User-Assisted Similarity Estimation for Searching Related Web Pages

Lin Li† Masaru Kitsuregawa‡

Department of Information and Communication Engineering, University of Tokyo†

Institute of Industrial Science, University of Tokyo‡

ABSTRACT

To utilize the similarity information hidden in the Web graph, we investigate the problem of adaptively retrieving related Web pages. 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 similar pages, 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 adaptively supply the ordering list of related pages according to an adjustable parameter. 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.

1. Introduction

In the context of Web search, 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 become an important issue. Bibliometrics measures such as Co-citation, bibliographic coupling, etc. are the 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. Moreover, the Web graph is dynamic and massive in nature, loading the whole Web graph to memory is unavailable for the present and some pages have good content but have not been linked by many authors [2]. Our idea is that the propagation of localized bibliometrics similarity on a global neighborhood graph of similarity will alleviate the problem. In addition, we should not overlook the subjectivity of similarity. Page-to-page similarity is not a fixed value, especially for cross-topic pages. If the 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. We propose a user-assisted estimation approach for similarity search in this paper, which makes use of three bibliometrics measures and the flexibility of the sequential, agglomerative, hierarchical, nonoverlaping (SAHN) clustering for adaptively ranking related pages.

2. Bibliometrics Measure

To determine how related two Web pages are, our measure is based on three different similarity measures derived from their link information: co-citation, bibliographic coupling, and Amsler measures [1]. We find that the three measures do not take into account the direct links between two pages. Therefore, we modify them by adding the effect of direct links between pages. Our measure is defined as:

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

Here, $C(p_1)$ represents the inlinks of the page p_1 in the citation measure, the outlinks of the page p_1 in the bibiographic coupling measure, or the combination of the inlinks and outlinks of the page p_1 in the Amsler measure. The value of the function $direct(p_1, p_2)$ is assigned 0 if there is no direct link between the two pages, or 2 if the two pages link each other. Otherwise the value is 1.

3. SAHN Clustering

To produce 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. 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 $d_1 \leqslant d_2 \leqslant \cdots \leqslant d_k$ 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. [3] derived a flexible method of the SAHN clustering by the constraint ($0<\alpha \leqslant 1$), defined as

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

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}. The constraint suggests a set of monotonic methods such that as α increases from 0 to 1, we can adaptively rank the related pages by combination distances.

4. Our Approach

Our approach consists of the following three steps.

Step1: Computation of Similarity Scores

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 our measure. Moreover, one observation from our experiments indicates that the neighborhood graph of similarity is an unconnected graph which may be subdivided into connected subgraphs.

Step2: Extraction of Subgraph 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 is added into the graph. Last, we extract all connected components, also called neighborhood subgraphs of similarity here.

Step3: 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. We last specify how to output the ordering list of related Web pages (Ties in the SAHN clustering are broken randomly). Given that the value of α is decided by a user, the SAHN clustering outputs a hierarchical structure where an input page and a candidate page will come together at the combination distance d_1, and at the distance d_2, the input page is merged with a group (page) in the first time, and the first merging for the candidate page is at a distance d_3. Then, the distance score between the two pages is estimated to $|d_2 - d_1| + |d_3 - d_1|$ to rank each candidate in the neighborhood subgraph. More details are in [4].

5. Experiments and Conclusions

To test our approach of identifying related pages, we first extracted 48 web pages included by three categories from Google Directory: Data Mining (C1), Knowledge Discovery (C2), and Machine Learning (C3), as our core set. For each page in the core set, we utilized the Google API to obtain its in-links with in-degree restricted to 50, and fetched all out-links. Necessary preprocessing and data statistics are in [4].

By choosing α = 0.5, we measure the number of correctly clustered pages in their corresponding categories as shown in Table 1. 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.

Table 1: Precision and Recall

	Precision %			Recall %		
	C1	C2	C3	C1	C2	C3
Cit	86.7	66.7	93.3	68.8	41.7	75.0
Bib	80.0	33.3	66.7	25.0	8.33	33.3
Ams	84.6	62.5	85.7	81.3	50.0	87.5

The co-citation measure outperformed the Amsler measure on precision, though the recalls under the cocitation 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. Moreover, the values of the recall under the bibliographic coupling measure are terribly low. This result demonstrates that the out-link information is sparse and noisy and that the in-link information is more reliable. In conclusion, our proposed approach used the combination distances of the SAHN clustering method to adaptively rank the related pages according to a parameter adjusted by users. The experimental results show that the co-citation measure, an in-link based method, generally outperformed the other two measures in precision.

6. References

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

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

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

[4] L.Li, Z.Yang, K. Somboonviwat, and M.Kitsuregawa. User-Assisted Similarity Estimation for Searching Related Web Pages. In Proc. Of Hypertext, Pages: 11 - 20 , 2007.