Rule-based automatic criteria detection for assessing quality of online health information

Wang, Yunli; Richard, R.

Publisher's version / Version de l'éditeur:

International Conference Addressing Information Technology and Communications in Health (ITCH) [Proceedings], 2007

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC·CNRC

Rule-based Automatic Criteria Detection for Assessing Quality of Online Health Information *

Wang, Y., and Richard, R.
February 2007

* Published at the International Conference Addressing Information Technology and Communications in Health (ITCH). February 15 - 18, 2007. Victoria, B.C., Canada. NRC 48803.

Canada

Rule-based Automatic Criteria Detection for Assessing Quality of Online Health Information

Yunli Wang, PhD, Rene Richard
Institute for Information Technology, National Research Council
46 Dineen Drive, Fredericton, NB, E3B 9W4, Canada
email:yunli.wang@nrc-cnrc.gc.ca

Keywords: quality assessment, health education standards, internet, quality control, metadata

Abstract

Automatically assessing the quality of health related Web pages is an emerging method for assisting consumers in evaluating online health information. We propose a rule-based method of detecting technical criteria for automatic quality assessment in this paper. Firstly, we defined corresponding measurable indicators for each criterion with the indicator value and expected location. Then candidate lines that may contain indicators are extracted by matching the indicator value with the content of a Web page. The actual location of a candidate line is detected by analyzing the Web page DOM tree. The expression pattern of each candidate line is identified by regular expressions. Each candidate line is classified into a criterion according to rules for matching location and expression patterns. The occurrences of criteria on a Web page are summarized based on the results of line classification. The performance of this rule-based criteria detection method is tested on two data sets. It is also compared with a direct criteria detection method. The results show that the overall accuracy of the rule-based method is higher than that of the direct detection method. Some criteria, such as author's name, author's credential and author's affiliation, which were difficult to detect using the direct detection method, can be effectively detected based on location and expression patterns. The approach of rule-based detecting criteria for assessing the quality of health Web pages is effective. Automatic detection of technical criteria is complementary to the evaluation of content quality, and it can contribute in assessing the comprehensive quality of health related Web sites.

1. Introduction

Search for health information is one of the most common tasks performed by Internet users [1-2]. The Pew reported in 2005 that 79% Americans with Internet access have used the Web to get health or medical information [3]. However, accuracy and completeness of health or medical information are common concerns among Internet users. Evaluating the quality of health information on the Web is particularly challenging and important. Many initiatives to regulate the quality and ethical standards for health information have been developed [4] and some of them are widely adopted, such as HONcode proposed by Health on the Net Foundation [5], OMNI (Organising Medical Networked Information) [6], and DISCERN [7]. Automatically assessing the quality of health Web pages is an emerging method for assisting consumers to evaluate online health information. Eysenbach et al. proposed a metadata based automatic downstream filtering [8], which is supported by metadata elements developed by MedCircle project. However, most Web pages do not contain metadata elements that could be used to evaluate quality, and manually editing metadata is the most common method to obtain metadata [9]. Price and Hersh made the first effort of detecting criteria by analyzing the content of Web pages [10]. Nevertheless, their results are too preliminary to answer the question of whether criteria can be automatically detected. Recently, Griffiths et al. proposed an automated quality assessment procedure to rank depression websites according to their evidence-based quality [11]. Although their method is promising, applying the method to any other medical domain requires generating training data sets, which is very time-consuming.

We proposed an automatic method for detecting indicators for technical criteria of online health information [12]. The method detects criteria by matching the content of Web pages with indicators, namely direct detection. The average precision and recall of the detection program can reach 98% and 93% respectively. Although this method is effective for most technical criteria, detection accuracy is low for some criteria such as author's name, credential and affiliation. In this paper, we describe the improvement of our method by using rule-based line classification and analyzing the structure as well as the content of Web pages. The results show that the overall detection accuracy of the rule-based method is higher than that of direct detection. It also shows a great increase in the detection accuracy for criteria that were difficult to detect using direct detection method.

2. Methods

Many organizations and institutes have published criteria for assessing quality of health information online. They can be classified into technical criteria, design, readability, accuracy and completeness [13], which represent different aspects of

quality. Technical criteria, design and readability are domain-independent criteria, while accuracy and completeness are domain-dependent criteria. Since we want to develop a domain-independent tool, we focus on technical criteria only in this study. Detecting criteria has three steps: defining measurable indicators, detecting indicators and detecting criteria. The first two steps will be discussed in section 2.1 and 2.2. The third step is the same as the previous study, please refer [12] for details.

2.1. Defining measurable indicators

In the previous work, we chose 18 criteria: *author's name, author's credentials, author's affiliation, reference provided, copyright notice, date of creation, date of last update, disclosure of editorial review process, disclosure of advertising policy, disclaimer, statement of purpose, privacy protecting, disclosure of sponsorship, disclosure of ownership, internal search engine present, feedback mechanism, site map and payment information*[12]. For each criterion, we defined corresponding measurable indicators with the attributes of criterion name, value and location. The indicator value is the symbolic representation of an indicator, and the location defines the HTML tag in which the indicator value may appear. Some criteria such as *copyright* and *privacy policy* can be reliably detected through matching the indicator values and locations with the content of Web pages. However, the detection accuracies of some criteria, such as *author's name, credentials and affiliation,* are much lower. Based on these observations, we found the method is the most effective for indicators in which the indicator values do not have much semantic or contextual ambiguity, for example, using "copyright" as the indicator value for *copyright*. Also, it is less effective for those indicators, in which the indicator values have some semantic and/or contextual ambiguity, such as using email as the indicator value for *feedback mechanism*. The method is the least effective for indicators that do not have direct indicator values, for instance, using "written by" as the indirect indicator value for *author's name*.

Realizing that relying on indicator values may not be reliable, we started to look at the possibility of detecting an indicator from where the indicator typically appears on a Web page. A typical Web page contains many information blocks. Beside the main content block, other blocks exist such as navigation panels, copyright, privacy notices and advertisements. Although these information blocks are represented in various formats, in general they can be combined into three sections: a top section, a main content section and a bottom section. Indicators are found in one of these three sections. The Web page section where a particular indicator may appear could be a valuable clue for detecting the indicator. Therefore, we describe an indicator from the aspects of criterion represented, indicator value, and location. An indicator value is a phrase that the criterion may be represented. We manually collected these indicator values with high occurrence frequencies from a large amount of Web pages for each criterion. The location is the section where the indicator value is supposed to appear on a Web page. For instance, the indicator value and location of an indicator for criterion *copyright* is "copyright", and "bottom section".

2.2. Detecting indicators

The process of detecting indicators is illustrated in figure 1. There are four main steps for detecting indicators: 1) obtaining candidate lines; 2) detecting the location of candidate lines; 3) detecting expression patterns of candidate lines; 4) rule-based line classification.

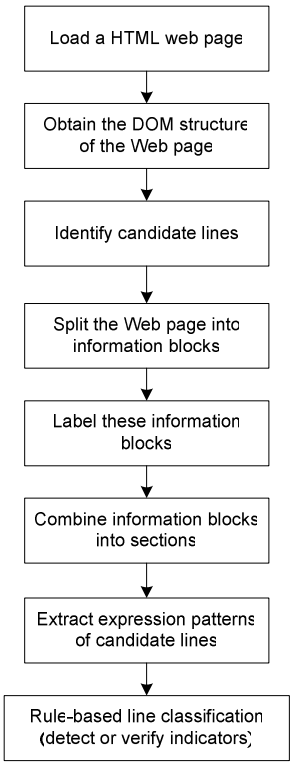

Figure 1 – Detecting indicators

2.2.1. Obtaining candidate lines

Obtaining the candidate lines that may contain indicators is a preprocessing step for detection. After a user submits a query to a search engine, many Web pages are retrieved. The DOM (Document Object Model) trees corresponding to these HTML Web pages are loaded using Cobra [14]. DOM provides a hierarchical structure for every Web page. HTML tags are internal nodes, and the detailed texts, images or hyperlinks are the leaf nodes. Each node of a DOM tree is represented by its line number, value and path. The line number of the node indicates the occurrence sequence of each node. The value of the node is the HTML tags or element content. The path of the node indicates the location of the node in the whole DOM tree. A tree traversal algorithm scans the Web pages and detects candidate lines by matching indicator values with the node values. For instance, a candidate line identified from nodes of a DOM tree is represented as "372", "copyright" and "#document:HTML:BODY:TABLE:TR:TD:TABLE:TR:TD:TABLE:TR:TD:P:A:#text:", corresponding to line number,

node value and node path respectively. The matched part of a specific node value is clustered and replaced with the general cluster name. For example, either the node value "copyright" or "all rights reserved" is represented as ":Copyright:".

2.2.2. Detecting the location of candidate lines

In general, we consider Web pages consisting of a top section, a main content section and a bottom section. Usually *search engine* appears at the top section; *author's name, author's credentials and updated date* are included in the main content section; *copyright,* and *privacy* appear at the bottom section. We notice that some criteria trend to appear at the top or bottom region of the main content section, such as *author's name,* so we further segment the main content section into top, middle and bottom regions. Although a DOM tree is sufficient to represent the layout or presentation style of a HTML page, its granularity is too fine to indicate where an in dictator may appear. Therefore, we performed HTML page segmentation to transform the fine-grained DOM tree into coarse-grained sections. Taking a DOM tree from a web page as the example, the identification process is showed in figure 2. We have 9 candidate lines that share the common upstream path "#document:HTML:BODY:TABLE:TR:TD:TABLE:TR:TD:TABLE:TR:TD:", but they all have their own downstream paths. These lines are represented as bolded rectangles in the DOM tree. We notice that copyright and privacy notice are usually represented as repetitive structures in a DOM tree, and their candidate lines share the same node path. Therefore, we split the DOM tree into a few information blocks, in each of which every node has the same path and expected section type. In figure 2, we identify 6 information blocks from 9 candidate lines. These information blocks are represented as rectangles with an index number in brackets. Each information block is labeled as the expected section of candidate lines. For instance, block 5, containing copyright and privacy is labeled as the bottom section because copyright and privacy is expected to appear at the bottom section of a Web page. The shaded boxes are nodes in the main content section, while other boxes are at the top or bottom sections. The boundaries of the main content section are detected through reliable indicators, which achieved high detection accuracies in the pervious study [12], such as copyright. These reliable indicators can be used to combine these information blocks into sections. For example, we know that "copyright" always appears at the bottom of a Web page so we merge block 5 and 6 into the bottom section. The identified HTML page sections are represented as triangles with section numbers inside in figure 2. After the main content section is identified, top, middle, and bottom regions of the main content section are determined through line numbers. If the line number of a candidate line falls into the range for the top region, it is considered in the top regions.

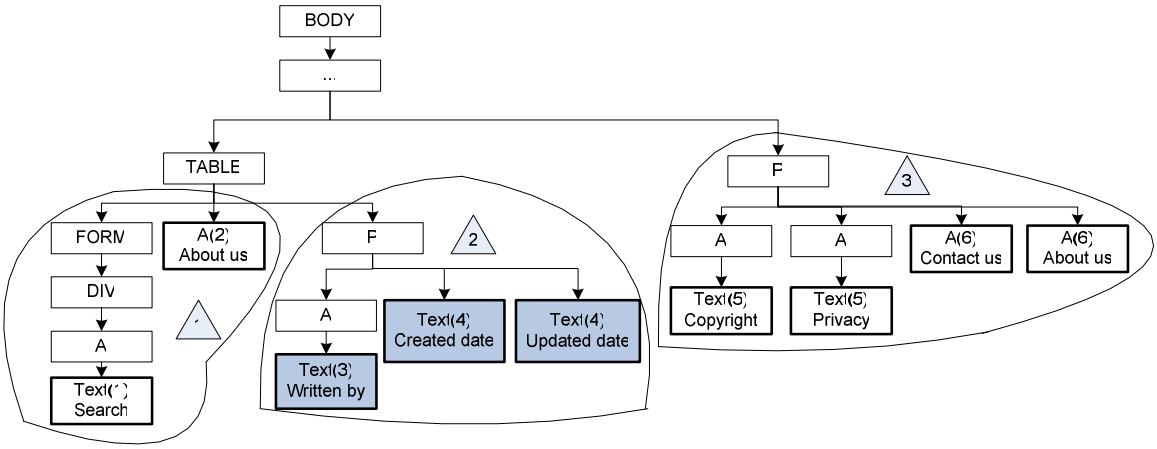

Figure 2 – Identifying sections from a DOM tree

For most indicators that do not have reliable indicator values, the location of the indicator plays an important role in determining whether the indicator exists. For example, we used "M.D." as an indirect indicator value for *author's name* and *author's credentials*. If the candidate line that contains "M.D." appears in the main content section of a Web page, the possibility of the line containing the author's name is larger than it would be in other sections.

2.2.3. Detecting expression patterns of candidate lines

We use the regular expressions to detect patterns from candidate lines. For example, regular expressions for date are used to extract patterns from candidate lines for *updated date,* and matched strings are represented as ":Date:". Using regular expressions allow us to precisely match indicators with expected patterns. For those criteria that do not have direct indicator values, domain databases were constructed to detecting patterns. We used the most common 18839 last names, 4275 female first names, and 1219 male first names from the US population during the 1990 census[15] as the domain database for

author's name. Names are extracted based on domain databases from candidate lines. These lines are represented with their extracted patterns. For example, a candidate line for *author's name and credentials* could be represented as ":FirstName: :LastName:, :AuthorDegree:, :FirstName: :LastName:, :AuthorDegree:".

2.2.4. Rule-based line classification

Where an indicator may appear on Web pages is quite flexible. For example, the author's information may be found under the title at the top region of the main content section on some Web pages, but may be found at the bottom region of main content section of other Web pages. Therefore we need to generate a few rules for classifying candidate lines into indicators. Table 1 shows some rules we used. For a candidate line to be considered as a "phone" indicator, it has to meet two conditions: the line must be in the main content section of the Web page and the indicator value ":Phone:" must appear before the ":PhoneNumber:". Some rules are complicated. If an indicator value for author's credential is detected in a line, but the author's name does not appear on the same line, it is not considered a valid indicator for author's name and credential. We also verify the appearance sequence of some related indicators. For instance, A*uthor's name, author's credentials and author's affiliation* usually appear at the main content section and in sequence of name, credentials and affiliation.

Table 1- Examples of rules used in rule-based line classification

Criteria	Indicator Value	Section	Pattern
Copyright	Copyright	Bottom	:Copyright:
Phone	Phone	Main content	:Phone:>:PhoneNumber:
Author's name	Written by	Top or Bottom of main content	:AuthorName: > :Firstname:|:LastName:
Author's credentials	M.D.	Top or Bottom of main content	:Firstname:|:LastName:> :AuthorDegree:
Author's affiliation	University	Top or Bottom of main content	:Firstname:|:LastName: > :AuthorDegree: :AuthorDegree: > :AuthorAffiliation:

"A > B" stands for A appears before B. "A|B" stands for A or B.

3. Results

We performed the evaluation of the effectiveness of the rule-based detection method. The first data set we used is based on an "Acne" data set obtained from the previous study [12]. From the top 30 Web pages retrieved with Google using the search term "Acne", only 20 of these can be loaded in DOM trees. We used these 20 web pages for measuring the overall detection accuracy of all criteria. We searched for indicators on these Web pages manually. The program also searched for indicators on the same data set. The number of indicators correctly detected by the program for all the criteria is shown in Table 2. Recall of the detection is the proportion of the correctly detected criteria from actual existing criteria. Precision is the proportion of the correctly detected criteria from detected criteria. We got 93% recall and 98% precision using the direct detection method in the previous study [12]. The recall of the rule-based method is slightly lower than that of the direct detection, but the precision reaches 100%. Please note that the comparison of the two methods is not considered as a strict one. Although we used the same Web pages, the number of Web pages has changed and also the number of criteria on the same Web page may have changed during the period of the pervious study and this experiment. The overall detection accuracy can not represent the detection accuracy for each criterion since the occurrence frequency of each indicator is different. Hence, we looked closely at the *author's name, credentials and affiliation*. These criteria were shown in our previous study to be most difficult to detect. There are only four Web pages out of these 20 containing *author's name*. The precision of detection is 100%, but the recall for *author's name* is only 50%. The detection accuracy obtained from these four web pages is not considered valid because of a very small sample size.

In order to further test the detection accuracy of these criteria, we collected a new data set in which each Web page contains at least one criterion related to author. We tested both the overall and individual detection accuracy of author's name, credential and affiliation using the new data set. We performed a search using "Acne" as the search term with Google. The top 50 accessible web pages that were manually verified to contain at least one criterion related to author were obtained. Only 29 out of these 50 Web pages can be loaded in DOM trees, so we used these 29 Web pages as the second data set. There are no overlaps between the first and second data set. The overall recall and precision are 94.02% and 99.16% respectively. The detection accuracies of the three criteria are showing in Table 3. In prior work, we got the precisions of 90%, 82% and 64% respectively for *author's name, author's credentials* and *author's affiliation* [12]. Using rule-based method, the precisions for detecting three criteria are 100%, 93.75% and 94.12%. The recalls, which were not tested in the previous study, also reach 89.66%, 100% and 94.12%.

The overall detection accuracy of the two data sets shows that the rule-based method is effective, in general, for detecting technical criteria. However, the program still missed some criteria. By examining the Web pages used in our study, we found a few reasons for such errors. Firstly, some indicators were generated by embedded java script code, so they were not explicitly showed in the source code of Web pages. The program is unable to detect such hidden indicators. Secondly, indicator values can not cover all scenarios that a criterion may appear in. For example, some disclaimer statements were written in the bottom of a Web page, but the particular indicator value "disclaimer" did not appear in the statements. In this paper we only include indicator values with high occurrence frequencies, so missing some indicator values can not be avoided.

With this new method, the detection accuracy for individual criterion is also greatly improved. The detection accuracy of *author's name, author's credential* and *author's affiliation* were lowest among other criteria using the direct detection method. Rule based method reaches much better performance for them, although their detection accuracies are still lowest among others. The detection accuracies for other criteria in two data sets are not showed in the paper. The improvement of precision for all criteria may be a consequence of using regular expressions and rule-based classification. The improvement of precision for *author's credential and affiliation* is largely due to identified sections obtained from DOM trees and the rule-based classification.

Table 2 – The overall detection accuracies of two data sets

Data Sets	Criteria found			Recall (%)	Precision (%)
	Human	Tool	Tool Correct		
Acne(20)	101	94	94	93.07	100
Acne(29)	251	238	236	94.02	99.16

Table 3 – The detection accuracies of three criteria

Criteria	Criteria found			Recall (%)	Precision (%)
	Human	Tool	Tool Correct		
Author's name	29	26	26	89.66	100
Author's credentials	15	16	15	100	93.75
Author's affiliation	17	17	16	94.12	94.12

4. Discussion

In this paper, we describe a rule-based method for detecting technical criteria. The method focuses on improving the detection accuracy by analyzing the structure and the content of Web pages. We proposed using DOM tree structural information for rule-based criteria detection. The method of combining structural and expression patterns achieves better performance than the direct detection method. The system of automatically detecting technical criteria could be used by consumers for evaluating the presentation quality of Web sites. However, other important criteria, such as accuracy, completeness, readability, design and usability still needs to be evaluated either by professionals or lay people. Also, the method we described in this paper can be used to assist experts to extract metadata elements from health related Web pages. The software detects candidate lines that contain metadata elements and experts can extract metadata from these lines.

There are some limitations of this work. Firstly, we need a more reliable HTML parser. To obtain the HTML structural information, HTML Web pages have to be loaded in DOM trees. For these two data sets, only 58% to 67% web pages can be loaded in DOM trees using the Cobra [15]. Second, a large scale validation of the method is needed. However, there are some difficulties in performing a large scale evaluation. Firstly, we need to store the Web pages because they are constantly changing and some Web pages may not be accessible later. Then, we have to manually label criteria on those Web pages. Also, the date set has to be large enough to cover each criterion and each indicator value. Evaluating recall is particularly time-consuming for some indicators with low occurrence frequencies.

For the future work, we would like to explore some technologies in related fields that might be helpful for analyzing structure or content of Web pages. Some methods for Web structure mining could be used to analyze the structure of Web pages. Lin et al. proposed methods for detecting informative blocks in news Web pages [16]. Yi et al. described a method for eliminating non-main content blocks on Web pages [17]. They detected the main content block of a Web page by comparing DOM trees corresponding to different Web pages on the same Web site. We focus on detecting sections by analyzing the DOM tree and context of each unique Web page. Their methods can not be directly adopted in our study, but they could be used to obtain fine grained Web structures.

Detecting criteria from Web pages can be considered as extracting specific metadata from health related Web pages. From the perspective of analyzing the content of Web page, Natural language processing (NLP) and machine learning are two potential methods for extracting metadata from Web pages. NLP was used to extract metadata elements from educational HTML web pages [18], but their specific method for educational Web page may not be general enough to extract health related metadata elements.

The effectiveness of machine learning methods for extracting metadata has been shown in some studies [19, 20].Han et al's method achieved high accuracy in extracting metadata elements including author's name, author's credentials and author's affiliation. However, adopting their methods in detecting criteria has some practical issues. Firstly, although the method is effective for detecting the information from document headers, it may not applicable for extracting criteria from health related Web pages. The effectiveness of machine learning methods relies on the selected features. The features used for extracting metadata from document headers and the method for extracting these features may be different from that of Web pages. Secondly, generating training data sets is the prerequisite for using a supervised machine learning method. To our knowledge, such datasets are not available for HTML web pages. Great effort is needed to generate such training data sets. Therefore, we used a rule-based detection method instead of machine learning method in this paper, but machine learning based method for detecting criteria is worth exploring in the future.

Acknowledgement

The authors would like to thank Kai Simon for his suggestions and Yong Liang for his help in programming work.

References

1. National Telecommunications and Information Administration, A nation online: Entering the broadband age. US Department of Commerce: Washington; 2004, available at: http://www.ntia.doc.gov/reports/anol/NationOnlineBroadband04.htm.
2. Statistics Canada, CANSIM, "Households using the Internet from home, by purpose of use". 2004, available at: http://www40.statcan.ca/l01/cst01/arts52b.htm.
3. S. Fox, "Health information online", Pew Internet & American Life Project Report 2005 May, available at: http://www.pewinternet.org/pdfs/PIP_Healthtopics_May05.pdf.
4. A. Risk, J. Dzenowagis, "Review of Internet Health Information Quality Initiatives", *Journal of Medical Internet Research,* vol. 3, no. 4, 2002, e28, available at: http://www.jmir.org/ 2001/4/e28/.
5. HON (Health On the Net Foundation), hhttp://www.hon.ch.
6. OMNI, http://www.omni.ac.uk/.
7. DISCERN, http://www.discern.org.uk/.
8. G. Eysenbach, C. Kohler, G. Yihune, K. Lamp, P. Cross, D. Brickley, "A metadata vocabulary for self- and third-party labeling of health Web sites: Health information disclosure, description and evaluation language (HIDDEL)", *Proceedings of the 2001 AMIA Annual Symposium*, 2001, pp.169-173.
9. L. F. Soualmiaa, S. J. Darmonia, "Combining different standards and different approaches for health information retrieval in a quality-controlled gateway", *International Journal of Medical Informatics,* vol. 74, no. 2-4, 2005, pp. 141-150.
10. S.L. Price, W.R. Hersh, "Filtering Web Pages for Quality Indicators: An empirical approach to finding high quality consumer health information on the World Wide Web". *Proceedings of the 1999 AMIA Annual Symposium*, 1999, pp. 911-915.
11. K.M. Griffiths, T.T. Tang, D. Hawking, H. Christensen, "Automated assessment of the quality of depression websites", *Journal of Medical Internet Research.* vol. 7, no. 5, 2005, e59, available at: http://www.jmir.org/ 2005/5/e59/.
12. Y. Wang, Z. Liu, "Automatic Detecting Indicators for Quality of Health Information on the Web", *International Journal of Medical Informatics*, in press
13. G. Eysenbach, J. Powell, O. Kuss, and E. Sa, "Empirical studies assessing the quality of health information for consumers on the World Wide Web". Journal of the American Medical Association, vol. 287, no. 20, 2002, pp. 2691-2700.
14. Cobra, http://html.xamjwg.org/cobra.jsp.
15. Most Common Names and Surnames in the U.S., http://names.mongabay.com/.
16. S.H. Lin, J.M. Ho, "Discovering informative content blocks from Web documents**"**, *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, SIG*KDD'02, 2002, pp. 588-593.
17. L. Yi, B. Liu, X. Li, "Eliminating noisy information in Web pages for data mining", *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, SIGKDD'03, 2003, pp. 296-305
18. O. Yilmazel, C.M. Finneran, E. D. Liddy, "MetaExtract: an NLP system to automatically assign Metadata", *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital libraries (JCDL'04)*, 2004, pp. 241-242.
19. H. Han, E. Manavoglu, H. Zha, K. Tsioutsiouliklis, CL. Giles, X. Zhang, "Rule-based word clustering for document metadata extraction", *ACM symposium on applied computing*, 2005, pp. 1049-1053.
20. H. Han, CL. Giles, E. Manavoglu, H. Zha, Z. Zhang, EA. Fox, "Automatic document metadata extraction using support vector machines", *Proceedings of the 2003 Joint Conference o Digital Libraries(JDCL'03)*, 2003, pp. 37-48.