Web Link Strength Analysis

ABSTRACT

Web link analysis is one of the core problems in the web search area, and one of the most popular web link analysis algorithms is PageRank, which has shown great potential in improving the web search performance. There are usually multiple web pages linking to a given web page, and a web page is usually linking to multiple web pages. Most existing algorithms treat these different links equally. However, according to our analysis, different hyperlinks between web pages have different strengths, and existing algorithms do not explicitly consider the strength of each web link. For example, the multiple outlinks or inlinks of a web page may have different levels of importance, and such importance levels are reflected by a few factors including content-based factors and vision-based factors. In this study, we define the web link strength analysis task and investigate how to automatically estimate the strength of a given web link. We solve the task by using support vector regression with a few useful features, and experimental results on a manually labeled dataset show the efficacy of the regression method for link strength estimation. We further incorporate the estimated link strengths into the weighted PageRank algorithm for improving the topic distillation results. Experiment results on the TREC datasets show promising results.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia

General Terms

Algorithms, Performance, Design, Human Factors.

Keywords

Link strength estimation, link analysis, weighted PageRank, topic distillation.

1. INTRODUCTION

Hyperlinks between web pages are very valuable for web search and web link analysis is one of the core problems in the web search area. PageRank [1] and HITS [2] are two of the most popular algorithms of link analysis, and almost all web search engines have adopted some link analysis algorithm for improving the web search performance. In particular, PageRank and its variants have been widely and successfully used for various web search tasks.

All link analysis algorithms rely on the hyperlinks between web pages. In most existing algorithms, different links between web pages are treated equally. For example, in the PageRank algorithm, the PageRank value of a web page is passed to its linked web pages uniformly. However, according to our analysis, different hyperlinks between web pages have different strengths, and existing algorithms do not explicitly consider the strength of each web link.

A given web page may have multiple outlinks to different web pages. These hyperlinks are usually displayed in different positions of the web page and the anchor texts for them are usually with different styles. Moreover, the linked web pages usually have different degrees of relevance with the given web page. Some linked web pages may be highly relevant with the given web page, but some linked web pages may be less relevant with the given web page. All these factors convince us that the multiple outlinks (or inlinks) of a web page may have different levels of strength, and the strength of a link reflects the importance of the link by considering both content-based factors and vision-based factors. To the best of our knowledge, none of previous studies has explicitly defined the link strength concept and investigated the link strength estimation task. We believe that the strength of web link is very useful in the web search area.

In this study, we define the web link strength analysis task and investigate how to automatically estimate the strength value of a given web link. We solve the task by using support vector regression with a few useful features, and experimental results on a manually labeled dataset show that the efficacy of the regression method for link strength estimation.

We further investigate making use of the estimated link strength values for the topic distillation task. We incorporate the link strength values into the weighted PageRank algorithm and compute a better PageRank value for each web page. The PageRank values are then used for improving the topic distillation results. Experiment results on a TREC dataset show promising results.

The rest of this paper is organized as follows: We briefly review related work in Section 2. The link strength analysis problem, technique and evaluation are described in Section 3. The use of link strength for topic distillation and its evaluation results are presented in Section 4. Lastly, we conclude this paper in Section 5.

2. RELATED WORK

Web link analysis has attracted much attention in the past decades and several advanced link analysis methods have been developed for improving web search performance. Most previous methods are extensions of the PageRank and HITS algorithms. For example, [3] segmented a web page into different blocks through VIPS and developed Block Level PageRank and Block Level HITS algorithms. [4] considered both the hierarchical structure and the link structure of the Web and proposed a Hierarchical Random Walk Model. [5] investigated temporal aspects as factors in assessing the authoritativeness of web pages and developed Time-Weighted PageRank (TWPR) algorithm. [8] incorporated the topical model within both PageRank and HITS without affecting the overall property and still rendered insight into topic-level transition. [9] brought out a very simple filtering algorithm to recognize and eliminate these unrelated links using Content Lexical and Positional analysis, and then apply standard the PageRank algorithm. None of the above works has explicitly defined and investigated the link strength analysis problem by considering various factors.

Another related research work is citation strength prediction [6], which aims to assign a strength value to each citation in a scientific paper to reflect the importance of the citation.

3. LINK STRENGTH ANALYSIS

3.1 Problem and Corpus

PageRank is a link analysis algorithm and it assigns a numerical weight to each element of a hyperlinked set of web documents, with the purpose of "measuring" its relative importance within the set. The PageRank value for a page *u* is dependent on the PageRank values for each page *v* in the set containing all pages linking to page *u* and the proportion of the link from *v* to *u*. The proportion of the link is usually calculated as one divided by the number of outbound links in page *v*, which means all the links in page *v* are treated equally. However, the links in a web page may have different levels of strength degree, and the strength of a link reflect the importance of the link by considering both content-based factors and vision-based factors. A link which exists in the central area of a web page and directs to a highly relevant web page is more important than a link which exists in a side area and directs to an irrelevant web page. For example, the links of 'about me', 'privacy' in a given web page usually direct to less important and less authoritative pages, because these links are usually exist in side areas and the web pages of 'about me' and 'privacy' are not very relevant with the major content of the given web page. Since most pages in a website often contain such links, and the web pages of 'about me' and 'privacy' can obtain higher PageRank values. If we can discriminate the important links and the unimportant links, we will obtain better PageRank values for the web pages. Figure 1 shows part of a sample web page[1] with multiple outlinks, and the links has different levels of strength degree, where the larger the value is, the more important the link is. The link strength analysis problem investigated in this paper is thus aiming at automatically estimate the strength value of each web link.

As far as we know, there exists no publicly available benchmark corpus for this problem. Therefore, we annotated our own corpus for this task. We first collected 256 pages from the DOTGOV corpus, including 1659 links. Two students were employed for link strength annotation. They were asked to label a score between 1 and 4 for each link after they carefully read the pages which the link is directed from and directed to. Here, "4" means "very strong and important", "3" means "moderately strong and important", "2" means "less strong and important", "1" means "not strong and not important".

Figure 1. A sample page with different kinds of links

Note that during the annotation process, we developed an annotation tool, which is showed in Figure 2, to assist our work. The tau-b between the two annotations is 0.56, and the Correlation Coefficient (ρ) is 0.67, which means that the annotation results are acceptable. Finally, the overall strength value of each link is the average of the scores provided by the two annotators. Figure 3 shows the distribution of overall link strength values, and we can see there is a considerable portion of links which are not important.

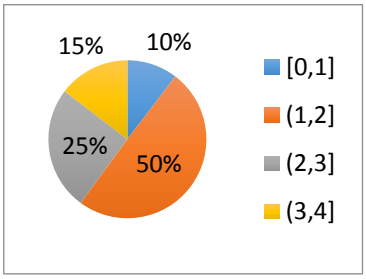

Figure 3. Distribution of link strength values in our annotated corpus

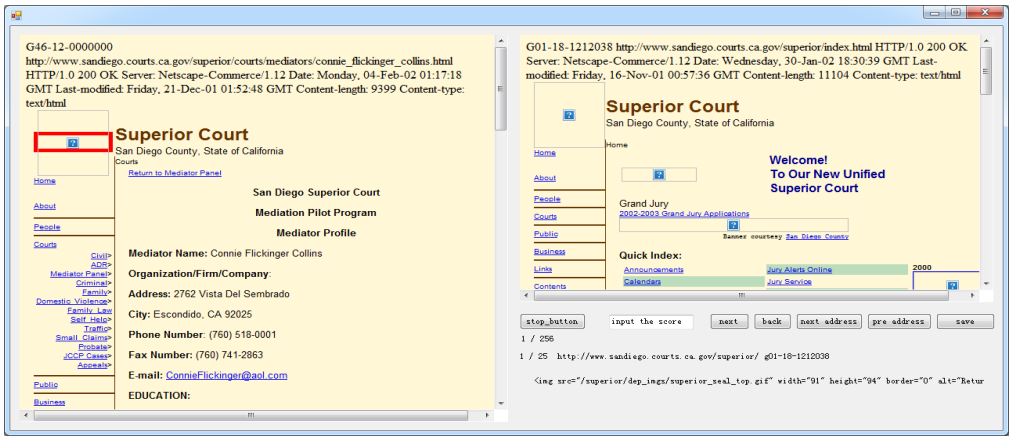

Figure 2. The annotation tool

[1] http://www.aaai.org/Conferences/AAAI/aaai15.php

3.2 Method

As mentioned above, link strength estimation is a task of mapping each link to a numerical value corresponding to the strength degree. The larger the value is, the more importance the link is. The task can be considered a regression problem and in this study, we adopt the support vector regression (SVR) method for addressing this prediction task. The SVR algorithm is firmly grounded in the framework of statistical learning theory (VC theory). The goal of a regression algorithm is to fit a flat function to the given training data points. More specifically, we use the LIBSVM tool with the RBF kernel for this regression task.

We develop the following two groups of features for each link, which are derived from different dimensions.

Vision-based features: Usually, the importance of a link depends on whether the link is visible and how the link attracts people's attentions, which lies on the position and size of the area the link exists in. For example, if a link is shown in a bigger area and lies in a more central position, the link is more important. We use the absolute and relative position and size of the link in the page, together with correlations between them, as eight numerical features.

Content-based features: In general, the more relevant the two pages linked by a hyperlink are, the stronger the link is. We obtain the TFIDF vector of each page and compute different relevance values of two pages as seven numerical features, such as the Euclidean distance, cosine similarity, etc.

We use the open-source WebKit[2] toolkit to browse the page and communicate with the browser to get the visual-based features. We compute the above feature values separately, and use all the feature values for regression. The provided svm-scale program is used to scale all the above feature values.

3.3 Evaluation Results

For evaluation, we randomly separated the labeled link set into four sets, and selected three of them as a training set and the remaining one as a test set. We then used the LIBSVM tool[3] for training and testing. The process was conducted for four times, and finally the results were averaged. Two standard metrics were used for evaluating the prediction results. The two metrics are as follows:

Mean Square Error (MSE): This metric is a measure of how correct each of the prediction values is on average, penalizing more severe errors more heavily.

Pearson's Correlation Coefficient (ρ): This metric is a measure of whether the trends of prediction values matched the trends for human-labeled data.

Table 1 shows the prediction results with the standard deviation values. Three simple baselines are used for comparison: a random baseline, one baseline using the cosine similarity between two linked pages, and one baseline using a link's relative position in the page. The results of the SVR method after removing every group of features are also reported.

We can see that the overall results of the SVR method with all features are very promising because it outperforms all baselines. We can also see the each feature group is beneficial to the overall prediction performance, because the performance values have a decline after removing any group of features.

Table 1. Prediction results

Method	MSE	ρ
All features	**0.3566** (stdev=0.0424)	**0.4867** (tdev=0.0885)
w/o vision-based features	0.5437 (stdev=0.0545)	0.2504 (tdev=0.0754)
w/o content-based features	0.6507 (stdev=0.0758)	0.3982 (tdev=0.1849)
Cosine Similarity Baseline	0.7105 (stdev=0.1524)	0.0089 (tdev=0.0013)
Relative Postion Baseline	0.6733 (stdev=0.0128)	0.1843 (tdev=0.0112)
Random Baseline	1.2009 (stdev=0.0796)	0.0003 (tdev=0.0002)

Finally, we apply the SVR method to predict the strength value of each link in the dataset. Each link e_{ij} is associated with a strength value $Strength(e_{ij})$.

4. USING LINK STRENGTH FOR TOPIC DISTILLATION

4.1 The Topic Distillation Task and Dataset

In order to show the usefulness of the estimated link strength, we chose the topic distillation task in web track of TREC2003 as the evaluation task [7]. The topic distillation task aims to find the key entry page of some broad topic. A good entry page for a topic needs to meet the following requirements:

• Is principally devoted to the topic;

• Provides credible information on the topic, and

• Is not part of a larger site also principally devoted to the topic.

For example, a good entry page for the topic of "science" is *www.nsf.gov/* .

The dataset used in this task is the DOTGOV dataset, and it contains 1,247,753 documents and more than 10 million hyperlinks. There are 50 topics defined for the topic distillation task, and the average number of relevant pages per query is 10.32. For simplicity, we use only the title of each topic as query.

We use the widely-used Lemur toolkit[4] to index all the DOTGOV documents, and then retrieve 1000 web pages for each topic. We use a very simple TFIDF retrieval method implemented in Lemur and it can get a reasonable relevance baseline with MAP of 0.1261 for later use. The parameters for the TFIDF method is set as follows: k1=2.6, b=0.9. Based on this retrieval method, each web page d_i is associated with a relevance value $Rel(d_i)$ with the given topic. The values exported by Lemur are usually larger than 10 and less than 100. In the experiments, we used the most popular MAP metric for evaluation.

4.2 Weighted PageRank Combination

We now want to compute a PageRank value for each web page by considering the estimated link strength, and then combine the PageRank value with the relevance value to obtain the final ranking values.

In order to incorporate the estimated link strength into the PageRank algorithm, we adopt the weighted PageRank algorithm by assigning the link strength to each edge in the graph. Formally, given a page set D, let $G=(V, E)$ be a directed graph to reflect the link relations

[2] http://www.webkit.org/

[3] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[4] http://www.lemurproject.org/

between web pages. V is the set of vertices and each vertex d_i in V is a web page in D. E is the set of edges. Each directed edge e_{ij} in E is associated with a weight $Strength(e_{ij})$. The weight is reflecting the strength degree of the link from d_i to d_j. We use a matrix M to describe G with each entry corresponding to the weight of an edge in the graph. $M = (M_{i,j})_{|V| \times |V|}$ is defined as $M_{i,j} = Strength(e_{ij})$. Then M is normalized to $\tilde{M}$ to make the sum of each row equal to 1.

Based on the transition matrix $\tilde{M}$, the weighted PageRank score $PR(d_i)$ for web page d_i can be deduced from those of all pages linking to it and it can be formulated in a recursive form as in the PageRank algorithm:

$$PR(d_i) = \mu \cdot \sum_{j \in InLink(i)} PR(d_j) \cdot \tilde{M}_{j,i} + \frac{(1-\mu)}{|V|}$$

where μ is the damping factor usually set to 0.85, as in the PageRank algorithm. *InLink(i)* means the set of web pages linking to page d_i.

Specifically, we adopt the graphchi [5] toolkit for computing the PageRank and Weighted PageRank values on a very large graph consisting of all the web pages. The toolkit can quickly compute the values on large graphs. The PageRank values exported by the toolkit is scaled and they ranges from 0 to several thousand, for example, the weighted PageRank value for G00-05-4074066(#www.usgs.gov/) is 2452.59.

Till now, we have obtained the weighted PageRank value and the relevance value of each web page, and we then combine the two values to get the final ranking value of each web page.

$$RankScore(d_i) = Rel(d_i)^{1-\lambda} \times PR(d_i)^{\lambda}$$

where $\lambda \in [0,1]$ is a parameter controlling the relative influences of the relevance factor and the authority factor. We tuned λ on the TREC2002 dataset and set λ to 0.05.

4.3 Preliminary Results

The comparison results are presented in Table 2. The methods in the table are described as follows:

Relevance Baseline: It rank the documents only by the relevance values.

PageRank Combination: It combines the relevance value and the basic PageRank value in the same way as our method. The basic PageRank values are computed without considering the strength degrees of web links.

Best TREC Run(csiro): It is the submitted run with the best MAP value. Documents are scored via a linear combination of link indegree, anchortext propagation, URL Length and BM25.

Weighted PageRank Combination: It makes use of the weighted PageRank value for combination and this is our proposed method.

As we expect, the use of PageRank is beneficial for the topic distillation task, because the PageRank combination method outperforms the Relevance baseline. Moreover, our proposed weighted PageRank combination method can outperform the PageRank combination method and the performance improvement is statistically significant with p-value<0.05 for t-test, which demonstrates the usefulness of the link strength for topic distillation. The proposed method also outperform the best submitted run on TREC 2003. In all, our proposed method has achieved promising performance.

Table 2. Comparison results on TREC 2003

	MAP
Relevance Baseline	0.1261
PageRank Combination	0.1498
Best TREC Run (csiro)	0.1543
Weighted PageRank Combination	**0.1596**

5. CONCLUSIONS AND FUTURE WORK

In this paper, we investigate how to automatically estimate the strength of a given web link. We solve the problem by using support vector regression with a few useful features, and experimental results on a manually labeled dataset show the efficacy of the regression method. We further incorporate the estimated link strengths into the weighted PageRank algorithm for improving the topic distillation results.

In future work, we will incorporate the link strength into other web search tasks to further validate its usefulness. We will also employ more features to improve the link strength estimation results.

6. REFERENCES

[1] Page, S. Brin, R. Motwani, and T. Winograd, The PageRank citation ranking: Bringing order to the web, *Technical report*, Stanford University, Stanford, CA, 1998.

[2] J. Kleinberg, Authoritative sources in a hyperlinked environment, *Journal of the ACM*, Vol. 46, No. 5, pp. 604-622, 1999.

[3] Cai, D., He, X., Wen, J. R., & Ma, W. Y. Block-level link analysis. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2004.

[4] Xue, G. R., Yang, Q., Zeng, H. J., Yu, Y., & Chen, Z. Exploiting the hierarchical structure for link analysis. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2005.

[5] Manaskasemsak, B., Rungsawang, A., & Yamana, H. Time-weighted web authoritative ranking. *Information Retrieval*, 14(2), 133-157, 2011.

[6] Wan, X., & Liu, F. Are all literature citations equally important? Automatic citation strength estimation and its applications. *Journal of the Association for Information Science and Technology*, 65(9), 1929-1938, 2014.

[7] Craswell, N., Hawking, D., Wilkinson, R., & Wu, M. Overview of the TREC 2003 Web Track. In *TREC* (Vol. 3, p. 12th), 2003.

[8] Nie, L., Davison, B. D., & Qi, X. Topical link analysis for web search. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 91-98). ACM, 2006.

[9] Zhang, Y., Wang, Y., Bing, L., & Zhang, Y. Weighting Links Using Lexical and Positional Analysis in Web Ranking. In *WAIM'08* (pp. 9-16). IEEE, 2008.

[5] https://dato.com/products/create/open_source.html