IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 12, December 2015

Web Focused Crawling based on Ontology

Asst. Prof. Snehal Mane[1], Asst. Prof. Poonam Gholap[2], Asst. Prof. Rakhee Kundu[3]

Computer Engineering, VESIT, Affiliated to Mumbai University, India[1, 2, 3]

Abstract: To retrieve information from web we use Google, Yahoo and MSN which are more famous search engines. Search engine is one tool to locate into on the www. It search for and identifies items in database with reference to keywords entered by user (where we get relevant data also which is not exact what we which is time consuming).For web crawling we use focused crawler which based on ontology architecture. Focused crawler search for web pages having more page rank for user requirement. Where ontology is a specification about domain. It is a part of artificial intelligence. Also web pages on site are liked with ontology structure. In pepper we use Focused Crawler and Ontology for exact service related site.

Key words: Focused Crawling, Ontology, Web services, Metadata, Digital Ecosystem.

I. INTRODUCTION

Information is in your hand, is today's need. We use web to search information frequently. Combination of focused crawling with ontology architecture is give me more accurate result which is time saving. Focused crawler which gives you exact match and ontology architected is use d for background designs.

For this system we must have detail domain knowledge of web services to create ontology structure. Where web pages are match with ontology structure to give you exact required output. This system support pdf, img, text data which is available in web page are converting in one unique data type to prepare report or documentation.
The Focused Crawler service type accepts as input and provides metadata for that service as the output. This process is divided into sub-processes like webpage fetcher, policy center, webpage pool, service metadata generator, service metadata, service metadata classifier, report generation etc.

The web page fetcher accepts service type as an input and fetches the webpage for the same. This web page is then sent to two sub-processes - policy center and webpage pool. The policy center extracts the URL from the web page to find accurate matches for its key words. The URLs of all such matches are sent as inputs for further processing. Next process is the webpage pool. This converts the web page in to plain text and stores it in the form of text file.

Before that, the webpage information is extracted using ontology markup languages (OML) to create services metadata [4]. This service Metadata is then stored in the Service Metadata database. The classification is done one base of the Ontology concept. It will get the Metadata from the Services Metadata database and classify the results of web pages. This means the predefine database which is build on Ontology structure, input resources will match the fetched pages and each page data will annotate on the user request and give a classification information for the pages.

This system does text mining on normal text, pdf, image as well as scanned text documents that include handwritten as well as printed scanned document for report generation.

II. EXISTING SYSTEM

Existing Business Ecosystems is based on digital ecosystem. Digital ecosystem growth is along with the evolution of business network and information technology. Digital Ecosystem can play client and server roles. Where we found some of the problems faced by the service provider are listed below:

1. The service provider will use some Semantic web mark up languages to describe the resources as uniform resources identifiers (URI). This will get some unnecessary data and duplicate data will get the resources.
2. Whenever we elicit information from online web pages for instant services, it is observed that such retrieval is very difficult from the existing pages and that such information is not annotated as the semantic services.
3. Some services providers work on the semantic annotation based on the domain knowledge. This search will provide very less useful information. Presently, there are no proper methods to do this work.

III. PROBLEM STATEMENT

Retrieval of information in a digital ecosystem is very complex as the data sources are distributed and hybrid in nature. Most of the existing work on digital ecosystem concentrates on single data source. Our approach concentrates on the multi data sources and multi data type. Here we use the data crawler approach to handle multiple data sources and text mining approach to handle multiple data types.

1. Objective

In order to address multiple data sources and multiple data types, in this work, we employ the semantic focused crawler. Several functions performed by the semantic focused crawler are

1. Get the data source list related to service.
2. Retrieve information regarding service entities from the web corresponding to the exact requirement of user.
3. Crawler should be able to filter and classify the service information by means of specific service domain knowledge, which corresponds to the functionality of service classification in Digital Ecosystems.

4. Applying the crawler on each and every data source.
5. Report generation from data source which contains information in any format like pdf, image, text, excel sheet.

2. Methodology

The semantic focused crawler implementation follows the following simple steps.

1. Before the crawler starts to work, users need to configure the database on base of the ontology structure of web services. (Usually the web sites' domain names) and the depth of exploring the services in database will give more accurate. It's name as, service metadata generator which generate, service metadata. Once the configuration has been completed user can submit service type as input to Webpage Fetcher for the web page crawling.
2. Once the Webpage Fetcher has downloaded a web page, it will extract the URLs in the web page and sends them to the Policy Centre for further analysis. The web page will be sent to the Webpage Pool for storage purposes.
3. When the Policy Centre receives the URLs from the Webpage Fetcher, it will determine whether they are within the crawling boundary, by analyzing their domain names. After that, the Policy Centre will discard pages which are repeated. Steps 2 and 3 are a recursive process until the user-defined web site exploration depth has been reached.
4. Once a web page has been passed to the Webpage Pool, all its embedded tags will be removed, and the web page will be stored in the form of plain texts.
5. The Service Metadata Classifier will compute the similarities between the plain text data and service metadata for each bottom-level ontology concept of a compatible ontology. If a similarity is above a threshold value, then annotation is take place to display report of matching requirement.
6. The semantic focused crawler will apply the multi data source on each and every individual data sources.
7. We can select any number of data source for report generation for services. In this, we apply text mining on data source which was generated by Webpage Pool.

IV. LITERATURE SURVEY

In this literature survey, the literature pertaining to this thesis is presented in different categories. Some of the categories are as follows: domain services knowledge providers, some web markup languages, existing documents process on the user query based, ontology based crawlers, and meta abstraction crawlers.

1. Web Markup Languages

A. OWL Web Ontology Language Overview

When a service provider publishes a service entity by means of Service Factories, the service entity can be annotated by alternative Semantic Web markup languages such as Resource Description Framework [or Web Ontology Language (OWL) [7], etc.

Resource Description Framework (RDF) is a framework for representing information about resources in a graph form. Since it was primarily intended for representing metadata about WWW resources, it is built around resources with Uniform Resource Interface (URI). To process the content of information instead of just presenting information of web sites OWL is used. OWL facilitates greater machine interpretability of Web content than that supported by XML, RDF, and RDF schema. It provides an introduction and capabilities of OWL by informally describing the features of each of the sublanguages of OWL. Some knowledge of RDF Schema is useful for understanding this process.

B. A Service Search Engine For The Industrial Digital Ecosystems

Here, these service entities are stored in the form of service metadata. Service entities are categorized by domain-specific ontology provided within Digital Ecosystems, by referencing the Uniform Resource Identifier (URI) of the service metadata to the ontological concepts[1],[3].

DE is combination of service provider and service requester, which is heterogeneous and distributed. There is some problem regarding reliable and trustworthy link between service providers and service requesters of DE. Proposed design in this research paper is a conceptual framework of a service-ontology-based semantic service search engine. Apart from the function of service search with a novel search model, this framework also provides a quality-of-services-based service evaluation and ranking methodology. To evaluate the feasibility of framework, they implement a prototype in the transport service domain, and compare the performance of the search model with three traditional information retrieval models.

2. Document Based User Query Evaluation

A. Mining Association Rules for Adaptive Search Engine Based On RDF Technology [4]

A method for mining association rules that reflect the behaviors of past users is proposed for an adaptive search engine. The logs of the users' retrieving behaviors are described with the resource description framework model, from which association rules that reflect successful retrieving behaviors are extracted. The extracted rules are used to improve the performance of a metadata-based search engine. The document repository with adaptive hybrid search engine is also developed based on the proposed method. The repository consists of a document registration module, hybrid search engine, and reasoning base. The document registration module is designed to reduce the cost of adding metadata to documents, and the hybrid search engine combines full-text search with metadata-based search engine to improve the recall of retrieval result. The reasoning base is implemented based on the association rule mining method, which contributes to improve both precision and recall of the hybrid search engine. Experiments are performed with a virtual user model, of which results show that appropriate rules can be extracted with the proposed method. The proposed technologies will contribute to realize the concept of humatronics in terms of establishing symmetric relation between humans and systems, as well as sharing

information, knowledge, and experiences via computer networks Software-reconfigurable e-learning platform for power electronics courses.

According to the existing literature, the emerging semantic focused crawlers can be categorized into two primary categories as follows: ontology-based focused crawlers and metadata abstraction focused crawlers.

3. Ontology Based Crawlers

A. State Of The Art In Semantic Focused Crawlers

A Web crawler is a software agent that can automatically browse and download Web pages from the Web. Web crawlers are usually deployed for retrieving and indexing Web documents for search engines, which enables search engines to rank the visiting priority of Web documents in terms of topics or user queries [5].

The research of focused crawler approaches the field of semantic web, along with the appearance of increasing semantic web documents and the rapid development of ontology mark-up languages. Semantic focused crawlers are a series of focused crawlers enhanced by various semantic web technologies. In this approach they discover eleven semantic focused crawlers from the existing literature, and classify them into three categories – ontology-based focused crawlers, metadata abstraction focused crawlers and other semantic focused crawlers. By means of a multi-dimensional comparison, they conclude the features of these crawlers and draw the overall state of the art of this field.

B. A Survey in Semantic Web Technologies-Inspired Focused Crawlers

Ontology-based focused crawlers refer to a group of focused crawlers that link Web documents with related ontology concepts, with the purpose of filtering and categorizing Web documents[6].

Crawlers are software which can traverse the Internet and retrieve WebPages by hyperlinks. In the face of the abundant spam Websites, traditional Web crawlers cannot function well to solve this problem. Semantic focused crawlers utilize semantic web technologies to analyze the semantics of hyperlinks and Web documents. This approach briefly reviews the recent studies on one category of semantic focused crawlers-ontology-based focused crawlers, which are a series of crawlers that utilize ontologies to link the fetched Web documents with the ontological concepts (topics). The purpose of this is to organize and categorize Web documents, or filtering irrelevant WebPages with regards to the topics. A brief comparison is made among these crawlers, from six perspectives - domain, working environment, special functions, technologies utilized, evaluation metrics and evaluation results. The conclusion with respect to this comparison is made in the final section.

C. Ontology-Based Web Crawler

Ganesh et.al.,[8] proposed an association metric, with the purpose of optimizing the order of visited URLs for web crawlers.

The requirement of a Web crawler that downloads most relevant pages is still a major challenge in the field of information retrieval systems. The use of link analysis algorithms like page rank and other importance-metrics have shed a new approach in prioritizing the URL queue for downloading higher relevant pages. The combination of these metrics along with a new metric called association-metric has been proposed. The association-metric estimates the semantic content of the URL based on the domain dependent ontology, which in turn strengthens the metric that is used for prioritizing the URL queue. In addition, after downloading the page, the association metric plays important role in estimating the relevancy of the links in that page. The proposed new metric solves the major problem of finding the relevancy of the pages before the process of crawling, to an optimal level.

D. THESUS: Organizing Web Document Collections Based On Link Semantics

THESUS aims to organize online documents by linking their URLs to hierarchical ontology concepts, which are seen as thematic subsets [9].

The requirements for effective search and management of the WWW are stronger than ever. Currently web documents are classified based on their content not taking into account the fact that these documents are connected to each other by links. They claim that a page's classification is enriched by the detection of its incoming links' semantics. This would enable effective browsing and enhance the validity of search results in the WWW context. Another aspect that is under addressed and is strictly related to the tasks of browsing and searching is the similarity of documents at the semantic level. The above observations lead us to the adoption of a hierarchy of concepts (ontology) and a thesaurus to exploit links and provide a better characterization of web documents. The enhancement of the documents characterization makes operations such as clustering and labeling very interesting. To this end, the authors devised a system called THESUS. The system deals with an initial set of web documents, extracts keywords from all pages' incoming links and converts them to semantics by mapping them to a domain's ontology. Subsequently, a clustering algorithm is applied to discover groups of web documents. The effectiveness of the clustering process is based on the use of a novel similarity measure between documents characterized by sets of terms. Web documents are organized into thematic subsets based on their semantics. The subsets are then labeled, thus enabling easier management (browsing, searching, querying) of the Web.

E. Semantic Web Services in Factory Automation: Fundamental Insights And Research Roadmap[2]

One of the significant challenges for current and future manufacturing systems is that of providing rapid re-configurability in order to evolve and adapt to mass customization. This challenge is aggravated if new types of processes and components are introduced, as existing components are expected to interact with the novel entities but have no previous knowledge on how to collaborate. This statement not only applies to innovative processes and devices, but is also due to the impossibility to incorporate knowledge in a single device about all types of available system components. This approach proposes the use of Semantic Web Services in order to overcome this

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 12, December 2015

challenge. The use of ontologies and explicit semantics enable performing logical reasoning to infer sufficient knowledge on the classification of processes that machines offer, and on how to execute and compose those processes to carry out manufacturing orchestration autonomously. A series of motivating utilization scenarios are illustrated, and a research roadmap is presented.

4. Metadata abstraction

A. State Of The Art in Metadata Abstraction Crawlers

Metadata abstraction focused crawlers are the focused crawlers that can abstract meaningful information from relevant web pages and annotate the information with ontology markup languages [10].

The research of crawlers moves closer to the semantic web, along with the appearance of increasing XML/RDF/OWL files and the rapid development of ontology mark-up languages. As an emerging concept, metadata abstraction crawlers are a series of crawlers that aim to abstract metadata from normal HTML documents, based on various semantic Web technologies. In this approach, they make a general survey of the current situation of metadata abstraction crawlers. Fourteen cases in this field are chosen as typical examples, and classified in five clusters.

B. Searching and retrieving legal literature through automated semantic indexing

Francesconi and Peruginelli [11] proposed a metadata abstraction focused crawler for a vertical portal system, which is a management system for legal documents.

Access to legal information and, in particular, to legal literature is examined in conjunction with the creation of a Portal to Italian legal doctrine. The design and implementation of services such as integrated access to a wide range of resources are described, with a particular focus on the importance of exploiting metadata assigned to disparate legal material. The integration of structured repositories and Web documents is the main purpose of the Portal: it is constructed on the basis of a federation system with service provider functions, aiming at creating a centralized index of legal resources. The index is based on a uniform metadata view created for structured data by means of the OAI approach and for Web documents by a machine learning approach. Subject searching is a major requirement for legal literature users and a solution based on the exploitation of Dublin Core metadata, as well as the use of legal ontologies and related terms prepared for accessing indexed articles have been implemented.

C. eBizSearch: A niche search engine for e-business

Giles et.al.,[12] proposed a metadata abstraction focused crawler for a niche e-business information search engine.

Niche Search Engines offer an efficient alternative to traditional search engines when the results returned by general-purpose search engines do not provide a sufficient degree of relevance. By taking advantage of their domain of concentration they achieve higher relevance and offer enhanced features. They discuss a new niche search engine, eBizSearch, based on the technology of Cite Seer and dedicated to e-business and e-business documents. They present the integration of Cite Seer in the framework of eBizSearch and the process necessary to tune the whole system towards the specific area of e-business. They also discuss how using machine learning algorithms they generate metadata to make eBizSearch Open Archives compliant. eBizSearch is a publicly available service and can be reached.

5.Text Mining

Text mining is the discovery of interesting knowledge in text documents. It is a challenging issue to find accurate knowledge (or features) in text documents to help users to find what they want [15].

V. IMPLEMENTATION

1. Software Requirement:

The language chosen for this project is

Vb.NETMicrosoft Office 2007

Microsoft Object Document Imaging(MODI)

OpenXML SDK Tool

2. Data Dictionaries

1.TABLE : Service_Cat_Metadata

[cat_id] [int] NULL ,

[serv_metadata] [varchar] (100)

2.TABLE : Service_Category

[cat_id] [int] NULL ,

[cat_name] [varchar] (50)

3.TABLE : Service_Metadata

[cat_id] [int] NULL ,

[sub_cat_id] [int] NULL ,

[service_link] [varchar] (100)

[page_name] [varchar] (50)

[service_desc] [varchar] (1500)

4.TABLE : Service_Relv

[serv_rel_metadata] [varchar] (100)

5.TABLE : Service_Sub_Cat_Metadata

[cat_id] [int] NULL ,

[sub_cat_id] [int] NULL ,

[service_metadata] [varchar] (100)

6.TABLE : Service_Sub_Category

[cat_id] [int] NULL ,

[sub_cat_id] [int] NULL ,

[sub_cat_name] [varchar] (100)

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 12, December 2015

3. Result Analysis

Process Interfaces Screenshot

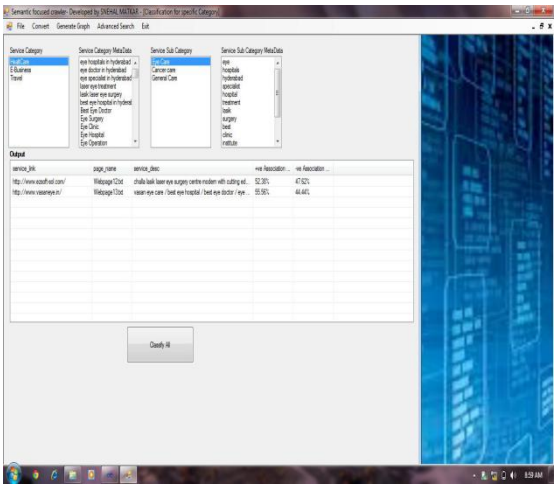

FIGURE 5.1 – Service domain with detail level selection for crawling relevant web page from web(here local server)

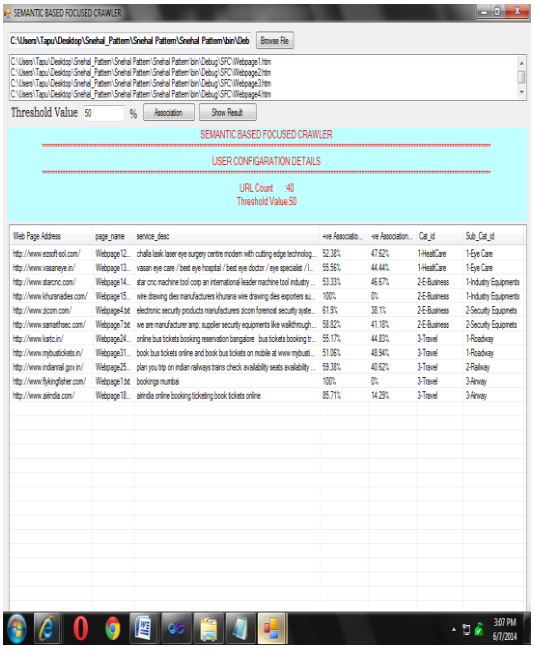

FIGURE 5.2 – Web Page Fetching Association Percentage with respect Ontology

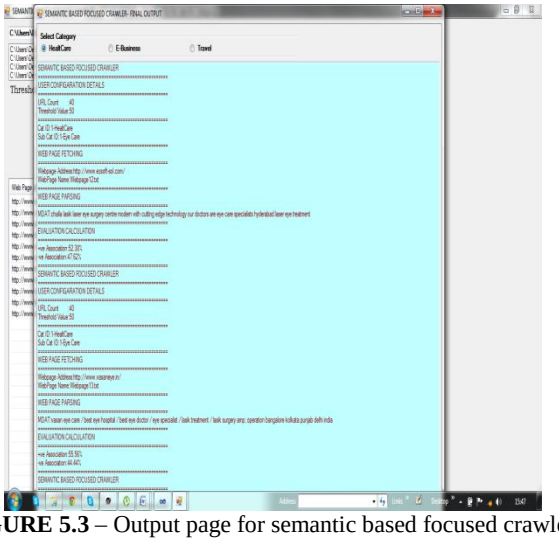

FIGURE 5.3 – Output page for semantic based focused crawler.

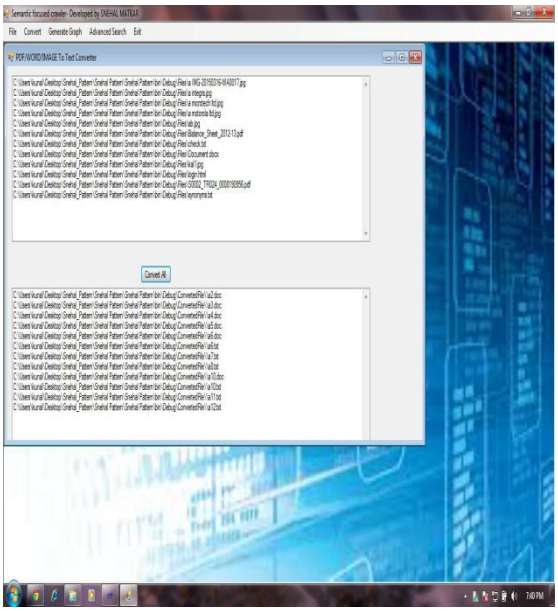

FIGURE 5.4 – Text File Generated in Webpage Pool

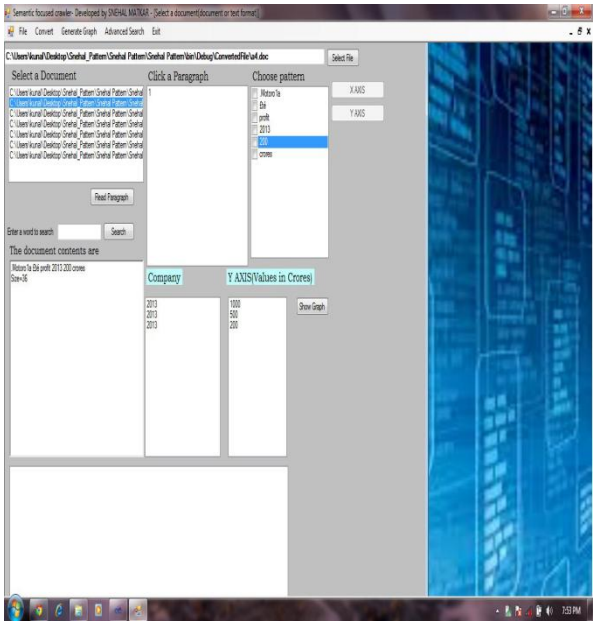

FIGURE 5.5 –Generated coordinates for output in form of graph

The above screenshot shows that the proposed mechanism for automatically categorizing and annotation for the services work very efficiently. We enhance the work with re-parsing categorized and annoted services for statistical data extraction for performance comparison as show in Figure – 5.5

VI. CONCLUSION

Web data is available in multi formats and multi data types. This causes problems in retrieving and unifying data for a unified representation. In this thesis, we presented ontology based and sentiment focused crawling which integrates web data of different formats like, PDF, Scanned Document, Images etc. Our system is implemented through the following five steps: (1) Semantic focused crawling for user query (2) Classification based on ontology for specific service (3)

Plain text conversion (4) Creation of service metadata and (5) Report generation.

This process is all about text mining on web pages in off line mode. In future I am expecting that text mining process must take place online, i.e., no need to download web page just submit URL to the process, for online application. Also we can add more text mining process for multimedia (i.e. images, audio, video, etc.) data available on web and extract information from it.
Proposed work will save time, memory and will guarantee efficient results.

REFERENCES

[1] T. Heistracher, T. Kurz, C. Masuch, P. Ferronato, M. Vidal, A. Corallo, G. Briscoe, and P. Dini, "Pervasive service architecture for a digital business ecosystem," in Proc. 18th ECOOP, Oslo, Norway, Pp. 71–80, 2004.
[2] J. L. M. Lastra and M. Delamer, "Semantic web services in factory automation: Fundamental insights and research roadmap," IEEE Trans. Ind. Informat., Vol. 2, No. 1, Pp. 1–11, Feb. 2006.
[3] P. Malone, "DE services in Ecosystem Oriented Architectures," in Digital Business Ecosystems, F. Nachira, P. Dini, A. Nicolai, M. L. Louarn, and L. R. Lèon, Eds: Eur. Commission, 2007.
[4] Y. Takama and S. Hattori, "Mining association rules for adaptive search engine based on RDF technology," IEEE Trans. Ind. Electron., Vol. 54, No. 2, Pp. 790–796, Apr. 2007. Web Data Unification Ontology based Text Mining
[5] H. Dong, F. K. Hussain, and E. Chang, "State of the art in semantic focused crawlers," in Proc. ICCSA, Yongin, Korea, 2009, Pp. 890–904.
[6] H. Dong, F. K. Hussain, and E. Chang, "A survey in semantic web technologies-inspired focused crawlers," in Proc. 3rd ICDIM, East London, U.K., 2008, Pp. 934–936.
[7] Jeff hefline,"An Introduction to the OWL web Ontology language",cse.
[8] S. Ganesh, M. Jayaraj, V. Kalyan, and G. Aghila, "Ontology-based web crawler," in Proc. ITCC: Coding Comput., Las Vegas, NV, 2004, Pp. 337–341.
[9] M. Halkidi, B. Nguyen, I. Varlamis, and M. Vazirgiannis, "THESUS: Organizing web document collections based on link semantics," VLDB J., Vol. 12, no. 4, Pp. 320–332, Nov. 2003.
[10] H. Dong, F. K. Hussain, and E. Chang, "State of the art in metadata abstraction crawlers," in Proc. IEEE ICIT, Chengdu, China, Pp. 1–6,2008.
[11] E. Francesconi and G. Peruginelli, "Searching and retrieving legal literature through automated semantic indexing," in Proc. ICAIL, Standford, CA, Pp. 131–138,2007.
[12] C. L. Giles, Y. Petinot, P. B. Teregowda, H. Han, S. Lawrence, A. Rangaswamy, and N. Pal, "eBizSearch: A niche search engine for e-business," in Proc. SIGIR, Toronto, ON, Canada, Pp. 413–414,2003.
[13] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. New York: Addison-Wesley, 1999.
[14] L. T. Su, "The relevance of recall and precision in user evaluation," J. Amer. Soc. Inf. Sci. Technol., Vol. 45, No. 3, Pp. 207–217, Apr. 1999.
[15] N. Cancedda, N. Cesa-Bianchi, A. Conconi, and C. Gentile, "Kernel Methods for Document Filtering," TREC, trec.nist.gov pubs/trec11/papers/kermit.ps.gz, 2002.