Content Extraction from Web Pages

Linhai Song[1, 2], Xueqi Cheng[1], Yan Guo[1], Guodong Ding[1]
1. Institute of Computing Technology, Chinese Academy of Sciences
2. Graduate School of the Chinese Academy of Sciences
songlinhai@software.ict.ac.cn, {cxq, guoy}@ict.ac.cn, dingguodong@software.ict.ac.cn

ABSTRACT
Web pages often contain some clutters, such as navigation bars, branding banners, and pop-up advertisements. These noises may distract users from actual content they are really interested in. Content extraction is defined as extraction of useful and relevant parts from web pages. It has many applications ranging from making better access to the web through mobile devices to providing data to web mining algorithms. In this paper, we introduce the content extraction module in our information system. This module is used to extract content from news pages, forums pages and blog pages. Three distinct algorithms are designed according to features of these three kinds of pages. Text-to-link ratio, title finding and page segmentation are core ideas we use to extract content from news pages, blog pages and forum pages separately. In order to further enhance the effect of content extraction, some heuristic rules are summarized in this module according to the fact that web pages we gather are all in Chinese. A demo is also provided to show the performance of our content extraction module.

Categories and Subject Descriptors
I.7.5 [**Document Capture**]: Document analysis, H.3.1 **[Content Analysis and Indexing]**

General Terms
Performance; Design

Keywords
Content extraction, Text-to-link ratio, Title finding, page segmentation.

1. INTRODUCTION
Web pages are often decorated with extraneous information which includes navigation bars, branding banners, JavaScript and advertisements. This kind of information may distract web users from actual content they are really interested in. Content extraction is utilized to remove these noises and gain useful parts from web pages.

There are at least two areas benefiting from technologies of content extraction. Firstly, content extraction can improve the experience of surfing the internet through mobile devices [2] [3]. Secondly, results of content extraction are used to strengthen the performance of web mining algorithms [1].

In our work, we need to upgrade an existing information system which gathers web pages from news sites, forums and blog sites, and runs many web mining applications on these pages. Formerly, html tags are removed and remainders are used as inputs of web mining algorithms. Apparently, there are so many noises in the input that effects of web mining algorithms are impaired. So we need content extraction to solve this problem.

In this paper, we introduce the content extraction module in the new version of our information system. According to requirements of web mining algorithms, we define our extraction targets as articles from news pages, posts from blog pages and posts and replies from forum pages. Ideally, we could find a general algorithm to realize content extraction from all these three kinds of pages. But in order to get better effects, we designed three distinct algorithms based on features of these three kinds of pages. Text-to-link ratio, title finding and page segmentation are key ideas we use to extract content from news pages, blog pages and forum pages separately. Because web pages our system gather are all in Chinese, some heuristic rules are also summarized based on this fact to enhance the performance of content extraction.

2. CHANLLENGES
The requirements from actual applications introduce several unique challenges for our work:

Effectiveness: Data used by web mining algorithms are provided by our content extraction module. If there are too many noises in results, effects of web mining algorithms must be impaired. Past experience with web mining tells us that both text-based precision and recall of content extraction must be above 90%.

Efficiency: The content extraction module can't be the bottleneck of the system it is integrated into. The crawler in our information system can gather 80 pages in one second, so the content extraction module has to keep up with this pace.

Less manual intervention: Some traditional methods use wrappers to extract information from web pages. Wrapper is a kind of extraction rules and can only be used to pages produced by the same template. Wrappers can be got through hand coding or training on training sets. These methods have two obvious drawbacks: 1) when getting an input page, you need to select the right wrapper; 2) and when page formats change, you need to update your wrappers. Considering ease of use, we hope that after starting our information system, there is no more manual intervention. So methods using wrappers are not suitable in this situation.

Better system integration: For better system integration, we need to simplify interfaces of our content extraction module. In the current version, the input of our module is a URL list, and outputs are content our module extracts. A configuration file is also provided, and you can adjust parameters with changes in the actual situation.

Unified Chinese character encoding: Input pages may be encoded in GBK, Unicode, Utf-8, or GB2312. But outputs of our module are required to be encoded in GBK.

Standardized output: According to requirements of some Chinese web mining algorithms, outputs of the content extraction module must have correct paragraphing and indenting information.

3. SULUSIONS

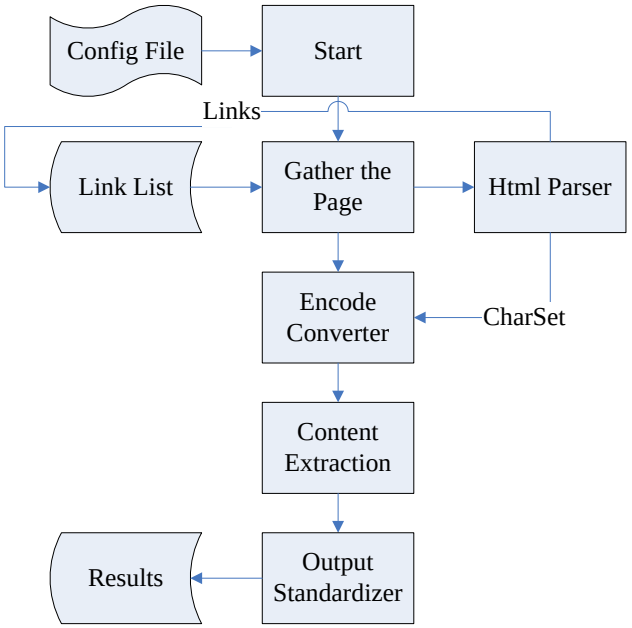

Figure 1. Content extraction from web pages.

In order to make our content extraction module more flexible, we design a configuration file to tell the module which type input pages belong to and other important configuration information. After one page is gathered, an html parser is used to get all links in the gathered page. Links are filtered and only useful links are added to the link list. The charset of the gathered page is also acquired by the html parser, and this information is used to perform transcoding. Content extraction is executed on pages in GBK, and a proper algorithm is chosen according to which type the gathered page belongs to. We use tag information to standardize extraction results, and correct paragraphing and indenting information are added to fulfill to requirements from web mining applications.

3.1 News Pages

The algorithm we use to extract articles from news pages is CoreEx, which is a text-to-link ratio based algorithm proposed by J. Prasad and A. Paepcke [1]. In this algorithm, a formula gives each node on the dom tree a value, and the node with highest is the node carrying the article of the input page.

There are two common features among Chinese news pages: firstly, noises in news pages don't contain full stops in Chinese; secondly, statements in news pages, which may not be dealt with by CoreEx correctly, always contain some fixed words. In the preprocess step, we realize two filters based on these two features to improve the performance of content extraction.

3.2 Blog Pages

Blog pages also have two common features: firstly, each blog page has one text node which indicate the title of the post, and this text node appear just before the post of the blog page; secondly, the text marked by "<title>" and "</title>" also contain the title of the post.

Based on these two observations, we design a two-stage blog post extraction algorithm: firstly, we calculate the similarity between each text node and the text marked by "<title>" and "</title>", and then we select the text node with largest similarity as the text node indicating the title of the post; secondly, we summary some patterns, and use these patterns to find the end of the post. We use all text nodes between the beginning and the end as the post of the input page.

3.3 Forum pages

How we design content extraction algorithm for forum pages is based on the fact that what we want to extract are posts and replies, they are not continuous but separated by some noises.

There are also two steps in this algorithm: firstly, we segment each forum page into several blocks; secondly, we select proper blocks and use text nodes in them as results.

Heuristic rules we use to segment forum pages are only about tag information, and we try our best to keep each block in a smaller size. Each block we select must fulfill these two conditions: the total length of text nodes it has must be longer than a given threshold and each of these text nodes must not follow in some noise patterns we summarize.

4. EXPERIMENTAL RESULTS

You can test our content extraction module at http://searchforum.org.cn/research/demos/slh/. The inputs are a URL and the type of the page you want to test. You can see the extraction result on the same page. Some heuristic rules are summarized for Chinese web pages, so our content extraction module may not work well on web pages in foreign languages.

Some content extraction results our information system produces are also provided. We organize these results according to their types and sites where they are from. You can see these results at http://searchforum.org.cn/research/demos/slh/results.html.

5. REFERENCES

[1] J. Prasad, and A. Paepcke, "CoreEx: Content Extraction from Online News Articles," In CIKM'2008: Proceedings of the 17th ACM Conference on Information and Knowledge Management, Napa Valley, California, USA, pp. 1391-1392..

[2] S. Gupta, G. K. Kaiser, P. Grimm, M. F. Chiang, and J. Starren, "Automating Content Extraction of HTML Documents," World Wide Web, 2005, vol. 8, pp. 179-224.

[3] S. Gupta, G. Kailer, D. Neistadt, and P. Grimm, "DOM-based Content Extraction of HTML Documents," In WWW'03: Proceedings of the 12th International Conference on World Wide Web, Budapest, Hungary, 2003, pp.207-214..