An Attractive Force Model for Weighting Links in Query-Dependant Web Page Ranking

Xinyue Liu[1,2]
1. School of Computer Science and Technology, Dalian University of Technology, Dalian, China
2. School of Software, Dalian University of Technology, Dalian, China
Email: xyliu@dlut.edu.cn

Hongfei Lin[1] and Liguo Zhang[2]
1. School of Computer Science and Technology, Dalian University of Technology, Dalian, China
2. School of Software, Dalian University of Technology, Dalian, China
Email: hflin@dlut.edu.cn

***Abstract*—Link weighting is crucial for performance of link-based web page ranking algorithms. Typically, a link is viewed as recommendation between pages, which is an unquantifiable term and existing approaches lack physical interpretations. In this paper, we view a link as the attractive force between pages, and map concepts of web (in/out degree, content similarity, etc.) to those of physics (mass, distance, attractive force). Inspired by Reilly's Law of Retail Gravitation, we propose a gravitation-like model for calculating the attractive force. We then implement a instance of our algorithm framework by taking some features of web pages into consideration. Experimental results show that this instance outperforms other typical algorithms (HITS, Randomized-HITS, and SALSA) with higher precision, better resistibility of TKC effect and no need of filtering intra-domain links.**

***Index Terms*—Link Analysis, Attractive Force Model, Link Weight, Web Information Retrieval**

I. INTRODUCTION

In modern web search engines, link-based ranking algorithms play an important role [1]. It is clear that different links are of different importance in contributing to the rank of a page on a given query. Thus a problem arises on how to assign weights to links. Till now, there is not a satisfying approach to the problem of weighting links since many features are involved, such as text in documents, link anchors, user feedback.

Typically, existing link analysis algorithms are based on the assumption that a link stands for a recommendation of one page to another. However, the term "recommendation" is hard to quantify, i.e., it is hard to determine the extent of recommendation from one page to another. This hardness essentially answers for the difficulty of weighting links and existing approaches lack physical interpretations.

The angle of view of recommendation lies at the start point of each link: a link from page A to page B exists is because that the author of A thinks B is important and makes a recommendation of B. However, from another angle of view, i.e., at the end point of each link, we can say that a link pointed to page B by page A is because that the importance of B causes the attention of A. From this point, instead of recommendation, a link can be interpreted as an attractive force between two pages, which is a quantifiable term. Note that web is a social phenomenon, and it has been observed by W.J.Reilly that the attractive force between cities follows Sir Isaac Newton's theory of gravitation (Reilly's Law of Retail Gravitation)[2], it is convincing to assume that the attractive force between pages follows a similar law.

From the above observations, in this paper, we propose a new framework for query-dependent link analysis in which each link is weighted by calculating the attractive force between its two associated pages. A mapping is built from concepts of web pages (in/out degree, content similarity, etc.) to those of physics (mass, distance, attractive force), and the attractive force is computed with formulas like theory of gravitation.

It is encouraging that we can really benefit from the nature since we have implemented an instance of the framework. Though only partial features are involved in this instance, the instance, which we call G-HITS, shows great advantages over other typical approaches such as HITS[3], Randomized-HITS[4] and SALSA[5]. Experimental results show that our instance algorithm outperform other algorithms in three aspects. 1) Better precision: G-HITS perform much better than other algorithms under the P@10 precision metric. 2) Resistance of TKC Effect: G-HITS is less vulnerable to the TKC effect than algorithms. 3) No need of filtering intra-domain links: Other algorithms need to filter intra-domain links to alleviate bias caused by mutually reinforcing effect. This is time consuming and may manslaughter some useful information. G-HITS without filtering intra-domain links perform as well as filtering, or even better.

The rest of the paper is organized as follows. Section 2 presents the background and some related works. The

Manuscript received December 28, 2010; revised March 1, 2011; accepted March 28, 2011.
Corresponding author: Xinyue Liu

doi:10.4304/jcp.7.1.124-129

attractive force model and the algorithm framework presented in section 3. In section 4 we implement a instance algorithm. The experiments and evaluations are given in section 5. Finally, we conclude the paper in section 6.

II. Background and Related Work

A. Background

Kleinberg [3] proposed that web documents had two important properties, called hubness and authority, as well as a mechanism to calculate them. In his Hyperlink-Induced Topic Search (HITS hubness) approach to broad topic information discovery, the score of a hub (authority) depended on the sum of the scores of the connected authorities (hubs). Kleinberg calculated these scores on the subset of the web that included top-ranked pages for a given query, plus those pages that pointed to or were referenced by that set.

Page and Brin [6] proposed an alternative model of page importance, called the random surfer model. In that model, a surfer on a given page i, with probability $(1 - d)$ chooses to select uniformly one of its outlinks, and with probability d to jump to a random page from the entire web. The PageRank score for node i is defined as the stationary probability of finding the random surfer at node i. PageRank is a topic-independent measure of the importance of a web page, and must be combined with one or more measures of query relevance for ranking the results of a search.

B. Related Works

Ideally, HITS is a good model for query-dependent page ranking. However, in the real web world, there are many negative characteristics that affect precision of ranking algorithms. Thus many works try to improve over HITS. Lempel and Moran [5] defined a tightly-knit community (TKC) as a small but highly connected set of sites. Even though such a community is not quite relevant to the query, it may still be ranked highly by link-based ranking algorithms. The authors proposed SALSA, a stochastic approach for link structure analysis, and demonstrated that it is less vulnerable to the TKC effect than HITS. Ng et al. [4] presented randomized HITS and subspace HITS algorithms to enhance the stability of the original HITS algorithm, etc. Though all of these algorithms adopt weighting links to address the problems, none give a clear answer or physical interpretation on why the links should be weighted in such a way. Bharat and Henzinger [7] proposed a number of improvements to HITS. One of the changes is an algorithm called imp, which re-weights links involved in mutually reinforcing relationships and drops links within the same host. They found that imp made a significant improvement over the original HITS. Chakrabarti et al. [8] extend HITS by increasing the weights of links whose anchor text incorporates terms from the query.

There are several works that try to provide general framework for links analysis. HITS emphasizes mutual reinforcement between authority and hub webpages, while PageRank emphasizes hyperlink weight normalization and web surfing based on random walk models. Ding et al [9] systematically generalized and combined these concepts into a unified framework. Xi et al [10] proposed a unified link analysis framework, called "link fusion", which considers both the inter- and intra-type link structure among multiple-type inter-related data objects and brings order to objects in each data type at the same time. Chen et al [11] proposed a unified framework that put both explicit and implicit link structures under a framework.

Finally, we note that physical models, especially gravitation model, have been successfully applied in related fields such as information retrieval and data mining. Shi et al [12] provided a gravitation-based model and bring explicit physical interpretation to formulas and concepts in information retrieval. Zhang et al [13] proposed a mechanical algorithm for data clustering which give explicit descriptions for movements of data points when falling into their genuine clusters.

III. Attractive Force Model and Algorithm Framework

A. Basic Concepts

1. Particle, Hub Mass and Authority Mass

In our model, the web pages are depicted as particles. Following Kleinberg [3], who associated every page with a hub weight and an authority weight, we associate every web page with a hub mass and an authority mass. Note that hub mass and authority mass are both query-dependent. Intuitively, a page with a high authority mass tends to attract more attention and a page with a high hub mass tends to convey more authority information. Thus, features like page content, page-query similarity and web-logs are factors of authority mass and hub mass.

We should note the more a page link to or linked by other pages, the more it can cause attention or convey information. Thus in-degree and out-degree are both main factors of hub/ authority mass.

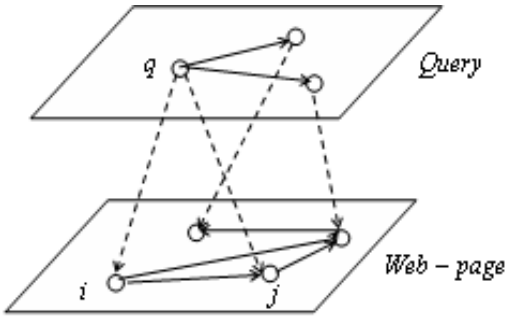

Figure 1. Query space and page space

2. Distance between Pages

The distance between two pages is a similarity metric. The WWW can be seen as a huge space, query and web page can be seen as two dimensions of the space. Most of the web link analysis research only analyzes the

hyperlinks within web pages, without considering the query dimension. However, in the real world, web pages and queries have close relationships. For example, see Fig 1, a query in the *Query* space can not only refer to other queries, but also refer to the web pages in the *Web-page* space.

3. Attractive Force between Two Pages

Assume there are two pages *i* and *j*, whose hub masses and authority masses are denoted by $M_{i,h}, M_{i,a}$ and $M_{j,h}, M_{j,a}$ respectively. There distance is denoted by $r_{i,j}$. Note that hub mass acts for conveying information and authority mass acts for attract attention, thus the attractive force between *i* and *j*, denoted by $P_{i,j}$ is dominated by $M_{i,h}$ and $M_{j,a}$, following a formula like theory of gravitation.

$$P_{ij} = \begin{cases} C\dfrac{M_{ih} \times M_{ja}}{r_{i,j}^{\ 2}}, & if \quad (i,j) \in E \\ 0, & otherwise \end{cases} \tag{1}$$

Where C is a constant. *E* is the set of the directed edges of the web graph *G= (V, E)*, derived from the given query.

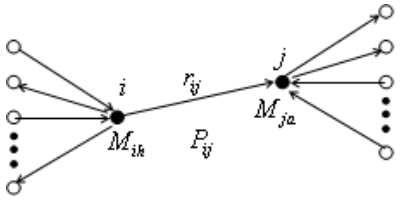

Figure 2. The attractive force between two pages

B. Algorithm Framework

1. Constructing a Sub Graph from the WWW

For a parameter *t*, the *t* highest-ranked pages for the query from a text-based search engine are collected. They are referred as the root set R_σ . The procedure is the same as HITS [3] algorithm. Then, the base set S_σ is obtained by growing R_σ to include any page pointed to by a page in R_σ and any page that points to a page in R_σ , with the restriction that every page in R_σ can bring at most *d* pages pointing to it into S_σ . So the collection of the hyperlinked pages in the S_σ can be viewed as a directed graph *G= (V, E)*.

2. Computing Hub Weights and Authority Weights

Like other typical link analysis algorithms, our algorithm framework is also based on the random walk process of Markov chains. Let there be a random surfer who follows hyperlinks in both the forward and backward directions. More precisely, the surfer starts from a randomly chosen page, and visits a new web page at every time step. Every time step, he tosses a coin with probability ε , and if the coin lands heads, he jumps to a new web page chosen at random. If the coin lands tails, then he checks if it is an odd time step or an even time step. If it is an odd time step, then he follows a randomly chosen forward link from the current page; if it is an even time step, then he traverses a random backward link of the current page. Thus, the random surfer alternately follows links in the forwards and backwards directions, and occasionally "resets" and jumps to a page chosen at random. That can be expressed in the matrix-vector terms as follows:

$$a^{(t+1)} = \varepsilon\vec{1} + (1-\varepsilon)P_r^T \times h^{(t)} \tag{2}$$

$$h^{(t+1)} = \varepsilon\vec{1} + (1-\varepsilon)P_c \times a^{(t+1)} \tag{3}$$

Where *P* is the adjacency matrix derived from the query-specific link graph, each element of the matrix *P* is the value of the attractive force that one page links to the other. $\vec{1}$ is the vector consisting of all ones, P_r is the same as *P* with its rows normalized to sum to 1, and P_c is *P* with its columns normalized to sum to 1.

IV. An Algorithm Instance

Implementation of our algorithm framework need concrete methods for computing masses and distances. In this section, we provided an instance of the algorithm in which part of the features are involved.

A. Computing hub mass and authority mass

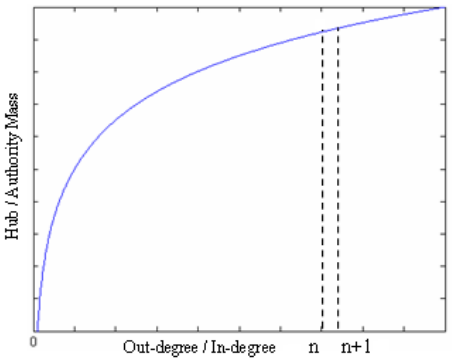

Figure 3. Increasing trends of hub/authority mass, with the increasing of the out-degree / in-degree

It has been analyzed that in-degree and out-degree are both main factors of hub/ authority mass. Let investigate how they affect the masses. Intuitively, the more a page is pointed by other pages, the higher authority, and it should be assigned a lager mass. In the beginning, when a page *i* has a small number of in-links, adding a new link to it is of much importance for it to gain more attention. However, when a page *i* already has a large number in-links, a new added link has little impact to its popularity. This means that a page's authority mass increase with its in-degree following a curve like Fig 3. The relationship between hub mass and out-degree is similar. We adopt logarithm functions to describe them. Note that out-degree also affect authority mass, since a page link to many other tends to be known more, but its impact is much less than in-degree. The relationship between authority mass and in-degree is similar.

The formulas of hub mass M_{ih} and authority mass M_{ia} are as follows. They are the base-e logarithm functions of $|F(i)|$ and $|B(i)|$, which represent out-degree and in-degree respectively.

$$M_{ih} = \alpha \times \ln(|F(i)|+1) + \beta \times \ln(|B(i)|+1) \quad (4)$$

$$M_{ia} = \alpha \times \ln(|B(i)|+1) + \beta \times \ln(|F(i)|+1) \quad (5)$$

Where α and β are the dependent factors. We set the factors as follows: α is 0.80-0.85, β is 0.15-0.20 and $\alpha + \beta = 1$.

B. Computing distance between pages

According to the relationship of queries and web pages in the space, distance $r_{i,j}$ between two pages can be reflected by the distance of pages *i* and *j* to query *q*, which can be indirectly reflected by the content similarities of pages *i* and *j* to query topic *q*. That can be expressed as follows.

$$r_{i,j} = \lambda \times f(s_{q,i}, s_{q,j}) \quad (6)$$

Where λ is the dependent factor and satisfies $0 < \lambda \le 1$. If λ is larger, the distance will be dependent on the content similarities more. The formula of content similarity $s_{q,j}$ of page *j* to the query topic *q* is showed in the below, $s_{q,i}$ is the same with it.

$$s_{q,j} = \frac{\sum_{k=1}^{t} w_{k,j} \times w_{k,q}}{\sqrt{\sum_{k=1}^{t} w_{k,j}^2} \times \sqrt{\sum_{k=1}^{t} w_{k,q}^2}} \quad (7)$$

Where $w_{k,j} = tf_{k,j} \times idf_k$, $w_{k,q} = tf_{k,q} \times idf_k$, and $tf_{k,j}$ is the frequency of the term *k* in document *j*, $tf_{k,q}$ is the frequency of the term *k* in query *q*, idf_k is an estimate of the inverse document frequency of the term *k* on the WWW.

Then, the formula of the distance is as follows.

$$r_{i,j} = \lambda \times (1/\sqrt{s_{q,i}^{\ 2} + s_{q,j}^{\ 2}}) \quad (8)$$

Formula (8) also reflects distance of page *i* to page *j* reasonably, we can assume that query *q*, page *i* and *j* can form a right-angled triangle in the data space. $r_{i,j}$, $s_{q,i}$ and $s_{q,j}$ are the edges of the triangle. The physical meaning of distance $r_{i,j}$ between two pages is that if content similarities $s_{q,i}$ and $s_{q,j}$ are larger, the distance of page *i* to *j* will be nearer.

C. The algorithm

The algorithm is a HITS-like algorithm with a gravitation model. We name it G-HITS. The algorithm is described as follows.

Algorithm 1: G-HITS algorithm to compute the hub weights and authority weights

Input: The sub graph *G (V, E)* derived from the given query
Output: *h*, *a*
Iteration (*G*)
F (i) denote the set of all pages *i* points to *i*
B (i) denote the set of all pages pointing to *i*
P_{ij} denote the attractive force from *i* to *j*

1. Let $z = [1, 1, \ldots, 1]^T$
2. Initialize *h* and *a*, set $h = z$, $a = z$
3. $h_i = \varepsilon \vec{1} + (1-\varepsilon) \sum_{j \in F(i)} a_j \times \frac{1}{|B(j)|} \times P_{ij}$
4. $a_i = \varepsilon \vec{1} + (1-\varepsilon) \sum_{j \in B(i)} h_j \times \frac{1}{|F(j)|} \times P_{ji}$
5. Normalize *h* and *a*
6. Obtaining $h_i = h_i / \|h\|$, $a_i = a_i / \|a\|$
7. Repeat 3, 4, 5, and 6 until convergence

Return *h*, *a*

The time complexity of the iteration algorithm above is $O(N^2)$. The element P_{ij} can be expressed as follows:

$$P_{ij} = f(M_{ih}, M_{ja}, r_{i,j}) \times W_{ij} \quad (9)$$

Where
$$W_{ij} = \begin{cases} 1, & if \ (i,j) \in E \\ 0, & otherwise \end{cases} \quad (10)$$

According to the Randomized HITS algorithm [10], the iteration will cause $a^{(t)}$ and $h^{(t)}$ converge to the odd-step and even-step stationary distributions.

V. Experiments and Evaluations

A. Experimental Methods

TABLE I.
THE SIMILARITY THRESHOLDS OF THE QUERIES

Queries	abortion	jaguar	geometry	genetic	bicycling	java
$s_{q,iMin}$	1.0E-06	5.0E-07	5.0E-07	5.0E-07	5.0E-07	1.0E-06

TABLE II.
THE DATASET DETAIL FOR EACH QUERY, FILTERING THE INTRA-DOMAIN URLS

Queries	abortion	jaguar	geometry	genetic	bicycling	java
B_σ size	1546	1097	2185	2117	1892	2851
Links	1018	1011	2190	2024	4327	9290

We performed experiments on G-HITS and three other typical algorithms with the following 6 different queries: abortion, jaguar, geometry, genetic, bicycling and java, which have been used in previous works [3]. We first obtained the root set R_σ and base set s_σ for every query like HITS [3] algorithm, and then we got the directed graph *G= (V, E)* derived from the s_σ .Secondly, we attempted to filter the noninformative URLs, which exist in the web pages, according to the distributions of the content similarities of the pages for each query. We set a similarity threshold $s_{q,iMin}$ for each query, showed in Table

1. Then we filtered the URLs whose content similarities are below $s_{q,iMin}$, and we got the new base set B_σ. Thirdly, we did the experiments under two conditions as follows:

(1) Filtering intra-domain links: In this experiment, we filtered intra-domain links, following HITS and SALSA algorithms. The size and the links of B_σ are presented in Table 2.

(2) No filtering the intra-domain links: In this experiment, we didn't filter the intra-domain links, since we want to have a comparison with the filtered experiments. The size and the links of B_σ are presented in Table 3.

Then, we set the factors α =0.80, β = 0.20, λ =0.95, C=1.00, t=200, d=50, the convergence threshold δ to be 1.0E-06, and the jump probability ε to be 0.20 for every query.

TABLE III.
THE DATASET DETAIL FOR EACH QUERY, WITHOUT FILTERING THE INTRA-DOMAIN URLS

Queries	abortion	jaguar	geometry	genetic	bicycling	java
B_σ size	1546	1097	2185	2117	1892	2851
Links	9536	9494	6704	11385	22846	49538

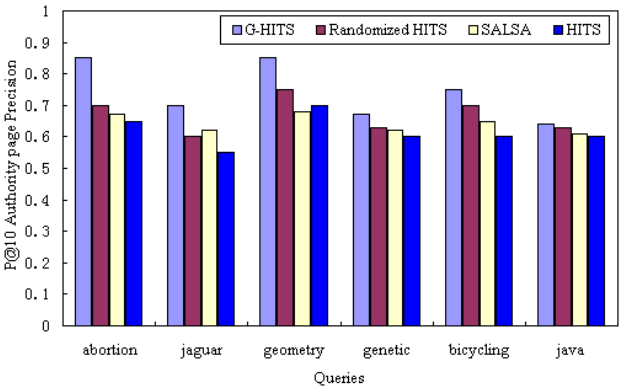

Figure 4. Comparison of P@10 Authority Page Precision of the G-HITS, Randomized HITS, SALSA, and HITS, filtering the intra-domain links

B. Results Analysis and Evaluations

1. Authorities Evaluation

P@10（Precision at 10）is the number of relevant documents in the top 10 documents in the ranked list returned for a topic. If the document is relevant, the score of it will be 1; else it will be 0.

We asked 100 volunteers to evaluate the result authorities according to our rules above. Then we got the average high relevant rate for each query topic, marked as P@10 Authority Page Precision. We can see the performance comparison of the four algorithms in Fig 4, the experiment has filtered the intra-domain URLs. And in Fig 5, the experiment is without filtering intra-domain links.

(1) The experiment with filtering intra-domain links.

We performed this experiment following HITS, SALSA and other algorithms to filter the intra-domain URLs. We can see that G-HITS algorithm get higher authorities than other three typical link analysis algorithms, in appendix 1 and 2. For example, the results authorities of Top1-Top10 for the query "abortion", showed in appendix 1. Seven URLs are quite popular with the volunteers in G-HITS algorithm, which are Top1 to Top6, and Top10. But there are four, five, and five URLs are popular, in HITS, SALSA, and Randomized HITS algorithms, respectively.

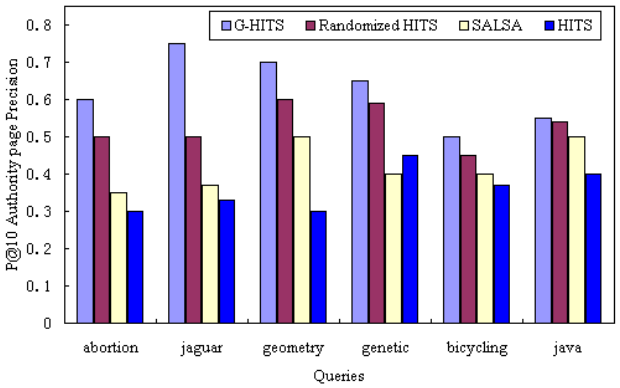

Figure 5. Comparison of P@10 Authority Page Precision of the G-HITS, Randomized HITS, SALSA, and HITS, without filtering the intra-domain links

(2) The experiment without filtering intra-domain links.

We can see the performance of different result authorities in appendix 3, with the query "jaguar". The authorities of G-HITS are much better than other three algorithms. We can compare the experimental results of the filtered and unfiltered intra-domain links in appendix 2 and appendix 3, with the same query "jaguar". We find that the unfiltered results of G-HITS in appendix 2 are even better than the filtered in appendix 3. But for the other three algorithms, it can be seen that the filtered results are better than unfiltered. Thus, we can get rid of the redundant time of filtering. This is because that G-HITS has strong resistibility against TKC effect, which is explained in the following sub-sections.

2. Resistance of TKC Effect

The TKC Effect occurs in HITS algorithm for the query topic "jaguar", showed in appendix 2, which has filtered intra-domain links. It is obviously that Top6 to Top10 are URLs from the same domain. That is the result of the Mutual Reinforcement Approach of HITS algorithm. Top8 to Top10 are same domain URLs of SALSA algorithm, which is less vulnerable than HITS in the TKC Effect. Randomized HITS also has TKC effect obviously. But see the authorities of G-HITS algorithm presented in appendix 2, we find that just Top4 and Top5 are in the same domain. Then, let's see appendix 3, which is results without filtering intra-domain links. The results of HITS, SALSA and Randomized HITS are quite bad because of the TKC Effect occurred. But the results of G-HITS are much better than them, and show the robustness against the TKC Effect. Similar phenomena are also found in the top ten authorities of query "genetic", "bicycling" etc. Thus, G-HITS algorithm is more resistant

to the TKC effect than other algorithms, no matter on the condition of the filtering intra-domain links, or no filtering intra-domain links.

We can give brief analysis to this phenomenon. In G-HITS algorithm, different links correspond to different attractive force values from physical perspective. For example, web page *k* links to both *i* and *j*, if the content similarity of page *i* to the given query *q* is larger than *j*, the distance of page *k* to *i* will be nearer than *k* to *j*; at the same time, if the authority mass of page *i* are not the same with the page *j*, then, attractive force P_{ki} will be different from P_{kj} .Thus, the mutually reinforcing effect alleviated and no page can benefit too much from its neighbor, which can lead to TKC Effect. While at the same time, not all intra-domain links are used for navigation, some are also used for recommendation. This explains why no filtering of intra-domain links for G-HITS sometimes performs better than filtering.

VI. Conclusions and Future Work

In this paper, an attractive force model for link weighting is proposed from the perspective of social science inspired by Reilly's Law of Retail Gravitation. An algorithm framework for base query-dependant web page ranking is then deduced and an instance of the framework is implemented. Experimental results show that it is encouraging that we can really benefit from the nature. An open problem is that how can we fuse all the features that affect page ranking into this model. Another problem is how we can develop a similar model for query-independent web page ranking.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grant No. 60873180 and the Fundamental Research Funds for the Central Universities.

References

[1] Allan Borodin et al, "Link Analysis Ranking Algorithms Theory and Experiments," *ACM Transactions on Internet Technologies*, 2005, vol. 5, pp. 231-297.

[2] William John Reilly, "The law of retail gravitation," *New York*, W. J. Reilly, 1931.

[3] J. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, 1999, vol. 46, pp. 604-632.

[4] A.Y.Ng, A.X.Zheng, and M. I. Jordan, "Stable Algorithms for Link Analysis," *Proc. 24th International Conference on Research and Development in Information Retrieval*, *New Orleans*, *Louisiana*, *USA*, 2001.

[5] R. Lempel and S. Moran, "The Stochastic Approach for Link-structure Analysis (SALSA) and the TKC effect," *In: Proceeding 9th International World Wide Web Conference*, *Amsterdam, Netherlands*, 2000.

[6] S. Brin and L. Page, "The Anatomy of a Large-scale Hyper textual Web Search Engine," *Computer Networks and ISDN Systems*, 1998, vol. 30, pp. 107-117.

[7] K. Bharat and M. R. Henzinger, "Improved algorithms for topic distillation in hyperlinked environments," *In Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998, pp. 104–111.

[8] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, S. Rajagopalan, "Automatic resource compilation by analyzing hyperlink structure and associated text," *Proceedings of the 7th ACM-WWW International Conference, Brisbane*, 1998, pp. 65-74.

[9] Chris Ding, Xiaofeng He, Parry Husbands, Hongyuan Zha and Host D. Simon, "PageRank, HITS and a unified framework for link analysis," *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 2002, pp. 353-354.

[10] Wensi Xi, Benyu Zhang, Yizhou Lu, Zheng Chen, Shuicheng Yan, Huajun Zeng, Wei-Ying Ma, and Edward A. Fox, "Link Fusion: A Unified Link Analysis Framework for Multi-Type Interrelated Data Objects," *The Thirteenth World Wide Web conference (WWW 2004)*, 2004, pp. 203-211.

[11] Zheng Chen, Li Tao, Jidong Wang, Liu Wenyin, Wei-Ying Ma, "A Unified Framework for Web Link Analysis," *Proceedings of the ACM Workshop Wireless Security*, 2002, pp. 63-72.

[12] Shuming Shi, Ji-Rong Wen, Qing Yu, Ruihua Song, Wei-Ying Ma, "Gravitation-Based Model for Information Retrieval," *The 28th Annual International ACM SIGIR Conference (SIGIR'2005)*, 2005.

[13] Xianchao Zhang, He Jiang, Xinyue Liu and Hong Yu, "A Clustering Algorithm Based on Mechanics," *Lecture Notes in Artificial Intelligence*, 2007, vol. 4426, pp. 367-378.

Xinyue Liu received the M.S degree in Computer Science and technology from Northeast Normal University, China, in 2006. She is currently working toward the Ph.D. degree in the School of Computer Science and Technology, Dalian University of Technology, Dalian, China. Her research interests include multimedia information retrieval, web mining and machine learning.

Hongfei Lin received the Ph.D degree from Northeastern University, China. He is a professor in the School of Computer Science and Technology, Dalian University of Technology, Dalian, China. His professional interests lie in the broad area of information retrieval, web mining and machine learning, affective computing.

Liguo Zhang received his M.S. degree in Software Engineering from Dalian University of Technology in 2008. His major interests lie in web link analysis.