Proceedings of the 2005 IEEE
International Conference on Information Acquisition
June 27 - July 3, 2005, Hong Kong and Macau, China

Web Structure Mining: An Introduction

Miguel Gomes da Costa Júnior Zhiguo Gong
Department of Computer and information Science
Faculty of Science and Technology
University of Macau
Av. Padre Tomás, S.J., Taipa, Macao S.A.R., China
{mcosta, fstzgg}@umac.mo

Abstract - Due to the increasing amount of data available online, the World Wide Web has becoming one of the most valuable resources for information retrievals and knowledge discoveries. Web mining technologies are the right solutions for knowledge discovery on the Web. The knowledge extracted from the Web can be used to raise the performances for Web information retrievals, question answering, and Web based data warehousing. In this paper, we provide an introduction of Web mining as well as a review of the Web mining categories. Then we focus on one of these categories: the Web structure mining. Within this category, we introduce link mining and review two popular methods applied in Web structure mining: HITS and PageRank.

Index Terms – Web structure mining, web mining, link mining.

I. INTRODUCTION

Nowadays, the World Wide Web has becoming one of the most comprehensive information resources. It probably, if not always, covers the information need for any user. However, the Web demonstrates many radical differences to traditional information containers such as databases, in schema, volume, topic-coherence. Those differences make it challenging to fully use Web information in an effective and efficient manner. Web mining is right for this need [1].

In fact, Web mining can be considered as the applications of the general data mining techniques to the Web. However, the intrinsic properties of the Web make us have to tailor and extend the traditional methodologies considerably. Firstly, even though Web contains huge volume of data, it is distributed on the internet. Before mining, we need to gather the Web document together. Secondly, Web pages are semi-structured, in order for easy processing, documents should be extracted and represented into some format. Thirdly, Web information tends to be of diversity in meaning, training or testing data set should be large enough. Even though the difficulties above, the Web also provides other ways to support mining, for example, the links among Web pages are important resource to be used.

Besides the challenge to find relevant information, users could also find other difficulties when interacting with the Web such as the degree of quality of the information found, the creation of new knowledge out of the information available on the Web, personalization of the information found and learning about other users. Web mining techniques could be applied to solve, partially or completely, the above cited problems. However, Web mining techniques are not the only tools to solve those problems. Other research communities such as database, machine learning and information retrieval, are also addressing the above mentioned difficulties. This situation creates confusion to determine what forms Web mining. In this paper, we firstly provide a survey on overall Web mining concepts and technologies, then, pay special attentions on Web structure mining in detail.

This paper is structured as follows. In section 2 we provide an overview of Web mining categories. In section 3 we discuss Web structure mining and introduce Link mining. In section 4, we review two well known algorithms: HITS and PageRank. Section 5 addresses Web page segmentation methodologies. And we conclude this paper in section 6.

II. WEB MINING OVERVIEW

To clarify the confusion of what forms Web mining. Kosala and Blockeel [2] had suggested a decomposition of Web mining in the following tasks:

1. *Resource finding*: the task of retrieving intended Web documents.
2. *Information selection and pre-processing*: automatically selecting and pre-processing specific information from retrieved Web resources.
3. *Generalization:* automatically discovers general patterns at individual Web sites as well as across multiple sites.
4. *Analysis*: validation and/or interpretation of the mined patterns.

In general, Web mining tasks can be classified into three categories [2; 3]: *Web content mining, Web structure mining and Web usage mining.* However, there are two other different approaches to categorize Web mining. In both, the categories are reduced from three to two: *Web content mining* and *Web usage mining*. In one, Web structure is treated as part of Web Content [11]; while in the other, Web usage is treated as part of Web Structure [12]. All of the three categories focus on the process of knowledge discovery of implicit, previously unknown and potentially useful information from the Web. Each of them focuses on different mining objects of the Web. Fig. 1

shows the Web categories and their objects. As follows, we provide a brief introduction about each of the categories.

Web content mining targets the knowledge discovery, in which the main objects are the traditional collections of text documents and, more recently, also the collections of multimedia documents such as images, videos, audios, which are embedded in or linked to the Web pages. Web content mining could be differentiated from two points of view: the agent-based approach or the database approach. The first approach aims on improving the information finding and filtering and could be placed into the following three categories [13]:

1. *Intelligent Search Agents*. These agents search for relevant information using domain characteristics and user profiles to organize and interpret the discovered information.
2. *Information Filtering/ Categorization.* These agents use information retrieval techniques and characteristics of open hypertext Web documents to automatically retrieve, filter, and categorize them.
3. *Personalized Web Agents*. These agents learn user preferences and discover Web information based on these preferences, and preferences of other users with similar interest.

The second approach aims on modeling the data on the Web into more structured form in order to apply standard database querying mechanism and data mining applications to analyze it. The two main categories are *Multilevel databases* and *Web query systems*. For further information about Web content mining please refer to [2; 5; 12].

Web structure mining focuses on the hyperlink structure of the Web. The different objects are linked in some way. Simply applying the traditional processes and assuming that the events are independent can lead to wrong conclusions. However, the appropriate handling of the links could lead to potential correlations, and then improve the predictive accuracy of the learned models [8]. Two algorithms that have been proposed to lead with those potential correlations: HITS [14] and PageRank [10], and Web structure mining itself will be discussed in the next section.

Web usage mining focuses on techniques that could predict the behavior of users while they are interacting with the WWW. Web usage mining collects the data from Web log records to discover user access patterns of Web pages. There are several available research projects and commercial products that analyze those patterns for different purposes. The applications generated from this analysis can be classified as personalization, system improvement, site modification, business intelligence and usage characterization [3].

The challenges involved in web usage mining could be divided in three phases [11]:

1. *Pre-processing.* The data available tend to be noisy, incomplete and inconsistent. In this phase, the data available should be treated according to the requirements of the next phase. It includes data cleaning, data integration, data transformation and data reduction.
2. *Pattern discovery.* Several different methods and algorithms such as statistics, data mining, machine learning and pattern recognition could be applied to identify user patterns.
3. *Pattern Analysis.* This process targets to understand, visualize and give interpretation to these patterns.

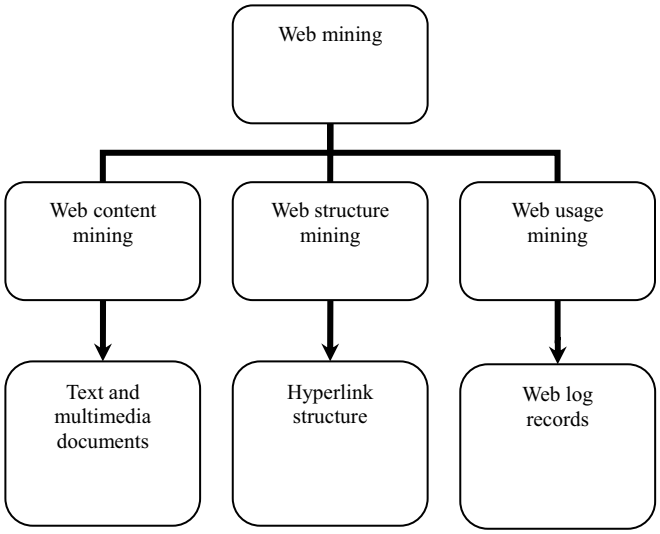

Fig. 1 Web mining categories and objects.

Web usage mining depends on the collaboration of the user to allow the access of the Web log records. Due to this dependence, privacy is becoming a new issue to Web usage mining, since users should be made aware about privacy policies before they make the decision to reveal their personal data. For further information about Web usage mining please refer to [3; 11; 13].

We should note that there is no clear boundary between the above categories. As we mentioned, the two or three category definitions are quite acceptable, showing that Web content mining, Web structure mining and Web usage mining could be used isolated or combined in an application.

III. WEB STRUCTURE MINING

The challenge for Web structure mining is to deal with the structure of the hyperlinks within the Web itself. Link analysis is an old area of research. However, with the growing interest in Web mining, the research of structure analysis had increased and these efforts had resulted in a newly emerging research area called Link Mining [8], which is located at the intersection of the work in link analysis, hypertext and web mining, relational learning and inductive logic programming, and graph mining. There is a potentially wide range of application areas for this new area of research, including Internet.

The Web contains a variety of objects with almost no unifying structure, with differences in the authoring style and content much greater than in traditional collections of text documents. The objects in the WWW are web pages, and links

are in-, out- and co-citation (two pages that are both linked to by the same page). Attributes include HTML tags, word appearances and anchor texts [8]. This diversity of objects creates new problems and challenges, since is not possible to directly made use of existing techniques such as from database management or information retrieval. Link mining had produced some agitation on some of the traditional data mining tasks. As follows, we summarize some of these possible tasks of link mining which are applicable in Web structure mining.

1. *Link-based Classification*. Link-based classification is the most recent upgrade of a classic data mining task to linked domains [7]. The task is to focus on the prediction of the category of a web page, based on words that occur on the page, links between pages, anchor text, html tags and other possible attributes found on the web page.
2. *Link-based Cluster Analysis*. The goal in cluster analysis is to find naturally occurring sub-classes. The data is segmented into groups, where similar objects are grouped together, and dissimilar objects are grouped into different groups. Different than the previous task, link-based cluster analysis is unsupervised and can be used to discover hidden patterns from data.
3. *Link Type.* There are a wide range of tasks concerning the prediction of the existence of links, such as predicting the type of link between two entities, or predicting the purpose of a link.
4. *Link Strength.* Links could be associated with weights.
5. *Link Cardinality*. The main task here is to predict the number of links between objects.

There are many ways to use the link structure of the Web to create notions of authority. The main goal in developing applications for link mining is to made good use of the understanding of these intrinsic social organization of the Web.

IV. HITS CONCEPT AND PAGERANK METHOD

In this section we review two approaches: HITS concept and PageRank method. Both approaches focus on the link structure of the Web to find the importance of the Web pages.

A. HITS: Computing Hubs and Authorities

In HITS concept, Kleinberg [14] identifies two kinds of pages from the Web hyperlink structure: authorities (pages with good sources of content) and hubs (pages with good sources of links). For a given query, HITS will find authorities and hubs.

According to Kleinberg [14], "*Hubs and authorities exhibit what could be called a mutually reinforcing relationship: a good hub is a page that points to many good authorities; a good authority is a page that is pointed to by many good hubs*". See Fig. 2. HITS associates a non-negative authority weight $x^{<p>}$ and a non-negative hub weight $y^{<p>}$. See Fig. 3. The weights of each type are normalized so that their squares sum to 1.

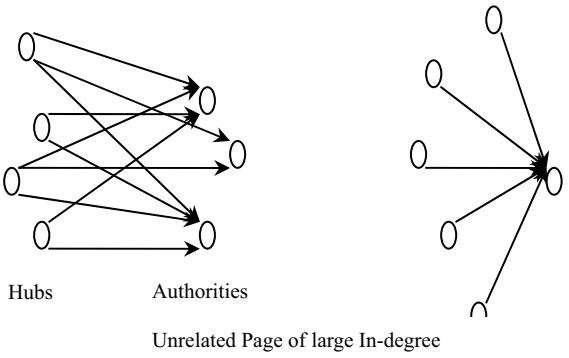

Fig. 2 A densely linked set of Hubs and Authorities (from [14])

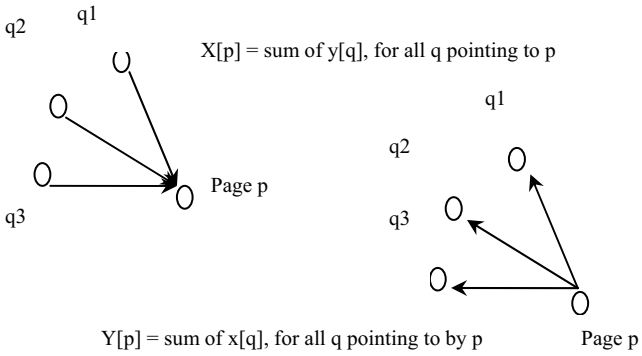

Fig. 3 The basic operations of HITS (from [14]).

According to Kleinberg [14], "Numerically the mutually reinforcing relationship can be expressed as follows: if p points to many pages with large x-values, then it should receive a large y-value; if p is pointed to by many pages with large y-values, then it should receive a large x-value. Given weights $x^{<p>}$,$y^{<p>}$, then the x-weights and y-value are as follows:" (See Fig. 4).

$$X^{<p>} \leftarrow \sum_{q:(q,p)\in E} y^{<q>} \qquad Y^{<p>} \leftarrow \sum_{q:(q,p)\in E} x^{<q>}$$

Fig. 4 X-weight and Y-weight (from [14]).

Although HITS provides good search results for a wide range of queries, HITS did not work well in all cases due to the following three reasons [5]:

1. *Mutually reinforced relationships between hosts.* Sometimes a set of documents on one host point to a single document on a second host, or sometimes a single document on one host point to a set of document on a second host. These situations could provide wrong definitions about a good hub or a good authority.
2. *Automatically generated links*. Web document generated by tools often have links that were inserted by the tool.
3. *Non-relevant nodes.* Sometimes pages point to other pages with no relevance to the query topic.

B. PageRank Model

L. Page and S. Brin [10;15] proposed the Page Rank algorithm to calculates the importance of web pages using the link structure of the web. In their approach Brin and Page extends the idea of simply counting in-links equally, by normalizing by the number of links on a page. The Page Rank algorithm is defined as [15]: "*We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor, which can be set between 0 and 1. We usually set d to 0.85. There are more details about d in the next section. Also C (A) is defined as the number of links going out of page A. The Page Rank of a page A is given as follows:*

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn)) \quad (1)$$

Note that the Page Ranks form a probability distribution over web pages, so the sum of all web pages' Page Ranks will be one." And "*The d damping factor is the probability at each page the "random surfer" will get bored and request another random page."*

Note that the rank of a page is divided evenly among its out-links to contribute to the ranks of the pages they point to. The equation is recursive, but starting with any set of ranks and iterating the computation until it converges may compute it. Page Rank can be calculated using a simple iterative algorithm, and corresponds to the principal eigen vector of the normalized link matrix of the web. Page Rank algorithm needs a few hours to calculate the rank of millions of pages [15].

C. Applications

HITS was used for the first time in the Clever [17] search engine from IBM, and PageRank is used by Google [18] combined with other several features such as anchor text, IR measures, and proximity.

The notion of authoritativeness comes from the idea that we wish not only to locate a set of relevant pages, but rather the relevant pages of the highest quality. However, the Web consists not only of pages but also of links that connect one page to another. This structure contains a large amount of information that should be exploited.

PageRank and HITS belong to a class of ranking algorithms, where the scores can be computed as a fixed point of a linear equation. Bianchini [16] noted that HITS and PageRank are used as starting points for new solutions, and there are some extensions of theses two approaches. There are other link-based approaches to be applied on the Web. For further information please refer to [3; 5; 9; 13; 14; 15; 16].

Beside being used for weighting Web pages, link resource can also be used for clustering or classifying Web pages. The principle is based on the assumption that (1) if page p_1 has a link to page p_2, p_1 should be similar to p_2 in content, and (2) if p1 and p2 are co-cited by some common pages, p_1 and p_2 should also similar. Web pages can be clustered into a lot of connected page communities with respect to their citation and co-citation strengths among the pages. In fact, Ziv Bar-Yossef and Sridhar Rajagopalan [19] put all the algorithms, which uses links, to three categories.

1. Relevant Linkage Principle: Links points to relevant resources.

2. Topical Unity Principle: Documents often co-cited are related, as are those with extensive bibliographic overlap. This idea is previous addressed by Kesseler for bibliographic information retrieval in [20].

3. Lexical Affinity Principle: Proximity of text and links within a page is a measure of the relevance of one to another.

Even though those link algorithms can always provide a good support for Web information retrievals, clustering and knowledge discoveries on the Web, authors also find problems associated with those technologies [19, 21, 22, 23].

Taher H. Haveliwala notices that original PangeRank algorithm, introduced by Page et al. [10], pre-compute ranking vector based on all the Web pages. This ranking vector is computed once and used for any queries later. The ranking is actually independent of the specific queries when using it. The authors tried to solve this limitation by computing a set of PageRank vectors, each biased with a different topic. In other words, for each topic, they assigned a weight for each page. Therefore, searches in different topics could select corresponding vectors for ranking.

Ziv Bar-Yossef and Srihar Rajagopalan [19], Deng Cai et al. [22], and Shian-Hua Lin et al. [23] addressed problems caused by topic drift of Web pages. In other words, it is easy to find a Web page with multiple topics. In such cases, two links from the same Web page may lead to different semantics if they are anchored in the areas of different topics. Or maybe some links only links to some advertisements. One effective solution for topic drift problems is page segmentation.

V. WEB PAGE SEGMENTATION

In this section, we will discuss some works on Web page segmentations.

Earlier works on document segmentation were on free texts and motivated by raising performances of the information retrieval. In information retrieval, documents are ranked with the values of similarities of the documents to the queries. One popularly used similarity model is calculated using cosine between vector of query terms and vector of document terms, as:

$$sim(q,d) = \frac{\sum w_{q,i} * w_{d,i}}{(\sum w_{q,i}^2)^{\frac{1}{2}} (\sum w_{d,i}^2)^{\frac{1}{2}}}$$

where $w_{q,i}$ and $w_{d,i}$ are the weights of term t_i in query q and document d respectively. From this metric function, it is clear that mutual affection of several topics in the same document may cause lower ranking of the document even it may contains a passage which can well covers the user's need. For this reason, documents are partitioned into a sequence of passages with respect to topics.

Text segmentation algorithms in general fall into three categories: The first one is to identify the locations of topic changes for text stream. Texts created from automatic speech recognition, newswire feeds, or television closed transcripts may contain cue-words as topic transitions [24]. However,

general texts may not contain noticeable topic change words. The second method is to partition texts with fixed-length windows. Though it is simple, such method improves retrieval performance effectively [25]. But it can not be coupled with topics in the document. The third method is based on semantic coherence of the text [26,27,28]. Such kind of methods partition texts by measuring semantic coherence between two consecutive blocks which are fine-grained text units such as sentences or even words.

Web documents are more likely to contain multiple topics than traditional free texts. Helpfully, more separators (different HTML tags) can be used in segmentations of the Web documents. In general, Web documents segmentations can be put into three categories including link-number based, visual layout based and semantic based. All the segmentation methods are on the top of DOM tree structures of the documents.

Ziv Bar-Yossef, et al. partitions a Web page into a sequence of semantic blocks called pagelet [19]. Actually, in their approach, sibling leaf nodes of the DOM tree are merged recursively until the number of overall links in the nodes exceeding certain value (3 in their work). Then, each leaf node in the transformed tree is called a pagelet.

Deng Cai et al. [22] created an algorithm called VIPS which partitions Web pages with respect to the layout structures of the Web documents. The author also combined VIPS with fix-length algorithm in order to raise the performance of information retrieval.

Shian-Hua Lin et al. [23] makes use of HTML table tags such as <TABLE>, <TD> and <TR> for the segmentations of the Web pages. Their work concentrated on Web pages with rich table tags, especially the Web pages generated by some template tools. Such tools often utilize table tags to arrange presentation structures of the Web page. And the Web pages on one site, created with such authoring tools, are often to be found homogeneous in structure, thus easy for the processing in such model.

Recently, we are trying to partition Web pages based on semantic similarities of the Web page elements. Our algorithm is carried out by merging leaf nodes of the DOM tree bottom-up recursively if the semantic coherence between two sibling blocks exceeding a threshold. Our measurement for semantic similarity is based on term co-occurrences in the collection which are extracted from our Web page database. In fact, the co-occurrence is described with two functions—support and confidence which are frequently used in data mining community.

Once Web pages are partitioned into blocks, Web page ranking and indexing can be performed on the granularity of semantic blocks other than on the whole pages. Therefore, ranking based on the information content can be improved [22, 23]. Furthermore, PageRank, HITS, and other link-based algorithms can be applied to page blocks. The basic ideas are: (1) Links from important blocks are assigned higher weights [23], (2) a block (other than page) is semantically similar to a page if there is a link anchored in this block to the page, (3) two pages are similar if there are co-cited via some common blocks (other than pages).

Deng Cai et al. [29] defined the importance factors of a block with respect to the size of the block and its position in the screen when browsing. Their experimental results reveals that block-level PageRank and HITS can improve retrieval performance significantly.

VI. CONCLUSION

In this paper we survey the research area of Web mining, focusing on the category of Web structure mining. We had introduced *Web mining*. Later in the paper when we had discussed Web structure mining, and introduced Link mining, as well as block-level link mining issues. We had also reviewed two popular algorithms to have an idea about their application and effectiveness. Since this is a huge area, and there a lot of work to do, we hope this paper could be a useful starting point for identifying opportunities for further research.

ACKNOWLEDGMENT

This work was done with the support of University Research Committee under Grant No. RG084/04-05S/GZG/FST.

REFERENCES

[1] O. Etzioni. The world wide web: Quagmire or gold mine. *Communications of the ACM*, 39(11):65-68, 1996.

[2] Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey, *ACM SIGKDD Explorations Newsletter*, June 2000, Volume 2 Issue 1.

[3] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pag-Ning Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, *ACM SIGKDD Explorations Newsletter,* January 2000, Volume 1 Issue 2.

[4] Jidong Wang, Zheng Chen, Li Tao, Wei-Ying Ma, Liu Wenyin, Ranking User's Relevance to a Topic through Link Analysis on Web Logs, *WIDM' 02*, November 2002.

[5] A. A. Barfourosh, H.R. Motahary Nezhad, M. L. Anderson, D. Perlis, Information Retrieval on the World Wide Web and Active Logic: A Survey and Problem Definition, 2002.

[6] G. Piatetsky-Shapiro, and W.J. Frawley, Knowledge Discovery in Databases. *AAAI/MIT Press*, 1991.

[7] Q. Lu, and L. Getoor. Link-based classification. In *Proceedings of ICML-03*, 2003.

[8] L. Getoor, Link Mining: A New Data Mining Challenge. *SIGKDD Explorations*, vol. 4, issue 2, 2003.

[9] S. Chakrabarti, B. E. Dom, D, Gibson, J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Mining the Link Structure of the World Wide Web. February, 1999.

[10] L. Page, S. Brin, R. Motwani, and T. Winograd. The Pagerank citation ranking: Bring order to the web. *Technical report*, Stanford University, 1998.

[11] Han, J., Kamber, M. Kamber. Data mining: concepts and techniques. Morgan Kaufmann Publishers, 2000.

[12] Wang Jicheng, Huang Yuan, Wu Gangshan, Zhang Fuyan. Web mining: knowledge discovery on the Web. *Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference* - on Volume 2, Page(s):137 - 141 vol.2 - 12-15 Oct. 1999

[13] Cooley, R.; Mobasher, B.; Srivastava, J.; Web mining: information and pattern discovery on the World Wide Web. *Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE* International Conference. Page(s):558 – 567 - 3-8 Nov. 1997.

[14] Kleinberg, J.M., Authoritative sources in a hyperlinked environment. *In Proceedings of ACM-SIAM Symposium on Discrete Algorithms, 1998*, pages 668-677 – 1998.

[15] Brin, S.; Page,L. The Anatomy of a Large-scale Hypertextual Web

Search Engine. *Proceedings of the Seventh International World Wide Web Conference*, 1998.
[16] Bianchini, M.; Gori, M.; Scarselli, F.; Inside PageRank. ACM Transaction on Internet Technology (TOIT), Volume 5 Issue 1 – February, 2005.
[17] http://www.research.ibm.com/topics/popups/innovate/hci/html/clever.html. Last accessed 15/04/2005.
[18] http://www.google.com/. Last accessed 15/04/2005.
[19] Ziv Bar-Yossef and Sridhar Rajagopalan. Template Detection via Data Mining and its Applications. In: Proceedings of WWW2002, May 7-11, 2002, Honolulu, Hawaii, USA. 580-591.
[20] M. Kessler. Bibliographic coupling between scientific papers. American Documentation, 14:10-25, 1963.
[21] Taher H. Haveliwala. Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No4, July/August 2003, 784-796.
[22] Deng Cai, Shipeng Yu, and et al. Block-based Web Search. In Proceedings of ACM SIGIR'04, July 25-29, 2004, Sheffield, South Yorkshire, UK. 465-463.
[23] Sian-Hua Lin and Jan-Ming Ho. Discovering Information Content Blocks from Web Documents. In: Proceedings of ACM SIGKDD'02, July 23-26, 2002, Edmonton, Alberta, Canada. 588-593.
[24] M. Hajime, H. Takeo, and O. Manabu. Text Segmentation with Mltiple Survace Linguistic Cues. In: Proceedings of COLING-ACL. 1998, 881-885.
[25] Callan, J. P., Passage-Level Evidence in Document Retrieval, In Proceedings of ACM SIGIR'94, Dublin, 1994. 302-310.
[26] Xiang Ji and Hongyuan Zha. Domain-Independent Text Segmentation Using Anisotropic Diffusion and Dynamic Programming. In: Proceedings of SIGIR'03, July 28-August 1, 2003, Toronto, Canada. 322-329.
[27] Marti A. Hearst. TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. Computational Linguistics, 23(1): 33-64, 1997.
[28] T. Brants, F. Chen, and I. Tsochantaridis. Topic-Based Document Segmentation with Probabilistic Latent Semantic Analysis. In: Proceedings of CIKM'02, November 4-9,2002, McLean, Virginia, USA. 211-218.
[29] Deng Cai, Xiaofei He, Ji-Rong Wen, and Wei-Ying Ma, Block-level Link Analysis, In Proceedings of ACM SIGIR'04, July 25-29, 2004, Sheffield, South Yorshire, UK. 440-447.