Measuring Semantic Relations of Web Sites by Clustering of Local Context

Carsten Stolz[1], Vassil Gedov[2], Kai Yu[3], Ralph Neuneier[3], and Michal Skubacz[3]

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

Abstract. Our contribution in this paper is an approach to measure semantical relations within a web site. We start with a web page description by key words. The implementation of structural and content information reduces the variety of key words. Thereby, the document-key-word-matrix is smoothend and similarities between web pages are emphasized. This increases the possibility of cluster key words and identif topics successfully. To do so, we implement a probabilistic clustering algorithm. To assess semantic relations, we introduce a number of measures and interpret them.

1 Introduction

Facing a huge, complex corporate web site, it is difficult to keep track of daily changing web pages and their relationship to the rest of the web site. Therefore, it is helpful to provide an insight into the content structure of a web site. With this semantic information it would be possible to improve the web site design and its usability.

In order to create a semantical overview of a web site, the first step should be the analysis of the web content. In a second step we should analyze the structure of the web site. Based on the assumption that the link structure provides useful semantic clues [2], it is important to integrate these implicit information. Additional insight into the semantical structure can be provided if each page is integrated into its context [3]: If a page is not regarded as a solitaire, the information of its neighborhood is relevant for its understanding.

In contrast to this approach, **related work**, done in this area of research[2] uses a pre-defined set of topics. Chakrabati et al. [2] argues that predefined topic taxonomies circumvent key word ambiguity. For their purpose, it is reasonable since they do not focus on huge corporate web sites. In contrast we do and aim to create a generic solution. Like them, we do not rely on well-kept meta information. Including an identification of hubs and authorities [1] has to be analyzed whether it can improve our results. Reasonable results of document clustering incorporating hyperlink structures can be found in He et al. [4]. The usage of the folder structure like Sun et al. describe in [5] is not possible for our target sites. Modern content management systems prevent from the use of the folder structure by cryptic URLs.

N. Koch, P. Fraternali, and M. Wirsing (Eds.): ICWE 2004, LNCS 3140, pp. 182–186, 2004.

2 Algorithm

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. In order to create single page descriptions, a crawler extracts key words and follows all web-site-internal links in a breadth-first mode. The words are stemmed in order to reduce key word diversity by applying the Porter Stemming Algorithm. 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. Very frequent key words like the company's name, occurring virtually on all pages, are pruned. In [3] we empirically determined the usable HTML tags, which we believe to reflect author's focus.

2.1 Local Context Page Description

In order to include the link structure, we combine a page's set of key words with the key words of its direct neighborhood like [1]. The latter is defined as the pages, having a direct link to (inedges) or from (outedges) the page in focus. We call it the **local context (LC)** of a page. We perform the combination of structure and content by generating a new set of key words. We add the key words from the LC to the processed page and accumulate the frequencies over the entire key word set including the anchor tags as key words, ordered by frequency. The number of words is kept proportional to the text length. We perform this process bottom-up with respect to the minimal click path. 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.

2.2 A Topic Discovery Algorithm

By clustering key words we are trying to identify topics. In other words, we consider word classes (topics) $\{z_j\}_{j=1}^{L}$, and model the likelihood of a document d (web pages) as follows,

$$p(\mathbf{w}_d) = \prod_w \big(\sum_z p(w|z,\beta)p(z|\theta_d) \big)^{n_{d,w}} \tag{1}$$

where β are parameters specifying the probability of words given topics, θ_d are document-specific parameters that indicate the topics mixture for document d, $\mathbf{w}_d$ is the word list of this document and $n_{d,w}$ is the number of occurrences of word w in $\mathbf{w}_d$. Then given a corpus of documents $d = 1, \ldots, n$, we have the following EM algorithm to estimate the parameters:

– E-step: for each document d, we estimate the posterior distribution of topics given each word w in it:

$$p(z|w,d) = \frac{p(w|z,\beta)p(z|\theta_d)}{\sum_z p(w|z,\beta)p(z|\theta_d)} \tag{2}$$

– M-step: we maximize the log-likelihood of "complete data" for the whole corpus:

$$\sum_d \sum_{\mathbf{Z}} \left(\prod_w p(z|w,d) \log \prod_w \left(p(w|z,\beta)p(z|\theta_d)\right)^{n_{d,w}} \right) \tag{3}$$

with respect to parameters β and $\{\theta_d\}_{d=1}^N$, which gives rise to

$$\beta_{i,j} = p(w_i|z_j) \propto \sum_d n_{d,w_i} p(z_j|w_i,d) \tag{4}$$

$$\theta_{j,d} = p(z_j|d) \propto \sum_w n_{d,w} p(z_j|w,d) \tag{5}$$

We perform the E-step and M-step iteratively and at the convergence obtain the final estimate of $p(w|z)$ and $p(z|d)$. $p(w|z)$ indicates the probability of occurrence of word w given topic z. The algorithm groups semantically related words into topics. Intuitively, if several words often occur together (in the same documents), then these words are likely to be associated with one topic. $p(z|d)$ indicates the probability of topic z for document d. The parameters transform documents, which are originally distributed in high-dimensional but low-level word space, into low-dimensional and high-level topic space.

3 Evaluation and Case Study

We describe measures to evaluate the improvements by the LC. Therefore, we evaluate each measurement for both key word descriptions: stand alone and the LC. We performed our clustering algorithm on our data with 100 topics (word clusters). Where the number was determined by applying SVD on our data. First, we provide some **basic definitions**: The relevant topics for a specific document are the topics with the highest probability $topn(p(z_j|d)$, given document d_i, called *topic mixture* for d. The number of documents in LC with respect to a particular document is $|LC(d_i)|$. We are also interested in the number of distinct topics which are most relevant for each document in the LC. The number may vary from 1 to $|LC(d_i)|$, respectively.

3.1 Evaluation of Topic Clustering

For our analysis we have used Siemens' corporate web site: *www.siemens.com*. We crawled 1569 web pages, extracted key word sets (stand alone description)and

derived key word set involving the LC. Regarding a particular document, we are now able to determine which topic is the most relvant one. Having this topic, we then select the documents which are most similar with respect to it. We observed in table 1, that the number of shared relevant topics has increased by using the LC: instead of one, now 3 of 4 topics are shared. **Topic mixture key words**: For a given document d_i, we determine the topic mixture. Each document d_i is represented by a topic vector $v_{d_i} = p(z|d_i) = \{z_1...z_n\}$. For all elements in the vector ($z_j \in v_d$) we select the topic key words. They are weighted with their corresponding probabilities for that topic. Finally, the probabilities of all topic key words are proportionaly accumulated which form the topic mixture key words: $\sum_z p(w|z)p(z|d_i)$. This mixture is supposed to describe the document which contains one to several topics with different importance. We are now interested in the most important key words for this mixture. A word is said to be a key word for the topic mixture, if it has a high accumulated value.

Table 1. Topic document rate in the LC for our sample.

	stand alone	LC
Number of documents of the LC	21	21
Number of distinct topics of the LC	20	4
Number of occurrences of topic 3 in the LC	-	11 of 21
Topic document rate	0,048	0,809

Topic document rate in LC: In order to evaluate the impact of the LC based key word sets, we compare the number of distinct topics for the pages in the specific LC. As we discussed above, the number of topics in a LC may vary from 1 to the number of documents in the local context. So, the topic document rate is defined as $1 - \frac{\#\text{topics } \in LC(d_i)}{|LC(d_i)|}$. Consecutively, the more topics are found the less is the rate. Note, that in both cases the pages within the LC are the same, but in the stand alone description the LC based algorithm from [3] is not applied. In table 1 we see that there are almost as many distinct topics as documents in the LC of our sample page. Furthermore, the most relevant topic of our test page does not occur at all. The number of distinct topics in the LC based description (table 1) has decreased significantly, and thus the topic document rate for the LC based description is very high. In addition, the most relevant topic for the sample page (topic ID 3) occurrs eleven times as most relevant topic for pages in the LC. Obviously, our LC based approach reduces key word diversity between linked documents.

3.2 Case Study for LC-Interpretation

By applying our measures on the data we have shown that the LC based description improves the result of our clustering algorithm. Now we want to exploit the

additional information revealed by the local context. The LC of a selected page consists of 21 pages. We compare the most relevant topics for each page of the LC - one topic per page. The table shows that from these 21 topics topic no. 3 is the most relevant topic. It occurred 11 times representing more than half of the pages. Regarding the variety of topics, one sees that the most relevant topics for 21 pages consist of 4 different topics. This observation is supported by analysis of the LC key words. Out of 215 key words one finds 65 distinct key words. Concluding these observations, one can characterize this LC as *homogeneous*. A manual inspection of the LC pages affirms this conclusion.

4 Conclusion

This paper describes and evaluates the concept of the local context (LC). First, we use the LC successfully to smoothen the stand alone key words what we have shown empirically. Furthermore, we have shown that by using the LC the web site structure gets incorporated into the key word description of a page. Consequently, the clustering algorithm benefits from these information which identifies topics successfully. In the case study, we have described one application of the results: We are able to characterize a web page and its surrounding pages in terms of their content homogenity. This is only a part of the potential applications. Consequently, we will compare structure and content of a web site intensively and combine them with an user behaviour analysis.

Since this approach does not rely on predefined topics, it can be used generically. Additionally, we attempt to give more emphasis to the web site structure by using an adjacency matrix in the cluster algorithm.

References

1. K. Bharat and M. Henzinger; Improved algorithms for topic distillation in a hyperlinked environment Proc. of SIGIR-98, p. 104–111, ACM Press, 1998
2. S. Chakrabarti; Mining the Web - Discovering Knowledge from Hypertext Data. Morgan Kaufmann, 2002
3. V. Gedov, C. Stolz, R. Neuneier, M. Skubacz, D. Seipel, Matching Web Site Structure and Content, WWW04, New York (2004)
4. X. He, H. Zha, C. Ding, and H. Simon Web document clustering using hyperlink structures Computational Statistics and Data Analysis, 41:19-45, 2002
5. A.Sun and E.-P. Lim. Web unit mining, In Proc. Int. Conf. on Information and Knowledge Manag. p. 108–115. ACM Press, 2003.
6. W. Wong and A. Fu, Incremental Document Clustering for Web Page Classification, IEEE 2000 Int. Conf. on Info. Society (IS2000)