A New Recognition Approach for Logical Link Blocks in Webpages

X.M. WANG[1, 2], Z.D. WU[1], Y.N. HUANG[3], Q. GU[4, 5*]

[1]Oujiang College, Wenzhou University, Wenzhou

[2]Network Research Institute of Wenzhou, Wenzhou
, Zhejiang, 325035, China

[3]kloudSmart, Inc., 1175 Eagle Cliff Court, San Jose, CA 95120, U.S.A.

[4]School of Mathematics and Computer Science, Hubei University of Arts and Science Xiangyang Hubei 441053, China

[5]Institute of Logic and Intelligence, Southwest University, Chongqing 400715, China

gujone@163.com

Journal of Digital Information Management

ABSTRACT: *Link block is a block structure widely existing in webpages. Existing approaches to link blocks recognition generally suffer from two drawbacks: 1) they are designed only aiming at link blocks of physical structure, and even only aiming at specific link blocks of block-level elements; and 2) the discovery and recognition of link blocks are based on analyzing HTML tag trees, consequently, often leading to high computing cost and thus making them fail to deal with the diversified non-standard webpages on the Internet. To this end, in this paper we propose the concept of logical link blocks and then present an effective approach to discover and recognize logical link blocks from webpages. In the approach logical link blocks are recognized through scanning HTML codes and calculating the distance between adjacent links, and then two distance thresholds are used to determine the final logical link blocks. As a result, the approach not only can be free from the limits of specific block-level link blocks, but also can greatly improve the robustness as the analysis on HTML tag trees is no longer required. Finally, experimental results demonstrate the effectiveness of the proposed approach, which not only provide a new way for the recognition of logical link blocks and text extraction, but also can be applied in other web information processing and mining fields due to less demanding for particle size control of link blocks.*

Categories and Subject Descriptors

I.2.7 [Natural Language Processing]: Text Analysis; **I.7 [Document and Text Processing]**

General Terms:
Information Extraction, Experimentation

Keywords: Web, Link block, Logical Link Block, Recognition

Received: 18 November 2014, Revised 22 December 2014, Acceted 27 December 2014

1. Introduction

World Wide Web is a super-large complex network constructed by a variety of links between webpages. Thus, links play a key role in web information organization, display, page navigation and so on. A Web crawler realizes the traversal crawl on the Internet based on links between webpages, and Internet users rely on the links between webpages to realize the cluster reading of the contents with the same topic. Links in webpages are usually organized according to different particle sizes. The finer the particle size of link blocks, the higher the topic correlation of links, i.e., with the increasing of block particle, the topic cohesion of link blocks gradually weakens. As shown in Figure 1, the page can be classified into three link blocks when the requirement of the particle size is fine. However, it would be regarded as one link block when the requirement of the particle size is not fine, and the use of the whole link block is navigation. In relevant studies of link blocks, the requirements of the fine degree of link blocks would change with different research objectives. In the analysis of link blocks, the requirement of the particle size usually is fine, for example, the link extraction of specific topic. However, in other

Figure 1. Particle size of link blocks

unspecialized studies on link blocks, the particle size requirement is not high, for example, the text extraction of webpages.

In terms of technological implementation, visual blocking usually corresponds to some block-level HTML tag elements[1] such as and <table>. Consequently, existing studies on link blocks focus mainly on this situation. However, considering the variety of web design technology and implementation, the implementation of visual blocking is not always realized by block-level HTML tags. Instead, it may be realized by inline HTML tags[1]. Therefore, the implementation model of link blocks used by designers cannot be known in advance. It needs to be based on the elaborate analysis on HTML tag attributes, which brings a lot of trouble to some automation applications based on massive web data.

The rest of this paper is organized as follows. Section 2 describes related work already done in this field. Section 3 presents an approach to realize the discrimination and recognition of logical link blocks. In Section 4, experiment results of the approach are reported and discussed.. Section 5 Concludes the paper.

2. Relevant Studies and Problems

Studies on link blocks have a long history and many webpage blocking or extraction methods have been proposed. In Literature [2], the webpage extraction methods were divided into 5 categories, i.e., 1) wrappers for content extraction, 2) template detection for extracting content, 3) content extraction using machine learning, 4) content extraction using visual cues and/or assumptions and 5) content extraction based on HTML features and/or statistics. These five features also can be applied to the blocking of webpage link blocks. Among them, the universality of wrappers for content extraction and template detection for extracting content method are poor and they usually need manual work and timely renewal and maintenance, consequently making them time-consuming and labor-consuming. Considering these factors, some researchers propose wrapper algorithms without template support or human supervision and have achieved good results[3-5]. Content extraction using machine learning needs proper training sets and appropriate features[6], so it is difficult to work without human supervision. VIPS[7] is a typical content extraction using visual cues and/or assumptions method, which owns high accuracy but its requirement of webpage analysis is too fine and the calculation is time-consuming. When it is used to deal with numerous non-standard webpages, the accuracy and robustness are difficult to be guaranteed. Moreover, if widely-used CSS[8] is adopted to control the visual presentation effects of HTML tags, relevant CSS still needs to be analyzed, which finally will lead to huge analysis tasks and lacking program robustness. Content extraction based on HTML features and/or statistics mostly are relevant to some heuristic rules [9,12,15,23] or statistical laws, whose university needs improvements.

Besides, researchers also put forward some extra methods, for example, a fuzzy-neural network[10] for the page blocking and MSS[11] page blocking method. Although relevant methods are various and have their respective features, we can conclude that relevant algorithms of link blocks recognition are basically based on HTML tag trees[12-16,21,22] or DOM[17]. Other methods are also generally based on the HTML tag trees [18,19].

Furthermore, in relevant studies of webpage blocking, some of them only specific to block-level HTML tags, for example and <table>. Due to diversity and power of table function[20], webpage layout, modification and content organization in the early stage are almost indispensable and correspondingly some literatures only consider webpages specific to table layout and fail to distinguish tables for layout and tables for content organization[21]. Specific to table-designed webpages, Son[21] distinguished and identified the two functions of tables. Experiments

demonstrated the effectiveness of the proposed method. However, limits of the processing mode only specific to tables were too large. Present webpage design basically coexists with div. Uzun[22] considered these two conditions and firstly obtained the blocking information according to div and td and then achieved good effects by combining with decision trees to generate extraction rules, especially obtaining the equivalent performance to manual rules in extraction speed. Wang[23] proposed BSU concept and based on this adopted cluster and heuristic rules to realize page information extraction, consequently, resulting in better performance compared to the results of div-based and table-based methods.

Current blocking algorithms of link blocks, especially various HTML tag tree-based methods, require webpages to obey strict standards. These strict standards include both HTML tag grammatical norms (for example, marriage relation of HTML tags) and norms of semantic design aspects (for example, if users see block-shaped contents on the browser, their corresponding codes are usually block-level tags, such as and <table>; if users see titles, their corresponding codes are usually h1 and h2 and other tags with title significance). In fact, in massive webpages, quite a few webpages don't observe HTML tag grammatical norms and semantic design norms. Although non-standard HTML tag grammars can be corrected by some existing webpage standardization programs, the accuracy is difficult to guarantee. The correcting difficulty of semantic design norms is larger. This determines that various methods based on HTML tag trees can achieve good results only in two kinds of environment such as standard web pages and nonstandard webpages with the ability of easy correction, i.e., for those haphazard web pages, they will fail to work.

Many existing relevant studies on webpage processing generally regard the corresponding code block of a block-level HTML tag as a block. This processing mode can greatly improve the processing effects of massive webpages, but in face of numerous and complicated webpages, this processing mode may bring about two consequences, i.e., misjudgment or detection error. For example, in many webpages, there are many non-block level advertisements. In the research field of webpage text extraction, these advertisement links cannot be detected according to the traditional block-level mode, shown as Figure 2.

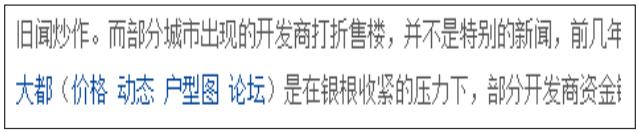

Figure 2. Non-block-level embedded advertisement links

3. Methods and Principles

For the convenience of the following expression, this paper firstly defines the following concepts. In the webpage code, there are two kinds of distances, which are respectively named code distance and text distance.

Definition 1 *(Code distance)*.The code distance between arbitrary two HTML tags refers to the length of all contents between HTML tag end mark of the former HTML tag, ">", and the HTML tag starting mark of the latter, "<". In the calculation of this paper, the attributes of each HTML tag shall be firstly removed and then the calculation of code distances can be implemented, for example, after the removal of HTML tag attributes, <A id=m>ABC</A>, we can obtain <A > ABC </A>.

Definition 2 *(Text distance)*. The text distance between arbitrary two HTML tags refers to the length of all texts between HTML tag end mark of the former HTML tag, ">", and the HTML tag starting mark of the latter, "<".

However, the calculation of text distance shall obey the following rules. (1) English and other letters take a word as a statistical unit, namely, one word length is noted as 1. (2) Chinese and other characters take one single character as a statistical unit, namely, the length of one Chinese character is noted as 1. (3) Number takes one complete digit as a statistical unit, namely, the length of one complete digit is noted as 1. For example, the length of Bei Jing 2008 is noted as 3. (4) Date string takes the whole date as a statistical unit. Namely, the length of a complete date is taken as 1. For example, the length of March 8th, 2014 is noted as 1. (5) The statistical rule of punctuation marks is the same as that of Chinese. If several adjacent punctuation marks are the same, the length is noted as 1.

Definition 3 *(Link distance)*. It refers to the distance between two adjacent links in webpages. Link distance can be measured by code distance or text distance.

Code distance: namely, the code distance between the former link "" and the latter link "<a>".

Text distance: namely, the text distance between the former link "" and the latter link "<a>".

Definition 4 *(Logical block)*. It refers to the continuous code area constituted of at least one and adjacent or nearby HTML tags. Logical block may be an HTML tag block or combined HTML tag block constituted of several adjacent or nearby HTML tag blocks and each HTML tag included in the logical block is not required complete and neither necessarily block-level HTML tag. Shown as Figure 3, A and B are two adjacent HTML tags, so they constitute a logical block. $A1$ and $A2$ both belong to the adjacent child-HTML tag of A, they also constitute a logical block. A2 and $B1$ belong to different parent HTML tags, but $A2$ and $B1$ are adjacent. Through the latter half code of A and the former half code of B, $A2$ and $B1$ can finally be a continuous code region. Thus, it is a logical block.

Definition 5 *(Logical link block)*. Suppose the number of links in one logical block is noted as C_{link}, and the distance between adjacent links in the logical block is noted as $(d_1, d_2, d_{Clink-1})$. If this logical block satisfies the

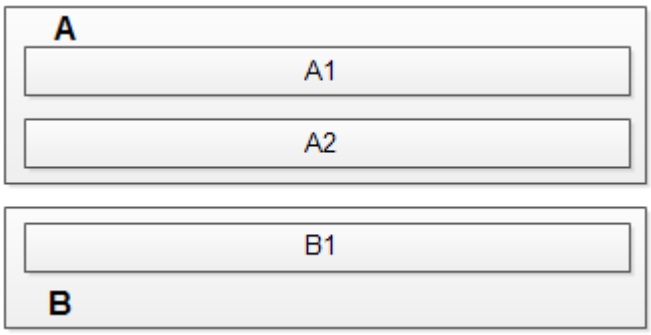

Figure 3. Logical block

following conditions, it is called a logical link block.

$C_{link} >= C_t$

$max\ (d_i) < d_t$

Where, C_t represents the minimum link number in link blocks; and d_t represents the maximum distance between adjacent links. It means that to be a logical link block, the link number shall be no less than C_t and the distance between adjacent links shall be no more than d_t.

The discovery of logical link blocks can be realized through scanning webpage codes from front to back and calculating the distance between adjacent links of the discovered links. When the distance is lower than the threshold d_t, record the link number and continue to scan afterward until the distance between adjacent links is over d_t. Judge whether the present accumulated link number exceeds C_t. If so, it indicates that the discovery of one link block ends and the discovery of the next link block starts. The advantage of this discovery approach lies in that there is no support of HTML tag trees, which means there is no need to cost massive computing resources on the analysis of HTML tag trees. This further avoids various problems of analyzing numerous and complicated codes lacking in standards.

At present, there is no available evaluation index about the recognition results of logical link blocks. This paper proposes two indexes, i.e., link coverage ratio (*LCR*) and code coverage ratio (*CCR*).

$$LCR = \frac{C_{BlockLinks}}{C_{PageLinks}},\ CCR = \frac{L_{Block}}{L_{Page}}$$

Where, $C_{BlockLinks}$ represents the total link number in logical link blocks that have been recognized; $C_{PageLinks}$ represents the total link number in the webpage; L_{Block} represents the total code length of the logical link blocks; L_{Page} represents the webpage code length.

4. Experiment Design and Result Analysis

4.1 Experimental Objectives

The following experiments aim at verifying the validity of the above-mentioned recognition approach of logical link blocks and exploring the effects and characteristics of this approach in case of processing index-type and content-type webpages.

4.2 Experimental Schemes

The original web data in the following experiments are randomly crawled from the Internet through program and two modes are adopted for sampling. 1) Manual screening. The webpage data of manual screening are from 20 well-known web portals, such as Netease, Sina, Chinanews, etc. 16 index webpages (a portal's home page) and 40 content webpages (detail page about one subject content, such as a news page, a blog page, a video display page, etc.) are selected from each website, in total 1,120 articles. 2) Random drawing. There are 184 index webpages and 1,024 content webpages that are randomly drawn, in total 1,208. Because webpage text extraction may be the most potential application of logical link blocks, different types of webpages are selected as many as possible when screening content webpages, including both long articles and short articles, both pure-text pages and pages with pictures and videos.

The experiments are divided into two groups. Each group respectively adopts code distance and text distance as the distance measurement indexes between links to test the link block recognition of index webpages and content webpages under different parameter configurations. For the convenience of the following expression, the link distance threshold based on text distance is noted as d_t^t and the link distance threshold based on code distance is noted as d_t^c. The experimental parameter configurations of two groups of experiments are as follows.

First group, suppose $C_t = 3$. In case of text distance, d_t^t = = $\{5, 10, ..., 60\}$; in case of code distance, $d_t^c = \{10, 20, ..., 120\}$. Second group, suppose $C_t = \{2, 3, ...,12\}$. In case of text distance, $d_t^t = 40$; in case of code distance, $d_t^c = 80$.

4.3 Experimental Results and Analysis

1) Influences of d_t^t on webpage link blocks

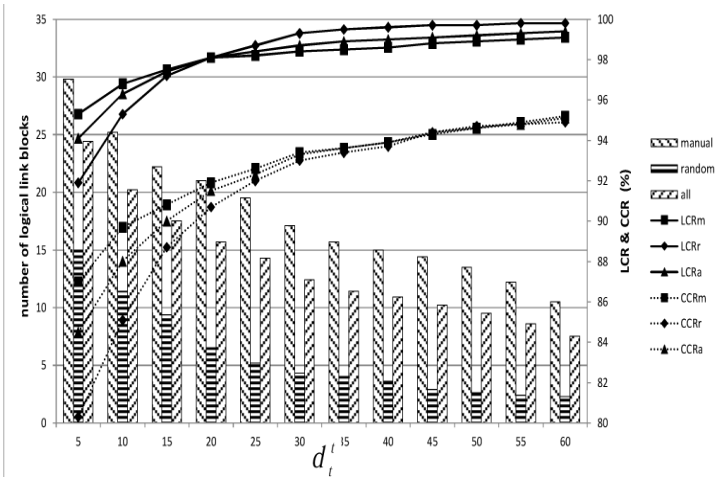

Figure 4. Influences of the link distance threshold, d_t^t, on logical link blocks-index webpages

For any webpage, it is not difficult to imagine that with the increasing distance threshold, d_t^t, the adjacent links will be easier to be involved in the same logical link block and the logical link block will be larger. In case of given total links, the number of logical link blocks will be less and correspondingly, the covered total links and code areas accumulated in each link block will be more, namely, the higher LCR and CCR will be. The experimental data in Figure 4 prove this. Among them, the subscripts, *m*, *r* and a in each figure respectively represent the manual group, the random group and the corresponding index of all data.

It can be seen from Figure 4 that, (1) index webpages contain numerous links, but pure texts in index webpages are extremely few. In no-pure text or area with extremely short text interval, all links will be involved in the same logical link block. Thus, when text distance is taken as link distance, the logical link blocks are extremely few, especially when d_t^t increases. (2) Manual sampling data are from portals with large webpage volume and complex structure. Due to abundant presented information and contents as well as complex columns, links are numerous. Most webpages in the random group are of normal size and the presented contents are relatively few and the columns are simple. Thus, the links are few. Furthermore, because webpages in the random group are relatively small and the long texts are few, the logical link blocks are obviously fewer than those in the manual group. (3) In index webpages, when $d_t^t = 5$, LCR is over 90%, indicating that the number of pure texts with the length of over 5 in the index webpages is few. These are the well-known conditions. Namely, various links are spread over index webpages, but hardly over pure texts. (4) When $d_t^t < 20$, there is a significant difference of CCR between the manual group and the random group and the difference of LCR curve is relatively small. The reasons are as follows.

When LCR is raised to a high level, isolate links or link blocks will become less. If d_t^t increases, its major function is to combine the small logical link blocks separated by some long texts into larger link blocks, instead of bringing isolate links or link blocks into logical link blocks to increase LCR. It manifests the phagocytosis effect of other codes out of links. In the merging process, logical link blocks become fewer and the original middle areas between logical link blocks are wholly brought into the new logical link blocks. Although no or few new links are involved in logical link blocks in this process, which may increase LCR, the involvement of codes in middle areas between logical blocks can significantly increase LCR. (5) Comparing LCR curve and CCR curve, it can be known that when $d_t^t \geq 20$, LCR basically maintains unchanged; while when $d_t^t \geq 45$, CCR will remain unchanged. This means that in index webpages, when $20 \leq d_t^t \leq 45$, the major contribution caused by the increase of d_t^t manifests on the phagocytosis effect of non-link codes; when $d_t^t \leq 25$, the increase of d_t^t can synchronously swallow links and codes between links, further manifesting the synchronous rise of LCR and CCR. (6) Comparatively speaking, the logical link blocks in the random group are easier to be influenced by d_t^t. The major reasons lie in that firstly the webpage links in the random group are few on the whole, usually falling between dozens and hundreds while portal webpages in the manual group usually contain thousands of links. Smaller cardinality of links makes it easier to be influenced. Secondly, pure texts in webpages in the random group are extremely few, the increase of d_t^t can rapidly cluster originally small logical link blocks into larger ones. Thus, logical link blocks are decreased dramatically, leading to the fluctuation of logical link blocks in the random group is more obvious.

Compared with index webpages, the experimental results of content webpages show a significant difference. (1) The number of logical link blocks significantly decreases, which is mainly caused by different functions of content webpages and index webpages. Index webpages undertakes the navigation function, including more links as much as possible to bear abundant information as much as possible. While content webpages focus on contents on one topic, which may be text, picture or video. These topic elements occupy numerous space, so the

number of links decreases dramatically, furthering leading to the number of the final logical link blocks decreasing dramatically. When d_t^t is large enough, the number of logical link blocks in webpages basically maintain between 2 to 3, among which mostly are 2. Namely, links before and after webpage body texts are respectively divided into one logical link block and in total there are 2 logical link blocks. Due to this feature, this method provides one new means for the extraction of webpage body texts. However, when dealing with Wikipedia, whose body texts are distributed various normal hyperlinks, extra judgments need to be done. In consideration of that common advertisement links are external links or realized by redirection, we can avoid misjudgment by judging whether hyperlinks direct or redirect to other sites. Of course, if judgments can be made through the context of link words and grammar and semantics, accuracy rate will be higher. This issue is not the focus of this paper, so here is no further explanation. (2) The difference of experimental results between the manual group and the random group is not significant. In the experimental results of index webpages, logical link blocks in random blocks are far smaller than those in the manual group. However, difference in content webpages varies slightly. Thus, it can be seen from content webpages, webpages in the manual group and the random group own similar structural features and textual features. (3) CCR significantly decreases. This mainly lies in that in content webpages, non-link blocks (for example, body texts) occupy sizable space and volumes of content webpages are far less than those of index webpages. (4) In the discovery of logical link blocks, body texts of webpages can be well preserved except pages with extremely short texts. This indicates that recognition method based on logical blocks can be applied to the text extraction of webpages. The recognition of logical link blocks needs no complex analysis on webpages and is free from the influences of irregular codes, which makes the text extraction method based on logical link blocks more robust and deserves further study. (5) For isolate links sporadically in texts of content webpages, distances between them are too far to be involved in link blocks, namely, the integrity of text blocks is free of influences. If the distances between links are short (shown as Figure 2), the method in this paper can classify them into logic link blocks and further correctly remove advertisement links embedded in texts. However, the traditional block recognition method based on block-level element cannot realize this. But if some isolate links happen to be close to an embedded advertisement area in texts, this may result in misjudgments. When d_t^t is small, the occurrence probability of this case is low. With the increase of d_t^t, the occurrence probability of this case will increase. This phenomenon needs further study and shall be avoided by more consideration or more exquisite factors. Specific to content webpages, the experiment results are shown as Figure 5.

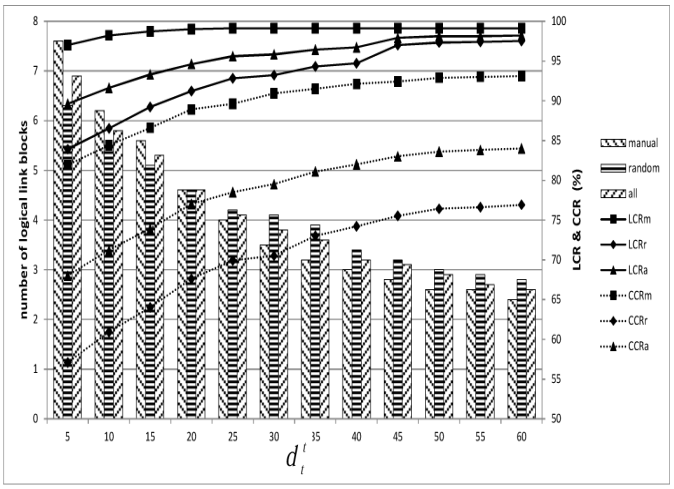

Figure 5. Influences of link distance threshold, d_t^t, on logical link blocks-content webpages

2) Influences of d_t^c on webpage link blocks

When text distance is taken as link distance, d_t^t is used to calculate texts between adjacent links. In webpages with few texts or with many texts which are dispersed into short text fragments, if the text is not long enough (namely, it cannot reach the threshold), the segmentation effect will fail to work, which may lead to adjacent links being classifying into the same logic link block. When code distance is taken as link distance, codes and texts synchronically segment logical link blocks. This means that webpages are segmented into more logical link blocks with code distance as link distance. At the same time, middle areas between link blocks also increase, which may lead to the decrease of CCR. Experiments prove that the above analysis is true. Compared with results in Figure 4, it is easy to see that in case of adopting code distance, the number of link blocks significantly increases while CCR and LCR decrease significantly. This phenomenon is especially significant when the link distance threshold is small. Results are shown as Figure 6.

It can be seen from Figure 6 that when d_t^c is small, and the number of link blocks in the data of the manual group is far more than those of the random group. With the increase of d_t^c, the difference between them gradually reduces. When $d_t^c > 90$, this difference hardly exists. This

means that in index webpages, the code distance between adjacent links basically falls in 90 whether in index webpages in portals in the manual group or regular index webpages in the random group. It is not difficult to see that the smaller d_t^c is , the finer the segmentation of webpages is. Otherwise, the larger d_t^c is, the rougher the segmentation of webpages is and the easier it is to highlight the macro structural features of webpages. Thus, whatever the size of the webpage scale is, there are certain similarities in macro structure, which is of great significance in some studies. This feature also exists in experiments specific to content webpages.

Similar to experimental results based on text distance, CCR specific to content webpages is significantly lower than that of index webpages and there is no significant difference in other aspects.

3) Influences of C_t on webpage link blocks-text distance

In the recognition process of logical link blocks, the link number threshold C_t determines the minimum link number of one logical link block for a logical block. Under the condition with determined d_t (d_t^c and d_t^t), the smaller C_t is, the easier it is for each logical block to satisfy the threshold conditions and become link blocks in the scanning and discovery process of logical link blocks and the more the total links that are involved in each link block are. The involvement of numerous links shall absorb more codes between links correspondingly. The corresponding LCR curve and CCR curve respectively reflect high LCR and high CCR. On the contrary, the larger of C_t , the more difficult to define a logical block as link blocks with the limit of the link number.

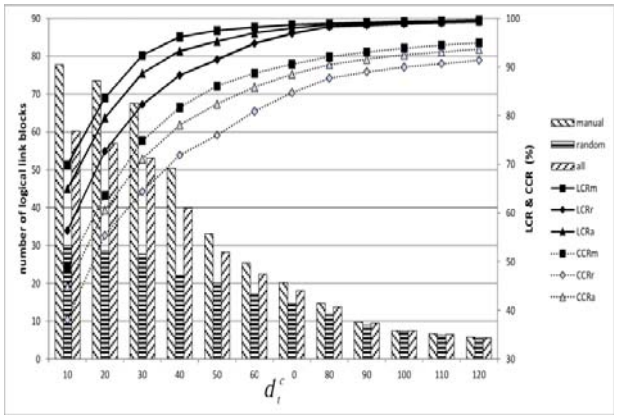

Figure 6. Influences of link distance threshold, d_t^c, on logical link blocks-index webpages

A logical block may contain a large number of links, but if the number of links is lower than C_t, it still cannot be regarded as a logic link block. Thus, more links fail to be classified into logical link blocks. Correspondingly, LCR and low CCR are low. At the same time, because many quasi-link blocks are abandoned, the number of the total logical link blocks decreases. The experimental results prove the above conclusion, shown as Figure 7.

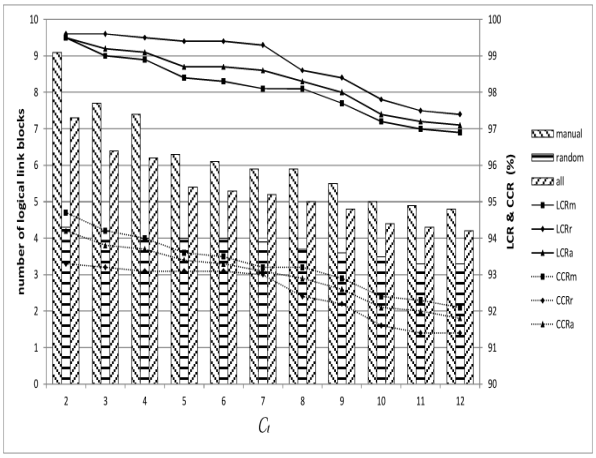

Figure 7. Influences of link distance threshold, C_t , on logical link blocks-index webpages

It can be seen from Figure 7 that logical link blocks in the manual group are significantly influenced by C_t while logical link blocks in the random group do not change sharply. The reason lies in d_t^t = 40 in this experiment. However, most webpages in the random group seldom have pure texts with the length of over 40, which lead to the fixed segmentation points of logical link blocks in webpages no matter what C_t is. In other words, few pure texts with length of over 40 act as segmentation point. It is not difficult to deduce that if d_t^t is small, this kind of pure texts will gradually increase and logical link blocks also present a large fluctuation. The experimental results

have proved this deduction.

Comparing the experimental results of content webpages and index webpages, major differences mainly manifest on, (1) few logical link blocks, basically below 4. This is mainly because pure texts in content webpages almost are clustered. Sometimes isolated hyperlinks occur in texts, but they may be too far to be involved with other link blocks. This exactly maintains the completeness of text blocks. Again, this demonstrates that text extraction method based on logical link blocks is feasible. If d_t^t keeps increasing, it may result in some short text blocks being classified into link blocks. (2) Logical link blocks present flat slide with the increase of C_t and there is no sharp fluctuation of link blocks in case of small C_t. The reason lies in the fixed number and positions of long texts in content webpages. In case of large and determined d_t^t, however C_t changes, it is long texts that play a key role in the segmentation of logical link blocks. The clustered long texts in content webpages determine that when d_t^t is large enough, most content webpages are classified into two link blocks, one link block before a body text and one link block after a body text. This conclusion has been proved in the experiment. (3) When C_t is small, LCR of content webpages is basically flat to LCR of index webpages. However, with the increase of C_t, the difference between LCR of content webpages and LCR of index webpages gradually increases. The reasons are as follows. In index webpages, link distribution is relatively intensive and uniform; while in content webpages, links are scattered. For example, links present scattered distribution in pages with few links in a body text, especially in blog pages and forum pages with some comments. When C_t is small, if scattered links are not far from each other or can occur in cluster form (typically are links about poster information around each reply in blog or forum), they are still regarded as logical link blocks. With the increase of C_t, more and more small link areas in cluster are excluded out of link blocks due to not satisfying the requirement of the minimum link threshold C_t and the adjacent link cluster also may be excluded due to the cutting-off of some texts. This condition is very few in index webpages. (4) CCR is far lower than index webpages. The essential reasons are large-length topic text blocks in content webpages. These text blocks basically are not involved in logical link blocks, which leads to significantly lower CCR of content webpages than those of index webpages. (5) LCR in the manual group is significantly higher than that in the random group. The major reason is stated as (3) that some longer posted articles in blog pages or forum pages usually have the segmentation function on pages. Due to these long posted articles, logical blocks with the link number lower than C_t cannot be classified as link blocks and further numerous links are abandoned. At last, LCR is decreased and correspondingly CCR is decreased. These phenomena can hardly be found in portal news pages in the manual group. (6) CCR in the manual group is significantly higher than that in the random group. The major reasons are as follows. First, pages in the manual group usually are longer than those in the random group. However, in terms of the length of body texts, there is no significant difference between them. According to the computation formula of CCR, it is easy to observe that this can lead to lower CCR of content webpages with short length. Second, stated as (5), the segmentation of some long posted articles can lead to lower CCR.

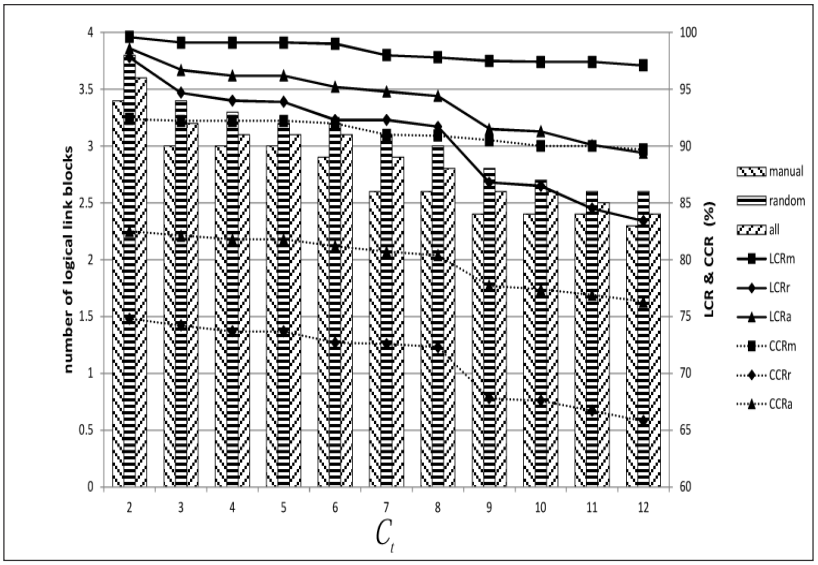

Figure 8. Influences of link distance threshold, C_t, on logical link blocks-content webpages

4) Influences of C_t on webpage link blocks-code distance In terms of the experimental results of index webpages, the differences between the approaches based on code distance and text distance mainly manifest on three aspects. (1) Link blocks based on code distance are more. Reasons are the same with the above and here is no need to repeat. (2) Lower CCR and LCR. Reasons are the same with the above and here is no need to repeat. (3)The difference in the number of logical link blocks between the random group and the manual group is not significant.

In terms of experimental results of content webpages, the differences between code distance-based method and text distance-based method are similar to those of index webpages.

Conclusion

Logical link blocks proposed in this paper extend the meaning of link blocks. The approach proposed in this paper avoids the indispensable analysis process of HTML tag trees in the traditional link block recognition, which means there is no need to cost massive computing resources on analyzing HTML tag trees and no various problems of analyzing numerous and complicated codes lacking in standards. Moreover, the discrimination rules of logical link blocks are simple and need no complex calculation. With once scanning of the webpage, the recognition of logical link blocks can be completed synchronically. The approach owns fast speed and strong anti-interference performance and it can better adapt to non-standard designed webpages. Also there is no requirement of high-link topic cohesion in link blocks. All these determine the potential application values of this method in text extraction of webpages and a bright application prospect in webpage information processing fields with lower fine particle sizes requirement of link blocks.

Acknowledgements

This work was supported by the Zhejiang Provincial Natural Science Foundation of China (LY13F010005), Wenzhou Science and Technology Project (R20130021), National Natural Science Foundation of China (61202171), the Science and Technology Support Program of Hubei Province (2014BKB068, 2014BDH124), and the Science and Technology Development Foundation of Xiangyang.

References

[1] W3C. (2014). HTML 4.01 Specification. http://www.w3.org/TR /html401/.

[2] Al-ghuribi, S M., Alshomrani, S. (2013). A comprehensive survey on web content extraction algorithms and techniques. *In*: IEEE Conference on Information Science and Applications (ICISA2013), Suwon, Korea: IEEE, p. 1–5.

[3] Wang, J F., He, X F., Wang, C., et al. (2009). News article extraction with template-independent wrapper. *In*: Proceedings of the 18th International Conference on World Wide Web, Madrid, Spain: ACM, p. 1085-1065.

[4] He, J., Gu, Y. Q., Liu, H. Y., et al. (2013). Scalable and noise tolerant web knowledge extraction for search task simplification. *Decision Support Systems*, (56)156–167.

[5] Wang, J. F, Chen, C., Wang, C., et al. (2009). Can we learn a template-independent wrapper for news article extraction from a single training site? *In*: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, Paris,France: ACM, p. 1345–1353.

[6] Peters, M., Lecocq, D. (2013). Content extraction using diverse feature sets. *In*: Proceeding WWW'13 Companion Proceedings of the 22nd International Cconference on World Wide Web companion, Republic and Canton of Geneva, Switzerland: ACM, p 89–90.

[7] Cai, D, Yu, S. P, Wen, J. R., et al. (2003). VIPS: a vision- based page segmentation algorithm, *Microsoft Technical Report*, MSR-TR-2003-79.

[8] W3C. (2014). Cascading Style Sheets (CSS) Snapshot 2010. http://www.w3.org/TR/CSS/.

[9] Xue, Y., Hu, Y., Xin, G., et al. (2007). Web page title extraction and its application. *Information Processing & Management*, 43 (5) 1332–1347.

[10] Caponetti, L., Castiello, C, Górecki, P. (2008). Document page segmentation using neuro-fuzzy approach. *Applied Soft Computing*, 8 (1)118–126.

[11] Pasternack, J., Roth, D. (2009). Extracting article text from the web with maximum subsequence segmentation. *In*: Proceedings of the 18th international conference on World Wide Web, New York, USA: ACM, p. 971–980.

[12] Ahmadi, H., Kong, J. (2012). User-centric adaptation of web information for small screens. *Journal of Visual Languages & Computing*, 23 (1)13–28.

[13] Cai, R., Yang, J. M., Lai, W., et al. (2008). iRobot: An intelligent crawler for web forums. In: *Proceedings of the 17th international conference on World Wide Web*, Beijing, China: ACM, p 447–456.

[14] Guo, Y., Tang, H. F., Song, L. H., et al. (2010). ECON: an approach to extract content from web news page. *In*: *IEEE 12th International Asia-Pacific Web Conference*, Busan, Korea: IEEE, p 314–320.

[15] Ji, X. W., Zeng, J. P, Zhang, S. Y, et al. (2010). Tag tree template for Web information and schema extraction. *Expert Systems with Applications*, 37 (12) 8492–8498.

[16] Wong, T .L, Lam, W(2009). An unsupervised method for joint information extraction and feature mining across different Web sites. *Data & Knowledge Engineering*, 68 (1)107–125.

[17] W3C. (2014). Document Object Model (DOM). http://www.w3. org/DOM/.

[18] Li, Zhiyi., Zhirui, Shen. (2013). Web Information Extraction Study Based On Natural Annotation. *Journal of the China Society for Scientific and Technical Information*, 32 (8) 853–859.(In Chinese)

[19] Álvarez, M., Pan, A., Raposo, J., et al (2008). Extracting lists of data records from semi-structured web p. *Data & Knowledge Engineering*, 64 (2) 491–509.

[20] Cafarella, M J., Halevy A., Wang D Z., et al. (2008).WebTables: exploring the power of tables on the web. *In*: Proceedings of the VLDB Endowment, Auckland, NewZealand: ACM, p 538–549.

[21] Son J-W., Park S-B. (2013). Web table discrimination with composition of rich structural and content information. *Applied Soft Computing*, 13 (1)47–57.

[22] Uzun E., Agun H V., Yerlikaya T. (2013). A hybrid approach for extracting informative content from web pages. *Information Processing & Management*, 49 (4) 928–944.

[23] Wang J Q., Chen Q C., Wang X L., et al. (2008). Basic semantic units based web page content extraction. In: *IEEE International Conference on Systems, Man and Cybernetics*,Singapore, Singapore:IEEE, p 1489–1494.