MatchSim: A Novel Neighbor-based Similarity Measure with Maximum Neighborhood Matching

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

ABSTRACT

The problem of measuring similarity between web pages arises in many important Web applications, such as search engines and Web directories. In this paper, we propose a novel neighbor-based similarity measure called MatchSim, which uses only the neighborhood structure of web pages. Technically, MatchSim recursively defines similarity between web pages by the average similarity of the maximum matching between their neighbors. Our method extends the traditional methods which simply count the numbers of common and/or different neighbors. It also successfully overcomes a severe counterintuitive loophole in SimRank, due to its strict consistency with the intuitions of similarity. We give the computational complexity of MatchSim iteration. The accuracy of MatchSim is compared with others on two real datasets. The results show that the method performs best in most cases.

Categories and Subject Descriptors: H.3.3 Information Search and Retrieval: Clustering; Information filtering

General Terms: Algorithms, Measurement

Keywords: Similarity Measure, Link Analysis, Web Mining, Graph Algorithm

1. INTRODUCTION

In the past few years, many popular Web applications have been requiring effective and efficient algorithms to automatically estimate web page similarities, such as search engines (e.g., Google's "similar pages" service) and web page classification services (e.g., Yahoo! Directory). According to the different kinds of input, there are basically two complementary approaches: *text-based* and *link-based*.

The *text-based* methods, originated from IR (information retrieval), use the textual content of web pages to extract similarities. The most notable are the *cosine similarity* and the *TFIDF* models [17]. New methods have been proposed for various of domains [3, 16, 5]. A problem of these methods is that they usually require large storage and long computing time due to the need for full-text comparison, which causes serious scalability problem when dealing with huge amount of and exponentially growing web pages. The *link-based* methods use the hyperlinks which are modelled by the *web graph*, with vertices corresponding to web pages and directed edges to the hyperlinks.

In this paper, we focus on the *neighbor-based* methods, a subset of the *link-based* methods, which share a simple intuition that "web pages are similar because they have similar neighbors". Therefore, the main task of these methods is to estimate the similarity between groups of neighbors. The other subset, the *graph-based* methods, consider the whole structure of the web graph. These methods include the *Maximum Flow/Minmum Cut* [14] that originate from graph theory, the Companion [4] which are derived from the HITS algorithm [10], and PageSim [13] which is based on the feature propagation of web pages, etc. Relatively, the *neighbor-based* methods are usually much easier to implement and faster in running time.

The motivation of this work is to develop efficient and effective neighbor-based similarity measures for the Web applications. The main contribution of the paper is that we propose a novel neighbor-based similarity measure called *MatchSim* which recursively defines the similarity between web pages by the average similarity score of the maximum matching between their neighbors. Moreover, experiments on two real datasets are conducted which demonstrates the good performance of our method.

The rest of the paper is organized as follows. Section 2 gives a brief review on related work. Section 3 describes the MatchSim algorithm in detail, including the definition, computation, and time complexity. Section 4 and 5 report the experimental results and conclude the work respectively.

2. RELATED WORK

The link structure of the Web, which has been greatly influenced by research in the fields of social network and citation analysis, has been widely used to exploit important information inherent in the Web. Successful link-based algorithms include PageRank [15] and HITS.

A number of link-based similarity measures have been proposed in the past few years. Traditional algorithms includes *Co-citation* [18], *bibliographic coupling* [9], and Jaccard Measure [7]. The *Maximum Flow/ Minimum Cut* and *Authority* algorithms were developed for measuring the similarity of scientific papers in a citation graph [14]. The SimRank algorithm was proposed to measure similarity of the structural context "in any domain with object-to-object re-

CIKM'09, November 2–6, 2009, Hong Kong, China.

lationships" [8]. It is a recursive refinement of co-citation, based on the assumption that "two objects are similar if they are referenced by similar objects". The *Jaccard measure* [7] and *Adamic/Ada* [1] were also applied to the link prediction problem underlying social network evolution using only link information in [12]. Interested readers may refer to [12], which contains an exhaustive list of link-based similarity measures.

A review, or even a listing of all the uses of similarity measures is impossible. We summarize the definitions of some well-known neighbor-based methods in Table 1, which are latter used to compare with MatchSim in performance experimentally. In the table, $sim(a,b)$ denotes the similarity score between pages a and b, and $I(a)$ and $O(a)$ the set of *in-link* and *out-link* neighbors of web page a respectively. The notations will be used throughout the paper.

Table 1: Neighbor-based similarity measures

Bibliographic Coupling	$\lvert O(a) \cap O(b) \rvert$
Co-citation	$\lvert I(a) \cap I(b) \rvert$
Jaccard Measure	$\frac{\lvert I(a) \cap I(b) \rvert}{\lvert I(a) \cup I(b) \rvert}$
SimRank	$\gamma \cdot \frac{\sum_{u \in I(a)} \sum_{v \in I(b)} sim(u,v)}{\lvert I(a) \rvert \lvert I(b) \rvert}$, where constant $\gamma \in (0,1)$

3. MATCHSIM ALGORITHM

3.1 The Basic Intuition

Traditional *neighbor-based* methods lack of flexibility since they only consider how many exactly same (and/or different) neighbors two pages have. The recent proposed SimRank makes an extension by taking also the similar neighbors into account. More precisely, SimRank defines the similarity between pages by the average of the overall similarity scores between their neighbors.

For example, given the graph snippet and known similarity scores in Fig. 1, we want to measure the similarity score $sim(a,b)$ between pages a and b. Traditional methods will report that $sim(a,b) = 0$ since the number of common neighbors is 0, which is inaccurate. SimRank outputs $sim(a,b) = \gamma \cdot \sum_{i=1,2} \sum_{j=1,2} sim(a_i,b_j)/4 = 0.4\gamma$, where $\gamma \in (0,1)$ is a decay factor. However, there is a severe loophole. If we remove the most similar neighbors (a_2,b_2), $sim(a,b)$ increases to $\gamma \cdot sim(a_1,b_1)/1 = 0.6\gamma$, which is obviously counterintuitive.

The problem results from the overall sum of similarity scores between neighbors. In fact, we can see that a and b are similar simply because their neighbors are quite *matched*, i.e., neighbors (a_i,b_i) $(i=1,2)$ are similar respectively.

This idea is inspired by the experience that people estimate how similar two objects are by the similarities of their pairwise "matched features" (same or similar features). Because measuring the similarity between one person's finger and others' hair makes no sense. In web graph, pages are objects, and their neighbors are their "features". The similarity between two neighbors reflects how matched they are. Therefore, measuring similarity between pages by the similarities of their (pairwise) *matched* neighbors would be more reasonable.

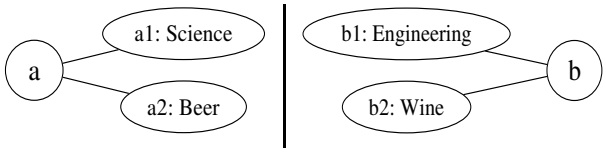

Figure 1: Measuring similarity between a and b based on their neighbors. ($sim(a_1,b_1) = 0.6$, $sim(a_1,b_2) = sim(a_2,b_1) = 0.1$, $sim(a_2,b_2) = 0.8$.)

Next question is how to find those *matched* neighbors. It can be easily modelled by the classic *weighted assignment problem*, in which two groups of neighbors form a bipartite graph, and the similarity scores between them are weights. The aim is to find a matching between the neighbors with maximum sum of similarity scores.

The assignment problem has been studied for many years, and various algorithms have been developed to implement it. This paper adopts the famous Kuhn-Munkres (K-M) algorithm [11] (also known as the Hungarian method). We refer to article [6] for a complete overview of finding maximum matchings in bipartite graphs.

3.2 The Definition

We model the Web graph 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 web pages. Given two pages a and b in a web graph of size n, we obtain a weighted bipartite graph $G_{a,b} = (I(a) + I(b), E, w)$, where $E = \{(u,v) | u \in I(a), v \in I(b)\}$ and $w(u,v) = sim(u,v)$. Based on the recursive intuition of "similar pages have similar neighbors", MatchSim measures the similarity between pages by "the average similarity of the *maximum matching* between their neighbors". Formally, the MatchSim score between two different pages a and b is defined by

$$sim(a,b) = \frac{\widehat{W}(a,b)}{max(I(a),I(b))}. \quad (1)$$

In the cases that $|I(a)| = 0$ or $|I(b)| = 0$, since there is no way to infer any similarity, we simply define $sim(a,b) = 0$. If $a = b$, we have $sim(a,b) = 1$, which is obviously.

In Eq. (1), $\widehat{W}(a,b)$ denotes the weight of maximum matching between $I(a)$ and $I(b)$, i.e.,

$$\widehat{W}(a,b) \triangleq W(m^*_{ab}) = \sum_{(u,v) \in m^*_{ab}} sim(u,v), \quad (2)$$

where m^*_{ab} is a maximum matching between $I(a)$ and $I(b)$.

$\widehat{W}(a,b)$ can be calculated using algorithms for the assignment problem. This paper adopts the Kuhn-Munkres (K-M) algorithm. Since the K-M algorithm always convert $I(a)$ and $I(b)$ to be "equally-sized" before computing m^*_{ab}, we define $l_{ab} \triangleq |m^*_{ab}| = max(I(a),I(b))$. Obviously, any matching between $I(a)$ and $I(b)$ is of size l_{ab}. Therefore, in Eq. (1), the factor $\frac{\widehat{W}(a,b)}{max(I(a),I(b))}$ is exactly the average similarity of a maximum matching between the neighbors of a and b.

3.3 MatchSim Score Computation

For a graph G of size n, we compute the n^2 MatchSim scores iteratively. For each iteration k, we can keep the n^2 scores $sim_k(*,*)$, where $sim_k(a,b)$ is the score between a

and b in iteration k. We successively compute $sim_{k+1}(*,*)$ based on $sim_k(*,*)$. That is, on each iteration $k+1$, we update the $sim_{k+1}(a,b)$ using the similarity scores from the precious iteration k. Formally speaking, we compute $sim_{k+1}(a,b)$ from $sim_k(*,*)$ as follows:

$$sim_{k+1}(a,b) = \frac{\widehat{W}_k(a,b)}{max(I(a),I(b))}, \quad (3)$$

where $\widehat{W}_k(a,b)$ is computed based on the scores $sim_k(*,*)$.

The MatchSim computation starts with $sim_0(a,b) = 1$ for $a = b$ and $sim_0(a,b) = 0$ for $a \neq b$. The MatchSim score between a and b is defined as $lim_{k\to\infty} sim_k(a,b)$. We proved that the limiting values exist and are unique, i.e., the MatchSim iteration converges. Due to space limitation, the detailed proof of convergency is omitted in this paper. In our experiments, the MatchSim computation converges within 15 iterations.

3.4 Complexity Analysis

Time Complexity. For any two pages a and b in a web graph $G = (V, E)$ of size n, we adopt the K-M algorithm to compute $\widehat{W}(a,b)$ in Eq. (1), and so the corresponding time complexity is l_{ab}^3, where $l_{ab} = |m_{ab}^*| = max(|I(a)|, |I(b)|)$. In each iteration, MatchSim invokes the K-M algorithm n^2 times. Supposing there are a total of K iterations and let $L = max_{a,b\in V}(l_{ab}) = max_{a\in V}(Ia)$, the time complexity of MatchSim is thus $O(Kn^2L^3)$.

Space Complexity. MatchSim has to store n^2 MatchSim scores. Moreover, the K-M algorithm invoked needs to store the similarity matrix of two pages, the size of which is $O(L^2)$. Therefore, the space complexity of MatchSim is $O(n^2) + O(L^2) = O(n^2 + L^2)$.

4. EXPERIMENTAL RESULTS

In the experiments, we compare the accuracy of MatchSim(MS) to those of several well-known neighbor-based methods, including Co-citation(CC), Bibliographic Coupling(BC), Jaccard Measure(JM), and SimRank(SR).

4.1 Datasets

We run the algorithms on the following two different kind of real-world datasets. All text in our datasets is in English.

1. **The Computer Web (CW) dataset** is a set of web pages crawled from the web site of our department, which contains 22,615 web pages and 120,947 hyperlinks linking them together. The average number of in-links of web pages is about 5.3.

2. **The Google Scholar (GS) dataset** is a citation graph crawled from Google Scholar [1], containing 20,000 articles and 87,717 citations among them. In the graph, a directed edge (u, v) exists if and only if article u cites v. The average number of in-link citations is about 4.4.

   To obtain this dataset, we first submitted keyword "web mining" to Google Scholar, which returned many related articles. Then we use the first 50 results as starting points and crawled the remaining articles by following the "Cited By" hyperlinks of the search results using the Breadth-First Search algorithm.

[1] http://scholar.google.com

4.2 Evaluation Metrics

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 symbol $score_{A,N}(v)$ denote the average score to v of the $top_{A,N}(v)$. We consider the average value 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, $\Delta(A,N) = (\sum_{v\in V} score_{A,N}(v))/n$.

A good evaluation of the similarity measures is difficult without performing extensive user studies or having a reliable ground truth. In this paper, we use two different evaluation methods as rough metrics of similarity to measure the accuracy of the algorithms. For the CW dataset, we use the cosine TFIDF, a traditional text-based similarity function. For the GS dataset, we use the "Related Articles" provided by Google Scholar. Certainly, neither of these metrics are guaranteed to be perfect, but based on our observation, they are satisfying generally.

(1) Cosine TFIDF Similarity: The cosine TFIDF similarity score of two web pages u and v is just the cosine of the angle between TFIDF vectors of the pages [2], 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.

(3) Overall Accuracy(OA) and Relative OA(ROA): OA and ROA are designed for measuring the "overal accuracy" of an algorithm A over the top N rankings, and the "relative overall accuracy" between two algorithms A and B, respectively. The definitions are as follows.

$$OA(A,N) = \frac{1}{N}\sum_{i=1}^{N}\Delta(A,i), ROA(A,B,N) = \frac{OA(A,N)}{OA(B,N)},$$

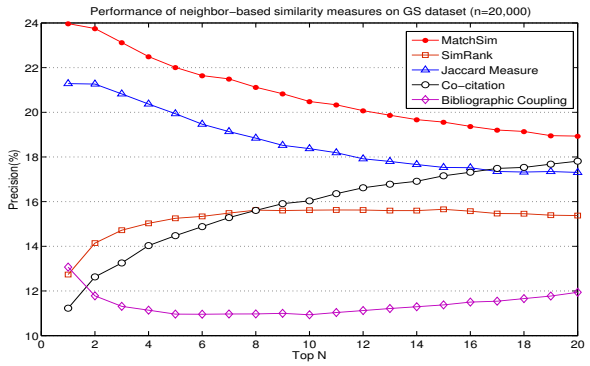

Figure 2: Accuracy curves of the neighbor-based similarity measures on the GS dataset

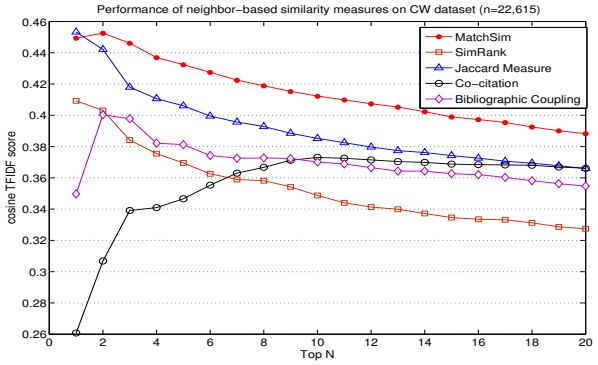

Figure 3: Accuracy curves of the neighbor-based similarity measures on the CW dataset

4.3 Performance of MatchSim

In this part, we compare the accuracy of MatchSim with other neighbor-based similarity measures on the CW and GS datasets. The definitions of the algorithms are given in Section 2. We set $\gamma = 0.8$ in SimRank.

Figures 2 and 3 plot the curves of $\Delta^P(A, N)$ and $\Delta^T(A, N)$ on the GS and CW datasets, respectively. To compare the overall accuracy of the algorithms with that of MatchSim, we also show the $ROA(*, MS, 50)$ values of the algorithms in Table 2. From the results, we can see that MatchSim outperforms all the other algorithms in almost all cases on both of the GS and CW datasets in term of accuracy.

Table 2: $ROA(*, MS, 50)$ of the algorithms

	BC	CC	JM	SR	MS
GS	0.55	0.76	0.89	0.73	1.00
CW	0.89	0.85	0.94	0.85	1.00

5. CONCLUSION AND FUTURE WORK

To effectively measure similarity between web pages, we propose a novel link-based method called *MatchSim*, which recursively defines the similarity between web pages by the average similarity of the *maximum matching* between their respective neighbors. Experiments on two different real-world datasets are conducted to show the effectiveness of the method.

There are a number of avenues for future work. (1) The efficiency of MatchSim needs to be improved to enable it to cope with the entirety of the Web. Possible approaches include neighborhood pruning and approximation algorithms. (2) MatchSim can be easily extended to the "bipartite" version, which is applicable to the recommender systems. In fact, this is our on-going work.

6. ACKNOWLEDGMENT

The work described in this paper is supported by grants from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project No.: CUHK 4128/08E and CUHK 4158/08E). This work is also affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies.

7. REFERENCES

[1] L. Adamic and E. Adar. Friends and neighbors on the Web. *Social Networks*, 25(3):211–230, 2003.

[2] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press/Addison-Wesley, 1999.

[3] H. Chim and X. Deng. A new suffix tree similarity measure for document clustering. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 121–130, NY, USA, 2007.

[4] J. Dean and M. R. Henzinger. Finding related pages in the World Wide Web. *Computer Networks*, 31(11–16):1467–1479, 1999.

[5] L. Friedland and J. Allan. Joke retrieval: recognizing the same joke told differently. In *CIKM '08: Proceeding of the 17th ACM Conference on Information and Knowledge Management*, pages 883–892, New York, NY, USA, 2008. ACM.

[6] A. Gupta and L. Ying. On algorithms for finding maximum matchings in bipartite graphs. In *Technical Report RC 21576 (97320)*. IBM T. J. Watson Research Center, 1999.

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

[8] G. Jeh and J. Widom. SimRank: a measure of structural-context similarity. In *KDD '02: Proceedings of the 8th ACM SIGKDD*, pages 538–543, NY, USA, 2002. ACM Press.

[9] M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14(10–25), 1963.

[10] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[11] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistic Quarterly*, 2:83–97, 1955.

[12] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM '03: Proceedings of the 12th International Conference on Information and Knowledge Management*, pages 556–559. ACM, November 2003.

[13] Z. Lin, I. King, and M. R. Lyu. Pagesim: A novel link-based similarity measure for the World Wide Web. In *WI '06: Proceedings of the 5th International Conference on Web Intelligence*, pages 687–693, Hong Kong, 2006. IEEE Computer Society.

[14] W. Lu, J. Janssen, E. Milios, and N. Japkowicz. Node similarity in networked information spaces. In *CASCON '01: Proceedings of the 2001 conference of the Centre for Advanced Studies on Collaborative research*. IBM Press, 2001.

[15] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

[16] T. Roelleke and J. Wang. Tf-idf uncovered: a study of theories and probabilities. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 435–442, New York, NY, USA, 2008. ACM.

[17] G. Salton. *Automatic Text Processing*. Addison-Wesley, 1989.

[18] H. Small. Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(265–269), 1973.