A DOMAIN-RELATED AUTHORITY MODEL FOR WEB PAGES BASED ON SOURCE AND RELATED INFORMATION

Liu Yang, Chunping Li and Ming Gu
Key Laboratory for Information System Security, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology, School of Software
Tsinghua University, Beijing, China

Keywords: Authority Model, Link Analysis, Source Information, Related Information, Finance.

Abstract: The Internet has become a great source for searching and acquiring information, while the authority of the resources is difficult to evaluate. In this paper we propose a domain-related authority model which aims to calculate the authority of web pages in a specific domain using the source and related information. These two factors, together with link structure, are what we mainly consider in our model. We also add the domain knowledge to adapt to the characteristics of the domain. Experiments on the finance domain show that our model is able to provide good authority scores and ranks for web pages and is helpful for people to better understand the pages.

1 INTRODUCTION

The Internet is playing an important role in our daily lives. People now can easily search and acquire almost everything on the web, for example, news, movies and pictures. With the numerous information and resources we are provided by the Internet, it is convenient to learn better about the world around us. However, sometimes we may find it difficult to judge the importance and authority of web pages. Among millions of web pages the users obtain from the Internet, it is easy to get lost.

PageRank (Brin and Page, 1998) and HITS (Kleinberg, 1999) are two algorithms which both provide good ways to evaluate web pages using link analysis. Following the idea of these two algorithms, there are some other algorithms proposed afterwards, and most of them also apply the idea of link analysis, e.g. PHITS (Cohn and Chang, 2000), and SALSA (Lempel and Morgan, 2001). With the help of these algorithms, people will know better about the importance of web pages.

However, in some practical situations, more specific and targeted evaluations on web pages would be better. The rankings of the traditional evaluations are general, and not limited to a specific domain. But pages considered important in some subject domains may not be considered important in others (Bharat and Mihaila, 2001). If the evaluation of web pages adapts to the domain which the pages belong to, it would be much better for users. For example, in finance domain, the web pages mostly have the characteristics of the domain, and their readers are limited to a certain group. Therefore they may not have good scores calculated by PageRank and HITS. But within the related domain, they may be important and authoritative. Since the scores given by normal link analysis are not appropriate for these pages, we should find out another way.

The goal of our research is to design and implement a domain-related authority model for web pages which is able to evaluate their authorities. Our authority model is based on link analysis and other characteristics of web pages, i.e. the source and related information, and will be applied to a specific domain. It will help the users to decide whether a web page is trustworthy. Because of the variety and complexity of web pages and the particularities of different domains, it is necessary to analyze the characteristics of the Internet and design a model which meets the domain-specific requirements.

The rest of the paper is organized as follows. In Section 2 we introduce some related work on the ranking of web pages. Our authority model based on the source and related information is presented in Section 3, followed by experiments and evaluations

Yang L., Li C. and Gu M. (2010).
A DOMAIN-RELATED AUTHORITY MODEL FOR WEB PAGES BASED ON SOURCE AND RELATED INFORMATION.
In *Proceedings of the 5th International Conference on Software and Data Technologies*, pages 245-253
DOI: 10.5220/0002998702450253

in finance domain described in Section 4. Finally we make our conclusions in Section 5.

2 RELATED WORK

Link analysis is introduced by PageRank and HITS, where hyperlink structures are used to determine the relative authority of web pages and produce improved algorithms for the ranking of web search results (Borodin et al., 2005). The existing algorithms can be divided into three classes, i.e. the algorithms based on Random Walk, the hub and authority framework and the probabilistic model.

Since first proposed in 1998, PageRank has been improved in these years. Eiron et al. (2004) refine the basic paradigm to take into account several evolving prominent features of the web, and propose several algorithmic innovations. The mathematical analysis of PageRank when the damping factor α changes is given in (Boldi et al., 2005), and an approach to accelerate the iterating computation of PageRank is proposed in (McSherry, 2005).

The hub and authority framework is proposed by Kleinberg in HITS, and used a lot in other similar algorithms, in which the framework is improved and combined with other information. For instance, in (Borodin et al., 2001) the authors introduce the Hub-Averaging-Kleinberg, Threshold-Kleinberg, and Breadth-First-Search based on the framework.

PHITS (Cohn and Chang, 2000) is a statistical hubs and authorities algorithm, and a joint probabilistic model of document content and hyperlink connectivity is suggested by Cohn and Hofmann (2000). An alternative algorithm, SALSA (Lempel and Morgan, 2001) combines ideas from both HITS and PageRank. A Bayesian algorithm is also introduced in (Borodin et al., 2001).

All the algorithms above use hyperlink structures to calculate the authority of web pages, and have good experiment results. Brian Amento points out in (Amento et al., 2000) that the result of link analysis algorithms is consistent with that of human experts, and there are no significant differences between different types of link analysis algorithms. However, none of the algorithms is perfect for all kinds of situations. In experiments people find out that different algorithms emerge as the “best” for different queries, while there are queries for which no algorithm seems to perform well (Borodin et al., 2001).

TruthFinder (Yin et al., 2008) is another algorithm which studies how to find true facts from a large amount of conflicting information on many subjects that is provided by various websites. This algorithm utilizes the relationship between websites and their information, and finds true facts among conflicting information and identifies trustworthy websites better than the popular search engines. The idea of TruthFinder is similar to link analysis algorithms, but the goal of the algorithm is different.

The algorithms mentioned above all provide the general evaluations on web pages, which are good for general requirements. Besides that, there are some topic-based algorithms coming up in recent years, whose rankings are more specific than those of the traditional algorithms. Topic-Sensitive PageRank (Haveliwala, 2002) is evolved from the traditional PageRank. It calculates a vector for every web page based on several topics. A score vector for each page is also applied in (Nie et al., 2006) to distinguish the contribution from different topics, using a random walk model that probabilistically combines page topic distribution and link structure. In these algorithms the score of a web page is not a single value, but a vector with regard to different topics. These researches consider the contents of web pages and provide the rankings on different topics. With the result of these researches, people are capable of knowing the importance of web pages in different areas. However, the rankings rely on the partition of topics in the algorithms. In our authority model, we aim to provide the domain-related ranking. To achieve this, the domain knowledge is added to our model, instead of partitioning the web pages into several topics. In this way, our authority model is able to be applied to different domains and provide the specific and targeted evaluations.

3 THE AUTHORITY MODEL FOR WEB PAGES

With the development of the Internet, people are provided with various information and resources. It is quite easy for us to acquire the information on the web. However, due to the variety of web pages, it is difficult for people to judge the quality of web pages, and decide whether to trust what the pages say. Link analysis is able to give the authority score for every web page. Following the idea of this technique, we can design and implement an authority model which aims to calculate the authority of web pages in a specific domain. In this way, users will know better about the trustworthiness of web pages when they browse them.

There are two aspects which are taken into account, the source information and related pages, which will be combined with link structure in our model, as shown in Figure 1. These two aspects are analyzed and extracted from the contents of web pages, and able to reflect the authority of web pages. The idea of our model is to calculate the authority of web pages by combining the information extracted from the contents with link analysis. In this way, our model will be more reasonable and effective. Details are described in the following sections.

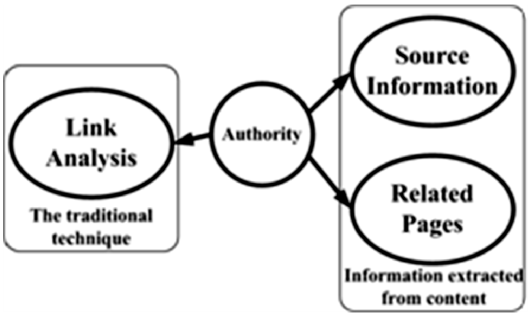

Figure 1: Aspects considered in our model.

3.1 Source Information

The authority of web pages partly lies in the source that releases the information. When we read web pages or newspapers, we will usually notice who the author is and judge the quality of the resource accordingly. Therefore, the source information is important to our authority model, and it accords with people's intuition.

The source information can often be found in the content of the web page, especially for news pages, which is denoted in Figure 2. Usually most of news web pages provide the name and the website address of their sources, which is convenient for our model to automatically extract them from the contents of web pages. For those pages whose sources are unable to extract, their authorities should be low from a human perspective. Therefore source information is not considered for these pages during the calculation of our model.

After extracting the source information from web pages, we need to find a way to obtain the importance score for every source. In this paper, we employ the traffic rank provided by Alexa (http://www.alexa.com/) as the importance scores for sources. Alexa is a website which provides the rankings for websites all over the world. The rankings in Alexa are based on the number of visits to websites, which are convenient and reasonable for source ranking. As a matter of fact, the reason why we choose Alexa is that the ranking data of websites is easy to automatically obtain from it and that its rankings are able to describe the importance of sources. In addition to the source importance obtained from Alexa, we also add domain knowledge to make our source rankings more appropriate for the domain, which will be described in Section 4.1.

Figure 2: Source information in a web page.

3.2 Related Web Pages

There are relationships existing among web pages. In our authority model, the authority of a web page is not only decided by itself and its source, but also influenced by related pages. Hyperlinks are considered as the relationships among pages in link analysis, but they are not enough for our model.

The relationship we take into consideration in our model is about the relativity of web pages. Nowadays most of the news web pages provide the related information through hyperlinks in the page, as shown in Figure 3. These lists of relative hyperlinks are organized by human editors, which list the topics and news related to the current page. This is reliable for the authority model, since the information is picked up and categorized by editors who are familiar with the news and its background. Therefore, it is helpful for computing the authority of web pages. Also, the related information is extracted from a specific section of the contents based on the properties of these pages, rather than all the hyperlinks in the pages. Therefore, it is related more closely to the contents of the pages, which is more reasonable for the calculation of authority and explains the difference between related pages and normal hyperlinks.

By automatically extracting the related information from web pages, we can utilize it in our model and calculate the authorities of web pages with the useful knowledge.

Figure 3: The related information in a web page from Yahoo!.

3.3 The Authority Model

There are two objects considered in our model, web pages and sources. The calculation of the authority model is on the basis of the relationships between sources and web pages, and between different pages. The relationships between sources and web pages are extracted from the contents of the pages, which describe the organizations which publish the articles. Hyperlinks are considered as the relationships between pages, and related information extracted from pages helps to establish the relativity of web pages.

In our authority model, link structure, source information and related pages are combined together to compute the authorities of web pages. The calculation is performed in two steps. In Step 1 we utilize the link structure in an iteration process. At first the authorities of all pages are initialized to 1. After that, the authorities are calculated iteratively. During each iteration, the authority of a web page is updated with the sum of the authorities of the web pages which point to it and are pointed to by it. The corresponding hyperlinks are named as in-links and out-links in the following sections. Normalization is done after each iteration. Considering that the web pages which point to the current page should have more influences on its authority than those pointed to by it, we set up two parameters, α and β, to adjust the weights of these two kinds of links. The iteration stops when the authorities of all web pages converge. The calculation in each iteration is shown in Equation (1).

$$a(p) = \alpha \times \sum_{q \to p} a(q) + \beta \times \sum_{p \to r} a(r) \qquad (1)$$

In Equation (1) p, q and r denote a web page respectively, $a(p)$ represents the authority of web page p, α and β are the adjusting parameters of in-links and out-links, and $0 < \beta < \alpha \leq 1$. $q \to p$ means the web page q has a hyperlink which points to p.

During the iteration process, the authorities of web pages are computed based on the link structure. In order to gain more appropriate and reasonable scores for web pages, we need to add the source and related information. Therefore in Step 2 the importance of the source and the authorities of related pages are added to the authority of the web page. The weights of source and related pages are adjustable by changing the values of the parameters.

$$a(p) = a \times a(p) + b \times s(p) + c \times \sum_{r \in RP(p)} a(r) \qquad (2)$$

In the above equation, $s(p)$ is the importance of the source of page p, $RP(p)$ represents the set of related pages for web page p, and a, b, c are the adjusting parameters for the three elements in Equation (2), $a + b + c = 1$. How to select the values of the parameters in Equation (1) and (2) will be presented in Section 4.

In Equation (2) the source importance and the authorities of related pages are added to the original authority with different weights. In this way, the two factors which influence the authority of a web page are combined into our model.

There are three factors we take into consideration in our model, link structure, source importance and related pages. Using link structure is the idea of traditional algorithms, while source importance and related pages are the information extracted from the contents of web pages. They are all important for the authorities of web pages. Calculating only one of them is insufficient. Therefore we combine them together to form a complete evaluation of web pages. With this model, we are able to precisely evaluate the authorities of web pages, and help the users to make a better judgment.

4 EXPERIMENTS AND EVALUATIONS

The effectiveness of our authority model is evaluated on the finance domain. The web pages in the domain are about financial news and comments, and their contents are professional and limited to the specific field. In order to adapt to the characteristics of finance domain, we need to add the domain knowledge to our model. Besides that, the calculation of web page authorities is the same as introduced in Section 3.

Our experiment is based on the web pages crawled from the Internet. Pages need to be processed after crawling to extract the necessary information for our experiment. Our authority model is applied to the web pages after that. In order to better evaluate the experiment result, we use a method to partition the authorities of web pages into different ranks, and a manually annotated set is used for evaluation. The detailed description and analysis are presented below.

4.1 Adding the Domain Knowledge

In our experiment, the authority model is applied to the finance domain. Therefore, the domain knowledge is quite necessary to judge the authority of web pages. The method of adding the domain knowledge to our model is mainly to adjust the importance of sources according to the features of the domain. In Section 3.1 we introduced our method of getting importance scores from Alexa, which are the general rankings on the basis of daily visits to websites. However, the area of finance has its own characteristics, which cannot be obtained simply from Alexa. For example, China Stock is a famous and professional website on finance in China, but its importance in Alexa is not ranked highly. Due to the specialty of financial websites and their limits of scopes, the websites usually do not have many visits, and their visitors are people who are interested in finance and have the background knowledge, rather than the normal Internet users. Hence, we may find those professional websites to be ranked lowly in Alexa, which should not represent their real rankings.

Therefore, adding the domain knowledge to our previous rankings is necessary for calculating the authorities of financial web pages. We find some resources about the rankings of Chinese finance newspapers, periodicals and websites. Based on the resources and the opinions of some domain experts, the importance of some sources is adjusted, i.e. the scores of some professional and important financial websites are increased, the less important financial websites are re-ranked lowly, and the scores of some well-known portals are decreased, since their main scopes are not finance. With the process of adjustment, we are able to build a database for source rankings. Moreover, more sources will be added to the database with the use of our model. Consequently, the database will contain more and more information about sources in the domain. This is useful knowledge for the authority calculation and can be reused in the future. Therefore the effort of adjustment is quite worthy. Through the adaptation to finance domain, the importance of sources accords more with the real situation within the domain, with which we will acquire more accurate result in our experiment.

4.2 Data Collection and Preprocessing

The process of data collection and preprocessing obtains the necessary information for our authority model, which includes link structure, source information and related information.

The process of getting link structure includes web page crawling, hyperlink extraction, hyperlink filtering and link relationship establishment. The web pages used in our experiments are crawled from Sina Finance (http://finance.sina.com.cn/), which contains thousands of financial news at home and abroad. These pages form the original set for our experiment. After the pages are downloaded from the Internet, their contents are analyzed, and the hyperlinks in them are extracted. In order to limit the web pages to the finance domain and research the relationships of financial pages, a filtering process is done after hyperlink extraction, which restricts the hyperlinks to Sina Finance and removes advertisement and navigation hyperlinks. In this way we make sure that all the web pages left are about finance. With the hyperlink lists of the original set, the corresponding pages of out-links are added. The in-links that point to the original set are also taken into consideration. These in-links are extracted from Site Explorer of Yahoo! (http://siteexplorer.search.yahoo.com/), and during the extraction, the number of in-links for every web page is limited to 50. For the new added pages of in-links and out-links, the link relationships among them are also established to completely form the link structure for all the web pages.

Besides that, the source information of web pages is also extracted from pages, and the importance scores are obtained from Alexa and then normalized. Then the process of adjustment is done to source importance to add the domain knowledge. Related hyperlinks in the web pages are also picked up and the corresponding relationships are established. The process of data collection and

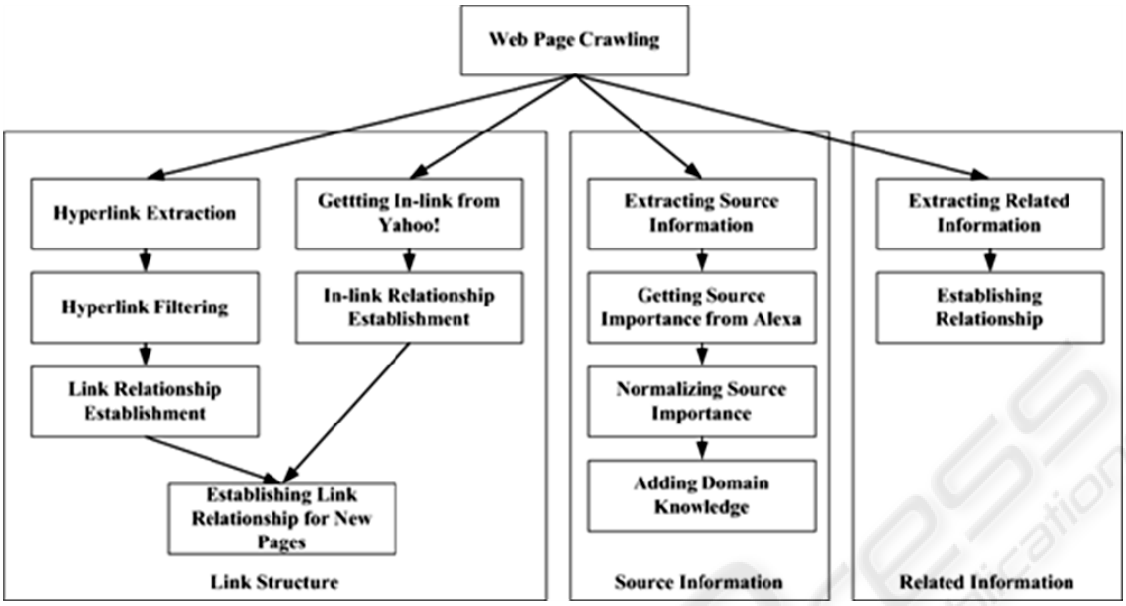

Figure 4: The process of data collection and preprocessing.

preprocessing is shown in Figure 4.

In our experiment we crawled 581 web pages from Sina Finance at the beginning, after establishing link relationships for these pages, the total number of web pages raises to 22558, and there are 860557 link relationships among the pages.

After data collection and preprocessing, our model is applied to the web pages. The iteration of Equation (1) converges after computing 19 times. In the authority model the parameters are set as follows. For Equation (1) α is 1 and β is 0.5. And in Equation (2) a is 0.6, b is 0.3 and c is 0.1. The analysis on choosing the values of parameters is described in Section 4.5.

4.3 Partitioning Authorities into Different Ranks

In practical application, users tend to accept and favor a rank level for each web page, which is easy and convenient for them to judge the pages. Partitioning the authorities of web pages into different ranks is also helpful for us to observe the distribution of the authorities in the result, and better evaluate the effect of our model.

We partition the experiment result into three different ranks according to the authorities of web pages. The first rank represents the very important and authoritative web pages, pages which belong to rank 2 are ordinarily authoritative, and pages in rank 3 are the least important. The partition uses a method which is similar to k-means (Lloyd, 1982). In the method, each authority of web pages is assigned to the rank whose center is the nearest to it during each iteration. The center of a rank is represented by the average value of the authorities of all the pages in that rank. The iteration stops when the ranks of all web pages no longer change. This method is able to adapt to the distribution of the authorities of web pages and partition the web pages into different ranks. With the method, we are able to better analyze and evaluate our result.

4.4 Evaluation Criterion and Result

We use a manually annotated set as a standard to evaluate the experiment result of the authority model. We randomly select 250 web pages from our collection, and ask three persons with finance knowledge to score these pages independently. The scores range from zero to 5.0. Higher scores indicate more authoritative pages. By averaging them we get the final score for each page. To compare our result with the manual annotation, we need to map the scores to the three ranks. The mapping between scores and ranks is shown in Table 1, in which *s* represents the score for a web page. The mapping is based on the distribution of the scores in the annotated set.

Table 1: Mapping between scores and ranks.

Score	Rank
2.4 <= s <= 5	1
1.2 <= s < 2.4	2
0 <= s < 1.2	3

Next, we need to compare the ranks given by our authority model with manual annotation. We calculate the number of pages whose ranks given by our model are the same as the manual annotation, and the precision of our model is calculated as below.

$$precision = \frac{Number\ of\ correct\ pages}{N} \quad (3)$$

Number of correct pages is the number of web pages whose ranks given by our model are equal to the manual annotation, and *N* is the cardinality of the annotated set.

Another evaluation criterion is *Average Rank Difference* (*ARD*), which is calculated using the equation below.

$$ARD = \frac{\sum_{i=1}^{N} |r_a(i) - r_m(i)|}{N} \quad (4)$$

In Equation (4), $r_a(i)$ represents the rank of page i given by our model, and $r_m(i)$ is the rank by manual annotation. *ARD* calculates the average difference of ranks between our model and manual annotation.

With the measurement of precision and ARD, we will have an objective evaluation of the authority model. After calculating the scores of the pages in the annotated set, mapping them to three ranks, and calculating precision and ARD, the result is as follows. *Precision* is 82.8%, and *ARD* is 0.172. The distribution of rank difference is shown in Figure 5, in which the horizontal axis denotes the rank difference, and the vertical axis denotes the number of pages. The pages whose rank differences are 0 and 1 cover the majority, which means that the result of our model is good and acceptable.

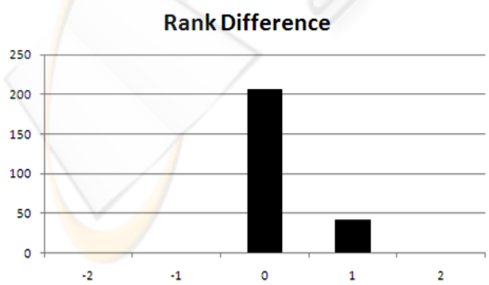

Figure 5: The rank difference distribution for the annotated set.

After analyzing the partitioning result and the corresponding pages, we have the following conclusions. The web pages which belong to rank 1 are from the important sources and have many related hyperlinks, in rank 2 the web pages are from less important sources and their related information is less too. As for rank 3, the pages do not have source information, or have less related information and fewer link relationships in them. This result is consistent with the design of our model, and reflects the influences of link structure, source and related information. These three factors are the characteristics of web pages which are easy to be extracted and quantitatively describe the authorities of web pages. By considering and combining them to our model, we are able to obtain the reasonable result in the experiment.

Also, the pages whose positions are high in the ranking list are often the financial news published by important sources. These pages usually cover the reports which most people concern. Therefore we can find more link relationships of the pages, and there are often more related pages. On the contrary, web pages which are ranked lowly are some commentary articles which express personal opinions. The ideas of these pages are subjective, thus their lower rankings given by our model are quite reasonable.

4.5 Determination of Parameters

In our authority model, there are two parameters, α and β in Equation (1), and three parameters, a, b and c in Equation (2). The values of these parameters will influence the effectiveness of our model. To compare the differences of results when using different parameters, we make several experiments on the same collection of web pages. The comparison of experiment results when different parameters are set is shown in Figure 6 and Figure 7. The experiments are made using the principle of exhaustion, with the step length 0.1. Due to the limits of pages, we only list a few results.

In the experiments, the best performance is achieved when α = 1, β = 0.5, a = 0.6, b = 0.3, c = 0.1. To achieve the best performance of our model, it is a better way to run the authority model on a small collection first, then applying the most suitable parameters to the whole collection.

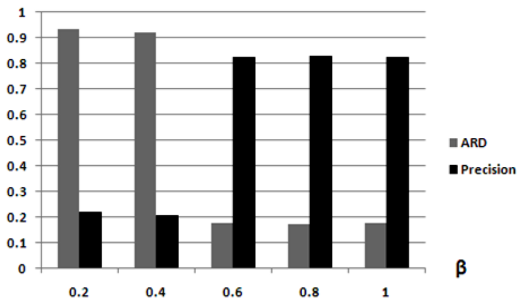

Figure 6: Comparison of experiment result. The parameter β changes from 0.2 to 1 when α = 1, a = 0.6, b = 0.3, c = 0.1.

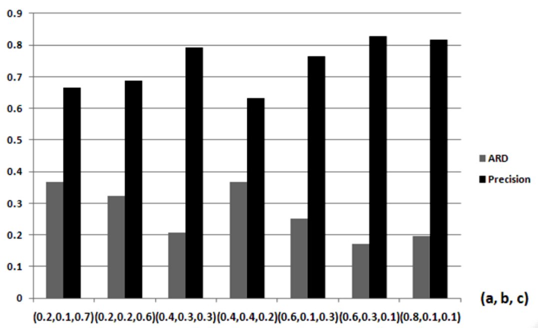

Figure 7: Comparison of experiment result. Parameters a, b and c change when α = 1, β = 0.5.

4.6 Comparing with Link Analysis

In order to illustrate the reason why we combine the three factors together into our model, we choose only the link structure and make the experiment on the same collection of web pages. When only link structure is considered in our model, it can be viewed as the traditional link analysis algorithm.

For link analysis, the *precision* is 50%, and *ARD* is 0.86. Comparing with traditional link analysis, our model is able to give more reasonable authorities and ranks. Therefore, combining source and related information in our model is important to calculate the authority of web pages in a specific domain.

With the authorities and ranks of the web pages, our model provides the users with great reference information whether to trust the pages. This is helpful for people who read the pages, and they will have a better judgment of the pages with our model.

5 CONCLUSIONS

In this paper we propose a domain-related authority model which is able to calculate the authorities of web pages in a specific domain. Three factors which will influence the authorities of web pages are taken into consideration, link structure, source information, and related pages. In order to adapt to the characteristics of the domain, we also add the domain knowledge to the model. Experiments show that our authority model is capable of providing good authority scores and ranks for web pages and facilitating people's reading experience as reference information. Compared with the traditional algorithms, the authorities of our model are more reasonable and appropriate. Therefore it has reached our expectations and met the requirement of the task. In the future we plan to extract some other characteristics of web pages with domain knowledge, and apply them to our model. We believe that with the characteristics which are able to precisely describe the authorities of web pages, our model will have a better result.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (Grant No. 90818021). We also would like to thank Naiqiao Du, Zixiao Yang and Bowen Zhang for their participation in manual annotation.

REFERENCES

Amento, B., Terveen, L., Hill, W., 2000. Does "authority" mean quality? Predicting expert quality ratings of web documents. In *23rd Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 296-303.

Bharat, K., Mihaila, G. A., 2001. When experts agree: Using non-affiliated experts to rank popular topics. In *Proceedings of the Tenth International Conference on World Wide Web*, 597-602.

Boldi, P., Santini, M., Vigna, S., 2005. PageRank as a function of the damping factor. In *Proceedings of the 14th International Conference on World Wide Web*, 557-566.

Borodin, A., Roberts, G.O., Rosenthal, J.S., Tsaparas, P., 2001. Finding authorities and hubs from link structures on the World Wide Web. In *Proceedings of the 10th International Conference on World Wide Web*, 415-429.

Borodin, A., Roberts, G.O., Rosenthal, J.S., Tsaparas, P., 2005. Link analysis ranking: algorithms, theory, and experiments. *ACM Transactions on Internet Technology*, 5(1), 231-297.

Brin, S., Page, L., 1998. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th International Conference on WWW*.

Cohn, D., Chang, H., 2000. Learning to probabilistically identify authoritative documents. In *Proceedings of the 17th International Conference on Machine Learning*. 167-174. Stanford University.

Cohn, D., Hofmann, T., 2000. The missing link - a probabilistic model of document content and hypertext connectivity. *Advances in Neural Information Processing Systems (NIPS)*, 13.

Eiron, N., McCurley, K.S., Tomlin, J.A., 2004. Ranking the web frontier. In *Proceedings of the 13th International Conference on World Wide Web*, 309-318.

Haveliwala, T.H., 2002. Topic-Sensitive PageRank. In *Proceedings of the 11th International World Wide Web Conference*, 517-526.

Kleinberg, J., 1999. Authoritative sources in a hyperlinked environment. *J. ACM*, 46.

Lempel, R., Moran, S., 2001. SALSA: The stochastic approach for link-structure analysis. *ACM Transactions on Information Systems*, 19(2), 131-160.

Lloyd, S.P., 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory*. IT-28(2), 129-137.

McSherry, F., 2005. A uniform approach to accelerated PageRank computation. In *Proceedings of the 14th International Conference on World Wide Web*, 575-582.

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

Yin, X., Han, J., Yu, P.S., 2008. Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering*, 20(6), 796-808.