SCIENCE CHINA
Information Sciences

• LETTER •

October 2019, Vol. 62 209202:1–209202:3
https://doi.org/10.1007/s11432-018-9498-6

Mining the rank of universities with Wikipedia

Zongjian LI[1], Cong LI[1,2*] & Xiang LI[1,2]

[1]Adaptive Networks and Control Lab, Department of Electronic Engineering, Fudan University, Shanghai 200433, China;
[2]Research Center of Smart Networks and Systems, School of Information Science Engineering, Fudan Univeristy, Shanghai 200433, China

Received 4 May 2018/Accepted 11 June 2018/Published online 29 March 2019

Citation Li Z J, Li C, Li X. Mining the rank of universities with Wikipedia. Sci China Inf Sci, 2019, 62(10): 209202, https://doi.org/10.1007/s11432-018-9498-6

Dear editor,
Wikipedia is a free-access and web-based multilingual encyclopedia that voluntaries from all around the world can write and edit. Previous studies on Wikipedia mainly focused on its collaborative systems, i.e., edit patterns [1, 2]. Besides the relation network of editors in Wikipedia, the articles of Wikipedia form a large-scale complex network, the Wikipedia article reference network (WARN). Articles are regarded as nodes, which are connected by the URL links. Gabrilovich and Markovitch [3] applied the machine learning techniques to explore the semantic relatedness of natural language texts with the concepts derived from Wikipedia. Medelyan et al. [4] designed an algorithm for the topic indexing with Wikipedia, which contains many article synonyms.

We study whether the reputation of a Wikipedia article represents the rank of the subject in its field. We take the rank of universities as the research subjects. Previous studies on university ranking mostly employed the core-driven ranking system [5, 6], which combines disparate indicators into a single total score. Although all these ranking systems are useful to evaluate the universities from different aspects, a ranking system without subjective assessment is still expected. We propose several reputation indicators for entries by utilizing the Wikipedia, and study the correlation between the reputation indicators of Wikipedia university entries and the Quacquarelli Symonds (QS) or Times Higher Education (THE) ranking. Moreover, we use the path length method and the local vertex connectivity method to discover the relation among subjects with Wikipedia data. Our results open up the possibility to utilize the Wikipedia data as a useful tool for the discovery of inherent relation among subjects in real world.

Data fetching and subjects selection. We firstly design a web crawler program to collect the raw data from English version of Wikipedia in this work. The obtained Wikipedia data can be mapped into a WARN and recorded into a graph database (Appendix A).

We select 114 universities as the subjects, which include the overlapping universities of the top 100 in QS ranking and the top 200 in THE ranking, as well as the overlapping universities of the top 100 in THE ranking and the top 200 in QS ranking. All the data of Wikipedia, the QS ranking and the THE ranking are collected by 2015 (Appendix B).

Methodology. We introduce three types of reputation indicators, i.e., the intuitive criterion, the potential criterion and the network-based criterion in this study. The intuitive criterion can be obtained directly when you read an article, for instance, the length of an article which is counted in bytes. The potential criterion is a kind of previous record of an article, i.e., the number of revisions of the study, the number of editors who have rewritten the article, and the number of times that an article has been edited in one year. The inter-edit time distribution of articles follows a double-power-law [7]. The network-based criterion is cal-

[*] Corresponding author (email: cong_li@fudan.edu.cn)

culated on the WARN, where the articles are regarded as nodes and are directly linked by the URLs. Note that this work only adapts articles and hyperlinks in English Wikipedia. The adjacency matrix of WARN is denoted by an $N \times N$ matrix A, consisting of elements a_{ij} that are either one or zero depending on whether there is a link from article i to article j. The in-degree reputation indicator is defined as the in-degree of an article $d_{\rm in}(i) = \sum_{j=1}^{N} a_{ji} = (u^{\rm T}A)_i^{\rm T}$ and the out-degree reputation indicator is defined as the out-degree of an article $d_{\rm out}(i) = \sum_{j=1}^{N} a_{ij} = (Au)_i$, respectively.

The relation between the reputation criterions and the QS or THE ranking is studied. Results show that the intuitive reputation criterion or the potential reputation criterion of universities in an English-speaking region is much higher than that of universities in a non-English-speaking region (Appendix C). It demonstrates that the influence of language should not be ignored when we evaluate the reputation of an entry with Wikipedia data. Moreover, it is found that the QS or THE ranking is more strongly linear correlated with the in-degree $d_{\rm in}$ reputation indicator than with the out-degree $d_{\rm out}$ reputation indicator (Appendix C). Inspired by [8], the sum of the degree of two-hopcount neighbors in WARN is calculated and compared with the QS or THE ranking. We find that the WARN has a small average shortest path length between any two articles, which is an essential small-world property [9]. Hence, we further consider the sum of in-degree and out-degree of the article itself and its 1-hopcount or 2-hopcount neighbors as the network-based criterions.

The linear correlation coefficients ρ between three types of criterions and the QS or THE ranking are calculated and shown in Figure 1(a). We find that for the universities located in English-speaking regions, the in-degree or the sum of in-degree are most strongly correlated with the QS (with $\rho > 0.5$) and THE university ranking (with $\rho > 0.6$). However, for the universities in non-English-speaking regions, the in-degree or the sum of the in-degree are most different from the QS ranking, while, the number of editors is most strongly correlated with the QS (with $\rho > 0.6$) and THE university ranking (with $\rho > 0.4$). Moreover, all potential criteria of universities located in non-English-speaking regions perform more closely to the QS ranking than that of universities located in English-speaking regions.

The WARN involves all articles in Wikipedia, however, only a small part of the articles is essential and interesting for the study in a particular field. We here call the small part articles as the targeted articles. It is a challenge to extract an effective Wikipedia article reference subnetwork (EWARS) from the WARN, which contains only the targeted articles. We design two methods to generate the EWARS, and take the EWARS of the 114 universities as an example.

It is obvious that we cannot just extract the 114 university articles and the URLs among them to generate the EWARS, because the indirect relations should not be ignored. For instance, two articles that are not connected directly may have an n-length path between them. The difficulty for generating the EWARS is how to pick up the indirect relations. We propose an EWARS generator which performs as follows:

Step 1. We generate the WARN, which contains all the articles and links in English Wikipedia;

Step 2. We generate the pre-EWARS by setting two parameters, i.e., the direction parameter and the depth parameter to pick up the neighborhood of the targeted articles;

Step 3. We calculate the relation weight between any two targeted articles, and set a threshold for the relation weight to obtain the EWARS.

In Step 2, the direction parameter is "incoming" or "out-going", which represents that the links to or from the 114 university articles will be considered separately. The depth parameter is the recursive depth that is used to limit the data mining depth. A larger recursive depth will allow more indirect relations to be added to the pre-EWARS, which makes the pre-EWARS a larger network. Note that the pre-EWARS is an unweighted network by adding inverse links to the picked directed links.

In Step 3, we propose to apply the shortest path length (PL) method or the local vertex connectivity (LVC) method to calculate the relation weight between any two targeted articles (Appendix D). The local vertex connectivity ν_{ij} is defined as the smallest number of nodes to remove that makes no path between vertices i and j.

The pre-EWARS of the 114 university articles with their out-going connected articles and recursive depth 1 has 29416 vertexes and 1898348 undirected edges. The EWARS generated by PL method is a connected network, where each pair of the university articles has a path length shorter than 4. There are 1900 pairs of university articles with a shortest path length of 1, while 4385 pairs of university articles have a shortest path length of 2, and 156 pairs of university articles have a shortest path length of 3. When we set the threshold of the average shortest path length as 1, the EWARS is already a dense connected network. The result

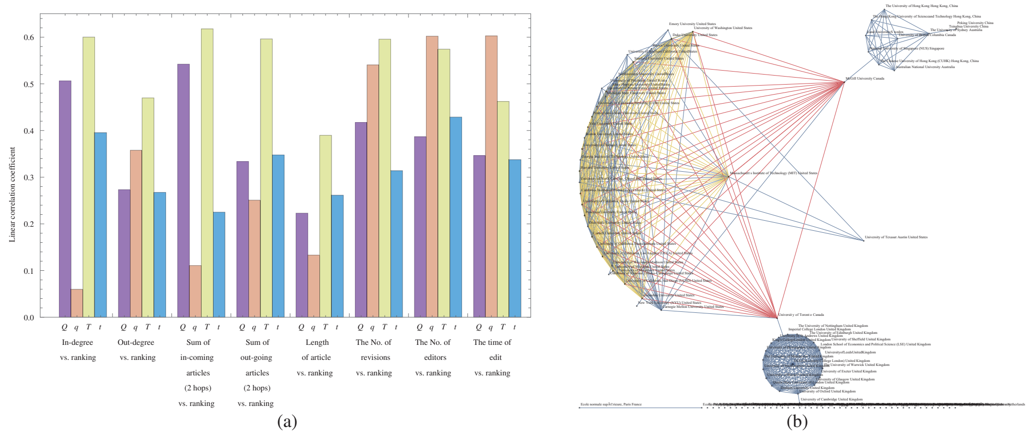

Figure 1 (Color online) (a) The linear correlation coefficient between the reputation indicators for university entries and the QS or THE university rankings; (b) the EWARS generated by utilizing the LVC method with a threshold $T = 172$.

displays that it is difficult to mine the deep relations among densely connected subjects, such as the university articles with the PL method. We further find that the direction parameter in PL method does not affect the distribution of the link weights by setting the direction parameter as in-coming.

Next, we apply the LVC method to generate the EWARS of the 114 university articles with the out-going links. Compared to the shortest path length, the local vertex connectivity has a large range of values, and approximately follows a binomial distribution. To see the detail of this EWARS, we also set a threshold to obtain an unweighted network. Edges in the EWARS with a higher weight than the threshold will be kept, otherwise disconnected. With the increasing of the threshold, the EWARS becomes more and more sparse. The EWARS with a threshold $T = 172$ is shown in Figures 1(b), where edges in the largest and the second largest cliques are colored in red and yellow, respectively. The left part consists of universities in America, two red dots in the middle are universities in Canada, the sphere at the bottom consists of universities in UK, and the right-top part is mainly occupied by universities in China, Australia and Singapore. Notice that the results for the EWARS of the 114 university articles with the in-coming links has similar results with the pervious discussion. The result demonstrates that the community property of the EWARS matches the geographic distribution of the universities. It implies that the deep relation between subjects can be discovered by mining the properties in the EWARS. Moreover, our result verifies the conclusion in [10] that the world university ranking has produced global geographies of higher education.

Acknowledgements This work was partly supported by National Natural Science Foundation of China (Grant Nos. 61603097, 71731004), and Natural Science Foundation of Shanghai (Grant No. 16ZR1446400).

Supporting information Appendixes A–D. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Pfeil U, Zaphiris P, Ang C S. Cultural differences in collaborative authoring of Wikipedia. J Comput Mediated Commun, 2006, 12: 88–113

2 Yasseri T, Sumi R, Kertész J. Circadian patterns of Wikipedia editorial activity: a demographic analysis. Plos One, 2012, 7: 30091

3 Gabrilovich E, Markovitch S. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In: Proceeding of the International Joint Conference on Artificial Intelligence, Hyderabad, 2007. 1606–1611

4 Medelyan O, Witten I H, Milne D. Topic indexing with Wikipedia. In: Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence, Chicago, 2008. 19–24

5 Bornmann L, Mutz R, Daniel H D. Multilevel-statistical reformulation of citation-based university rankings: the Leiden ranking 2011/2012. J Am Soc Inf Sci Technol, 2013, 64: 1649–1658

6 Docampo D, Cram L. On the internal dynamics of the Shanghai ranking. Scientometrics, 2014, 98: 1347–1366

7 Zha Y L, Zhou T, Zhou C S. Unfolding large-scale online collaborative human dynamics. Proc Natl Acad Sci USA, 2016, 113: 14627–14632

8 Li C, Li Q, van Mieghem P, et al. Correlation between centrality metrics and their application to the opinion model. Eur Phys J B, 2015, 88: 65

9 Watts D J, Strogatz S H. Collective dynamics of 'small-world' networks. Nature, 1998, 393: 440–442

10 Jöns H, Hoyler M. Global geographies of higher education: the perspective of world university rankings. Geoforum, 2013, 46: 45–59