Gray Tunneling Based on Joint Link for Focused Crawling

DONG Wei[1, a *], NI Hong[2,b], DENG Haojiang[3,c] and TUO Liheng[4,c]

[1] National Network New Media Engineering Research Center, Institute of Acoustics, Chinese Academy of Science, Beijing 100190, China; University of Chinese Academy of Science, Beijing 100049,China

[2] University of Chinese Academy of Science, Beijing 100049,China

[3] University of Chinese Academy of Science, Beijing 100049,China

[4] University of Chinese Academy of Science, Beijing 100049,China

[a]dongw@dsp.ac.cn, [b]nih@dsp.ac.cn, [c]denghj@dsp.ac.cn, [d]tuolh@dsp.ac.cn

Keywords: Focused Crawling, Gray Tunneling, Web Link Machine Learning, Q Learning

Abstract. Tunneling problems of the topic-multiplicity of a web page makes the relevance of the highly relevant page to be weakened. In this paper, we proposed a novel relevance prediction for focused crawling to solve gray tunneling. Our approach is based on calculating the relevancy score of web page based on its block relevancy score with respect to topics and calculating the URL score based on its parent pages and its anchor contexts, and we joins the context similarity and the link similarity which is based on Q feedback learning. Experimental results showed that the proposed method outperformed the Link-Contexts, Best-First and Breadth-First for all test data sets.

Introduction

In 2014, burgeoning OTT TV business continues rapid development, and people's TV consumption habits change further. Video websites have become an important source of family watching video programs. However, with the growth of the Web, crawling the Web quickly and entirely is an expensive, unrealistic goal because of the required hardware and network resources. Focused crawling can get popular, time-sensitive video, and also save crawling time, reduce resource usage, and improve user's interest. However, some kinds of web pages are multiple topics. It causes that portion of the web pages are related with the topic. To solve this kind problem, we proposed a novel relevance prediction for focused crawling to solve gray tunneling. Our approach is based on the context similarity and the link similarity which is based on Q feedback learning[1].

The rest of the paper is organized as follows. Section II reviews related work. Section III describes our approach to relevance prediction that is the key to the performance of a focused crawler. Section IV shows our experimental result, and finally Section V concludes our work.

Related work

Under the Web environment, most of the related pages often link to each other, but this does not mean that all the related pages are linked together. Therefore, in order to be able to find more pages which are relevant to the subject, focused crawling algorithm should have the ability to across the tunnel to get more related web pages. Dbnna et al.[2] introduced the nugget and dud concept through the tunnel. Pant and others[3] took advantage of the individual greed parameters and Best-N-First method to control the crawling to explore the process of the algorithm. Peng et al.[4] discusses the gray and black tunnels. Yohanes et al[5] proposed Genetic Algorithm to improve the recall ratio of the crawling. Xu et al.[6] made a research on a dynamic adjust crawling algorithm for guiding the topic crawler through tunnels. Zhang et al.[7] used a joint link similarity evaluation based method for crawling the resources on web. Hao et al.[8] proposed TFIDF+LSI algorithm to guide the crawling. Jianzhong Q. et al.[9] improved Best-First Search Algorithm Based Focused Crawling Research. Fu T. et al[10] took advantage of the graph to construct Sentimental spider.

PROPOSED APPROACH

1) Proposed Architecture

The proposed architecture is depicted below in Fig. 1.

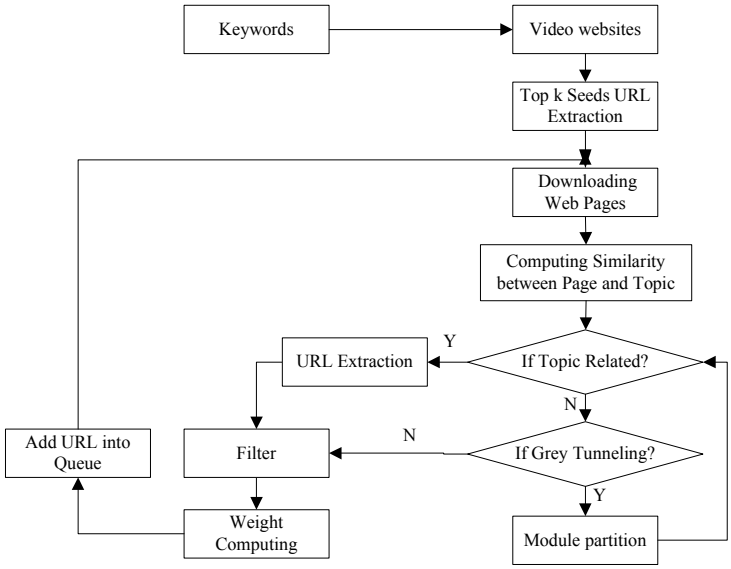

Fig 1. The Proposed Architecture

The details of our proposed approach are described below.

A. Seed URL Extraction：According to the user entering the keywords, collect the top k web pages of the video websites as the seed URLs.

B. Downloading Web Pages：We download the web pages from the Internet according to the rank in the queue. When the page has the higher priority, the page is downloaded earlier.

C. Queue：It is initialized by seed URLs. The unvisited URLs are put into the queue. A URL which has higher URLs score is given higher priority.

D. Similarity Computing and Similarity Analysis:When the web page is downloaded, extract the terms and the hyperlinks. According to OCSVM, calculate the similarity between the page and t topic.

E. Page Segmentation：In page segmentation, according to Peng et al.[3] proposed page segmentation algorithm, the web page is constructed into a DOM tree, and then segmented into several blocks.

2) Proposed Algorithm

Assume that the threshold of the similarity between the topic and the page is K, and the minimum marginal value of the similarity is $K_{\min}$. Sim is the similarity between the topic and the page. S represents the relevance and P is the path. Therefore,

$$S=\begin{cases}1 & Sim \geq K \\ 0 & K > Sim \geq K_{\min} \\ -1 & K_{\min} > Sim\end{cases} \tag{1}$$

The steps of the algorithm of the gray tunnel through are as follows:

Step1: $P=0$. If $S=1$,extract the URLs and check whether the url is not visited, and calculate the weight of the unvisited url, and put the unvisited url into the queue according to the weight.

$$Sim(URL)_1 = max\{Sim_{parent}, Sim_{anchor}\}, P(URL)=0 \tag{2}$$

Step2: If $S=0$, the page is segmented into the block, and check whether the block of the web page is topic related. Then the block related with the topic is dealt as the web page related with the topic.

$$Sim(URL)_1 = max\left\{\frac{Sim_{parent}+Sim_{block}}{2}, \frac{Sim_{parent}+Sim_{anchor}}{2}\right\}, P(URL)=P_{parent}+1(P<P_{max}) \tag{3}$$

Step3: Construct web link figure as figure 2. The dark node is the leaf node.

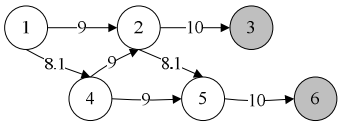

Figure 2. Web link figure

Step4: According to Markov Decision Process, any node'evaluation function $Q(u_i,<u_i,u_j>)$ in the web link figure is computed as: $Q(u_i,<u_i,u_j>)=r(u_i,<u_i,u_j>)+\gamma \max_{<u_j,u_m>\in E} Q(u_j,<u_j,u_m>)$

And finally, $Q(u_i)=\max_{<u_i,u_j>\in E} Q(u_i,<u_i,u_j>)$ (4)

In this paper, the γ adopts 0.9, and the Q will be normalized.

Step5: Final similarity is aquired based on the discuss above.

$Sim(URL)=\alpha Sim(URL)_1+\beta Q(URL)_{normalized}, \alpha+\beta=1(0<\alpha<1,0<\beta<1)$ (5)

Experiments and results

We compared our method with three other crawling methods, the conventional Link-Contexts [6] , the Best-First [11] , and Breadth-First [10].

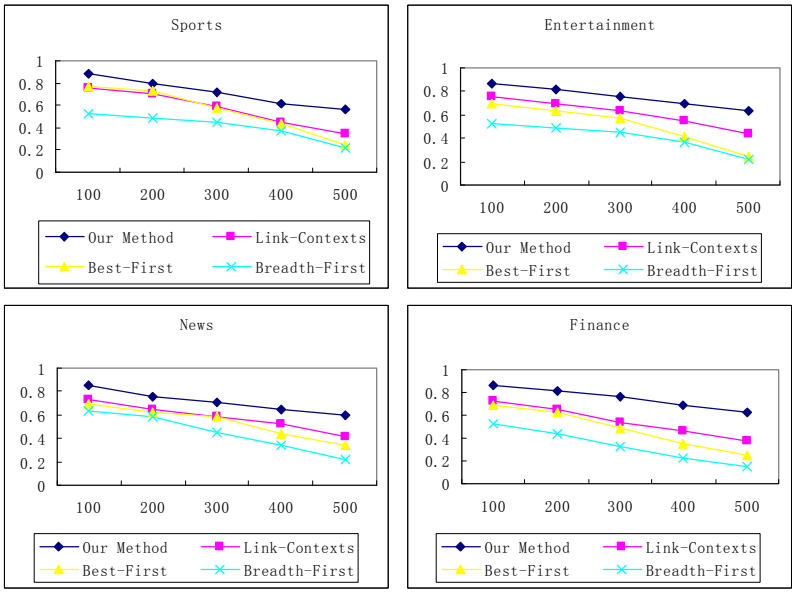

Fig.3: The precision curves on four topics

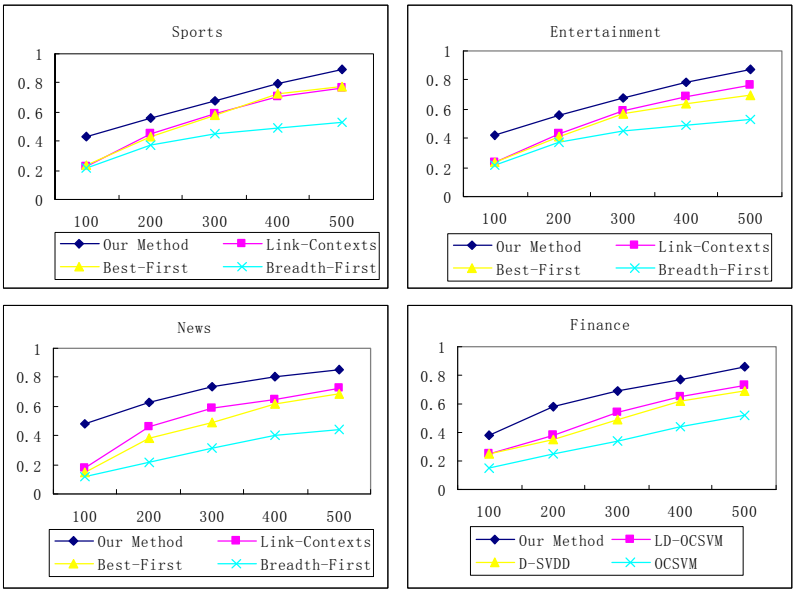

Fig.4: The recall curves on four topics

Figure 3 depicts the precision curves obtained by our method and other techniques. For these data sets, the performance of our proposed method maintains good stability. Along with the increase in data collecting, precision falls slightly. Link-Contexts and Best-First maintain high accuracy under the circumstances of acquisition of less data, while the precision falls greater than our method's along with the increase in collecting data. The precision of Link-Contexts is slightly better than Breadth-First's. The accuracy of conventional Breadth-First is much lower than three other methods.

Figure 4 depicts the recall curves on four topics. Our proposed method shows better recall than other method. The recall of Link-Contexts is close to Best-First. The performance of conventional Breadth-First is much lower than others.

Summary

In this study, we proposed a novel relevance prediction for focused crawling to solve gray tunneling. Our approach is based on calculating the relevancy score of web page based on its block relevancy score with respect to topics and calculating the URL score based on its parent pages and its anchor contexts, and we joins the context similarity and the link similarity which is based on Q feedback learing. Experimental results showed that the proposed method outperformed the Link-Contexts, Best-First and Breadth-First for all test data sets.

Thanks for the National Science and Technology Support Plan Subject (2012BAH73F02) and the Forerunner of the Chinese Academy of Sciences Subject(XDA06040501).

References

[1]Whitehead S D. Complexity and cooperation in Q-learning[C]//Proceedings of the Eighth International Workshop on Machine Learning. 2014: 363-367.

[2]Bergmark D, Lagoze C, Sbityakov A. Focused crawls, tunneling, and digital libraries[M]//Research and Advanced Technology for Digital Libraries. Springer Berlin Heidelberg, 2002: 91-106.

[3] Pant G, Srinivasan P, Menczer F. Exploration versus exploitation in topic driven crawlers[C]//WWW02 Workshop on Web Dynamics. 2002.

[4]Peng Tao, Meng Yu, Zuo Wanlil, Wang Ying, Hu Liang. Tunneling Techniques for Focused Web Crawling[J]. Journal of Computer Research and Development, 2010, 47(4)：628—637.

[5]Yohanes B W, Handoko H, Wardana H K. Focused Crawler Optimization Using Genetic Algorithm[J]. TELKOMNIKA (Telecommunication Computing Electronics and Control), 2013, 9(3): 403-410.

[6]Xu C, Jian-guo X, Bin J. Research on a dynamic adjust crawling algorithm for guiding the topic crawler through Tunnels[C]//E-Business and E-Government (ICEE), 2011 International Conference on. IEEE, 2011: 1-4.

[7]Zhang N Z, Li S J, Yu W, et al. Using a joint link similarity evaluation based method for crawling the resources on Web[J]. Jisuanji Xuebao(Chinese Journal of Computers), 2010, 33(12): 2266-2280.

[8]Hao H W, Mu C X, Yin X C, et al. An improved topic relevance algorithm for focused crawling[C]//Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on. IEEE, 2011: 850-855.

[9]Jianzhong Q. An Improved Best-First Search Algorithm Based Focused Crawling Research[J]. New Technology of Library and Information Service, 2013: Z1.

[10]Fu T, Abbasi A, Zeng D, et al. Sentimental spidering: leveraging opinion information in focused crawlers[J]. ACM Transactions on Information Systems (TOIS), 2012, 30(4): 24.