2007 IEEE/WIC/ACM International Conference on Web Intelligence

Extending Link-based Algorithms for Similar Web Pages with Neighborhood Structure

Zhenjiang Lin, Michael R. Lyu and Irwin King
Department of Computer Science and Engineering
The Chinese University of Hong Kong
{zjlin, lyu, king}@cse.cuhk.edu.hk

Abstract

The problem of finding similar pages to a given web page arises in many web applications such as search engine. In this paper, we focus on the link-based similarity measures which compute web page similarity solely from the hyperlinks of the Web. We first propose a simple model called the Extended Neighborhood Structure (ENS), which defines a bi-directional *(in-link and out-link) and* multi-hop *neighborhood structure. Based on the ENS model, several existing similarity measures are extended. Preliminary experimental results show that the accuracy of the extended algorithms are significantly improved.*

1 Introduction

Unlike the *keyword searching* which searches web pages related to the query keywords provided by users, *instance searching* searches by instance. That is, it takes a web page as the input and returns a list of related (or similar) web pages to this page. For example, for a query such as "www.cnn.com", the searching result would be such web pages related to news as "www.usnews.com" or "news.bbc.co.uk". One advantage of instance searching is that users can find the related web pages to the web page they are interested in, without having to worry about selecting the right keywords.

The problem of finding similar pages to a given web page arises in many web applications. One example is the "similar pages" service of Google. Each time users click on the 'Similar Pages' link for a search result, Google automatically searches the Web for pages that are related to this result. Web document classification (such as Yahoo! Directory) categorizes web pages into a hierarchical structure according to the degree of similarity between web pages.

In the fields of information retrieval (IR) and recommender systems, the problem of finding similar objects has been studied extensively for many years. In traditional IR, text-based similarity measures based on *matching text* may be employed to find similar documents to a given query in a document corpus. For collaborative filtering in a recommender system, similar users may be grouped by *users' preferences*. In particular, several link-based algorithms have been suggested to exploit the similarity information hidden in the link structure of graph, such as Co-citation, bibliographic coupling [5], SimRank [2], and PageSim [4]. Further methods arise from graph theory, such as similarity measure that is based on network flows.

One major problem of the link-based algorithms is that they usually do not make full use of the structural information of the graph. For example, Co-citation only considers direct neighbors. SimRank is a multi-hop algorithm, but it considers only one direction. We believe that a well-designed algorithm should take into account as much link information as possible to produce high quality results.

Motivation and Contributions: To develop effective and flexible similarity measures which make full use of the structural information of the Web to produce high quality results motivates our research work. In this paper, we focus on the link-based similarity measures which compute similarity between web pages solely from the hyperlink structure modelled by the web graph, with vertices representing web pages and directed edges representing hyperlinks between pages. The main contributions of this paper are as follows.

1. We propose the Extended Neighborhood Structure (ENS) which defines a *bi-directional* (in-link and out-link) and *multi-hop* neighborhood structure. This model is designed for helping link-based algorithms make full use of the link information of a graph.

2. Several similarity measures are extended based on the ENS model. The accuracy of the extended algorithms improves significantly, which illustrates the effectiveness of the ENS.

0-7695-3026-5/07 $25.00 © 2007 IEEE
DOI 10.1109/WI.2007.125

2 Extended Neighborhood Structure Model

Recent research has suggested that there are large amounts of valuable information hidden in the vast link structure of the Web. For example, a web page linking to another page usually implies some kind of relationship between them. This is because the fact that generally authors of web pages would like to link their pages to those pages which they think are related to theirs.

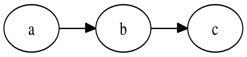

Figure 1. Interpretation of the ENS model

Consider the graph in Figure 1, the basic information hidden in the fact that a links to b is that a knows b. Certainly, b may not know a since there's no link from b to a. It is very much like the relationship between people. Therefore, a web page may have two kinds of neighbors: *in-link neighbors* (those who know it) and *out-link neighbors* (whom it knows). In Figure 1, a is b's in-link neighbor and b is a's out-link neighbor. Now, we can come up with a straightforward intuition on web page similarity: *similar web pages have similar neighbors*. On the other hand, page c is a 2-hop indirect out-link neighbor of a, which implies page a may not be so familiar with c as with b. This can be thought as the familiarity decreases along links.

Therefore, the concept of neighborhood is now extended in two aspects: *bi-direction* and *multi-hop*. Although the intuition of similarity is still "similar web pages have similar neighbors", its meaning is generalized, since the "neighbors" here refer to the bi-directional and multi-hop neighbors instead of single-direction or direct neighbors. This model is based on the natural hypothesis that a link-based algorithm likely improves its accuracy by considering more structural information of the graph. Throughout the paper, notation $Sim(a, b)$ represents the similarity score of web pages a and b produced by the similarity measures.

3 Extending Similarity Measures

3.1 Web Graph Model

We model the Web as a directed graph $G = (V, E)$ with vertices V representing web pages $v_i(i = 1, 2, \cdots, n)$ and directed edges E representing hyperlinks among the web pages. $I(v)$ denotes the set of in-link neighbors of pages v and $O(v)$ denotes the set of out-link neighbors of page v. Let $path(u_1, u_s)$ denote a sequence of vertices $u_1, u_2, \ldots, u_s$ such that $(u_i, u_{i+1}) \in E(i = 1, \cdots, s-1)$ and u_i are distinct. It is called a **path** from u_1 to u_s. Let $length(p)$ denote the **length** of path p, and define $length(p) = |p| - 1$, where $|p|$ is the number of vertices in path p. Let $PATH(u, v)$ denote the set of all possible paths from page u to v.

3.2 Extended Co-citation and Bibliographic Coupling

Co-citation and bibliographic coupling are two 1-hop single-directional algorithms. in co-citation, the more common in-link neighbors two pages have, the more similar they are. Therefore, $Sim(a, b) = |I(a) \cap I(b)|$. In bibliographic coupling, the more common out-link neighbors two pages have, the more similar they are. Therefore, $Sim(a, b) = |O(a) \cap O(b)|$.

We can construct a bi-directional algorithm called Extended Co-citation and Bibliographic Coupling (ECBC): the more common neighbors two pages have, the more similar they are. Accordingly,

$$Sim(a, b) = \alpha|I(a) \cap I(b)| + (1-\alpha)|O(a) \cap O(b)|,$$

where $\alpha \in [0, 1]$ is a user-defined constant.

3.3 Extended SimRank

SimRank is a fixed point of the recursive definition: *two pages are similar if they are referenced by similar pages*. Numerically, for any web page u and v, this is specified by defining $Sim(u, u) = 1$ and

$$Sim(u, v) = \gamma \cdot \frac{\sum_{a \in I(u)} \sum_{b \in I(v)} Sim(a, b)}{|I(u)||I(v)|}$$

for $u \neq v$ and $\gamma \in (0, 1)$. If $I(u)$ or $I(v)$ is empty, then $Sim(u, v)$ is zero by definition. The SimRank iteration starts with $Sim(u, v) = 1$ for $u = v$, and 0 otherwise.

SimRank is a multi-hop algorithm, but it is not bi-directional. We extend the intuition of SimRank to be "two pages are similar if they have similar neighbors". Accordingly, SimRank can be extended to $Sim(u, v) =$

$$\gamma \cdot (\sum_{a \in I(u)} \sum_{b \in I(v)} Sim(a, b) + \sum_{a \in O(u)} \sum_{b \in O(v)} Sim(a, b))$$

$$\times(|I(u)||I(v)| + |O(u)||O(v)|)^{-1}.$$

3.4 Extended PageSim

PageSim can be regarded as a "weighted multi-hop" version of Co-citation algorithm. First, it takes the common in-link information of 1-hop as well as multi-hop neighbors into account to improve the quality of the result. Moreover, PageSim also considers the importance of the web pages. PageSim can be described as: each web page propagates its "feature information" to its (multi-hop) out-link neighbors along hyperlinks. Once the propagation of all web pages finish, the similarity between two pages is measured by the "feature information" they have in common.

For a web page, the volume of its "feature information" is measured by its PageRank score or other score that represents the importance of this page. Obviously, each web page has to preserve a space, which is called the *feature vector*, to store the "feature information" of its own as well as all the others.

Definition 1 *Let $PR(v)$ denote the PR score of page v. Let $PG(u, v)$ denote the PR score of page u that propagates to page v through $PATH(u, v)$. We define*

$$PG(u,v) = \begin{cases} \sum_{p\in PATH(u,v)} \frac{d\cdot PR(u)}{\prod_{w\in p, w\neq v} |O(w)|} & v \neq u, \\ \\ PR(u) & v = u, \end{cases} \tag{1}$$

where $d \in (0, 1]$ is a decay factor and $u, v \in V$.

Definition 2 *Let $\overrightarrow{FV}(v)$ denote the Feature Vector of page v, we have $\overrightarrow{FV}(v) = (PG(v_i, v))^T = (PG_i(v))^T, i = 1, \cdots, n$, where $v, v_i \in V$.*

Definition 3 *Let $PS(u, v)$ denote the PageSim score of pages u and page v. We define*

$$PS(u,v) = \sum_{i=1}^{n} \frac{min(PG_i(u), PG_i(v))^2}{max(PG_i(u), PG_i(v))}, \tag{2}$$

where $u, v \in V$.

The detailed explanations of the above definitions are given in [4]. In short, there are two stages in the PageSim algorithm: *PR score propagation stage* and *PS score computation stage*. Equations (1) and (2) correspond to the processes in these two stages respectively.

Extended PageSim (EPS): In PageSim, the "feature information" of web pages propagate along only out-links, and the consequent PS scores are actually "out-link" PS scores. In EPS, we also propagate along in-links (with decay factor $1-d$) and produce the "in-link" PS scores. This is because we consider the in-links complement to out-links. The EPS score of two pages is hence defined by the sum of "in-link" and "out-link" PS scores of them. We denote the EPS score of pages u and v by $EPS(u, v)$.

Moreover, we adopt the Jaccard measure [1], which is commonly used in IR to measure the similarity between two vectors, to calculate the Extended PageSim (EPS) scores. For example, to calculate the "out-link" PS score of pages u and v, we use Equation (3) instead of Equation (2).

$$PS(u,v) = \frac{\sum_{i=1}^{n} min(PG(v_i,u), PG(v_i,v))}{\sum_{i=1}^{n} max(PG(v_i,u), PG(v_i,v))}, \tag{3}$$

where $u, v \in V$.

4 Experimental Results

We have proposed the ENS model and extended several link-based similarity measures, including Co-citation, bibliographic coupling, SimRank, and PageSim. In this section, we report on some preliminary experimental results. The primary purpose is to show that the ENS model indeed helps link-based similarity measures improve their accuracy.

4.1 Datasets

We tested the algorithms on two datasets:

CSE Web (CW) dataset is a set of web pages crawled from *http://www.cse.cuhk.edu.hk*, containing about 22,000 web pages and 180,000 hyperlinks.

Google Scholar (GS) dataset contains a citation graph of 20,000 articles which were crawled through public interface of Google Scholar search engine, with vertices representing articles and directed edges representing citations between articles. To obtain this dataset, we first submitted keyword "web mining" to the Google Scholar which returned 50 related articles as a result. Then we crawled the remaining articles by following the "Cited By" hyperlinks of the search results using Breadth-First Search algorithm.

4.2 Ground Truth and Evaluation Methods

For any vertex v in graph G, a similarity measure A would produce a list of top N vertices most similar to v (excluding v itself), which is denoted by $top_{A,N}(v)$. Let the number $score_{A,N}(v)$ denote the average score to v of the $top_{A,N}(v)$. Thereby, we consider the average number of $score_{A,N}(v)$ for all $v \in V$ as the quality of the top N results produced by algorithm A, which is denoted by $\Delta(A, N)$. That is, $\Delta(A, N) = (\sum_{v\in V} score_{A,N}(v))/n$.

In this paper, we use two different evaluation methods. For the CW dataset, we use the cosine TFIDF, a traditional text-based similarity function, as rough metrics of similarity. For the GS dataset, we use the "Related Articles" provided by Google Scholar as ground truth.

(1) Cosine TFIDF Similarity: The cosine TFIDF score of two web pages u and v is just the cosine of angle between TFIDF vectors of the pages [3], which is defined by

$$TFIDF(u,v) = \frac{\sum_{t\in u\cap v} W_{tu} \cdot W_{tv}}{\|u\| \cdot \|v\|},$$

where W_{tu} and W_{tu} are TFIDF weights of term t for web pages u and v respectively. $\|v\|$ denotes the length of page v, which is defined by $\|v\| = \sqrt{\sum_{t\in v} W_{tv}^2}$.

Therefore, for the CW dataset, we define

$$score_{A,N}(v) = \frac{1}{N} \sum_{u\in top_{A,N}(v)} TFIDF(u,v),$$

and $\Delta^T(A, N) = \Delta(A, N)$ which measures the average cosine TFIDF score of top N similar web pages returned by algorithm A.

(2) Related Articles: For an article v in citation graph G, the list of its "Related Articles" returned by Google Scholar is denoted by $RA(v)$. We define $related_N(v) = \{\texttt{top N related articles } v_i | v_i \in RA(v) \cap V\}$.

The precision of similarity measure A at rank N is

$$precision_{A,N}(v) = \frac{|top_{A,N}(v) \cap related_N(v)|}{|related_N(v)|}.$$

Therefore, for the GS dataset, we simply define $score_{A,N}(v) = precision_{A,N}(v)$, and $\Delta^P(A, N) = \Delta(A, N)$ which measures the average precision of algorithm A at top N.

4.3 Performance Evaluation of Algorithms

In this part, we evaluate the algorithms mentioned in this paper on the CW and GS datasets. These algorithms include Co-citation (*CC*), Bibliographic coupling (*BC*), Extended Co-citation and Bibliographic Coupling (*ECBC*), SimRank (*SR*), Extended SimRank (*ESR*), PageSim (*PS*), and Extended PageSim (*EPS*). The parameter settings of the algorithms are listed in Table 1.

Table 1. Parameter Settings

ECBC	SR	ESR	PS	EPS
$\alpha = 0.5$	$\gamma = 0.8$	$\gamma = 0.8$	$r = 3$, $d = 0.5$	$r = 3$, $d = 0.6$

Due to the space limitation, we only show the detailed result on the CW dataset, and summarize the accuracy improvement across all N for each algorithm-pair in Table 2. Figure 2 plots the curves of $\Delta^T(A, N)$ for different algorithms on the CW dataset. Figure 3 shows the the average values of the curves in Figure 2. We can see that the accuracy of the extended algorithms are significantly improved in all test cases and EPS performs best. Similar conclusion can be drawn from the result on the GS dataset.

Table 2. Improvement of Overall Performance

Dataset	ECBC/CC	ECBC/BC	ESR/SR	EPS/PS
CW	18.87%	14.39%	7.86%	4.73%
GS	20.61%	164.21%	36.94%	27.44%

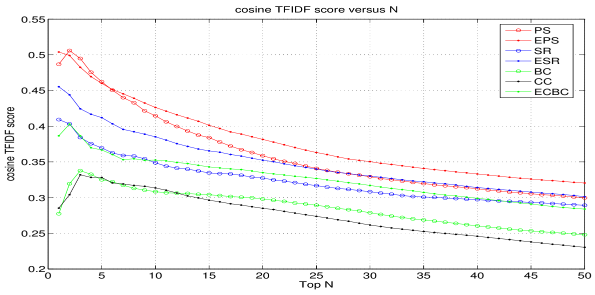

Figure 2. Accuracy of the algorithms on CW

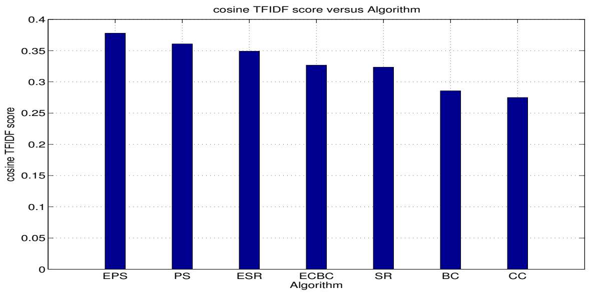

Figure 3. Overall accuracy on CW

5 Conclusion and Future Work

In this paper, we proposed the Extended Neighborhood Structure (ENS) model and extended several link-based similarity measures based on the model. Experimental results show that the accuracy of the extended algorithms are significantly improved.

Acknowledgment This work is supported by grants from the Research Grants Councils of the HKSAR, China (Project No. CUHK4205/04E and Project No. CUHK4235/04E) and is affiliated with the VIEW Technologies Laboratory and the Microsoft-CUHK Joint Laboratory for Human-centric Computing & Interface Technologies.

References

[1] A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., NJ, USA, 1988.

[2] G. Jeh and J. Widom. SimRank: a measure of structural-context similarity. In *Proceedings of KDD '02*, pages 538–543, 2002.

[3] T. Joachims. A probabilistic analysis of the rocchio algorithm with TFIDF for text categorization. In *Proceedings of ICML '97*, pages 143–151, 1997.

[4] Z. Lin, I. King, and M. R. Lyu. Pagesim: A novel link-based similarity measure for the world wide web. In *Proceedings of WI '06*, pages 687–693, 2006.

[5] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1986.