MONITORING NETWORK STRUCTURE AND CONTENT QUALITY OF SIGNAL PROCESSING ARTICLES ON WIKIPEDIA

Tao C. Lee and Jayakrishnan Unnikrishnan

School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
Email: {tao.lee, jay.unnikrishnan}@epfl.ch

ABSTRACT

Wikipedia has become a widely-used resource on signal processing. However, the freelance-editing model of Wikipedia makes it challenging to maintain a high content quality. We develop techniques to monitor the network structure and content quality of Signal Processing (SP) articles on Wikipedia. Using metrics to quantify the importance and quality of articles, we generate a list of SP articles on Wikipedia arranged in the order of their need for improvement. The tools we use include the HITS and PageRank algorithms for network structure, crowdsourcing for quantifying article importance and known heuristics for article quality.

Index Terms— Signal processing, Wikipedia, network analysis, crowdsourcing, information quality metrics.

1. INTRODUCTION

Wikipedia has become one of the most widely-used resources of signal processing. Therefore, maintaining a high quality standard for SP articles on Wikipedia is without doubt important. This research is devoted to monitoring the network structure and content quality of SP articles on Wikipedia. The goal is to perform automatic analysis of the network structure and content quality of SP articles on Wikipedia and to generate a list of articles ordered in terms of their need for improvement.

The wide popularity of Wikipedia has attracted the attention of researchers to explore its link structure and information quality. The two main research directions adopted in existing literature are : (i) Exploring how the link structure of Wikipedia reflects the importance of its articles [1] [2] and (ii) Quantifying the information quality of the various articles on Wikipedia [3] [4] [5]. For (i), content relevance algorithms such as HITS [6] and PageRank [7] are applied to the directed graph formed by Wikipedia articles and the ranking results are examined for different categories of articles. The two algorithms typically capture distinct aspects of the network structure [1]. For (ii) various heuristics are adopted to measure article quality. For example, in [3] seven different information quality metrics are proposed, each computed using various statistical characteristics of the articles such as article length, number of links, number of editors and so on. These metrics are then used in experiments to identify featured articles from randomly chosen articles on Wikipedia. Successful classification is reported with few misclassified instances [3].

In this paper, we implement the HITS and PageRank algorithms for network analysis, the quality heuristics from [3] for estimating article quality and crowdsourcing to estimate article importance for the articles under the SP category and its subcategories on Wikipedia. Using the results of these algorithms we generate a list of articles ordered in terms of their need for improvement. Such a list could serve as a valuable tool for potential contributors to Wikipedia when selecting articles for editing. Due to lack of space in the paper we have made most of our results available online [8].

For implementing these algorithms, we downloaded the content and revision histories of all Wikipedia articles classified under the SP category and its subcategories [9] in XML format. We used free software from Wikipedia [10], [11], [12], [8] to download and process these XML files.

This paper is organized as follows. We introduce importance ranking and network analysis in Section 2 and provide a comparison to crowdsourcing in Section 3. In Section 4, we present results of information quality ranking and analysis. In Section 5 we present differential analysis and our methodology of generating an improvement list for SP articles. In Section 6, we summarize our results and present avenues for future work.

2. IMPORTANCE RANKING & NETWORK ANALYSIS

2.1. Overview

Since the rise of the Google search engine, ranking the importance of hyper-linked articles has become an active research topic [1], [6], [7], [13], [14]. Two popular algorithms are

This research was supported by ERC Advanced Investigators Grant: Sparse Sampling: Theory, Algorithms and Applications - SPARSAM - no. 247006.

the HITS and PageRank algorithms. In applying these algorithms the hyper-linked articles are represented as nodes on a directed graph, with a directed edge from node i to j indicating a hyperlink from article i to article j.

The HITS algorithm defines authority and hubness scores for each node, with authority score defined as the normalized sum of hubness scores of neighboring nodes, and hubness score as the normalized sum of authority scores of neighboring nodes. The algorithm is implemented via an iterative process that converges to a fixed point of authority and hubness scores. The PageRank algorithm, on the other hand, uses a stochastic square matrix to represent the link structure in the network with the rows and columns of the matrix representing nodes in the network. The (i, j)-th entry of the matrix represents the probability that a reader on page j moves to page i assuming she picks any outgoing link with equal probability. The PageRank score of node i is the i-th entry of the eigenvector of the matrix corresponding to an eigenvalue of 1. The property of the square stochastic matrix allows one to compute its eigenvector using power iteration. To overcome some singular cases, the inventors of the PageRank algorithm [7] proposed a random walk model to assure that the computation process converges to a meaningful fixed point.

For performing network analysis we need to select the extent of the network relevant to signal processing. A straightforward but memory-intensive approach is to perform network analysis on the entire Wikipedia network. However, we adopt a simpler approach restricting our network to articles from the signal processing category and its subcategories as shown in Figure 1. While this is computationally less intense, this approach ignores links from articles outside signal processing category, which could potentially provide more information about the importance of various SP articles.

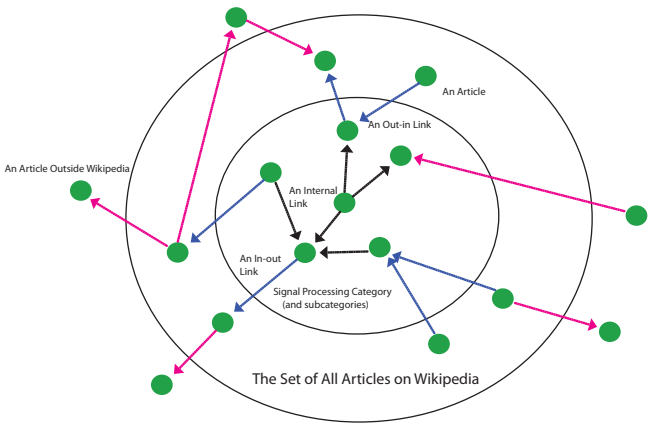

Fig. 1. Network structure of signal processing articles on Wikipedia.

Two types of links deserve more attention: broken links and in-out links. Broken links are links that point to articles that do not yet exist in Wikipedia. A snapshot of the current Wikipedia articles naturally contain many broken links. Our implementation of ranking algorithms considers these broken links in network analysis, but filters out these links when generating the importance rankings. Similar considerations are also applied to in-out links, which are links that point from articles within SP categories to articles outside SP categories.

2.2. Experimental results and observations

The directed graph formed by SP articles on Wikipedia has 1157 nodes and 1762 edges. The complete SP graph in vector graphics format can be downloaded from [8].

We ranked the articles in the order of their HITS and PageRank scores. The top-100 ranking lists for both the HITS and the PageRank scores are available at [8]. In the HITS list we excluded articles with zero authority scores. The following are the top-10 articles in HITS scores:

1. Dirac delta function 2. Dirac comb 3. Nyquist-Shannon sampling theorem 4. Whittaker-Shannon interpolation formula 5. Nyquist frequency 6. Fourier analysis 7. Discrete Fourier transform 8. Digital signal processing 9. Fast Fourier transform 10. LTI system theory.

We observed that the HITS and PageRank algorithms produce different ranking lists. In Section 3, we see how these rankings compare with the importance rankings obtained via crowdsourcing. A number of interesting observations can be made by comparing the HITS and the PageRank rankings. Notably the following three observations are highlighted.

Self-reinforcing links: the case of *Itakura-Saito distance*

One interesting observation after examining the ranking lists is that some less-known articles have high PageRank ranking but have low HITS ranking (not even on the list of top-100). We highlight the example of *Itakura-Saito distance* because it is not a well-recognized top-100 article by signal processing researchers.

A closer look at *Itakura-Saito distance* shows that its local network structure is formed by seven nodes with bidirectional links pointing to each other. These seven nodes form a supernode that is isolated from the rest of the network. In the random walk model assumed by the PageRank algorithm, this supernode has higher probability to be traversed, and each node within this supernode gets higher probability to be traversed across iterations as they are connected by bidirectional links.

The missing top-100: where is *image denoising*?

A careful examination of the top-100 rankings reveals that some important SP articles are missing. *Image denoising*, for instance, is widely recognized by signal processing researchers, but does not even make to the top-100 in both rankings. A closer look at its local network structure is shown in

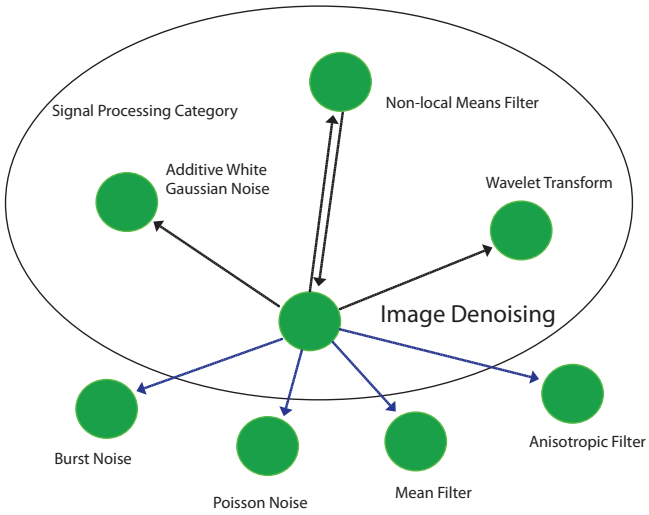

Fig. 2. Local network structure of *image denoising*.

Fig 2. We see that this article acts as a parent node for (i.e., links to) several other nodes. But among its children only *non-local means filter* links back to *image denoising*. This leads to the low PageRank and HITS scores. However, the importance scores of *image denoising* can be easily improved by adding a link to it from the article on *wavelet transform* which already has high importance scores.

The underestimated supernodes

One of the problems we encountered in our analysis is the existence of distinct articles on closely-related topics. For instance there are distinct articles on *image processing* and *digital image processing*. The HITS and PageRank algorithms rank these articles separately and this leads to the importance dilution of these articles. Other examples of such dilution can be seen in some articles on wavelets, e.g., the articles on *wavelet transform* and *continuous wavelet transform*. A potential solution to this problem is to bunch up such closely related articles as supernodes in the network and consider them as a single entity when performing network analysis.

3. CROWDSOURCING

While automatic ranking algorithms are attractive, they cannot be used as reliable indicators of article importance unless these rankings match the importance as seen by signal processing researchers. For this reason, we carried out a crowdsourcing experiment among SP researchers from EPFL and elsewhere. Nineteen researchers were presented with a list of top-100 articles ranked by the HITS algorithm, and asked to select the top-20 articles based solely on their own judgement and not on how these articles are written on Wikipedia. We collected the results and ranked the articles according to the frequency of the article being selected in researchers' top-20 lists. The entire list can be downloaded from [8]. The following are the top-10 voted articles:

1. Convolution 2. Fast Fourier transform 3. Nyquist-Shannon sampling theorem 4. Sampling (signal processing) 5. Fourier analysis 6. Filter (signal processing) 7. Kalman filter 8. Cross-correlation 9. Wavelet transform 10. Wiener filter

Several interesting observations are made as follows. (i) Eight articles recognized by more than half of the researchers are also ranked as top-20 articles by the HITS ranking, whereas only three articles of them are ranked as top-20 by the PageRank ranking. This demonstrates the accuracy of the HITS ranking. (ii) Top-20 rankings differ quite significantly even for researchers from the same laboratory as only 12 out of 75 articles are selected by more than half of the researchers. (iii) *MATLAB* is the only commercial article that comes in the top-20 list and is recognized by one-fifth of the researchers. (iv) *Audio signal processing* and *speech signal processing* are not recognized by both the HITS ranking and SP researchers as top-20, whereas *image processing* is widely recognized by both. (v) We observed that the problem of importance dilution still persists here. For instance, some researchers voted for *image processing* while some others preferred *digital image processing*, although both these topics are essentially identical. Hence the real importance of *image processing* gets diluted because of this split of votes.

In conclusion, the HITS ranking reflects the crowdsourced importance ranking better than the PageRank ranking and hence we choose the HITS ranking as a proxy for the importance of the articles.

4. INFORMATION QUALITY RANKING

4.1. Overview

Among seven independent metrics proposed in [3], we only adopt two of them, the *reputation* metric and the *completeness* metric, because they are more applicable to our research. The definitions of these two metrics are given in Equation 1. Experimental results have shown that *completeness* is a more reliable information quality metric.

$$\begin{aligned} reputation &= 0.2 * (\text{number of editors}) \\ &\quad + 0.2 * (\text{number of edits}) + 0.1 * (\text{connectivity}) \\ &\quad + 0.3 * (\text{number of reverts}) \\ &\quad + 0.2 * (\text{number of external links}) \\ &\quad + 0.1 * (\text{number of registered user edits}) \\ &\quad + 0.2 * (\text{number of anonymous user edits}) \\ completeness &= 0.4 * (\text{number of internal links}) \\ &\quad + 0.6 * (\text{article length}) \end{aligned} \tag{1}$$

4.2. Experimental results and observations

Our top-100 ranking lists for the *reputation* ranking and the *completeness* ranking can be found here [8].

Since we use heuristics-based metrics to measure the quality, they can never be fully foolproof. For example, the top-20 list of the *completeness* ranking is biased towards long articles. Some of these articles are: *Geophysical MASINT, Avizo (software)*, and *JPEG 2000*. Short articles with many links to other articles have low *completeness* scores, and *speech signal processing* is a good example.

5. AN IMPROVEMENT LIST

Using both importance rankings and information quality rankings, we want to highlight a list of articles that are most in need of improvement. For this purpose, we perform differential analysis between rankings to highlight articles having high importance ranking but low information quality ranking. We use the *difference* score, defined by the difference of the HITS ranking and the *completeness* ranking, as the metric.

In experiments we observed that the ranking list generated by differential analysis is noisy. Hence, we instead use a Need For Improvement (NFI) score as defined in Equation 2. We first use two threshold parameters to select articles that have a low *completeness* score and simultaneously a high *difference* score. For this purpose, we define the θ function and δ function in Equation 2. We then define the NFI score of such an article as the product of the difference between the total number of articles and the HITS ranking of the article, the θ function of *difference* score and the δ function of *completeness* score. In effect, articles with high NFI scores have high importance ranking, high ranking difference between importance and information quality, and are still incomplete.

$$\text{NFI score} = \Gamma * \theta(\textit{difference}\ \text{score}) * \delta(\textit{completeness}\ \text{score}) \tag{2}$$

where

$$\Gamma = (\text{total articles} - \text{the HITS ranking})$$

$$\theta(s) = \begin{cases} s & : s > \textit{difference}\ \text{threshold} \\ 0 & : \text{otherwise} \end{cases}$$

$$\delta(s) = \begin{cases} s & : s < \textit{completeness}\ \text{threshold} \\ 0 & : \text{otherwise} \end{cases}$$

We compute the NFI scores for all SP articles, and rank them in terms of their NFI scores [8]. From the results we observed that articles related to speech and audio signal processing tend to have high NFI scores, suggesting that these articles are the most in need for improvement.

6. CONCLUSIONS AND FUTURE WORK

A number of observations have been made in this research about the current status of SP articles on Wikipedia.

First, we saw that some important articles, e.g., *image denoising*, do not get high scores in the HITS and PageRank algorithms because there are not enough articles referring to these important articles. This can be improved by adding more links.

Second, our differential analysis revealed that some important SP articles have high importance rankings but low information quality rankings and hence have a high need for improvement. Some of these articles are: *digital signal processing, image processing, spectral density estimation, statistical signal processing, video processing* and *time-frequency analysis*. We suggest to invest efforts in improving the content quality of these important articles.

Third, articles related to *audio signal processing* and *speech signal processing* are highlighted for improvement. Both the importance rankings and information quality rankings are far behind articles in the image processing category. Thus we suggest the SP community invest efforts in strengthening the content of articles related to audio and speech signal processing.

Fourth, some topics have multiple articles dealing with the same topic, e.g., *image processing* and *digital image processing*, *wavelet transform* and *continuous wavelet transform*, etc. These articles may be merged to avoid redundancy.

Finally, several interesting research questions are worth exploring in the future. (i) We observed some SP articles have community structures. Notably *wavelet* and *image processing* have many closely-related articles. Automatic detection of such community structures of SP articles would be worthwhile to investigate. (ii) It would be worthwhile to enlarge the scope of network analysis to include the entire Wikipedia network. This would give more accurate estimates of the network structure of SP articles. (iii) It would be worthwhile to enlarge the scope of crowdsourcing to include information quality evaluation by a larger group of researchers. (iv) Web interfaces have become popular for performing online crowdsourcing and displaying research results [15]. We think that a web interface that performs realtime monitoring of network structure and content quality of SP articles on Wikipedia would be a useful tool for contributors to Wikipedia.

Acknowledgement

We thank Martin Vetterli for several helpful suggestions, and anonymous contributors to crowdsourcing.

7. REFERENCES

[1] F. Bellomi and R. Bonato, "Network analysis for Wikipedia," in *Proceedings of Wikimania 2005, The First International Wikimedia Conference*, 2005.

[2] Ulrik Brandes, Patrick Kenis, Jürgen Lerner, and Denise van Raaij, "Network analysis of collaboration structure in Wikipedia," in *Proceedings of the 18th international conference on World wide web*, New York, NY, USA, 2009, WWW '09, pp. 731–740, ACM.

[3] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser, "Assessing information quality of a community-based encyclopedia," *JASIST*, vol. 58, pp. 1720–1733, 2007.

[4] J. E. Blumenstock, "Automatically assessing the quality of Wikipedia articles," *Technical Report, School of Information, University of Califonia, Berkeley*, 2008.

[5] Sara Javanmardi and Cristina Lopes, "Statistical measure of quality in Wikipedia," in *Proceedings of the First Workshop on Social Media Analytics*, New York, NY, USA, 2010, SOMA '10, pp. 132–138, ACM.

[6] J. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, pp. 604–632, 1999.

[7] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, vol. 30, pp. 107–117, 1998.

[8] Tao C. Lee and Jayakrishnan Unnikrishnan, "SP Wiki software tools, ranking lists and graphs," `http://lcav.epfl.ch/page-87349.html`, 2012.

[9] Wikipedia, "Wikipedia signal processing category," `http://en.wikipedia.org/wiki/Category:Signal_processing`, 2012.

[10] Wikipedia, "Wikipedia database download web page," `http://en.wikipedia.org/wiki/Wikipedia:Database_download`, 2012.

[11] Wikipedia, "Wikipedia export web page," `http://en.wikipedia.org/wiki/Special:Export`, 2012.

[12] Wikipedia, "Mwdumper web page," `http://www.mediawiki.org/wiki/Mwdumper`, 2012.

[13] S. Nomura, S. Oyama, T. Hayamizu, and T. Ishida, "Analysis and improvement of HITS algorithm for detecting web communities," in *Applications and the Internet, 2002. (SAINT 2002). Proceedings. 2002 Symposium on*, 2002, pp. 132 –140.

[14] D. Austin, "How Google finds your needle in the web's haystack," `http://www.ams.org/samplings/feature-column/fcarc-pagerank`, 2012.

[15] Jonas Arnfred, "Trailhead: a graphical representation of articles," `http://trailhead.epfl.ch/`, 2012.