.org/wiki/List_of_academic_databases_and_search_engines

[5] Wikipedia's List of search engines, http://en.wikipedia.org/wiki/List_of_search_engines

[6] Deep Web Research Resources and Sites http://deepwebresearch.blogspot.com/

[7] Search Tools News, http://www.searchtools.com/info/database-search.html, 2010

[8] Q. Peng, W. Meng, and Hai He, "WISE-Cluster: Clustering E-Commerce Search Engines Automatically", *WIDM'04*, November 12-13, 2004, Washington, DC, USA, 2004

[9] http://citeseer.ist.psu.edu/

[10] A. S. Tigelaar,, D. Hiemstra, "QueryBased Sampling: Can we do Better than Random?", CTIT Technical Report , 2009

[11] K. C. Chang, B. He, C. Li, M. Patel, and Z. Zhang." Structured databases on the web: Observations and implications", *SIGMOD Record*, 33(3):61–70, Sept. 2004.

[12] M. Shokouhi, L. Azzopardi, P. Thomas, "Effective Query Expansion for Federated Search", SIGIR'09, July 19–23, , Boston, Massachusetts, USA, 2009

[13] A. S. Tigelaar , D. Hiemstra, "Query-Based Sampling: Can we do Better than Random?", CTIT Technical Report, http://wwwhome.cs.utwente.nl/~hiemstra/papers/tr-ctit-10-04.pdf, 2004

[14] L. Soldatova1, R. D. King, " An Ontology of Scientific Experiments", Journal of the Royal Society Interface, December 22; 311): 795–803. , 2006

[15] Web site to OAI-PMH compliant digital repositories in the world, http://www.openarchives.eu/home/home.aspx

[16] WorldWideScience.org

[17] Science.gov

[18] Scitopia.org