Article

JIS

Journal of Information Science
2014, Vol. 40(6) 761–778

DOI: 10.1177/0165551514545429
jis.sagepub.com
SAGE

C-Rank and its variants: A contribution-based ranking approach exploiting links and content

Dong-Jin Kim
NHN Institute for the Next Network, Seongnam, Korea

Sang-Chul Lee
Department of Electronics and Computer Engineering, Hanyang University, Seoul, Korea

Ho-Yong Son
Department of Electronics and Computer Engineering, Hanyang University, Seoul, Korea

Sang-Wook Kim
Department of Electronics and Computer Engineering, Hanyang University, Seoul, Korea

Jae Bum Lee
NHN Corp., Seongnam, Korea

Abstract
This paper addresses the problem in Web page ranking of effectively combining link and content information with efficiency high enough to be applicable to real-world search engines. Unlike previous surfer models, our approach is based on the viewpoint of a Web page author. Based on this viewpoint, we formulate the concept of contribution score, which indicates the amount to which a term in each page is utilized by other pages. To improve efficiency without loss of effectiveness, we exploit the expectations of both a Web page author and a Web search engine user on retrieval results, and restrict candidate terms that can contribute to other pages to a set of keywords of each page. In this paper, we propose three contribution-based models: C-Rank, PC-Rank and HC-Rank. Experimental results show that C-Rank provides the best precision among the models and is very effective for topic distillation tasks on the .GOV collection in TREC. Most importantly, the proposed models are efficient enough to be applicable to real-world search engines.

Keywords
Content and link ranking; contribution based ranking; contribution constraints; C-Rank; Web information retrieval

1. Introduction

Link structure is the most characteristic feature in Web information retrieval compared with traditional information retrieval. However, the most direct evidence that a page is relevant to a given query is the contents of the page [1]. To effectively combine content and link information, a number of approaches have been suggested [2–28].

Page *et al*. suggested PageRank based on the novel concept that a page has a high rank if the sum of the ranks of its backlinks is high [13]. However, the original concept cannot be implemented because of the problem of rank sinks, which

Corresponding author:
Sang-Wook Kim, Department of Computer Software, Hanyang University, 17 Haengdang1-dong Seongdong-gu, Seoul, Korea
Email: wook@hanyang.ac.kr

occur when there exists a cycle or a page with no outgoing links. To overcome this problem, they adopted the concept of a random surfer [3–8, 13, 26] making a random jump. If the random surfer is on a page with no outgoing links or getting into a small cycle of pages, he/she jumps to a randomly chosen page, thereby solving the problem. PageRank indicates a global authority score of every page that is independent of a query. However, Page *et al*. did not mention how to combine PageRank with the relevance score of a page to a query.

Most of the following works that tried to combine content evidence with PageRank can be classified into two categories. The first one focuses on the probability of a link to be chosen by a surfer, depending on the contents of a page such as a topic [3, 5, 8, 11, 12, 14, 18, 19]. The second one independently computes both the relevance score of a page to a query and the global authority score of the page, and regards the weighted sum of the relevance and global authority scores as the final rank of the page [24, 29–31].

Probabilistic approaches based on the random surfer model measure the authority score of a page by summing the authority scores currently flowing into the page when the flow through every page is in equilibrium, that is, the amount of incoming flow is equal to the amount of outgoing flow. A real Web graph consists of an enormous number of pages and links. Nevertheless, the insertion/deletion of a page into/from a real Web graph breaks the equilibrium state and, consequently, affects the authority score of every page. Even though the variance may be small, the insertion/deletion affects the authority score of a page regardless of how far it is from the inserted/deleted page. That is, even if the path from one page to another is long, for example, 20 links away, the pages still have an effect on one another. We do not think this effect properly reflects the real Web environment.

Compared with PageRank, which works on all pages, Shakery and Zhai suggested the relevance propagation model [20] that works on a subset of pages. This model is similar to PageRank, but it is query dependent. The relevance propagation model and its variants have been known to be very effective [6, 17, 20–22]. However, they can rarely be applied to real-world search engines since, at querying time, the models need to construct a subset of pages and propagate relevance scores among pages in the subset until the scores reach convergence.

In this paper, we introduce a new approach for ranking Web pages that effectively combines content and link information: C-Rank, PC-Rank and HC-Rank, standing for contribution-based rank, probabilistic C-Rank and hybrid C-Rank, respectively. Our proposed approach aims to discover a bunch of Web pages that have not only a high authority score, but also good content relevant to the topic in a given query. This task can be classified as the topic distillation task [32, 33].

Unlike previous surfer models, our approach is based on the viewpoint of a Web page author rather than a Web surfer. Based on this viewpoint, we formulate the concept of a contribution score that indicates the amount to which a page is utilized by other pages on a given topic. A term in a page is regarded as a topic of that page, and the contribution score of a page on a term is considered the contribution score of the term in the page to other pages. Then we can compute the contribution score of every term in every page in preprocessing time. This computation might look infeasible. However, there exists an additional worthwhile feature based on the expectations of a Web page author and a Web search engine user. We propose an efficient method that utilizes this feature. This idea is briefly described below and detailed in Section 3.1.

Usually, a Web page author creates a page with one or a few topics, and the page can be summarized by a small number of terms. In this paper, we define these terms as keywords of the page. To improve efficiency without loss of effectiveness, we note the expectations of a Web page author and a Web search engine user on retrieval results, and thus consider only those keywords as the candidate terms that can contribute to other pages. As a result, the contribution score of a term that is not a keyword is set to zero.

During preprocessing time, for each term in a page, we combine a contribution score of the term in the page to other pages and a relevance score of the page to the term. The combined score is assigned as the final score of the term. Given a query of multiple terms, the contribution score of a page under the query is determined by the sum of the final scores of query terms for the page. Consequently, at querying time, our approach requires the same retrieval time as other models using only a relevance score.

To clarify the difference between our approach and previous probabilistic approaches, we compare their ranking measures. The ranking measure of probabilistic approaches based on the surfer model is the amount of authority scores currently flowing into a page at the time that all pages reach an equilibrium state. If we analyse our approach from the aspect of the flow of authority scores, our ranking measure corresponds to the amount of authority scores absorbed by a page until there is no flow of an authority score through the page.

In our experiments, we evaluate the effectiveness and efficiency of our approach according to the number of keywords of a page and the maximum length of a path through which a page contributes to others. To evaluate our approach, we conduct extensive experiments on the .GOV collection with two sets of topic distillation task topics [34, 35]. Experimental results show that C-Rank provides the best precisions among our three models and it is very effective on

the .GOV collection for topic distillation tasks. It is remarkable that the results of C-rank with 10 keywords and the maximum length 3 are comparable to other results of C-Rank using the larger number of keywords and the longer maximum length. Also, compared with previous relevance propagation models on the topic distillation tasks in both P@10 and MAP, our results show that C-Rank is the best or very close to the best in effectiveness. Our models are also notable in terms of efficiency. In contrast with previous relevance propagation models, which can rarely be applied to real-world search engines owing to a lack of efficiency, our models are efficient enough to be applicable to real-world search engines.

The rest of this paper is organized as follows: in Section 2, we review related works. Section 3 explains our approach in detail. In Section 4, our experimental results are shown. We conclude in Section 5.

2. Related work

The original concept of PageRank is the following: a page has a high rank if the sum of the ranks of its backlinks is high [13]. This concept can be formularized by equation (1):

$$P(p) = \sum_{q \in inlink(p)} P(q)/N_q \tag{1}$$

where p and q are Web pages, $P(p)$ is the ranking of page p, $inlink(p)$ is the set of pages that point to page p and N_q is the number of outlinks from page q. Equation (1) has the problem of *rank sinks* occurring by a cycle or a page without outlinks. To overcome this problem, Page *et al*. introduced a random jump component. PageRank can be described as below:

$$P(p) = \lambda \sum_{q \in inlink(p)} \frac{P(q)}{N_q} + (1 - \lambda)/N \tag{2}$$

where N is the total number of pages and $1 - \lambda$ is the probability of jumping to a random page. $P(p)$ is computed iteratively until it reaches convergence.

Richardson and Domingos proposed the query-dependent PageRank depending on content of pages and query terms that a surfer is looking for [19]. The resulting probability distribution over pages under query term t is computed as below:

$$P(p) = \lambda \sum_{q \in inlink(p)} P_t(q) P_t(q \to p) + (1 - \lambda) P'_t(p) \tag{3}$$

where $P_t(q \to p)$ is the probability that a surfer on page q moves to page p through an outlink of page q under term t. $P'_t(p)$ specifies the probability that a surfer moves to page p with a random choice. Richardson and Domingos combined PageRank with a relevance score by defining the two probabilities as the following:

$$P_t(q \to p) = \frac{R_t(p)}{\sum_{r \in outlink(q)} R_t(r)}, P'_t(p) = \frac{R_t(p)}{\sum_{r \in W} R_t(r)} \tag{4}$$

where $R_t(p)$ is the relevance score of page p to term t and W is the set of all pages. The relevance score of a page to a query is computed by summing relevance scores of the page to the query terms.

As a method of combining link and content information, Shakery and Zhai proposed a relevance propagation model [20, 21]. Given a query, this model constructs a *working set* that consists of the topmost relevant pages to the query and other relevant pages to the query that have links to/from the topmost relevant pages. For every page in the working set, the relevance score of a page is initially assigned by its content-based relevance score to a query. Then, each page propagates its relevance score into its neighbour pages in the working set according to the weight of a link that is proportional to the initial relevance score of a page pointed to by the link. Once all the pages receive relevance scores from their neighbours, each page resets its relevance score to the sum of the currently received relevance scores and again propagates the new relevance score to its neighbour pages. This process is repeated until the relevance score of every page reaches convergence. The converged relevance score of each page is regarded as the final rank of the page. Formally, the rank of page p is defined as follows [21]:

$$P(p) = \sum_{i=1}^{k} \lambda_i \sum_{q \in D} P(q) P_i(q \to p) \tag{5}$$

where D is the working set, k is the number of neighbour sets, λ_i indicates the probability of choosing a particular type of a neighbour set when propagating a relevance score, $P(p) = \sum_{i=1}^{k} \lambda_i = 1$, $P_i(q \to p)$ is the probability of propagating a relevance score of page q to page p in the chosen neighbour set, and $\sum_{q \in D} P(q) P_i(q \to p) = 1$.

Qin *et al*. [17] proposed a hyperlink-based term propagation model (or HT model), merging the concepts of the relevance propagation model [20] and the sitemap-based term propagation model [22]. The HT model also constructs a working set of pages and propagates the frequency of a query term in a page in the working set. Then, a relevance weighting algorithm such as BM25 [36] is used to rank the pages. The HT model is formularized as below:

$$f_t^{k+1}(p) = \lambda f_t^0(p) + (1 - \lambda) \sum_{q \in D_p} f_t^k(q) P_t(q \to p) \tag{6}$$

where $f_t^k(p)$ is the occurrence frequency of term t in page p after the kth iteration, $f_t^0(p)$ is the original occurrence frequency of term t in page p, D_p contains the pages which point to or are pointed to by page p, and $P_t(q \to p) \propto f_t^0(p)$.

The previous relevance propagation model and its variants provide a certain improvement in effectiveness. However, they can rarely be applied to real-world search engines since the models require much computational overhead at querying time for constructing a working set and propagating scores.

3. Proposed approach

3.1. Overview

Our approach starts from investigating the viewpoint of a Web page author rather than a Web surfer. A typical Web page author does not place random links in his/her pages (with the possible exception of banner advertising), but instead tends to create links to pages on related topics [37]. We note that the main purpose of creating links is to supplement the contents of a page. Therefore, a page can be considered to contribute to other pages pointing to the page by a link or through a chain of links.

Consider two pages connected as $q \to p$. To measure the amount of contribution of page p to page q under a given topic, we give the following intuition in contribution between page p and page q:

(1) As page q gets more valuable on the topic, the possibility that page p has good information on the topic becomes higher, thereby contributing to page q more on the topic. Conversely, if page q is not related to the topic, the amount of contribution of page p is small. This intuitive description is similar to the original concept of PageRank formularized as equation (1).
(2) As page p gets more valuable on the topic, page p contributes to page q more on the topic. It is natural that a better page contributes more. A similar concept is used in the intelligent surfer as equations (3) and (4).
(3) If page q is more valuable than page p on the topic, it means that the Web page author of page q created a link to page p to supplement insufficient information partially on the topic. Therefore, even though page p is valuable on the topic, the amount of contribution is not that much. To the best of our knowledge, our approach is the first to adopt this concept.
(4) As page q points to more pages related to the topic, the amount of contribution of page p to page q on the topic becomes smaller. This makes sense since, if page p gives complete information on the topic, the Web page author might not create links to other pages. Equations (1) and (4) show that this concept has been similarly used in previous surfer models.

In this paper, to construct a formula representing the amount of contribution, we use the above four types of intuition as the contribution constraints of the formula.

The contribution score of a page under a term can be considered as the contribution score of the term in the page. Based on the above contribution constraints, we carefully evaluate the amount of contribution of every term in every page in preprocessing time. It might look infeasible. However, there exists an additional worthwhile feature based on

the expectations of a Web page author and a Web search engine user on retrieval results. We propose a novel technique that utilizes those expectations.

Usually, a Web page author creates a page with one or a few topics, and the page can be summarized by its keywords. Meanwhile, a Web search engine user makes query terms to describe a page the user wants to retrieve. Many pages containing the query terms might exist. However, the pages relevant to the query are highly likely to have the query terms as keywords.

Now, we focus on the expectations of both a Web page author and a Web search engine user simultaneously. We believe that there exist two kinds of expectations: first, a Web page author expects that his/her page is ranked high for query terms matched to the keywords of the page, but the author might not mind the ranking of the page for query terms not matched to the keywords of the page; second, a Web search engine user rarely expects a page that does not contain the query terms as its keywords to be ranked high. There is no special reason for considering a page, the author of which disregards the page under a given topic, to be valuable on the topic.

Regarding a keyword of a page as the topic of the page, our approach computes the contribution score of every topic of every page to other pages. The contribution score of a term that is not a keyword is considered to be zero. Then, the final score of a term in a page is determined by combining the relevance score of the page to the term and the contribution score of the term to other pages. This computation is conducted offline in preprocessing time. Given a query of multiple terms, the final score of the page to the query is determined online by summing the final scores of all the query terms for the page.

3.2. C-Rank and its variations

A *C-Rank* of *a term in a page* is determined by the relevance score of the page to the term and a contribution score of the term to other pages. In a Web graph constructed with a large number of pages and links, the insertion of a new link into the Web graph affects the amount of contribution of a term in a page connected by the link or through a path containing the link. In this section, we describe C-Rank, PC-Rank and HC-Rank, which differ in the strategy of how a link affects a contribution score.

The first model considers only the effect of inlinks. The larger the number of inlinks of a page is, the higher the contribution score of a term in the page becomes. In the second model, the contribution score of a term in a page is represented by probability, thereby keeping the sum of contribution scores over all pages invariant. The third model is a hybrid of the first and the second models. In the hybrid model, the contribution score of a term in a page is not reduced to less than the relevance score of the page to the term even though an addition of an outlink causes its contribution score to decrease.

3.2.1. Basic model: C-Rank. In the basic model, a C-Rank of a term in a page is defined by the sum of both the relevance score of the page to the term and a portion of the contribution score of the term to other pages. Let $D(p, d)$ be the set of pages, each of which is a starting page of a path of length d that ends at page p. Here, a path means a sequence of *distinct* pages such that each of pages in the path has a link to the next page in the sequence [38]. Therefore, the path in our case does not allow a cycle. For example, consider a graph consisting of three pages connected as $r \rightarrow q \rightarrow p$. Here, $D(p, 1) = \{q\}$, $D(p, 2) = \{r\}$ and $D(q, 1) = \{r\}$. Then, the C-Rank of term t in page p can be formulated as below:

$$CR_t(p) = R_t(p) + \lambda \sum_{d} \sum_{q \in D(p,\, d)} C_t(p, q) \tag{7}$$

where $R_t(p)$ corresponds to the relevance score of page p to term t, $C_t(p, q)$ is the contribution score of term t in page p to page q, and $\lambda(\geq 0)$ is the ratio of a portion of the contribution score to be added to a C-Rank. Since we are interested in a rank rather than a score, equation (7) can be converted into equation (8):

$$CR_t(p) = \lambda R_t(p) + (1 - \lambda) \sum_{d} \sum_{q \in D(p,\, d)} C_t(p, q),\ 0 \leq \lambda \leq 1 \tag{8}$$

To determine a contribution score $C_t(p, q)$, we utilize the contribution constraints described in Section 3.1. Various kinds of contribution score formula satisfying the constraints can be constructed. In this paper, to satisfy the first constraint, we adopt a simple formula, as below:

$$C_t(p, q) = \alpha_t^d(p, q) R_t(q) \tag{9}$$

Table 1. Summary of notation.

Symbol	Description
$CR_t(p)$	C-Rank of term t in page p
$R_t(p)$	Relevance score of page p to term t
$NR_t(p)$	Normalized relevance score of page p to term t
$C_t(p,q)$	Contribution score of term t in page p to page q
$\alpha_t^d(p,q)$	Contribution ratio of term t in page p to page q by a path of length d
$\beta_t(p)$	Ratio of an absorption ratio of PC-Rank or HC-Rank to an absorption ratio of C-Rank
$\gamma_t(p)$	Preservation ratio of PC-Rank

where $\alpha_t^d(p, q)$ is the *contribution ratio* of term t in page p to page q when page q points to page p by a path of length d indirectly for $d \geq 2$ and directly for $d = 1$. For two pages directly connected by a link, the remaining three constraints can be satisfied by equation (10):

$$\alpha_t^1(p, q) = \frac{R_t(p)}{R_t(q) + \sum_{r \in outlink(q)} R_t(r)} \tag{10}$$

If d $\geq$ 2, then $\alpha_t^d(p, q)$ can be determined with contribution ratios of links in the path. Let us consider a path of $q \rightarrow r_1 \rightarrow \cdots \rightarrow r_{d-1} \rightarrow$ p. The contribution ratio of term t in page p to page q depends on contribution ratios of links in the path under term t. We define $\alpha_t^d(p, q)$ for $d \geq 2$ by equation (11):

$$\alpha_t^d(p, q) = \alpha_t^1(p, r_{d-1}) \times \prod_{i=1}^{d-2} \alpha_t^1(r_{i+1}, r_i) \times \alpha_t^1(r_1, q) \tag{11}$$

Finally, the basic model, $CR_t(p)$, is defined by equation (12):

$$CR_t(p) = \lambda R_t(p) + (1 - \lambda) \sum_d \sum_{q \in D(p, d)} \alpha_t^d(p, q) R_t(q) \tag{12}$$

For the reader's convenience, we summarize symbols mainly used in this paper in Table 1.

3.2.2. Probabilistic model: PC-Rank. The sum of all C-Ranks over all terms in all pages increases whenever a link is newly inserted. Accordingly, C-Rank cannot be explained by probability. To construct a probabilistic model satisfying the contribution constraints, we modify the basic model (C-Rank) into the probabilistic model (PC-Rank), adopting two strategies that keep the sum of all C-Ranks invariant.

Consider a page p, another page q that has a path to page p, and a set of the third pages to which page p has paths. The first strategy considers the sum of contribution scores to page q of both page p and the third pages. The goal of the first strategy is to keep the sum of contributions scores invariant for inserting/deleting a link to/from page p and the third pages. Equation (7) can be expressed as follows: $CR_t(p) = R_t(p) + \lambda C_t(p)$ where $C_t(p) = \sum_d \sum_{q \in D(p, d)} \alpha_t^d(p, q) R_t(q)$. To reflect the first strategy, we modify $C_t(p)$ as follows:

$$C_t(p) = \beta_t(p) \sum_d \sum_{q \in D(p, d)} \alpha_t^d(p, q) R_t(q) \tag{13}$$

where

$$\beta_t(p) = \frac{R_t(p)}{R_t(p) + \sum_{r \in outlink(p)} R_t(r)} \tag{14}$$

In the aspect of a score propagation method, equations (13) and (14) mean that page p absorbs the portion of $\lambda\beta_t(p)$ from its contribution score to accumulate the portion into a PC-Rank, and propagates only the remaining score to other pages pointed to by page p.

To show that equation (13) reflects the first strategy, let us consider a simple Web graph of $q \rightarrow p$. The contribution score of term t in page p to page q is $\alpha_t^1(p, q)R_t(q)$. Now, assume that we insert page o pointed to by page p as $q \rightarrow p \rightarrow o$. Then, the contribution score of term t in page o to page q is $\alpha_t^1(o, p)\alpha_t^1(p, q)R_t(q)$. According to equation (13), $C_t(p) = \beta_t(p)\alpha_t^1(p, q)R_t(q)$. Then, the sum of the two contribution scores to page q is equal to $\alpha_t^1(p, q)R_t(q)$ since $\alpha_t^1(o, p) + \beta_t(p) = 1$ according to equations (10) and (14). Consequently, the sum of contribution scores is not changed by the additional link insertion. This induction can be easily generalized to the case where page p has multiple outlinks and inlinks and also to the case where pages are connected by a path.

Compared with the first strategy that treats the contribution score of a page to other pages, the second strategy focuses on the contribution scores of the third pages to the page. According to the first strategy and equation (10), given a page p, the sum of contribution scores of the third pages to page p, $U_t(p)$, is computed as equation (15). Initially, a PC-Rank of a term in page p is assigned by the relevance score of page p to the term. Then, we reduce the PC-Rank by $\lambda U_t(p)$. Accordingly, the sum of the final PC-Ranks of all pages under term t is the sum of relevance scores of all the pages to term t.

$$\begin{aligned} U_t(p) &= \frac{\sum_{r \in outlink(q)} R_t(r)}{R_t(p) + \sum_{r \in outlink(q)} R_t(r)} R_t(p) \\ &= [1 - \beta_t(p)] \times R_t(p) \end{aligned} \tag{15}$$

Based on the two strategies, C-Rank can be modified as below:

$$CR_t(p) = [1 - \lambda(1 - \beta_t(p))] \times R_t(p) + \lambda\beta_t(p) \sum_d \sum_{q \in D(p, d)} \alpha_t^d(p, q)R_t(q) \tag{16}$$

Keeping consistency with $CR_t(p) = R_t(p)$ for $\lambda = 1$ in equation (12), we define the probabilistic model, PC-Rank, as follows:

$$PCR_t(p) = \gamma_t(p)NR_t(p) + (1 - \lambda)\beta_t(p) \sum_d \sum_{q \in D(p, d)} \alpha_t^d(p, q)NR_t(q) \tag{17}$$

where $0 \leq \lambda \leq 1$, $\gamma_t(P) = 1 - (1 - \lambda) \times (1 - \beta_t(p))$ and $NR_t(p) = R_t(p) / \sum_{s \in W} R_t(s)$.

3.2.3. Hybrid model: HC-Rank. The probabilistic model causes some loss of a relevance score so that the importance of a term in a page might be underestimated. To prevent such underestimation, we can think of the hybrid model (HC-Rank) which combines the basic and the probabilistic models as below:

$$HCR_t(p) = R_t(p) + \lambda\beta_t(p) \sum_d \sum_{q \in D(p, d)} \alpha_t^d(p, q)R_t(q) \tag{18}$$

where $\lambda \geq 0$.

The sum of all HC-Ranks increases whenever a link is added to a Web graph. However, according to the first strategy used in the probabilistic model, the sum of all contribution scores is bounded above by $\sum_{p \in W} R_t(p)$. To keep consistency with $CR_t(p) = R_t(p)$ for $\lambda = 1$ in equations (12) and (17), we define the hybrid model, HC-Rank, by the following:

$$HCR_t(p) = \lambda R_t(p) + (1 - \lambda)\beta_t(p) \sum_d \sum_{q \in D(p, d)} \alpha_t^d(p, q)R_t(q) \tag{19}$$

where $0 \leq \lambda \leq 1$.

3.3. Contributing terms

Consider two pages, p and q, connected as $q \rightarrow p$. According to equation (9), the contribution score of term t in page p to page q depends on the relevance score of term t to page q. Therefore, if page q does not contain term t, term t in page p does not contribute to page q. In this paper, we do not consider the relation between synonyms such as 'car' and 'automobile'. The extension to the synonym issues will be tackled in future work.

The goal of a Web search engine is to provide a user with the most relevant pages to a query. Usually, both a Web page author and a Web search engine user expect a page that has the query terms as its keywords summarizing the page to be ranked high. They might not care about those pages that do not contain the query terms as their keywords. Therefore, we consider only the keywords of each page as candidate terms that can contribute to other pages and also safely ignore other non-keyword terms when computing contribution scores. According to this restriction and the contribution constraints, keyword t in page p contributes to page q only if (1) there is a path from page q to page p and (2) every page on the path contains keyword t as one of its keywords. In this paper, we call this path the *contribution path of term t in page p*.

3.4. C-Rank computation

Computations of C-Rank, PC-Rank and HC-Rank follow the same procedure. In this paper, we explain the procedure on C-Rank. The C-Rank of every term in every page is computed offline. Then, given a query Q possibly consisting of multiple terms, the C-Rank for page p under query Q is determined online by summing the pre-computed C-Ranks of all the query terms for page p as follows: $CR_Q(p) = \sum_{t \in Q} CR_t(p)$. In this section, we describe the way of computing a C-Rank of a term in a page.

A C-Rank can be computed by score propagation. Consider three pages connected by links as $r \rightarrow q \rightarrow p$. According to equations (9) and (11), the contribution score of keyword t in page p to page r is $\alpha_t^1(p, q)\alpha_t^1(q, r)R_t(r)$. For keyword t in page q to page r, it is $\alpha_t^1(q, r)R_t(r)$. Consequently, under keyword t, the contribution score of page p to page r can be determined by propagating the contribution score of page q to page r, $\alpha_t^1(q, r)R_t(r)$, attenuating it by the ratio of $\alpha_t^1(p, q)$. This computation can be generalized to the two pages indirectly connected by a path of length > 1. Therefore, it is possible to compute all contribution scores by considering pairs of pages only connected by a link and propagating their contribution scores. We restrict the maximum length of paths through which a page is allowed to contribute to others. In this paper, we denote it by the *cutoff path length*.

The C-Rank computation algorithm consists of three stages: initialization, iteration and finishing. In the initialization stage, first, the relevance scores of all the pages to each term are computed and C-Ranks of all terms in each page are initialized to zero. Second, a set of keywords of each page is determined according to the relevance scores. Finally, for each pair of page q and page p connected as $q \rightarrow p$, their common keywords are retrieved and then, for each keyword t in the common keywords, the contribution ratio of $\alpha_t^1(p, q)$ and the contribution score of $\alpha_t^1(p, q)R_t(q)$ are computed.

In the iteration stage, first, for each page q, the contribution score of every page p pointed to by page q, which is computed in the initialization stage, is transferred to page p. Consequently, every page receives its contribution score. Let the *absorption ratio* be the ratio of a contribution score accumulated into the final score to a received contribution score. The absorption ratios of keyword t of page p of C-Rank, PC-Rank and HC-Rank are $1 - \lambda$, $(1 - \lambda)\beta_t(p)$ and $(1 - \lambda)\beta_t(p)$, respectively. Second, each page accumulates the received contribution scores multiplied by its absorption ratio. Third, each page p propagates the received contribution scores of each keyword of page p to each page o connected by an outlink, after being multiplied by the contribution ratio of $\alpha_t^1(o, p)$. Consequently, the ratio of the propagated scores to the received scores becomes $1 - \beta_t(p)$. The second and third steps are repeated until (1) the number of repetitions reaches a given number or (2) all the pages stop propagating contribution scores. Each page stops propagating if the sum of received contribution scores of every keyword is below a given threshold.

Now, we explain the finishing stage. Let the preservation ratio be the ratio of the portion of a relevance score added into the final score to a relevance score. For keyword t, the preservation ratios of C-Rank, PC-Rank and HC-Rank are λ, $\gamma_t(p)$ and λ, respectively. For a term which is not a keyword, they are λ, 1 and λ, respectively. For each keyword and each non-keyword in each page, its relevance score multiplied by its preservation ratio is added to its final score.

The above procedure does not show how to construct a path that consists of distinct pages. To construct the path, we attach a tag on each propagation score. The tag contains the path information through which the contribution score is propagated. Before a contribution score is propagated to another page, we check whether the insertion of the page to the path stored in the tag induces a cycle. If a cycle occurs, we shall not propagate the contribution score to the page.

3.5. Comparison of ranking measure

To clarify the difference between our approach and previous probabilistic models described in Section 2, we compare the ranking measures of our probabilistic model (PC-Rank) with the previous probabilistic models. As described in Section 3.4, a PC-Rank can be computed by score propagation. Let $P_t^I(p)$ be the sum of received contribution scores of keyword t in page p over all iteration steps until the repetition of propagation stops. On PC-Rank, a received contribution score of keyword k in page p is divided to two parts: absorbed score and transferred score. The absorbed score is

accumulated into PC-Rank and the transferred score is re-propagated into other pages. Also, a PC-Rank of keyword k in page p is decreased by contribution scores which page p gives to other pages corresponding to $(1-\lambda)(1-\beta_t(p))NR_t(p)$ in equation (17). Therefore, PC-Rank of equation (17) can be expressed as follows:

$$P_t(p) = P_t^0(p) - (1-\lambda)P_t^0(p)(1-P_t(p\to p)) + (1-\lambda)P_t^I(p)P_t(p\to p) \tag{20}$$

$$P_t^I(p) = \sum_{q\in inlink(p)} [P_t^0(q) + P_t^I(q)] \times P_t(q\to p) \tag{21}$$

where $P_t^0(p) = NR_t(p)$, $P_t(q\to p) = \alpha_t^1(p,q)$, $P_t(p\to p) = \beta_t(p)$, and $P_t(q\to q) + \sum_{p\in outlink(q)} P_t(q\to p) = 1$. Equation (20) shows that $P_t(p)$ depends on $P_t(p\to p)$, that is, the amount of absorbed scores from contribution scores passing through page p.

To compare PC-Rank with PageRank, let us consider the special case that, in all the pages, all relevance scores are completely disregarded and all the received contribution scores are propagated to other pages without absorption. Then, in equation (20), $P_t^0(p) = 0$ and $P_t(p\to p) = 0$. Therefore, $P_t(p)$ is zero for every term in every page. It means that PC-Rank cannot be used as the ranking measure of pages in the special case. Meanwhile, equation (21) becomes concise as below:

$$P_t^I(p) = \sum_{q\in inlink(p)} P_t^I(q)P_t(q\to p), \quad \sum_{p\in outlink(q)} P_t(q\to p) = 1 \tag{22}$$

Equation (22) corresponds to the original concept of PageRank of equation (1). As described in Section 2, to overcome the problem of rank sinks, Page *et al*. [13] suggested the random surfer model of equation (2) containing the random jump component. However, the ranking measure of PC-Rank does not require the random jump.

To compute PageRank, equation (2) is repeated until the variances of scores of all the pages at each iteration step are under a given threshold. This means that the scores reach convergence. Meanwhile, the PC-Rank computation stops when newly received contribution scores are under a given threshold. In conclusion, the ranking measure of PageRank is the amount of current probability flow through each page when the probability flow through every page is in equilibrium, that is, the amount of probability flow is not changed further. However, the PC-Rank measure is the amount of absorption of each page from the probability flow until no probability flow occurs. C-Rank and HC-Rank are not probabilistic models, but these models use ranking measures similar to that of PC-Rank, except using different absorption ratios and preservation ratios.

Similar to probabilistic approaches based on the surfer model, relevance propagation models [17, 20, 21] use the amount of flow as the ranking measure even though these models propagate relevance scores of pages to a given query instead of an authority score and utilize both inlinks and outlinks. Unlike the previous surfer models, our approach does not have a random jump component and the amount of score flow is attenuated by the ratio of $1-\beta_t(p)$ at each iteration step. Consequently, those pages that are connected by a long path or are not connected do not have an effect on one another.

3.6. Discussions

C-Rank and its variants have two practical advantages. First, our approach is robust against a cycle or a spam farm created by a spammer. The spammer tries to manipulate ranks of some Web pages by boosting their authority scores intentionally, using a weakness of previous approaches. The weakness is that most previous approaches allow the authority score of a Web page to be propagated to itself at least once when it is directly or indirectly connected to itself. However, our approach could avoid such boosting of the authority scores because it does not allow a Web page to contribute to itself. Also, when Web pages in a cycle or a spam farm are not related or less related to one another, our approach rarely propagates a relevance score of a Web page to the other Web pages in a cycle or a spam farm. Additionally, both a PC-Rank and an HC-Rank of a Web page become smaller as the Web page has more outlinks. That is, no matter how many outlinks a Web page has in a cycle or a spam farm, the Web page does not heighten its authority score. These properties of our approach can counteract the gains from a cycle or a spam farm.

Second, our approach can dynamically update ranks of Web pages. When some Web pages are newly added or removed, most previous approaches re-compute the ranks of all the pages in a Web graph. On the other hand, our approach needs to re-compute the ranks of only some Web pages associated to the updated (added or removed) Web pages rather than all ones. This is owing to the two properties of our approach: the cutoff path length and no random jump. The cutoff path length allows a Web page to contribute to other Web pages that are directly or indirectly

connected by a path of length less than or equal to cutoff path length. In addition, since our approach does not perform the random jump, it requires the re-computations of ranks only for those Web pages whose path lengths to the updated one are less than or equal to the cutoff path length. Using those properties, we expect to develop a dynamic update mechanism to keep up-to-date ranks of all the Web pages effectively. We think that this dynamic update issue could be quite an interesting topic as the future work.

4. Experimental results

4.1. Experimental environment

For our experiments, we used the .GOV collection that is constructed with 1.25 million pages and 11.2 million links. As our query sets, we used two sets of topic distillation task topics in Web Track of TREC-2003 and TREC-2004 whose numbers of topics are 50 and 75, respectively.

Our experiments were conducted on the public search engine-Lucene.[1] The baseline accuracy was evaluated by pure relevance scores, provided in Lucene, of terms to a page containing all query terms. As performance measures, precision at 10 (P@10) and mean average precision (MAP) were used. The set of N keywords of a page is constructed with N terms that have the topmost relevance scores in the page.

4.2. Effectiveness evaluation

Precisions of C-Rank and its variants are affected by three factors: the weight value of λ, the number of keywords and the *cutoff path length*. The cutoff path length indicates the maximum length of paths through which a page is allowed to contribute to others. It decides the number of repetitions for computing contribution scores.

4.2.1. Precision depending on three factors. To describe the influence of the three factors on precision, we first show the experimental results conducted with C-Rank. Figure 1 shows C-Rank precisions with 10 keywords according to the cutoff path length 3, 5, 7 and 9. In almost all cases, the cutoff path length 3 gave the best results. As two pages are further apart, it is less likely that the two pages have a common topic. Therefore, it is reasonable to set the cutoff path length to a value between 3 and 5.

Figure 2 shows C-Rank precisions of in P@10 and MAP with the cutoff path length 3 according to 10, 20, 30, 40, 50 and 60 keywords. Overall, P@10 with 10 keywords was better than the others. Meanwhile, for MAP, using 20 or 40 keywords generated better accuracies. From these results, we can infer two things: first, as more keywords are used, more relevant pages move upward in ranking; second, concurrently, non-relevant pages move upward and, consequently, P@10 decreases. The keywords of a page were chosen with those terms of the topmost relevance scores in the page. We do not think that the keywords currently used are the best terms summarizing the page. Therefore, we expect that the accuracies can be more improved if those keywords are carefully constructed by employing more sophisticated methods [35].

Behaviours of precisions for combinations of the number of keywords and the cutoff path length are analogous. We list in Table 2 the best MAP and P@10 in each combination. For example, P@10 = 0.110 in Table 2, which is for the case of 10 keywords and cutoff path length 3 in TREC-2003 dataset, indicates the maximum value of P@10 for $L3$ in Figure 1(a) where the value of λ is set as 0.8. The number in parentheses shows the percentage improvement over the baseline. We observe that C-Rank greatly improves the effectiveness compared with the baseline. As the cutoff path length or the number of keywords becomes larger, more computation time is required. Therefore, we recommend the combinations of the cutoff path length set to 3–5 and the number of keywords set to 10–30 to apply C-Rank to real-world search engines.

4.2.2. Comparison of C-Rank, PC-Rank and HC-Rank. To compare C-Rank and its variants, we conducted experiments on .GOV with 10 keywords and the cutoff path length 3. Figure 3 shows that C-Rank gave the best results both on TREC-2003 and TREC-2004. Meanwhile, PC-Rank was the worst. This result coincides with those in other studies discovering that un-normalized scores often work better than normalized ones [39–41]. We conjecture that this is caused by the preservation ratio $\gamma_t(p)$ of PC-Rank which generates the case that, if a page of a high relevance score has a small number of inlinks and a large number of outlinks, its PC-Rank might go down less than a PC-Rank of another page of a low relevance score which has a small number of outlinks.

4.2.3. Precision decrease by cycles. A contribution score might be overestimated if we admit a term in a page to contribute to the page multiple times through a cycle. Figure 4 shows the dependency of precisions on cycles. Even though a cycle

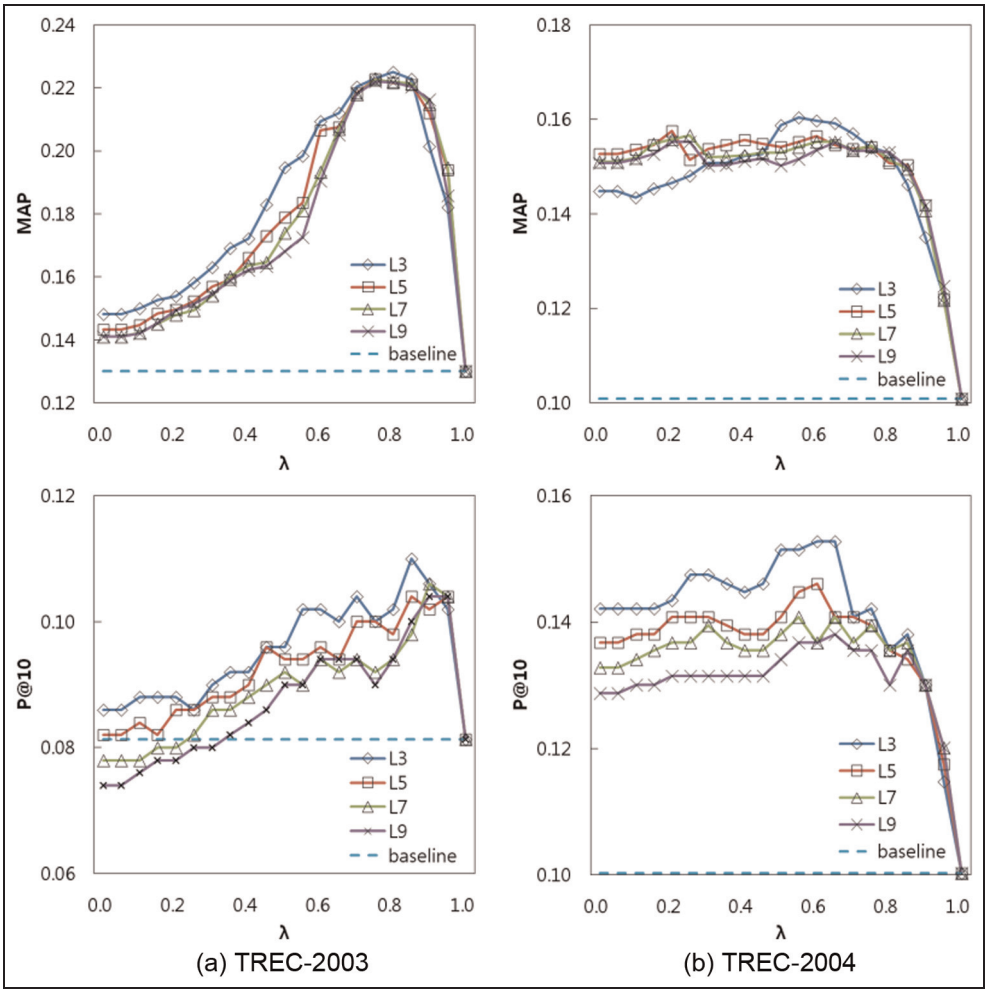

Figure 1. C-Rank precisions according to the cutoff path length with 10 keywords. L3, L5, L7 and L9 mean the cutoff path lengths 3, 5, 7 and 9, respectively.

is admitted, the C-Rank precision does not go down below the baseline precision. However, the decrease in precision is significant.

The computational cost for cycle-break is highly dependent on the cutoff path length. However, we should note that, according to our experimental results, using the cutoff path length 3–5 provides good precisions. Therefore, a cycle-break operation is applicable to real-world search engines with this small cutoff path length.

4.2.4. Precision comparison with previous models. Relevance propagation models have been most effective among previous models exploiting link and content information [6, 17, 21]. However, the models require too much computation time, at querying time, to be applied to real-world search engines [6]. To show that C-Rank is effective, we compare C-Rank with relevance propagation models in terms of effectiveness.

Figure 5 shows that the results of C-Rank with 10 keywords are the best or close to the best on each set of topics in both P@10 and MAP. In particular, the results of C-Rank are very dominant in MAP on TREC-2003. Meanwhile, the results of WIO were the best on both TREC-2003 in P@10 and TREC-2004 in MAP and close to the best on TREC-2004 in P@10. The results of HT-WI were not bad, but it did not generate the best results. HT-WO gave results that were worse than others. From those results, we can observe that C-Rank and WIO are comparable to each other in effectiveness.

Finally, we compare C-Rank results with official runs submitted to TREC [34, 35]. Table 3 shows the three best precisions among both official runs and C-Rank results obtained from Table 2. In MAP on TREC-2003, C-Rank results

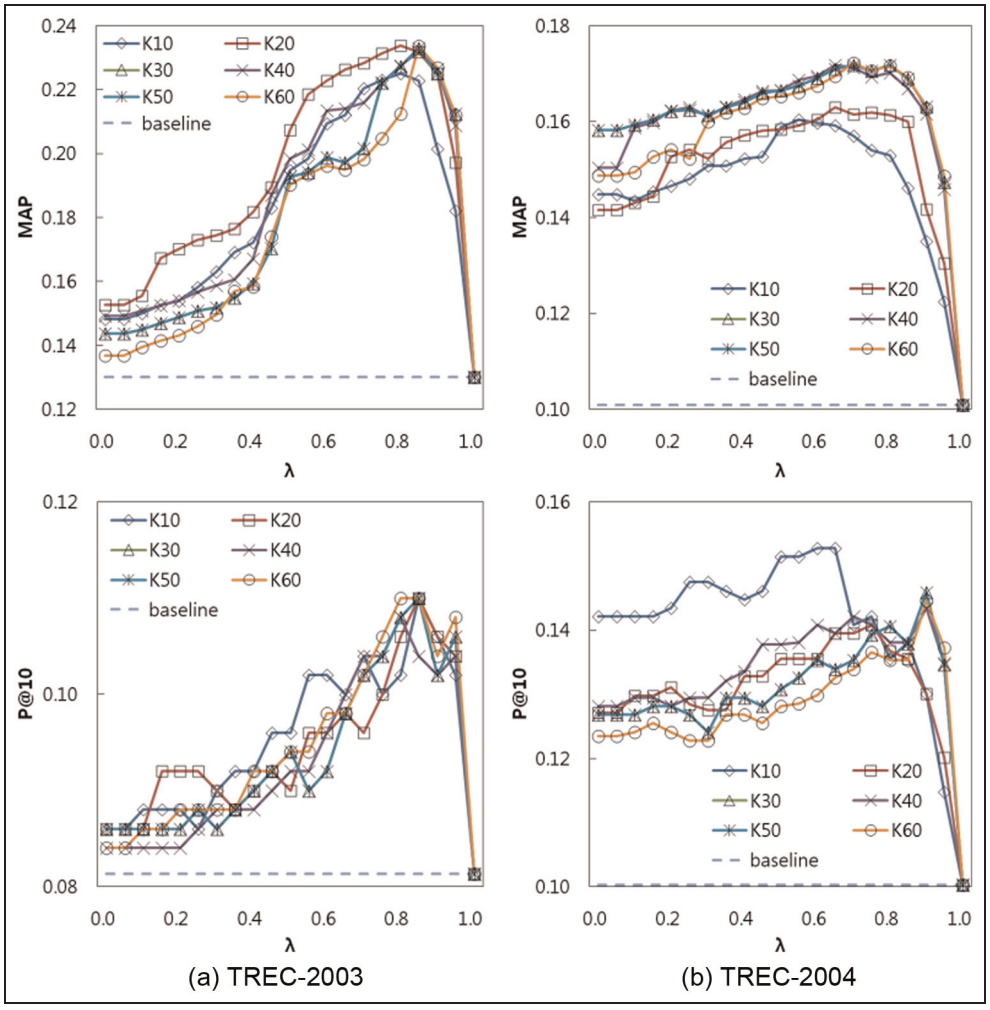

Figure 2. C-Rank precisions according to the number of keywords with the cutoff path length 3. K10, K20, K30, K40, K50 and K60 mean 10, 20, 30, 40, 50 and 60 keywords, respectively.

Table 2. Best performances of C-Rank according to the cutoff path length *L* and the number of keywords *K*

Model		TREC-2003		TREC-2004	
		Best P@10	Best MAP	Best P@10	Best MAP
Baseline		0.081	0.130	0.100	0.101
L = 3	*K* = 10	0.110(35%)	0.225(73%)	0.153(52%)	0.160(58%)
	K = 20	0.110(35%)	0.234(80%)	0.141(40%)	0.162(60%)
	K = 30	0.110(35%)	0.233(79%)	0.146(46%)	0.172(70%)
	K = 40	0.108(33%)	0.232(78%)	0.144(43%)	0.172(70%)
	K = 50	0.110(35%)	0.233(79%)	0.146(46%)	0.172(70%)
	K = 60	0.110(35%)	0.234(80%)	0.145(44%)	0.172(70%)
L = 5	*K* = 10	0.104(28%)	0.223(71%)	0.146(46%)	0.158(56%)
	K = 20	0.110(35%)	0.233(79%)	0.138(38%)	0.163(61%)
	K = 30	0.106(30%)	0.234(80%)	0.138(38%)	0.166(64%)
	K = 40	0.104(28%)	0.226(74%)	0.144(43%)	0.170(68%)
	K = 50	0.106(30%)	0.228(75%)	0.143(43%)	0.171(69%)
	K = 60	0.108(33%)	0.228(75%)	0.139(39%)	0.171(69%)

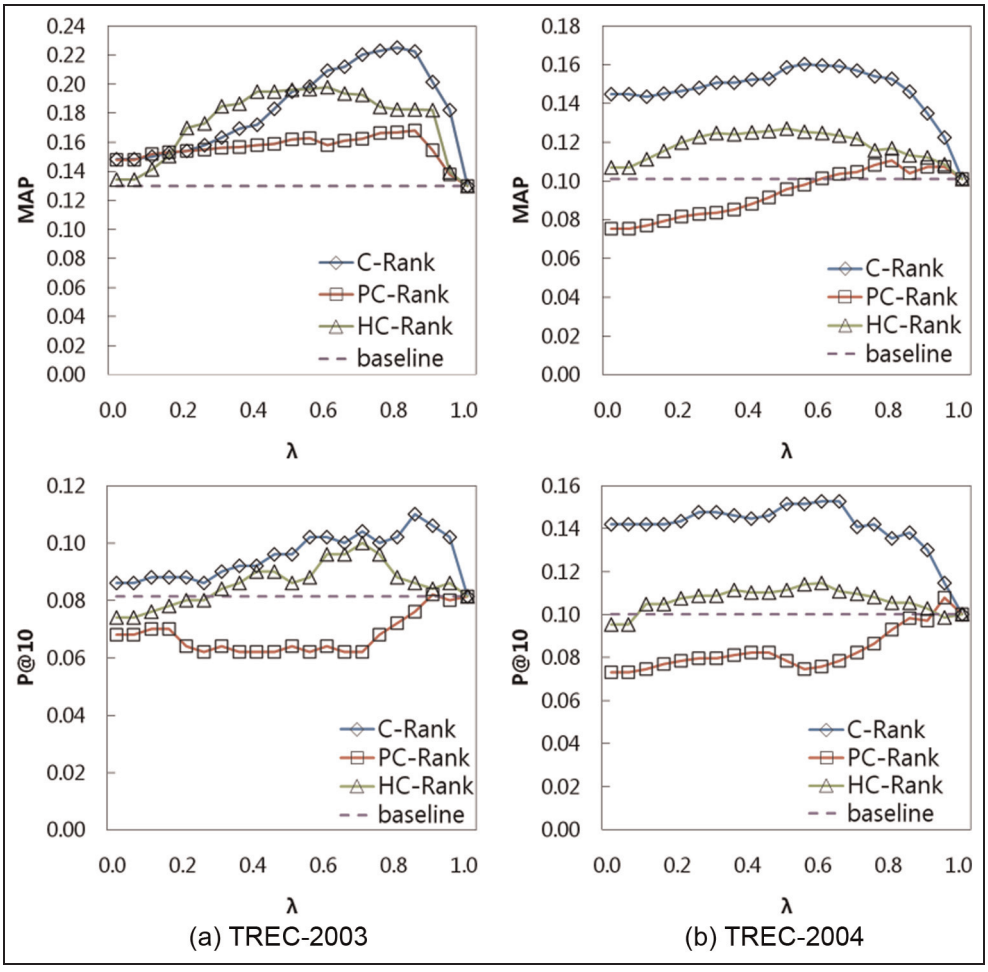

Figure 3. Precision comparison among C-Rank and its variants with 10 keywords and the cutoff path length 3.

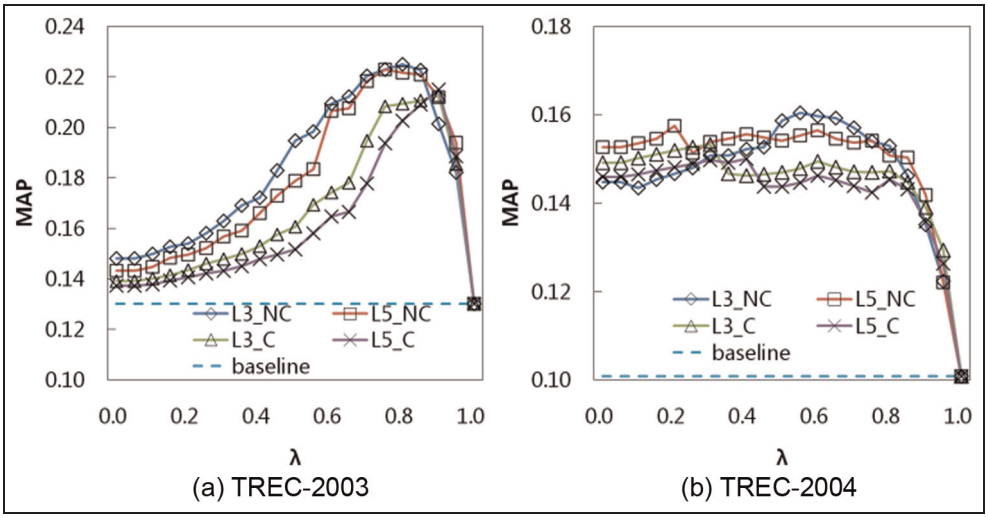

Figure 4. Precision comparison of cycle-break and cycle-admission with 10 keywords and the cutoff path length 3 or 5 on C-Rank. NC and C mean cycle-break and cycle-admission, respectively.

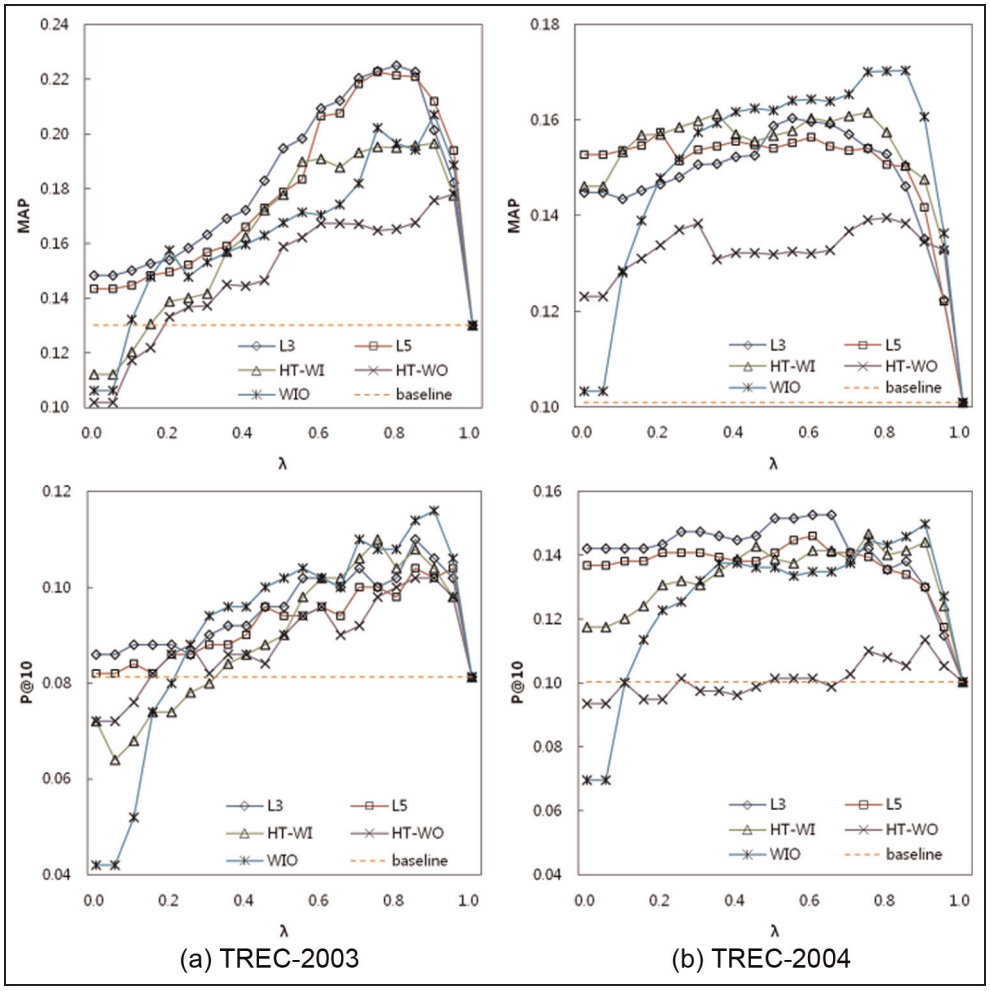

Figure 5. Performance comparison of C-Rank and relevance propagation models.

Table 3. Best performances comparison

	Model	Three best P@10	Three best MAP
TREC-2003 (Topic Distillation)	C-Rank	0.110–0.110	0.234–0.234
	Official runs	0.122–0.128	0.134–0.154
TREC-2004 (Topic Distillation)	C-Rank	0.146–0.153	0.172–0.172
	Official runs	0.231–0.249	0.165–0.179

greatly overwhelm the three best official runs. Both in MAP on TREC-2004 and P@10 on TREC-2003, C-Rank results are close to the three best official runs. Meanwhile, in MAP on TREC-2004, C-Rank results are worse than the three best official runs. Official TREC runs utilized a variety of extra features such as url types, anchor texts, a page structure and so on. However, in this paper, we did not employ those features to focus on combining links and content. Therefore, we expect the effectiveness of C-Rank to be much improved if the above features are exploited in C-Rank.

4.3. Efficiency evaluation

To be used in real-world search engines, efficiency is another important factor besides effectiveness. In this section, we investigate efficiencies of C-Rank, HT-WI, and WIO in two aspects of computation: online and offline.

Table 4. Online average computation time per a term

Model	C-Rank	HT-WI	WIO
Time usage (s)	0.00	5.92	18.16

Table 5. Offline computation time of C-Rank with the cutoff path length 3

Number of keywords	10	20	30	40
Time usage (h)	0.11	0.30	0.69	1.72

The C-Rank of every term in each page is computed offline. Therefore, given a query, C-Rank spends the same time as the baseline method using only a relevance score. Consequently, C-Rank requires no additional online time consumption for exploiting link and content information. Meanwhile, relevance propagation models require a lot of extra time online for constructing a working set of pages relevant to a given query and propagating a relevance score or a term frequency to the relevant pages. To measure efficiency, we logged a time usage on 1.6 GHz CPU and 8 GB memory.

Table 4 shows average time usages of C-Rank, HT-WI and WIO for computing one term score, excluding I/O access time and score sorting time, which are identical over all models. As an advantage to the relevance propagation models, the number of iterations for propagating a relevance score or a term frequency is restricted below 10. That is, if the number of iterations becomes 10, the models stop even though relevance scores do not reach convergence. Additionally, we assume all pages in a working set are located in the same server to ignore the data transfer time among servers for propagating relevance scores or term frequencies. As a set of test queries, 50 terms in TREC-2003 and TREC-2004, were chosen at random. The zero average time usage for C-Rank in Table 4 indicates that C-Rank does nothing besides I/O access and score sorting. HT-WI and WIO spent 5.92 and 18.16 s per term on average, respectively. It is obvious that C-Rank is more efficient than HT-WI and WIO in online computation. Also, the results show that the relevance propagation models are rarely feasible for application in real-world search engines.

The offline computation time of C-Rank depends on the number of keywords and the cutoff path length, assuming that all the relevance scores of pages to each term are pre-computed. As shown in Figure 1, the cutoff path length 3 gave best results. Therefore, we measured offline time usage in computing C-Rank according to the number of keywords with the cutoff path length 3 on the .GOV collection. From Table 5, we conclude that C-Rank is efficient enough to be applied to real-world search engines.

Compared with the baseline method, the relevance propagation models require heavy online computation, but they do no additional offline computation. One might think about the way of pre-executing such online computation. However, a query can consist of multiple terms so that the number of candidate queries becomes almost infinite. Therefore, it is impossible for the relevance propagation models to build rankings of pages offline under all candidate queries.

5. Conclusions

In this paper, we proposed the contribution-based ranking approach exploiting link and content information with efficiency high enough to be applicable to real-world search engines: C-Rank, PC-Rank and HC-Rank. Based on the viewpoint of a Web page author, we suggested the contribution constraints, and then formulated the contribution score of a term in a page satisfying them.

Our approach, to the best of our knowledge, is the first to adopt the viewpoint of a Web page author. As a result of moving a viewpoint from a Web surfer to a Web page author, a contrived modification such as a random jump used in random surfer models is not required. A Web page author creates his/her Web page without considering all the Web pages in the Internet. Thus, newly created Web pages affect the scores of only these pages associated with their links rather than all the pages in the Internet. Therefore, we claim that our approach reflects the real-world Web environment better than random surfer models.

To clarify the difference between our approach and previous probabilistic models such as random surfer models, we compared the ranking measures of PC-Rank with PageRank. The ranking measure of probabilistic approaches based on the surfer model is the amount of authority scores currently flowing into a page at the time all the pages reach an equilibrium state. Our ranking measure corresponds to the amount of authority scores absorbed by a page until there is no flow

of an authority score through the page. If insertion/deletion of a page into/from a real-world Web graph occurs, probabilistic models should update the scores of all the pages while our approach updates the scores of only the pages near the page inserted/deleted.

To improve efficiency without loss of effectiveness, we utilize the expectations of both a Web page author and a Web search engine user on retrieval results simultaneously, and restrict candidate terms that can contribute to other pages to a small number of keywords of each page. C-Rank, PC-Rank and HC-Rank are computed by combining the contribution score of a term in a page and the relevance score of the page to the term. The computations for every term in every page are done offline so that the computational cost at querying time is the same as the cost of the models using only relevance scores.

Compared with relevance propagation models that construct a working set at querying time in order to simultaneously exploit link and content information, our approach utilizes the expectations of both a Web page author and a Web search engine user. As a result, no additional cost is required at querying time. This means that our approach clearly solves the performance problem of relevance propagation models while maintaining the merits of exploiting link and content information simultaneously.

To show the effectiveness and efficiency of our approach, we conducted extensive experiments on the .GOV collection with two sets of topic distillation task topics. Experimental results show that C-Rank gives the best precisions among the proposed models. For effectiveness, C-Rank with 10 keywords and the cutoff path length 3 performed the best or was very close to the best compared with the relevance propagation models in both P@10 and MAP. From these results, we understand that our basic concepts of focusing on keywords and adopting the scores of only the pages associated by links are effective. For efficiency, the experimental results show that our approach is very efficient, but the relevance propagation models comparable to C-Rank in terms of effectiveness can rarely be applied to real-world search engines owing to their lack of efficiency. From our results, we conclude that the proposed contribution-based ranking approach is quite effective and also efficient enough to be applicable to real-world search engines.

There are several interesting directions for future works. First, our approach does not differentiate Intranet from Internet. Meanwhile, the previous works [17, 22] show that a site-map based model is effective in a topic distillation task for finding homepages. We think it is worthwhile to employ site-map information in our approach. Second, in this paper, we extracted the keywords of a page in a naive way of selecting those terms having the topmost relevance scores to the page. However, we do not think that these keywords are the best terms summarizing the page. We expect that the effectiveness of C-Rank can be much more improved if keywords are chosen with more sophisticated techniques [42]. Third, treating the relation between synonyms such as ‘car’ and ‘automobile’ for a contribution score might be helpful in improving the effectiveness. Fourth, our experimental results show that it is enough to consider relations among pages that are close to one another. It means that it is possible to insert/delete a page into/from a Web graph without recalculating scores of all pages. In a real Web environment, many pages are daily generated and deleted. Therefore, it is very worthwhile to develop a method that dynamically updates the scores of only the pages associated with inserted/deleted pages.

Acknowledgements

This research was supported by the Information Technology Research Center support program (NIPA-2014-H0301-14-1022) supervised by the National IT Industry Promotion Agency, a National Research Foundation of Korea grant funded by the Korea government (MEST; no. 2011-0029181), and Semiconductor Industry Collaborative Project between Hanyang University and Samsung Electronics.

Note

1. Lucene, http://lucene.apache.org.

References

[1] Kostoff R. Expanded information retrieval using full-text searching. *Journal of Information Science* 2010; 36(1): 104–113.
[2] Almpanidis G, Kotropoulos C and Pitas I. Combining text and link analysis for focused crawling – An application for vertical search engines. *Information Systems* 2007; 32(6): 886–908.
[3] Baeza-Yates R, Boldi P and Castillo C. Generalizing PageRank: Damping functions for link-based ranking algorithms. In: *Proceedings of ACM SIGIR conference*, Seattle, WA, 6–11 August 2006, pp. 308–315.
[4] Bidoki A and Yazdani N. DistanceRank: An intelligent ranking algorithm for Web pages. *Information Processing and Management* 2008; 44(2): 877–892.

[5] Boldi P, Santini M and Vigna S. PageRank as a function of the damping factor. In: *Proceedings of international World Wide Web conference*, Chiba, Japan, 10–14 May 2005, pp. 557–566.
[6] Chibane I and Doan B. Relevance propagation model for large hypertext document collections. In: *Proceedings of recherche d'information assistee par ordinateur*, Pittsburgh, PA, 30 May to 1 June 2007, pp. 585–595.
[7] Guo Y. MixPR – An approach of combining content and links of Web page. In: *Proceedings of international conference on fuzzy systems and knowledge discovery*, Haikou, China, 24–27 August 2007, pp. 456–460.
[8] Haveliwala T. Topic-sensitive PageRank. In: *Proceedings of international World Wide Web conference*, Honolulu, HI, 7–11 May, 2002, pp. 517–526.
[9] He B, Huang J and Zhou X. Modeling term proximity for probabilistic information retrieval models. *Information Sciences* 2011; 181(14): 3017–3031.
[10] Macdonald C and Ounis I. The influence of the document ranking in expert search. *Information Processing and Management* 2011; 47(3): 376–390.
[11] Nie L, Davison B and Qi X. Topical link analysis for Web search. In: *Proceedings of ACM SIGIR conference*, Seattle, WA, 6–11August 2006, pp. 91–98.
[12] Nie L and Davison B. Separate and Inequal: Preserving heterogeneity in topical authority flows. In: *Proceedings of ACM SIGIR conference*, Singapore, 20–24 July 2008, pp. 443–450.
[13] Page L, Brin S, Motwani R and Winograd T. The PageRank citation ranking: Bringing order to the Web. Technical Report, Stanford Digital Libraries, 1998.
[14] Pal S and Narayan B. A Web surfer model incorporating topic continuity. *IEEE Transactions on Knowledge and Data Engineering* 2005; 17(5): 726–729.
[15] Plachouras V and Ounis I. Usefulness of hyperlink structure for query biased topic distillation. In: *Proceedings of ACM SIGIR conference*, Sheffield, 25–29 July 2004, pp. 448–455.
[16] Plachouras V, Ounis I and Amati G. The static absorbing model for the Web. *Journal of Web Engineering* 2005; 4(2): 165–186.
[17] Qin T, Liu T, Zhang X, Chen Z and Ma W. A study of relevance propagation for Web search. In: *Proceedings of ACM SIGIR conference*, Salvador, Brazil, 15–19 August 2005, pp. 408–415.
[18] Rafiei D and Mendelzon A. What is this page known for? Computing Web page reputations. In: *Proceedings of international World Wide Web conference*, Amsterdam, 15–19 May 2000, pp. 823–835.
[19] Richardson M and Domingos P. The intelligent surfer: Probabilistic combination of link and content information in PageRank. In: *Proceedings of neural information processing systems*, Vancouver, 3–8 December 2001, pp. 1441–e1448.
[20] Shakery A and Zhai C. Relevance propagation for topic distillation UIUC TREC-2003 Web track experiments. In: *Proceedings of the text retrieval conference*, Gaithersburg, MD, 17–22 November 2003, pp. 673–677.
[21] Shakery A and Zhai C, A probabilistic relevance propagation model for hypertext retrieval. In: *Proceedings of the international conference on information and knowledge management*, Arlington, VA, 6–11 November 2006, pp. 550–558.
[22] Song R, Wen J, Shi S, Xin G, Liu T, Qin T, Zheng X, Zhang J, Xue G and Ma W, Microsoft Research Asia at Web track and terabyte track of TREC 2004. In: *Proceedings of the text retrieval conference*, Gaithersburg, MD, 16–19 November 2004, pp. 1–13.
[23] You G and Hwang S. Search structures and algorithms for personalized ranking. *Information Sciences* 2008; 178(20): 3925–3942.
[24] Weninger T, Bisk Y and Han J. Document-topic hierarchies from document graphs. In: *Proceedings of the international conference on information and knowledge management*, Maui, HI, 29 October to 2 November 2012, pp. 635–644.
[25] Wu D and Mendel J. A comparative study of ranking methods, similarity measures and uncertainty measures for interval type-2 fuzzy sets. *Information Sciences* 2009; 179(8): 1169–1192.
[26] Kurland O and Lee L. PageRank without hyperlinks: structural re-ranking using links induced by language models. In: *Proceedings of ACM SIGIR conference*, Salvador, Brazil, 15–19 August 2005, pp. 306–313.
[27] Yoon S, Kim J, Kim S and Lee C. TL-Rank: A blend of text and link information for measuring similarity in scientific literature database. *Information Processing Letters* 2012; E95-D(10): 2556–2559.
[28] Hamedani M, Lee S, Kim S and Kim D. On exploiting content and citations together to compute similarity of scientific papers. In: *Proceedings of ACM international conference on information and knowledge management*, San Francisco, CA, 27 October to 1 November 2013, pp. 1553–1556.
[29] Craswell N, Robertson S, Zaragoza H and Taylor M. Relevance weighting for query independent evidence. In: *Proceedings of ACM SIGIR conference*, Salvador, Brazil, 15–19 August 2005, pp. 416–423.
[30] Upstill T, Craswell N and Hawking D. Query-independent evidence in home page finding. *ACM Transactions on Information Systems* 2003; 32(3): 286–313.
[31] Wu M, Hawking D, Turpin A and Scholer F. Using anchor text for homepage and topic distillation search tasks. *Journal of the American Society for Information Science and Technology* 2012; 63: 1235–1255.
[32] Wu M, Scholer F and Turpin A. Topic distillation with query-dependent link connections and page characteristics. *ACM Transactions on the Web (TWEB)* 2011; 5(2): 6.
[33] Tsikrika T and Lalmas M. Best entry pages for the topic distillation task. Queen Mary, Technical Report, University of London, Department of Computer Science, 2005.

[34] Craswell N, Hawking D, Wilkinson R and Wu M. Overview of the TREC 2003 Web track. In: *Proceedings of the Text Retrieval Conference*, Gaithersburg, MD, 17–22 November 2003, pp. 78–92.
[35] Craswell N and Hawking D. Overview of the TREC 2004 Web track. In: *Proceedings of the Text Retrieval Conference*, Gaithersburg, MD, 16–19 November 2004, pp. 1–9.
[36] Jones K, Walker S and Robertson S. A probabilistic model of information retrieval: Development and comparative experiments (parts 1 and 2). *Information Processing and Management* 2000, 36(6): 779–840.
[37] Davison B. Topical locality in the Web. In: *Proceedings of ACM SIGIR conference*, Athens, 24–28 July 2000, pp. 272–279.
[38] Bondy J and Murthy U, *Graph Theory with Applications*. Amsterdam: Elsevier, 1976.
[39] Bashir S and Rauber A. On the relationship between query characteristics and IR functions retrieval bias. *Journal of the American Society for Information Science and Technology* 2011; 62(8): 1515–1534.
[40] Pal A and Counts S. Identifying topical authorities in microblogs. In: *Proceedings of the fourth ACM international conference on Web search and data mining*, Hong Kong, 9–11 February 2011, pp. 45–54.
[41] Rudoy D and Zelnik-Manor L. Viewpoint selection for human actions. *International Journal of Computer Vision* 2012; 97(3): 243–254.
[42] Wan X, Yang J and Xiao J. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In: *Proceedings of the annual meeting of the association of computational linguistics*, Prague, 23–30 June 2007, pp. 552–559.