IJICTR International Journal of Information & Communication Technology Research

ITRC

Volume 6- Number 1- winter 2014 (43-54)

Web Page Streams and Relevance Propagation for Topic Distillation

Mohammad Amin Golshani
Department of Electrical and Computer Engineering
Yazd University
Yazd, Iran
Golshani.ma@yahoo.com

AliMohammad ZarehBidoki
Department of Electrical and Computer Engineering
Yazd University
Yazd, Iran
AliZareh@Yazduni.ac.ir

Received: February 24, 2012- Accepted: September 18, 2013

Abstract— **Over the past decade, several studies in field of relevance propagation models have been proposed to improve quality of web search, which include hyperlink-based score propagation, hyperlink-based term propagation and popularity-based relevance propagation models; however, all of them have used low precision content similarity functions in the propagation process and their throughputs are not entirely satisfactory. In this paper, two stream-based content similarity functions that could be used to derive new relevance propagation models were introduced. In the proposed content similarity functions, the web page was split to different streams with different degrees of importance and the text of each web page was divided between these streams. To evaluate the proposed relevance propagation models, Letor 3.0 (including two standard web test collections) was used in the experiments. It was concluded that splitting web pages as different streams could provide significant improvement in relevance propagation models.**

Keywords- Web page streams; relevance propagation; topic distillation; information retrieval; search engine; web page ranking

I. INTRODUCTION

Information retrieval is a computer science subfield, the goal of which is to find all documents relevant to a user querying a given collection of documents. When a user sends a query to a search engine, the search engine returns URLs of the documents matching all or one of the terms, depending on both the query operator and the algorithm used by the search engine. Ranking is the process of ordering the returned documents in decreasing order of relevance; i.e. the "best" answers are on the top. To make the web more interesting and productive, an effective and efficient ranking algorithm is needed to present more appropriate results to users. In general, the three major components of a search engine are: crawler, indexer and searcher [1, 2]. Many studies have been done about challenges in web search engines, such as web page ranking, web crawling, freshness, spam detection, etc. [3-15]. Ranking process as the primary part of the searcher module has always been a challenging issue of every search engine. Different from traditional information retrieval, the web contains both content and link structures that have provided many new dimensions for exploring better IR techniques.

In this paper, a comprehensive study of relevance propagation technologies (including 24 propagation algorithms) was done for web information retrieval. The existing works [16-20] were extended by introducing two stream-based content similarity functions to provide 4 new relevance propagation models (including 12 new relevance propagation methods). Also, both theoretical and experimental

evaluations were conducted over these models to compare them against the old ones. To evaluate the proposed relevance propagation models, LETOR 3.0 benchmark collection was used in the experiments. Based on the experimental results, it was found that splitting web pages to different streams can improve accuracy in relevance propagation technologies.

Organization of this paper is as follows. In Section II, the existing relevance propagation models are reviewed. In Section III, two content similarity functions that can be used to derive new propagation models are introduced. Then, effectiveness and efficiency of the relevance propagation models are examined in Sections IV and V, respectively. Conclusions and future works are given in Section VI.

II. RELATED WORK

Currently, there are three major categories of ranking algorithms based on content and connectivity as follows [21]:

- **Content-based**. In traditional IR, the evidence of relevance is thought to reside within the text content of documents. Consequently, the system tries to find documents corresponding to the user query. The fundamental strategy of traditional IR is to rank documents according to their estimated degree of relevance based on measures such as term similarity or term occurrence probability. In order words, for each query, the documents with more similar content to the query will be selected as more relevant ones. Examples of the content-based ranking algorithms are TF-IDF [22] and BM25 [23].

- **Connectivity-based**. In the web setting, information can reside outside textual content of documents. For example, links between pages can be used to increase the term-based estimation of document relevance. Furthermore, hyperlinks, being the most important source of evidence in web documents, have been the subject of many researches which explore retrieval strategies based on link analysis. Connectivity-based algorithms use the links between web pages and assign numerical weighting to each element of a hyperlinked set of documents in order to measure its relative importance within the set. Instances of the connectivity-based ranking algorithms are PageRank [24] and DistanceRank [25].

- **Combinational**. Using either content-based or connectivity-based algorithms independently leads to a low-precision ranking function which cannot fully satisfy users' demands in the web [26]. Therefore, combination algorithms which use both content and link structures have been introduced. In fact, they combine content and connectivity information together. These methods can be divided into two groups: one is to enhance link analysis with the assistance of content information, such as HITS and topic-sensitive PageRank [27-35] and another is relevance propagation, which propagates content information with the assistance of the web structure [2, 16, 36 and 37]. In recent years, relevance propagation methods as one of the salient combinational algorithms has attracted IR researchers' attention. In the relevance propagation models [16, 23, 24, 34 and 37], content-based score or query terms are propagated through hyperlinks from one page to another.

Okapi BM25 is a ranking function used by search engines to rank matching documents according to their relevance to a given search query. It is based on the probabilistic retrieval framework developed in the 1970s and 1980s by Stephen E. Robertson, Karen Spärck Jones and others [38]. BM25 is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document. It is not a single function, but actually is a whole family of scoring functions with slightly different components and parameters. One of the most prominent instantiations of the function is as follows.

Given a query Q, containing keywords $q_1,...,q_n$, the BM25 score of a document D is:

$$score(D,Q)=\sum_{i=1}^{n}\frac{N-n(q_i)+0.5}{n(q_i)+0.5}\times\frac{f(q_i,D).(k_1+1)}{f(q_i,D)+k_1.(1-b+b.\frac{|D|}{avgdl})} \quad (1)$$

TABLE I. BM25 PARAMETERS.

Variable	Definition
$f(q_i,D)$	q_i's term frequency in the document D
$\|D\|$	Length of the document D in words
$avgdl$	Average document length in the text collection from which documents are drawn
k_1, b	Tuning parameters, $k_1 \in [1.2, 2.0]$, $b = 0.75$
N	The total number of documents in the collection
$n(q_i)$	# of documents containing q_i
$Score(D,Q)$	Similarity between query Q and doc D

Many relevance propagation methods have been proposed to propagate content information through link structure in order to increase the number of document descriptors. These methods were grouped as shown in Table II. Score-level and term-level models propagate content similarity between web pages and submitted queries as BM25 score and term frequency through the link structure, respectively. As an example from Table II, the PSH model as a score-level method propagates BM25 score and popularity measure (PageRank score) of the web pages in the relevance propagation process.

TABLE II. RELEVANCE PROPAGATION MODELS AND THEIR ABBREVIATIONS.

	Model	Abbreviation
Score-level	Hyperlink-based score propagation model [16]	HS
	Popularity-based Relevance Propagation model [19]	PSH
Term-level	Hyperlink-based term propagation model [18]	HT
	Popularity-based Relevance Propagation model [19]	PTH

TABLE III. SPECIAL CASES OF THE RELEVANCE SCORE PROPAGATION MODEL (HS MODEL).

Special case	Abbreviation	Model formulation	
Weighted In-Link	HS-WI	$h^{k+1}(p)=\alpha S(p)+(1-\alpha)\sum_{p_i\to p} h^k(p_i)\omega_I(p_i,p)$	(2)
Weighted Out-Link	HS-WO	$h^{k+1}(p)=\alpha S(p)+(1-\alpha)\sum_{p\to p_j} h^k(p_j)\omega_O(p,p_j)$	(3)
Uniform Out-link	HS-UO	$h^{k+1}(p)=S(p)+(1-\alpha)\sum_{p\to p_j} h^k(p_j)$	(4)

TABLE IV. SCORE PROPAGATION RESULT OF FIGURE 1 (HS-WI METHOD).

Iteration	P1	P2	P3	P4
0	$h^0(p_1)=\alpha S_1$	$h^0(p_2)=\alpha S_2$	$h^0(p_3)=\alpha S_3$	$h^0(p_4)=\alpha S_4$
1	$h^1(p_1)=\alpha S_1$	$h^1(p_2)=\alpha S_2+(1-\alpha)\alpha S_1\omega_1$	$h^1(p_3)=\alpha S_3+(1-\alpha)\alpha S_1\omega_1$	$h^1(p_4)=\alpha S_4+(1-\alpha)\alpha S_2\omega_1$
2	$h^2(p_1)=\alpha S_1$	$h^2(p_2)=\alpha S_2+(1-\alpha)\alpha S_1\omega_1$	$h^2(p_3)=\alpha S_3+(1-\alpha)\alpha S_1\omega_1$	$h^2(p_4)=\alpha S_4+(1-\alpha)(\alpha S_2+(1-\alpha)\alpha S_1\omega_1)\omega_1$
3	$h^3(p_1)=\alpha S_1$	$h^3(p_2)=\alpha S_2+(1-\alpha)\alpha S_1\omega_1$	$h^3(p_3)=\alpha S_3+(1-\alpha)\alpha S_1\omega_1$	$h^3(p_4)=\alpha S_4+(1-\alpha)(\alpha S_2+(1-\alpha)\alpha S_1\omega_1)\omega_1$

Shakery et al. [16] considered how to use web structure to further improve relevance weighting. They propagated relevance score of a page to another page through a hyperlink between them (web structure) and 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 pointing to it (in-link pages) and a weighted sum of the hyper relevance scores of all the pointing pages (out-link pages). According to these definitions, their relevance propagation model can be written as:

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

$$\text{where } \alpha+\beta+\gamma=1,\ h^0(p)=S(p),\quad \omega_I(p_i,p)\propto S(p) \text{ and } \omega_O(p,p_j)\propto S(p_j) \tag{5}$$

$h^k(p)$ is the hyper relevance score of page p after the k-th iteration, $S(p)$ is content similarity between page p and query (BM25 score) and ω_I and ω_O are weighting functions for in-link and out-link pages, respectively. For implementation, they presented three special cases of this model: weighted in-link (WI), weighted out-link (WO), and uniform out-link (UO) (Table III).

An example is given in Figure 1. This is a website with only 4 pages. Table IV shows the propagation results according to the iterative version of the HS-WI method. From this table, it can be seen that 2 iterations are taken to converge because the scores after the 2nd and 3rd iterations are the same.

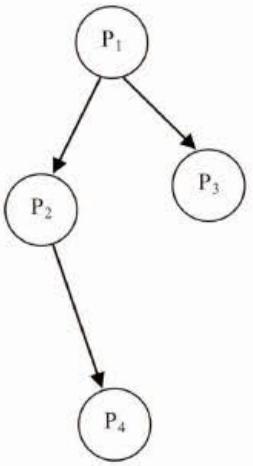

Figure 1. A website with 4 pages.

QIN et al. [18] proposed another relevance propagation model, called HT[1] model, which needed to propagate the frequency of query term (TF[2]) in a web page before adopting relevance weighting algorithms to rank the document. In fact, HT model was an extended version of the HS model and similar to HS, it had three special cases as given in Table V, where $f_t^k(p)$ is term frequency of term t in page p in *k-th* iteration.

[1] Hyperlink-based term propagation model
[2] Term Frequency

TABLE V. SPECIAL CASES OF THE RELEVANCE TERM PROPAGATION MODEL (HT MODEL).

Special case	Abbreviation	Model formulation
Weighted In-Link	HT-WI	$f_t^{k+1}(p)=\alpha f_t^0(p)+(1-\alpha)\sum_{p_i\to p} f_t^k(p_i)\omega_I(p_i,p)$ where , $\omega_I(p_i,p)\propto f_t^0(p)$ (6)
Weighted Out-Link	HT-WO	$f_t^{k+1}(p)=\alpha f_t^0(p)+(1-\alpha)\sum_{p\to p_j} f_{tc}^k(p_j)\omega_O(p,p_j)$ where $\omega_O(p,p_j)\propto f_t^0(p_j)$ (7)
Uniform Out-link	HT-UO	$f_t^0(p)=f_t^0(p)+(1-\alpha)\sum_{p\to p_j} f_t^k(p_j)$ (8)

TABLE VI. POPULARITY-BASED PROPAGATION MODELS AND THEIR ABBREVIATIONS.

Model		Abbreviation	Corresponding method
Popularity-based score propagation using hyperlink (PSH model)	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 (PTH model)	Weighted in-link	PTH-WI	HT-WI
	Weighted out-link	PTH-WO	HT-WO
	Uniform out-link	PTH-UO	HT-UO

TABLE VII. SPECIAL CASES OF THE POPULARITY-BASED RELEVANCE PROPAGATION MODELS.

Method	Method formulation
PSH-WI	$h^{k+1}(p)=\alpha S(p)+(1-\alpha)\sum_{p_i\to p} h^k(p_i)\omega_I(p_i,p)P(p_i)$, $P(p_i)=\frac{-\gamma}{\log(PR(p_i))}$, $\omega_I(p_i,p)\propto S(p)$ (9)
PSH-WO	$h^{k+1}(p)=\alpha S(p)+(1-\alpha)\sum_{p\to p_j} h^k(p_j)\omega_O(p,p_j)P(p_j)$, $P(p_j)=\frac{-\gamma}{\log(PR(p_j))}$, $\omega_O(p,p_j)\propto S(p_j)$ (10)
PSH-UO	$h^{k+1}(p)=S(p)+(1-\alpha)\sum_{p\to p_j} h^k(p_j)P(p_j)$, $P(p_j)=\frac{-\gamma}{\log(PR(p_j))}$ (11)
PTH-WI	$f_t^{k+1}(p)=\alpha f_t^0(p)+(1-\alpha)\sum_{p_i\to p} f_t^k(p_i)\omega_I(p_i,p)P(p_i)$, $P(p_i)=\frac{-\gamma}{\log(PR(p_i))}$, $\omega_I(p_i,p)\propto f_t^0(p)$ (12)
PTH-WO	$f^{k+1}(p)=\alpha f_t^0(p)+(1-\alpha)\sum_{p\to p_j} f_t^k(p_j)\omega_O(p,p_j)P(p_j)$, $P(p_j)=\frac{-\gamma}{\log(PR(p_j))}$, $\omega_O(p,p_j)\propto f_t^0(p_j)$ (13)
PTH-UO	$f_t^{k+1}(p)=f_t^0(p)+(1-\alpha)\sum_{p\to p_j} f_t^k(p_j)P(p_j)$, $P(p_j)=\frac{-\gamma}{\log(PR(p_j))}$ (14)

TABLE VIII. RELEVANCE PROPAGATION MODELS, THEIR STRUCTURES AND ABBREVIATIONS.

Model		Score-level	Term-level	Popularity-measure	Links		Abbreviation
					Inlink	outlink	
Hyperlink-based score propagation [16] (HS model)	WI	√			√		HS-WI
	WO	√				√	HS-WO
	UO	√				√	HS-UO
Hyperlink-based term propagation [18] (HT model)	WI		√		√		HT-WI
	WO		√			√	HT-WO
	UO		√			√	HT-UO
Popularity-based score propagation using hyperlink [19] (PSH model)	WI	√		√	√		PSH-WI
	WO	√		√		√	PSH-WO
	UO	√		√		√	PSH-UO
Popularity-based term propagation using hyperlink [19] (PTH model)	WI		√	√	√		PTH-WI
	WO		√	√		√	PTH-WO
	UO		√	√		√	PTH-UO

Mousakazemi et al. [19] extended HT and HS models and proposed new propagation models (PSH and PTH models, Table VI). In fact, they used popularity measure of the web pages (PageRank score) in the propagation process of the relevance propagation methods (Table VII). PageRank is a popular ranking algorithm used by Google to measure the importance of web pages. PageRank weights each link based on the importance of the document from which it is originated and the number of outlinks in the origin document. It models users' browsing behaviours as a random surfer model [2, 39]. In this model, a person surfs the web by randomly clicking links on the visited pages. When s/he (PageRank) reaches a web page that does not have any outward link, s/he will be randomly jumped to another page. PageRank assumes that a user either follows a link from the current page or jumps to a random page on the web graph. Rank of page j is then computed by the following equation:

$$r(j) = \frac{1-d}{n} + d * \sum_{i \in B(j)} r(i)/o(i) \qquad (15)$$

where n is the number of web pages, $O(i)$ denotes the number of outgoing links from page i and $B(j)$ shows the set of pages that point to page j. Parameter d, damping factor, is used to guarantee the convergence of PageRank and removes effects of sink pages (pages with no outputs).

For simplicity, the reviewed models, their structures and abbreviations are listed in Table VIII.

III. PROPOSED ALGORITHMS

All of the reviewed algorithms have supposed that document is a single body of text, unstructured and undifferentiated. However, it is commonplace in search systems to assume at least some minimal structures for documents. In this section, documents which are structured into a set of fields or streams are considered. That is, there is a global set of labelled streams and text of each document is split between these streams. An obvious example could be a title/abstract/body structure such as the one seen in scientific papers or in the web context, in which a web page can be split to body/anchor/title/URL streams. The general idea is that some streams may be more predictive of relevance than others. For example, a query match on the title might be expected to provide stronger evidence of possible relevance than an equivalent match on the body text. It is now well known in the web context that matching on anchor text is a very strong signal. In the following two subsections, two stream-based content similarity functions which can be applied in the propagation models are introduced.

A. Multiple streams and BM25F

There is a set of $\boldsymbol{S}$ streams and the intention is to assign relative weights v_s to them. For a given document, each stream has its associated length (total length of the document would be normally the sum of the stream lengths). Each term in the document may occur in any of the streams, with any frequency, the total across streams of these term-stream frequencies would be the usual term-document frequency. The entire document becomes a vector of vectors [40] (Table IX).

TABLE IX. WEBPAGE STRUCTURE.

Variable	Definition
Streams	s =1, ..., S
Query terms	Q = $q_1, q_2, \dots, q_n$
Stream lengths	sl_s
Average stream length	$avsl_s$
Stream weights	v_s
Document	(tf_1, ..., $tf_{|V|}$) vector of vectors
tf_i vector	(tf_{1i},..., tf_{Si}) -where tf_{si} is the frequency of term *i* in stream *s*
$w_{Q,D}^{BM25F}$	BM25F score of the query terms for document D

BM25F is an extension of the BM25 ranking function adapted to score structured documents. Here, appropriate version of BM25F was presented [40]:

$$tf_i = \sum_{s=1}^{S} v_s \frac{tf_{si}}{B_s} \tag{16}$$

$$B_s = ((1-b_s) + b_s \frac{sl_s}{avsl_s}), 0 \le b_s \le 1 \tag{17}$$

$$w_i^{BM25F} = \frac{tf_i}{k_1 + tf_i}.w_i^{IDF}, \qquad w_{Q,D}^{BM25F} = \sum_{q \in Q} \sum_{i \in s} w_i^{BM25F} \tag{18}$$

$$w_i^{IDF} = \log \frac{N - n_i + 0.5}{n_i + 0.5} \tag{19}$$

Table X shows BM25F parameters. To get the best results, a trial-and-error procedure was used to tune the BM25F parameters.

TABLE X. MB25F PARAMETERS.

Parameter	value
k_1	32.0
b_{Title}	0.95
b_{Body}	0.9
$b_{AnchorText}$	0.1
v_{Title}	18.0
v_{Body}	1.0
$v_{AnchorText}$	46.0

B. Stream-based term frequency (STF)

Term frequency (TF) is the earliest and most important retrieval signal in retrieval models [41-46]. Use of TF can be dated back to Luhn's pioneer work on automatic indexing [47]. Here, a content similarity function, called *STF*, is proposed. STF is a modification of TF in which web page is considered to be composed of several streams (*title/body/anchor text and URL*) with different degrees of importance and text of the page is split between these streams. Similar to BM25F, each term in the document may occur in any of the streams, with any frequency. Eq. 20 shows the main formula of the STF method.

$$STF_{Q,D} = \sum_{q_t \in Q} \sum_{i \in s} v_{s_i} \times TF(q_t, s_i) \tag{20}$$

where $STF_{Q,D}$ shows STF score of query terms for document D, v_{s_i} is weight of stream *i* and $TF(q_t, s_i)$ shows the number of times that term t occurs in stream *i*. Table XI shows stream weights. The term frequency of each term in the streams is based on the weights assigned to these streams; for example, according to Table XI if a term occurs twice in each of the body and URL fields, its frequency for the body and URL fields will be equal to 1 and 10 respectively. A trial-and-error procedure was also used for tuning STF parameters.

TABLE XI. STF PARAMETERS.

Parameter	value
v_{Title}	2.0
v_{Body}	0.5
$v_{AnchorText}$	2.0
v_{URL}	5.0

Then, the old content similarity functions (BM25 and TF) in the reviewed propagation models were replaced with the introduced ones (BM25F and STF). For ease of reference, structure and abbreviations of the proposed and their corresponding models are shown in Table XII.

TABLE XII. NEW EXTENDED RELEVANCE PROPAGATION MODELS, THEIR STRUCTURES, ABBREVIATIONS, AND CORRESPONDING MODELS.

Model		Content similarity function	Abr[3]	Crs_mthd[4]
Stream hyperlink-based score propagation (SHS model)	WI	BM25F	SHS-WI	**HS-WI**
	WO	BM25F	SHS-WO	**HS-WO**
	UO	BM25F	SHS-UO	**HS-UO**
Stream hyperlink-based term propagation (SHT model)	WI	STF	SHT-WI	**HT-WI**
	WO	STF	SHT-WO	**HT-WO**
	UO	STF	SHT-UO	**HT-UO**
Stream popularity-based score propagation using hyperlink (SPSH model)	WI	BM25F	SPSH-WI	**PSH-WI**
	WO	BM25F	SPSH-WO	**PSH-WO**
	UO	BM25F	SPSH-UO	**PSH-UO**
Stream popularity-based term propagation using hyperlink (SPTH model)	WI	STF	SPTH-WI	**PTH-WI**
	WO	STF	SPTH-WO	**PTH-WO**
	UO	**STF**	**SPTH-UO**	**PTH-UO**

[3] Abbreviation
[4] Corresponding method

IV. EMPIRICAL EVALUATIONS

In this section, performance and effectiveness of the extended models are evaluated against the old ones. First, experimental settings, some implementation issues and evaluation measures are investigated; then, the results of effectiveness evaluation are shown.

A. Experimental settings

For the purpose of "effectiveness evaluation", the ".GOV" corpus of the LETOR 3.0 [48] was used. LETOR is a benchmark collection for the research on learning to rank for IR, released by Microsoft Research Asia (MSRA). LETOR 3.0 contains standard features, relevance judgments, data partitioning, evaluation tools and several baselines for the OHSUMED and .GOV data collection. Version 3.0 was released in December, 2008. The .GOV corpus, which was 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. For the present query set, topic distillation task was used 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 websites.

B. Constructing the working set

Following other researchers [16, 18- 20], instead of running the experiments on the whole set of data, for each query, first, a working set was constructed. To construct the working set, we first found the top *400* pages with the highest score as the core set. Then, the core set was expanded to the working set by adding the pages pointing to the pages in core set (Citing Set) and the pages were pointed by the pages in the core set (Cited Set) (Figure 2).

$$WorkingSet = (CoreSet \cup CitingSet \cup CitedSet) \cap Relevant Set$$

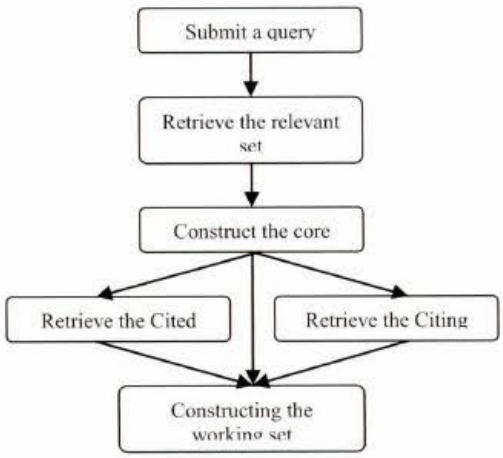

Figure 2. The flowchart of the working set construction.

In the working set construction phase of the proposed models, *BM25F* and *STF* were used as the relevance weighting function for score-level and term-level models, respectively.

C. Evaluation measures

For the purpose of evaluation, a number of evaluation measures which are commonly used in information retrieval, namely Precision at n (P@n) [1], Mean Average Precision (MAP) [1] and Normalized Discount Cumulative Gain (NDCG) were applied [49].

1) Precision at n (P@n)

As quoted in [1], precision at n measures relevance of the top *n* documents in the ranking list with respect to a given query:

$$p@n = \frac{\#\text{of relevance docs in top n results}}{n} \quad (21)$$

2) Mean average precision (MAP)

Average precision (AP) [1] of a given query is calculated as Eq. (22) and corresponds to average of *p@n* values for all relevant documents:

$$AP = \frac{\sum_{i=1}^{N} (\text{P@i} * \text{rel(i)})}{\#\text{total relevant docs for this quer}} \quad (22)$$

where N is the number of retrieved documents and *rel(n)* is a binary function that is evaluated 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.

3) Normalized discount cumulative gain (NDCG)

For a single query, the NDCG value of its ranking list in position *n* is computed by Eq. (23):

$$NDCG(n) = Z_n \sum_{j=1}^{n} \begin{cases} 2^{r_j} - 1, j = 1 \\ \frac{2^{r_j} - 1}{\log(j)}, j > 1 \end{cases} \quad (23)$$

where *r(j)* is 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 Letor 3.0, there are two ratings {0, 1} corresponding to "relevant" and "not relevant" in order to compute NDCG scores.

4) Effectiveness Evaluation

Experimental evaluations of the proposed models against their corresponding models are provided in Figures 3 and 4, which demonstrate performances of the relevance propagation models on the ".GOV" corpus with the TD2003 & TD2004 query sets. As can be seen, all of the stream-based models (with proper parameters) boosted the retrieval performance compared to the old ones.

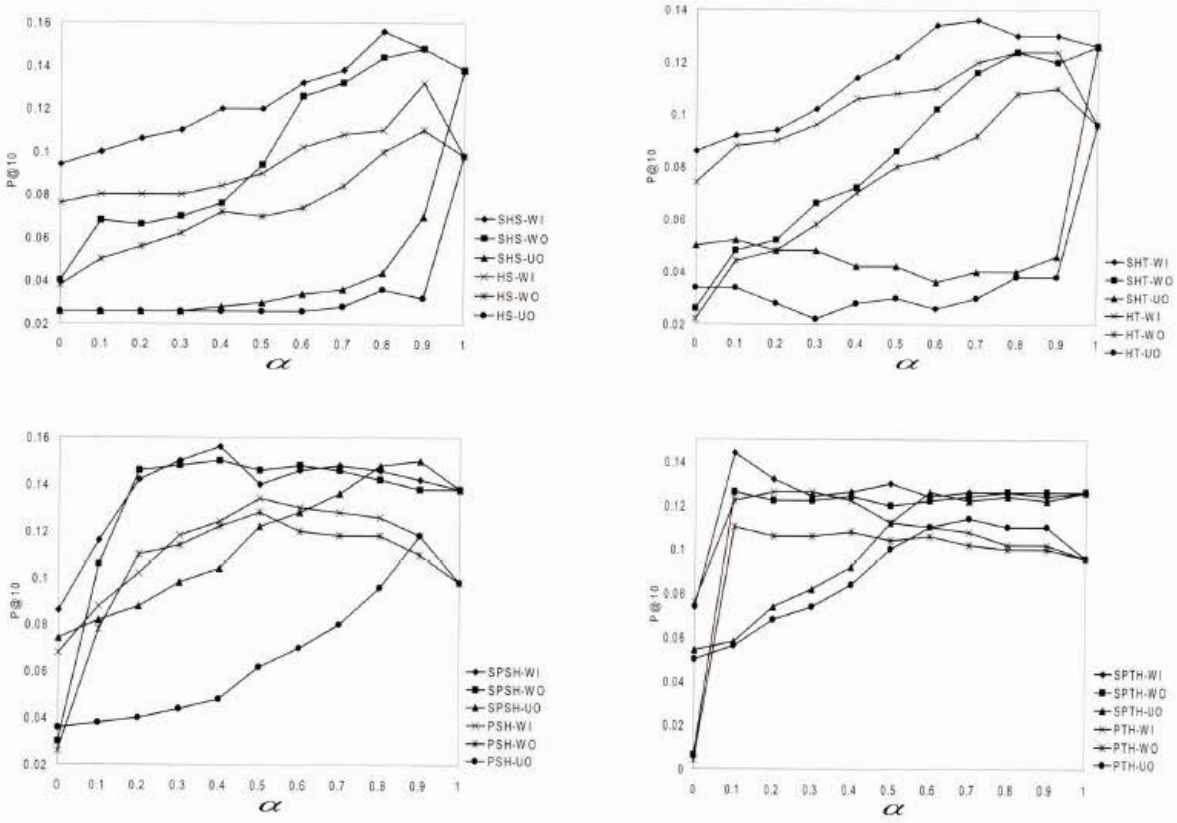

Figure 3. Performance on the ".GOV" corpus with TD2003 (in terms of P@10).

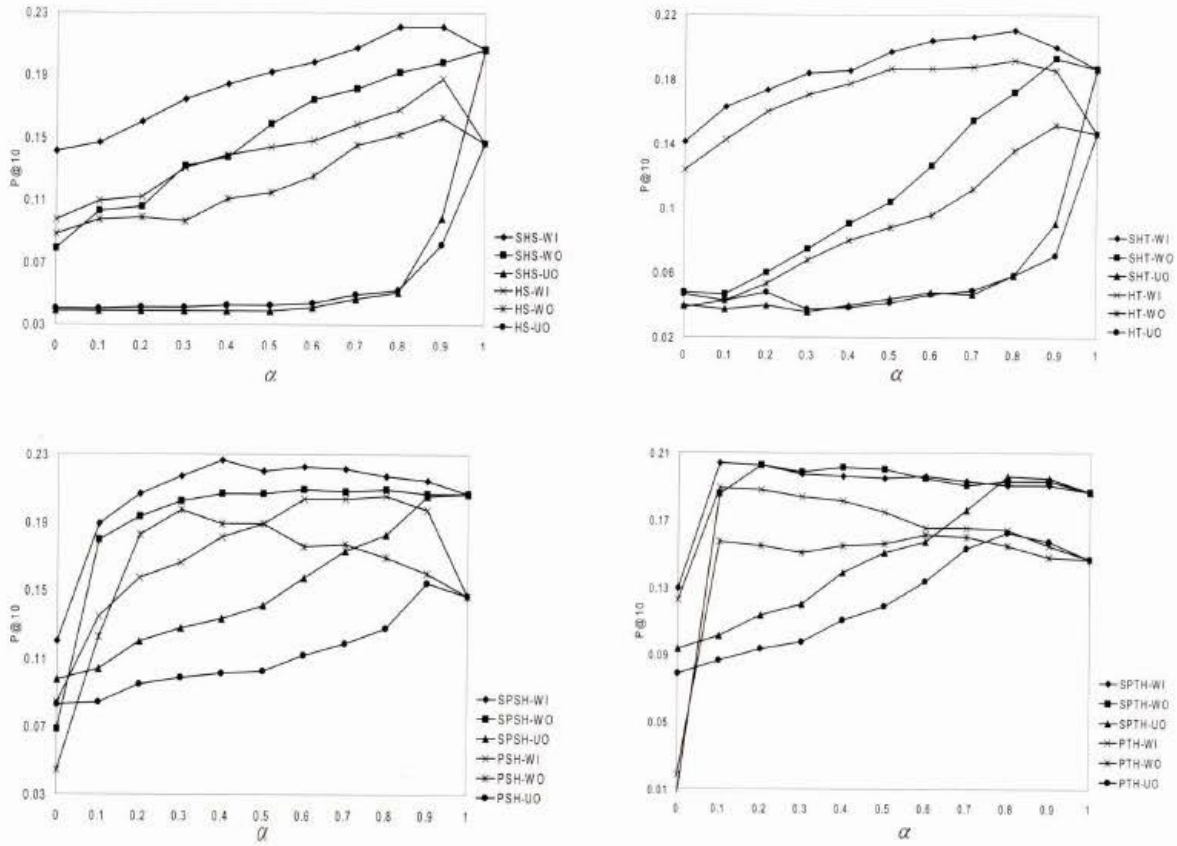

Figure 4. Performance on the ".GOV" corpus with TD2004 (in terms of P@10).

To gain more understanding of the performance comparison, Tables XIII and XIV list the best MAP and NDCG@(1, 2, 3, and 10) for each algorithm. Using these tables, it can be found that stream-based models outperformed others. For example, the best MAP of the SPSH-WI method was 0.241, which was not only 26% better than the corresponding method (PSH-WI), but also won the best result reported in TREC 2003. Similarly, in TREC 2004, the best NDCG@10 of the SHS-WI method was 0.289, which outperformed the HS-WI (corresponding method) by 12%.

TABLE XIII. PERFORMANCE COMPARISON OF NEW MODELS AGAINST THE OLD ONES.

Model		GOV with TD-2003			GOV with TD-2004		
		α	Best MAP	Improvement	α	Best MAP	Improvement
Content similarity functions	**BM25F**	-	**0.186**	*40%*	-	**0.184**	*28%*
	BM25	-	0.133		-	0.144	
	STF	-	**0.169**	*26%*	-	**0.17**	*18%*
	TF	-	0.134		-	0.144	
SHS and HS models	**SHS-WI**	**0.95**	**0.240**	*23%*	**0.95**	**0.196**	*11%*
	HS-WI	0.97	0.196		0.97	0.177	
	SHS-WO	**0.9**	**0.218**	*44%*	**0.9**	**0.183**	*23%*
	HS-WO	0.9	0.152		0.9	0.148	
	SHS-UO	**0.97**	**0.222**	*67%*	**1**	**0.184**	*28%*
	HS-UO	1	0.133		1	0.144	
SHT and HT models	**SHT-WI**	**0.85**	**0.215**	*15%*	**0.85**	**0.194**	*9%*
	HT-WI	0.80	0.187		0.80	0.178	
	SHT-WO	**0.95**	**0.183**	*21%*	**0.95**	**0.171**	*18%*
	HT-WO	0.85	0.151		0.85	0.145	
	SHT-UO	**1**	**0.169**	*26%*	**1**	**0.169**	*17%*
	HT-UO	1	0.134		1	0.144	
SPSH and PSH models	**SPSH-WI**	**0.7**	**0.241**	*26%*	**0.7**	**0.197**	*11%*
	PSH-WI	0.80	0.192		0.80	0.177	
	SPSH-WO	**0.5**	**0.217**	*21%*	**0.45**	**0.186**	*11%*
	PSH-WO	0.4	0.179		0.4	0.167	
	SPSH-UO	**0.85**	**0.237**	*34%*	**0.9**	**0.185**	*27%*
	PSH-UO	0.9	0.174		0.9	0.146	
SPTH and PTH models	**SPTH-WI**	**0.1**	**0.236**	*16%*	**0.1**	**0.184**	*5%*
	PTH-WI	0.1	0.204		0.1	0.175	
	SPTH-WO	**0.5**	**0.186**	*19%*	**0.5**	**0.179**	*18%*
	PTH-WO	0.3	0.156		0.3	0.152	
	SPTH-UO	**0.8**	**0.186**	*18%*	**0.8**	**0.178**	*14%*
	PTH-UO	**0.8**	**0.157**		**0.8**	**0.156**	

TABLE XIV. BEST PERFORMANCE OF EACH ALGORITHM (IN TERMS OF NDCG@N).

Model		GOV with TD-2003					GOV with TD-2004				
		α	NDCG @1	NDCG @2	NDCG @3	NDCG @10	α	NDCG @1	NDCG @2	NDCG @3	NDCG @10
Content similarity functions	**BM25F**	-	**0.24**	**0.25**	**0.238**	**0.256**	-	**0.32**	**0.266**	**0.269**	0.259
	BM25	-	0.14	0.13	0.157	0.179	-	0.24	0.226	0.213	**0.189**
	STF	-	**0.18**	**0.21**	**0.203**	**0.232**	-	**0.266**	**0.273**	**0.262**	0.236
	TF	-	0.14	0.13	0.152	0.177	-	0.253	0.226	0.210	**0.188**
SHS and HS models	**SHS-WI**	**0.95**	**0.40**	**0.33**	**0.29**	**0.296**	**0.95**	**0.36**	**0.353**	**0.338**	0.289
	HS-WI	0.97	0.30	0.28	0.279	0.259	0.97	0.28	0.28	0.265	**0.259**
	SHS-WO	**0.9**	**0.32**	**0.31**	**0.275**	**0.279**	**0.9**	**0.346**	**0.326**	**0.315**	0.264
	HS-WO	0.9	0.18	0.19	0.185	0.206	0.9	0.266	0.206	0.198	**0.197**
	SHS-UO	**0.97**	**0.28**	**0.26**	**0.25**	**0.261**	**1**	**0.32**	**0.266**	**0.269**	0.257
	HS-UO	1	0.14	0.13	0.157	0.179	1	0.24	0.226	0.213	**0.189**
SHT and HT models	**SHT-WI**	**0.85**	**0.34**	**0.30**	**0.277**	**0.273**	**0.85**	**0.373**	**0.326**	**0.312**	0.273
	HT-WI	0.8	0.26	0.27	0.264	0.251	0.8	0.333	0.26	0.264	**0.247**
	SHT-WO	**0.95**	**0.22**	**0.19**	**0.205**	**0.232**	**0.95**	**0.28**	**0.26**	**0.258**	0.241
	HT-WO	0.85	0.10	0.15	0.164	0.203	0.85	0.253	0.186	0.186	**0.188**
	SHT-UO	**1**	**0.18**	**0.21**	**0.203**	**0.232**	**1**	**0.266**	**0.273**	**0.262**	0.236
	HT-UO	1	0.14	0.13	0.152	0.177	1	0.253	0.226	0.210	**0.188**
SPSH and PSH models	**SPSH-WI**	**0.7**	**0.400**	**0.340**	**0.299**	**0.301**	**0.7**	**0.346**	**0.346**	**0.337**	0.288
	PSH-WI	0.80	0.28	0.28	0.274	0.255	0.80	0.28	0.273	0.273	**0.260**
	SPSH-WO	**0.5**	**0.28**	**0.30**	**0.267**	**0.277**	**0.45**	**0.333**	**0.333**	**0.333**	0.272
	PSH-WO	0.4	0.26	0.24	0.207	0.234	0.4	0.333	0.286	0.278	**0.243**
	SPSH-UO	**0.85**	**0.30**	**0.29**	**0.284**	**0.287**	**0.9**	**0.32**	**0.346**	**0.327**	0.273
	PSH-UO	0.9	0.20	0.23	0.212	0.229	0.9	0.226	0.22	0.223	**0.201**
SPTH and PTH models	**SPTH-WI**	**0.1**	**0.36**	**0.33**	**0.308**	**0.299**	**0.1**	**0.333**	**0.273**	**0.287**	0.261
	PTH-WI	0.1	0.28	0.25	0.266	0.251	0.1	0.346	0.313	0.295	**0.252**
	SPTH-WO	**0.5**	**0.24**	**0.19**	**0.195**	**0.229**	**0.5**	**0.32**	**0.28**	**0.264**	0.248
	PTH-WO	0.3	0.12	0.15	0.162	0.197	0.3	0.20	0.24	0.233	**0.202**
	SPTH-UO	**0.8**	**0.22**	**0.23**	**0.224**	**0.241**	**0.8**	**0.32**	**0.28**	**0.267**	0.250
	PTH-UO	**0.8**	**0.18**	**0.18**	**0.195**	**0.208**	**0.8**	**0.266**	**0.246**	**0.238**	**0.214**

Similar conclusions can be drawn from these tables to those from Figures 3 and 4 that splitting web pages to different streams provides significant improvement in relevance propagation models.

V. EFFICIENCY EVALUATION

In the previous section, effectiveness of the relevance propagation models was investigated. However, for real-world applications, efficiency is another important factor besides effectiveness. In this regard, efficiency of the models is evaluated in this section to see their potential of being used in search engines.

Roughly speaking, typical architecture of a search engine has three components [1, 2]: crawler, indexer and searcher. If relevance propagation technologies are to be integrated into search engine, these three components should be considered. Clearly, relevance propagation could be only embedded in the second or third components. Since the search engine indexes the web offline and implements the search operation online, efficiency of relevance propagation will be discussed for the online and offline cases, respectively.

A. Online complexity

Due to the algorithm descriptions, all the relevance propagation models have two kinds of computations. The first one is to retrieve the relevant pages and rank them by relevance weighting functions. Actually this is also needed by existing search engines. The second is the additionally-introduced complexity, including working set construction, relevance propagation and so on. This will be the major concern when integrating these models into the search engines. In this regard, we will focus on the analysis of these additional computations in this section. According to the model formulation and the implementation issues, we can get the following estimations on the online complexity of

the relevance propagation models. Note that the time complexity we estimate here is for one query.

1. For each step of iteration in the score-level models (HS, PSH, SHS, and SPSH), we need to propagate the relevance score of a page along its in-link or out-link in the sub graph of the working set. Note that the source and destination pages of the hyperlink should be both in the working set, and so the average numbers of in-links and out-links per page are equal to each other. We denote this number by l. If we further use c_h to indicate the time complexity of propagating an entity from a page to another page along hyperlinks, we can get that the complexity of each step of iteration in the score-level models is wlc_h. Where w is the size of the working set. If it takes t iterations for the propagation to converge, the overall complexity will be $twlc_h$.
2. Similar to the analysis of the score-level models, complexity of all the term-level models (HT, PTH, SHT and SPTH) can be obtained as $twlc_h$.

B. *Offline complexity*

Since a real search engines should handle hundreds of queries per second [1, 2], it will be very difficult to implement these propagation techniques online. So offline implementation is much more preferred if we want to apply them in real-world applications. Search engines usually build offline invert and forward indices to store the information of each term (including frequency, position and so on) in web pages [1, 2]. Then it is easily understood that term-level propagation models can well match this mechanism and we only need to refine the offline index files. To illustrate it, let us take the SHT-WI method for example. Suppose the parent pages of page p contain a particular word, and we need to propagate the occurrence frequency of this word to page p. If p already contains this particular word, we only need to modify its frequency; while if p does not contain the word, we need to add its ID to the forward index [2] of page p, and then update its term frequency. Comparatively, the score-level propagation models could hardly be integrated into search engines, because scores do not exist in the offline indices but are dependent on the online relevance ranking algorithm used in the search engine.

VI. CONCLUSIONS

In this paper, a comprehensive study was conducted on relevance propagation in web information retrieval. In particular, two generic stream-based content similarity functions were proposed and they were shown to be used to derive new relevance propagation models. Then, effectiveness of the propagation models were investigated using experimental verifications. To evaluate the propagation models, the Letor 3.0 benchmark collection was used in the experiments. The following conclusions were drawn from the presented work:

1. Splitting web pages to different streams can boost accuracy of relevance propagation models (especially for P@1 and NDCG@1).
2. Among relevance propagation models, stream-based propagation models outperform others.
3. Among the stream-based models, SHS and SPSH models obtain the best results.

There is one interesting direction for further research that, other than neighbor sets derived from the explicit link structure of the web, other types of neighbors can be also defined. In general, propagation models allow for definition of any set of documents with a specific characteristic as a neighbor set. As an example, the set of pages with similar content can be defined as a neighbor set [50]. It is interesting to see whether exploiting these types of neighbors can further improve retrieval accuracy.

REFERENCES

[1] R. Baeza-Yates & B. Ribeiro-Neto, "Modern Information Retrieval". *ACM Press/Addison Wesley*, 1999.

[2] S. Brin, L. Page, "The Anatomy of a Large Scale Hypertextual Web Search Engine", *Proc. 7th WWW*, 1998.

[3] MA. Golshani, A.M. ZarehBidoki, "IECA: Intelligent Effective Crawling Algorithm for Web pages", *International Journal of Information & Communication Technology Research (IJICTR)*, 2012.

[4] S. Mukherjea, "Discovering and analyzing World Wide Web collections", *Knowledge and Information Systems*, Vol. 6, No. 2, pp. 230-241, March 2004.

[5] Z. Gong, Leong, and C.W.Cheang., "Web image indexing by using associated texts", *Knowledge and Information Systems*, Vol. 10, No. 2, pp. 243-264, August 2006.

[6] F. Xia, et al., "Ranking with decision tree", *Knowledge and Information Systems*,Vol. 17, No. 3, pp. 381-395, December 2008.

[7] M. Al-Kabi, et al, "Content-based analysis to detect Arabic web spam", *Journal of Information Science*, Vol. 38, No. 3, pp. 284-296, June 2012.

[8] J.C. Na,. and T.T. Thet, "Effectiveness of web search results for genre and sentiment classification", *journal of Information Science*, Vol 35, No. 6, pp. 709- 726, December 2009.

[9] N. Hochstotter and M. Koch, "Standard parameters for searching behavior in search engines and their empirical evaluation", *Journal of Information Science*, Vol. 35, No. 1, pp. 45-65, February 2009.

[10] A. Uyar, I"nvestigation of the accuracy of search engine hit counts", *Journal of Information Science*, Vol. 35, No. 4, pp. 469-480, August 2009.

[11] P. Huntington, D. Nicholas, H.R. Jamali, "Employing log metrics to evaluate search behaviour and success: case study BBC search engine", *Journal of Information Science*, Vol. 33, No. 5, pp. 584-597, October 2007.

[12] A. Uyar, "Google stemming mechanisms", *Journal of Information Science*,Vol. 35, No. 5,pp. 499-514, October 2009.

[13] Y.-L Chen, and X.-H. Chen, "An evolutionary PageRank approach for journal ranking with expert judgements", *Journal of Information Science*, Vol. 37, No. 3, pp. 254-272, June 2011.

[14] D.Lewandowski, "A three-year study on the freshness of web search engine databases", *Journal of Information Science*, Vol. 34, No. 6, pp. 817-831, December 2008.

[15] S. Kwon, Y.-G. Kim, and S. Cha, "Web robot detection based on pattern-matching technique", *Journal of Information Science*, Vol. 38, No. 2, pp.118-126 , February 27, 2012.

[16] A. Shakery, & C.X. Zhai, "Relevance Propagation for Topic Distillation UIUC TREC 2003 Web Track Experiments", *In Proceedings of the TREC Conference*, 2003.

[17] A. Shakery, & C.X. Zhai, "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.

[18] T. Qin, T.Y. Liu, X.D. Zhang, Z. Chen, & W.Y. Ma, "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] E. Mousakazemi, M.A. Saram, A.M. ZarehBidoki, "Popularity-based relevance propagation", *Journal of Web Engineering*, Vol . 1, No. 4, pp. 350-364, 2012.

[20] M.A. Golshani and A.M ZarehBidoki, "Slash-Based relevance propagation model for Web page retrievals", *journal of Web engineering*, 2012 (in progess).

[21] A.M. ZarehBidoki, P. Ghodsnia, N. Yazdani, F. Oroumchian, "A3CRank: an adaptive ranking method based on connectivity, content and click-through data", *Information Processing and Management*, Vol. 46, No. 2, pp. 159-169, 2010.

[22] S. Robertson, K. Jones, "Relevance Weighting of Search Terms", *Journal of the American Society of Information Science*, pp. 129-146.

[23] G. Salton, C. Buckley, "Term weighting approaches in automatic text retrieval", *Information Processing and Management*, Vol. 24, No. 5, pp. 513-523, 1988.

[24] L. Page, S. Brin, R. Motwani, T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", *Technical report, Stanford University, Stanford, CA*, 1998.

[25] A.M. ZarehBidoki, N. Yazdani, "DistanceRank: An intelligent ranking algorithm for web pages", *Information Processing and Management*,Vol. 44, No. 2, pp. 877-892, 2008.

[26] Liu T, Xu J, Qin T, Xiong W, and Li H. Letor: Benchmark dataset for research on learning to rank for information retrieval. Proc. of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval 2007

[27] B. Amento, L. Terveen L, W. Hill, "Does Authority Mean Quality? Predicting Expert Quality Ratings of Web Documents", *Proc. the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 296-303, 2000.

[28] E. Amitay, D. Carmel, A. Darlow, R. Lempel, A. Soffer, "Topic Distillation with Knowledge Agents", *Proc. of the Eleventh Text Retrieval Conference (TREC-11)*, 2002.

[29] K. Bharat K, M. Henzinger M, "Improved algorithms for topic distillation in a hyperlinked environment", *Proc. of SIGIR-98, 21st {ACM} International Conference on Research and Development in Information Retrieval*, pp. 104-111, 1998.

[30] K. Bharat, G. Mihaila,."When Experts Agree: Using Non-affiliated Experts to Rank Popular Topics", *ACM Transactions on Information Systems*, Vol. 20, No. 1, pp. 47-58, 2002.

[31] S. Chakrabarti, "Integrating the Document Object Model with Hyperlinks for Enhanced Topic Distillation and Information Extraction" , *Proc. Proceedings of the 10th international conference on World Wide Web*, pp. 211-220, 2001.

[32] S. Chakrabarti, M. Joshi, V. Tawde, "Enhanced Topic Distillation using Text, Markup Tags, and Hyperlinks", *Proc of the 24th annual international ACM SIGIR conference on Research and development in information retrieva*l, pp. 208-216, 2001.

[33] N. Craswell, D. Hawking, R. Wilkinson, M. Wu, "Overview of the TREC-2003 Web Track", *Proc. of TREC-2003*, 2003.

[34] T. Haveliwala, "Topic-Sensitive Pagerank", *Proc. of the 11th WWW*, pp. 517-526, 2002.

[35] J. Kleinberg, "Authoritative sources in a hyperlinked environment", *Journal of the ACM*, Vol. 46, No. 5, pp. 604-632, 1999.

[36] O. McBryan, "Tools for Taming the Web", *Proc. of the First International World Wide Web Conference (WWW)*, 1994.

[37] R. Song, J. Wen, S. Shi, G. Xin, T. Liu, T. Qin, X. Zheng, J. Zhang, G. Xue, W. Ma, "Microsoft Research Asia at Web Track and Terabyte Track of TREC 2004", *Proc. the 13th TREC*, 2004.

[38] S. Robertson, "Overview of the Okapi Projects", *Journal of Documentation*, Vol. 53, No. 1, pp. 53: 3-7, 1997.

[39] S. Pandey and C. Olston C, "User-centric Web crawling" , *In 14th international conference on World Wide Web*, pp. 401-411, 2005.

[40] S. Robertson, and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond", *Foundations and Trends in Information Retrieval*,Vol. 3, No. 4, pp. 333-389, 2009.

[41] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, M. Gatford, "Okapi at trec-3", *In TREC'94*, pp. 109–126, 1994.

[42] A. Singhal, C. Buckley , and M. Mitra, "Pivoted document length normalization", *In SIGIR '96*, pp.21–29, 1996.

[43] J. Ponte and W. Croft, "A language modeling approach to information retrieval", *In SIGIR*, pp. 275–281, 1998.

[44] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to ad hoc information retrieval", *In SIGIR*, pp. 334–342, 2001.

[45] Y. Lv , C. Zhai, "Adaptive term frequency normalization for bm25", *In Proceedings of the 20th ACM international conference on Information and knowledge management, pp.* 1985-1988, 2011.

[46] H. Fang, T. Tao, C. Zhai, "A formal study of information retrieval heuristics", *In SIGIR Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 49-56, 2004.

[47] H.P. Luhn, "A statistical approach to mechanized encoding and searching of literary information", *IBM Journal of Research and Development*, Vol. 1, No. 4, pp. 309-317, October 1957.

[48] T Qin, T.Y Liu, J. Xu, & H. Li, "Letor: A benchmark collection for research on learning to rank for information retrieval", *Information Retrieval Journal*, Vol. 13, No. 4, pp. 346-374, 2010.

[49] K. Jarvelin and J. Kekalainen, "Cumulated Gain-based Evaluation of IR Techniques", *ACM Transactions on Information Systems*, Vol. 20, No. 4, pp. 422–446, 2002.

[50] O. Kurland and L. Lee. "Pagerank without hyperlinks: structural re-ranking using links induced by language models", *In Proceedings of ACM SIGIR*, pp. 306–313, 2005.

Mohammad Amin Golshani received his B.Sc. degree in computer engineering from Department of Electrical and Computer Engineering of Zanjan University, Iran, in 2009 and his M.Sc. degree in Information Technology (IT), under the supervision of assistant professor ZarehBidoki in Yazd University, Yazd, Iran, 2011. His current research interests include web crawling, web ranking, and spam detection.

AliMohammad ZarehBidoki got his B.Sc. degree in computer engineering from Isfahan University of Technology in 1999. He also received his M.Sc. in Computer Architecture from University of Tehran in 2002. Furthermore he graduated as a Ph.D. in Computer Engineering at University of Tehran in 2009. His research interests include Web Information Retrieval, Searc engines and Data Mining. He has published more than 30 papers in international journals and conference. Currently, he is an assistant professor in Yazd University.