

INTERNET ARCHIVE
WayBackMachine BETA
http://smi.ucd.ie/hyppia/publications/DELOS_workshopAF/node3.html
Go
DEC 23 2003
17 captures
27 Oct 02 - 12 Apr 06
n prev

Next: [Opinion Classification](#) Up: [Fact or fiction: Content](#) Previous: [Research Issues](#)

Body Text Extraction

Articles published on the WWW often contain extraneous clutter. Most articles consist of a main body which constitutes the relevant part of the particular page. Surrounding this body is irrelevant information such as copyright notices, advertising, links to sponsors, etc. By identifying this main body of the article and basing our classifiers on features that occur within this area we anticipate improving the accuracy of classifiers. For example, images and links occurring within this body are likely to be very relevant to the article in question, while images and links outside this body are far less likely to provide interesting information about the article. Similarly when extracting text from a web page for use in classification, text extracted from the article body is much less likely to mislead the classifier than text extracted from the surrounding clutter. Fig. 2 shows a typical news article with the main body of text marked. Below the body is a generic copyright notice, while on each side are links to advertisers and unrelated articles.



Figure 2: A typical news article.

Automatically identifying the region of the web page that contains the main body of text is useful when building a classifier. We found that text from outside the body of an article was less likely to be relevant to the document and therefore more likely to mislead a text classifier.

Identifying the main body of a web page in a general robust manner is a difficult information extraction problem. Our approach is to view a web page as consisting of two kinds of tokens: HTML tag tokens and text tokens (i.e., words). Thus an HTML page can be represented as a sequence of bits B , with a $B_n = 0$ indicating that the n 'th token is a word, and $B_n = 1$ indicating a tag. Fig. 3 shows a graph of the cumulative distribution of tags, as a function of the position in the document, for the article from Fig. 2. Our extraction algorithm is based on the idea that the ends of the "plateau" in the middle of the distribution correspond to the points at which the main body of the article begins and ends.

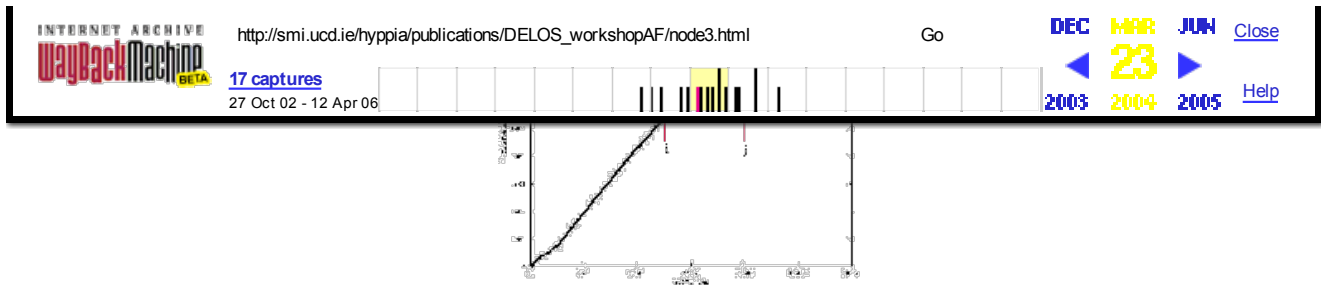



Figure 3: The cumulative distribution of tags, as a function of the position in the document. The central ``plateau'' corresponds to the body of the text highlighted in Fig. 2.

The problem can now be viewed as an optimization problem. We must identify points i and j such that we maximize the number of tag tokens below i and above j , while simultaneously maximizing the number of text tokens between i and j . The text is only extracted between i and j . We chose a simple objective function $T_{i,j}$ where:

$$T_{i,j} = \sum_{n=0}^{i-1} B_n + \sum_{n=i}^j (1 - B_n) + \sum_{n=j+1}^{N-1} B_n$$

The advantage of this method of text extraction is that it does not require any parameters, and can extract text from different web sites without the need for site specific wrappers.

To summarize, we use this text extraction algorithm to extract the body of text from web pages  for use by our classifiers. This was motivated by the observation that text from outside the main body often misled the classifiers.

 [n](#)  [prev](#)

Next: [Opinion Classification](#) **Up:** [Fact or fiction: Content](#) **Previous:** [Research Issues](#)
 aidan 2001-06-27