Matching Web Site Structure and Content

Vassil Gedov[1] Carsten Stolz[2]

Ralph Neuneier[3] Michal Skubacz[3] Dietmar Seipel[1]

{gedov,seipel}@informatik.uni-wuerzburg.de, University of Würzburg, Germany[1]
carsten.stolz@ku-eichstaett.de, University of Eichstätt-Ingolstadt, Germany[2]
{ralph.neuneier,michal.skubacz}@siemens.com, Siemens AG, Corporate Technology, Germany[3]

ABSTRACT

To keep an overview of a complex corporate web sites, it is crucial to understand the relationship of contents, structure and the user's behavior. In this paper, we describe an approach which is allowing us to compare web page content with the information implicitly defined by the structure of the web site. We start by describing each web page with a set of key words. We combine this information with the link structure in an algorithm generating a context based description. By comparing both descriptions, we draw conclusions about the semantic relationship of a web page and its neighborhood. In this way, we indicate whether a page fits in the content of its neighborhood. Doing this, we implicitly identify topics which span over several connected web pages. With our approach we support redesign processes by assessing the actual structure and content of a web site with designer's concepts.

General Terms: Algorithms

Categories and Subject Descriptors: H.3.3 Information Systems Information search and retrieval; I.5.4 Computing methodologies Applications

Keywords: Web Structure, Web Content Mining, Semantic Description

1. INTRODUCTION

Facing the complexity of dynamically generated corporate web sites, it becomes increasingly difficult to understand user navigation patterns without deep knowledge of the content and the semantical linkage of the pages. Another challenge for web site owners as well as web strategists is to keep track of the information structures on the web site in the content management system. Following recent research concerning integration of web structure, content and usage mining [1] [2], we intend to provide an insight into the relationship between the structure and the content. We first try to extract the key content from each page taking its hypertext markup into account. Then, we reason about the role of the web structure[3] since we believe that the linkage created by the web designer conceals a semantic connection between web pages.

WWW2004, May 17–22, 2004, New York, NY USA.
ACM 1-58113-912-8/04/0005.

2. COMBINING STRUCTURE AND CONTENT INFORMATION

Before associating structure and content with each other, we will try to define them. The **structure** of a web site is determined by the link structure between its web pages. A link in this context is considered to be a link encoded as a HTML tag. We do not consider the anchor text of a link as structural information since we believe it belongs to the content information. Excluding self-references, we focus on hard coded inter-page links represented by a directed cyclic graph.

The **content** of a web site may consist of text, hypertext, meta information or multimedia content like pictures, figures, video or sound. Here, we concentrate on text and html since they contain most of the information. As more and more web sites use different media types it could be reasonable to include those in future work .

In order to combine structure and content, we have to **map structural and content information** to a common concept. We achieve it by putting them in a common data structure, in order to make them comparable. Next, we combine the textual content of a web page with linked contents, and thus we implicitly create a content-structure graph.

3. ALGORITHM

In this section we describe how to create a content - structure graph and compare it with the structure graph of a web site. Thereafter, we present a way to measure the distance between both.

3.1 Structure and Content Description

From each page we extract the links within a web site generating an implicit directed cyclic structure graph. We consider site-internal links for each web site, respectively. In order to describe the content, we extract all words and stem them. Further more, stop words are filtered as we are only interested in content, and not in style or grammatical information. Next, we calculate word frequencies per web page. We assume that the most frequent words define the content of the web page. The highest ranked words are said to be key words, where their number is proportional to the text length. Not all key words contribute to the semantic: very frequent key words like the company's name, occurring virtually on all pages, are pruned. Whereby the pruning threshold must be determined empirically. Additionally to

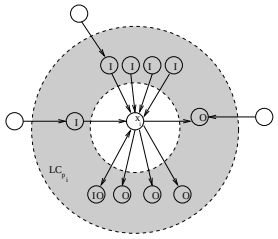

Figure 1: The Local Context LC

the key words, phrases from HTML tags are taken into account since we believe that they reflects author's focus. This could be for example, <h1> which typically describes the major topic.

3.2 Mapping

For a single page, we combine its set of key words with the key words of its direct neighbourhood. The latter is defined as the pages, having a direct link to (inedges) or from (outedges) the page in focus (see figure 1). We call it the local context (LC) of a page x_i where:

$$LC(x_i) = \{x_j | x_j \in (\text{inedges}(x_i) \cup \text{outedges}(x_i))\} \qquad (1)$$

We perform the combination of structure and content by generating a new set of key words. The new set is initiated with the own key words of the processed web page. Now, involving the site structure, we add the key words from the LC and accumulate the frequencies over the entire key word set. Additionally, the anchor tags from the LC are included as key words. To avoid overweighting of the LC, we assigned low constant frequencies to the key words originating from it. Then, we arrange the set according to the accumulated frequency. The number of words, which we keep, is again proportional to the text length. We perform this process bottom-up with respect to the minimal click path from the entry point of the web site. For already processed pages we use the new combined descriptions.

We interpret the results of our algorithm as an intersection between the contents of the LC and the particular page. Moreover, we assume that salient words ascend towards the root page. In the empirical study, we will discuss the relationship between the key word set which describes the page alone content and the new combined set.

4. EMPIRICAL STUDY

We selected corporate web sites of Bayer, Motorola and Siemens in order to evaluate the performance of the above introduced algorithm. For a random set of test pages, we compared the automatically generated key word sets with manually compiled abstracts.

Studying the results, we observed that the page key words (like those in the left column of table 1) cover nearly all topics mentioned in the abstracts which was desired.

Next, we analyzed the results of the mapping from section 3.2. An example is shown in the right column of table 1. Comparing this combined set with the manual abstract as well as with page key words, we observed that some key words - topics fell out. These key words carried page specific information. In contrast, the newly inserted words reflect topics present in the LC of the page in focus. A manual review of the LC confirmed the relevance of the new words.

Page key words		LC key words	
Network	13	Network	32.03
revenue	7	Operators	14.11
Wireless	6	GPRS	11.9
Quality	6	Support	10.3
Operators	6	CDMA	8.14
subscribers	6	Brochure	5.92
Infrastructure	5	GSM	5.53
Operability	1	UMTS	4.22
Capacity and Coverage	1	Contact Motorola	4

Table 1: www.motorola.com/networkoperators

Regarding the two different sets of key words, the differences between them can be interpreted as follows. In the case of a homogeneous LC with topics matching the content of the page in focus, we observed no or only slight changes in its key word set. On the other hand, pages dealing with different topics than its homogeneous LC, showed significant differences. Whereas a heterogeneous LC, regardless of the topic of the page it is unlikely to have any significant influence on the resulting set.

While evaluating the mapping, we observed several effects which led us to improvements: In contrast to the page-wise extracts, the HTML tag based key words were removed from mapping in order to rule out local information, specific only to one page in the LC. On the other hand, we kept the anchor tags linking to this page as they enhance the relation between structure and content.

Our experiments showed that by excluding inedges from the LC, the key words reflected the relationship of the page and the subsequent contents more suitably. By 'subsequent' we imply that, except for navigation links, the outgoing links lead to more detailed information on the parent topic.

5. CONCLUSIONS

We showed that our approach is able to make structure and content comparable. In this way, our approach can indicate whether content of a page fits to content of the web pages in its neighborhood whereas the neighborhood is defined by the link structure. Additionally, we implicitly identify topics which span over several connected web pages. Which in a way leads us to discovering of semantical relationships. With our algorithm we can support web designers and strategists by comparing their intentions with the actual structure and content of a web site. A subsequent redesign of a web site incorporating our results can improve user perception and customer retention.

Our future research includes developing more differentiating measurements for the structure and content analysis. Furthermore, we find it interesting to combine our approach with usage patterns to improve the topic identification capabilities of our algorithm.

6. REFERENCES

[1] K. Bharat and M. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment Proceedings of SIGIR-98, p. 104–111, ACM Press, 1998

[2] A. Sun and E.-P. Lim. Web unit mining: finding and classifying subgraphs of web pages. In Proceedings 12th Int. Conf. on Information and knowledge management, p. 108–115. ACM Press, 2003.

[3] Soumen Chakrabarti. Mining the Web - Discovering Knowledge from Hypertext Data. Morgan Kaufmann, 2002.