Extracting Pure Text from Web Pages

Helmut POSCH*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
xposch@stuba.sk

Extraction of pure text from web pages is the first step for successful analysis and next processing of web page content. Pure text of web page can be useful for recommendation or search systems. It can be displayed on mobile devices, which haven't enough space on display for extensive information. These days, extraction of pure text from web pages is a nontrivial task. Each web page can have different structure and content. Web page developers insert to their web pages elements, e.g., advertisement, navigation bars, header or footer, which aren't from the information value point of view necessary useful for users.

Our work is based on the approach Content Extraction from Web Pages Based on Chinese Punctuation Number [1]. This approach is based on observation that large number of punctuation is in main content, especially full-stops. Nodes of web page (node is some HTML tag) are evaluated by full-stops count and ratio of ASCII text and anchor text of node content. Method can extract only text in Chinese.

We extended this approach by using punctuation, which is used in Latin languages (as Slovak, English etc.). Because of this, we computed statistics of punctuation occurrence in these languages, to confirm the punctuation location on web pages. Compared to referenced method, we found that intensity of full-stops occurrence in main content isn't so high. The most frequent punctuation char in main content in Latin languages is comma. Because of this, we switched from full-stop to comma in our approach. Our first improvement of referenced method is the usage of punctuation, which mostly isn't in main content. Based on our experiments and statistics, we found, that anchor full-stops and anchor commas meets the requirements. For node evaluation we use the following equation:

$$score = \frac{CS^3 * \frac{CL}{ACL}}{ACFL^2}$$

where *CS* refers to Commas Sum, *CL* is Content Length (of node and without HTML tags), *ACL* is Anchor Content Length and *ACFL* is Anchor Commas and Full-stops Length. Power of parameters indicates importance of the parameter. After evaluation

[*] Supervisor: Michal Kompan, Institute of Informatics and Software Engineering

of all web page parts, any part of the web page, which has score above the average is preliminary extracted as main content.

In proposed approach, the part of webpage represents HTML tag DIV. For DIV evaluation and text extraction, we take only text, which is directly in the DIV (no in child DIVs). In this way of evaluation DIVs we obtain better results.

The second improvement of referenced method is position control of preliminary extracted DIVs in web page tree hierarchy. Key part of this process is direct parent of DIV with best score. If some other preliminary extracted DIV isn't anywhere in direct parent tree hierarchy, then is ruled out of final text extraction. Otherwise content of DIV is extracted as the main content. Process of searching direct parent of best evaluated can fail due to difficult string comparing. In this case, all parts of web page with score above of average are selected as main content. This, second part of our method is helpful, when the web page has a lot of not anchor advertisements or copyrights around main content.

Proposed approach focuses on web pages with some coherent text (news etc.). We expect comparable and better results than referenced method in Chinese. Among other datasets, for experimental testing we will use part of dataset, which has been used by authors of CETR [2] method.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG 1/0646/15.

References

[1] Mingqiu S., Xintao W.: Content Extraction from Web Pages Based on Chinese Punctuation Number. In: Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007, International Conference. pp. 5573 – 5575.

[2] Wenninger T., Hsu W.H., Han J.: CETR: content extraction via tag ratios. In: Proceedings of the 19th international conference on World wide web, 2010, pp. 971-980.