Web Site Summarization using Document Associations

K. Selçuk Candan Wen-Syan Li

C&C Research Laboratories, NEC USA, Inc.
110 Rio Robles, M/S SJ100, San Jose, CA 95134, USA
Email:candan,wen@ccrl.sj.nec.com Tel:408-943-3028 Fax:408-943-3099

ABSTRACT

Hypermedia has emerged as a primary means for storing and structuring information. Yet, due to the continuously increasing size of these infrastructures, it is getting ever difficult for users to understand and navigate through such sites. We see that in order to overcome this obstacle, it is essential to use techniques that recover the Web authors' intentions and superimpose it with the users' retrieval contexts in summarizing Web sites. Therefore, in this paper, we present a framework which uses implicit associations among Web documents which considers three factors: (1) document separation (by the number, type, and content of the likes); (2) connectivity; and (3) document content.

Keywords

Web mining, association, link analysis, random walk, topic distillation, connectivity, summarization

1. INTRODUCTION

In this paper, we present a framework for discovering implicit associations among Web documents and describe its use for creating Web site summarizations. What differentiates our work from related work in the literature is that, *we are not only interested in discovering the existence document associations, but also in inducing the reasons why they are associated.* Knowing these reasons, among other things, is essential in (1) in creating Web site maps that matches Web designers' intentions and (2) in superimposing the logical structure of a Web site with the context provided by a visitors interests.

Given two web pages, A and B, we see that following two guidelines, along with the actual content of the pages, can be used to identify why they are associated:

1. Pages on a shorter path between A and B are stronger indicators than others to reflect why A and B are associated.
2. Pages which appear on more paths between A and B should be stronger indicators than others to reflect why A and B are associated.

A web page which satisfies both of the above criteria (i.e. near seed URLs and with high connectivity) would be a good representative for the association.

Based on this motivation, we develop a Web mining technique, based on a *random walk algorithm*, which considers three factors: (1) document distances by link; (2) connectivity; and (3) document content and we use this algorithm to construct of summarizations of Web neighborhoods, which can be viewed and used as a Web site map.

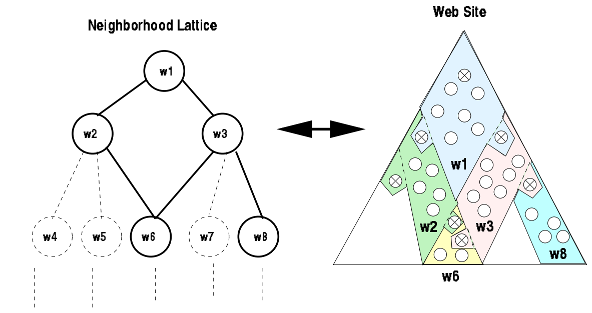

Figure 1: The crossed circles denote the entry points of neighborhoods. Each sub-neighborhood contain one entry point per its parent neighborhoods. Each parent neighborhood includes the entry points of its sub-neighborhoods.

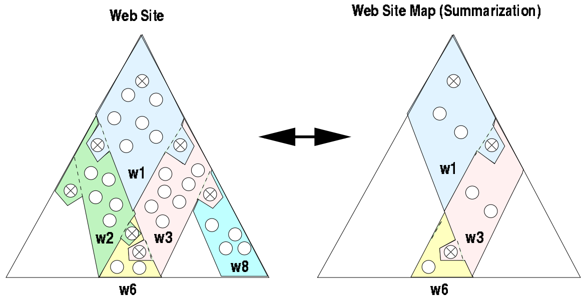

Figure 2: Summarization

2. WEB SITE SUMMARIZATION

We see that corporate sites, and most of the Web space, is composed of two types of neighborhood: physical and logical. However, as (1) the foci of users may differ from each other, (2) the focus of a single user may shift from time to time, in order to create a dynamic site map, it is imperative to use both physical and logical neighborhoods of a corporate site simultaneously.

Physical neighborhoods are decided by the link structures of the Web sites. In [1], we described algorithms to discover logical neighborhoods. Here we will assume that a corporate Web site, W, is already partitioned into logical neighborhoods. We denote this portioning as a partially ordered lattice $\mathcal{W}$ as shown in Figure 1.

Intuitively, the lattice corresponds to a hierarchy of neighborhoods. At the highest level, we have a neighborhood consisting of high-level corporate pages and the entry pages of lower neighborhoods. Similarly, each neighborhood consists of a set of high-level pages and the entry-pages of all its sub-neighborhoods. Consequently, summarization of

```
1. Let $G^N(V^N, E^N) \subseteq G^u$ be the neighborhood graph;
2. $V^\sigma = \emptyset$; $E^\sigma = \emptyset$;
3. Let $\mathcal{K} \subseteq V_{G^N}$ be the set of $k$ vertices with the highest dominance values;
4. $V^\sigma = \mathcal{K}$;
5. For each $v_i$ and $v_j \in V^\sigma$
   (a) $V_{temp} = V^N - \{V^\sigma - \{v_i, v_j\}\}$;
   (b) $E_{temp} = \{\langle v_k, v_l \rangle \mid (\langle v_k, v_l \rangle \in E^n) \wedge (v_k \in V_{temp}) \wedge (v_l \in V_{temp})\}$
   (c) If $sp(v_i, v_j)$ is the shortest path in $G_{temp}(V_{temp}, E_{temp})$ between $v_i$ and $v_j$ then
         i. Let the length of the path $sp(v_i, v_j)$ be $\Delta$
        ii. $E^\sigma = E^\sigma \cup \{e = \langle v_i, v_j \rangle\}$;
       iii. If $v_j$ is reachable from $v_i$ in the directed graph $G$ through the vertices in $sp(v_i, v_j)$, but if $v_i$ is not
            reachable from $v_j$, then $\delta(e) = \langle \Delta, right \rangle$
        iv. If $v_i$ is reachable from $v_j$ in the directed graph $G$ through the vertices in $sp(v_i, v_j)$, but $v_j$ is not
            reachable from $v_i$, then $\delta(e) = \langle \Delta, left \rangle$
         v. If $v_i$ and $v_j$ are reachable from each other in the directed graph $G$ through the vertices in $sp(v_i, v_j)$,
            then $\delta(e) = \langle \Delta, bi \rangle$
        vi. If neither $v_i$ nor $v_j$ is reachable from the other in the directed graph $G$ through the vertices in $sp(v_i, v_j)$,
            then $\delta(e) = \langle \Delta, none \rangle$
```

Figure 3: Algorithm for constructing a summary

W involves of two tasks: (1) identification of which nodes in the partially ordered lattice $\mathcal{W}$ will be shown to the user (i.e., focusing on the neighborhoods) and (2) summarization of each focussed neighborhood based on user interest. Figure 2(b) shows this process:

First, $w1$ is summarized. The entry point of $w3$ remains in the focus, whereas the entry point of $w2$ is out of focus. Since the entry point of $w2$ is out of focus, $w2$ is not summarized. Next, $w3$ is summarized. This time, the entry point of $w6$ remains in the focus, whereas the entry point of $w8$ is out of focus. Finally, $w6$ is summarized. Note that $w6$ has two entry points. However, since the entry point from $w2$ is out of focus, the summarization should be done with respect to the entry point from $w3$.

3. IDENTIFICATION OF FOCUS NEIGHBORHOODS

In order to identify the focus neighborhoods, we need to start from the root neighborhood, w_1 of $\mathcal{W}$. This neighborhood contains, the high-level pages of the given corporate site along with the entry level pages of it sub-neighborhoods. Let us assume that we are interested in a predetermined number, k, of focus points in this top neighborhood:

- In order to identify the k focal points of the neighborhood w_1 of $\mathcal{W}$, summarize w_1 into a graph of size k. The k remaining pages are in user focus.
- Let us assume that $F = \{v_1, v_2, \ldots, v_k\}$ are the k pages in the summary of w_1. If $v_i \in F$ is an entry-page of a sub-neighborhood, w_i, then repeat the same process for the sub-neighborhood w_i.

4. SUMMARIZATION OF A NEIGHBORHOOD

In order to summarize a given neighborhood, we first have to identify the pages that are important. In this case, the entry pages of a neighborhood (from parents in focus) are relatively important as they will connect the web site maps of the neighborhoods. Note also that the entry pages of the sub-neighborhoods are also important as they will extent the map downwards in the hierarchy, given that the lower neighborhoods are also in focus.

Therefore, given a neighborhood, w_i, the set, $\mathcal{E}$, of focussed entry pages from its parents, and the set, $\mathcal{L}$, of entry-pages to its sub-neighborhoods, we can create a set of seed pages (for summary) $\mathcal{S} = \mathcal{E} \cup \mathcal{L}$. Then our goal is,

- given the set, $\mathcal{S}$, of seed (entry) Web pages,
- potentially a content-description,
- a Web neighborhood, $G^N = w_i$ which contains these seeds, and
- an integer k,

to create a *summary*, with k pages, of the neighborhood with respect to the seed pages.

Observation: The concept of summarization is related to the concept of *document associations*. Since, the Web site map is a set of representative nodes in a Web site, the nodes in a Web site map needs to satisfy the following criteria: (1) high connectivity so that users can navigate from these Web site map nodes to other nodes easily; and (2) the contents of these Web site map nodes need to be representative.

An algorithm based on these observations is given in Figure 3. Note that Step 3 of the algorithm requires the identification of the k most dominant vertices (or the vertices which describe the document associations the best) in the graph with respect to the seed vertices. In [2] we have describe an algorithm that finds such dominant vertices which describe document associations. Those entry pages which are still in the map after the summarization are called *focussed entry-pages*, and they point to the other logical domains that have to be further explored and summarized. Therefore, we recursively apply the summarization algorithm described above for those domains who have at least one *focused entry-page*.

5. CONCLUSION

In this paper, we present a framework for site map construction and Web page summarization. For this purpose, we have built on an algorithm we developed in'[2] for mining implicit associations among Web documents, induced by Web link structures and document contents.

6. REFERENCES

[1] Wen-Syan Li, Okan Kolak, Quoc Vu, and Hajime Takano. Defining Logical Domains in a Web Site. In *Proceedings of the 11th ACM Conference on Hypertext*, pages 123–132, San Antonio, TX, USA, May 2000.

[2] K. Seluk Candan and Wen-Syan Li. Using Random Walks for Mining Web Document Associations. In *Proceedings of 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining Conference*, pages 294–305, Kyoto, Japan, April 18-20, 2000