A Case Study on Linked Data Generation and Consumption

Jianqiang LI Yu ZHAO

NEC Laboratories China
14F, Bldg.A, Innovation Plaza, Building 1, Tsinghua Science Park
No.1 Zhongguancun East Road, Haidian District, Beijing, China 100084
+86-10-6270-5180

{lijianqiang, zhaoyu}@research.nec.com.cn

ABSTRACT

The availability of large amounts of interlinked semantic data is a fundamental prerequisite of the Semantic Web. At present, almost all the usable ontological data is built manually or by directly transforming certain (semi-)structured data sources into certain formats of semantic data. To solve the "isolated data island" problem of the Semantic Web caused by this situation, the Linking Open Data initiative was launched to unite these machine readable datasets as a Web of Data. In this paper, an alternative method to make a contribution to this trend is proposed. By extracting the linked data from the existing Web of Documents in an automatic way, the inherent statements implied by the hyperlinks can be made available on the Web. This realizes the maximized reuse of the current Web (In most cases, the current Web is rendered as Web of Documents) and provides a bridge between the Web of Documents and the Web of Data. An experimental study utilizing the resultant linked data for Small Web search is introduced, in which the ontological information about the webpages serves as the metadata and this provides the potential to enhance the web search quality. The results of this study show that, compared with the existing solutions, the precision of Small Web search is significantly improved.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Design, Experimentation

Keywords

Linked Data, Semantic Web, Ontology, Metadata, Web Mining

1. INTRODUCTION

The existence of large amounts of interlinked semantic data is a prerequisite for making the Semantic Web become a reality. Although many ontology learning algorithms [1] have been studied, it is difficult to use them in a real application due to the complex format of the recommended data models for knowledge representation, which usually contain many items that are difficult to fill in, even by a human. Current usable semantic data has mainly been built by human- or template-based translation from existing (semi-)structured data sources into certain formats of semantic data. [3] The result of this fact is that each of them might be an isolated island of data. Recently, following the Linked Data principles [3] outlined by Tim Berners-Lee, the W3C community project called Linking Open Data [3] is pursuing the rectification of this situation, and has released several large, interlinked RDF datasets (e.g., DBpedia, Geonames, WordNet, and the DBLP bibliography). Obviously, their adopted approach for linked data construction relies heavily on the already existing (structured) data sources and the efforts made by the data publishers.

Considering the fact that the existing Document Web provides an unprecedented opportunity and fertile ground for information analysis and knowledge discovery, this paper proposes an alternative method to make a contribution to the Web of Data. The method is to extract the inherent statements implied in the hyperlinks as a form of semantic data and make the data available to be consumed by different Semantic Web applications.

The global Web is constituted of multiple publicly-accessible local websites or intranets, i.e., so-called Small Webs. Correspondingly, the semantic relations between the topics of the webpages can be roughly classified into two types, i.e., hierarchical and reference relationships:

1. The website builder generally adopts a hierarchical structure for web page organization [5] [6]. Hierarchical relationships between webpages are mainly hold for intra-websites hyperlinks. In a Small Web, a large amount of well-structured hyperlinks for organizing the collection of webpages are created through an administrative way. They collectively reflect a common view or understanding through a hierarchical mode. Intuitively, each website corresponds to a distinct semantic dataset with hierarchical relations.
2. The web page author usually utilizes reference hyperlinks for web page citation. They implicitly make a statement on the target web page (and then the corresponding topic in the website) and convey the recommendation or reference relations between the topics which are distributed over various websites. Reference relationships between webpages are mainly hold for inter-websites hyperlinks. It is possible to extract such reference relations to tie multiple datasets of hierarchies together as a large pool of linked data.

This paper describes our exploratory research to construct the semantic datasets reflecting the common views or understandings. Since the resulting semantic data inherit the hyperlink relations, they comply naturally with the Linked Data principles.

LDOW2008, April 22, 2008, Beijing, China.

Basically, the extracted semantic data can be shared as a new semantic data resource (like a sitemap) to be browsed by a user [7]. However, the more important role of this machine-understandable data is to be consumed by certain web applications to provide improved performance. This paper reports a case study on consuming the resultant linked data as metadata of corresponding webpages to improve Small Web search.

The rest of the paper is organized as follows. Section 2 describes a method to extract the linked data from the Document Web. Then, a case study for consuming the resultant interlinked semantic data for Small Web search is introduced in Section 3, where the experimental results are presented. Section 4 concludes the paper with a discussion.

2. LINKED DATA GENERATION FROM THE DOCUMENT WEB

Web documents discuss various topics. Each topic could be represented by its portal webpage. For example, the topic "Department of Computer Science, Stanford University" could be identified by the homepage of the department, http://cs.stanford.edu. As mentioned above, the relationships between topics can be divided into two types: hierarchical relations and reference relations. Hierarchical relations link the main topics and subtopics together, while reference relations reflect the flat association between topics. Within the Document Web, the topic relations correspond to the web structure.

The hierarchical relations are mainly implied in the Small Web, i.e., the website's internal hyperlink structures, while the reference relations are mostly embodied in the global Web, i.e., the inter-site hyperlinks. Therefore, we can extract a topic hierarchy from each website and form a semantic dataset. The linkages of these distinct datasets, namely reference relations, can be derived directly from the inter-site hyperlinks.

2.1 Hierarchical relationship extraction

We argue that the hierarchical relationships could be exploited from a certain type of hyperlinks, called *hierarchical hyperlinks* (HLs). For example, the webpage http://cs.stanford.edu, identifying the topic "Department of Computer Science", includes the hyperlink anchor "*Faculty*" which directs to the page http://cs.stanford.edu/People/faculty, identifying the topic "CS's faculties". Obviously, this hyperlink implies the hierarchical relationship between these two topics, which is that the latter topic is a sub-topic of the former.

Therefore, our goal is to identify the HLs, or to remove the non-HLs from all the hyperlinks. Since most HLs are intra-site, i.e., the source and target webpages of the hyperlink belong to the same website, we only investigate intra-hyperlinks here.

We divide non-HLs into two sub-types, syntactic non-HLs and semantic non-HLs. The algorithm to remove the non-HLs and obtain the HLs is described in the following.

2.1.1 Syntactic non-HL removal

A syntactic non-HL is defined as a hyperlink whose target page can be judged as the superior source page, by comparing their URLs syntactically. For example, if a hyperlink has the source page http://www.abc.com/d/file.html, and the target page is http://www.abc.com, it can be judged directly that the target is the parent page of the source within the website hierarchy, by analyzing the two URLs syntactically. It is certainly the same for the corresponding topic relationship. In such a case, the hyperlink is a not an HL but a pure navigational hyperlink intended only for the user's convenience when site browsing. Namely, the hyperlink from http://www.abc.com to http://www.abc.com/d/file.html is a HL, and the inverse is not a HL.

Generally, we can build several rules to identify syntactic non-HLs. A standard HTTP-based URL string could be formatted as: *http://[host]/[path]/[file]#[fragment]*; the source page, target page, source page URL, and target page URL of a hyperlink l are denoted by $s(l)$, $t(l)$, u_s, and u_t, respectively; then, if l meets any one of the following conditions, l is judged as a syntactic non-HL: 1) t is the homepage of the corresponding website; 2) $s(l)=t(l)$; 3) t is the homepage of $host(u_t)$, and $host(u_s)$ is a sub-host of $host(u_t)$; 4) t is the homepage of $path(u_t)$, $host(u_t)=host(u_s)$, and $path(u_s)$ is a sub-path of $path(u_t)$; 5) $host(u_t)=host(u_s)$, $path(u_t)=path(u_s)$, $file(u_t)=file(u_s)$, $fragment(u_s)\neq$**null**, and $fragment(u_t)=$**null**.

2.1.2 Semantic non-HLs removal

Distinct from the syntactic non-HLs, those non-HLs that cannot be identified explicitly are called semantic non-HLs. Our observation is that most implicit semantic non-HLs, including the hyperlinks between sibling pages and the upward hyperlinks from the low level pages to the high level pages, can be extracted in an analysis of the relationships among the hyperlinks. Especially if the target page set P_1 comes from the same Link Collection (a Link Collection represents a group of hyperlinks within the same semantic block of a webpage), and the pages have a common outbound page set P_2, then there is a high possibility that P_1 are the sibling pages at the same hierarchical level, and the hyperlinks from P_1 to P_2 are non-HLs.

Based on above assumption, an algorithm for identifying semantic non-HLs is designed. We denote the whole intra-site hyperlink set as L and currently identified non-HL set as L'. Within the scope of L-L', the outbound hyperlinks and pages and the inbound hyperlinks and pages for each page P are denoted as OL_p, OP_p, IL_p and IP_p, respectively. LC_p is a partition of OL_p by Link Collections, and each $lc \in LC_p$ is a link collection. A Link Collection could be generated by various means. The method we adopt is to group the hyperlinks which are siblings and of the same DOM path (HTML tag sequence starting from the root element, i.e., <*html*> tag) on the page as a link collection. The outbound page set of link collection lc is defined as $OP_{lc} = \{t(l) \mid l \in lc\}$. We can thus get the common outbound pages of OP_p and OP_{lc}, denoted by C_p and C_k, respectively. Then we can define the principle for judging: for $\forall l \in lc$, if $t(l) \in (C_p \bigcup C_{lc}) \bigcap (OP_p \bigcup \{p\})$, l is judged as a semantic non-HL. The semantic non-HL set is denoted by L''.

Thus, for a website, we can identify the HL set as $HL=L-L'-L''$. Each HL l represents a hierarchical relationship instance between the topic of $s(l)$ and the topic of $t(l)$, which is that the latter is a sub-topic of the former.

2.2 Publishing as linked data

With the above algorithm, the hierarchical relations between topics are extracted. Also, the reference relations can be directly identified from all the inter-site hyperlinks. The next step is to make these semantic relations available to be shared on the Web.

Firstly, we have defined a set of vocabularies, Web Document as a Topic (WDT), with RDFS. WDT includes two core concepts, one

is the topics which the documents describe and another is the relations between documents' topics. WDT defines two new class terms for these concepts, "wdt:Topic" and "wdt:TopicRelation". The topic relation class has a sub-class for hierarchical relations (wdt:HierarchicalRelation) and a sub-class for reference relations (wdt:RefererenceRelation). A topic instance is generated automatically from the corresponding portal web document, whose title becomes the topic's label. We reuse a FOAF property, "foaf:isPrimaryTopicOf", to link the topic and the document (a web document is a "foaf:Document"-type resource). For a topic relation instance, its label property value can also be generated automatically from the anchor texts of the corresponding hyperlinks. Each topic relation instance is a bilateral relationship between two topics. For a hierarchical relation, the two correlative topics are called the main topic (wdt:mainTopic) and the subtopic (wdt:subTopic). The main topic is considered the "parent" topic in the hierarchical relation, while the subtopic is the "child". Comparatively, for a reference relation, we call them the referee topic (wdt:refereeTopic) and the referred topic (wdt:referredTopic). Figure 1 shows the WDT framework.

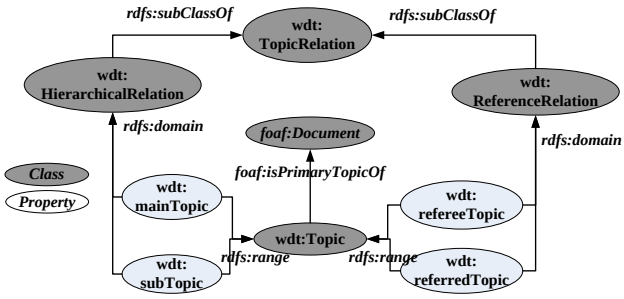

Figure 1. The WDT Vocabularies Framework

We generate a separate WDT/RDF file as a data source for each topic hierarchy which comes from a separate website, including the topic instances and the hierarchical relation instances. These separate data sources are interlinked with the reference relations. A reference relation instance is defined within the data source where the definition of the referee topic is located, and then the reference topic is hyperlinked by a URI.

```
# Topic "Protégé"
<http://www.nec.com.cn/lab/WDT/data/stanford.edu#34211>
    rdf:label "The Protégé Ontology Editor and Knowledge
Acquisition System" ;
    rdf:type wdt:Topic ;
    foaf:isPrimaryTopicOf <http://protege.stanford.edu> .
# Topic "Overview of Protégé"
<http://www.nec.com.cn/lab/WDT/data/stanford.edu#34212>
    rdf:label "What is Protégé?" ;
    rdf:type wdt:Topic ;
    foaf:isPrimaryTopicOf
<http://protege.stanford.edu/overview/> .
# Hierarchical relation between above two topics
<http://www.nec.com.cn/lab/WDT/data/stanford.edu#34302>
    rdf:label "OVERVIEW" ;
    rdf:type wdt:HierarchicalRelation ;
    wdt:mainTopic <
http://www.nec.com.cn/lab/WDT/data/stanford.edu#34211>  ;
    wdt:subTopic <
http://www.nec.com.cn/lab/WDT/data/stanford.edu#34212>  .
```

Figure 2. A segment of the topic hierarchy of *stanford.edu*

Here we give a simple example to illustrate the generated linked data. We extract topics and topic relations from two websites, *stanford.edu* and *w3.org*, respectively. For each website, a topic hierarchy is generated based on the intra-site hyperlink structures. Figure 2 shows a segment of the Turtle representation of the topic hierarchy of *stanford.edu*. Similarly, a segment of the topic hierarchy of *w3.org* is shown in Figure 3.

```
# Topic "OWL"
<http://www.nec.com.cn/lab/WDT/data/w3.org#1419>
    rdf:label "Web Ontology Language OWL / W3C Semantic Web
Activity" ;
    rdf:type wdt:Topic ;
    foaf:isPrimaryTopicOf <http://www.w3.org/2004/OWL/> .
# Topic "OWL Guide"
<http://www.nec.com.cn/lab/WDT/data/w3.org#1421>
    rdf:label "OWL Ontology Web Language Guide" ;
    rdf:type wdt:Topic ;
    foaf:isPrimaryTopicOf <http://www.w3.org/TR/owl-guide/> .
# Hierarchical relation between above two topics
<http://www.nec.com.cn/lab/WDT/data/w3.org#1507>
    rdf:label "OWL Ontology Web Language Guide" ;
    rdf:type wdt:HierarchicalRelation ;
    wdt:mainTopic <
http://www.nec.com.cn/lab/WDT/data/w3.org#1419>  ;
    wdt:subTopic <
http://www.nec.com.cn/lab/WDT/data/w3.org#1421>  .
```

Figure 3. A segment of the topic hierarchy of *w3.org*

The inter-site hyperlinks of these two separate websites are exploited to generate the reference relation instances to connect the above two data sources. Figure 4 is an example of the reference relation.

```
# Reference relation between protégé and OWL
<http://www.nec.com.cn/lab/WDT/data/stanford.edu#34311>
    rdf:label "OWL Ontology Web Language Guide" ;
    rdf:type wdt:ReferenceRelation ;
    wdt:refereeTopic <
http://www.nec.com.cn/lab/WDT/data/stanford.edu#34212>  ;
    wdt:referredTopic <
http://www.nec.com.cn/lab/WDT/data/w3.org#1421>  .
```

Figure 4. An example of a reference relation

If we surf the whole Document Web to generate WDT data, it is apparent that each web document would generate a topic instance within WDT data sources. Thus, the web document resources would be rendered as the bridge between WDT data sources and other data sources defined by other communities. For example, within the OWL schema definition data (http://www.w3.org/2002/07/owl), the links to the web document resources can be also found, as shown in Figure 5.

```
<rdfs:isDefinedBy rdf:resource="http://www.w3.org/TR/2004/REC-
owl-semantics-20040210/" />
```

Figure 5. Link from data to document

Then this data source is connected to the WDT data source of *w3.org* with these web documents. Such interlinks between different data sources accord with the linked data principles.

3. CONSUMING THE LINKED DATA FOR SMALL WEB SEARCH

The main goal of the semantic data generation is to make them understandable and consumable by a real web application. This section proposes a novel webpage retrieval method, called PathRank, for consuming the WDT data for Small Web search, where the semantic statements, i.e., topic relations, within WDT

data serve as the metadata of the corresponding webpage to improve the ranking results.

To characterize our novel web search approach, we describe a usage scenario to show its importance and usefulness in Small Web search. Statistical surveys indicate that more than 24 percent of web search queries are navigational queries, i.e., search for several specific sites or pages that the users assume exist. The most representative examples of navigational queries are to search for the homepages of such subjects as a specific person, a project, or an organization. Also included are situations of retrieving the introductory pages of a subject denoted by a query, where corresponding sub-subjects should be also recalled at the same time. Although existing predominant search algorithms such as HITS or PageRank, which take the page contents and anchor texts as indexes, can retrieve the seemingly correct results for some navigational queries, in practice they assume the queries are informational searches, i.e., they retrieve and rank the webpages by explicitly including the query's keywords within the content or anchor texts. Thus, in many cases, they do not obtain satisfactory results for navigational queries, especially in the scenario that the targeted pages or even their relative anchor texts do not include the keywords syntactically. For example, if a user wants to obtain the alumni list pages of the Department of Computer Science at Stanford University, he might submit a query with the keywords "computer science alumni" to the stanford.edu website search engine. However, the top-100 search results do not contain any direct links to alumni list pages. Instead, a user can find alumni list pages for "Undergrads", "Masters" and "PhDs" by manual navigation from the homepage of the Department of Computer Science. None of these three pages contain any textual clues about "computer science"; however, such an implicit sense could be deduced semantically from the navigational context, i.e., the department's homepage links. That is to say, from the perspective of the user, because there is a navigation path from the department's homepage to each alumni page, the alumni pages are virtually tagged with the topic "homepage, computer science".

We define a semantic path as the path from the root page of a website to a destination page and its associated hyperlinks through the navigation hierarchy. Semantic paths can give important indicative or contextual information about the content of the destination webpages.

3.1 Semantic path generation

After a topic hierarchy for a website is extracted, the semantic paths for each web document within the website can then be generated from a sequential hierarchical relation instance list without a circular path.

For a web document p, its semantic paths would be: $<h,HR,p>$; and $HR=\{hr_i\}$, $i=1,2,\ldots,n$, where hr_i is a hierarchical relation instance, $isPrimaryTopicOf(mainTopic(hr_1))$ is the homepage h of the website, $mainTopic(hr_{i+1})=subTopic(hr_i)$, $isPrimaryTopicOf(subTopic(hr_n))=p$, and for any $i\neq j$, there are $hr_i\neq hr_j$, $mainTopic(hr_i)\neq mainTopic(hr_j)$ and $subTopic(hr_i)\neq subTopic(hr_j)$.

3.2 Path-based method for webpage retrieval

Each semantic path includes a list of text nodes. In our initial implementation, the corresponding URL, page title, and anchor text are extracted to constitute the text node.

We have built a two-step procedure for realizing webpage ranking in a Small Web by utilizing the semantic paths: Step 1) Computing the rank value $R_{path}(q)$ for each semantic path *path* according to its located Small Web and the query q; Step 2) The PathRank value R_{page} of a webpage *page* is determined by all its corresponding semantic paths (or together with the page's content-based score).

Intuitively, each text node is about a topic. A semantic path comprises multiple levels of statements and covers multiple topics. A query might also cover one or several topics. Step 1 realizes the similarity computing between the query and the semantic path by comparing how many equivalent topics they cover.

There might be several semantic paths in a Small Web that point to the same webpage. Step 2 merges their similarities with the query as the final score to rank the similarity between the query and the pages.

3.3 Experiment in Small Web search

Two experiments were conducted, one for informational querying and one for navigational querying, to evaluate our application – PathRank-based web search.

3.3.1 Data set and query set

We utilize the web document collection from the Stanford University website to simulate a web search, where the inter-subdomain hyperlinks are utilized to rank the subdomain site within stanford.edu. Based on the consideration that breadth-first search crawling yields high-quality pages, a corresponding strategy is configured in our crawler, whereby the maximum number of hyperlink hops is set to 15. We collected about 2 million webpages from 2,980 subdomains of stanford.edu. After pruning away webpages from the subdomains that contain less than 20 pages and duplicate webpages, about 1.4 million unique pages from 768 distinct subdomains remain.

For navigational queries, we randomly select 50 objects from the departments, research groups, events, people and specific subjects, and use each of their names as the queries for testing whether their homepages could be retrieved with top rank. We should note here that this query set includes the "Computer Science alumni" as mentioned above. For informational queries, we randomly select 50 names of research disciplines as the query keywords. As such, the search result pages are expected to be directly related to the corresponding discipline.

3.3.2 Evaluation criteria

Because topic hierarchies play a key role in the subsequent webpage retrieval, in this experiment, we evaluate not only the search results, but also the topic hierarchy extraction algorithm. For the evaluation of topic hierarchy extraction, we calculate the precision, recall and F-measure of the obtained hierarchical relation instances, where the ground truth is identified by hand for the sample set. Then, for the web search, a precision evaluation is the main goal of the experiment. The following criteria are utilized in our evaluation of the approach:

S@5 (S@50) for the navigational query: This is the proportion of queries for which one of the correct answers is ranked in the top 5 (50) in the ranked list returned for the query.

P@10 (P@20) for the informational query: This is the proportion of relevant webpages in the top 10 (20) webpages in the ranked list returned for the query.

SP: This is used to evaluate the overall quality of the approach for the website search. It is an average of the precision of the navigational and informational queries:

$$SP = \frac{\gamma(S@5) + (1-\gamma)(P@10)}{2},$$

where γ reflects the weighting of the navigational query versus the informational query. For simplicity, it is set to 0.5 here.

3.3.3 Experiment results

Firstly, we randomly select a subdomain from the whole website document set to evaluate our topic hierarchy extraction algorithm. All possible hierarchical relation instances within this subdomain are collected by hand from the hyperlink set. This results in 1,321 hierarchical relation instances for 478 documents.

The topic hierarchy extraction is conducted within the scope of the whole *stanford.edu* site. After generating the topic hierarchy for the whole site, the hierarchical relation instances within the sample subdomain are sorted out, and these amount to 1342. Then the corresponding criteria are calculated, and whether a resultant semantic path is correct is judged by a human observer. The amount of the correct relation instances is 1123, so the recall is 85%, the precision is 83.6%, and the F-measure is 84.3%. These results are satisfactory.

Then for evaluating the PathRank-based web search results, we use the local website search engine at *stanford.edu* as the baseline. Since it is Google-based search engine, it is assumed that the website search engine represents the state-of-the-art small web search technology. We conduct the topic hierarchy extraction using two methods: PathRank1 regards the whole stanford.edu as a website, then all the site rank values R_W are set as 1 when calculating the PathRanks; PathRank2 regards each subdomain of stanford.edu as a distinct website, and so the topic hierarchies are extracted within each subdomain, and the site rank values R_W are calculated from the reference relations across the subdomains. For both methods, the maximum length for semantic paths is set to 7, which is to balance the tradeoff between the result quality and the calculation cost. In the entire stanford.edu site, the number of the resulting semantic paths is 7.5 million and 6.9 million from PathRank1 and PathRank2, respectively. They have reasonable coverage, at 78% and 80% respectively for the whole 1.4-million-page data set. Additionally, we combine the traditional content-based retrieval method together with PathRank to investigate the effects of this combination. Thus we obtain a total of five sets of experiment results, shown in Table 1.

Tabel 1. The evaluation of PathRank based web search

	S@5	S@50	P@10	P@20	SP
***stanford.edu* search**	64%	74%	82%	79%	73%
PathRank1	78%	86%	75%	69%	77%
PathRank1+content	76%	90%	81%	72%	78%
PathRank2	85%	89%	88%	71%	81%
PathRank2+content	88%	92%	86%	77%	87%

These results illustrate that PathRank can achieve better search quality than the baseline, and especially, the quality of navigational queries is remarkably improved. Generally, the precision of search results from our approach can show improvements by as much as 14% compared with that of the *stanford.edu* website search engine.

4. CONCLUSIONS

Our research originated from the idea that the existing Document Web (WWW) is a mirror of the real world. As an important knowledge and communication resource, the Document Web provides descriptions and reflects linkages to real world entities. Deriving from this idea, our research was to construct linked semantic data from the current Document Web, through which to make a contribution to the Semantic Web of Data. Based on the similarity of hyperlinks and RDF links, which provide the linkages between two webpages and two pieces of data, respectively, an intuitive method to extract the statements implied in the Document Web is proposed. Also, a flexible model for describing such statements is given.

To show the benefit of the semantic data in improving the performance of the solution based totally on the Document Web, a case study exploiting the semantic dataset for Small Web search is conducted. The semantic path generated from the semantic datasets includes summary information about a set of hierarchically related webpages. Collectively, these inside texts give "virtual tags" to the destination pages. From an ontological point of view, the semantic inference functionality is implicitly incorporated into the webpage retrieval process. The relationships between multiple keywords in a query are exploited to realize word sense disambiguation and webpage filtering in the web search process. The technique proposed in this paper is especially suitable for the current stage in which the traditional Document Web and Semantic Web would co-exist for a long time (while bridging the gap between them).

5. REFERENCES

[1] Asunción Gómez-Pérez, David Manzano-Macho. A survey of ontology learning methods and techniques. Deliverable of IST Project IST-2000-29243 OntoWeb.

[2] Christian Bizer, Richard Cyganiak, and Tobias Gauß. The RDF Book Mashup: From Web APIs to a Web of Data, Scripting for the Semantic Web, SFSW2007.

[3] Tim Berners-Lee. Linked data, 2006. http://www.w3.org/DesignIssues/LinkedData.html.

[4] Chris Bizer, Tom Heath, Danny Ayers, and Yves Raimond. Interlinking Open Data on the Web. In Demonstrations Track, 4th European Semantic Web Conference (ESWC2007).

[5] Jie Han, Yong Yu, etc., "A Hierarchical Model of Web Graph", in Proc. of the 2nd International Conference on Advanced Data Mining and Applications, LNCS 4093/2006.

[6] Nan Liu, Christopher C. Yang: A link classification based approach to website topic hierarchy generation. WWW 2007: 1127-1128.

[7] Tim Berners-Lee et al. Tabulator: Exploring and analyzing linked data on the semantic web. In Proceedings of the 3rd International Semantic Web User Interaction Workshop, 2006. http://swui.semanticweb.org/swui06/papers/Berners-Lee/Berners-Lee.pdf