2016 International Symposium on Nonlinear Theory and Its Applications,
NOLTA2016, Yugawara, Japan, November 27th-30th, 2016

On ranking and mining universities with the encyclopedia Wikipedia

Zongjian Li†, Cong Li†‡, Xiang Li†‡

†Adaptive Networks and Control Lab, Department of Electronic Engineering, Fudan University, Shanghai 200433, P.R.China
‡Research Center of Smart Networks and Systems, School of Information Science Engineering, Fudan Univeristy, Shanghai 200433, P.R.China
Email: {zjli13, cong_li, lix}@fudan.edu.cn

Abstract—The large data of *Wikipedia* has motivated new research branches, such as the evaluation of the reputation of an entry in its field by utilizing the *Wikipedia* data, where we can dig the relationship of entries via the analysis of *Wikipedia* data. In this work, we extensively evaluate the university entries in the Wikipedia to rank the universities over the whole world. Several reputation indicators for *Wikipedia* entries are introduced and compared with the QS and THE university ranking, while the in-degree reputation indicator has a strong correlation with QS and THE ranking. We propose two data mining methods to generate effective *Wikipedia* article reference subnetworks and we find that the community property of the university article reference subnetworks can reflect the geographic distribution of the universities.

1. Introduction

Wikipedia is a free-access and web-based multilingual encyclopedia that voluntaries from all around the world can write and edit. The English *Wikipedia* with more than 5 million articles is the largest one among the 291 Wikipedia editions, and has become one of the most popular public collaborative information repository [1]. Previous study on *Wikipedia* most focus on its collaborative systems, i.e. edit patterns [2–5]. Besides the relationship network of editors in *Wikipedia*, the articles of *Wikipedia* form a large-scale complex networks, the *Wikipedia* article reference networks (WARN). Articles are regarded as nodes, which are connected by the URL links. The trustreputation management system [5] of *Wikipedia* guarantees that the article jumps are based on the reputation of articles, in other words, articles with more links are high-reputation articles.

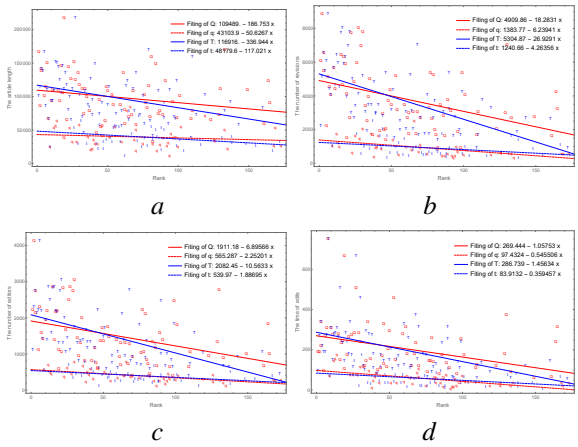
The massive data from *Wikipedia* provides us with a new resource to dig the relationship between things. One interesting question is whether the reputation of an article represents the rank of the entry of the article in its field. For instance, does the article of "google" entry in *Wikipedia* have more links than that of the article of an unnameable company? In this paper, we study this question by taking the rank of universities as the research subjects, since the rank of universities has been widely discussed in our daily life, and are associated to people's life. We analyze the relationship between the reputation of the university articles in *Wikipedia* and the human-defined university ranking, including the World University Rankings by Quacquarelli Symonds (QS) and by Times Higher Education (THE) magazine. The list of 114 chosen universities includes 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*, QS ranking and THE ranking are collected by 2015.

This paper is organized as follows. In Section 2, we study the relation between the reputation indicators for *Wikipedia* entries with the QS and THE rankings. In Section 3, we propose two methods to mine the effective *Wikipedia* article reference subnetworks (EWARS). In Section 4, we apply the EWARS to investigate the relationship between universities. Finally, we conclude in Section 5.

2. Reputation indicators for *Wikipedia* entries

The reputation of the entries can be evaluated by different indicators, such as the length of an article and the number of the revisions of the article. Three types of indicators, the intuitive criterion, the potential criterion and the deep seated criterion are discussed in this work. 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 the article, such as the number of the revisions of the article, the number of editors who have rewritten the article, and the time of the edits of the article in one year. The relation between these two types of criterions and the QS or THE ranking are shown in Figure 1. The "T" and "t" ("Q" and "q") marks denote that their x-axis is the THE (QS) ranking. Marks in upper-case letter mean that the universities locate in an English-speaking region, and the lower-case letter marks represent the universities in a non-English-speaking region. The 114 universities are located in 20 countries and regions, and 6 of them are English-speaking countries. Totally, 72 universities are in the English-speaking region.

The deep-seated criterion takes into account not only the inherent properties of an article but also the relationship between the article and others. Articles are regarded as nodes and the URL links as connections between articles. This study only adapts articles in English *Wikipedia*, while the external links to other web sites, such as non-



a *b* *c* *d*

Figure 1: Relation between the intuitive reputation criterion or the potential reputation criterion of univerisities in *Wikipedia* and their QS or THE ranking.

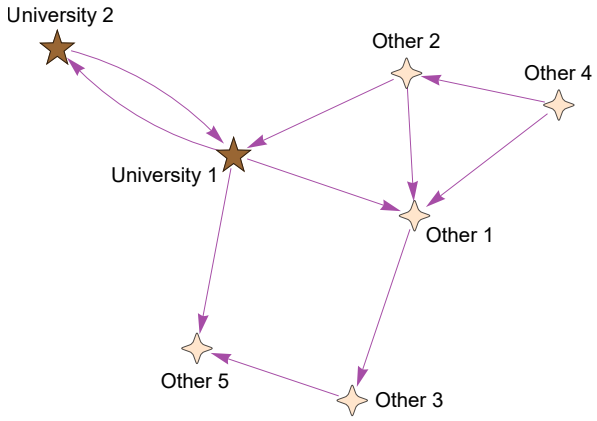
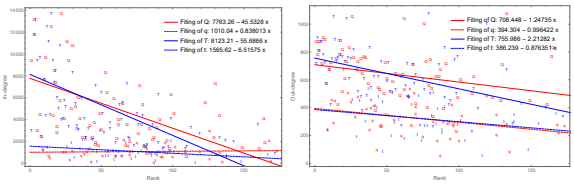


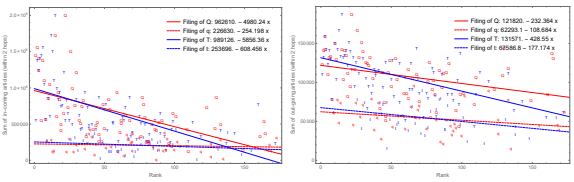
Figure 2: An example of the *Wikipedia* article reference networks, including 7 articles and 2 of them are articles of university entry. For "University 1", there are 3 out-going links involving 4 articles within 1 hop, while there are 5 out-going links involving 5 articles within 2 hops.

English Wikipedia sites, non-article page and self-loop site are not counted in. The *Wikipedia* article reference networks (WARN) are directed networks (see Figure 2). Outgoing links direct to another article of a key-word that appears in current article, however, the sources of in-coming links can be any article in *Wikipedia*. Figure 3 shows that the QS or THE ranking is more strongly linear correlated with the in-degree reputation indicator than with the out-degree reputation. Inspired by [6], the sum of the degree of two-hopcount neighbors in WARN is calculated and compared with the QS or THE ranking (See Figure 4). Same with existing research [7], we find that the WARN has small-world properties, i.e. small average shortest path and large clustering coefficient [8]. This phenomenon implies that a relative large hopcount of a node may cover most part of the whole network. Hence, we only consider the



a *b*

Figure 3: Relation between the in-degree and out-degree of univerisity articles in *Wikipedia* and QS or THE ranking.



a *b*

Figure 4: Relation between (a) the sum of in-degree and (b) the sum of out-degree of the articles in 2-hopcout from a university and its QS or THE ranking.

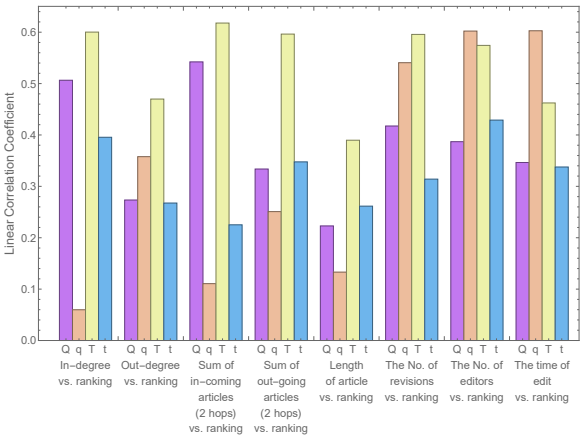


Figure 5: The linear correlation coefficient between the reputation indicators for university entries and the QS or THE university rankings.

sum of in-degree and out-degree of the node itself and its 1-hopcount or 2-hopcount neighbors.

The linear correlation coefficients between three types of criterions and the QS or THE ranking are shown in Figure 5. For the universities located in English-speaking countries, the deep-seated reputation criterion performs better than other types of criterions in characterizing the university ranking.

3. Two data mining methods to generate effective *Wikipedia* article reference subnetworks

The *Wikipedia* article reference network (WARN) involves all articles in *Wikipedia*, however, only a small part of the articles is essential and interesting for researchers from different fields. It is a challenge to extract the effective *Wikipedia* article reference subnetworks (EWARS), which contain only the needed articles. In this work, we design two methods to generate the EWARS, and take the EWARS of the 114 universities as an example.

3.1. Path length (PL) Method

First, we generate a pre-EWARS, which is composed of the articles of the 114 universities along with their 1-hopcount neighbor articles connected by the in-coming links (or out-going links), and all the connections between these articles. Second, we convert this directed network to an undirected pre-EWARS by adding inverse links. Note that the WARN is a directed network which contains in-coming links and out-going links. The study in Section 2 has shown that the neighbours connected by in-coming links and the out-going links of the universities are considerably different, thus, there are two types of pre-EWARS. Third, the shortest path length between any two university articles is calculated, and the length represents the strength of the correlation between two universities. Then, the shortest path length threshold will be set to obtain the EWARS. Two university articles with an equal or smaller shortest path length than the threshold will be connected by an undirected link, otherwise disconnected.

3.2. Vertex connectivity (VC) Method

Similarly to PL Method, we first generate the pre-EWARS, and then remove connections between university articles and connections between non-university articles. In other words, only undirected connections between a university article and a non-university article are kept (see Figure 6). Next, the vertex connectivity between any two university articles are calculated to characterize the correlation between two universities. The vertex connectivity is defined as the smallest number of nodes to remove that makes no path between two nodes. Finally, a vertex connectivity threshold will be set to generate the EWARS. Two university articles with a vertex connectivity over the threshold will be connected by an undirected link, otherwise disconnected.

4. Application of the EWARS

We obtain an undirected pre-EWARS of the 114 university articles with their out-going connected articles. There are 29,416 nodes and 1,898,348 undirected links in this pre-EWARS. With the EWARS generated by using PL method, there are 1,900 pairs of university articles have shortest path length of 1, while 4,385 pairs of university articles have shortest path length of 2, and 156 pairs of university articles have shortest path length of 3. That means each pair of the 114 university articles has a path length shorter than 4, so this pre-EWARS is a strongly-connected network. When we set a threshold of the shortest path length as 1, we get an EWARS of the university articles (see Figure 7). The (red) nodes in the middle are the largest clique [9], which has 40 universities including 37 American universities and 2 Canadian universities. However, we find the EWARS is already a dense network even when we choose the smallest threshold of 1. The detail relationships between the universities are difficult to be digged.

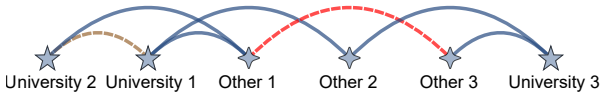


Figure 6: The pre-EWARS with removing picked links (marked in dashed lines) in VC method.

We use the VC method to generate the EWARS of the 114 university articles with their out-going connected articles. Compared to the shortest path, the vertex connectivity between any two articles has a large range of values, and approximately follows binomial distribution (see Figure 8). The quartiles of the vertex connectivity thresholds are $T = 69, 91, 117$. EWARS with a threshold $T = 172$ is shown in Figures 9. In Figure 9, university articles are automatically divided in groups by their relation with other university articles by cartographic software, so the groupings are just for demonstration. Specially, links in the largest and the second largest clique are colored as 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, the right-top part is mainly occupied by universities in Hong Kong, China, Australia and Singapore. Notice that the results for the EWARS of the 114 university articles with their in-coming connected articles has similar results with the above discussion. The results demonstrate that the community property of the EWARS generated by the VC method matches the geographic distribution of the universities. It implies that the relationship between entries in *Wikipedia* can be reflected by the properties in EWARS.

5. Conclusion

In this work, we study several reputation indicators for entries in *Wikipedia*. We take the university articles as an example, and compare them with the QS and THE university rankings using internal and external properties of articles. The linear correlation coefficient between the indicators and the QS and THE rankings are also calculated. We find that the in-degree reputation criterion has the most strong correlation with the QS or THE rankings for the u-

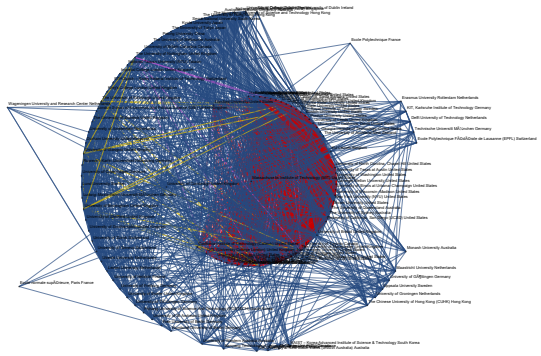


Figure 7: The EWARS of the university articles generated by PL method with threshold $T = 1$.

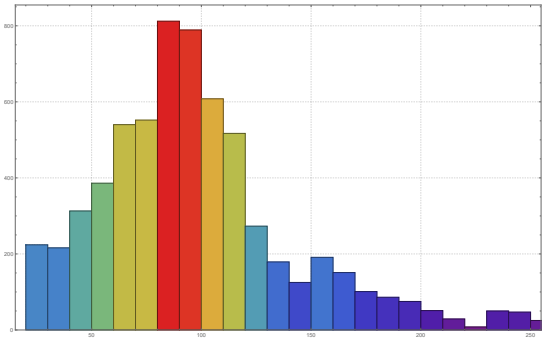


Figure 8: The distribution of vertex connectivities between any two articles in pre-EWARS generated by VC method.

niversities located in English-speaking countries. We then propose two data mining methods, based on the shortest path length and the vertex connectivity, to generate effective *Wikipedia* article reference subnetworks. An interesting finding is that the community property of the effective university article reference subnetworks well matches the geographic distribution of the universities.

Acknowledgments

This work was partly supported by the Natural Science Fund for Distinguished Young Scholar of China (No. 61425019), the National Natural Science Foundation (No. 61273223), and Natural Science Foundation of Shanghai (No. 16ZR1446400).

References

[1] *https://en.wikipedia.org/wiki/Wikipedia*.

[2] U. Pfeil, P. Zaphiris and C.S. Ang, *Cultural Differences in Collaborative Authoring of Wikipedia*, Journal of Computer-Mediated Communication 12 (1): 88, 2006.

[3] T. Yasseri, R. Sumi and J. Kertész, *Circadian Patterns of Wikipedia Editorial Activity: A Demographic Analysis*, PLoS ONE 7 (1): e30091, 2012.

[4] R. Sumi, T. Yasseri, A. Rung, A. Kornai, J. Kertész, *Characterization and prediction of wikipedia edit wars*, Proceedings of the ACM WebSci' 11: 1C3, 2011.

[5] S. Javanmardi, C. Lopes, P. Baldi, *Modeling user reputation in wikis*, Statistical Analysis and Data Mining 3: 126C139, 2010.

[6] C. Li, Q. Li, P. Van Mieghem, H. E. Stanley and H. Wang, *Correlation Between Centrality Metrics and Their Application to the Opinion Model*, European Physical Journal B, Vol. 88, No. 3, article 65, 2015.

[7] V. Zlatić, M. Božičević, H. Štefančić and M Domazet, *Wikipedias: Collaborative web-based encyclopedias asZBSD complex networks*, Physical Review E, 74(1): 016115, 2006.

[8] Watts D. J and Strogatz S. H, *Collective dynamics of "small-world" networks*, 1998 Nature, **393**, 440-442.

[9] C. Li, H. Wang and P. Van Mieghem, *New Lower Bounds for the Fundamental Weight of the Principal Eigenvector in Complex Networks*, Third International IEEE Workshop on Complex Networks and their Applications, in Proceedings-10th International Conference on Signal-Image Technology and Internet-Based Systems, SITIS 2014, p 317-322, April 7, 2015.

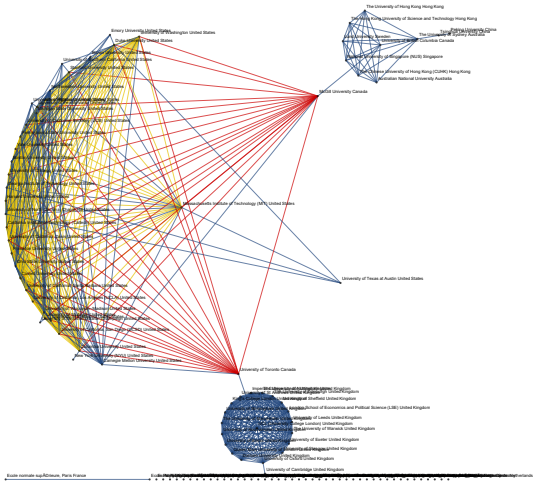


Figure 9: The EWARS obtained by utilizing the VC method with a vertex connectivity threshold $T = 172$.