User-Assisted Similarity Estimation for Searching Related Web Pages

Lin Li, Zhenglu Yang, Kulwadee Somboonviwat, and Masaru Kitsuregawa
Dept. of Info. and Comm. Engineering, University of Tokyo
4-6-1 Komaba, Meguro-Ku, Tokyo 153-8305, Japan
Tokyo 153-8305, Japan
{lilin, yangzl, kulwadee, kitsure}@tkl.iis.u-tokyo.ac.jp

ABSTRACT

To utilize the similarity information hidden in the Web graph, we investigate the problem of adaptively retrieving related Web pages with user assistance. Given a definition of similarities between pages, it is intuitive to estimate that any similarity will propagate from page to page, inducing an implicit topical relatedness between pages. In this paper, we extract connected subgraphs from the whole graph that consists of all pairs of pages whose similarity scores are above a given threshold, and then sort the candidates of related pages by a novel rank measure which is based on the combination distances of a flexible hierarchical clustering. Moreover, due to the subjectivity of similarity values, we dynamically supply the ordering list of related pages according to a parameter adjusted by users. We show our approach effectively handles a set of pages originating from three related categories of Web hierarchies, such as Google Directory. The experiments with three similarity measures demonstrate that using in-link information is favorable while using a combination measure of in-links and out-links lowers the precision of identifying similar pages.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*clustering, search process*; G.2.2 [**Discrete Mathematics**]: Graph Theory—*Graph algorithms*

General Terms

Algorithm, Experiments, Performance

Keywords

graph partitioning, similarity search, clustering

1. INTRODUCTION

Modern search engines store the link structure of a large part of the Web to serve users in two main ways, namely through "keyword" and "related pages" searches [1]. Both broad-topic and short queries will retrieve thousands of "hits" which may prove difficult for users to locate pages from. Therefore, query expansion and query recommendation techniques have been studied to help users formulate better queries [3, 9]. On the other hand, it often happens that a user is already familiar with some Web pages and needs to find more related ones. Thus, the task of identifying similar pages on the Web has also become an important issue.

Recently the link-based similarity search has gained much attention. Co-citation[26] and bibliographic coupling [19] are two of the more fundamental measures used to evaluate the similarities between two scientific papers. When applied in the Web, these bibliometrics measures can be thought of as local in nature because they typically consider only the local link properties between two pages which exist inside a narrow area of the Web graph. Kumar et al. [21] used bipartite subgraphs to recognize the cores of the communities and empirically concluded that a large fraction is in fact topically coherent. These approaches are not good at finding large related subsets of the Web graph due to their localized structures. PageRank [4], HITS [20], SimRank [18], Random Walk with Restart [27] are more global since they work by iteratively transmitting weights through the Web graph. The weights reflect the similarity approximation between pages and, therefore, can search large collections of related and valuable Web pages. However, if the Web page collection comes within several topics, they may only cover the most prevalent topics and leave out the less prevalent ones. In a solution for HITS, one has to compute multiple eigenvectors in order to extract the smaller community that one is interested in. Calado et al. [5] compared the bibliometrics and HITS-based similarity measures, from which the co-citation measure presented the best performance in determining if two pages are related.

Alternatively, similarity metrics such as co-citation and bibliographic coupling can be used along with clustering techniques, thus identifying related pages in a cluster that is a collection of pages which are similar between them and are dissimilar to the pages belonging to other clusters. The fundamental problem in attempting to separate the Web into clusters based upon its link structure and any kind of normative definition of a cluster is that the size of the Web means that any standard approach based upon the whole Web graph would be computationally infeasible. Flake et al. [12] figured out this problem in another way by developing a heuristic-based algorithm that iteratively extends any seed set of pages into a larger interconnected one.

HT'07, September 10–12, 2007, Manchester, United Kingdom.

However, the Web graph is dynamic and massive in nature, which brings about three limitations during computing similarities: 1) links of pages have not yet been crawled; 2) loading the whole Web graph to memory is unavailable for the present; 3) some pages have good content but have not been linked by many authors, which has already been addressed by Dean et al. [10]. They also pointed that their algorithm does not claim to find all relevant pages. Due to the limited information is available, we consider how to estimate relatedness between pages Our idea is that the propagation of localized bibliometrics similarity on a globalized neighborhood graph of similarity that is explained in Section 5.1 will alleviate the problem. For example, using co-citation as a similarity measure, if a page A is with the score of 0.5, and similar to a page B that is similar to a page C with the score of 0.4, it is intuitive for us to infer that A is also similar to C with some probability (e.g., 0.5∗0.4=0.2), though the co-citation between A and C is 0. The score of 0 does NOT mean the two pages are definitely dissimilar. They may only be not popular enough to be linked by many other pages, or their link information may still be un-crawled or unloaded. We think that the similarity scores will propagate from page to page to compensate for these limitations of missing information. One more observation is that if there is a page D similar to C with the score of 0.01, our inference is that A is more similar to C than D. However, because the link information in the Web is noisy, the score between C and D (i.e., 0.01) is too small to be credible.

In addition, we should not overlook the subjectivity of similarity. SimRank [18] and Random Walk with Restart [27], as link-based methods, computed the static similarity scores which are sensitive to specific choices made by the authors of Web pages. Different users, however, have different standards to measure relatedness. For example, given a page about knowledge discovery, some people point to database technologies for useful information, while others prefer the topic of Artificial Intelligence. Sometimes they cannot precisely judge whether a page is related or not by themselves. Therefore, page-to-page similarity is not a fixed value, especially for cross-topic pages. If the ordering list of related pages may be able to be changed under some constraints, then the users will be supplied with more candidates of related pages.

Based on the above facts, we propose a user-assisted estimation approach for similarity search in this paper. Our approach uses three bibliometrics measures, and has three major steps: computing similarity scores of pairs of pages, extracting connected subgraphs from the neighborhood graph of similarity, and applying sequential, agglomerative, hierarchical, nonoverlaping (SAHN) clustering on these subgraphs for ranking related pages. The flexibility in the SAHN clustering makes it possible for users to take part in the process of searching related pages. The main contributions of this paper are as follows.

(1) We put forward the problem of similarity estimation for searching related pages.

(2) A novel rank mechanism is proposed to put in order the related pages based on the monotonicity of the successive combination distances in the SAHN clustering.

(3) Through our approach, users can become active participants during the search instead of just being passive acceptors of results.

The rest of this paper begins with a review of the related work. We then describe three bibliometrics measures in Section 3. In Section 4, concerns regarding the SAHN clustering method are addressed. The details of our approach are given in Section 5. Finally, we report on the experimental results and follow them up with our conclusions and directions towards future works.

2. RELATED WORK

A considerable amount of research has focused on examining collections of hyper-linked pages and structures. These works have been very cross-disciplinary, touching the hypermedia, the World Wide Web, Sociology, bibliometrics, and even culture and the communication fields. We summarize a small part of these works to put our own study within the proper context.

2.1 Content-Based Analysis

Chakrabarti et al. [6] developed text-based methods for automatically classifying hypertext into a given topic hierarchy. Haveliwala et al. [15] presented a technique for automatically evaluating strategies for answering Related Pages queries by using Web hierarchies, such as open directory, in place of user feedback. Zheng et al. [28] built an offline clustering tool to allow the topical clustering of the entire Web. Clustering Web pages by content, however, may require storage and manipulation of an overwhelming amount of data, and is not generally scalable to an online clustering of the whole Web. Moreover, these text-based methods are not applicable, at least in principle, in settings including: non-text pages like multimedia (image) files, Usenet archives and documents in non-HTML file formats such as PDF and DOC documents, pages with limited access like sites that require registration, and dynamic pages which are returned in response to a submitted query or accessed only through a form. As pages are modified, the text-based methods are more susceptible to placing a URL in different clusters than link-based methods.

2.2 Link-Based Analysis

Kleinberg's HITS [20], a topic distillation algorithm, is applied in [7, 10, 13] to identify groups of similar pages. Dean et al. [10] described the Companion algorithm to find related pages by building a weighted graph around the starting page and then running a modified version of HITS. Chirita et al. [7] help users to find hubs related to a given initial set of pages on the original Web graph. On the other hand, random walk based methods are also alternatives. SimRank [18] analyzes similarity between graph nodes by constructing a node-pairs graph with SimRank similarity scores. The scores are based on the theory that two objects are similar when they are referred by similar objects and reflect node similarities at a pair level. Sun et al. [27] computed the relevance score for each node using random walk with restarts and graph partitioning to identify similar nodes and anomaly nodes.

In a different way, He et al. [16] explored textual information, hyperlink structure, and co-citation relations and applied the normalized-cut graph partitioning to the task of clustering. If time and space complexity issues were irrelevant, then the approach proposed in [16] could identify tightly coupled communities. Flake et al. [11] calculated the Web community in the maximum-flow and minimumcut framework based on the Web graph that is locally stored to yield results fast. Ino et al. [17] proposed Web page partitioning with the aid of equivalence relation to overcome boundary ambiguity of a Web community defined by [11].

2.3 Bibliometrics

In bibliometrics, a range of metrics have been widely used to assess the similarities of documents. Co-citations [26] were proposed under the assumption that two articles that are frequently cited together are likely to have something important in common. Bibliographic coupling [19] is a complementary measure. Two papers share one unit of bibliographic coupling if both cite a same paper. Although the purpose of most of these techniques is to provide the primary mechanism to map and traverse the intellectual structure of scientific information space, they have also been explicitly cited as motivation for computer link based methods and have been applied to the problem of mining Web pages [24]. Pitkow et al. [25] clustered sets of hypertext Web pages, transferring the concept of scientific publication citations to hypertext links on the Web. However, applying these methods to systems such as the Web with at least 11.5 billion pages [14], would obviously be challenging, if not at all daunting.

3. BIBLIOMETRICS MEASURE

To determine how related two Web pages are, we used three different similarity measures derived from their link information: co-citation, bibliographic coupling, and Amsler measures [5] that were all introduced in bibliometrics as measures of how related two scientific papers. In this paper, we evaluate how they perform in the Web graph. Hyperlinks are a generalized form of citation, acting diverse roles such as advertising, in-site navigation, providing access to pages that are the results of database queries, and so on. Therefore, links are a less reliable source of evidence, when used as an indicator of similarity between Web pages. However, we assume that it is still promising to regard Web links as analogous to conventional citations.

3.1 Co-citation

A Web page author will insert links to pages related to his own page. In this case, by treating links as citations, we say that two pages are co-cited if a third page has links to both of them. To further make it clear, let p be a Web page and let $I(p)$ be the set of pages that link to p, called the in-links of p. The co-citation similarity between two pages p_1 and p_2 is defined as:

$$cit(p_1, p_2) = \frac{|I(p_1) \cap I(p_2)|}{|I(p_1) \cup I(p_2)|} \,. \tag{1}$$

Equation 1 shows that the more in-links of p_1 and p_2 have in common, the more similar they are. This value is normalized by the total set of in-links, so that the co-citation similarity varies between 0 and 1. If both $I(p_1)$ and $I(p_2)$ are empty, we define the co-citation similarity as zero.

3.2 Bibliographic Coupling

Two authors of Web pages on the same topic tend to insert links to the same pages. More formally, let p be a Web page. We define $O(p)$ as the set of pages that p links to, also called out-links of p. Bibliographic coupling between two pages p_1 and p_2 is defined as:

$$bib(p_1, p_2) = \frac{|O(p_1) \cap O(p_2)|}{|O(p_1) \cup O(p_2)|} \,. \tag{2}$$

According to Equation 2, the more out-links in common page p_1 has with page p_2, the more related they are. This value is normalized by the total set of out-links, ranging from 0 to 1. If both $O(p_1)$ and $O(p_2)$ are empty, we set the bibliographic coupling similarity as zero.

3.3 Amsler

Calado et al. [5] introduced a measure of similarity called Amsler that combines both co-citation and bibliographic coupling in an attempt to take the most advantage of the link information available between pages. On the principle of Amsler, two pages p_1 and p_2 are related if 1) p_1 and p_2 are linked by a third page, 2) p_1 and p_2 link the same page, or 3) p_1 links a third page p_3 that links p_2. p is denoted as a Web page, let $I(p)$ be the set of in-links of p, and let $O(p)$ be the set of out-links of p. The Amsler similarity between two pages p_1 and p_2 is defined as:

$$ams(p_1, p_2) = \frac{|(I(p_1) \cup O(p_1)) \cap (I(p_2) \cup O(p_2))|}{|(I(p_1) \cup O(p_1)) \cup (I(p_2) \cup O(p_2))|} \,. \tag{3}$$

Equation 3 shows that the more links (either in-links or out-links) p_1 and p_2 have in common, the more they are related. The measure is normalized by the total number of in-links and out-links. If neither p_1 nor p_2 have any in-links or out-links, the similarity is assigned to zero.

3.4 Our Measure

We find that the above three measures do not take into account the direct links between two pages. For example, if an in-link of a page p_1 is from a page p_2, the common in-links between p_1 and p_2 do not include p_2. But p_2 definitely contributes to the similarity score. The same problem occurs in bibliographic coupling and Amsler measures. Therefore, we modify them by adding the effect of direct links between pages. We give a generalization for the above three measures, listed as follows:

$$sim(p_1, p_2) = \frac{|C(p_1) \cap C(p_2)| + direct(p_1, p_2)}{|(C(p_1) \cup C(p_2) \cup p_1 \cup p_2|} \,, \tag{4}$$

where $C(p)$ represents $I(p)$ in Equation 1, $O(p)$ in Equation 2, or $I(p) \cup O(p)$ in Equation 3. In the rest of our paper, we still call them "co-citation", "bibliographic coupling", and "Amsler" for simplicity. In Equation 4, $direct(p_1,p_2)$ is defined as

$$direct(p_1, p_2) = \begin{cases} 0 & \text{if no direct links between } p_1 \text{ and } p_2 \\ 1 & \text{if } p_1 \text{ links } p_2 \text{ OR } p_2 \text{ links } p_1 \\ 2 & \text{if } p_1 \text{ links } p_2 \text{ AND } p_2 \text{ links } p_1 \end{cases} \,.$$

4. SAHN CLUSTERING

Before producing an ordering list of related pages, we cluster and rank the candidates by the SAHN clustering which outputs a hierarchy, more informative than the unstructured set clusters in flat clustering algorithms like k-means and EM. The SAHN clustering treats each object as a singleton group at the beginning, and then merges pairs of groups iteratively until all groups have been merged into a single group structured as a hierarchy that contains all objects. The fundamental assumption is that the best possible merger is found at each step. Furthermore, it may have an interesting property that suggests that distance measures associated with successive merge operations could be monotonic; if d_1, d_2, $\cdots$, d_k (the definition will be expressed soon) are successive combination distances of the SAHN clustering, then $d_1 \le d_2 \le \cdots \le d_k$ must hold. Urged by the monotonic property, we think that pages which have the shortest distances will be merged first. At each remaining step in the hierarchy, the next closest pair of pages (or groups) should be merged. The sequence of merge operations scores the relevance of two pages and produces an ordering of related pages for a specific page.

Lance et al. [22, 23] derived a flexible method of the SAHN clustering by the constraint $(0 < \alpha \le 1)$, defined as

$$d_{hk} = \alpha d_{hi} + \alpha d_{hj} + (1 - 2 * \alpha) d_{ij} \, , \tag{5}$$

where (h), (i), and (j) are three groups, containing n_h, n_i, and n_j elements, respectively, with inter-group distances already defined as d_{hi}, d_{hj}, and d_{ij}. They further assume that the smallest of all distances still to be considered d_{ij}, so that (i) and (j) fuse to form a new group (k), with n_k $(=n_i+n_j)$ elements. The constraint suggests a set of monotonic methods such that as α increases from 0 to 1, the hierarchy changes from an almost completely "chained" system to one with increasingly intense clustering. A given set of pages may now, by varying the parameter α, be made to appear as sharply clustered as a user may desire. Thus, we can adaptively rank the related pages by combination distances that vary with α as well.

5. OUR APPROACH

After the collection and preprocessing of Web link information that are presented in Section 6, our approach consists of the following three steps.

5.1 Step 1: Computation of Similarity Scores

Given that the in-links and out-links data are stored in search engines, should we have to compute similarities for all pairs of pages that will count a quadratic number of values? Notice that we are interested in pairs of pages whose similarity is above a specified threshold, a high quality collection. The latest work [2] addressed this scalability issue without relying on approximation methods or extensive parameter tuning. Inspired by their work, we describe a monotone minimum size constraints on the number of in-links (out-links) of candidate pages before computing the similarity scores. Given a threshold δ, the following inequalities are established:

$$\begin{aligned} \frac{|C(p_i) \cap C(p_j)| + d(p_i, p_j)}{|(C(p_i) \cup C(p_j) \cup p_i \cup p_j|} \quad &< \quad \frac{|C(p_i) \cap C(p_j)| + 2}{|(C(p_i) \cup C(p_j)|} \\ &\le \quad \frac{|C(p_i)| + 2}{|C(p_j)|} \le \delta \, . \end{aligned}$$

The above equation tells us if the $(|C(p_i)|+2)/|C(p_j)| \le \delta$, their similarity score does not need to be stored to reduce the amount of candidate pairs. Furthermore, we sort pages in the decreasing order of $C(p)$ to save on computation time. This preprocess means if $(|C(p_i)| + 2)/|C(p_j)| \le \delta$ is met, the pages p_i to p_n can be skipped without doing similarity computation with p_j (the total number of all pages is n), thus accelerating our approach.

Finally, we eliminate the similarity scores between pages which are larger than 0.95 because of duplicated pages (e.g., mirror sites, different aliases for the same page). The remain pages compose the whole neighborhood graph of similarity, denoted as G, which is quietly different from the original hyper-link graph. Its nodes correspond to pages, but edges are weighted according to a distance score measured by $dis(p_1, p_2) = 1 - sim(p_1, p_2)$. An simple example is shown in Figure 2.

Applying the SAHN clustering to cluster pages in the whole graph G is a direct way, however, its complexity is at least quadratic in the number of pages because of the distance matrix of all pairs of pages. Moreover, one observation from our experiments indicates that the neighborhood graph of similarity is an unconnected graph which may be subdivided into connected subgraphs. These encourage our approach to partition the original neighborhood graph of similarity to connected subgraphs, and then perform the SAHN clustering only on the subgraphs.

5.2 Step 2: Extraction of Neighborhood Subgraphs of Similarity

To obtain connected subgraphs from G, each vertex in G is first in its own set on the basic initialization of the disjoint-sets structure. We then calculated connected subgraphs based on the edges in G, embedding the results in the disjoint-sets data structure. The disjoint-sets structure is updated when an edge (p_1, p_2), whose similarity score is above the given thresholds, is added into the graph. Last, we extract all connected components, also called neighborhood subgraphs of similarity here. Refer to [8] for the disjoint-sets structure in detail.

5.3 Step 3: Ranking Related Pages with User Assistance

Given a page, we run the SAHN clustering on the neighborhood subgraph containing the input page. All the pages in the subgraph are candidates for related pages. Then, we rank them to form an ordering list. The pseudo code of this step is described in Table 1. In each iteration, the two most similar clusters are merged (Line 11 ∼ Line 13) and the rows and columns of the merger cluster i in D are updated (Line 14 ∼ Line 18). Ties in the SAHN clustering are broken randomly. The process of the SAHN clustering is stored as an N by 2 matrix in M, where N is the number of candidates. Row i of M describes the merging of clusters at the step i of the clustering. If a number j in the row is negative, then the single page $|j|$ is merged at this stage. If j is positive, the merger is with the cluster formed at stage j of the algorithm. I indicates which clusters are still available to be merged. H stores the combination distances between merging clusters at the successive stages.

We last specify how to output the ordering list of related Web pages. Given that the value of α in Equation 5 is decided by a user, the SAHN clustering outputs a hierarchical structure where an input page (ip) and a candidate page (cp) will come together at a combination distance ($d_{ip,cp}$) (Line 27 ∼ Line 28), and at a distance (d_{ip}), the input page is merged with a group (page) in the first time (Line 21 ∼ Line 22), and the first merging for the candidate page is at a distance (d_{cp}) (Line 25 ∼ Line 26). Then, the distance score between the two pages is estimated to $|d_{ip} - d_{ip,cp}| + |d_{cp} - d_{ip,cp}|$ which ranks each candidate in the neighborhood subgraph (Line 29). The clustering structure is illustrated in Figure 3, where "Height", the label of the or-

Table 1: SAHN clustering and Rank Mechanism

```
Input: N candidates in the neighborhood subgraph of
       similarity, $\alpha$ in the Equation 5 chosen by a user
Output: an ordering of related pages for the input page
Distance matrix
for $k$=1 to $N$
    for $l$=1 to $N$
        $D[k][l] = dis(p_k, p_l)$
Initialization
$H[N]$ (for combination distances)
$M[N][2]$ (for collecting merge sequence)
$O[N]$ (for the ordering list)
for $k$=1 to $N$
    $I[k] = 1$ (keeps track of active cluster)
Compute clustering
for $k$=1 to $N$
Begin Loop
    $(i,j) = argmin_{(i,j)l \neq m, I[i]=I[j]=1} D[i][j]$
    $M$.append($< i, j >$)
    $H$.append($(i, j)$)
    for $h$=1 to $N$
    Begin Loop
        $d_{h(i,j)} = \alpha * D[h][i] + \alpha * D[h][j] + (1-2*\alpha) * D[i][j]$
        $D[i][h] = D[h][i] = d_{h(i,j)}$
    End Loop
    $I[j] = 0$ (deactivate cluster)
End Loop
Rank Mechanism
if $M[i][j] == -ip$
    $d[ip] = H[i]$ (the first merge for $ip$)
for $cp$=1 to $N$ ($cp \neq ip$)
Being Loop
    if $M[i][j] == -cp$
        $d[cp] = H[i]$ (the first merge for $cp$)
    if row $i$ of $M[N][2]$ are clusters that include $cp$
                    and $ip$ respectively
        $d[(ip, cp)] = H[i]$
    $O[cp] = |d[ip] - d[(ip, cp)]| + |d[cp] - d[(ip, cp)]|$
End Loop
Sort $O[N]$ in increasing order
```

dinate axis, means the combination distance at each merging operation. In this way the similarity propagation is implicitly and adaptively realized by the rank mechanism. Give a simple example shown in Figure 1 and 2. If our approach computes $sim(a,b) = 0.5$, $sim(b,c) = 0$, $sim(a,c) = 0.4$, $sim(d,a) = 0.3$, and $sim(d,b) = 0.2$ in the first step, the page a and the page b are merged in the first combination under $\alpha = 0.02$. The page c will be merged with $a(b)$ in the second combination, and then the page d will be merged in the third combination. The ordering of related pages of the page b is a, c, and d, despite of $sim(b,c) = 0$ and $sim(b,d) = 0.2$. And if the user chooses $\alpha = 0.5$, the ordering list becomes a, d, and c.

6. EXPERIMENTS

6.1 Experimental Setup

The notion of "similarity" is subjective and difficult to measure. To make experimental results more objective, we used the Google Directory as a form of "ground truth" for relatedness to evaluate three similarity measures discussed in Section 3.

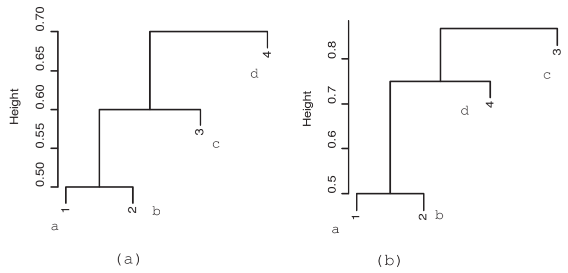

Figure 1: (a) $\alpha = 0.02$ and (b) $\alpha = 0.5$ in Equation 5

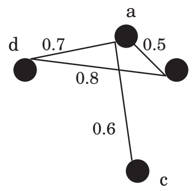

Figure 2: An Example of similarity propagation. The edges are weighted by $1 - sim$. (e.g., 0.7=1-0.3)

6.1.1 Datasets

To test our approach of identifying related pages, we first extracted 48[1] web pages related to three categories: Computers − > Software − > Databases − > Data Mining, Reference − > Knowledge Management − > Knowledge Discovery, and Computers − > Artificial Intelligence − > Machine Learning, as our core set on which we focused the following analyses. The three categories were chosen to pose a more difficult clustering problem: there is a large vocabulary overlap and cross-linkage between the categories. For each page in the core set, we utilized the Google API[2] *link:* query to obtain its in-links with in-degree restricted to 50 (how we determined 50 will be described in Section 6.3), and downloaded its HTML to fetch all out-links. The pages in the core set and their in-links and out-links are a level one expansion, which we called the Expansion1 dataset. Then we expanded the Expansion1 dataset by including the top 50 in-links and all out-links of each page in the Expansion1 dataset, which resulted in the Expansion2 dataset. In some cases there was an insufficient level of in-links with a page to provide adequate result. If there were not at least 10 in-links of a page, we then used the page's URL with one path element removed. If the resulting URL was invalid, we continued to chop elements until we were left with just a hostname, or we found a valid URL [10].

6.1.2 Data Preprocessing

We counted the number of out-links for each page. Once the number was larger than 1000, we eliminated the page

[1] There are 53 pages from the three categories, but 5 of them are unaccessible now.

[2] http://code.google.com/apis/

because it was very likely to be a portal and we also wanted to keep the overall dataset within a reasonable size. Furthermore, as we know, a large fraction of web pages point to popular sites like Google or Yahoo!, even though the topic of these pages may be completely unrelated. Dean et al. [10] used a stoplist STOP of URLs to ignore all the URLs on the stoplist when forming the vicinity graph. Keeping an updated stoplist is troublesome in the dynamic web. Besides a stoplist, our measure mechanism has cushioned us from the effects of popular sites. Bibliometrics measure is based on the frequency of common in-links or out-links, which reduces the likelihood of the computation being dominated by a single page (e.g., a popular site). In other words, although there are many in-links for a popular site, the frequency of common in-links with another page is relatively low enough to be omitted by a given threshold. After performing data preprocessing, we ran our three-step algorithm as described in Section 5.

6.2 Statistics about Datasets

Table 2 and Table 3 summarize the statistics about the distinct in-link and out-links of our two datasets. In the tables, “E1(50,1000)” means that the maximal numbers of in-links and out-links of a page are 50 and 1000, respectively in the Expansion1 dataset. “IL”, “IPP”, “OL”, “OPP”, and “ALL” are short for “in-links”, “in-links per page”, “out-links”, “out-links per page”, and “all pages”, respectively in each data set.

Table 2: In-links and Out-links of Expansion1

Data set	IL	IPP	OL	OPP	ALL
E1(10,1000)	538	12.2	1490	33.9	2072
E1(20,1000)	620	14.1	1490	33.9	2154
E1(50,1000)	1001	22.8	1490	33.9	2535

E2$_c$ in Table 3, represents the chopped Expansion2 dataset where each page URL was chopped to its hostname to eliminate the navigational links. This operation unified pages that are on the same host. Thus, the number of out-links per page was greatly reduced (e.g., dropping from 24.3 to 5.5 in the E2(50,50) data set). On the other hand, the number of in-links per page were kept relatively stable (e.g., dropping from 5.7 to 3.7). It also tells us that allowing pages with the same hostname to remain separate can greatly mar the results. Moreover, our experiments on the Expansion2 dataset observed that most of the top related pages are navigational links. In terms of quantity the out-link exceeds the in-link, especially for the Expansion2 dataset. The next experiments will check whether the out-links superior in number are more informative than the in-links.

6.3 Results and Discussion

6.3.1 Results of Step 1 of our approach

We explain the notations used in Table 4 and Table 5 that are statistics about neighborhood graph of similarity. “Cit(10)” means that the number of in-links used to compute co-citations is 10 at the most. “Bib(1000)” represents that the number of out-links used to compute bibliographic coupling is 1000 at the most. “Ams(10,1000)” means that the numbers of in-links and out-links used to compute Amsler measure are 10 and 1000 at the most, respectively. The rest of the measures in the “Measure” column are explained in the same way. If the number pages is N, the total number of pairs of pages will reach $(N*N-N)/2$ that divides the number of pairs of pages in the neighborhood graph of similarity. The result of the division is denoted as the “Percentage”. “Average Similarity” is computed by using $(N*N-N)/2$ to divide the the sum of the similarity scores.

Table 3: In-links and Out-links of Expansion2

Data Set	IL	IPP	OL	OPP	ALL
E2(50,50)	16326	5.7	69,025	24.3	88,194
E2(50,100)	16326	5.7	78,431	27.6	97,600
E2(50,500)	16326	5.7	90,579	31.8	109,748
E2(50,1000)	16326	5.7	92,967	32.7	112,136
E2$_c$(50,50)	8845	3.7	13,172	5.5	24,405
E2$_c$(50,100)	8845	3.7	18,441	7.7	29,674
E2$_c$(50,500)	8845	3.7	27,409	11.5	38,642
E2$_c$(50,1000)	8845	3.7	30,017	12.6	41,250

N is 48, in Table 4. From Table 2, the out-links per page (i.e., 33.9) is much more than the in-links per page (the maximum is 22.8). However, the size of the neighborhood graph similarity under the bibliographic coupling measure is terribly small, only 2.4%, and not to be compared with the size of the graph under co-citation or Amsler (the maximal size is 19.9%). This observation exists in Table 5 as well, indicating that the out-link information is more sparse and less informative than the in-link information. It motivates us to investigate whether small quantities of similarities between pages are enough to properly retrieve the pages in the core set. We present the experiment results in the next part.

Table 4: Neighborhood Graph of Similarity of E1

Measure	Pairs of Pages	Percentage	Average Similarity
Cit(10)	107	11.3%	0.0632
Cit(20)	126	13.3%	0.0547
Cit(50)	175	18.4%	0.0381
Bib(1000)	23	2.4%	0.0301
Ams(10,1000)	121	12.8%	0.0314
Ams(20,1000)	142	15.0%	0.0298
Ams(50,1000)	188	19.9%	0.0236

Furthermore, we further processed Expansion1 dataset by omitting pairs of pages whose similarity scores are larger than 0.95 and smaller than 0.004. The lowerbound (i.e., 0.004) is selected at 10% of the average similarity score because we trust that the larger values are more reliable in determining the relatedness. The upperbound (i.e., 0.95) is used to delete the duplicated pages. Processed in the same way as the Expansion1 dataset, the upperbound and the lowerbound are 0.95 and 0.002 respectively for the chopped Expansion2 dataset. Here N is 2843 for the Expansion2 data set and 2388 for the chopped one in Table 5.

From the above tables, the data were worked with lowerbound thresholds like 0.002 or 0.004. Increasing the values of the thresholds can trade off recall against precision. The absolute values of similarity are small, but our algorithm is interested in the ordering decided by the values, not the actual values themselves.

Table 5: Neighborhood Graph of Similarity of E2$_c$

Measure	Pairs of Pages	Percentage	Average Similarity
Cit(50)	10331	0.36%	0.0423
Bib(50)	16385	0.57%	0.0325
Bib(100)	20318	0.71%	0.0267
Bib(500)	23761	0.83%	0.0230
Bib(1000)	24083	0.84%	0.0227
Ams(50,50)	36351	1.28%	0.0228
Ams(50,100)	38385	1.35%	0.0216
Ams(50,500)	40998	1.44%	0.0201
Ams(50,1000)	41281	1.45%	0.0199

6.3.2 Results of the Step 2 of Our Approach

We report the results of the following data sets: E1(20,1000), E1(50,1000), E2$_c$(50,50), and E2$_c$(50,100) after completing Step 2 of our approach in Table 6. The core set consists of pages from three categories of the Google Directory. We regard pages in the same category as related pages. Our goal is to retrieve as many pages in the core set as possible. Taking the three categories as a whole, we evaluate the recall over all categories performed on the above four data sets, defined as:

$$Recall = \frac{|\{\text{pages in the core set}\} \cap \{\text{retrieved pages}\}|}{|\{\text{pages in the core set}\}|},$$

which considers the proportion of pages in the core set are retrieved out of all pages in the core set by the three different measures.

Table 6: Overall Recalls of E1 and E2$_c$

	E1(20,1000)			E1(50,1000)		
Measure	Cit	Bib	Ams	Cit	Bib	Ams
Recall (%)	72.73	38.64	84.09	79.55	38.64	88.64
	E2$_c$(50,50)			E2$_c$(50,100)		
Measure	Cit	Bib	Ams	Cit	Bib	Ams
Recall (%)	85.42	79.17	93.75	85.42	79.17	93.75

For the Expansion1 dataset, we keep the number of outlinks of each page unchanged, and increase the number of in-links during computing the co-citation and the Amsler similarity scores (the Bibliographic coupling measure only needs the out-link information, so it is not influenced by the augmentation of the in-link information). The results show that the recall has been improved from 72.73% to 79.55% under the co-citation measure, and from 84.09% to 88.64% under the Amsler measure, and that the in-link based measure performed much better than the out-link based one (the recall of bibliographic coupling is only 38.64%). As a result, we choose 50 as the in-degree restrict to obtain the Expansion2 dataset.

The number of pages in the chopped Expansion2 dataset is much larger than that in the Expansion1 dataset, which supplies more chances to bridge the pages in the core set. Therefore, the chopped Expansion2 dataset performed better than the Expansion1 dataset on all three measures. We especially saw a big improvement of the bibliographic coupling measure, from 38.64% to 79.17%. In addition, we kept the number of in-links of each page unchanged, and increase dthe number of out-links from 50 to 100 during the computation of the Bibliographic coupling and the Amsler scores, but the recall did not go up. Even after raising the number to 1000, the improvement had been very limited in our experiments.

In summary, the results in Table 6 tell us that the in-link information and the size of candidates of related pages have much more effect on recall than the out-link information.

6.3.3 Results of the Step 3 of Our Approach

Using the Expansion1 dataset as an example and choosing $\alpha = 0.5$ in Equation 5, we explain the clustering results of Step 3 in Table 7. Our goal is not only to retrieve relevant pages, to also cluster pages from the same category. Therefore, recall alone is not enough but we need to measure the number of correctly clustered pages in their corresponding categories. Recall and precision scores were necessary to be computed for individual categories. Here, precision is defined as the proportion of correctly clustered pages in the set of all pages clustered to a category. Recall is defined as the proportion of correctly clustered pages out of all the pages having the category. They are given as follows:

$$Precision_i = \frac{|\{\text{pages in category i}\} \cap \{\text{retrieved pages in cluster i}\}|}{|\{\text{retrieved pages in cluster i}\}|},$$

$$Recall_i = \frac{|\{\text{pages in category i}\} \cap \{\text{retrieved pages in cluster i}\}|}{|\{\text{pages in category i}\}|},$$

where i are $c1$, $c2$, or $c3$. The notations, c1, c2, and c3 are three categories, namely Data Mining, Knowledge Discovery, and Machine Learning, respectively.

Table 7: Precision and Recall of Each Category of E1

E1(20,1000)	Precision (%)			recall (%)		
	c1	c2	c3	c1	c2	c3
Cit	81.82	62.50	82.35	50.00	33.33	68.75
Bib	80.00	0.00	66.67	25.00	0.00	33.33
Ams	66.67	57.14	78.57	56.25	41.67	81.25
E1(50,1000)	Precision (%)			recall (%)		
	c1	c2	c3	c1	c2	c3
Cit	86.67	66.67	93.33	68.75	41.67	75.00
Bib	80.00	33.33	66.67	25.00	8.33	33.33
Ams	84.62	62.50	85.71	81.25	50.00	87.50

Compared with the E1(20,1000) data set, the E1(50,1000) data set holds better performance on both precision and recall for almost all three categories (the bibliographic coupling measure is an exception because of the same reason addressed in the above part). The co-citation measure outperformed the Amsler measure on precision, though the recalls under the co-citation measure are inferior to that under the Amsler measure. This observation showed that mixing the in-link information and the out-link information generally hurts precision while helping recall. The bibliographic coupling measure outperformed the Amsler measure on precision for the E1(20,1000) data set (we explain the reason that $c2$ is an exception in next paragraph). However, when adding more in-links to the Amsler measure in the E1(50,1000)data set, the bibliographic coupling measure failed in precision. Moreover, although we included much more out-links than in-links in the dataset, the values of the recall under the bibliographic coupling measure are terribly low. This fact further demonstrates that the out-link information is sparse and noisy and that the in-link information is more reliable.

The pages from the Knowledge Discovery category present unsatisfactory precision and recall. It is a complex topic and has cross-links between the Data Mining and Machine Learning categories. As we know, the knowledge discovery process takes the raw results from data mining and transforms them into useful and understandable information uncovered through the use of AI techniques such as machining learning, and so on. These closely related categories pose a more difficult clustering problem. We are now combining content-based measures to solve it.

The neighborhood subgraph of similarity extracted from the E1(50,1000) data set under the co-citation measure is shown in Figure 4. We also easily found many related pages are not adjacent because their similarity scores are zero, such as page x and page z. But our SAHN clustering can identify them through propagating similarity via page y. Moreover, the structures of the cluster changed in accordance with the variable α, shown in Figure 3. As α increases from 0.06 to 0.80, the hierarchy changes from an almost completely "chained" system to one with increasingly intense clustering. Thus, it is obvious that the ordering in the list of related pages partly changed as well. We represented this changing in Table 8 and compared the ordering lists produced by co-citation counts with those computed by our rank mechanism (here our subgraph is based on the Co-citation measure, but we rank the candidates by the combination distances). The simple co-citation counts can only sort pages with similarity scores of non-zero. But in our context, all the pages in the same category are related. The similarity estimation based on our rank mechanism adaptively recommends more related pages some of which are with similarity scores of zero.

6.3.4 Preliminary Results of Cross-Topic Page

The pages from the Data Mining (c1) and Machine Learning (c3) categories form into densely connected subgraphs respectively while the pages from the Knowledge Discovery (c2) category are built like a bridge between the other two categories, circled by the broken line in Figure 4 (we omitted some nodes to see it clearly). If coarsely regarding them as one cluster, the concept of related pages is generalized that pages are related if they are related to related topics, not the same topic. This generalization contributes to finding cross-topic pages.

We now describe a new idea to estimate the likelihood that a page is cross-topic. Based on the ordering list of related pages for a page A, we can easily obtain the similarity scores for pairs of pages in the final ordering list through our rank mechanism. Among these scores, if there exists a low similarity score between a page B and a page C, it means that the two pages, B and C, with a high probability are related to different topics. Thus the page A is likely a cross-topic page. Experimenting on pages from the Knowledge Discovery category, we found that some estimated scores from the ordering lists are nearly zero. We are planning to do further experiments on this idea. Our purpose of this part is not to present the best algorithm for identifying cross-topic pages, but to show the potential of using estimated similarity scores in a new way without expensive analysis overhead.

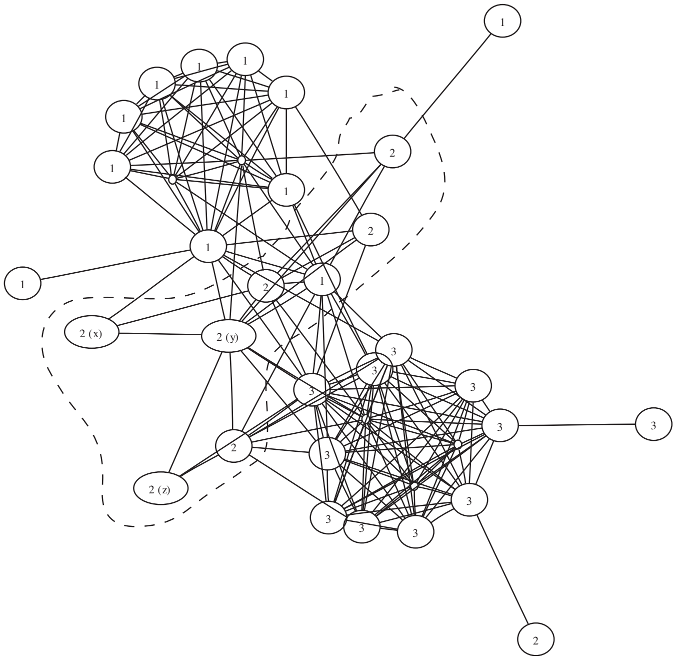

Figure 4: 1: Data Mining; 2: Knowledge Discovery; 3: Machine Learning.

7. CONCLUSION

In this paper, we put forward the problem of the similarity estimation for searching related pages. To solve this problem, our proposed approach efficiently computed the similarity graph above a specified threshold and used the combination distances of the SAHN clustering method to adaptively rank the related pages according to a parameter adjusted by users. Our experimental results show that the co-citation measure, an in-link based method, generally outperformed the other two measures in precision. Exploring larger out-links on the web can be futile and dangerous. The results may be offset by incorrect information collected in the search. In addition, the flexible strategy of the SAHN clustering makes users pick up different related pages in terms of their requirements. Finally, we described the potential of similarity scores used in a novel way to identify cross-topic pages. We are now crawling much larger amount of web page information to further test our approach. In the future, scalable evaluation strategies should be studied, especially when tuning parameters is needed.

8. REFERENCES

[1] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan. Searching the web. *ACM Trans. Internet Techn.*, 1(1):2–43, 2001.

[2] R. J. Bayardo, Y. Ma, and R. Srikant. Scaling up all pairs similarity search. In *Proc. of the 16th International Conference on World Wide Web (WWW'07)*, pages 131–140, 2007.

[3] D. Beeferman and A. L. Berger. Agglomerative clustering of a search engine query log. In *Proc. of the Sixth ACM SIGKDD International Conference on Knowledge discovery and data mining (KDD'00)*, pages 407–416, Boston, MA, USA, 2000.

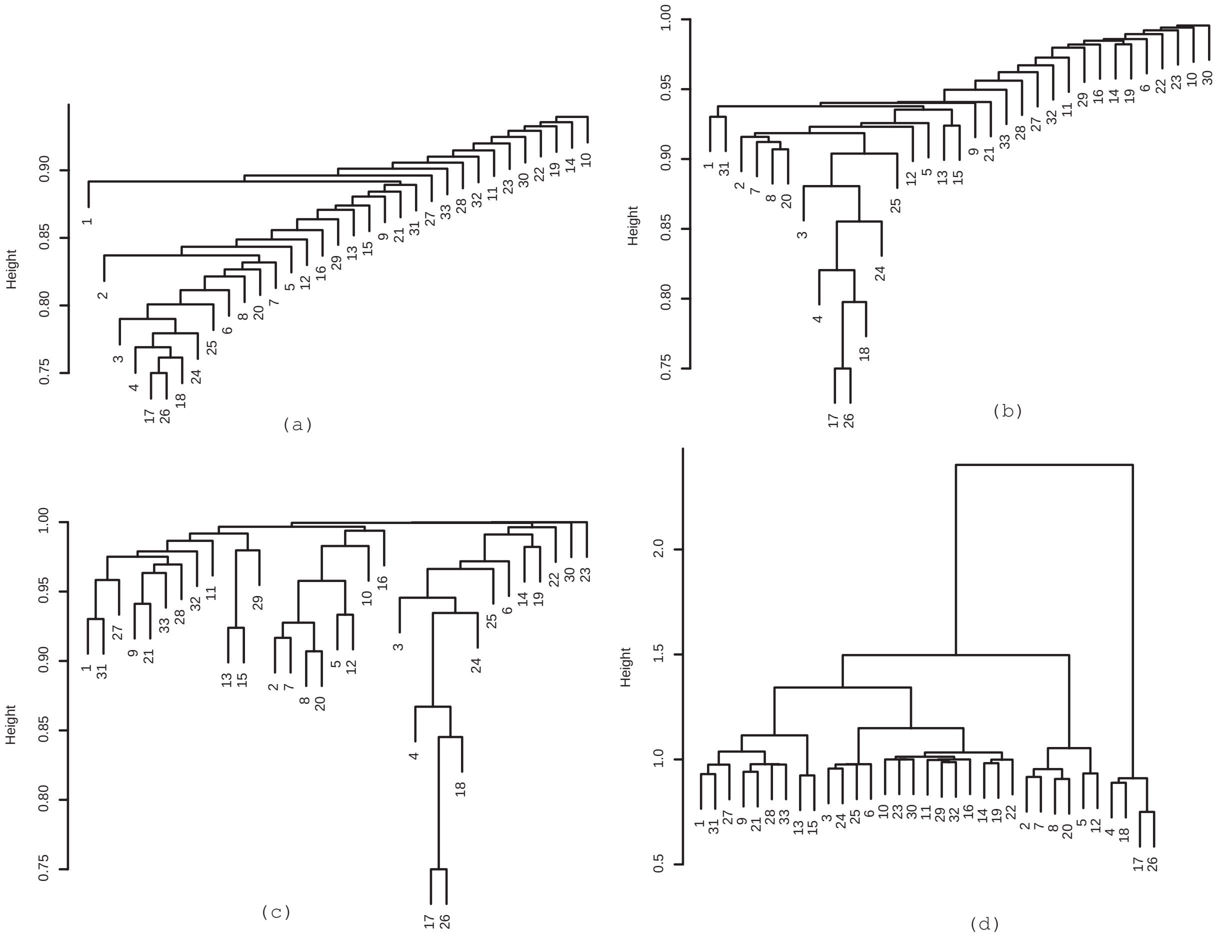

Figure 3: (a) $\alpha = 0.06$ (b) $\alpha = 0.25$ (c) $\alpha = 0.50$ (d) $\alpha = 0.80$. "Height" is the combination distance at each merging operation.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7):107–117, 1998.

[5] P. Calado, M. Cristo, M. A. Gonçalves, E. S. de Moura, B. A. Ribeiro-Neto, and N. Ziviani. Link-based similarity measures for the classification of web documents. *JASIST*, 57(2):208–221, 2006.

[6] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In *Proc. of ACM SIGMOD International Conference on Management of Data (SIGMOD'98)*, pages 307–318, Seattle, Washington, USA, 1998.

[7] P.-A. Chirita, D. Olmedilla, and W. Nejdl. Finding related pages using the link structure of the www. In *Proc. of IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)*, pages 632–635, Beijing, China, 2004.

[8] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. Introduction to Algorithms. McGraw-Hill, 1990.

[9] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma. Query expansion by mining user logs. *IEEE Trans. Knowl. Data Eng.*, 15(4):829–839, 2003.

[10] J. Dean and M. R. Henzinger. Finding related pages in the world wide web. *Computer Networks*, 31(11-16):1467–1479, 1999.

[11] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *Proc. of the Sixth ACM SIGKDD International Conference on Knowledge discovery and data mining (KDD'00)*, pages 150–160, Boston, MA, USA, 2000.

[12] G. W. Flake, S. Lawrence, C. L. Giles, and F. Coetzee. Self-organization and identification of web communities. *IEEE Computer*, 35(3):66–71, 2002.

[13] D. Gibson, J. M. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *Proc. of the Ninth ACM Conference on Hypertext and Hypermedia (HT'98)*, pages 225–234, Pittsburgh, PA, USA, 1998.

[14] A. Gulli and A. Signorini. The indexable web is more than 11.5 billion pages. In *Proc. of the 14th international conference on World Wide Web (WWW'05)- Special interest tracks and posters*, pages 902–903, Chiba, Japan, 2005.

[15] T. H. Haveliwala, A. Gionis, D. Klein, and P. Indyk. Evaluating strategies for similarity search on the web. In *Proc. of the Eleventh International World Wide*

Table 8: The Ordering lists of page 17 and page 1. "No.": corresponding page numbers in Figure 3; "X": non-related pages; other numbers are positions in the ordering list.

			Page 17 from c1					Page 1 from c3				
Category	Page URL	No.	Cit	$\alpha =$ 0.06	$\alpha =$ 0.25	$\alpha =$ 0.50	$\alpha =$ 0.80	Cit	$\alpha =$ 0.06	$\alpha =$ 0.25	$\alpha =$ 0.50	$\alpha =$ 0.80
c1	http://www3.primushost.com/...	2	X	11	7	X	X	X	X	12	X	X
c1	http://www.web-datamining.net/	3	5	5	5	5	X	X	X	X	X	X
c1	http://www.tysonsoftware.co.uk/	4	3	3	3	3	3	X	X	X	X	X
c1	http://www.the-data-mine.com/	5	X	12	9	X	X	X	X	9	X	X
c1	http://www.tdan.com/i010ht01.htm	6	6	7	X	7	14	X	X	X	X	X
c1	http://www.kluweronline.com/issn/...	14	X	X	X	8	12	X	X	X	X	X
c1	http://www.kdnuggets.com/gpspubs...	17	0	0	0	0	0	6	X	X	X	X
c1	http://www.dsslab.com/	18	2	2	2	2	2	X	X	X	X	X
c1	http://www.crm-forum.com/library/...	22	X	X	X	10	4	X	X	X	15	X
c1	http://www.ccsu.edu/datamining/...	24	4	4	4	4	X	X	X	X	X	X
c1	http://www.bgu.ac.il/ ratsaby/	25	7	6	6	6	15	X	X	X	X	X
c1	http://web.tiscali.it/wiseminosse/...	26	1	1	1	1	1	X	X	X	X	X
c1	http://home.kimo.com.tw/misforto/...	30	X	X	X	11	7	X	15	X	X	15
c2	http://www.secondmoment.org/	8	X	8	11	X	X	X	X	14	X	X
c2	http://www.modelandmine.com/	10	X	X	X	X	5	X	X	X	X	13
c2	http://www.megaputer.com/	12	8	13	8	14	X	X	14	10	X	X
c2	http://www.kdnuggets.com/tools.html	16	X	14	X	13	11	X	12	X	14	12
c2	http://www.db2mag.com/db_area/...	19	X	X	X	9	13	X	X	X	X	X
c2	http://kdd.ics.uci.edu/	29	X	X	X	X	9	10	10	X	9	10
c3	http://www-anw.cs.umass.edu/rlr/	1	X	X	15	X	X	0	0	X	0	0
c3	http://www.stormloader.com/gmdh/...	7	X	10	10	X	X	X	X	13	X	X
c3	http://www.proto-mind.com/	9	X	X	X	X	X	5	4	4	7	5
c3	http://www.miislita.com/	11	X	X	X	X	8	11	11	X	6	9
c3	http://www.learningtheory.org/	13	X	X	13	X	X	3	9	2	10	7
c3	http://www.kernel-machines.org/	15	X	15	14	X	X	X	7	3	11	8
c3	http://www.cs.monash.edu.au/ dld/mixture...	20	X	9	12	X	X	2	X	15	13	X
c3	http://www.cs.iastate.edu/ honavar/...	21	X	X	X	X	X	X	3	5	8	6
c3	http://www.cis.upenn.edu/ ais/	23	X	X	X	12	6	X	13	X	12	14
c3	http://satirist.org/learn-game/	27	X	X	X	X	X	4	2	8	2	2
c3	http://people.revoledu.com/kardi/tutorial/...	28	X	X	X	X	X	7	6	7	3	3
c3	http://home.earthlink.net/d̃waha/research/...	31	X	X	X	X	X	1	1	1	1	1
c3	http://cgm.cs.mcgill.ca/ godfried/...	32	X	X	X	15	10	9	8	11	5	11
c3	http://athos.rutgers.edu/ml4um/	33	X	X	X	X	X	8	5	6	4	4

Web Conference (WWW'02), pages 432–442, Honolulu, Hawaii, USA, 2002.

[16] X. He, H. Zha, C. H. Q. Ding, and H. D. Simon. Web document clustering using hyperlink structures. *Computational Statistics & Data Analysis*, 41(1):19–45, 2002.

[17] H. Ino, M. Kudo, and A. Nakamura. Partitioning of web graphs by community topology. In *Proc. of the 14th International Conference on World Wide Web (WWW'05)*, pages 661–669, Chiba, Japan, 2005.

[18] G. Jeh and J. Widom. Simrank: a measure of structural-context similarity. In *Proc. of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02)*, pages 538–543, Edmonton, Alberta, Canada, 2002.

[19] M. Kessler. Bibliographic coupling between scientific papers. *Journal of American Documentation*, 14(1):10–25, 1963.

[20] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proc. of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'98)*, pages 668–677, 1998.

[21] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. *Computer Networks*, 31(11-16):1481–1493, 1999.

[22] G. N. Lance and W. T. Williams. A generalized sorting strategy for computer classifications. *Nature*, 212:218, 1966.

[23] G. N. Lance and W. T. Williams. A general theory of classificatory sorting strategies: 1. hierarchical systems. *The Computer Journal*, 9:373–380, 1967.

[24] R. Larson. Bibliometrics of the world wide web: An exploratory analysis of the intellectual structure of cyberspace. In *Ann. Meeting of the American Soc. Info. Sci.*

[25] J. E. Pitkow and P. Pirolli. Life, death, and lawfulness on the electronic frontier. In *Proc. of the CHI 97 Conference on Human Factors in Computing Systems*, pages 383–390, Atlanta, Georgia, USA, 1997.

[26] 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(4):265–269, 1973.

[27] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In *Proc. of the 5th IEEE International Conference on Data Mining (ICDM'05)*, pages 418–425, Houston, Texas, USA, 2005.

[28] A. X. Zheng, A. Y. Ng, and M. I. Jordan. Stable algorithms for link analysis. In *Proc. of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01)*, pages 258–266, New Orleans, Louisiana, USA, 2001.