International Journal of Database Theory and Application
Vol.6, No.5 (2013), pp.45-52
http://dx.doi.org/10.14257/ijdta.2013.6.5.05

# A Novel Approach of Calculating Information Entropy in Information Extraction

Rong Li* and Hongbin Wang

*Department of computer*
*Xinzhou Teachers' University, Xinzhou 034000, China*
*lirong_1217@126.com, wanghb@163.com*

***Abstract***

*Noise data of web page is easy to cause the topic drift problem in web information extraction. To improve the accuracy of web information extraction effectively, a novel calculation method of mixing entropy is presented, which can more accurately reflect the topic information of web page. The information block is discussed under the multi-page site environment. The impacts of information within local page and the same information distribution between web pages generated by template are all considered so as to ensure the precision of calculating information entropy. The method is verified by calculating the entropy of information block in information extraction. Compared with other methods, the simulation results indicate that the novel method shows great superiority over other traditional methods in both the accuracy of information entropy calculation and discrimination between topic-related information blocks and topic-unrelated information blocks.*

***Keywords:*** *web hybrid entropy, information extraction, information distribution, feature word*

## 1. Introduction

Current Web page contents have become increasingly diverse, a lot of contents irrelevant to topics, including pictures, scripts, codes, links and etc, have also appeared. These contents lead to effect data mining for web page and the display of web contents on a small screen (*e.g.*, mobile phone). So topic information extraction from Web pages has become a focus for researchers, and the application of information entropy in Web information extraction application has been extensively studied [1-4]. The information entropy theory which is raised by American mathematician, C.E. Shannan, is to measure the uncertainty of information, to select the measuring method of uncertainty degree according to the information source and to determine that information characterizes the uncertainty degree of the information source. For instance, based on a page set, theme documents and related links of web pages were extracted for calculation of information entropy in [5, 6]. Based on the information within page, the degree of correlation between theme information was measured for calculation of information entropy in [7, 8].

In order to extend web sites flexibly and easily, many webmasters use the predefined templates to generate Web pages. Experimental results show that web pages containing templates account for about 43 percent of all web pages. Among different pages in some sites, there is some similar information distribution such as the consistent navigation bar, the same

* The Corresponding Author

ISSN: 2005-4270 IJDTA

advertising links etc. Thus, the impact of website templates on web content is very large. The calculation of Web page information entropy not only needs to consider the information inside page, but also should not ignore the same information distribution among a lot of pages because of the use of template. The distribution of page information among different Web pages was mainly considered and the role of information within page was ignored for the calculation of the information entropy in [5, 6]. On the contrary, The information distribution within page was only considered and the influence of the information (*e.g.*, the similar navigation bar and advertisements in different pages generated by template) was completely ignored for calculating the information entropy in [7, 8]. It leads to the result that the discrimination between topic-related information blocks and topic-unrelated information blocks is not great and the information extraction is greatly affected by parameters.

Therefore, in this paper, we propose a novel method for calculating the information entropy of Web page, hybrid entropy, which discusses information blocks under the multi-page site environment and adds the information within page to improve the accuracy of information entropy. In addition, the new method also takes into account the influence of the similar information of each Web page generated by a Web site template.

## 2. Calculation of Hybrid Information Entropy of Web Page

Crawled pages from Internet are pretreated and may be made standard web pages, Then according to their HTML tags, the standard web pages can be represented as a tree, that is DOM tree structure [9, 10]. Each node in the DOM tree represents a page tag which contains the name information, property, all characters between markers (denoted as innertext). The innertext of Node *N* includes all the characters in the subtree whose root is Node *N*. An innertext of a page root node includes all characters whose tags have being removed.

In the paper, we extract key words or phrases as feature words from the root innertext in the web page DOM tree. As Chinese is different from English, we should first implement the Chinese word segmentation, remove stop words, and extract the corresponding features words, and then calculate the corresponding hybrid entropy for the extracted features words

### 2.1. Information Entropy Theory

The related definitions and properties may be described as follows:

Definition 1. The self-information quantity of the arbitrary random events may be defined as the negative logarithm value of the probability of occurrence. Set the probability of Event $x_i$ to be $p(x_i)$ ,and its definition of self-information quantity may be described in Eq. (1).

$$I(x_i) \equiv -\log p(x_i) \tag{1}$$

Definition 2. On Set *X*, the mathematical expectation of the random variable *I(x)* is defined as the average self-information quantity in Eq. (2).

$$H(X) = E[I(x_i)] = E[3/\log p(x_i)] = \sum_{i=1}^{q} p(x_i) \log p(x_i) \tag{2}$$

The average self-information quantity of Set *X* is also known as the information entropy of Set *X*, referred to as entropy.

Property 1. The information entropy has no negative value, and the smaller the entropy, the better the system stability. While the information entropy of page node

 

being calculated and if the node entropy value is less than a certain threshold, the information contained in the node can be considered as page theme information.

Property 2. The information entropy has additivity. According to the property, the information entropy of web page node can be regarded as the sum of the information entropy of each feature word, which affords a theoretical basis to determine whether the node is related to topic based on the node information entropy.

### 2.2. Calculation for the Information Entropy based on Page Set

The formula of entropy value of each feature word based on page set, *ENS*, is stated in Eq. (3).

$$ENS(term_i) = -(\sum_{i=1}^{n}(w_{ij})log_n w_{ij}), \quad w_{ij} > 0 \; and \; n = |D| \tag{3}$$

We denote Character *D* as a page set and $W_{ij}$ as the normalized word frequency of feature $term_i$ appeared in Page *j* In Eq. (3). The information entropy of feature word is calculated mainly from the page set perspective in Eq. (3). The same feature words, generated by the template, among different pages have the character, that is, the more evenly feature words distributed in more pages, the larger the entropy value of *ENS* is, the lesser contained information is.

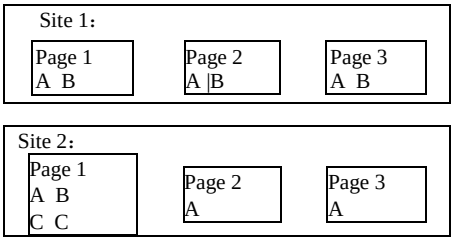


**Figure 1. Feature Word Distribution**

As shown in Figure 1, the feature word distributed proportionately in Site 1 and Site 2 is Term *A*, the feature word only distributed proportionately in Site 1 is Term *B*, and the feature word only in Page 1 of Site 2 is Term *C*. By Eq.(3), we calculate the information entropy *ENS* of each feature as follows.

ENS(A)=-6/12*log6(1/12)=0.69,ENS(B)=-4/12*log6(1/12)=0.46,

ENS(C)=-1/6*log6(1/6)=0.17

As can be seen from the calculation results of *ENS*, the information entropy of Term *A* which appears in each page of both Site 1 and 2 generated by template is larger, and the information quantity is smaller. The information entropy of Term *C* only appearing in Page1 of Site 2 is smaller, and the information quantity is larger. Term *B* appearing in each page of Site 1 generated by template contains less information. However, the amount of information of Term *B* in Site 2 is significantly different from that of *B* generated by template in Site 1. By Eq. (3), the difference can not be distinguished.

### 2.3. Calculation for the Information Entropy based on Page Content

The formula of entropy value of each feature word based on page content, *ENI*, is stated as follows.

$$ENS(term_i) = -(1/Cr_i)log_c(C_i/n) \quad (4)$$

In the above Eq.(4), $Cr_i$ is denoted as the number of occurrences of $term_i$ in page, $C_i$ as the number of occurrences of $term_i$ in node, and $n$ as the total number of occurrences of feature word. The information entropy of feature word is mainly calculated based on the frequency of occurrence of feature word in the page node In Eq. (4). When the feature words appear more frequently, it is closer to the theme information, its information entropy value, *ENI*, should be smaller. Compared with the standard Eq. (2), Eq.(4) changes the factor $C_i/n$ to be $1/Cr_i$, so as to reduce quickly the information entropy of the feature word containing more theme information.

We calculate the corresponding information entropy *ENI* for feature of Page 1 in Site 2 in Figure 1, the results are stated as follows.

*ENI(A)*=-*log4(1/4)=1, *ENI(B)*=-*log4(1/4)=1, *ENI(C)*=-(1/2)*log4(2/4)=0.25

As can be seen from the calculation results of *ENI*, the number of occurrences of Term *A* and *B* are fewer, their information entropy value is bigger and the contained information content is less. Term *C* appears is more frequently, its information entropy value is smaller and the contained information content is bigger. However, Term *B* is not generated by template, not evenly distributed in each page in site and maybe more related to topic. By (4), Term *A* and *B* can not be obviously distinguished.

### 2.4. Calculation for the Hybrid Entropy of Web Page

In summary, the calculation of information entropy individually based on page content only considers the internal information content of page, and ignores the huge effect of Web template. The calculation of information entropy separately based on page set only considers the information content of node from template and ignores the function of information within page. In order to optimize the calculation of information entropy, we propose a new calculation method of information entropy-mixing entropy *ENA*, which adds the information inside pages to improve the accuracy of information entropy, and meanwhile, takes into account the effect of the same information generated by template. The specific calculation formula of mixed entropy for each feature word is stated in Eq.(5).

$$ENA(term_i) = \alpha ENS1(term_i) + (1-\alpha)ENI(term_i) \quad (5)$$

In the above Eq.(5), $ENS1(term_i) = -(\sum_{i=1}^{n}(1/Cr_i)log_n(w_{ij}), w_{ij} > 0, n = |D|$, where Character $D$ is denoted as a page set, $W_{ij}$ as the normalized word frequency of feature $term_i$ appearing in Page $j$, and $Cr_i$ as the frequency of occurrence of $term_i$ in page. And the calculation of information entropy *ENS1* is also considered from page set. The information entropy value *ENS1* of feature word which is generated by template and evenly distributed in many pages is bigger and the contained information is less. Compared with Eq.(3), Eq. (5) changes the factor $w_{ij}$ to be $1/Cr_i$ so as to take into account the effect of feature words of more occurrence frequency and make the information entropy of feature word containing more theme information in the node smaller. The calculation of *ENI* information entropy (see Eq.(4))is based on page content. The frequently occurring term in the whole document shows that the feature word is close to topic information and the *ENI* information entropy is smalle. $\alpha = km_i/n_i$, where $m_i$ is denoted as the page number of the similar advertisement generated by template in site where the feature word $term_i$ locates, and $n_i$ is denoted as the total page number in the site where $term_i$ locates. Because of at least one page in any site, $n_i$ can not be zero. And $k$ is denoted as

 

the adjustment parameter of weight value. Obviously, by definition, a website has only one $\alpha$ value, $\alpha$ value indicates that the page proportion of the same advertisements generated by template in site. Through statistical analysis, each site can be established weight vector.

As can be Observed and analyzed by experiments, many large portal sites use template and generate a lot of advertisement evenly distributed which has little to do with its theme information, its template function is obvious, and $\alpha > 0.5$. In the calculation of mixed entropy, the function of *ENS1* value is higher than *ENI* value, the hybrid information entropy of $term_i$ in the information node that is generated, evenly distributed and irrelative to topic. On the contrary, in non-template generated Web pages (such as a university website), the page of the same advertisement content is less, the template effect is relatively smaller, and $\alpha<0.5$. In the calculation of mixed entropy, the function of *ENI* value is higher than *ENS1* value, the calculation of the hybrid entropy is mainly based on page content. When the same advertising content page generated by template is zero, $\alpha=0$, then Eq.(5) is unnecessary to consider the template effect, so the calculation of information entropy is completely based on the page content.

The corresponding hybrid entropy *ENA* for feature word of Page 1 in Site 2 in Figure.1 is calculated as follows.

*E*NA(A)=6.86, *EN*A(B)=4.65, *E*NA(C)=0.45

As can be seen from the *ENA* calculation results, Term *A* is generated by template in Site 2, the information entropy of Term *A* is bigger, and its contained information quantity is small. However Term *B* is generated by template in Site 2, its information entropy is smaller than that of Term *A*, also its small amount of information entropy takes the middle position. The information entropy of Term *C* is relatively very small, and its amount of information is very prominent.

Compared with Eq.(3), Eq.(5) is based on page set for the information entropy calculation and considers the effect of the feature word which occurs more and is closer to theme information. Hence the hybrid entropy of the feature word which is generated by template and distributed evenly in non-topic information in each page relatively increases. while the hybrid entropy of the feature word, which occurs more and is closer to theme information, relatively reduces. Therefore, the difference between them increases.

Compared with Eq.(4), Eq.(5) is based on page content for the information entropy calculation, and considers the effect of a lot of distributed evenly advertisement content which has little to do with self theme information. So the difference between the hybrid entropy of the feature word containing non-topic information and that of the feature word containing topic information all increases.

Therefore, Eq.(5) makes the difference of feature word between the large amount of information and the small amount of information more obvious and helps to the extraction of theme information.

## 3. Experimental Results

In order to test the validity of information entropy of Web pages, we combine the hybrid information with the information entropy in [4]. (information entropy 1) and that in [7].(information entropy 2). In the experiment, the node entropy takes average entropy of the contained feature word.

After taking the corresponding page, parsing HTML, filtering, and extracting feature word, we can get the feature word information of each node. The number behind the following feature word indicates the occurrence number how many this word appears in this node.

R:{Legend:3; computer:2; function:3; code:3; laboratory:1; security:6; figure:2; vulnerabilities:1; battery:1; memory:1; patent:1; adviser:1; terrace:1; developer:1: step:1; violate:1; filed:1; safety lock:1; user:1; jurisdiction:1; telephone:1; number:1; multimedia message:1; ring:1; cartoon:1}, telephone, number, multimedia message, ring and cartoon are generated by site template among them.

$N_1${Legend:1;computer:2;number:1;security:1; function:1}

$N_2${Legend:1; battery:1; code;3; laboratory:1; jurisdiction:1; memory:1}

$N_3${Legend:1; patent:1; security:4; adviser:1; terrace :1; developer:1; step:1}

$N_4${safety:1; function:2; violat:1; field:1; number:1; safety lock:1; user:1; jurisdiction :1}

$N_5${telephon:1; number:1}

$N_6${multimedia message:1; ring:1; cartoon:1}

The 3 different calculation results of information entropy are shown in Table 1.

**Table 1. Information Entropy of Web Information Block**

| Node number | Information entropy | | |
|---|---|---|---|
| | Information entropy 1 | Information entropy 2 | Hybrid entropy |
| R | 2.07 | 0.41 | 3.11 |
| $N_1$ | 1.08 | 0.32 | 0.99 |
| $N_2$ | 2.12 | 0.26 | 1.19 |
| $N_3$ | 2.02 | 0.31 | 1.10 |
| $N_4$ | 2.52 | 0.33 | 1.40 |
| $N_5$ | 3.56 | 0.46 | 13.37 |
| $N_6$ | 3.56 | 0.51 | 13.37 |

The experimental results show that compared with [4] and [7], the hybrid entropy calculation in the paper causes the topic-related information block and the topic-unrelated information block to have better discrimination, further helps to extract theme information and can improve the precision of information extraction.

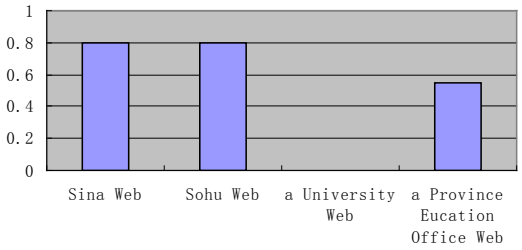


**Figure 2. *α* Values of 4 Web Sites**

Figure 2 shows the comparison of $\alpha$ value from 5 sites. 200 Web pages from each Web site are arbitrarily extracted for calculating $\alpha$ value if the inherent page information such as copyright information being ignored.

As can be seen from Figure 2, $\alpha$ value of a larger portal site is larger, because such web sites generate more pages which have a lot of similar ,having little to do with theme, advertisement content generated by template. But $\alpha$ value of provincial education network is smaller, because the page proportion of which such Web sites generate the

 

similar advertisement page is less than that of portal sites. And $\alpha$ value of college sites is zero, because such Web sites don't generate similar advertisement pages by template.

## 4. Conclusions

In the paper, a calculation method for the information entropy of Web page is proposed. Compared with the existing research method, this method not only considers the impact of widely used website templates, but also considers the impact of the information within page. This method puts the information blocks under a Web site environment for discussion, through adding the page content to improve the accuracy of information entropy. In addition, the method simultaneously takes into account the effect of the same information generated by template in each page. The Web page information entropy by using the method can obviously improve the discrimination between the topic-related information node and the topic-unrelated information node, and this helps to extract theme information and improves the precision of information extraction. The next step of work is to apply the calculating method of Web information entropy to the information extraction from Web page.

## Acknowledgements

We would like to thank to the reviewers for their helpful comments. This work was financially supported by the Natural Science Foundation of China (#1072166), the Higher School Science and Technology Development Project in Shanxi Province of China, and the key discipline construction project of Xinzhou Teachers' University (# ZDXK201204).

## References

[1] A. McCallum, D. Freitag and F. Pereira, "Maximum entropy markov models for information extraction and segmentation", Proceedings of the Seventeenth International Conference on Machine Learning, Stanford, California, **(2000)**,pp. 591-598.
[2] A. L. Berger, S. A. Della-Pietra and V. J. Della-Pietra, "A maximum entropy approach to natural language processing. Computational Linguistics, vol. 22, no. 1, **(1996)**, pp. 39-71.
[3] S. H. Lin and J. M. Ho, "Discovering informative content blocks from Web documents", Proceedings of the 8th ACM SIGKDD International Conference on Knowledge discovery and data mining, **(2002)**, pp. 588-593.
[4] H. Y. Kao, J. M. Ho and M. S. Chen, "WISDOM: Web intrapage informative structure mining based on document object model", IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 5, **(2005)**, pp. 614-630.
[5] H. Y. Kao, J. M. Ho and M. S. Chen, "Entropy-based link analysis for mining web informative structures", Proceedings of the eleventh international conference on Information and knowledge management, McLean, Virginia, USA, **(2002)**, pp. 574-581.
[6] H.-C. Zhu and Y. Tang, "An Entropy-Based Approach for News Article Extraction from Web Page New Technology of Library and Information", (in Chinese), vol. 2, no. 4, **(2007)**, pp. 48-51.
[7] Z.-P. He and X.-Z. Xu, "Extracting topic information of Web page based on entropy computer engineering and applications", (in Chinese), vol. 43, no. 4, **(2007)**, pp. 164-166.
[8] H. Li and J. Shen, "Reviews Discovery and Opinions Extraction Based on Segment and Entropy Application", Research of Computers, (in Chinese), vol. 24, no. 2, **(2007)**, pp. 269-272.
[9] S. Gupta, G. Kaiser and D. Neistadt, "DOM-based content extraction of HTML documents", 12th International World Wide Web Conference, Budapest, Hungary, **(2003)** May.
[10] A. F. R. Rahman, H. Alam and R. Hartono, "Content extraction from HTML documents", 1st International Workshop on Web Document Analysis, Seattle, Washington, USA, **(2001)** September 8.
[11] K. Fujinami, "A Case Study on Information Presentation to Increase Awareness of Walking Exercise in Everyday Life", IJSH, vol. 4, no. 4, **(2010)** October, pp. 11-26.
[12] K. Mitra, D. Bhattacharyya, and T.-H. Kim, "Urban Computing and Information Management System Using Mobile Phones in Wireless Sensor Network", IJCA, vol. 3, no. 1, **(2010)** March, pp. 17-26.

[13] S. Kianpisheh, S. Jalili and N. Moghadam Charkari, "Predicting Job Wait Time in Grid Environment by Applying Machine Learning Methods on Historical Information", IJGDC, vol. 5, no. 3, **(2012)** September, pp. 11-22.
[14] L. Meng, J. Gu and Z. Zhou, "A New Model of Information Content Based on Concept's Topology for Measuring Semantic Similarity in WordNet", IJGDC, vol. 5, no. 3, **(2012)** September, pp. 81-94.

## Authors

**Rong Li** received her Master's degree in School of Computer and Information Technology from Shanxi University in Taiyuan, China in 2007. She is currently an associate professor in the Department of Computer at Xinzhou Teachers' University in Xinzhou, China. Her main research interests include intelligent information processing and machine learning. She has published over 20 research papers in scholarly journals and international conferences in the above research areas.

**WANG Hong-Bin** received his M.S. and Ph.D. degrees in College of Information Engineering from Taiyuan University of Technology in Taiyuan, China. He is currently a professor in the Department of Computer at Xinzhou Teachers' University in Xinzhou, China. His main research interests include intelligent information processing and information security. He has published over 30 research papers in scholarly journals and international conferences in the above research areas.

 