Journal of Web Engineering, Vol1. 1, No.4 (2010) 350-364

POPULARITY-BASED RELEVANCE PROPAGATION

EHSAN MOUSAKAZEMI

MEHDI AGHA SARRAM   ALI MOHAMMAD ZAREH BIDOKI

Department of Electrical & Computer Engineering, Yazd University, Yazd, Iran
ehsan.mousakazemi@gmail.com
{mehdi.sarram, alizareh}@yazduni.ac.ir

Received November 14, 2011
Revised May 14, 2012

It is evident that information resources on the World Wide Web (WWW) are growing rapidly with unpredictable rate. Under these circumstances, web search engines help users to find useful information. Ranking the retrieved results is the main challenge of every search engine. There are some ranking algorithms based on content and connectivity such as BM25 and PageRank. Due to low precision of these algorithms for ranking on the web, combinational algorithms have been proposed. Recently, relevance propagation methods as one of the salient combinational algorithms, has attracted many information retrieval (IR) researchers' attention. In these methods the content-based attributes are propagated from one page to another through web graph. In this paper, we propose a generic method for exploiting the estimated popularity degree of pages (such as their PageRank score) to improve the propagation process. Experimental results based on TREC 2003 and 2004 gathered in Microsoft LETOR 3.0 benchmark collection show that this idea can improve the precision of the corresponding models without any additional online complexity.

Key words: Information Retrieval, Ranking, Relevance Propagation, Popularity Measure
Communicated by: B. White & J. Freire

1 Introduction

Nowadays using WWW as the main resource of obtaining information has had an increasing growth. Search engines as the most useful tool in this huge and diverse environment, play a vital role in our everyday life. According to the published information by Alexa web site, Google and Yahoo have attained the first and fourth rank in the web traffic respectively [1].

In general, the three major components of a search engine are: crawler, indexer and searcher [2][3]. Ranking process as the primary part of the searcher module has always been a challenging issue of every search engine. In short, ranking problem is to sort the retrieved documents in response to users' queries. Many different ranking algorithms have already been proposed which can be divided into two broad categories: content-based and connectivity-based algorithms [29].

In traditional IR [2], the ranking strategy is solely based on the content of documents. In other words, for each query the documents with more similar content to the query will be selected as the more relevant ones. Examples of the content-based ranking algorithms are TF-IDF [21], BM25 [20] and etc.

WWW has a hyperlink structure between web pages which creates a massive graph called web graph. Connectivity-based algorithms such as PageRank [17], DistanceRank [28] and HostRank [27] exploit this link structure to estimate the web page popularity (importance) [4] regardless of its content. It is fair to say that all link analysis algorithms are based on this assumption that a link from page A to page B can be considered a recommendation of page B by the author of A.

Using either content-based or connectivity-based algorithms independently, leads to a low-precision ranking function which cannot fully satisfy the users' demands in the web [16]. Therefore combination algorithms which use both content and link structure were introduced. However early efforts in this context considered web content and link structure in two separate stages. For instance [10][5][6] use TF-IDF of the query term in the page to compute a relevance score, and use hyperlinks to compute a query-independent importance score (e.g. PageRank). Finally the rank of retrieved documents is calculated by combining these two scores.

In recent years, relevance propagation methods as one of the salient combinational algorithms, has attracted many IR researchers' attention. In the relevance propagation methods [14, 18, 22, 23, 24], the content-based score or query terms are propagated through hyperlinks or sitemap tree from one page to another. In this paper, we propose a generic method which uses the estimated popularity degree of source page (such as PageRank score) to improve the propagation process. We compare six different models derived from the popularity-based propagation method with their corresponding models on two standard TREC web test collections. Our results show that all the six models can outperform the baseline models without introducing additional online complexity.

The rest of the paper is organized as follows. Section 2 reviews some of the connectivity-based algorithms and relevance propagation methods. In section 3 we investigate our new method theoretically and in section 4 we evaluate its derived models experimentally. At last, in section 5 we summarize our main contributions and discuss some possible further improvements on our proposed method.

2 Related Work

In section 2.1, we first provide an outline of state-of-the-art link analysis algorithms such as PageRank, DistanceRank and HostRank, which can be used as the popularity measure in our work. PageRank is reviewed deeper than other algorithms, because it is the first and simplest choice in the experiments. Then, in section 2.2, we conduct a review of the latest representative relevance propagation algorithms have already been introduced.

2.1 Popularity Measure

In 1998, Page et al. [17] proposed a new link analysis algorithm, called PageRank, which was the first algorithm employed by Google search engine [8]. The key idea behind PageRank is that a web page is important if several other important pages point to it. Intuitively, if a page is itself very important, then its author's opinion on the importance of other pages is more reliable; and if a page links to a lot of pages, the importance score it confers to each of them are decreased. Therefore the PageRank score $PR(i)$ of page i is a weighted function of the PageRank score of its parents (the pages point to the page i):

$$PR(i) = d\left(\sum_{j \in B(i)} \frac{PR(j)}{O(j)}\right) + \frac{1-d}{n} \tag{1}$$

where $d < 1$ is the damping factor, n is the total number of pages and $B(i)$ and $O(j)$ are the set of pages linking to i and the out-degree of the page j, respectively. The division by $O(j)$ captures the intuition that pages which point to page i evenly distribute their rank boost to all of the pages they point to. The presence of the damping factor is necessary, because the web graph is not a strongly connected graph (SCG). By considering the damping factor there will be a virtual link between the page i and all other pages. Hence the sink pages (pages with no out-link) problem is resolved and the convergence of algorithm is guaranteed.

The definition of PageRank lends to an interpretation based on random walks called Random Surfer Model. In this model a user starts form a random page on the web and at each step she randomly chooses an out-link until gets bored and makes a jump to a random page (Enters the URL manually). Then the probability of the user visiting each page is equivalent to the PageRank of that page computed using the above formula.

In the language of linear algebra, simple PageRank (the equation (1) with $d = 1$) can be written as $r = A^T * r$, where r is the n-dimensional rank vector $[r(1), r(2), \ldots, r(n)]$ and the element a_{ij} of the matrix A is defined in this way:

$$a_{ij} = \begin{cases} \dfrac{1}{O(i)} & \text{if } i \text{ points to } j \\ 0 & \text{O.W.} \end{cases} \tag{2}$$

With this definition of simple PageRank, its computation is equivalent to computing the principal eigenvector of the matrix A^T corresponding to the eigenvalue 1. One of the simplest methods of computing the principal eigenvector of a matrix is called power iteration. In the power iteration, an arbitrary initial vector is repeatedly multiplied with the given matrix until it converges to the principal eigenvector [7]. Therefore PageRank (equation (1)) can also be computed iteratively. In a strict mathematical sense, iterations should run to convergence. Most of the time, however, we are more interested in relative ordering of the pages than the actual PageRank values. Thus it is more common to terminate the power iteration once the ordering of the pages becomes reasonably stable. Let $|V|$ and $|E|$ denote the number of vertices and edges in the web graph, respectively. Then theoretically the algorithmic complexity of PageRank is $O(|V| * |E|)$ and practically is $O(100 * |E|)$ i.e. 100 iterations for an acceptable ranking is sufficient.

In [9], Haveliwala developed a personalized and topic-dependent scheme of PageRank called Topic-Sensitive PageRank (TSPR) to improve personalized web search. TSPR computes several PageRank vectors biased using a set of representative topics which capture the page importance with respect to a particular topic. In other words, a different PageRank value is computed for each page and for each topic.

Zareh et al. [28] proposed a connectivity-based ranking algorithm, called DistanceRank, based on reinforcement learning [25] in which the distance between pages are considered as punishment. The distance between two pages i and j is defined as the logarithm of i 's out-degree when i points to j (i.e. the number of ‘‘average clicks’’ between two pages [13]). The main objective of algorithm is to minimize the sum of received punishments (distance) by the user agent so that the pages with the low distance will have a higher rank. This algorithm tries to model a real user surfing the web. When a user randomly browses the web, she selects the next pages based on her background from the last pages and the current status of the web page.

Unlike the aforementioned algorithms which consider a flat web graph, HostRank [27] and BlockRank [12] are two other link analysis algorithms which try to leverage the hierarchical structure on the web graph and employ modified stochastic transition matrices to compute the PageRank scores. HostRank algorithm starts from computing the importance of a host and then the hierarchy structure of the host is used to distribute the host’s importance to web pages within the host. BlockRank use the link structure hierarchy, e.g. domain, hosts, and pages, to build the personalization vectors.

The key point about all the above algorithms is that they are query-independent and their major computations are offline (not at query time) which has a direct impact on the algorithmic complexity of our algorithm.

2.2 Relevance Propagation Methods

Many relevance propagation methods were proposed to propagate content information through the link structure to increase the number of document descriptors. For example [3] propagates anchor text from one page to another to expand the feature set of web pages. Shakery et al. [23] consider how to use web structure to further improve relevance weighting. They propagate the relevance score of a page to another page through hyperlink between them. They defined the hyper relevance score of each page as a function of three variables: its content similarity to the query (self-relevance), a weighted sum of the hyper relevance scores of all the pages that point to it (in-link pages), and a weighted sum of the hyper relevance scores of all the pages it points to (out-link pages). According to these definitions, their relevance propagation model can be written as:

$$h^{k+1}(p) = \alpha S(p) + \beta \sum_{p_i \to p} h^k(p_i)\omega_I(p_i, p) + \gamma \sum_{p \to p_j} h^k(p_j)\omega_O(p, p_j) \quad (3)$$

$$\text{where } \alpha + \beta + \gamma = 1,\ h^0(p) = S(p),\ \omega_I(p_i, p) \propto S(p) \text{ and } \omega_O(p, p_j) \propto S(p_j)$$

$h^k(p)$ is the hyper relevance score of page p after the k -th iteration, $S(p)$ is the content similarity between page p and the query and ω_I and ω_O are weighting functions for in-link and out-link pages, respectively. For implementation, they have given three special cases of this model: weighted in-link (WI), weighted out-link (WO), and uniform out-link (UO) (Table 1). Their experimental results show that relevance propagation generally performs better than using only content information. However, the amount of improvement is sensitive to the document collection and the tuning of parameters.

Song et al. [24] proposed another propagation algorithm which propagates query term frequency from child pages to parent pages in the sitemap tree. Firstly the sitemap of each website is constructed based on URL analysis, and then the query term frequency is propagated along the parent-child relationship in the sitemap tree as follow:

$$f_t^{'}(p) = (1+\alpha) f_t(p) + \frac{(1-\alpha)}{|Child(p)|} \sum_{q \in Child(p)} f_t(q) \tag{4}$$

where $f_t(p)$ and $f_t^{'}(p)$ are the occurrence frequencies of term t in page p before and after propagation, q is the child page of p and α is a weight which controls the contribution of the child pages to their parent. After propagation of term frequency, any content-based algorithm (such as BM25, TF-IDF, etc.) can be used to rank the pages.

Table 1. Special cases of the relevance score propagation model.

Special case	**Model formulation**
WI	$h^{k+1}(p) = \alpha S(p) + (1-\alpha) \sum_{p_i \to p} h^k(p_i) \omega_I(p_i, p)$ (5)
WO	$h^{k+1}(p) = \alpha S(p) + (1-\alpha) \sum_{p \to p_j} h^k(p_j) \omega_O(p, p_j)$ (6)
UO	$h^{k+1}(p) = S(p) + (1-\alpha) \sum_{p \to p_j} h^k(p_j)$ (7)

At first glance, the two latter aforementioned algorithms seem very different. However, Qin et al. [18] proposed a generic relevance propagation framework, which brings together techniques from [23] and [24], as well as different propagation methods. We call this framework GRPF. In this framework, the relevance score propagation model proposed by Shakery et al. [23] and the iterative version of sitemap-based term propagation model proposed by Song et al. [24] have renamed hyperlink-based score propagation model (HS model) and sitemap-based term propagation model (ST model) respectively. In addition to this two models, two new models can be derived from this generic framework: hyperlink-based term propagation model (HT model) and sitemap-based score propagation model (SS model) (Table 2). Similar to the HS model, the HT model also has three special cases: WI, WO and UO. Our work is actually a contribution to the models of this framework, especially the hyperlink-based models (HS and HT).

Table 2. New models extracted from the generic relevance propagation framework: HT and SS.

Model	**Model formulation**
HT	$f_t^{k+1}(p) = \alpha f_t^0(p) + \beta \sum_{p_i \to p} f_t^k(p_i) \omega_I(p_i, p) + \gamma \sum_{p \to p_j} f_t^k(p_j) \omega_O(p, p_j)$ where $\alpha + \beta + \gamma = 1$, $\omega_I(p_i, p) \propto f_t^0(p)$ and $\omega_O(p, p_j) \propto f_t^0(p_j)$ (8)
SS	$h^{k+1}(p) = \alpha S(p) + \frac{(1-\alpha)}{\lvert Child(p) \rvert} \sum_{q \in Child(p)} h^k(q)$ (9)

3 Popularity-based Relevance Propagation

Our idea for the popularity-based relevance propagation is relatively simple: Intuitively, we can conclude that the amount of content features propagation (score or term), from one page to another in the relevance propagation methods, is extremely influenced by the popularity of the source page. In other words, the more popular the source page, the more influence it has in the propagation process. For example, if the source page is a page from Wikipedia web site [26], it should be considered more important than a regular page (Figure 1).

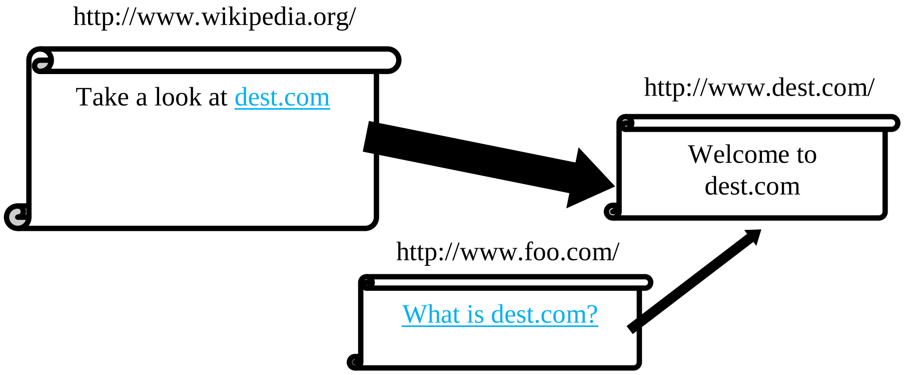

Figure 1. An example of the popularity-based relevance propagation method.

As mentioned in the section 2.1, the popularity measure can be any connectivity-based score which is computed independent of the query and for the whole web graph. The simplest choice is the PageRank score of a page, but other scores such as its DistanceRank score or HostRank score are also acceptable and may improve precision of the proposed method.

Table 3. Popularity-based models, their abbreviations and their corresponding models in GRPF.

Model		**Abbreviation**	**Corresponding GPRF models**
Popularity-based score propagation using hyperlink	Weighted in-link	PSH-WI	HS-WI
	Weighted out-link	PSH-WO	HS-WO
	Uniform out-link	PSH-UO	HS-UO
Popularity-based term propagation using hyperlink	Weighted in-link	PTH-WI	HT-WI
	Weighted out-link	PTH-WO	HT-WO
	Uniform out-link	PTH-UO	HT-UO
Popularity-based score propagation using sitemap		PSS	SS
Popularity-based term propagation using sitemap		PTS	ST

The idea of using popularity measure can easily be applied to the eight models of GRPF: HS-WI, HS-WO, HS-UO, HT-WI, HT-WO, HT-UO, SS and ST. For ease of reference, the abbreviations of our proposed models and their corresponding GRPF models are shown in Table 3.

Let $P(p)$ denote the popularity degree of page p. The main question is how to affect amount of propagation by the source page based on its popularity? Since popularity of a page has similar role to the propagation weight in relevance propagation models, we can multiply it to the current hyper score of the propagator page. In this regard, model formulation of different popularity-based relevance propagation models is shown in Table 4.

Table 4. Model formulation of popularity-based relevance propagation models.

Model	Model formulation
PSH-WI	$h^{k+1}(p)=\alpha S(p)+(1-\alpha)\sum_{p_i \to p} h^k(p_i)\omega_I(p_i, p)P(p_i)$ (10)
PSH-WO	$h^{k+1}(p)=\alpha S(p)+(1-\alpha)\sum_{p \to p_j} h^k(p_j)\omega_O(p, p_j)P(p_j)$ (11)
PSH-UO	$h^{k+1}(p)=S(p)+(1-\alpha)\sum_{p \to p_j} h^k(p_j)P(p_j)$ (12)
PTH-WI	$f_t^{k+1}(p)=\alpha f_t^0(p)+(1-\alpha)\sum_{p_i \to p} f_t^k(p_i)\omega_I(p_i, p)P(p_i)$ (13)
PTH-WO	$f^{k+1}(p)=\alpha f_t^0(p)+(1-\alpha)\sum_{p \to p_j} f_t^k(p_j)\omega_O(p, p_j)P(p_j)$ (14)
PTH-UO	$f_t^{k+1}(p)= f_t^0(p)+(1-\alpha)\sum_{p \to p_j} f_t^k(p_j)P(p_j)$ (15)
PSS	$h^{k+1}(p)=\alpha S(p)+\frac{(1-\alpha)}{\lvert Child(p)\rvert}\sum_{q \in Child(p)} h^k(q)P(q)$ (16)
PTS	$f_t^{k+1}(p)=\alpha f_t^0(p)+\frac{(1-\alpha)}{\lvert Child(p)\rvert}\sum_{q \in Child(p)} f_t^k(q)P(q)$ (17)

4. EMPIRICAL EVALUATIONS

In this section we are going to evaluate the performance and efficiency of the proposed models (except sitemap base models duo to no URL accessibility in order to construct the sitemap) against the corresponding models of GPRF. Firstly, we investigate experimental settings, some implementation issues and the evaluation measures and then the results of the effectiveness evaluation are shown. Finally, we conduct a review of the models efficiency.

4.1 Experimental Settings

For the purpose of "Effectiveness Evaluation", we used the ".GOV" corpus of the LETOR 3.0 [19]. LETOR is a benchmark collection for the research on learning to rank for IR, released by Microsoft Research Asia (MSRA) [15]. LETOR3.0 contains standard features, relevance judgments, data partitioning, evaluation tools, and several baselines, for the OHSUMED and the .GOV data collection. Version 3.0 was released in December, 2008. The .GOV corpus, which is crawled from the

.gov domain in January, 2002, has been used as the data collection of Web Track since TREC 2002. There are totally 1,053,110 pages with 11,164,829 hyperlinks in it. As our query set, we used the topic distillation task in Web Track 2003 and 2004 (with 50 and 75 queries, respectively). Topic distillation aims to find a list of entry points of good websites principally devoted to the topic. The focus is to return entry pages of good websites rather than the web pages containing relevant information, because entry pages provide a better overview of the websites.

Both in construction of the working set (section 4.2) and ranking the documents after propagation in hyperlink based term propagation methods (PTH and HT), we used BM25 as the relevance weighting function:

$$S(Q,D)=\sum_{t\in Q}\left(\log\frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)}\right)\frac{(k_1+1)tf}{K+tf}\frac{(k_3+1)qtf}{k_3+qtf}+k_2|Q|\frac{avdl-dl}{avdl+dl} \qquad (18)$$

The parameters of the equation are shown in Table 5. In our experiments, we set $k_1=2.5$, $k_2=k_3=0$ and $b=0.8$ [19].

Table 5. Parameters of BM25 model.

Variable	Definition
r	# of relevant docs containing term t for query Q
R	# of relevant docs for query Q
n	# of docs containing term t
N	# of all docs
tf	Frequency of term t in doc D
qtf	Frequency of term t in query Q
$avdl$	Average doc length
dl	Doc D length (# of its terms)
$\|Q\|$ or nq	# of query Q terms
b,k_1,k_2,k_3 and K $K=k_1((1-b)+b(dl/dl_{avg}))$	Tuning parameters
$S(Q,D)$	Similarity between query Q and doc D

We chose PageRank as the popularity measure in the popularity-based models (PSH and PTH). However, the original PageRank score of a page is usually very small and thus it will ruin the propagation process. Therefore instead of using the original PageRank score ($PR(p)$), we used a function of it as the popularity measure in our experiments (Eq. (19)). In this equation γ is a tuning parameter, which according to our experiments can be set to 1.4.

$$P(p)=\frac{-\gamma}{\log(PR(p))} \qquad (19)$$

4.1 Constructing the Working Set

Following Shakery et al. [23], instead of running our experiments on the whole set of data, for each query, we first construct a working set. To construct the working set, we first find the top 400 pages with the highest BM25 score as the core set. Then we expand the core set to the working set by adding the pages point to the pages in core set (parent pages) and the pages are pointed by the pages in the core set (child pages).

4.1 Evaluation Measures

For the purpose of evaluation, we use a number of evaluation measures commonly used in information retrieval, namely Precision at n (P@n) [2], Mean Average Precision (MAP) [2], and Normalized Discount Cumulative Gain (NDCG) [11].

4.3.1 Precision at n (P@n)

As it has been quoted in reference [2], precision at n measures the relevance of the top n documents in the ranking list with respect to a given query:

$$P@n = \frac{\text{\# of relevant docs in top } n \text{ results}}{n} \tag{20}$$

4.3.2 Mean Average Precision (MAP)

The average precision (AP) [2] of a given query is calculated as Eq. (21), and corresponds to the average of $P@n$ values for all relevant documents:

$$AP = \frac{\sum_{n=1}^{N}\left(P@n * rel(n)\right)}{\text{\# of relevant docs for this query}} \tag{21}$$

where N is the number of retrieved documents, and $rel(n)$ is a binary function that evaluates to 1 if the n-th document is relevant, and 0 otherwise. Finally, MAP is obtained by averaging the AP values over the set of queries.

4.3.3 Normalized Discount Cumulative Gain (NDCG)

For a single query, the NDCG value of its ranking list at position n is computed by Eq. (22):

$$NDCG@n = Z_n \sum_{j=1}^{n} \frac{2^{r(j)} - 1}{\log(1 + j)} \tag{22}$$

where $r(j)$ is the rating of the j-th document in the ranking list, and the normalization constant Z_n is chosen so that the perfect list gets NDCG score of 1. For the TREC 2003 and TREC 2004 datasets, there are two ratings {0, 1} corresponding to "relevant" and "not relevant" in order to compute NDCG scores.

4.4 Effectiveness Evaluation

In this section, we present an experimental evaluation of the proposed models, their corresponding models and two other algorithms: BM25 as the content-only algorithm and PageRank as the connectivity-only algorithm (our popularity measure). In the following figures, there is a separate figure for each category of propagation methods: hyperlink-based score propagation methods (PSH and SH) and hyperlink-based term propagation methods (PTH and HT).

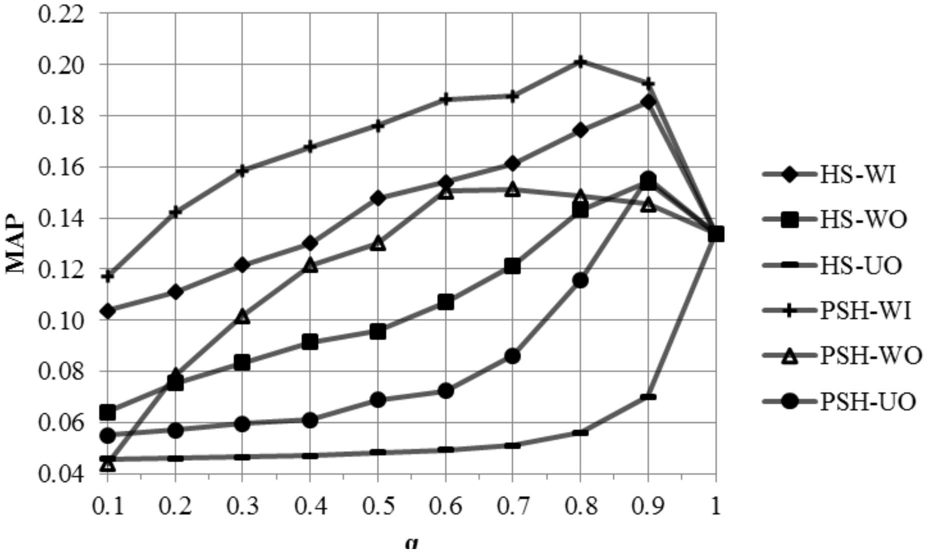

Figure 2. Evaluation of hyperlink-based score propagation models at different α s in TREC 2003.

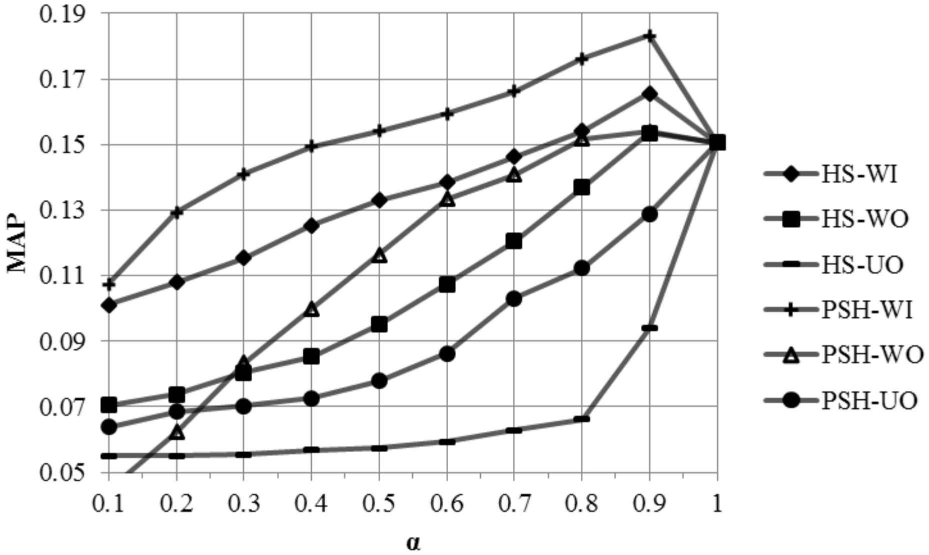

Figure 3. Evaluation of hyperlink-based score propagation models at different α s in TREC 2004.

Figures 2 and 3 show the performance of the all hyperlink-based score propagation models at different α s on TREC 2003 and TREC 2004 datasets, respectively. As can be seen, all popularity-based models (PSH models) boosted the retrieval performance against their corresponding models on both datasets, except for PSH-WO which its boosting interval is limited to $0.2 < \alpha < 0.8$ and $0.3 < \alpha < 0.9$ on TREC 2003 and TREC 2004 respectively.

PSH-WI has the best performance with MAP = 0.2039 (for $\alpha = 0.85$, which of course cannot be seen because the horizontal axis is plotted at 0.1 intervals) and MAP = 0.1832 (for $\alpha = 0.9$) on TREC 2003 and TREC 2004, respectively.

Similar to hyperlink-based score propagation models, popularity-based term propagation models also improve their corresponding models of GPRF, almost for all values of α . However, PTH-WI has higher performance when $\alpha < 0.6$ on both datasets (Figures 4 and 5). The boosting retrieval of PTH-UO is also limited to $\alpha < 0.85$ on TREC 2003. PTH-WI has the best MAP among all other models. Its best MAP is 0.2015 and 0.l837 (corresponding to $\alpha = 0.1$) on TREC 2003 and TREC 2004, respectively.

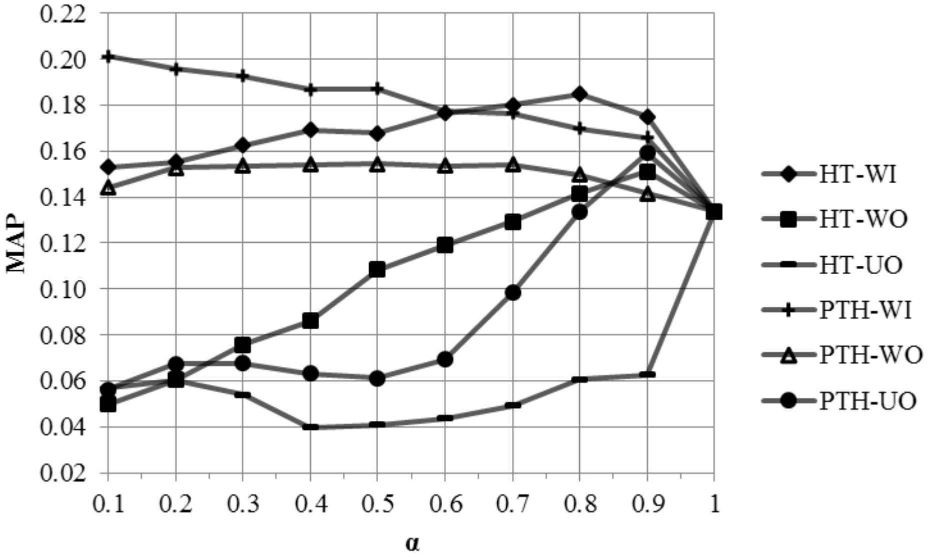

Figure 4. Evaluation of hyperlink-based term propagation models at different α s in TREC 2003.

Table 6 depicts the best P@10, MAP and NDCG@10 of each algorithm over different α s on both datasets (BM25 and PageRank are exceptions). From this table, we can found that the popularity-based weighted in-link methods (PSH-WI and PTH-WI) can generally outperform their corresponding models (HS-WI and HT-WI) by 10% and 5% respectively. However the performance of our models is usually higher in TREC 2003 compared to TREC 2004. We can draw similar conclusions about the popularity-based uniform out-link models. The popularity-based weighted out-link models (PSH-WO and PTH-WO), by contrast, would not have impressive improvement against their corresponding models (HS-WO and HT-WO).

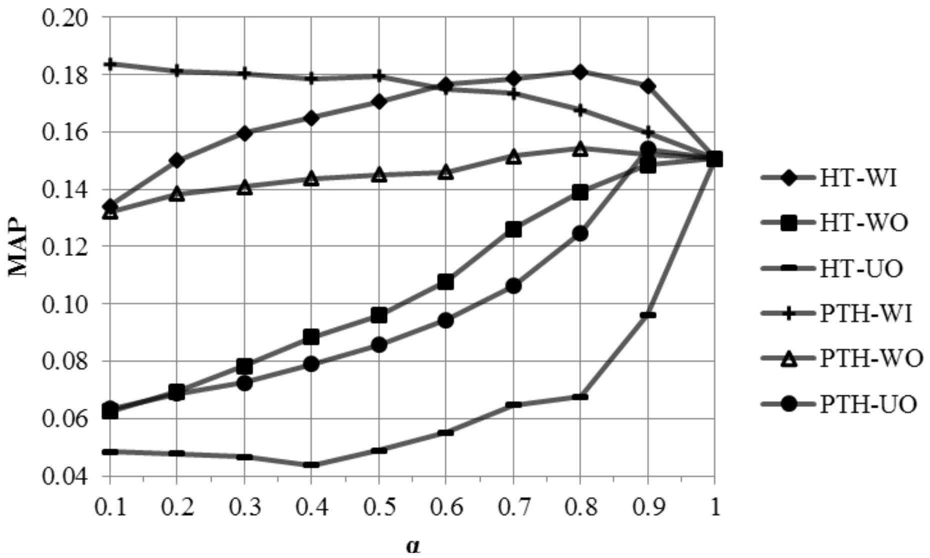

Figure 5. Evaluation of hyperlink-based term propagation models at different α s in TREC 2004.

Table 6. The best performance of each algorithm.

Model	The best P@10		The best MAP		The best NDCG@10	
	TREC 2003	TREC 2004	TREC 2003	TREC 2004	TREC 2003	TREC 2004
BM25	0.1020	0.1587	0.1337	0.1502	0.1858	0.1976
PageRank	0.0540	0.1107	0.0658	0.0915	0.0857	0.1296
HS-WI	0.1380	0.1960	0.1854	0.1657	0.2446	0.2415
PSH-WI	0.1340	0.2160	**0.2039**	**0.1832**	0.2667	0.2700
HS-WO	0.1140	0.1707	0.1560	0.1535	0.2127	0.2076
PSH-WO	0.1140	0.1640	0.1512	0.1539	0.2072	0.2032
HS-UO	0.1020	0.1640	0.1338	0.1506	0.1859	0.2026
PSH-UO	0.1020	0.1640	0.1556	0.1506	0.1859	0.2026
HT-WI	0.1300	0.1973	0.1849	0.1812	0.2539	0.2510
PTH-WI	0.1380	0.1987	**0.2015**	**0.1837**	0.2809	0.2555
HT-WO	0.1220	0.1640	0.1520	0.1508	0.2171	0.2026
PTH-WO	0.1180	0.1640	0.1546	0.1543	0.2146	0.2042
HT-UO	0.1020	0.1640	0.1338	0.1508	0.1859	0.2026
PTH-UO	0.1160	0.1640	0.1597	0.1540	0.2165	0.2060

The question which should now be answered is whether the performance of the popularity-based weighted out-link methods can be improved or not? Results of an experiment show if we use the popularity of the propagator page (child page) instead of the popularity of the propagation destination (current page) in out-link popularity-based methods (PSH-WO, PSH-UO, PTH-WO and PTH-UO), we would get better results. For example, Table 7 includes the best MAP of these methods after applying the above change. This result is intuitively understandable. Since the importance of the in-links is more than out-links and the propagation direction is from out pages to the current page, so using the popularity of the propagator page does not enhance the precision. Therefore, in these models, the more popular the destination page, the more score it should achieve.

Table 7 shows that PSH-WO, PSH-UO, PTH-WO and PTH-UO have the potential to enhance the precision of their corresponding methods by about 16%, 10%, 7% and 3% on average respectively.

Table 7. Best MAP after considering the popularity of propagation destination page.

Dataset	**PSH-WO**	**PSH-UO**	**PTH-WO**	**PTH-UO**
TREC 2003	0.1883	0.1638	0.1705	0.1658
TREC 2004	0.1724	0.1506	0.1556	0.1597

4.5. Efficiency Evaluation

In the previous section, the effectiveness of the proposed models and their corresponding models were discussed. In this section, we are going to explore the algorithmic complexity of the models which is another important factor when they are subject of being employed in real world application like search engines. Because of the close similarity between our work and GRPF, we refer to the same description given in [18]. In GRPF [18], the efficiency of different relevance propagation models, has been divided into two cases: online complexity and offline complexity. The only overhead introduced by the popularity-based relevance propagation models is the popularity measure computation which is an offline overhead. For example if PageRank is used as the popularity measure, the offline complexity equals to $O(100 * |E|)$ where $|E|$ denotes the number of edges in the web graph. Let w, c, q, l and t indicate the size of the working set, the time complexity of propagating an entity from a page to another page along hyperlinks, the average number of query terms in a page in the working set, the average number of in-links and out-links per page and the number of iteration for propagation convergence, respectively. So in addition to the popularity measure computation, the offline complexity of PSH and PTH models is $twlc$ and $twlcq$, respectively.

In contrast to PSH models which cannot do their computations offline (because the scores doesn't exist until the user query is processed), the offline implementation of PTH models is possible and their online complexity is $\overline{q}tnlc$, where $\overline{q}$ is the average number of unique words per page and n is the total number of pages in the document corpora.

5. Conclusion and Future Work

In this paper, a new idea for using of the popularity measure of pages in the propagation process of the relevance propagation methods is proposed. We argue that this idea is applicable to most of the current propagation methods, especially the models proposed in [18]. We used the popularity degree of the propagation source as a weight in the propagation process. Our experiments showed that this technique can greatly boost the precision of the weighted in-link models (PSH and PTH). However, in the case of popularity-based out-link propagation models (PSH-WO, PSH-UO, PTH-WO and PTH-UO), exploiting the popularity measure of the propagation destination page (current page), will provide more improvement as compared to the popularity of the propagator page (child page).

We used PageRank as the popularity measure in our experiments. Future work will explore utilizing other choices such as DistanceRank, HostRank and etc. We also plan to test the performance of the proposed methods under using the TSPR score of a page which most probably will show better results than the original PageRank for a particular topic.

In this work a simple function of PageRank is used in the experiments. It is important to explore a more appropriate function if we are going to adopt PageRank as the popularity measure.

Acknowledgements

This work was supported by the Iran Telecommunication Research Center (ITRC). The authors would like to acknowledge the many valuable suggestions made by Mohammad Amin Golshani, M. Sc. student at Yazd University.

References

1. Alexa the Web Information Company, http://www.alexa.com/, 2011.
2. Baeza-Yates, R. & Ribeiro-Neto, B. Modern Information Retrieval. ACM Press/Addison Wesley, 1999.
3. Brin, S. & Page, L. The Anatomy of a Large Scale Hypertextual Web Search Engine. In Proceedings of the 7th World Wide Web Conference, 1998.
4. Cho, J. & Roy, S. Impact of search engines on page popularity. In proceeding of the International World-Wide Web Conference, 2004.
5. Craswell, N. & Hawking, D. Overview of the TREC 2003 Web Track. In the 12th TREC, 2003.
6. Craswell, N. & Hawking, D. Overview of the TREC 2004 Web Track. In the 13th TREC, 2004.
7. Golub, G. & Van Loan, C. Matrix Computations. John Hopkins Press, 1989.
8. Google Search Engine, http://www.google.com/, 2011.
9. Haveliwala, T. Topic-Sensitive PageRank. In Proceedings of the 11th International World-Wide Web Conference, 2002.
10. Hawking, D. Overview of the TREC-9 Web Track. In the 9th TREC, 2002.
11. Jarvelin, K. & Kekalainen, J. Comulated Gainbased Evaluation of IR Techniques. ACM Transactions on Information Systems, Vol. 20 No. 04, pp. 422–446, 2002.
12. Kamvar, S. D., Haveliwala, T. H., Manning, C. D. & Golub, G. H. Exploiting the block structure of the web for computing. Technical report, Stanford University, Stanford, CA, 2003.
13. Matsuo, Y., Ohsawa, Y. & Ishizuka, M. Average-clicks: A new measure of distance on the World Wide Web. Journal of Intelligent Information Systems , pp. 51–62, 2003.
14. Mcbryan, O. GENVL and WWW: Tools for tamping the web. In Proceedings of the 1st WWW, 1994.
15. MicroSoft Research Asia, http://research.microsoft.com/en-us/labs/asia/default.aspx, 2011.
16. Najork, M., Zaragoza, H. & Taylor, M. J. Hits on the web: how does it compare? In Proceedings of SIGIR'07, pp. 471-478, 2007.
17. Page, L., Brin, S., Motawni, R. & Winogard, T. The PageRank citation algorithm: Bringing order to the web. Technical report, Standford Digital Library Technologies Project, 1998.
18. Qin, T., Liu, T. Y., Zhang, X. D., Chen, Z., & Ma, W. Y. A study of relevance propagation for web search. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 408–415, 2005.

19. Qin, T., Liu, T., Xu, J. & Li, H. Letor: A benchmark collection for research on learning to rank for information retrieval. Information Retrieval Journal, pp. 346-374, 2010.
20. Robertson, S. E, Walker, S., Jones, S, M.Hancock-Beaulieu, M. & Gatford, M. Okapi at TREC-3. In Harman, D. K., editor, The Third Text REtrieval Conference (TREC-3), pp. 109-126, 1995.
21. Salton, G. & Buckley, C. Term-weighting approaches in automatic text retrieval. Information Processing and Management: an International Journal, Vol. 24 No. 5, pp. 513-523, 1988.
22. Shakery, A. & Zhai, C. X. A probabilistic relevance propagation model for hypertext retrieval. In Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM), pp. 550-558, 2006.
23. Shakery, A. & Zhai, C. X. Relevance Propagation for Topic Distillation UIUC TREC 2003 Web Track Experiments. In Proceedings of the TREC Conference, 2003.
24. Song, R., Wen, J. R., Shi, S. M., Xin, G. M., Liu, T. Y., Qin,T., Zheng, X., Zhang, J. Y., Xue, G. R. & Ma, W. Y. Microsoft Research Asia at Web Track and Terabyte Track of TREC 2004. 13th TREC, 2005.
25. Sutton, R. S. & Barto, A. G. Reinforcement learning: An introduction. Cambridge, MIT Press, 1998.
26. Wikipedia, http://www.wikipedia.org/, 2011.
27. Xue, G. R., Yang, Q., Zeng, H. J., Yu, Y., & Chen, Z. Exploiting the hierarchical structure for link analysis. In SIGIR, August 2005.
28. Zareh Bidoki, A. M. & Yazdani, N. DistanceRank: An Intelligent Ranking Algorithm for Web Pages. Information Processing & Managament, Vol. 44, No. 2, pp. 877-892, March 2008.
29. Zareh Bidoki, A. M. Effective Web Ranking & Crawling. Ph.D. Thesis, University of Tehran, May 2009.