Ranking Universities Using Linked Open Data

Rouzbeh Meymandpour
School of Information Technologies
The University of Sydney
Sydney, NSW 2006
Australia
rouzbeh@it.usyd.edu.au

Joseph G. Davis
School of Information Technologies
The University of Sydney
Sydney, NSW 2006
Australia
joseph.davis@sydney.edu.au

ABSTRACT

Ranking of universities represents a complex endeavor which involves gathering, weighting, and analyzing diverse data. Emerging semantic technologies enable the Web of Data, a giant graph of interconnected information resources, also known as Linked Data. A recent community effort, Linking Open Data project, offers the possibility of accessing a large number of semantically described and linked concepts in various domains. In this paper, we propose a novel approach to take advantage of this structured data in the domain of universities to develop proxy measures of their relative standing for ranking purposes. Derived from information theory, our approach of computing the Information Content for universities and ranking them based on these scores achieved results comparable to the international ranking systems such as Shanghai Jiao Tong University, Times Higher Education, and QS. The metric we developed can also be used for innovative semantic applications in a range of domains for entity ranking, information filtering, and multi-faceted browsing.

Categories and Subject Descriptors

H.1.1 [**Models and Principles**]: Systems and Information Theory – *information theory, value of information*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering, retrieval models*; I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods – *semantic networks*; J.1 [**Computer Applications**]: Administrative Data Processing – *education*;

General Terms

Measurement

Keywords

Linked Data, Semantic Web, Entity Ranking, Web of Data, University Ranking, Linking Open Data, Partitioned Information Content, Informativeness Measurement.

1. INTRODUCTION

Universities and other academic institutions increasingly see their presence, visibility, and footprint on the World Wide Web as central to their reputation and international standing. This extends beyond the provision of information regarding researchers, publications, and prizes to include means for scholarly communication. Increasingly, the academic web is evolving into more than a vehicle for communicating scientific and cultural achievements; information content on the universities is seen more and more as a reflection of the overall organization and performance of the universities [1]. However, traditional bibliometric methods and publication and citation analyses have not scaled up to the challenges of multidimensional assessments that are required for comprehensive university ranking systems. We argue that Linked Open Data (LOD) opens up the possibility for such a project and demonstrate how the computation of Information Content based on a number of critical semantic information that signify individual university performance, can be carried out. The proposed metric measures the quality and value of these semantics to identify the relative position of universities world-wide. A preliminary evaluation by correlating them with existing, well-established rankings shows that the results are comparable, with the additional advantages of high reproducibility and the low cost of data collection.

2. UNIVERSITY-RELATED LINKED OPEN DATA

Emerging Semantic Web technologies extend the traditional Web of Documents and enable the Web of Data, also known as Linked Data. The graph structure of Linked Data consists of information resources described and connected by semantic relations. This new data model not only provides access to a large amount of structured data sources but also enables machines and software agents to automatically analyze this semantic knowledge. Thanks to the Linking Open Data community project, datasets in a wide range of domains, from encyclopedic knowledge bases to scientific data sources, are now semantically described and connected to each other [2]. The Linking Open Data cloud[1] currently provides access to more than 295 datasets in various areas such as Media, Geography, Publications, Government, and Life Sciences. However, there is a lack of specialized university-related Linked Data.

The primary source of semantic data related to universities is DBpedia[2] [3], the most essential part of the Linked Open Data. Its main aim is to provide a structured representation of Wikipedia according to the LOD guidelines. Some of the key relations describing universities in DBpedia include ‘dbo:almaMater’[3] and ‘dbo:education’ that provide information about university graduates. Several relations link universities to their members, such as ‘dbo:training’, ‘dbo:occupation’, and ‘dbo:employer’. Particular specifications are also provided, such as the location of universities as well as number of students, staff, and faculty members. Some of these relations are listed in Table 1.

LDOW2013, May 14, 2013, Rio de Janeiro, Brazil.

[1] http://lod-cloud.net/

[2] http://dbpedia.org/

[3] ‘dbo:’ is the prefix for http://dbpedia.org/ontology/

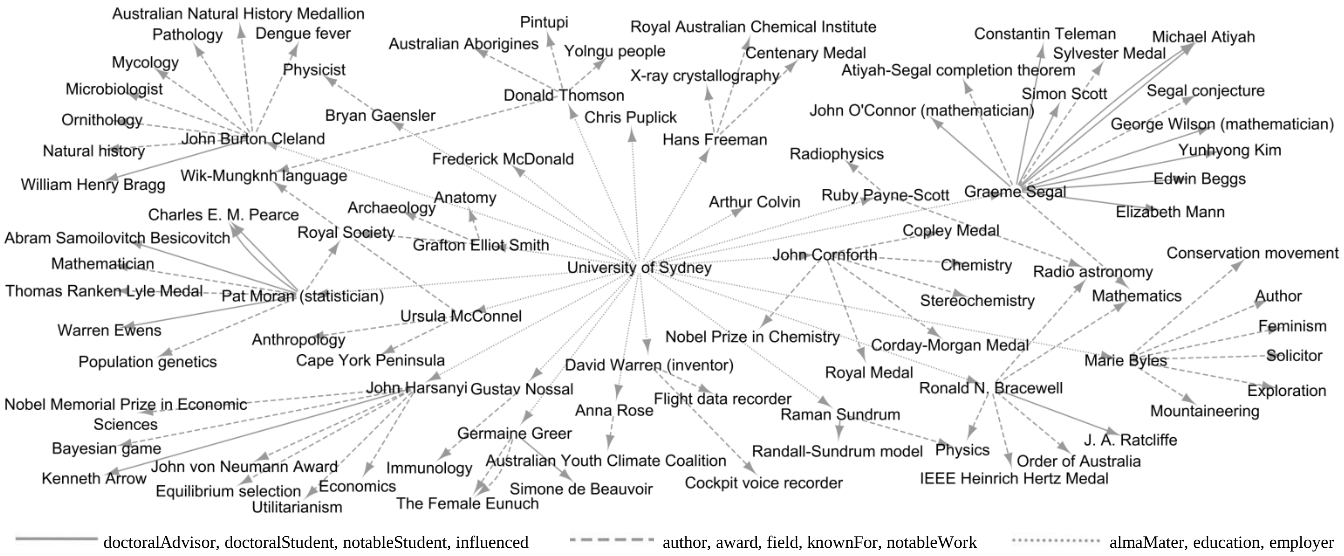


Figure 1. A small portion of Linked Open Data graph regarding The University of Sydney

Table 1. Facts provided by DBpedia related to universities[a]

dbo:affiliation (in)	dbo:numberOfPostgraduateStudents (out)
dbo:almaMater (in)	dbo:numberOfStudents (out)
dbo:campus (out)	dbo:numberOfUndergraduateStudents (out)
dbo:chancellor (out)	dbo:occupation (in)
dbo:city (out)	dbo:president (out)
dbo:college (in)	dbo:publisher (in)
dbo:dean (out)	dbo:staff (out)
dbo:education (in)	dbo:team (in)
dbo:employer (in)	dbo:training (in)
dbo:facultySize (out)	dbo:viceChancellor (out)
dbo:head (out)	

[a] The direction of each relation is expressed inside parentheses (in/out).

In addition to the relations characterizing universities, DBpedia offers massive amount of valuable semantics on university professors and researchers. These nodes are mainly linked through ‘dbo:almaMater’, ‘dbo:education’, ‘dbo:employer’, and ‘dbo:training’ relations to universities. Details available regarding university alumni and faculty members include their awards and prizes, doctoral students, notable work, and other key contributions (see Figure 1 and Table 2).

This semantic knowledge extracted from Linked Data can be employed as indicators for estimating the relative standing of world universities. By extracting and analyzing these facts, our ranking methodology is designed to rate and rank universities based on the Information Content of the semantic data.

Table 2. Facts provided by DBpedia related to alumni and university researchers

dbo:author (out)	dbo:field (out)
dbo:award (out)	dbo:influenced (in/out)
dbo:designer (out)	dbo:keyPerson (in)
dbo:developer (out)	dbo:knownFor (out)
dbo:doctoralAdvisor (in/out)	dbo:notableStudent (in/out)
dbo:doctoralStudent (in/out)	dbo:notableWork (out)
dbo:foundedBy (in)	

3. RANKING METHODOLOGY

Ranking systems consider a variety of indicators in their methodologies. In the case of university rankings these indicators could be the excellence of publications and number of citations, Nobel and other prizes won by staff, proportion of international staff and students, the faculty to student ratio, etc. In this section, we describe how the public semantic data available through Linked Open Data can be exploited to develop proxy measures for university reputation and standing.

Drawn from information theory, we propose a novel metric to compute the informativeness of semantics (resources and relations) that signify the universities in Linked Data. We proceed to develop and experiment with a ranking metric which is based on the aggregated Information Content of each of the universities.

3.1 Information Content Measurement

The notion of informativeness can be described as the value of information associated with a given entity. In information theory, Information Content (IC), also referred to as Entropy or Self-Information, is the amount of bits required to reconstruct the transmitted information source [4-6]. Based on probability theory, Information Content is computed as a measure of generated amount of surprise [7]:

$$IC(a) = -\log(\pi(a)) \qquad (1)$$

such that $\pi(a)$ is the probability of appearance of the term or concept a in its context. The logarithm in this equation is usually calculated to the base 2 and therefore, the unit of Information Content is denoted by bits.

Based on this definition, Information Content of an entity has a negative relation with its probability. More common terms in a given corpus with higher chance of occurrence cause less surprise and accordingly, carry less information, whereas infrequent ones are more informative. The concept of Information Content can be used to rank each entity, term, or alphabet in the corpus.

3.2 Partitioned Information Content

In order to extend the idea of Information Content to measure the informativeness of resources in Linked Data and to present the IC-based ranking metric, we first describe the Web of Data as a semantic network of resources connected together using a wide range of relations:

Definition 1 (Linked Data): Linked Data is a labeled directed graph (LD), defined as $\langle R, L\rangle$, where $R = \{r_1, r_2, \ldots, r_{|R|}\}$ is the set of resources, and $L = \{l_1, l_2, \ldots, l_{|L|}\}$ is the set of links, in which l_i is the relation, defined as $\langle r_1, l_i, r_2\rangle$, connecting resource r_1 to resource r_2.

Based on this definition, each resource in the Linked Data graph can be described using its incoming and outgoing edges:

Definition 2 (Resources in Linked Data): A resource such as $r \in R$ in a Linked Data (LD) is defined as its features set F_r:

$$\begin{aligned} \mathrm{F_r} = &\{\forall\langle l_i, r_i, Out\rangle, l_i \in L, r_i \in R \mid \exists r_i \in R, \exists\langle r, l_i, r_i\rangle \in LD\} \\ &\cup \{\forall\langle l_i, r_i, In\rangle, l_i \in L, r_i \in R \mid \exists r_i \in R, \exists\langle r_i, l_i, r\rangle \in LD\} \end{aligned} \quad (2)$$

In this definition, the semantics of relations, including its type, target, and direction, are considered to represent the resource and its features. For instance, *(almaMater, University of Sydney, In)*, *(award, Nobel Memorial Prize in Economic Sciences, Out)*, and *(knownFor, Bayesian game, Out)* are some features of *John Harsanyi* in Figure 1.

By drawing on the formal explanation of the Web of Data and the theoretic foundations of Information Content measurement, we present the Partitioned Information Content (PIC) to assess the informativeness of resources in Linked Data [8]:

Definition 3 (Partitioned Information Content for Linked Data): Information Content of a resource $r \epsilon R$ in Linked Data, represented as its set of features $F_r = \{f_1, f_2, \ldots, f_{|F_r|}\}$, is defined based on the amount of surprise evoked by its features:

$$PIC(F_r) = -\log\big(\pi(F_r)\big) \quad (3)$$

$$= -\log\Big(\pi(f_1)\,\pi(f_2)\cdots\pi\big(f_{|F_r|}\big)\Big) \quad (4)$$

Thus, we have

$$PIC(F_r) = \sum_{\forall f_i \in F_r} IC(\mathrm{f}) \quad (5)$$

where IC of each feature is calculated by their probability in Linked Data:

$$IC_{\forall f_i \in F_r}(f_i) = -\log\left(\frac{\varphi(f_i)}{N}\right) \quad (6)$$

such that $\varphi(f_i)$ is the frequency (number of occurrence) of the feature a_i and N is the frequency of the most popular feature in a given Linked Dataset.

The PIC of a resource is computed based on the Information Content conveyed by its features. A key characteristic of the metric is that it places more emphasis on distinctiveness of the features rather that their popularity. Hence, resources with more unique and valuable features acquire more PIC and therefore, rank higher. This metric also automatically eliminates invaluable relations that carry less information than others. For instance, broad relations such as *(University of Sydney, is member of, Universities in Australia)* have less impact on the ranking than more specific ones such as *(University of Sydney, is member of, The Group of Eight)*.

3.3 PIC-Based Ranking Metric

The informativeness of each resource is a measure of the quality of related semantics available in Linked Data. Informativeness measurement using the Partitioned Information Content-based metric (Equation (5)) can be used to rank resources. We devised the metric to adjust the influence of different relations on the ranking score by assigning weightings to each link:

$$WPIC(F_r) = \sum_{\forall f_i \in F_r} w_i\, IC(f_i) \quad (7)$$

It can also be extended to include further information in the ranking computation. Adjacent nodes linked to the main entities through an intermediate node can also be employed to express them:

$$WPIC(F_r)_2 = \sum_{\forall f_i \in F_r} w_i \left[IC(f_i) + \sum_{f_j \in F_{f_i}} w_j IC(f_j) \right] \quad (8)$$

such that F_r is the directly connected features set and F_{f_i} is the set of features linked to the resource r through $f_i \in F_r$.

It incorporates the PIC of distant neighbors into the ranking and adjusts their impact based on the types of links that connect them to the main resource. We can recursively extend Equation (8) to obtain more features in deeper layers:

$$WPIC(F_r)_{\mathrm{k}} = WPIC(F_r) + \sum_{\forall f_i \in F_r} w_i\, WPIC\big(F_{f_i}\big)_{\mathrm{k}-1} \quad (9)$$

$$\mathrm{k} > 1$$

4. EVALUATION

Having described the ranking methodology and its theoretical basis, we present the evaluation of the proposed metric by comparing our results with existing university rankings.

4.1 Experimental Context

The main Linked Dataset employed in our experiments was DBpedia 3.8, released on August 2012. We downloaded its English dump files for a faster local processing. These datasets were loaded into a Java-based RDF store, Jena TDB[4], and processed using Jena API[5].

We evaluated the accuracy of our Linked Data-based ranking approach according to its similarity with well-known world university ranking systems. We compared our list with the 2012-13 rankings provided by Shanghai Jiao Tong University (SJTU)[6], QS World University Rankings (QS)[7], and Times Higher Education (THE)[8].

In order to achieve some degree of control and to avoid unforeseen errors and noise in the DBpedia dataset, we limited the number of relations to those that are closely related to the university performance. We also double-checked the rdf:type[9] of neighbors connected to the universities. For example, all nodes connected through 'dbo:almaMater' relation have to be a 'dbo:Person'. It guarantees that incorrect links do not influence the accuracy of results.

We employed Equation (9), the weighted PIC-based metric, for up to two levels of depth. Table 3 lists the relations employed to perform the ranking as well as their relative weightings which

[4] http://jena.apache.org/documentation/tdb/

[5] http://jena.apache.org/

[6] http://www.shanghairanking.com/

[7] http://www.topuniversities.com/

[8] http://www.timeshighereducation.co.uk/

[9] 'rdf:' is the prefix for http://www.w3.org/1999/02/22-rdf-syntax-ns#

were assigned by an expert who has served as a reviewer for THE and QS rankings.

In order to provide more insight into the effect of considering various types of relations in the ranking, we also evaluate a basic variation of PIC (Equation (5)) that only takes immediate neighbors into account and does not incorporate the weights into the ranking score. The PIC (Basic) metric considers all kinds of relations in the first level, without any restriction.

Table 3. Employed relations and assigned weightings

University (First Depth)			
dbo:almaMater	1	dbo:president	1
dbo:education	1	dbo:chancellor	1
dbo:team	1	dbo:dean	1
dbo:training	1	dbo:viceChancellor	1
dbo:occupation	1	dbo:head	1
dbo:employer	1	dbo:publisher	1
Person (Second Depth)			
dbo:award	4	dbo:keyPerson	2
dbo:knownFor	2	dbo:foundedBy	2
dbo:doctoralAdvisor	1	dbo:doctoralStudent	1
dbo:influenced	2	dbo:notableWork	2
dbo:notableStudent	2	dbo:designer	2
dbo:author	2	dbo:developer	2
Publication (Second Depth)			
dbo:academicDiscipline	1	dbo:author	1
dbo:editor	1		

4.2 Evaluation Metrics

We conducted two experimental evaluations. In the first experiment, we matched universities in each individual ranking with their corresponding DBpedia URI by computing the Levenshtein distance [9] of university names in the ranking lists and 'rdfs:label'[10] of universities in DBpedia. We then manually double-checked the results. Finally, the Pearson Correlation and the Spearman Rank Correlation coefficients were computed to identify the correspondence of aggregate scores computed by our Linked Data-based approach with the total scores provided by the other ranking systems.

The second experiment was carried out to measure the similarity between our ranked list of universities and other rankings. To obtain the list of universities, we retrieved the top 493 universities in the QS list. By adding missing universities from the top 100 items of other lists, we ended up with a list of 500 universities using which we performed the ranking. Finally, because ranking lists are non-conjoint, i.e. one item may be ranked in one list but not in the other, we exploited the non-conjoint ranking similarity metrics, namely Overlap (O) and Average Overlap (AO, also referred to as Average Accuracy), in order to measure the intersection of our ranked list of universities and the rankings provided by other organizations [10-12]:

Overlap:

$$O(R_1, R_2)_k = \frac{|R_{1(k)} \cap R_{2(k)}|}{k} \qquad (10)$$

Average Overlap:

$$AO(R_1, R_2)_k = \frac{1}{k}\sum_{i=1}^{k} \frac{|R_{1(k)} \cap R_{2(k)}|}{k} \qquad (11)$$

where $O(R_1, R_2)_k$ and $AO(R_1, R_2)_k$ are the overlap and average overlap between top-k items of R_1 and R_2 lists. Unlike Overlap, Average Overlap is top-weighted, i.e. the top of the ranking is more important than the rest.

4.3 Results

Table 4 shows the main criteria and their contribution to the total ranking score. This includes the total Information Content engendered by adjacent nodes connected to the universities via these relations. We can see that 'dbo:education' and especially 'dbo:almaMater' are dominant relations that generate the most Information Content of each university. For the full list of top 100 global universities ranked based on Linked Open Data, refer to the Appendix 1.

Table 4. Top 5 universities and the PIC obtained by each relation

	Harvard University	*Princeton University*	*Massachusetts Institute of Technology*	*Columbia University*	*Stanford University*
dbo:almaMater	114,387.1	68,121.6	65,404.4	48,694.0	39,707.7
dbo:education	9,745.1	2,535.4	1,682.5	10,484.6	4,652.5
dbo:employer	917.8	211.6	238.7	453.0	446.7
dbo:occupation	97.5	60.9	137.4	839.8	157.6
dbo:president	21.2				21.2
dbo:publisher	76.3	159.4	78.4	58.2	21.2
dbo:team	99.5	175.8		55.8	56.1
dbo:training	634.8	41.3	493.8	2,078.2	863.5
Total	125,979.3	71,306.0	68,035.2	62,663.6	45,926.4

Table 5 below compares the correlation between different ranking scores. The result shows a high correlation with others rankings. In terms of Pearson correlation coefficient, our ranking scores are closer to the Shanghai Jiao Tong ranking (0.848, $p < 0.01$). The proposed metric also demonstrated high correlations with other lists, with more than 0.67 and 0.68 of correspondence with Time Higher Education and QS rankings, respectively. Based on Spearman rank correlation of nearly 0.64 ($p < 0.01$), our LOD-based approach also shows the strongest association with the QS ranking.

The improvements of PIC over PIC (Basic) are also significant both for Pearson Correlation and Spearman Rank Correlation. The highest increase can be seen in the Spearman Ranking Correlation with QS rankings, jumping from 0.44 to 0.64.

Table 5. The correlation between the LOD-based rankings and others

	Pearson Correlation		*Spearman Rank Correlation*	
	PIC (Basic)	PIC	PIC (Basic)	PIC
SJTU	0.788	0.848	0.515	0.585
QS	0.553	0.680	0.439	0.643
THE	0.650	0.672	0.552	0.619

We also evaluated the similarity of our approach with the others by ranking 500 universities. Figure 2 illustrates that the degree of match between LOD-based ranks and others is higher for the top one hundred universities (50% intersection) and tends to diverge beyond that.

[10] 'rdfs:' is the prefix for http://www.w3.org/2000/01/rdf-schema#

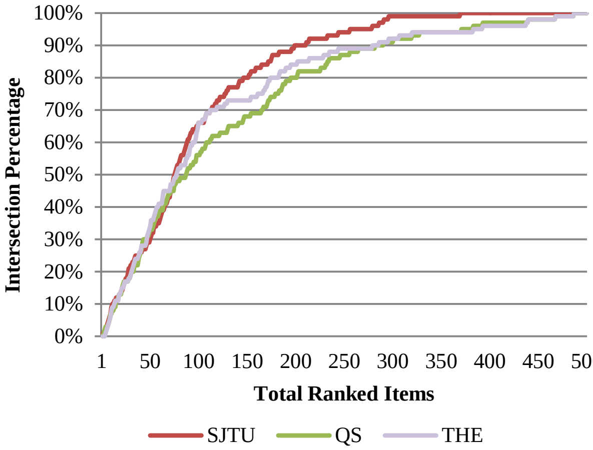


Figure 2. Intersection between LOD-based ranking and others

To measure similarity, we computed the Overlap (O) and Average Overlap (AO) for the top 100 items in each list. It is observable from the results that the Average Overlap similarity between our rankings and others is more than 62% (see Table 6). These results again show higher similarity with the Shanghai Jiao Tong rankings (67%).

These results also show 5% to 12% increase in the overall similarity score of PIC against its basic version.

Table 6. The similarity between the LOD-based rankings and others

	Overlap		*Average Overlap*	
	PIC (Basic)	PIC	PIC (Basic)	PIC
SJTU	0.610	0.660	0.616	0.669
QS	0.510	0.560	0.511	0.628
THE	0.600	0.660	0.573	0.638

In order to provide better understanding of the results, the Average Overlap similarity between all pairs of ranking systems were evaluated. Table 7 reveals that all the rankings, including ours, are 63% to 73% similar to each other in the list of top 100 universities.

Table 7. Pairwise Average Overlap similarity of rankings

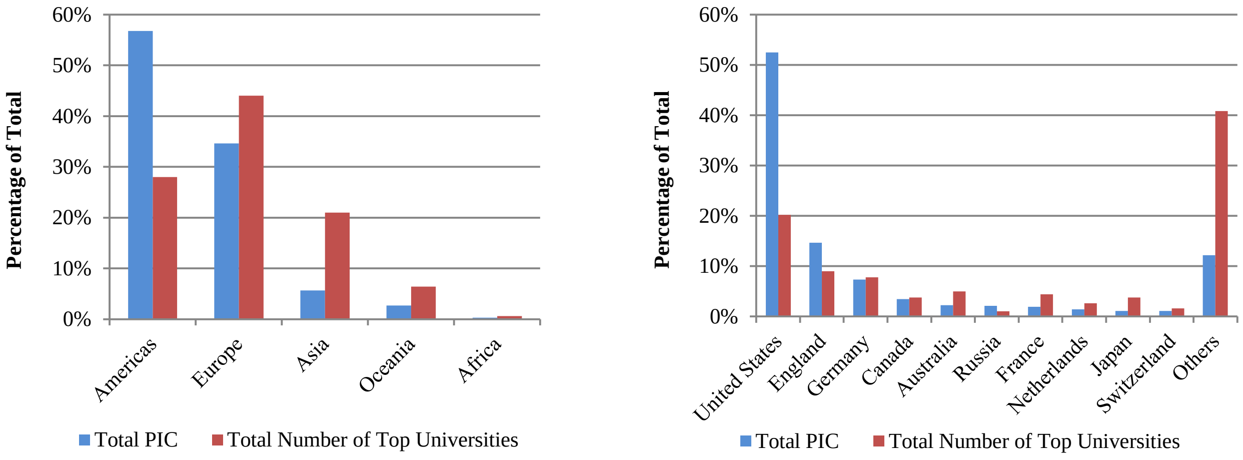
	PIC	*SJTU*	*QS*	*THE*
PIC	1	0.669	0.628	0.638
SJTU	0.669	1	0.627	0.728
QS	0.628	0.627	1	0.721
THE	0.638	0.728	0.721	1

We also computed the distribution of university-related semantic data across countries and continents (see Figure 3). The results show that although there are more European universities in the top 500 list, Information Content generated by universities in Americas is noticeably higher. It can be observed that the U.S. with more than 100 entries is not only the country with the highest number of high-ranked universities, its share of semantic content is disproportionately higher compared to other countries and even Europe, Asia, and Australasia. This can be viewed as a particular manifestation of digital divide.

5. DISCUSSION

As presented in the previous section, the results obtained by applying the proposed Linked Open Data-based method are comparable to other ranking systems, with the extra benefit of its low-cost data acquisition methodology.

It can be observed from Table 7 that the PIC-based ranking is closer to the Academic Ranking of World Universities provided by Shanghai Jiao Tong University. This can be explained by looking at the ranking methodologies employed in different systems. While SJTU adopts an approach based on objective data, others use a combination of objective and subjective indicators. The SJTU ranking aims to measure the quality of research and researchers by including Nobel Prizes, Fields Medal, publications, and citations. In contrast, QS gives a weighting of 40% to a survey conducted among academics around the world aiming to measure the academic reputation. Likewise, THE ranking methodology assigns up to 30% weighting on the quality of teaching and learning environment measured by several indicators, including a reputation survey. As we adopted what can be considered to be relatively objective indicators in the ranking methodology, the similarity of results to the Shanghai Jiao Tong University's rankings is not surprising.



(a)

(b)

Figure 3. Distribution of Information Content regarding top 500 universities across continents (a) and countries (b)

The average of 8.5 percent difference between PIC (Basic) and PIC rankings indicates that at least eight universities that are not in the top 100 list of other systems appeared in the PIC (Basic) rankings. This difference becomes even more pronounced after that. This indicates that including additional details in deeper levels improves the overall accuracy of the metric. However, considering the fact that PIC (Basic) ranked universities without any restriction on or weights attached to link types and by including only the first level neighbors, its ranking performance with Average Overlap of 51 to 62 percent is encouraging.

It can also be seen that the ranking of the first one hundred universities are more than 60% similar across all the rankings. This suggests that having 60% to 70% Average Overlap with other rankings is reasonable for our metric. Moreover, our intention is not to have very similar results to others. Rather, this ranking is complement to other systems and can be used in conjunction with them, measuring the presence of universities on the Semantic Web.

Further experiments revealed that at least 52% of Information Content available in the Linked Open Data regarding top universities belong to American institutions. This reveals a significant divide between the visibility of world universities across various regions within the LOD. This may encourage them to keep track of their presence on the Web and routinely publish more valuable information on the World Wide Web. Academic institutions can have policies to contribute to Wikipedia and regularly publish articles about their scholars and their achievements.

6. PROBLEMS AND LIMITATIONS

One of the limitations of this study is the lack of specific Linked Open Data related to universities and research conducted at academic institutions. Although online research publication repositories such as ACM, IEEE, and DBLP are available through Linked Open Data cloud[11], publications are not fully described. For example, institution of the authors is not covered by the ontologies. Hence, it is not accurately possible to incorporate these datasets into the ranking procedure. Moreover, some data quality problems, such as low levels of completeness and consistency, have been observed in Linked Open Data. The key difficulty is that not all specifications of universities and details about researchers are published on and covered by the ontology of DBpedia (or Wikipedia).

Other data quality issues of Linked Open Data, such as duplicate and incorrect property-values as well as misspelling errors, have to be precisely controlled. Although the metric automatically eliminates redundant information to improve the ranking accuracy, further manual and semi-automatic processing are also required to reduce the noise.

Another problem with DBpedia arises when retrieving the list of universities. We could execute a simple query and have a list of more than 14,000 worldwide universities. However, many constituent colleges and schools are also expressed as university in DBpedia. For example, Harvard Business School which is a part of Harvard University is also described with the *(rdf:type, dbo:University)* relation. We tried to create a comprehensive list of all universities based on the *(rdf:type, yago:University[12])* relation but this does not cover several universities in the ranking lists. To resolve the problem, we used the limited QS ranking list of universities and found their matches in the DBpedia.

7. RELATED WORK

Although extensive research has been carried out on the ranking problem in general and on the Web in particular, a few researchers have attempted to exploit Web content for ranking universities. The Webometric proposed by Aguillo et al. [1, 13] provides ranking of universities based on four indicators including number of Web pages, rich files (pdf, ps, doc, etc.), external links, and articles on Google Scholar repository related to the university being ranked. In contrast, our method of ranking is based on the quality of semantics regarding the university researchers, graduates, and their publications. However, the authors found the same digital divide between American universities and European institutions on the Web. In Aguillo et al. [14], they also discovered high similarities between various university ranking systems, especially for the top 100 list.

A number of ranking techniques have been proposed based on the link structure of networks, such as PageRank [15], SimRank [16], and HITS [17]. They are widely used for filtering results in search engines and ranking Web pages. The main drawback of applicability of these methods on Linked Open Data graph is that they do not take into account the semantics of the links, i.e. diverse types of links. In the World Wide Web structure, hyperlink is the only type of link connecting Web pages, while in Linked Data various kinds of links are used to express the relations between resources.

Several studies have attempted to incorporate the semantics of links into the ranking metric [18-23]. ObjectRank [18] enabled ranking in directed labeled graphs by extending PageRank. Bamba and Mukherjea [19] applied PageRank for organizing the results of Semantic Web queries. Franz et al. proposed the TripleRank algorithm [22], which is a generalization of the HITS method in the context of Linked Data. It was evaluated for faceted browsing and filtering semantic relations for better Linked Data exploration experience. In addition to centrality-based features extracted using PageRank and HITS algorithms, a recent work by Dali et al. [24] also contributed some statistics of RDF graph into the ranking method. By obtaining features such as number of subjects and objects as well as the diversity of incoming and outgoing relations reachable at different distances from the main node, their main aim was to rank results of RDF entity search. The main contribution of our metric is that it is completely based on the semantics of links and the statistical characteristics of Linked Data, whereas the importance of nodes in these approaches is computed according to the link structure of the Web of Data.

In Meymandpour and Davis [8], the authors proposed the notion of informativeness measurement for Linked Data. They experimented with the metric in several applications, such as entity and property ranking as well as faceted browsing and Linked Data quality analysis. Another related work is a random walk model presented by Kasneci et al. [25] for extracting the most informative sub-graphs in a labeled graph. Similar to the feature frequency used in Meymandpour and Davis [8], in this approach, the rank score of nodes is computed using frequency-based edge weights.

8. CONCLUSION AND FUTURE WORK

We have presented an innovative ranking metric that takes into account the informativeness of entities on the Web of Data. By computing the quality of facts available publicly in Linked Open

[11] See http://lod-cloud.net/

[12] The URI is http://dbpedia.org/class/yago/University108286163

Data, we measured the relative footprint of world universities on the Web. It can also be modified and applied in other domains and in a variety of Linked Data-based applications, such as information filtering, data visualization, multi-faceted browsing, and semantic navigation.

We highlight the need for a Linked Open Data providing university- and research-related semantics. As a structured and reliable source of semantic data, it can offer significant benefits for a low-cost and accurate performance analysis of global universities. The method proposed in this paper can result in more accurate rankings with such a resource. It provides academic institutions with a useful representation of the relative quality of their footprint on the World Wide Web.

As a part of our future work, we will focus more on the accuracy of the ranking by capturing more semantics from LOD cloud and by eliminating any trace of redundancy. A panel of academicians will be asked to give weightings to the indicators and the main metric will be evaluated against its variations to examine the impact of varying weights and considering diverse kinds of semantics on the final result. We will also keep updating and releasing the rankings on an annual basis[13].

9. ACKNOWLEDGMENTS

We thank the three anonymous reviewers for their helpful comments and suggestions.

10. REFERENCES

[1] Aguillo, I. F., Ortega, J. L. and Fernández, M. 2008. Webometric Ranking of World Universities: Introduction, Methodology, and Future Developments. *Higher Education in Europe*, 33, 233-244. Routledge.

[2] Bizer, C., Heath, T. and Berners-Lee, T. 2009. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5, 3 (2009), 1-22.

[3] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R. and Ives, Z. 2007. DBpedia: A Nucleus for a Web of Open Data. *The Semantic Web*, 4825, 722-735. Springer Berlin / Heidelberg.

[4] Gray, R. M. 2009. *Entropy and Information Theory*. Springer-Verlag, New York.

[5] Edwards, S. 2008. Elements of information theory, 2nd edition. *Information Processing & Management*, 44, 1 (Jan 2008), 400-401.

[6] Shannon, C. E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(1948), 379–423, 623–656.

[7] Ross, S. M. 2002. *A First Course in Probability*. Prentice Hall.

[8] Meymandpour, R. and Davis, J. G. 2013. Linked Data Informativeness. *Web Technologies and Applications*, 7808, 629-637. Springer Berlin Heidelberg.

[9] Levenshtein, V. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10, 8 (1966), 707-710.

[10] Fagin, R., Kumar, R. and Sivakumar, D. 2003. Comparing top k lists. In *Proceedings of the Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms* (Baltimore, Maryland, 2003). Society for Industrial and Applied Mathematics.

[11] Wu, S. and Crestani, F. 2003. Methods for ranking information retrieval systems without relevance judgments. In *Proceedings of the 2003 ACM symposium on Applied computing* (Melbourne, Florida, 2003). ACM.

[12] Webber, W., Moffat, A. and Zobel, J. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28, 4 (2010), 1-38.

[13] Aguillo, I., Ortega, J., Fernández, M. and Utrilla, A. 2010. Indicators for a webometric ranking of open access repositories. *Scientometrics*, 82, 3 (2010/03/01 2010), 477-486.

[14] Aguillo, I., Bar-Ilan, J., Levene, M. and Ortega, J. 2010. Comparing university rankings. *Scientometrics*, 85, 1 (2010/10/01 2010), 243-256.

[15] Brin, S. and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the seventh international conference on World Wide Web 7* (Brisbane, Australia, 1998). Elsevier Science Publishers B. V.

[16] Jeh, G. and Widom, J. 2002. SimRank: a measure of structural-context similarity. In *Proceedings of the Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (Edmonton, Alberta, Canada, 2002). ACM.

[17] Kleinberg, J. M. 1999. Authoritative sources in a hyperlinked environment. *J. ACM*, 46, 5 (1999), 604-632.

[18] Balmin, A., Hristidis, V. and Papakonstantinou, Y. 2004. Objectrank: authority-based keyword search in databases. In *Proceedings of the Thirtieth international conference on Very large data bases - Volume 30* (Toronto, Canada, 2004). VLDB Endowment.

[19] Bamba, B. and Mukherjea, S. 2005. Utilizing Resource Importance for Ranking Semantic Web Query Results. *Semantic Web and Databases*, 3372, 185-198. Springer Berlin / Heidelberg.

[20] Ding, L., Pan, R., Finin, T., Joshi, A., Peng, Y. and Kolari, P. 2005. Finding and Ranking Knowledge on the Semantic Web. *The Semantic Web – ISWC 2005*, 3729, 156-170. Springer Berlin / Heidelberg.

[21] Hogan, A., Harth, A. and Decker, S. 2006. ReConRank: A Scalable Ranking Method for Semantic Web Data with Context. In *Proceedings of the Second International Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS 2006), in conjunction with International Semantic Web Conference (ISWC 2006)* (2006).

[22] Franz, T., Schultz, A., Sizov, S. and Staab, S. 2009. TripleRank: Ranking Semantic Web Data by Tensor Decomposition. *The Semantic Web - ISWC 2009*, 5823, 213-228. Springer Berlin / Heidelberg.

[23] Delbru, R., Toupikov, N., Catasta, M., Tummarello, G. and Decker, S. 2010. Hierarchical Link Analysis for Ranking Web Data. *The Semantic Web: Research and Applications*, 6089, 225-239. Springer Berlin / Heidelberg.

[24] Dali, L., Fortuna, B., Duc, T. T. and Mladenić, D. 2012. Query-Independent learning to rank for RDF entity search. In *Proceedings of the 9th international conference on The Semantic Web: research and applications* (Heraklion, Crete, Greece, 2012). Springer-Verlag.

[25] Kasneci, G., Elbassuoni, S. and Weikum, G. 2009. MING: mining informative entity relationship subgraphs. In *Proceedings of the 18th ACM conference on Information and knowledge management* (Hong Kong, China, 2009). ACM.

[13] Rankings are available on http://sydney.edu.au/engineering/it/~rouzbeh/university-rankings/

APPENDIX 1: LOD-based Top 100 Universities

Rank	*University*	*PIC Score*
1	Harvard University	125,979.3
2	University of Cambridge	115,418.5
3	Princeton University	71,306.0
4	Massachusetts Institute of Technology	68,035.2
5	Columbia University	62,663.6
6	University of California, Berkeley	61,787.8
7	Yale University	60,686.7
8	University of Oxford	48,677.2
9	University of Chicago	47,178.7
10	Stanford University	45,926.4
11	University of Michigan	33,817.3
12	Humboldt University of Berlin	33,404.1
13	California Institute of Technology	33,037.6
14	Moscow State University	32,053.8
15	Cornell University	31,193.8
16	University of Göttingen	28,620.9
17	University of Edinburgh	24,242.3
18	University of Pennsylvania	23,990.7
19	New York University	23,742.6
20	University of Wisconsin–Madison	22,797.0
21	University of Toronto	22,612.4
22	University of Illinois at Urbana–Champaign	21,963.7
23	University College London	21,268.0
24	University of California, Los Angeles	21,195.0
25	École Normale Supérieure	19,554.9
26	University of Vienna	19,232.3
27	Brown University	17,288.0
28	Johns Hopkins University	16,082.3
29	Ludwig Maximilian University of Munich	15,993.6
30	Northwestern University	14,446.2
31	Saint Petersburg State University	14,335.9
32	University of Minnesota	14,056.4
33	University of Florida	13,672.1
34	Imperial College London	13,442.8
35	University of Texas at Austin	12,983.5
36	University of Southern California	12,712.0
37	Hebrew University of Jerusalem	12,656.7
38	McGill University	12,594.9
39	University of Tokyo	12,400.8
40	King's College London	12,114.5
41	University of Melbourne	11,962.1
42	University of Manchester	11,961.8
43	Dartmouth College	11,839.1
44	Heidelberg University	11,701.8
45	University of Glasgow	11,472.6
46	Carnegie Mellon University	11,307.7
47	Duke University	11,229.0
48	Leiden University	11,126.9
49	University of Utah	11,008.6
50	ETH Zurich	10,680.9
51	London School of Economics	10,627.0
52	University of Leipzig	10,204.9
53	University of Sydney	9,995.6
54	University of Washington	9,519.2
55	Ohio State University	9,276.1
56	Georgetown University	9,240.5
57	University of Bonn	9,201.8
58	Uppsala University	8,901.6
59	University of Iowa	8,848.1
60	Rutgers University	8,804.4
61	Trinity College, Dublin	8,803.7
62	University of Notre Dame	8,782.6
63	Boston University	8,616.3
64	École Polytechnique Fédérale de Lausanne	8,569.5
65	George Washington University	8,536.1
66	University of Warsaw	8,315.8
67	University of Virginia	8,101.3
68	Brandeis University	7,708.0
69	University of Maryland, College Park	7,702.5
70	University College Dublin	7,599.5
71	University of British Columbia	7,503.2
72	University of Rochester	7,370.0
73	University of Birmingham	7,314.0
74	Pennsylvania State University	7,243.2
75	University of North Carolina at Chapel Hill	6,901.0
76	Florida State University	6,838.6
77	University of Pittsburgh	6,754.2
78	University of Arizona	6,693.8
79	Rice University	6,672.4
80	Tufts University	6,642.9
81	University of Oslo	6,592.4
82	Georgia Institute of Technology	6,563.9
83	Syracuse University	6,462.3
84	Berlin Institute of Technology	6,450.5
85	University of Colorado at Boulder	6,335.7
86	Stockholm University	6,331.9
87	Utrecht University	6,154.5
88	Charles University in Prague	6,117.4
89	University of Bristol	6,086.5
90	University of Manitoba	5,966.1
91	Durham University	5,865.3
92	Purdue University	5,837.7
93	University of California, Santa Cruz	5,781.0
94	Queen's University	5,756.6
95	University of Marburg	5,649.4
96	University of Kansas	5,647.6
97	University of Adelaide	5,372.0
98	Washington University in St. Louis	5,364.8
99	University of Missouri	5,357.7
100	Michigan State University	5,258.4