Linked Open Data Search Engine

Conference Paper · March 2016

DOI: 10.1145/2905055.2905075

Hiteshwar Kumar Azad
VIT University

Linked Open Data Search Engine

Hiteshwar Kumar Azad
Department of CSE
National Institute of Technology
Patna, Bihar, India
azad07it17@gmail.com

Akshay Deepak
Department of CSE
National Institute of Technology
Patna, Bihar, India
akshayd@nitp.ac.in

Kumar Abhishek
Department of CSE
National Institute of Technology
Patna, Bihar, India
kumar.abhishek@nitp.ac.in

ABSTRACT

Linked data indicate a manner of publishing and interlinking structured data on the web. The basic hypothesis behind the concept of linked data is that the value and importance of data increases more when it is interlinked with different data sources. This interlinked web of data is termed as the Linked data. Searching data and providing the most relevant information in linked data is a big challenge. A search engine's utility depends upon the relevance of the search results it returns. Traditional search engines are made to search data on the World Wide Web, where the data are not interlinked. On the other hand, Linked Data based search engine will have to operate over an interlinked web of data. Yet another challenge is to rank the search results. The searched term or phrase can be present in numerous web pages. The usefulness of information present in some pages may be greater than others. So in order to provide the most relevant data, Search engines need to apply various ranking methods on their search results. However the methodology used for ranking data cannot be used in the same manner as it is used in traditional search engines, because the probability of a random user to visit a particular link is not equally likely. In this manuscript a methodology for ranking linked data has been proposed. Also, we have categorized the search into two basic types as Forward search and Backward search. The aim of this bifurcation is to minimize search delays and to provide the end user the data that he or she is most probably looking for.

Keywords

Linked open Data, Linked Data, LOD Search Engine, Semantic Search engine, Semantic web.

1. INTRODUCTION

The term Linked Data [5] was first coined up by T. Berners-Lee and provide the principals for publishing and connecting the structured data on the web. Technically, Linked data is practically utilizing the HTTP and RDF to publish and interlink the structured data on the web to different data sources. We can say the linked data is an extension of the semantic web because many features of semantic web, exists in linked data. However, searching, optimizing and ranking are big challenges in Linked Open Data Search Engine.

The first paper, describing PageRank [10, 15], was published in 1998. The measure of probability of the tendency that a user randomly clicking on links will reached at a particular page, is termed as PageRank. At the beginning of the PageRank calculation the probability distribution is equally distributed among all documents in the group. However, the probability of following a particular outgoing link on the linked data is not uniform. Hence, the distribution cannot be divided equally between all documents in the group. Therefore, the random surfing model cannot be employed for the linked data.

This manuscript tries to formulate a new ranking methodology that can be used for the linked data. The search results are generated after undergoing three steps. First, the result set is formed by collecting data from various domains in linked data using a simple text based search. Second, the result set obtained after collecting data from domains is optimized. The optimization is performed to remove any duplicate data that may have been collected from different sources. Finally, the ranking methodology is applied. In this paper, we have described this ranking methodology. This paper also describes two types of searches, Forward search and Backward search, which is classified on the basis of the end results desired by the user.

This manuscript concisely introduces the term RDF and Ontology used throughout this manuscript.

RDF: Linked data depends upon documents containing data in RDF (Resource description Framework) format. The RDF data model is a graph based data model which is used to publish data on the web. The statements about the resources are constructed in the form of subject, predicate,object. The subject denotes the resource, object denotes the value of a particular characteristic of the resource, and predicate denotes the relationship between the subject and the object. For instance, consider this statement, "The car has the color black". Here the car is the subject, black is the object, and 'has the color' is the predicate. The RDF model depicts data in the form of triples as subject, predicate, object. Resources are identified using Uniform Resource Identifiers (URI). The subject can be a URIs identifies a resource.The object can be a URI or an actual string. The predicate is also represented by a URI and is used to

ICTCS '16, March 04-05, 2016, Udaipur, India
© 2016 ACM. ISBN 978-1-4503-3962-9/16/03…$15.00
DOI: http://dx.doi.org/10.1145/2905055.2905075

specify the relation between the subject and the object.
Ontology: Ontology define the concept and relationship to enable data integration. It is used to describe and stage an area of interest. It is a collection of URIs (Uniform Resource Identifiers) with meanings. Ontology is also an RDF document. The primary role of ontology is to classify things in terms of semantics or meaning. This is usually achieved by the description of Individuals, Classes, Attributes, and Relations. Individuals are the instances or basic objects, Classes are the collection of concepts, Attributes are the properties or characteristics of that object, and Relations are the ways in which Classes and Individuals can be related to each other.

The rest of the manuscript is organized as follows, section 2: introduces a number of related works, section 3: briefly describes types of search, section 4: explain a number of approaches for collecting data into a single data set called the result set, section 5: describes optimization of result set, section 6: describes the ranking methodology, section 7: conclusive remarks.

2. RELATED WORK

In the present scenario large number of linked data being published on the web day by day, numbers of researchers are working to build a platform that uses the web of linked data. Examples of Linked Open Data Search Engines, that crawl the linked data from semantic web search engine (SWSE), Swoogle [9], Falcons [6], Sindica and Watson. Google and Yahoo have also begun to crawl web of linked data from its triple store. Google uses the crawls triple linked data for its public chart API to improve the search result fragmentation for people, reviewers and products. Yahoo grants access to crawl linked data over its BOSS API and uses it within searchMonkey to enhance the search result. PageRanks [10,15] and HITS [12] algorithm measure the web page's utility by analyzing their link structure. In these approaches the probability distribution is uniformly distributed among all web pages in the group. However, when the probability of following a particular link is not uniform, such as semantic web the above mentioned approaches fail to provide the desired result. Recent works [3,16] have now started to consider this non-uniform probability of following a link. Some of these works have been described briefly below.

Pop-Rank [14] is an object-level link analysis algorithm. It automatically accredits a popularity propagation factor for every type of relations. Pop-Rank is domain independent; rank is calculated for each and every page.

ObjectRank [4] system perform as authority-based ranking, it uses keyword search to rank objects in the database that are semantically related. In authority based algorithm, initially the authority is generated at the node (object) which contains the keyword and then this authority flows to other objects according to their semantic relations.

The Swoogle search engine [9] uses OntoRank, a version of PageRank for the Semantic Web. The OntoRank algorithm employs the link analyze method. The importance of ontology is evaluated in a static manner, and the user query is not considered as an effective component for ranking the results. Swoogle uses a rational random surfing model which describes the different types of links which can occur between semantic web documents. They compute popularity of resources by applying link analysis at query time.

The Falcons [6] Search engine searches for entities over RDF data. It maps, keyword phrases to query relations between entities, and quickly restricts initial results using class hierarchies.

Ding [8] offers a semantic ranking of RDF datasets, and is based on VoID (Vocabulary of Interlinked Datasets) descriptions of the datasets. It analyses the link between datasets by utilizing the information furnished by the VOID explanations. It takes into account the types of relationship and number of link sets. They use an automatic weighting scheme to assign appropriate weights for every relation type. However, currently there are not many VoID descriptions available, so their approach is theoretically less extensible.

Azad et al. [1, 2] propose the semantic synaptic web mining(SSWM) which is the most organized and ideal forms of the web, that combine the semantic web and synaptic web at low entropy and provide the most relevant and machine understandable data known as linked data. It also presents the measurement technique of web content and algorithm for SSWM.

3. SEARCHING TRIPLES

When a certain term is searched, the information intended is either attributes of the searched term, or information on terms that has the searched term mentioned as its attribute. In linked Open Data search engine, text based search is performed on triplestore databases of all the domain present on the linked open data and the output represents as RDF triples, where the triple's subject belongs to one server and object belongs to another. The Triple's predicate determines the types of linked between them. In this manuscript, block diagram has three domains (Domain A, Domain B and Domain C) as shown in figure 1 and figure 2. Each Domain splits into three columns as subject, predicate and object as like triple. Based on the requirement of the user, search has been categorized into two categories.
Forward Search: When the user wants to search about a term and its attributes, the user has triple's subject and searches for a triple's object, then the search can be identified as a Forward Search. In this search, we have to select those Domain which has searched term available at triplestore and collect the triples that contain the searched term as the subject. For example, if a user is searching for a "disease" and its symptoms, then the search that user proceeds with, is a Forward Search because it select those triple that contain "disease" as the subject column.
Backward Search: When the user wants to search about a term present as an attribute of other terms, the user has triple's object and search for a triple's subject, then the search can be referred to as a Backward Search. In this search, we have to select those Domain which has searched term available at triplestore and collect the triples that contain the searched term as the object. For example, when a user searches for "symptoms" and the diseases that have those symptoms, then such backtracking of information can be termed as Backward Search because it select those triple that contain "symptoms" as the object column.

4. DATA EXTRACTION

First step is to retrieve the most relevant data from the search query and the extracted data from different data sources is stored in the result set in triple form. In this section, various methods for data extraction from different

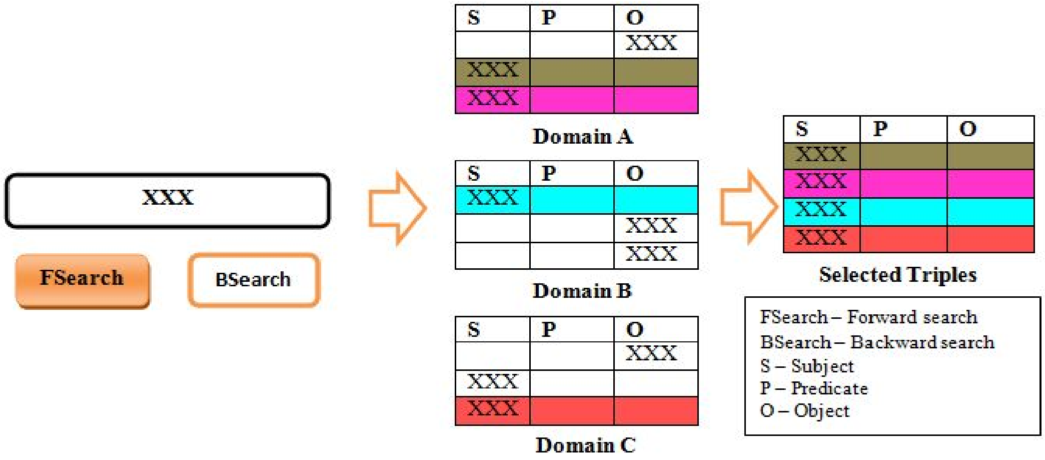

Figure 1: Block diagram of forward search.

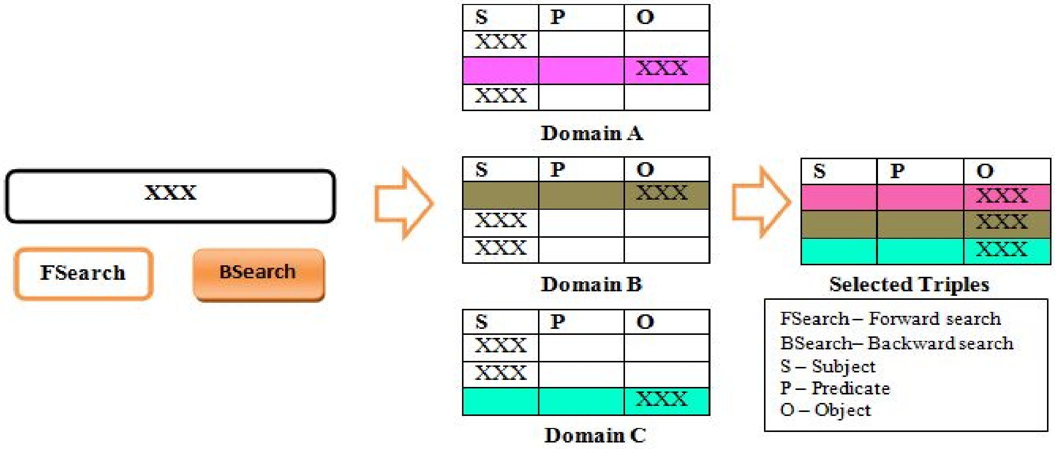

Figure 2: Block diagram of backward search.

domains will be discussed.

Appending datasets: The desired datasets are retrieved from triplestores of various domains using queries. The extracted datasets are then placed one after the other same as the triple of a dataset are placed after another dataset's triples. This is the simplest way of gathering data from different data warehouse.

Extraction of triples using SPARQL: SPARQL command is used to extract the desired data from RDF based large data sources. Linked Open Data can be queried in three different ways. The first way is to extract data from distinct data sources by querying them independently and then merging the extracted data. The second method works in the following way: According to the output of the earlier queries and substituting the output in the place of the query pattern, consequently queries are employed over different data source. The third method is the application of the present SPARQL restriction that grant us to access different types of suitable data origin.

Traversing HTML web pages: In this method a desired domain is determined and based on some standard of appropriate data; desired data are extracted from available data sets using a data extractor that automatically traverses them. Finally the extracted data from different datasets are merged together. In a different approach, the data extractor initially selects a dataset for collection of related data from it and then traverses other appropriate datasets by using information acquired from previous extracted data. As, for example by using owl:seeAlso or owl:sameAs predicates [13].However the types of method chosen, the data extractor must be compatible with different data collection that have distinct ontologies and their architecture. Also, the complication for the inaccessibility of contents of a dataset must be resolved by traverser.

5. OPTIMIZING THE RESULT SET

Ontology Mapping: Past few years,the web has an explosive growth of ontology publicly available and accessible on the web, so we need an application to use them. Various numbers of ontologies required to access from various applications. A mapping could provide a common platform for exchanging the information and provide the semantic meaning of the information. Most of the datasets use one or more ontologies for defining the data. Thus, each dataset normally has different ontologies. Hence, it may be the same data with distinct names or distinct data with the same name present in the result set. One solution to this problem is to use ontology mapping concepts [7, 11]. There are various approaches to perform ontology mapping, however, we will not be discussing those approaches here.

Duplicated Data: Duplicated data are generated because of two reasons. Firstly, when there are two alike subject and predicate in distinct datasets with distinct object value. Secondly, when there are various object values for a predicate of a particular subject. So there is only one of the duplexed value (the genuine value) must be picked from the result set. The solution to this problem is to select the most authentic data and exclude the other duplexes. In this paper, we have solved the problem by choosing data from specific domains and removing additional duplex data.

6. RANKING

The proposed methodology for ranking the RDF dataset is query dependent, such as the scores resulting from the link analysis is influenced by the search terms. A simple text based search is used for collecting triples to form the result set. The ranking methodology is applied on the result set after its optimization. Two ranking parameters are used to provide the most relevant results.

6.1 Domain Rank

Currently the domain in the linked open data is divided into nine major areas; publications, geographic data, life sciences, media, user generated content, cross domain, social networking, government and linguistic data source. It is more likely that data from a specialized domain will contain more relevant data than other domains. For example, if we are searching for symptoms of a disease, then a domain which specializes in the field of life sciences can provide the most valid data. Hence, we have categorized data initially on the basis of the domain. A simple algorithm for categorizing domains on the basis of relevance is as follows. It is a non-iterative ranking algorithm, as the ranks are influenced by the search terms. The goal of the algorithm is to categorize the domains into three groups High-Rank, Mid-Rank, and Low Rank.

Table 1: Algorithm for domain ranking

Assumption:
Statement 1: It returns maximum number of triples.
Statement 2: The domain has maximum number of Ontologies, which define the searched term.
Statement 3: In the Ontologies of a particular domain, the total number of distinct properties present in the subclass of the searched term is maximized.
Statement 4: It is linked by the High-Rank group domains for further information on the queried term.

```
begin

if (statement 1 || statement 2 || statement 3) then
A domain is placed in a High-Rank group.
else if (statement 4) then
A domain is placed in a Mid-Rank group.
else
A domain is placed in a Low-Rank group.
end
```

6.2 Page Rank

The purpose of Domain's rank was to provide the most valid source of data, but it is not sufficient. Within a domain there can be multiple documents providing information regarding the searched term, and so ranking them becomes evident. In order to rank these documents we have used Swoogle's ranking method for SWDs (Semantic Web Documents).

Swoogle [9] employs a rational random surfing model which describes different types of links existing between SWDs (Semantic Web Documents). Let us consider two Semantic Web Documents P and Q, the four classifications of inter-SWD links are as follows:

1. Imports (P, Q), P import all content of Q.
2. Uses-term (P, Q), P uses some of terms determined by Q without importing Q.
3. Enhances (P, Q), P enhances the definitions of terms determined by Q.
4. Asserts (P, Q), P makes condition about the individual determined by Q.

The probability of following these links are not uniform. Hence different weights are assigned to the four groups of inter-SWD relations. Also, the number of references is taken in consideration for computing the rank, because the probability that a server will visit the link from P to Q is higher if the number of terms in Q referenced by P is greater. Based on the above conditions, the rank of SWD m is determined using the following equation.

$$Rk(X) = (1-d) + d \sum_{n \in S(m)} Rk(n) \frac{f(n,m)}{f(n)} \tag{1}$$

$$f(n,m) = \sum_{l \in links(n,m)} weight(l) \tag{2}$$

$$f(n) = \sum_{m \in U(n)} f(n,m) \tag{3}$$

Where S(m) is the set of Semantic Web Documents that links m, U(n) is the set of Semantic Web Documents that x links to d is the damping factor (similar to Page Rank's direct visit component), n is the current Semantic web database, m is the Semantic Web Documents that n links to f(n, m) which is the sum of all link weights from n to m, and f(n) is the sum of the weights of all outgoing links from n.

From above equations we can see that the PageRank's direct visit component (d) is retained, but the link is chosen with different probability $-\frac{f(n,m)}{f(n)}$, where n is the current Semantic web database, m is the Semantic Web Documents that n links to f(n, m) which is the sum of all link weights from n to m, and f(n) is the sum of the weights of all outgoing links from n.

Final Result: The result set is divided into three major groups, High-Rank, Mid-Rank and Low-Rank group. There can be more than one domain present in the above groups. To arrange these domains, we calculate the average of Page Ranks of the semantic web documents of each domain. The domains are then arranged in descending order of their average Page Rank value.

The final result set is first arranged according to the domain rank groups (High-Rank group followed by Mid-Rank group and finally Low-Rank group), within the groups the domains are arranged in descending order of their average Page Rank value, and within the domain the semantic web documents are arranged in descending order of their Page Rank.

7. CONCLUSIONS

Current page rank algorithms do not work well with RDF documents since the semantics of link refers to a different probability by following a particular outgoing link. This manuscript presents a new approach to rank RDF data, distributed across various domains in order of relevance. The ranking of RDF datasets is query dependent. The result set is formed by collecting data using various methods, and then optimizing the final result set. The ranking methodology is applied to this result set. For Ranking the result set manuscript uses the Domain Rank groups and Swoogle's Page Rank method. Domain rank groups are formed to find the source of the most relevant data. To Page Rank different weights are assigned, which depicts the probability of following that type of links, to the four classes of inter-SWD relations.

This manuscript also describes two approaches for searching, Forward Search and Backward Search. The main goal behind the use of these approaches is to provide the most relevant result by making the search efficient and less time consuming.

8. REFERENCES

[1] Hiteshwar Kumar Azad and Kumar Abhishek. Entropy measurement and algorithm for semantic-synaptic web mining. In *Data Mining and Intelligent Computing (ICDMIC), 2014 International Conference on*, pages 1–5. IEEE, 2014.

[2] Hiteshwar Kumar Azad and Kumar Abhishek. Semantic-synaptic web mining: A novel model for improving the web mining. In *Communication Systems and Network Technologies (CSNT), 2014 Fourth International Conference on*, pages 454–457. IEEE, 2014.

[3] Ricardo Baeza-Yates and Emilio Davis. Web page ranking using link attributes. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, pages 328–329. ACM, 2004.

[4] Andrey Balmin, Vagelis Hristidis, and Yannis Papakonstantinou. Objectrank: Authority-based keyword search in databases. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 564–575. VLDB Endowment, 2004.

[5] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data-the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pages 205–227, 2009.

[6] Gong Cheng and Yuzhong Qu. Searching linked objects with falcons: Approach, implementation and evaluation. 2009.

[7] Namyoun Choi, Il-Yeol Song, and Hyoil Han. A survey on ontology mapping. *ACM Sigmod Record*, 35(3):34–41, 2006.

[8] Renaud Delbru, Nickolai Toupikov, Michele Catasta, Giovanni Tummarello, and Stefan Decker. Hierarchical link analysis for ranking web data. In *The Semantic Web: Research and Applications*, pages 225–239. Springer, 2010.

[9] Li Ding, Rong Pan, Tim Finin, Anupam Joshi, Yun Peng, and Pranam Kolari. Finding and ranking knowledge on the semantic web. In *The Semantic Web–ISWC 2005*, pages 156–170. Springer, 2005.

[10] Taher Haveliwala. Efficient computation of pagerank. 1999.

[11] Yannis Kalfoglou and Marco Schorlemmer. Ontology mapping: the state of the art. *The knowledge engineering review*, 18(01):1–31, 2003.

[12] Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[13] Deborah L McGuinness, Frank Van Harmelen, et al. Owl web ontology language overview. *W3C recommendation*, 10(10):2004, 2004.

[14] Zaiqing Nie, Yuanzhi Zhang, Ji-Rong Wen, and Wei-Ying Ma. Object-level ranking: bringing order to web objects. In *Proceedings of the 14th international conference on World Wide Web*, pages 567–574. ACM, 2005.

[15] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[16] Wenpu Xing and Ali Ghorbani. Weighted pagerank algorithm. In *Communication Networks and Services Research, 2004. Proceedings. Second Annual Conference on*, pages 305–314. IEEE, 2004.