Original Article

The Feasibility of Using Classification and Identification Techniques to Auto-Assess the Quality of Health Information on the Web

Polun Chang[1], Fan-Pin Huang[1,2], Min-Ling Lai[1,3]

Institute of Biomedical Informatics, National Yang-Ming Univ.[1],
Armed Forces Taichung General Hospital[2],
Dept. of Information Technology, Taipei City Government[3]

Abstract

Objective: An automatic detection tool was created for examining health-related webpage quality we went further by examining its feasibility and performance. **Methods:** We developed an automatic detection system to auto-assess the authorship quality indicator of an health-related information webpage for governmental websites in Taiwan. The system was integrated with the Chinese word segmentation system developed by the Academia Sinica in Taiwan and the SVM^{light}, which serve as an SVM (Support Vector Machine) Classifiers and a method of information extraction and identification. The system was coded in Visual Basic 6.0, using SQL 2000. **Results:** We developed the first Chinese automatic webpage classification and information identifier to evaluate the quality of web information. The sensitivity and specificity of the classifier on the training set of webpages were both as high as 100% and only one health webpage in the test set was misclassified, due to the fact that it contained both health and non-health information content. The sensitivity of our authorship identifier is 75.3% ,with a specificity of 87.9%. **Conclusion:** The technical feasibility of auto-assessment for the quality of health information on the web is acceptable. Although it is not sufficient to assure the total quality of web contents, it is good enough to be used to support the entire quality assurance program. ***(Journal of Korean Society of Medical Informatics 15-3, 247-254, 2009)***

Key words: Health Information Quality, Automatic Assessment, Categorization, Information Extraction

Received for review: September 29, 2009; **Accepted for publication:** September 29, 2009

Corresponding Author: Polun Chang, Associate Professor, Institute of BioMedical Informatics, National Yang-Ming University, Rm. 520, Library and Information Building, #155, Li-Nong St., Sec. 2, Taipei 11221, Taiwan/ROC
Tel: +886-2-2826-7238, **Fax:** +886-2-2820-2508, **E-mail:** polun@ym.edu.tw

DOI:10.4258/jksmi.2009.15.3.247

I. Introduction

The search of better health information has been a very important issue for people to take care of their health well-being today. The latest Pew Internet Project survey estimated that between 75% and 80% of internet users had looked online for health information in 2009 and 75% of e-patients with a chronic condition said their last health search affected a key decision about how to treat an illness or condition[1]. Health information includes information for staying well, preventing and managing disease, and making other decisions related to health and health care[2]. Many users would like to query better information, to be better prepared when meeting the doctor, or to search for support, alternative answers or reassurance[3]. The Internet is changing how people give and receive health information and health care.

The health information on the net appears to be a very valuable but worrisome source. Information on the net is much easy to access but hard to assure its quality. Patients, health care professionals and administrators, researchers, those who create or sell health products or services, and other stakeholders, must join together to create a more safe environment and enhance the value of the Internet for meeting health care needs[4].

Different organizations have proposed various models for quality guiding and to assist the common people to distinguish quality internet information from the bad. Hi-Ethics proposes self regulation. Some governments also take an active role for regulation[2][5]. The HON Foundation, a Non-Governmental Organization and well known internationally for its pioneering and significant work in pursuing the health information ethics and quality, has promoted the Health on the Net Foundation Code of Conduct (HONcode) for medical and health Websites for long[2][5][6]. It is believed to be the most widely used evaluation criteria for health related websites. The cores of these models can be categories into three groups: the use (1) codes of conduct, (2) third-party certification, and (3) tool-based evaluation[6].

From the perspective of economy and efficiency, a tool-based, automatic approach appears very attractive. Some researchers have developed automatic detection tools to evaluate webpage like using HTML document, the components was: text, comment, simple tags, and ending tags, or created the actual location of a candidate line detected by analyzing the Web page DOM tree[7][8]. In this study, we developed a crawler to integrate with artificial intelligence techniques to implement an automatic identification of health related information on the webs and to assess the feasibility of automatics evaluation for a very basic quality issue of web health information, which is the authorship, through information extraction.

II. Materials and Methods

Due to the stability and better maintenance quality of web information contents, we used the governmental websites in Taiwan in 2008 as our sample websites for demonstration. The systems were coded in Visual Basic 6.0 and used a SQL 2000 database. A segmentation module from the Taiwan Academia Sinica[9] and A SVM (Support Vector Machine)[10] Classification module developed by Thorsten Joachims[11] were used as system components.

Two main techniques were used to develop the cores of systems: text categorization technique, which was used to distinguish health-related websites from non-health ones, and information extraction technique, which was used to parse the key quality content information for further examination, as shown in Figure 1. Sample websites, including both health and non-health ones were collected and separated into Testing and Training groups to develop a health website classifier. The classifier was developed through a process of segmenting, feature parsing, indexing and testing for accuracy fine-tuning. The feature parsing only kept meaningful and important information contents. The indexing score was

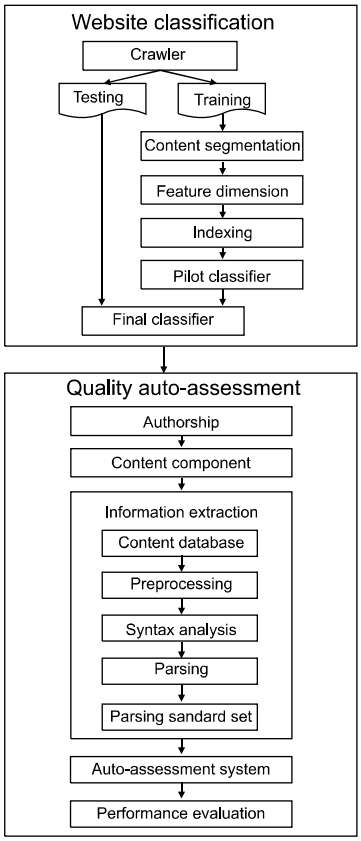

Figure 1. Process of system operations which consist two key components: classifier to auto-identify health websites and extractor to identify quality contents for assessment.

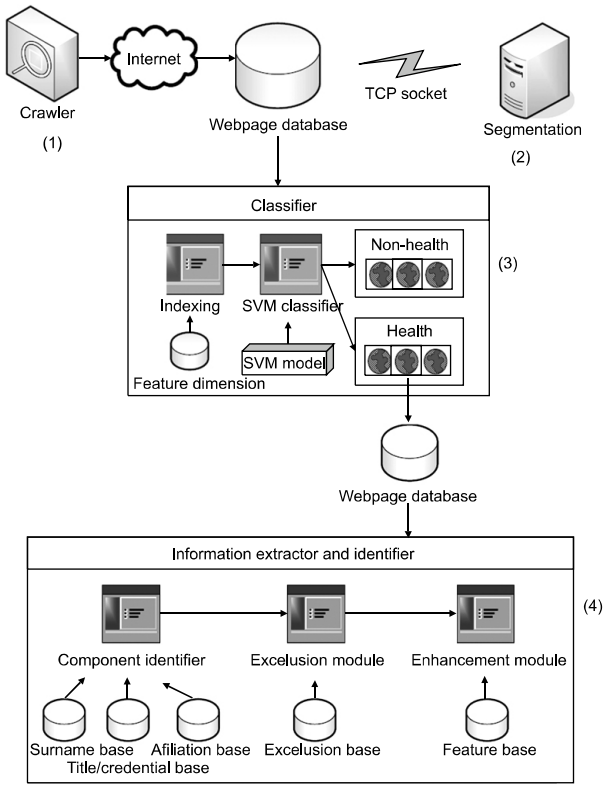

Figure 2. The system structures.

calculated using the TF-IDF approach[12]. The classifier was built using the SVM method. The fine-tuning of final classifier was based on the performance of sensitivity and specificity of pilot ones, using experts' judgment as gold standards. The information extractor to identify the authorship components consists of a database of standard set which lays out the possible layout of authorship components.

The systems, as shown in Figure 2, were composed of 4 key parts: (1) a crawler module to automatically collect the homepage contents of all governmental websites and saved to a local database, (2) a Chinese word segmentation module[13] to parse homepage contents for categorization (3) a SVM^{light} machine-learning engine module[11] to classify and index the health and non-health websites in which the health one was reserved for further quality evaluation, and (4) a quality examination module using the authorship as example, which was parsed into the components of name of authors, credentials, and affiliated organizations for assessment.

The crawler module was design to collect governmental websites in a "www.x.gov.tw" format, in which the x was determined to be any combination of 1-7 a-z alphabet characters in random order after examining the naming of sample governmental websites. The found but non-Chinese websites were excluded because their intended readers were foreign people. Inside the Information Extraction module, the Surname database mainly consisted of the most common family name in Chinese society; the title and credentials database was collected from the list collected by the Taichung Veterans General Hospital[14] and from the sample WebPages; and the affiliation organization list was collected from the National Health Insurance Bureau[15], which is our national

insurance payer and maintain the most current health institutes and organizations. The exclusion module is used to exclude some terms which are similar to the keywords but have different meaning implication and the enhancement module will highlight the most likely terms when there are more than one group of authorship found simultaneously.

III. Results

1. The systems

The screenshot of the crawler is shown in Figure 3. Once the user presses the start button (1), the crawler will start to search through the internet for governmental websites (2) with their first layer of sub-websites (3) and save the web contents of websites into database (4). The crawler will stop when use pressed the stop button (5).

The snapshot of classifier is shown in Figure 4. The system retrieved web contents from a website (1) and parsed to keep body contents (2) in order for segmentation (3). The result of classifying into either health or non-health websites was saved and displayed at the bottom (4).

The snapshot of authorship identifier is shown in Figure 5. The assessment result with statistics were

Figure 3. The screenshot of crawler.

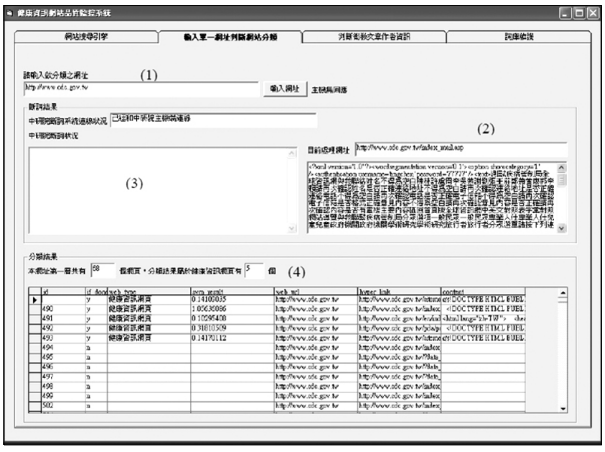

Figure 4. The snapshot of classifier.

Figure 5. The snapshot of authorship identification.

Figure 6. The snapshot of web contents after auto-assessment.

saved and displayed at the bottom of system. Users could examine and browse the web contents by taping the link, which was shown in Figure 6. The terms of corpus could be maintained and examined in the system too, as shown in Figure 7.

2. The performance of health website classifier

There were 520 governmental websites as a total, and among these, 45 were health-related. A total of 468 were used as the training set for developing the classifier, and among these, 40 were health-related.

The content of homepages of all 468 websites was segmented and there were 23,058 terms obtained before refinement. Four types of terms were first reserved: Noun, Action Intransitive Verb, Status Intransitive Verb and Adjective. Then, based on the document frequency approach, those occurred only once in webpage database, such as sidewalk and metropolitan area, were removed and the number of the set of terms left at this stage were 6,321. Next, a manual approach was used to examine these 6,321 terms and another 346 terms, such as Windows and street names, were regarded as non-related and removed. Therefore, combined with the 40 health and 428 non-health WebPages with this final set of 5,975 terms, the classifier using the SVM^{light} engine was trained.

Figure 7. The snapshot of corpus maintenance.

The sensitivity and specificity, to represent the accuracy, of the classifier on the training set of WebPages were both as high as 100%. The corresponding values for the classifier on the testing set were 98% and 100%.

3. The performance of authorship auto-assessment identifier

Contents of health information and education materials from 2,114 health-related URLs on health education, out of 26 health websites were collected for developing the auto-assessment of authorship quality. Sources from 1,901 URLs, randomly selected, were used for training the system and others, as the test set. Examined manually, 810 of 1,901 training URLs and 130 out of 213 test URLs did not contain any author information. The test set appeared to have worse authorship quality. It is interesting to see that as low as 55.5% of governmental URLs on health education contained the authorship information.

With the similar process used to build the term set for the classifier, we built five set of core terms used to identify components of authorship and to make adjust-

Table 1. The outcome of auto-assessing the authorship quality (N=213)

Auto-assessing \ Standard	Containing authorship				Not containing authorship	Total
	Completely correct	Partially correct	Incorrect	Total		
Containing authorship	61	5	3	69	16	85
Not containing authorship		12			116	128
Total		81			132	213

ment. The final sizes of the term sets for surname, title, affiliation, exclusion and enhancement were 484, 33, 22658, 2098 and 17 respectively. Therefore the final size of the authorship identification corpus was 25,290.

The outcomes of using the final corpus to auto-assess the authorship quality in the test set are shown in Table 1. It can be seen that only 69, out of 81, URLs containing the author information were identified but only 61 were completely correctly identified in terms of name, title and affiliation. The sensitivity of our authorship identifier is 75.3% with the specificity, 87.9%.

IV. Discussion

This is the first study to develop an automatic Chinese webpage classification and assessment tool for the web information quality and many interesting lessens can be learned though tools for English contents have been done[8)16)17)]. This challenge comes from the difficulty of segmenting the Chinese sentences and paragraphs into smaller units for language processing. Each Chinese character follows the previous one and is followed by the next one closely without a space which is used in English and many other languages to separate word from word.

The fundamental work for this study is the Chinese Word Segmentation System which has a as high as 95%-96% of accuracy[13)]. Though it is not 100% perfect, but the performance of webpage classification based on the results, combined with a SVM engine, was very satisfying. For the training set, both sensitivity and specificity are 100%. For the test set, the sensitivity is 98% due to one health webpage was mistakenly classified into non-health one. This error was mainly caused by the fact that that webpage happened to contain both health education contents and lots of administrative documentation materials. Too much non-health materials just diluted the features of health-related information of that webpage.

The combination of both health and non-health information in the same webpage appears to bring up an interesting issue of how to define a "health webpage," or a "health website" which can be then targeted for quality assessment. Though the model of building a health portal for the public could resolve this difficulty and all information inside that portal could be regarded as health related, there are still many people searching the web information through search engines using keywords. A way to distinguish the health related from the non-health might still be useful. The metadata model used in the Australian HealthInsite might be helpful[18)]. They require all providers of health information to use tags to represent the features of web information to make searching and management easier. Therefore, we believe that the satisfying classification of webpage in the auto-assessment of quality should be technically achievable.

The challenge of auto-assessing the quality of health web information might come from two factors: the feature of quality indicator and the completeness of terms in corpus. In this study, we used the authorship indicator as example. However it is only one of many quality indicators of health web information and more technically structured. Kim et al. obtained a list of 165 various quality indicators and categorized those into 13 groups, such as the content, design, authorship/financial support/owner, etc[19)]. Eysenbach et al. further categorized 86 quality indicators from literature review into 5 groups: technical nature, design, readability, accuracy and completeness[20)]. Among these groups, only technical indicators could be auto-assessed by machines but still represent a significant proportion of indicators. In Taiwan, we used as high as 50 quality indicators for our National Quality Health Website Awards program and among them, as high as 71%, in terms of number, were technical in nature.

There are only three components to define the authorship: author name, title and affiliation. A quality authorship of a webpage is required to have these three components at the same time. In our study, we built a

corpus database for each component. The largest size is 22,658 for the organization and the smallest size, 33 for the title. To improve identification performance we also built two subsets of corpus: exclusion set and feature set.

Though there are only three components to define the authorship and we had built a corpus to extract key information from the web contents, we didn't achieve a high accuracy rate yet. In this study, the sensitivity was 75.3% and we found the reason came from there was no specific term in the corpus. Therefore, we believe the sensitivity could be improved by keeping adding new terms into corpus. The actually challenge might come from the specificity, 87.9%, in which many webs without authorship were inaccurately identified as containing the information. The reason was mainly caused by the system to identify the name, title and organization information from the web contents which are not the real component of authorship, but the content itself. Therefore, there should be an extra rule to distinguish the authorship information from the body context. Once more, the Australian metadata approach appears to be a good solution that the provider of information should clearly include the authorship information in the tags.

In this study, we used the governmental websites as our samples which might overestimate the performance of auto-assessment techniques. The governmental websites tended to be more structuralized and well maintained compared to those from unmanageable internet, which composed of commercial and noncommercial, well-funded and ill-funded, live and dead websites.

Therefore, the model of building a health portal, which can be managed by a creditable web master such as a governmental institute, for the general public might be a more effective way to control the quality of web information. And through the internal publication guidelines and quality assurance mechanism, the outcome might be better than the model of auto-assessment.

However, the auto-assessment approach is still worth of pursuing because the majority of quality indicators are still technical in nature. We could build systems to assure the quality of 70% of total indicators and focus the more precious manpower resources on manually examining the other indicators such like readability, completeness and design.

Therefore, in this study we concluded that the technical feasibility of auto-assessment for the quality of health information on the web is very acceptable. Though it cannot be used to assure the total quality of web contents, it is good enough to be used to support the entire quality assurance program. Moreover, we recommend that it would be an excellent combination if the auto-assessment techniques could come with the design of metadata tags for each webpage.

REFERENCES

1. Fox S. The Engaged E-patient Population. Pew Internet & American Life Project August 26, 2008. Available at: http://www.pewinternet.org/Reports/2008/The-Engaged-Epatient-Population.aspx. Accessed on August 18, 2009.
2. eHealth Code of Ethics. Available at: http://www.hi-europe.co.uk/files/2000/ehealth_code.pdf. Accessed on August 18, 2009.
3. Rozmovits L, Ziebland S. What do patients with prostate or breast cancer want from an internet site? A qualitative study of information needs. Patient Education and Counselling 2004;53:57-64.
4. Gaudinat A, Ruch P, Joubert M, et al. Health search engine with e-document analysis for reliable search results. International Journal of Medical Informatics 2006;75(1):73-85.
5. Kemper DW. Hi-Ethics: Tough principles for earning consumer trust. URAC/Internet Healthcare Coalition, 2001.
6. HON code. Available at: http://www.hon.ch/HON-code/. Accessed on August 18, 2009.
7. Wang Y, Liu Z. Automatic detecting indicators for quality of health information on the web. International Journal of Medical Informatics 2007;76(8):575-582.
8. Wang Y, Richard R. Rule-based automatic criteria detection for assessing quality of online health information. The International Conference Addressing Information Technology and Communications in

Health (ITCH) 2007;15-18.
9. The Natual Language Processing Lab, National Taiwan University, Taiwan. Available at: http://nlg.csie.ntu.edu.tw/advisor.html. Accessed on September 25, 2009.
10. Support Vector Machine. Available at: http://en.wikipedia.org/wiki/Support_vector_machine. Accessed on Septermber 25, 2009.
11. SVM^{light} (2006) Available at: http://svmlight.joachims.org/. Accessed on August 18, 2009.
12. TF-IDF. Available at: http://en.wikipedia.org/wiki/ Tf-idf. Accessed on September 25, 2009.
13. The Chinese Word Segmentation System. Available at: http://ckipsvr.iis.sinica.edu.tw/. Accessed on September 25, 2009.
14. Taichung Veterans General Hospital. Available at: http://www.vghtc.gov.tw/. Accessed on Septermber 25, 2009.
15. The Taiwan National Health Insurance Bureau. Available at: http://www.nhi.gov.tw/. Accessed on September 25, 2009.
16. Price SL, Hersh WR. filtering web pages for quality indicators: an empirical approach to finding high quality consumer health information on the world wide web. Proceeding of AMIA Symposium 1999;911-915.
17. Wang Y, Liu Z. Automatic detecting indicators for quality of health information on the web. International Journal of Medical Informatics 2007;76(8):575-582.
18. A Quality Health Information Gateway for Australian. Available at: http://www.healthinsite.gov.au/. Accessed on September 25, 2009.
19. Kim P, Eng TR, Deering MJ, et al. Published criteria for evaluating health related websites: review. BMJ 1999;318:647-649.
20. Eysenbach G, Powell J, Kuss O, et al. Empirical studies assessing the quality of health information for consumers on the world wide web: a system review. JAMA 2002;287(20):2691-2700.