Automatic Extraction of Textual Elements from News Web Pages

Hossam Ibrahim, Kareem Darwish, Abdel-Rahim Abdel-sabor
Cairo University
5 Ahmed Zuwail St., Orman, Giza, Egypt
E-mail: hossamsi@yahoo.com, kareem@darwish.org, abdelrahim@alzoa.com

Abstract

In this paper we present an algorithm for automatic extraction of textual elements, namely titles and full text, associated with news stories in news web pages. We propose a supervised machine learning classification technique based on the use of a Support Vector Machine (SVM) classifier to extract the desired textual elements. The technique uses internal structural features of a webpage without relying on the Document Object Model to which many content authors fail to adhere. The classifier uses a set of features which rely on the length of text, the percentage of hypertext, etc. The resulting classifier is nearly perfect on previously unseen news pages from different sites. The proposed technique is successfully employed in Alzoa.com, which is the largest Arabic news aggregator on the web.

1. Introduction

Since we live in an increasingly global community that motivates the rapid and accelerating investment in information infrastructure, there is an urgent need to help users find information of interest. Much of the available information exists on the web.

Despite the increasing ubiquity of standards in the development of web pages through the use of technologies such as Really Simple Syndication (RSS), many websites containing important information fail to adhere to such standards. This necessitates the development of automatic tools that read such websites and convert the information contained therein to standard formats that are easily shared and distributed.

This paper will focus on the extraction of key items, namely titles and body texts from web pages of news sites to produce the same news in RSS format, which is a popular format for disseminating news. News sites typically contain many different news topics belonging to different categories, such as politics, sports, economics, and news items are continually added or updated. The extraction process is done via the use of specialized programs which will henceforth be referred to as parsers. Two strategies are possible to build such parsers. The first strategy involves building rule-based parsers, which employ the consistency of style in individual websites. Such parsers are unfortunately site specific and breakdown when the site maintainers change the style of pages. The other approach is to build universal parsers that "guess" where the different elements to be extracted are. The latter approach clearly seems advantageous over the first.

This paper presents a supervised machine learning technique for extracting important elements of news web pages. The technique uses a Support Vector Machine (SVM) classifier which uses strictly structural features to extract the required elements. The technique is employed within Alzoa.com, which is the largest aggregator of Arabic news on the web. Alzoa aggregates news every hour from more than 160 Arabic news sites, many of which do not follow proper DOM tree standards and provide no RSS feeds and a few of which are authored via the editing of static HTML without the use content management systems.

The rest of the paper is organized as follows: Section 2 surveys related work; Section 3 describes the new extraction technique and experimental setup; Section 4 reports results; and Section 5 concludes the paper.

2. Background

Much of the work of automatic content extraction from web pages has focused on the structural analysis of HTML. This has led to the development of a myriad of HTML parsers which programmatically allow the navigation of structures in HTML.

One of the more advanced HTML parsers is the Vision based Page Segmentation (VIPS) which was developed by Microsoft Research (Cai, et al., 2003). VIPS attempts to extract the semantic structure of the web page based on its visual presentation as opposed to a page's DOM representation, because many HTML authors fail to adhere to DOM. Another approach attempts to extract meaningful blocks from HTML by selecting predefined HTML blocks that typically contain useful content such as <TD> and <DIV> and then using the HITS algorithm, which is a graph-based method for identifying hubs and authorities, to compute the value of a block (Kao et al., 2004). Diao et al. (2000) focused on identifying relevant segments in webpages in response to queries as opposed to whole pages. In their work, they segment a web page using its HTML data elements, partitioning the document into four major segments types: paragraphs, tables, lists, and headings. These segments are consequently tagged with attributes such as "content," "description," and "hyperlink" and the tagging is done using a Bayesian classifier. Yi et al. (2003) build a so-called Site Style Tree (SST) to capture the common presentation styles in a website. Consequently, each page is mapped to the SST to determine which portions of a page is "noise" as

Figure 1: Example page from Al-Ahram newspaper

opposed to "content." Wong et al. attempt to heuristically construct the structure into a tree of segments and to discover which tags define the structure of the page. Consequently segments are assigned labels based on the discovered tree structure.

3. Proposed technique

The proposed technique uses a supervised learning to detect if an HTML textual element is a title or an article full-text associated with the news stories based strictly on structural features, which alleviates the use of a page's DOM tree, which often does not adhere properly to standards.

The technique initially extracts all HTML blocks that may contain the desired textual elements, namely elements contained within the tags <DIV>, <TD>, <P>,
, and various headers <H1 … H6>. For each element, certain features are extracted and the element is classified as a title, an article full-text, or other. The technique relies on the following features of textual elements:

1. Length of the text, which is the number of words in the element. Perhaps this can be the number of characters for languages such as Chinese where words are connected. The intuition is that long sequences of text are likely to be body text segments.
2. Percentage of hypertext, which is the percentage of text bound by the <a> tag within the text. Typically side bars have a high percentage of hypertext (such as related articles, etc.), body text has limited hypertext, and titles do not have any hypertext.
3. Percentage of meta-script text, which is the percentage of text bound by <Meta> and <script> tags. Body text and title blocks typically contain limited or no text bound by these tags.
4. Percentage of decoration data, which is the percentage of text bound by <Marquee>, <Input>, <Select>, or <Option> tags. Similar to feature 4, body text and title blocks typically don't contain such tags.
5. Percentage of image text, which is the percentage of text bound by the <Img> tag. Title blocks typically don't contain the <img> tag and body text is likely to contain a limited number of such a tag.
6. Is Date-Copyright, which is a binary feature based the percentage of dateline or copyright words, such as month names, names of days, a.m. and p.m., copyright words, etc. We have collected a dictionary of such words. The following are examples of Date-Copyright elements where dateline and copyright words are

underlined:
الأحد 20 من ربيع الأول 1428هـ 8 ابريل 2007 السنة 131-العدد 43952
6:11AM EDT, Sun 8 Apr 2007
© 2006 BigCharts Inc. All rights reserved. Please see our Terms of Use.
If these words constitute more than 30% of the total number of words in the element, it is considered as a Date-Copyright element. Date-Copyright elements generally don't show in either body text or title blocks.

The SVM classifier was trained using 100 pages from three different categories (hot news, politics, and sports) from 10 different Arabic news sites such as CNN, Reuters, AFP, Al-Jazeera.net, etc. We used the TinySVM implementation of the classifier with a linear kernel. The classifier was tested on 216 randomly selected pages from 10 different categories from 30 different websites, 25 of them are Arabic sites and 5 are English sites. None of the sites from which we extracted training pages was used to extract pages for testing. Some of the websites that were included in the test set were sites that used inconsistent styles and we believe were edited by hand without the use of content management systems. These sites include: http://www.akhbarelyom.org.eg whose style varies greatly from page to page and from day to day.

4. Results

Table 1 shows the accuracy of extraction of titles and body text for the different sources used in testing the automatic extraction technique. Basically, if both title and body text from an article are extracted correctly, then this is considered correct. Otherwise, it is considered incorrect. The results show the effectiveness of the proposed approach with accuracy ranging between 95-100%. Further, extraction from most sources has an accuracy of 100%. All the sources used for testing were local or regional and were the most likely to contain inconsistent formatting.
The proposed technique is effective in extracting titles and body text. The sources for which extraction accuracy is less than 100% generally showed the most variety in HTML formatting.

Source	Accuracy
Akidaty: http://www.algomhuria.net.eg/akidaty	100 %
Akhbar-Alkhaleej: http://www.akhbar-alkhaleej.com	100 %
Akhbar Alyom: http://www.akhbarelyom.org.eg	95%
Al-Akhbar: http://www.akhbar.tn	100%
BAB: http://www.bab.com	100%
Al-Manara: http://almanara.org	98%
Al-Wafd: http://www.alwafd.org	100%
Rose-Al-Yousef: http://www.rosaonline.net	100%
Al-Araby: http://www.al-araby.com	100%
Islam Memo: http://www.islammemo.cc	100%
Islam Online: http://www.islamonline.net	100%

Table 1: Sources used for testing and the extraction accuracy

5. Conclusion

The paper presents a technique for automatically extracting textual elements from news articles that include the headline and the body text. The technique uses a supervised learning technique to perform the task. It uses an SVM classifier that is trained on strictly structural features to identify the desired structural elements. The technique is robust in extracting such elements for previously unseen website designs and websites that don't conform to proper standards.

6. Acknowledgements

The authors would like to thank the Egyptian Ministry of Communication and Information Center of Excellence for Data Mining for partially supporting this work.

7. References

Cai, D., S. Yu, J. R. Wen, and W. Y. Ma. (2003). VIPS: a vision-based page segmentation algorithm. *Microsoft Technical Report* (MSR-TR-2003-79).

Diao, Y., H. Lu, S. Chen, and Z. Tian. (2000). Towards learning based web query processing. *In Proceedings of International Conference on Very Large Databases*, pp. 317-328.

Kao, H. Y., S. H. Lin, J. M. Ho, M. S. Chen. (2004). Mining web informative structures and contents based on entropy analysis. *IEEE Transactions on Knowledge Discovery and Engineering*, January 2004: Vol. 16, No. 1 pp. 41-55.

Yi, L. and B. Liu. (2003). Eliminating noisy information in web pages fro data mining. *In ACM Conf. on Knowledge Discovery and Data Mining (SIGKDD)*.

Wong, W. and A. W. Fu. (2000). Finding structure and characteristics of web documents for classification. *In ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD)*, Dallas, TX., USA.