DSRank: A New Hyper-Linked Based Method to Rank Datasets in LOD Cloud

Hamidreza Fardad[1], Hadi Khosravi-Farsani[2], Mohammadali Nematbakhsh[3]
1- Department of electrical and computer engineering, Majlesi Branch, Islamic Azad University, Isfahan, Iran
Email: hfardad@iaumajlesi.ac.ir
2- Computer Engineering, Shahrekord University, Shahrekord, Iran
Email: Khosravi81@gmail.com
3- Computer Engineering, University of Isfahan, Iran
Email: nematbakhsh@eng.ui.ac.ir

Received: April 2013 Revised: July 2013 Accepted: Sep. 2013

ABSTRACT:
The increase of available datasets in web of data, causes the ranking of the datasets become very important. The present article, the famous PageRank algorithm is extended and a new link-based method is proposed for ranking the datasets in web of data. In this method, the number of links to the dataset, the type of the links, and the number of each type of the links has been considered and a new hyper-linked based approach name as DSRank is proposed. The suggested algorithm has been implemented on datasets through collecting from the web amounting to 20 GB. All of the datasets are arranged by using suggested method. In order to evaluate, the access log files of Dbpedia, DBTune, and Dog Food are used. The number of requests by users in one day for these datasets are calculated and then datasets are organized based on user's opinion. The results of comparing our suggested algorithm with the number of the requests by the users in a day indicate that the order of the assigned ranks in the dataset through using the proposed method is correct.

KEYWORDS: Ranking, Linked Data, Web of Data, Semantic Web.

1. INTRODUCTION

The current web can be described as a file system with each file encompasses so many concepts. The lowest level of accessibility in the current web is documents which are linked through hyperlinks. The links among web pages are of no specific type and an individual can easily recognize the meaning behind a link between two pages i.e. the meaning behind links and the content of the pages are implicitly understood. Moreover, web pages have weak structure and hence it is not possible to automatically find and extract data thereof using a machine. In this structure, HTML and XML documents have been linked through un-typed links; yet, there is no understanding of the meaning of the pages and their links by the machine.

The challenges of the current web can be divided into three main categories (heath, 2009). The first problem pertains to the simplicity of displaying information. Having unstructured data, is containing links without a specific type, as well as disintegrated data all lead to simple publication of data. The second problem has to do with the lack of integrity in various databases. In other words, the required information is available in several databases and the user has to look for the information himself. The third problem of the web is searching which is, at the present time, based on a set of keywords. In this case, the needs of the user are not properly satisfied.

In 2007, Berners-lee, the inventor of the web, proposed Linked Data to overcome the problems (Berners-Lee, 2006). In the presented structure called Web 3, documents are not linked but data are linked through meaningful links. This structure has been shown in Figure 1. Unlike the previous method, entities are the lowest level of accessibility that pointed to other entity through typed links. In fact, various entities are defined in different databases and are linked through pre-defined links. The machine understands the meaning and the relationships among entities. Hence, this structure promotes the web to a distributed database in which finding of precise information can be easily reached. Each of the references (Bizer et al., 2009; Bizer et al., 2008; Heath, 2009; Bizer et al., 2009; Shinavier, 2009) explains the different aspects of the linked data.

By now, a number of organizations and companies have published their data to the form of linked data. To publish data in this format, three basic steps are required. First a URI is chosen for each of the entities (Berners-Lee et al., 2005). Then the required ontology for publishing is determined. It has been suggested to use standard ontology in this respect. In the final step, a link is made from the considered dataset to other datasets. The link enables the search engines and users to access more information about the defined entities in the datasets in question. Figure 2 shows a part of the published datasets in linked data. For example, Wikipedia database has been transformed to linked data shown with DBPedia (Bizer et al., 2009; Auer et al, 2009) which function as the central hub. The other famous datasets which can be seen in the Figure include Geonamies consisting of geographical concepts[1], WordNet linguistic ontology (Miller, 1996), Flicker entailing music data, DBLP encompassing articles [2].

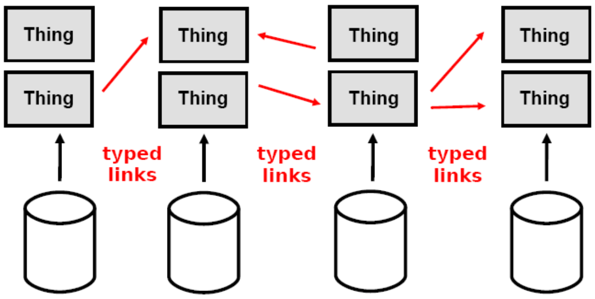


Fig. 1. The structure of the Linked Data

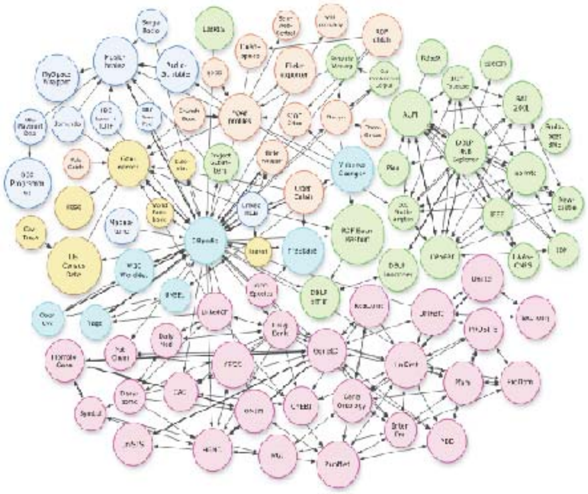


Fig. 2. Various Datasets in the Linked Data

One of the fundamental parts of the search engines is ranking, whose function is finding high quality pages and organizing the output based on the priorities set by the user. This search engine requires an efficient algorithm for ranking web pages (Brin et al., 1998; You et al., 2008). If we describe the documents as nods and the links among them as edge, we will have a very large graph. A node with high-importance in-edge has high rank. The most effective methods of document ranking in the present web apply the same graph structure in which the links among the documents are analyzed and a number indicating its ranking is reported.

In the new web structure, the data presentation, search engine, publishing tools, and other technologies which were used for the traditional web must be reexamined and modeled. In spite of, a large number of previous models and algorithms would be expanded and revised for the new web structure.

One of the influential parameters in ranking an entity in web of data is the source in which the entity is defined or the dataset which defined that entity. The present article is an attempt to propose a method for ranking datasets. Ranking the dataset in web of data is of prime importance. Once an organization decides to publish its data in LOD cloud, it is essential that it is aware of the dataset which is the most important. In this case, it can link its entities to the entities available in a more importance dataset. Furthermore, since the dataset ranking is used by search engines, the entities in a more important dataset will be more important.

In order to estimate the rank of a dataset, the number of links to the dataset, the type of the links, and the number of each type of link have been used. Section 2 deals with the review of the literature. In section 3, the weights of all types of the links are taken to be the same and datasets are ranked accordingly. In section 4, each link is given a different weight and ranking is repeated. The evaluation of the suggested methods on the basis of the number of requests from the web server of some of the datasets is presented in section 5. Finally, section 6 is devoted to the conclusions and further studies.

2. LITERATURE REVIEW

Information retrieval systems fetch a list of pages based on the given query and considering page title, content, etc and select some relevant pages. Then, they use ranking algorithms for arranging the documents according to their similarities with the given query. To this end, both the documents and the query are converted into a format which can be processed by computers. VSM is used to model the document and the query in that they are displayed as a vector of the index terms. Assume that a total number of *t* index terms are available in the whole set, then document *D* and query *Q* are shown as follows:

$$D = (W_{d1}\ , W_{d2}, W_{d3}, \ldots, W_{dt})$$
$$Q = (W_{q1}, W_{q2}, W_{q3}, \ldots, W_{qt})$$

[1] http://www.geonames.org/ontology/

[2] http://www4.wiwiss.fu-berlin.de/dblp/

In which W_{di} is the weight assigned to different terms for document *D*. The similarity between a document and a query is estimated through the COS of the angle between the two vectors as you can see in formula 1 (Salton et al., 1998):

$$Similarity\ (Q,D) = \frac{\sum_{i=1}^{t} W_{qi} * W_{di}}{\sqrt{\sum_{1}^{1} (W_{qi})^2 * \sum_{1}^{t} (W_{di})^2}} \qquad (1)$$

The experiments reveal that merely using the content of the documents is not fruitful in ranking the documents (Fan et al., 2004). At the present time, link structure with content of document is generally used to rank documents. Page rank algorithm was first used by (Page et al., 1999; Ridings et al., 2002) which is the most famous ranking algorithm for the time being and used in Google search engine. The popularity of this algorithm comes from its being independent of the user's query and assessing the quality of a page only by analyzing its links. Google first selects a list of pages based on page title, page content, etc. and then applies PR to arrange the results according to their importance. Simply put, PR algorithm shows that if there are important links to a page, the links referring from that page to other pages will be important, too. PR is also able to take backlinks into account and publish the ranking according to the links. A page is highly important provided that the sum of its backlink ranking is high. Formula 2 show the simple PR algorithm for estimating the rank of page *u* in which N_v shows the number of the total output links from *v* an B(*u*) represents the number of the pages referring to page *u*.

$$PR(u) = c \sum_{v \to B(u)} \frac{PR(v)}{N_v} \qquad (2)$$

The ranking of the pages of a web can be estimated repeatedly and by beginning from a specific page of the web. It is possible that two or more pages of a website are linked and form a cycle. If these pages do not refer to anywhere and yet referred to by pages outside the cycle, they have accumulated the rank and not distributed it, the problem referred to as *rank sink* (Page et al, 1999).

To overcome this problem, the user's activities can be checked. What can be seen is that the users do not follow the links. For example, the user might decide not to follow the links after seeing page *a* and decide to directly enter page *b* i.e. he types the address for *b* in the search bar. In this case, the rank pertaining to *b* has resulted from page *a* even if these two pages are not directly linked. Accordingly, the PageRank algorithm will be formula 3 in which *d* is the probability of following the link by the user. *d* is called Damping Factor which is valued as 0.85 for the time being.

$$PR(u) = (1-d) + d \sum_{v \to B(u)} \frac{PR(v)}{N_v} \qquad (3)$$

Although the standard PR algorithm (Exp. 3) is famous and highly efficient, researchers have aimed at enhancing its efficiency and speed. Weighted PR (Xing et al., 2004), N-step PR (Bao et al., 2007), Probabilistic PR, and WLRank (Baeza et al., 2004) are some versions whose specifications can be found in the references.

Another algorithm named HITS was proposed by Chakrabarti for ranking web pages in 1999 (Chakrabarti et al., 1999). This algorithm analyzes the web pages based on the input and output links. In this algorithm, web pages referred to by many hyperlinks are called Authority and those referred to by many pages are called Hub. These two types of pages are shown in the Figure 3.

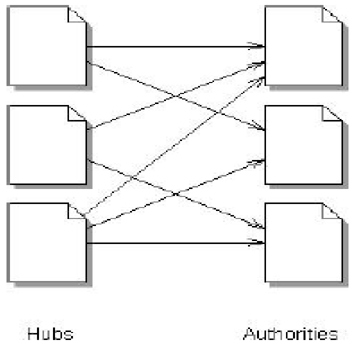


Fig. 3. Authorities and Hubs

A page can be ranked in the following way. The values for authority and hub are estimated for each page. An authority referred to by many hubs can be an importance one. Similarly, a hub referring to many strong authorities is an importance one. Assume that a_p and h_p indicate the authority and hub of page p. Then, B(p) is the number of referring pages and I(p) is the number of the citations. Using 4 and 5 one can estimate a_p and h_p pertaining to page p. Figure 4 shows the way these values can be estimated for page p (Cohn et al., 2000; Ding et al., 2004; Kleinberg, 1999).

$$a_p = \sum_{q \to B(P)} h_q \qquad (4) \qquad h_p = \sum_{q \to I(p)} a_q \qquad (5)$$

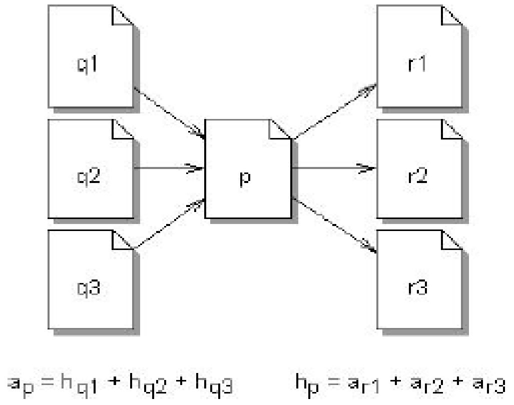

q1
q2
q3
p
r1
r2
r3
$a_p = h_{q1} + h_{q2} + h_{q3}$
$h_p = a_{r1} + a_{r2} + a_{r3}$

Fig. 4. Values of Hub and Authority pertaining to a page

Unlike a web of documents, in a web of data an Object Retrieval System is designed and performed. Mr. Pound in (Pound et al., 2010) defines such a system as follows:

Input: the user's query *q* (structureless search) of type *t*, intention *z*, and data graph G.

Output: an organized list of the sources O = (o_1, o_2, … , o_k) is given in which any of the o_i's are presented in the data graph.

Evaluation: a number is assigned to any entity o_i independent of the others reflecting the extent of its relation to the user's query and his intention.

The type of the query by the user in web of data is essentially different from web of documents. Different types of the query in an entity retrieval system can be divided into the following:

- Entity query: the user looks for a specific entity in this query.
- Finding instances of a class: the entities which are instances of a specific class are looked for.
- Feature query: the aim is to find values of a feature from an entity.
- Relational query: the aim is to find how two or more entities are related.
- Other query: any queries which cannot be categorized as the above.

Different types of search exert a great influence on ranking since each query requires different results and is evaluated separately. Hence, any attempt to map the entity retrievals to document retrievals is doomed to failure. In these cases, the concepts used in document retrieval systems such as evaluation criteria can be applied (Pound et al., 2010). Mapping the data graph of the new web to the traditional one and using traditional web ranking methods are used in article (Harth et al., 2005). To get a better appreciation of this method, take in to account Figure 5. The entity Jim is defined in source A, entity Tim in source B, and entities Mary and Bob in source C. The context graph has been shown in Figure 6. Source A has used Bob in source C which indicates a vote from A to C. Similarly, C has used Tim defined for B indicating that a vote is from C to B. If such link-based ranking algorithms as PageRank or HITS are used, the ranking of each of the datasets can be estimated and transferred to the available entities and therefore the importance of each of the entities can be estimated.

Article (Hogan et al., 2008) has ranked the entities based on two concepts. The first concept, linked graph, shows the structure of the links among the entities (entities include those which have appeared as Subject at least once). Figure 5 represents the links among the Entities. The second concept is the context graph which shows the links among different sources. Regarding these concepts the following elements can be taken into account in ranking the entities:

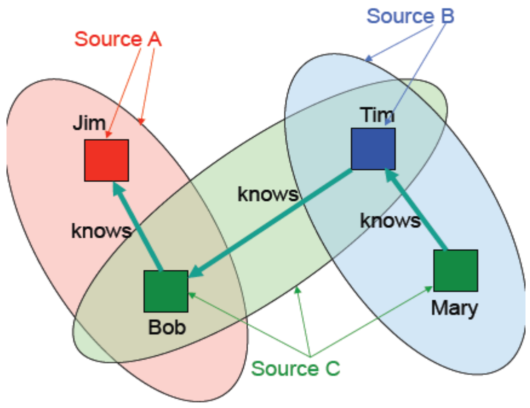


Fig. 5. A part of the Web of Data graph

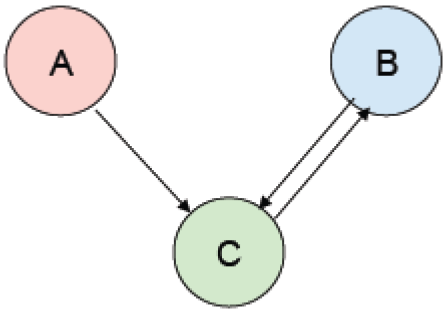


Fig. 6. The context graph derived from figure 6

1. The importance of a context hinges upon the data and the sources describing thereof.
2. The resources appearing in important contexts are important.
3. The resources appearing in several contexts are important as well.

The datasets in which one entity is defined exert extensive influence in the importance of that entity. By the same token, Toupikov has ranked the datasets using void[3] descriptions (Toupikov et al., 2009). Void is a standard format to describe datasets (Alexander et al., 2009). Another important paper for ranking dataset is the one by Renaud (Delbruet al., 2010). They proposed layered approach to rank entities. In the first layer, PageRank has been used to rank datasets. After that, this value is combined with entity rank which is calculated at the second layer.

The main weakness of the previous methods of ranking can be accounted for in this way: in a number of the previous methods, no dataset in which an entity is defined has been taken into account for ranking the entities. In some other methods, the type of the link used for citing the entities is not accounted for. Moreover, the number of the links from one dataset to another is ignored. In order to rank the entities in the present article, datasets are primarily ranked and to this end two methods are proposed. In the first method, datasets are converted into a graph in which the nods show datasets and among them an edge presents the number of links among them. In this case, PageRank algorithm ranking has been extended and another algorithm named DataSetRank (we here call it, DSRank) has been used. In the second method, the type and the number of links from one dataset to another has been accounted for. This way, a graph can be made in which the nods show the datasets. Also, there are different types of edge between two nodes which shows the number of various attributes. Section 3 and 4 elaborate these methods accordingly.

3. THE PROPOSED METHOD FOR THE CASE WHERE CITATION ATTRIBUTES ARE EQUALLY WEIGHTED

An entity in a web of data has multiple triples. A triple includes Subject, Attribute, and Value. Subject and Attribute are usually a URI. Value can be a numerical one or a URI. We assume in the current implementation, if the attribute of a *type* is in one of the triples pertaining to a URI, it is regarded as an entity. The attribute of *type* is defined in *http://www.w3.org/1999/02/22-rdf-syntax-ns*.

The entities in a web of data are defined in a variety of datasets. For instance, the entity *http://dbpedia.org/page/Berlin* is defined in *dbpedia* dataset. In the implementation carried out, according to the entity of a URI, its dataset is derived. It must also be pointed out that in the collected triples from the web; a huge number of entities have not defined in datasets and appearing in other formats. For example, RDFa (Adida et al., 2008) and Microformat documents (Çelik et al., 2008) can define entities. Such entities do not form datasets. It is assumed in this article that the number of the entities in a dataset must be larger than a specific threshold. In all the tests carried out, the minimum number of entities to for a dataset is regarded as 50.

In a web of data, different types of attributes have been used to link. In the first column of table 1, a few numbers of the most prominent attributes have been presented. For example, the attribute Sameas defined in w3.org ontology is used to express the equality of two entities. The attribute See Also is defined in w3.org. Once an entity uses this link and refers to another entity, it implies the user can use the pointed entity to understand the first one. The field for the value of a triple which has used the citation attributes is always the address appertaining to another entity in the web.

Table 1. Association Attributes

Association Name	Weight on the first Experiment	Weight on the Second Experiment
http://dbpedia.org/property/reference	1	2
http://www.w3.org/2002/07/owl#sameAs	1	2
http://www.rdfabout.com /rdf/schema/usbill/relatedTo	1	1
http://www.w3.org/2000/01/rdf-schema#seeAlso	1	1
http://www.rdfabout.com/ rdf/schema/usbill/identicalTo	1	0.5
http://purl.uniprot.org/core/source	1	3
http://purl.uniprot.org/core/locatedOn	1	1
http://purl.org/dc/terms/license	1	1
http://www.w3.org/2002/07/owl#imports	1	0.5

For modeling the datasets in a web of data, take Figure 7 into account. There are three datasets named A, B, and C in this example. The number of the citation attributes has been shown by a number on the edges. For example, number 5 is on the edge between datasets A and C which implies that dataset A entails five distinct entities referring to the entities in C. Using these citation attributes reflects the ranking transference from dataset A to dataset C. In this case, dataset ranking is changed into nod ranking in the graph presented in Figure 7.

[3] Vocabulary of Interlinked Dataset

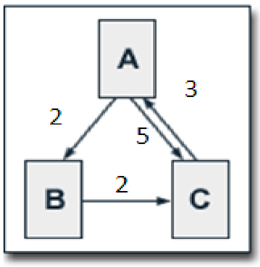


Fig. 7. The sets of data and the number of links among the dataset

In the first proposed method named DS Rank, algorithm 6 has been used to rank datasets. Ranking of the dataset d_i is interpreted in this way: the more high-rank citation attributes refer to a dataset, the higher the rank of the dataset. In Formula 6, *d* is damping factor and shows the possibility of remaining in d_i. Link (d_j,d_i) shows the number of links from dataset d_j to dataset d_i. LinkOut(d_j) indicates the total number of output links from d_j.

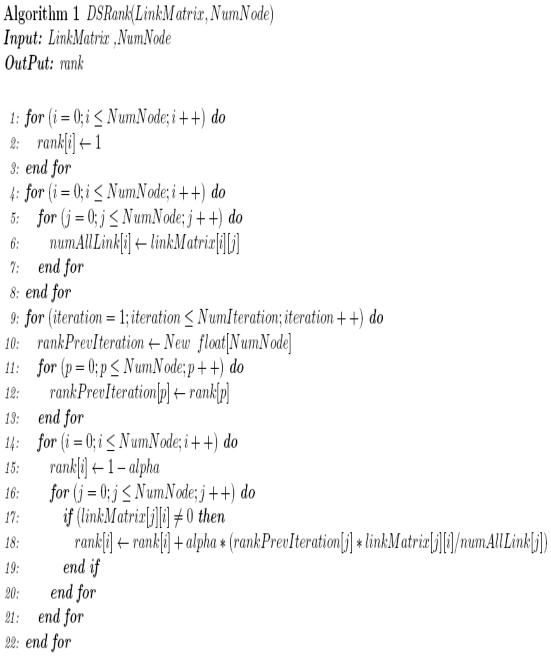
$$DR(d_i) = (1-d) + d\left(\frac{|Link\ (d_1, d_i)|}{|LinkOut\ (d_1)|} * DR(d_1) + \ldots + \frac{|Link\ (d_n, d_i)|}{|LinkOut\ (d_n)|} * DR(d_n)\right) \quad (6)$$

If algorithm 6 is used to estimate the rank of each of the datasets in Figure 7, the first column of table 2 will result. The second column in this table represents the final rank of each of the entities. As it can be seen, C is larger than A and A is larger than B.

Table 2. The formula for estimating the rank of datasets irrespective of the link types

$DR(A) = 0.5 + .5\,DR(C)$	DR(A) = 1.113636
$DR(B) = 0.5 + .5(\frac{2}{7} * DR(A))$	DR(B) = 0. 659090
$DR(C) = 0.5 + .5(\frac{5}{7} * DR(A) + \frac{2}{2} * DR(B)$	DR(C) = 1.227272

Java has been applied to perform the above algorithm. The required data have been collected from triples by a crawler from challenge.semanticweb.org. The size of these data amount to 700GB downloadable in 376 files each having the size of about 2GB. Due to the high overhead of processing of these data and lacking the required bandwidth, the tests have been carried out only on 20GB thereof. It is worth mentioning the format of the files is Quad. Each file is formed of so many lines and each line encompasses four columns with the first indicating subject, the second showing attribute, the third representing value, and the fourth showing the context in which each triple is defined.



```
Algorithm 1 DSRank(LinkMatrix, NumNode)
Input: LinkMatrix ,NumNode
OutPut: rank

for $(i = 0; i \leq NumNode; i++)$ do
  $rank[i] \leftarrow 1$
end for
for $(i = 0; i \leq NumNode; i++)$ do
  for $(j = 0; j \leq NumNode; j++)$ do
    $numAllLink[i] \leftarrow linkMatrix[i][j]$
  end for
end for
for $(iteration = 1; iteration \leq NumIteration; iteration++)$ do
  $rankPrevIteration \leftarrow New\ \ float[NumNode]$
  for $(p = 0; p \leq NumNode; p++)$ do
    $rankPrevIteration[p] \leftarrow rank[p]$
  end for
  for $(i = 0; i \leq NumNode; i++)$ do
    $rank[i] \leftarrow 1 - alpha$
    for $(j = 0; j \leq NumNode; j++)$ do
      if $(linkMatrix[j][i] \neq 0$ then
        $rank[i] \leftarrow rank[i] + alpha * (rankPrevIteration[j] * linkMatrix[j][i]/numAllLink[j])$
      end if
    end for
  end for
end for
```

Fig. 8. A java script program for ranking datasets in web of data

The number of the datasets increased as the number of triples increased. In the tests carried out, regarding the values of the parameters mentioned above, the number of datasets rapidly increases when size of processing data augment. Using only 4GB of the triple resulted in 80 datasets. This implies that there are 80 datasets whose number of entities have been larger than 50. Using sizes of 8, 12, 16, and 20 GB respectively lead to 131, 171, 206, and 245 datasets. These results have been presented in Table 3. In addition, the lowest, the highest, and the average rank have been presented for the ranks of the datasets. It should be mentioned that the highest rank comes from semantic-mediawiki.org dataset.

Table 3. The formula for estimating the rank of datasets irrespective of the link types

Processing Data (GB)	4 GB	8 GB	12 GB	16 GB	20 GB
The number of datasets	80	131	171	206	245
The lowest rank of the datasets	0.5	0.5	0.5	0.5	0.5
The highest rank of the datasets	4.49	7.053	7.25776	7.7241597	8.25
The average rank of the datasets	0.689	0.692673	0.67201	0.67922	0.674523

In majority of datasets, the triples increase as the data size increases. In table 4, the ranks appertaining to some of the datasets are shown. Figure 9 represents table 4 in form of graphs. As it can be seen, the rank of dbpedia.org increases as the number of triples increases. It can be said that the number of the triples referring to this dataset has increased and that the rank of the datasets referring to dbpedia.org has increased as well. The rank of uniprot.org decreases as the number of triples increases. This dataset is related to protein and molecule data whose decrease can be explained this way: as the number of triples increase, the number of the triples that refer to this dataset decreases and the number of links referring to this dataset grows, and consequently the backlink nods of this dataset are divided into the output links and ultimately its rank falls.

Table 4. Ranks for the dataset

DataSet Name	Size = 4 GB	Size = 8 GB	Size = 12 GB	Size = 16 GB	Size = 20 GB
http://dblp.l3s.de	1.36	1.55	1.70	1.98	2.15
http://www.uniprot.org	0.50	0.51	0.51	0.51	0.50
http://purl.org	0.83	1.16	1.28	1.27	1.44
http://www.bibsonomy.org	0.65	0.66	0.71	0.73	0.74
http://dbpedia.org	1.27	1.32	1.59	1.77	1.77
http://dbtune.org	0.70	0.81	1.08	1.07	1.09
http://data.semanticweb.org	0.5	0.5	0.5	0.55	0.54

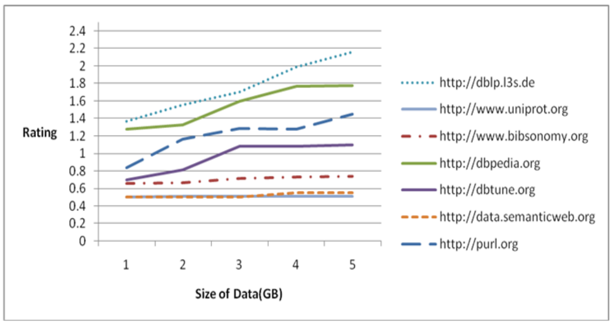


Fig. 9. The ranks of the datasets where citation features are equally weighted

4. THE PROPOSED ALGORITHM WHERE CITATION ATTRIBUTES ARE UNEQUALLY WEIGHTED

The type of the link referring to a dataset exerts influence in ranking transfer from the first dataset to the second. For instance, Same as and See Also in table 1 have different weights at the time of transfer meaning that in using Same as, more credit will be transferred into the second dataset. The third column in table 1 shows the weight chosen for each of the citation attributes. It should be noted that there has been no specific principle in choosing the weight and in all the following tests, the same weight as third column in Table 1 is assigned. For instance, the weight for Sameas is equal to 2 implying that in the ranking, the number of such links will be multiplied by 2.

Figure 10 is the Figure used for the previous case with the only difference being the type and the number of the links. In this Figure, three datasets named A, B, and C is accounted for. For the sake of simplicity, only two links are focused. The solid lines show *Sameas* relationships and the dotted lines show *identicalto* relationships. As it can be noted, there are two *Sameas* links and one *identicalto* link from C to A. Other relationships and their numbers can be seen in the Figure.

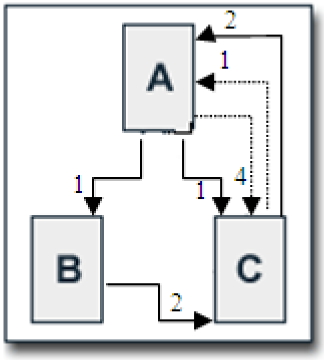


Fig. 10. Datasets and the distinct links between them

Algorithm 7 is proposed for ranking datasets where unequal weights are envisaged for the links. Dataset d_i is explained as: the more citation attributes of more reputation with larger weight refers to a dataset, the higher the rank of that dataset. In formula 7, *d* is damping factor and shows the probability of remaining in dataset d_i. WLink(d_j,d_i) shows the mentioned weighted links from dataset d_j to dataset d_i. Formulae 7 and 8 show the way the total weighted links of a dataset can be estimated. α represents the link weight which is substituted from the third column in table 1. Moreover, WLink (d_j) is equal to the total weight of all of the output links from dataset d_j.

$$DR(d_i) = (1-d) + d\left(\frac{|WLink(d_1, d_i)|}{|WLink(d_1)|} * DR(d_1) + \ldots + \frac{|WLink(d_n, d_i)|}{|W(d_n)|} * DR(d_n)\right) \tag{7}$$

$$/Link\ (d_j, d_i) = \sum_{t \to (Associatio\ \ nAttribtes\ )} \alpha_t * |link_t(d_j, d_i)|$$

$|link_t(d_j , d_i)|$ = Number of link of type t from DataSet d_j to d_i

α_t = weight of the link to t'th of Association Attribute

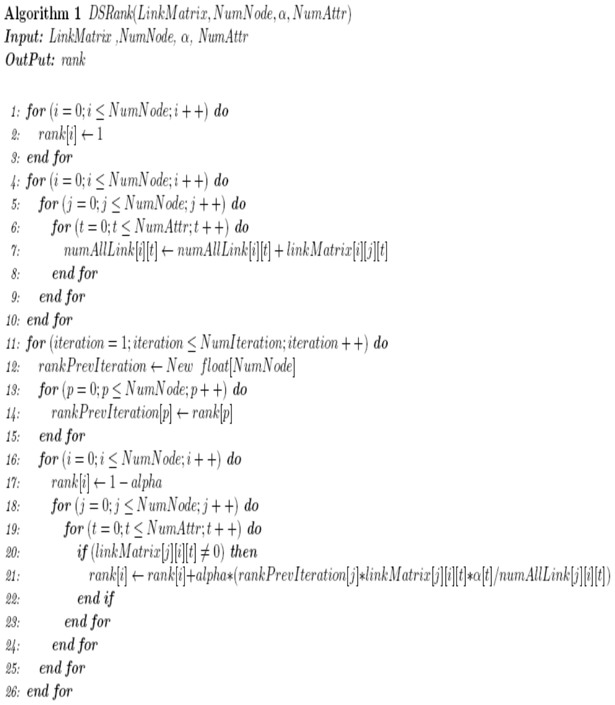
(8)

Using the above relationships for estimating the Dataset ranks in figure 10, we will have the formulae in the first column of table 5. Having estimated these relationships, the estimated ranks for each of the datasets has been presented in the second column of table 5. As it can be noted, C has a higher rank than A and A has reached a higher rank than B. Yet, compared with table 2, C has a lower rank than the previous case. The reason is that 5 relationships from A to C in the previous case had the coefficient of 1 while there are 4 relationships which have the coefficient of 0.5 and one relationships which has coefficient of 2. By the same token, the rank in dataset B has increased and in A dramatically decreased.

Table 5. The formula for estimating the dataset rank regarding the type and the weight of the link

DR(A) = 0.5+0.5DR(C)	DR(A) = 1.105216
DR(B) = 0.5+ 0.1666 *DR(A)	DR(B) = 0.684129
DR(C) = 0.5+0.5DR(B) +0.3333DR(A)	DR(C) = 1.2104

As in the previous case, a repetitive method has been applied to estimate the rank. In implementation shown in figure 11 only the ranking method has been presented in which the primary rank for each of the datasets has been assumed to have the value of 2 and *d* to have the value of 0.5. In this test after 60 repetitions, dataset ranks reached the accuracy of more than 0.0000001.



```
Algorithm 1 $DSRank(LinkMatrix, NumNode, \alpha, NumAttr)$
Input: $LinkMatrix$ ,$NumNode$, $\alpha$, $NumAttr$
OutPut: $rank$

for $(i = 0; i \leq NumNode; i++)$ do
  $rank[i] \leftarrow 1$
end for
for $(i = 0; i \leq NumNode; i++)$ do
  for $(j = 0; j \leq NumNode; j++)$ do
    for $(t = 0; t \leq NumAttr; t++)$ do
      $numAllLink[i][t] \leftarrow numAllLink[i][t] + linkMatrix[i][j][t]$
    end for
  end for
end for
for $(iteration = 1; iteration \leq NumIteration; iteration++)$ do
  $rankPrevIteration \leftarrow New\ \ float[NumNode]$
  for $(p = 0; p \leq NumNode; p++)$ do
    $rankPrevIteration[p] \leftarrow rank[p]$
  end for
  for $(i = 0; i \leq NumNode; i++)$ do
    $rank[i] \leftarrow 1 - alpha$
    for $(j = 0; j \leq NumNode; j++)$ do
      for $(t = 0; t \leq NumAttr; t++)$ do
        if $(linkMatrix[j][i][t] \neq 0)$ then
          $rank[i] \leftarrow rank[i] + alpha * (rankPrevIteration[j] * linkMatrix[j][i][t] * \alpha[t] / numAllLink[j][i][t])$
        end if
      end for
    end for
  end for
end for
```

Fig. 11: The Java script program for dataset ranking with weighted links

Table 6 presents the statistical results of the ranks pertaining to a number of datasets. In this case, the ranking of most of the datasets has increased as the size of the data or the number of triples has increased. In Table 6, the same datasets have been chosen and their ranks have been presented. Figure 12 shows the results in table 4 in terms of graphs. As it can be seen, the rank for *dbpedia.org* has increased as the number of triples has grown compared to the case of equal weight citation attributes. The rank for *uniprot.org* has decreased as the number of triples has increased. Furthermore, the rank for *dbpedia.org* has had an overall increase indicating that the links of higher weight have referred to this dataset. Also, the rank for *uniprot.org* has remained the same meaning that the type of the links referring to this dataset has been the same, both using a link with weight of 1.

Table 6. The weight of different datasets

Dataset name	Size = 4 GB	Size = 8 GB	Size = 12 GB	Size = 16 GB	Size = 20 GB
http://dblp.l3s.de	1.34	1.50	1.64	1.91	2.07
http://www.uniprot.org	0.50	0.51	0.51	0.51	0.50
http://purl.org	0.76	0.99	1.07	1.07	1.23
http://www.bibsonomy .org	0.68	0.68	0.72	0.73	0.74
http://dbpedia.org	1.42	1.49	1.84	2.13	2.13
http://dbtune.com	0.68	0.78	1.03	1.02	1.04
http://data.semanticwe b.org	0.5	0.5	0.5	0.55	0.55

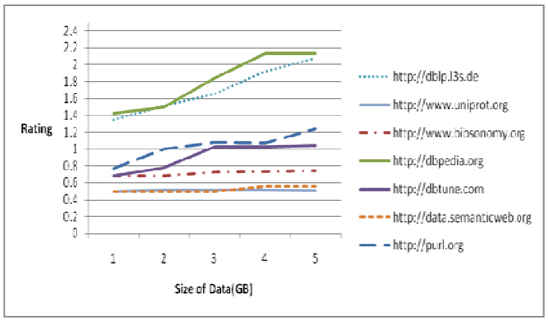


Fig. 12. Different weights for citation specifications

5. EVALUTION

To evaluate the results, the log accesses of 3 of the web servers have been used. For security reasons, log files could not be accessed therefore the results of article server (Möller et al., 2010) have been used. In this statistical article the requests to three datasets Dbpedia.org,
dbtune.org, and data.semanticweb.org have been presented. These statistics have been presented in figure 13 again. In what will follow, we will discuss these results.

dbpedia.org: The DBpedia site serves both RDF and HTML documents about its resources. For DBpedia, we had access to server log dating from 30/06/2009 - 25/10/2009 (i.e., 118 days). Number of request to this dataset for a day can be calculated from dividing number of request in 118 day by 118. It leads to 739011 request per day is made from DBpedia web server (Möller et al., 2010).

DBTune: DBTune is a meta-site which hosts different (currently 10) sub-datasets of linked data for a number of musicrelated non-LOD datasets, such as MusicBrainz, MySpace or Jamendo. We had access to log file in about two month which nearly has 7 million requests in two month. It can be seen that 122411 requests per day are made from DBTune (Möller et al., 2010).

Dog Food: The smallest dataset in our analysis in terms of RDF triples (80,000 RDF) is served through the Semantic Web Conference metadata site (SWC or \Dog Food"). We had 14117 requests per day for this data source (Möller et al., 2010).

Table 7. The number of requests to the web servers

DataSet Name	# triples	# days	Total # hits	**# hits per a day**
Dog Food	79,175	597	8,427,967	**14117.2**
DBpedia	109,750,000	118	87,203,310	**739011.1**
DBTune	74,209,000	61	7,467,125	**122411.9**

As it can be seen in table 7, the number of the requests has been normalized in column 5. The average number of the requests has been 14117 for Dog Food dataset, 739011 for DBPedia, and 22411 for DBTune. DBpedia has the largest requests followed by DBTune and then by Dog Food. The results of analyzing the available links in figures 9 and 12 indicate the same order as well.

6. CONCLUSIONS AND FURTHER WORKS

Ranking datasets in web of data has numerous applications. In search engines, the rank of any of the datasets can be generalized to its defined entities. When publishing an organization's datasets, higher rank datasets can be chosen and links can be made to their entities. In the previous studies of dataset ranking, the type and the number of links had not been taken into account.

In this article, a method of ranking datasets was presented based on the number of input links. In the first case, the weight of the citation attributes is taken to be of the same value, PageRank formula has been expanded and DSRank is brought forward so that dataset ranks can be estimated. In the second case, the type of the link to a dataset is accounted for and a specific weight is given to each of the citation attributes. Similarly, here PageRank formula is expanded and accordingly the DSRank have been estimated.

The next step includes using the proposed algorithms in a search engine so that we can precisely appreciate their efficiency. Moreover, we intend to determine the weights for each of the links which were formerly selected randomly. Ultimately, we intend to rank the entities so that each of the entities in a dataset is assigned a specific rank. In this ranking given for the entities, the rank of the data set will certainly be influential.

REFERENCES

[1] B., Adida, M., Biberick, "**RDFa Primer**", *Available At http://www.w3.org/TR/xhtml-rdf*, Accessed 20 November 2010.

[2] K., Alexander, R., Cyganiak, M., Hausenblas, J., Zhao, "**Describing Linked Datasets - On the Design and Usage of voiD, the Vocabulary of Interlinked Datasets**", *In Linked Data on the Web Workshop (LDOW09) Workshop at 18th International World Wide Web Conference*, Spain, 2009.

[3] S., Auer, C., Bizer, G., Kobilarov, J., Lehmann, and Z., Ives, "**DBPedia: a nucleus for a web of open data**", *In 6th International conference of Semantic Web*, Busan, Korea, 2009.

[4] R., BaezaYates, E., Davis, "**Web Page Ranking using Link Attributes**", *In Conference of World Wide Web*, New York, USA pp. 328-329, 2004.

[5] R., Baeza-Yates, B., Ribeiro-Neto, "**Modern Information Retrieval**", Addison Wesley Longman, 2001.

[6] Y., Bao, Z., Ming, and Y., Shang, "**Some Recent Results on Ranking Webpages and Websites**", *Technical Report*, Chinese Academy of Sciences, 2007.

[7] T., Berners-Lee, Available At "**http://www.w3.org/ DesignIssues/LinkedData.html**", Accessed 20 November 2010.

[8] T., Berners-Lee, R., Fielding, and L., Masinter, "**Uniform Resource Identifier: Generic Syntax. Request for Comments: 3986**", *IETF Network Working Group*, 2005.

[9] C., Bizer, T., Heath, D., Ayers, and Y., Raimond, "**Interlinking Open Data on the Web**", *In Proceedings Poster Track*, ESWC, Innsbruck, Austria, 2007.

[10] C., Bizer, T., Heath, and T., Berners-Lee, "**Linked Data-The Story So Far**", *International Journal on Semantic Web and Information Systems (IJSWIS)*, Special Issue on Linked Data, Vol. 5, No. 3, pp. 1-22, 2009.

[11] C., Bizer, J., Lehmann, G., Kobilarov, S., Auer, C., Becker, R., Cyganiak, and S., Hellmann, "**DBpedia - A crystallization point for theWeb of Data**", *Journal of Web Semantics: Science*, Services and Agents on theWorld Wide Web, 2009

[12] C., Bizer, T., Heath, and T., Berners-Lee, "**Linked Data: Principles and State of the Art**", in 17th *International World Wide Web Conference*, in Beijing, China, 2008.

[13] D., Brickley, R. V., Guha, "**RDF Vocabulary Description Language 1.0: RDF Schema**", Available At http://www.w3.org/TR/rdf-schema/, Accessed 20 November 2010, 2004.

[14] S., Brin, L., Page, "**The anatomy of a large-scale hypertextual Web search engine**", In 7th *Conference of World Wide Web*, Australi, 1998.

[15] S., Chakrabarti, B., Dom, S., Kumar, P., Raghavan, S., Rajagopalan, A., Tomkins, D., Gibson, and J., Kleinberg, "**Mining the Web's link structure**", In *Journal of IEEE Computer*, Vol. 32, No. 8, pp. 60-67, 1999.

[16] D., Cohn, H., Chang, and M., Kaufmann, "**Learning to probabilistically identify authoritative documents**", in 17 *International Conference on Machine Learning*, San Francisco pp. 167–174, 2000.

[17] R., Delbru, N., Toupikov, M., Catasta, G., Tummarello, and S., Decker, "**Hierarchical Link Analysis for Ranking Web**", In *7th Extended Semantic Web Conference*, Greece, 2010.

[18] C., Ding, H., Zha, X., He, P., Husbands, and H., Simon, "**Link analysis: hubs and authorities on the world wide web**", *SIAM Review*, Vol. 46, No. 2, pp.256-268, 2004.

[19] W., Fan, M., Gordon, and P., Pathak, "**A generic ranking function discovery framework by genetic programming for information retrieval**", In *Journal of Information Processing And Management*, Vol. 40 , No. 4, pp. 587–602, 2004.

[20] A., Harth, S., Decker, "**Optimized Index Structures for Querying RDF from the Web**", In *Proceedings of the Third Latin American Web Congress*, Buenos Aires, pp. 71-80, 2005.

[21] T., Heath, "**An Introduction to Linked Data**", *Available At http://tomheath.com/slides/2009-02-austin-linkeddata-tutorial.pdf*, Austin, Texas, (Accessed 20 November 2010), 2009.

[22] A., Hogan, A., Harth, and S., Decker, "**ReConRank: A Scalable Ranking Method for Semantic Web Data with Context**", In 2nd *Workshop on Scalable Semantic Web Knowledge Base Systems*, 2008.

[23] J. M., Kleinberg, "**Authoritative sources in a hyperlinked Environment**", *Journal of the ACM*, Vol. 46, No. 5, pp. 604-632, 1999.

[24] G., Miller, WordNet. Available At http://wordnet.princeton.edu (Accessed 20 November 2010).

[25] K., Möller, M., Hausenblas, R., Cyganiak, S., Handschuh, and G., Grimnes, "**Learning from Linked Open Data Usage: Patterns & Metrics**", In *Proceedings of the Web Science Conference*, Raleigh, NC, April, 2010.

[26] L., Page, S., Brin, R., Motwani, and T., Winograd, "**The Page Rank citation ranking: Bringing order to the web**". *Stanford Digital Libraries*, Technical Report, 1999.

[27] J., Pound, P., Mika, and H., Zaragoza, "**Ad-hoc Object Ranking in the Web Of Data**", In *International Conference of World Wide Web*, ACM, Raleigh, North Carolina, 2010.

[28] C., Ridings, M., Shishigin, "**Pagerank uncovered**", *White Paper, Available At: http://www.voelspriet2.nl/PageRank.pdf* (Accessed 20 November 2010), 2002.

[29] J., Shinavier, "**The State of The art in Linked Data**", *Literature Review, Avalable At www.slideshare.net/joshsh/the-state-of-the-art-in-linked-data* (Accessed 20 November 2010), 2009.

[30] T. Çelik. H., Calendar, "**Microformat specification**", *Available at http://microformats.org/* (Accessed 20 November 2010), 2008.

[31] N., Toupikov, J., Umbrich, "**DING! Dataset Ranking using Formal Descriptions**", In *Linked Data on the Web Workshop (LDOW09) Workshop at 18th International World Wide Web Conference*, Spain, 2009.

[32] G., You, S., Hwang, "**Search structures and algorithms for personalized ranking**", In *Journal of Information Science*, Vol. 178, No. 20, pp. 3925-3942, 2008.

[33] Xing, W., Ghorbani, A.(2004), "Weighted PageRank Algorithm", In Proceedings of the 2nd Annual Conference on Communication Networks and Services Research, pp. 305 – 314.