Extracting Semantic Structure of Web Documents Using Content and Visual Information

Rupesh R. Mehta
Dept. Comp. Sc. Engg.
Indian Institute of Technology
Kanpur 208016, India
rrmehta@iitk.ac.in

Pabitra Mitra
Dept. Comp. Sc. Engg.
Indian Institute of Technology
Kanpur 208016, India
pmitra@iitk.ac.in

Harish Karnick
Dept. Comp. Sc. Engg.
Indian Institute of Technology
Kanpur 208016, India
hk@iitk.ac.in

ABSTRACT

This work aims to provide a page segmentation algorithm which uses both visual and content information to extract the semantic structure of a web page. The visual information is utilized using the VIPS algorithm and the content information using a pre-trained Naive Bayes classifier. The output of the algorithm is a semantic structure tree whose leaves represent segments having unique topic. However contents of the leaf segments may possibly be physically distributed in the web page. This structure can be useful in many web applications like information retrieval, information extraction and automatic web page adaptation. This algorithm is expected to outperform other existing page segmentation algorithms since it utilizes both content and visual information.

Categories and Subject Descriptors: H.5.4 [Information Interfaces and Presentation]: Hypertext/Hypermedia

General Terms: Algorithms, Design.

Keywords: DOM, VIPS, Naive Bayes classifier, Topic hierarchy, Page segmentation

1. INTRODUCTION

Web documents are often heterogeneous and contain multiple topics which do not have relation among each other. They also contain navigational and contact information links which are irrelevant to the topic of the web page. A better retrieval performance can be achieved by considering the page not as an indivisible unit but as having an underlying semantic structure with topically coherent segments as atoms.

2. VISION BASED PAGE SEGMENTATION

The initial attempts to utilize the semi-structured nature of web documents were through document object model (DOM) trees extracted from the markups. However, web authoring is often casual and therefore DOM poorly reflects the actual semantic structure of a page. Visual page layout structures are more faithful to the semantic partitioning of a page. The vision based page segmentation (VIPS) [2] algorithm utilizes the fact that semantically related contents are often grouped together, and the entire page is divided into different regions using implicit or explicit visual separators such as images, lines, font sizes, blank areas, etc. VIPS iteratively uses DOM structure and visual cues for block extraction, separator detection and content structure generation. Based on these visual perceptions each node of the VIPS-tree [2] is assigned a 'Degree of Coherence' (DoC) value (ranging between 1-10) to indicate how coherent the content in the block is. A higher DoC value signifies that a segment is more homogeneous.

3. DETERMINING TOPIC COHERENCY OF PAGE SEGMENTS

The content homogeneity of a page block can be determined from its degree of belongingness to different categories in a topic hierarchy. The open directory project (ODP) [4] topic hierarchy is considered for this purpose. The degree of belongingness to a category is measured in terms of the 'posterior' probability of classifying a segment into the category using a pre-trained naive Bayes classifier [3]. If the probability is above Th $(= 0.4$, say) for some category, the segment is considered as belonging to that category.

The basic principle of naive Bayes classifier is to use the joint probabilities of words and categories to estimate the probabilities of categories given a document. The primary reason for using this method instead of a vector space cosine similarity approach are as follows: Page segments are usually of short length and the vector space model is not a good representative of their semantic content. By using a trained classifier one exploits the statistical properties of the training corpus in the coherency determination process. Also, the classifier categorizes a segment into a unique topic belonging to the finite set of topics in the hierarchy. This segment category is later used as its index while generating the semantic structure tree. It may be noted that purely content based document segmentation without considering any markup structure has been a well studied problem in information retrieval [1]. The main hindrance in this approach is the difficulty in identifying segment boundaries.

4. PROPOSED ALGORITHM

In this article, we propose a page segmentation algorithm which uses both visual and content information to obtain semantically meaningful blocks. The VIPS algorithm is initially used to obtain fine grained segments. However the depth of the semantic tree is controlled by their topic coherency as measured by a pre-trained naive Bayes classifier.

WWW 2005, May 10–14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

This is followed by a merging phase where topically similar segments are combined together. The output of the algorithm is a *semantic structure tree.* The leaves of the tree represent segments having unique topic and the contents of the leaf segment may possibly be distributed across the web page. The two phases of the algorithm are described below.

Split Phase: In this phase, VIPS algorithm [2] with a permitted degree of coherence (pDoC), is applied to a web page to get an initial content structure tree. Considering each leaf node of a tree as a segment, we check whether the segment belongs to more than one category in topic taxonomy [4] using the trained naive Bayes classifier. If it belongs to more than one category, it implies that it is still not semantically coherent and needs further segmentation. We therefore apply VIPS algorithm recursively by increasing the permitted degree of coherence, till each segment belongs to exactly one category or number of words in the segment are below some threshold. In this recursive process the content structure tree is modified as specified by the following algorithm.

```
Algorithm: Splitting (page, Parent, pDoC)
i. Apply VIPS with permitted degree of coherence (pDoC) on a page to get the initial content structure tree.
ii. Add the content structure tree to Parent, if Parent is not null.
iii. For each leaf node (segment) in content structure tree:
    (a.) Compute the probabilities of belonging to each category at some predefined level of topic taxonomy, using naive Bayes classifier.
    (b.) If for a segments the probability value is higher than some threshold for more than one classes, apply Splitting (page, Parent, pDoC+1) recursively.
    (c.) Else (if segment belongs to only one category) do not fragment that segment.
```

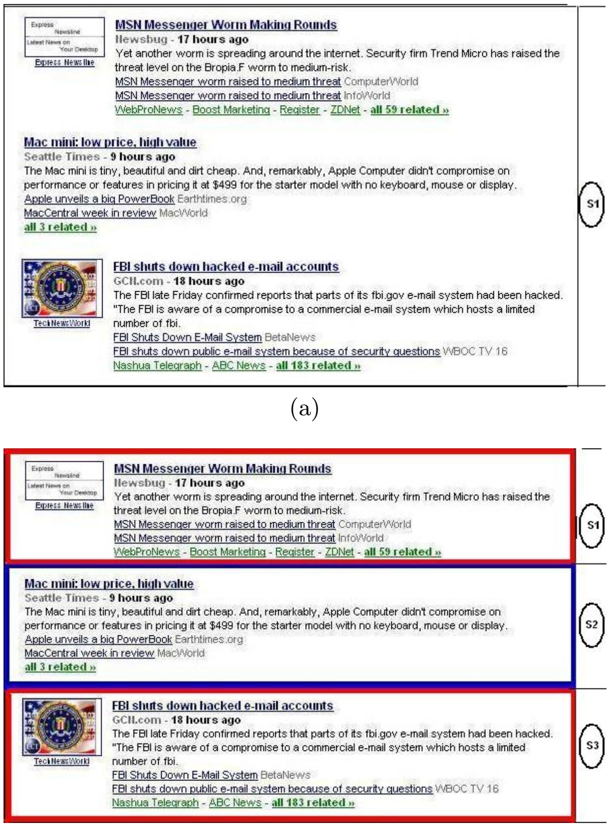
Merge Phase: After the Split phase, we get a content structure tree containing fine grained segments at leaf level. Each segment at leaf level of a tree belongs to exactly one category. In this phase, segments belonging to same category are identified and merged to form a new segment. Parent of newly formed node is a least common ancestor of all merged nodes in content structure tree. This helps in merging the semantically related contents which are distributed across the web page.

5. EXPERIMENTAL RESULTS

Experiments were performed on several pages with heterogeneous content. In most cases, the proposed algorithm produced segments which were semantically more meaningful. Fig. 1 shows, as an example, the output of VIPS and the proposed algorithm on a part of a Google Sci./Tech. news page. Fig. 1.a is the output of VIPS with pDoC=6. It treats whole part as a single segment (S_1), whereas Fig. 1.b is output of the proposed algorithm where segments (S_1 and S_3) related to computer security category are merged together to form a single block inspite of their physical distribution in the page. The unrelated segment (S_2) lying between these two blocks is extracted as a separate block. Experimental studies on information retrieval tasks are currently being performed.

6. CONCLUSION AND DISCUSSIONS

A method for extracting semantic structure of web pages based on content and visual cues is proposed. In this algorithm, topic coherency determines the granularity of the individual segments and thus circumvents the need to choose a DoC threshold as in the VIPS algorithm. The assumption of VIPS that semantically related contents are grouped together in a web page is not always true. Semantically related segments may be distributed across the web page intentionally or unintentionally. In our approach, physically distributed but semantically homogeneous blocks are treated as a single unit. In future, incorporating topic hierarchy in the semantic structure tree of a web page would make it more amenable to semantic web mining. Such a semantic structure of a web page can be useful in enhancing the performance of various web applications like information retrieval, information extraction and automatic web page adaptation.



MSN Messenger Worm Making Rounds
Newsbug - 17 hours ago
Yet another worm is spreading around the internet. Security firm Trend Micro has raised the threat level on the Bropia.F worm to medium-risk.
MSN Messenger worm raised to medium threat ComputerWorld
MSN Messenger worm raised to medium threat InfoWorld
WebProNews - Boost Marketing - Register - ZDNet - all 59 related »

Mac mini: low price, high value
Seattle Times - 9 hours ago
The Mac mini is tiny, beautiful and dirt cheap. And, remarkably, Apple Computer didn't compromise on performance or features in pricing it at $499 for the starter model with no keyboard, mouse or display.
Apple unveils a big PowerBook Earthtimes.org
MacCentral week in review MacWorld
all 3 related »

FBI shuts down hacked e-mail accounts
GCN.com - 18 hours ago
The FBI late Friday confirmed reports that parts of its fbi.gov e-mail system had been hacked. "The FBI is aware of a compromise to a commercial e-mail system which hosts a limited number of fbi.
FBI Shuts Down E-Mail System BetaNews
FBI shuts down public e-mail system because of security questions WBOC TV 16
Nashua Telegraph - ABC News - all 183 related »

(a)

MSN Messenger Worm Making Rounds
Newsbug - 17 hours ago
Yet another worm is spreading around the internet. Security firm Trend Micro has raised the threat level on the Bropia.F worm to medium-risk.
MSN Messenger worm raised to medium threat ComputerWorld
MSN Messenger worm raised to medium threat InfoWorld
WebProNews - Boost Marketing - Register - ZDNet - all 59 related »

Mac mini: low price, high value
Seattle Times - 9 hours ago
The Mac mini is tiny, beautiful and dirt cheap. And, remarkably, Apple Computer didn't compromise on performance or features in pricing it at $499 for the starter model with no keyboard, mouse or display.
Apple unveils a big PowerBook Earthtimes.org
MacCentral week in review MacWorld
all 3 related »

FBI shuts down hacked e-mail accounts
GCN.com - 18 hours ago
The FBI late Friday confirmed reports that parts of its fbi.gov e-mail system had been hacked. "The FBI is aware of a compromise to a commercial e-mail system which hosts a limited number of fbi.
FBI Shuts Down E-Mail System BetaNews
FBI shuts down public e-mail system because of security questions WBOC TV 16
Nashua Telegraph - ABC News - all 183 related »

(b)

Figure 1: Page segmentation using (a) VIPS, (b) proposed algorithm

7. REFERENCES

[1] D. Beeferman, A. Berger, and J. Lafferty. Statistical models for text segmentation. *Machine Learning*, 34(1-3):177–210, 1999.

[2] J.-R. W. D. Cai, S. Yu and W.-Y. Ma. Extracting content structure for web pages based on visual representation. In *Proc. 5th Asia Pacific Web Conf*, Xi'an, China, 2003.

[3] T. Mitchell. *Machine Learning*. McGraw-Hill, NY, 1997.

[4] Open directory project. http://dmoz.org/.