Non-topical classification of healthcare information on the web

Hirokawa, Sachio
Research Institute for Information Technology, Kyushu University

Ishita, Emi
Kyushu University Library, Kyushu University

http://hdl.handle.net/2324/1457794

出版情報：Frontiers in Artificial Intelligence and Applications. 262, pp.237-247, 2014-01-01. IOS Press
バージョン：
権利関係：

Non-Topical Classification of Healthcare Information on the Web

Sachio HIROKAWA [a], Emi ISHITA [a]
[a] *Research Institute for Information Technology, Kyushu University*
[b] *Kyushu University Library, Kyushu University*

Abstract.
The present paper collected the asthma related 4,762 Web pages from 1,759 sites using 6 queries. Each site is manually categorized by the standard topics of description and information dissemination, diary and idle talk and Q&A. By careful analysis, it turned out that the pages can be classified in non-topical categories such as "reading level", "objectivity/subjectivity" and "reliability". The manually assigned labels of non-topical categories are then used as learning data to apply SVM (support machine vector). The prediction performance (F-measure) were below 50% with the naive application of SVM. However, the prediction performance was improved over 50% by feature selection except for reading level.

1. Introduction

Concern about health is increasing by progress of the aging society. The needs for medical information and health information are growing. By the spread of Webs everyone can obtain information now comparatively easily. In order to meet these needs, medical information and health information are offered in some public libraries. Tokyo Metropolitan Library, for example, has a Web page of health and medical information corner, where we can see the lists of related books and magazines [1]. Tottori prefecture Library and many other libraries have the medical and healthcare corners and the special library on struggle against diseases The information needs are high in the U.S. as well. There are varieties of practices on service about health and medical information. The public librarian's guide [2] covers the useful sources on users, reference services, Web information and service samples on medical and healthcare information, particularly in the U.S. public libraries. The Beacon Hill branch at Seattle of Washington State public library is offering a different kind of service, where they provide a free consulting session on healthcare every month. They collaborate with SHIBA (Statewide Health Insurance Benefits Advisors) and CISC (Chinese Information and Service Center) to help non-native English speakers in introducing insurance system and hospitals which can respond in languages other than English. This is a good sample to respond diversified needs in medical and health care.

Compared to those special services on the libraries, Web and search engines are widely used to obtain variety of data, from professional sources to general information. Both specialized agencies and ordinary people use the Web as useful media for information dissemination. Now, the Web itself is one information service and acquiring information from a Web is performed as a general act.

In the present paper, we collected Web pages related to the medical treatment and healthcare, and analyzed them based on the topics of the pages. Firstly, we manually classified the Web pages according to the standard topics such as portal site, Q&A, diary and blog. Secondly, we considered the non-topical views [9] such as objectivity and reliability, to classify Web pages. Those criteria is much more crucial as medical and healthcare information. Manual classification is not effective for the huge number of real Web pages. So, we considered applying the machine learning method to characterize those Web pages with non-topical features. In this paper, we applied SVM (support vector machine) with feature selection. We used the manually labeled Web pages as the learning data and constructed the model whose prediction performance are over 50%, which is not perfect but satisfactory as the first trial.

2. Related Work

2.1. Web Contents of Medical Institutions

Ikari[3,4] and Hashimoto et. al. [5] studied the contents of Web pages containing medical and healthcare information in Japan. They analyzed the Web pages of medical institutions in Tokyo, Kanagawa and Chiba prefectures listed on the Kanto-region hospital list of names 2001/2003. Hashimoto et. al. [5] reported the progress of information dissemination for raising the accessibility of the hospital to patients and for producing a brand of each hospital. On the other hand, they confirmed that only 20% hospitals provided the medical information and the patient education information. Here they use the term “medical information” to represent the doctors’ career and school education, staffs organization, medical facilities, medical results and expense information. The term “patient education information” represents the link to disease information, prevention / life management information, the medical treatment method information, the patient meeting information, other medicine sites and the inter-regional association. We use the term “medical and healthcare information”, in the present paper, meaning close to the term “patient education information” of [5].

Toyama et. al. [6] surveyed the contents of Web pages of dental care organizations. Setoyama and Nakayama [7] studied the contents of public health centers’ Web pages. Sasaki and Nagamatsu [8] reported concerning medical information for foreigners on the local government Web pages.

Most of these reports are centered on the medical institutions and are published in 2002 and 2003. Even the latest ones [7,8] are in 2008. In those days, Web pages were created mainly for information dissemination of companies and organizations and not for individual users. The number of Web pages was not huge as it is today, where ordinary people write their blogs and send their message on the twitter. These information is easily available through search engines. In a word, the situation of Web in 2002 and 2003 were quite different to that of nowadays.

2.2. Non-Topical Categories

This paper analyzes the Web pages containing medical treatment and healthcare information, focusing on asthma. However, the analysis viewpoint is not in medical or healthcare but in non-topical categories. Following the previous work [9], we consider the following four categories – “target reader”, “reading level”, “objectivity” and “reliability”. From a careful inspection of the collected Web pages, we believe that this viewpoint is useful.

The name of a site is a useful clue to estimate the reading level, i.e., the required knowledge to understand the content of the page. However, we found several pages that require middle level knowledge which cannot be guessedd from the names of the sites. The reading level and the readability are closely related. Sakai [10] listed the syntax, the vocabularies and the structure of a sentence as the factors that affect the document readability of the medical treatment and the healthcare information.

The proper evaluation of the reliability would not be able unless we are expert. However, there are some proposals of automatic judgement of the reliability. Miyamori et.al. [9] proposes the four criteria – the information content, the information source, the information appearance and the social evaluation, as the standards to estimate the reliability. The reliability of the information source is determined by the reliability of the institution or of the person who delivered the information. The information appearance is measured if and how the source and the contact is shown clearly. They evaluated the reliability by combining the four measures. Terajima et.al. [12] proposes to evaluate the reliability of a Web site by the four items – the producer, the information basis, the creation and revised date, and the contact.

3. Medical Treatment and Healthcare Information on the Web

3.1. Search Queries to Collect URLs

We expected that there would be too much information on medical treatment and healthcare on the Web. The target of our analysis is not in the diversity of diseases, but on the diversity of description of those information. So, we limited the target pages to a specific disease. Firstly, we collected the URLs of Web pages which contains keywords related to medical treatment or healthcare information. We sent search queries to the search engine google in February and March 2013. As a query, we used the name of a disease and a another word considered to be used at the same time. We prepared the following 6 queries: “asthma treatment”, “asthma condition”, “asthma cause”, “asthma prevention”, “asthma experience” and “child asthma”. Those are the English translation of the original Japanese queries. The collected pages are written in Japanese. As the result of search, we obtained 894, 851, 951, 1037, 890 and 842 URLs respectively to each query, which amounts to 5,465 URLs in total. We manually operated this process without using API. By removing the duplicated URLs, we obtained 4,762 URLs.

3.2. Categorization of Web Sites

Table 1 shows the top 17 sites categorized by topics. The second author chose the topic of each site. The name of the sites are authors' English translation of the original Japanese names of the sites. The sites of the topic "Q&A" contain the frequently asked questions and the answers for the questions. Anyone can ask and respond on the sites with unspecified anonymity. A "portal" site covers a general information such as the cause, the condition, the treatment and the way to search the hospitals on a specific disease.

	topic	name	URL	# pages	%
1	Q&A	Yahoo! Chiebukuro	detail.chiebukuro.yahoo.co.jp	132	2.8
2	portal	Happy Husband and Wife	happyfu-fu.com/zensoku	91	1.9
3	Q&A	OKWave	okwave.jp	64	1.3
4	blog	Livedoor blog	blog.livedoor.jp	53	1.1
5	blog	Ameba blog	ameblo.jp	52	1.1
6	portal	Change Asthma!	naruhodo-zensoku.com	49	1.0
7	portal	Comprehensive Asthma Site	zensoku.jp	49	1.0
8	portal	Hospital Navi Momo	momo365.jp	48	1.0
9	portal	Cause and Treatment of Asthma	ppaapp.com/zensoku	41	0.9
10	shopping	Prevention of Hay Fever	kahunbousinm.blog.so-net.ne.jp	39	0.8
11	Q&A	Yahoo! Chiebukuro	chiebukuro.yahoo.co.jp	38	0.8
12	blog	goo blog	blog.goo.ne.jp	38	0.8
13	Q&A	Teach Me! goo	oshiete.goo.ne.jp	37	0.8
14	shopping	Amazon	www.amazon.co.jp	34	0.7
15	shopping	Google book s	books.google.co.jp	32	0.7
16	blog	Yahoo! blog	blogs.yahoo.co.jp	32	0.7
17	Q&A	AskDoctors	asthma.askdoctors.jp	32	0.7

Table 1. List of Sites of Medical Treatment and Healthcare Information

"Yahoo! Chiebukuro" appears at the top and at the 11-th and contains many pages. "Yahoo! Chiebukuro" contains the questions such as the cause of the asthma which does not come out at all when sleeping night and the cause and prevention of bronchial asthma. The 17-th "AskDoctors" is a Q&A site where the visitors can ask to the doctors. The doctors respond to the questions of ordinary people. In addition to Q&A, there are items, such as illness and a condition encyclopedia, a medicine encyclopedia, and hospital search, in this site. High ranked Q&A sites, such as "OKWave" and "Teach Me goo", are realized by many and unspecified questioner and respondents. There are 303 Q&A sites, in the top 17, which cover 6.4%. The ratio of Q&A sites is larger than that of blog pages.

"Happy Husband and Wife" is the top site among the portal sites, where all searched pages are on the directory "/zensoku". An ordinary person who is not a medical staff is managing this site. This person provides a bulletin board where anyone can use for information exchange. The 6-th ranked site of "Change Asthma" is a comprehensive site about the right asthmatic knowledge and medical treatment run by two drug companies. The doctors are supervising the contents. There is an item of the cure for asthmatic, the medicine for asthmatic, the point of prevention, etc. "Comprehensive Asthma Site" is also run by a drug company. The site offers the information dissemination and illness hospital search about adult asthma and child asthma. "Hospital Navi Momo" is a portal site where they provide a nationwide search service of hospitals and clinics for women. The site is ranked among the top sites, because many information of the hospitals which named asthma as the medical examination item were contained in the search results.

As a summary of the topical analysis of the sites, we found that most of the medical treatment and healthcare information come either from information dissemination by many and unspecified people, such as Q&A and blogs, or from the portal sites which offer information synthetic about a specific disease.

3.3. Manual Inspection of Web pages

The second author and a student made a detailed topic categorization of search results which were retrieved with the queries "asthma experience", "asthma cause" and "asthma treatment". The number of Web pages are 842, 951 and 60, respectively. There were 1,837 pages after removing the duplicated pages. We excluded some URLs, such as the link to a movie, which were not accessible. We compiled 1,759 pages. Table 2 displays the content based topic categorization. We classified some pages of portal sites into Q&A pages, when the pages contains the questions and the answers.

category	# pages	ratio(%)
information dissemination	486	27.6
diary and idle talk	383	21.8
Q&A	184	10.5
bulletin board	90	5.1
experiences	90	5.1
institution introduction	87	4.9
goods and products	74	4.2
news	64	3.6
books	53	3.0
Q&A by doctors	51	2.9
other	197	11.2
total	1,759	100.0

Table 2. Contents of Pages

The category of "information dissemination" has the largest number of Web pages which give description and explanation about the condition, the cause, and the cure of asthma. Most of those pages are in the portal sites of Table 1. But we found several sites of clinics where asthma is explained as their medical examination. We found many pages in the medium size portals where various information on asthma are provided. The second largest category is "diary and idle talk". The blogs describing individual contents are included in this category. But, the pages introducing individual experience and experience of many people collectively was included in the category of "experiences". The pages introducing of instruments, such as medicine for asthmatic medical treatments, programs, books, inhalers, and the pages recommending purchase are classified in “goods and products”. There are conspicuos pages where the cure experiences of asthma were introduced and then the medical treatment programs are recommended.

On the whole, it turned out that there are two types of Web pages. The Web pages of the first type give the description and the information about the disease. The second type consists of the categories “diary and idle talk”,”bulletin board” and “experiences” whose purpose is to share personal information. A typical page of the second type is created to form a community of patients for sharing information. In another words, those pages are splitted into the pages for offering the objective fact about medical treatment and health and the pages for sharing individual experiences and opinions.

4. Non-Topical Categorization

4.1. Variety of Search Purpose

The search engines return a variety of information as a ranked list of Web pages. We only have to send a query. However, the ranking of search results is determined by the relationship between the query and the main contents of each Web page and the popularity of the page. Before we read each of the search results, we do not know the assumed reader of the page, the reliability of the page and the subjectivity or the objectivity of the page. There might be a diversity on the contents of Web pages even if they are dealing with the same topic depending on who writes the page, why and how is the page is written. Imagine a topic “asthma treatment”, for example. A page might be written by a patient who writes his experience. Another page might be written by an medical institution or by an expert to explain the latest treatment of the disease. The former page is valuable for a patient who is looking for other patients’ experience of the same disease. Sharing the experience would be more important than the reliability of the information. On the other hand, the later page is more important than the former, for a person who is looking for the latest treatment. The reliability and the quality of the page is crucial in this case. That is, even if they are searching in the same topic, the information they need differs according to the viewpoint. Thus, in order to enable various search, it is necessary to assign the additional information corresponding to those viewpoints, which are not necessarily contained explicitly. The present paper considers the categorization of Web pages based not on the standard topics of healthcare and medical treatment, but on the non-topical categories. We first manually assign the labels. Then we apply a machine learning method for classification to those non-topical categories.

4.2. Manual Assignment of Non-Topical Category

Based on the previous work [9], we selected four categories – “target reader”, “reading level”, “objectivity” and “reliability” as non-topical categories. We manually inspected the 1,759 pages as we did in the previous section. We used the labels of "ordinary person", "expert" and "child" as the label of target reader. We used 1,2,3,4 and 5 to evaluate the level for the other three categories. We asked three subjects to label the Web pages. Before starting evaluation, we asked them to discuss some sample pages to negotiate the evaluation so that the judgement should be stable. Then each subject evaluated the pages.

Concerning to "target reader", 96.2% of 1,759 Web pages are for ordinary person. Only 26 pages (3.7%) are for experts. Most of the pages are the research reports and bibliographic information of research reports. There are only 2 pages for child. The pages are the quiz pages in a site where children can learn about asthma. There are many Web pages dealing with child asthma, since we searched with "child asthma". But the most of the pages are aimed at their parents and not for children.

84.8% pages were labeled as the pages which slightly require advanced knowledge with respect to the reading level. Most of those pages are evaluated to target the ordinary persons. The pages labeled as to require "advance knowledge" are the scientific articles in CiNii [1] and the description pages of the condition which the medical association offers, in which technical terms are used. The Q&A sites by doctors are labeled as to require "a little advanced knowledge".

Concerning to the subjectivity and the objectivity, the largest group of Web pages are labeled as "subjective" which have 464(26.5%) among 1,759 pages. The similar number of pages are classified as "objective" or "neutral". There are many pages of patients' experiences in "subjective" pages. Most pages judged as"objective" are labeled so, mainly because they seem to be written by an medical institution or by a doctor. The purpose of those pages are information dissemination.

	level	# pages
1	unreliable	315
2	slightly unreliable	428
3	neutral	466
4	slightly reliable	161
5	reliable	389
	total	1,759

Table 3. Reliability

	level	# pages
1	subjective	464
2	slightly subjective	428
3	neutral	452
4	slightly objective	119
5	objective	296
	total	1,759

Table 4. Subjectivity and Objectivity

	level	# pages
1	no knowledge	11
2	slightly required	29
3	medium degree	1,492
4	a little advanced	123
5	advanced	104
	total	1,759

Table 5. Reading Level

[1] http://ci.nii.ac.jp

5. Classification by Support Vector Machine with Feature Selection

When we ordinary people search for information on a disease and on a treatment of the disease, we use the keywords that represent the disease, the condition and the treatment. We use the query such as “asthma treatment”, “asthma condition” or “asthma cause”. However, the evaluation of a Web page obtained as the search result depends not only those topical contents of the page, but also on the person who wrote the page, on whether the page is reliable or not, and on whether the page describes the subjective facts or the objective opinions and experiences. In the present paper, we call such a viewpoint as a non-topical category. It is not a trivial task to choose appropriate keywords to describe those non-topical categories. If we know the characteristic words that determine a non-topical category, we can select the documents we want.

We apply the machine learning method SVM (support vector machine) to determine the non-topical Web pages. SVM is used in many domains as the technique of machine learning with high flexibility. The strength of SVM is in its applicability to the high dimension data which have a huge number of attributes. It has been believed that we do not have to focus on the particular attributes to obtain the best performance of classification. On the other hand, there have been researches on the feature selection where an appropriate set of feature are pursued to interpret the meaning of the classification. It has been reported that there are some cases where the feature selection greatly improves the classification performance [13].

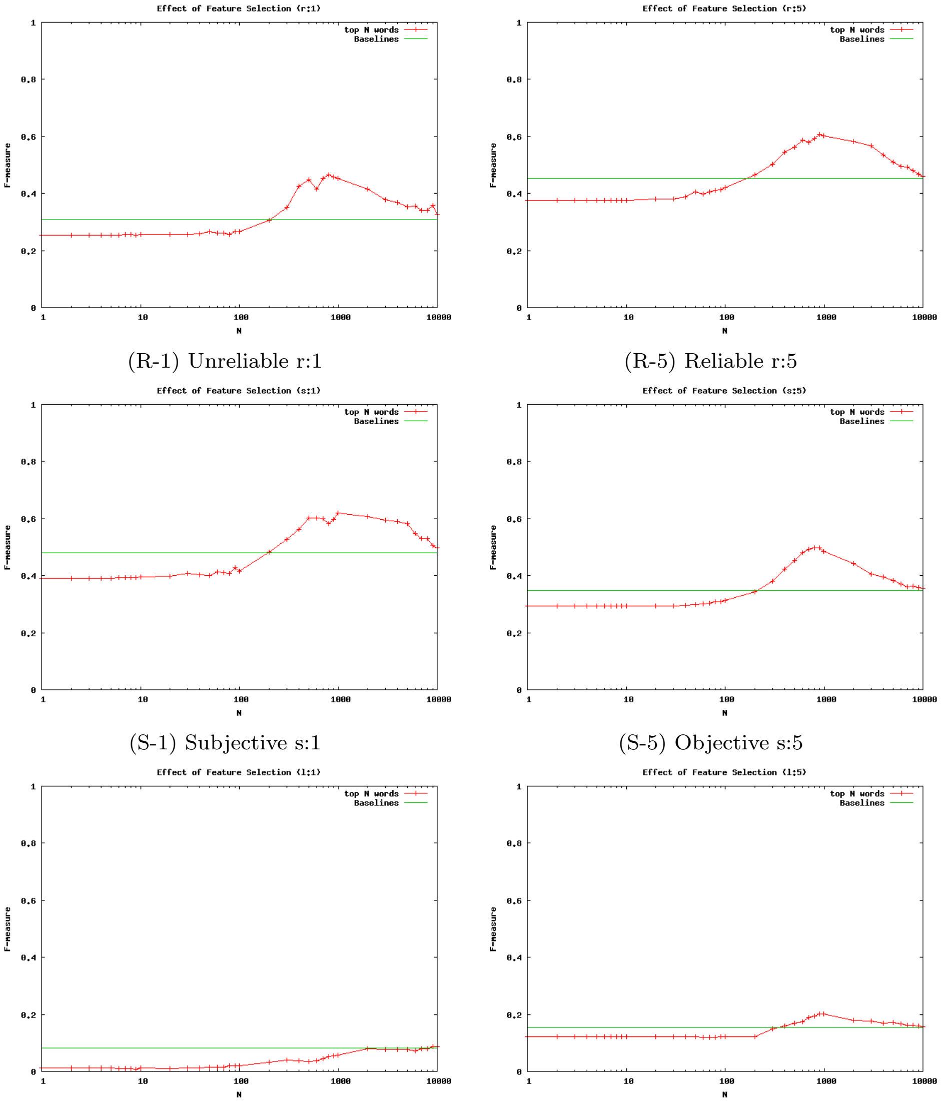
c	eval	prec	rec	F	acc
r	1	0.1877	0.8808	0.3089	0.4240
r	5	0.2962	0.9640	0.4521	0.4601
s	1	0.3223	0.9410	0.4798	0.5069
s	5	0.2133	0.9620	0.3486	0.3811
l	1	0.0437	0.8000	0.0825	0.6974
l	5	0.0845	0.9351	0.1545	0.3301

Table 6. Naive SVM

c	eval	N	prec	rec	F	acc
r	1	800	0.3221	0.8344	0.4642	0.7195
r	5	900	0.4610	0.9039	0.6061	0.7271
s	1	1000	0.4775	0.8786	0.6183	0.7373
s	5	900	0.3525	0.8565	0.4982	0.7042
l	1	9000	0.0466	0.8000	0.0878	0.7049
l	5	900	0.1145	0.8743	0.2014	0.5413

Table 7. Optimal Feature Selection

We apply the method of [13] to the learning data which we explained in the previous section. Table 6 displays the prediction performance by naive application of SVM, where each page is vectorized with all words that appear in the page. The first column denoted by “c” shows the non-topical categories. “r” stands for the reliability, “s” stands for the subjectivity and the objectivity, and “l” stands for the reading level. The performance for the reading level is very low, because the number of pages of the level=1 or level=5 are very small, as we can see in Table 5. The F-measures for the reliability and the subjectivity are between 30% and 50%. Table 7 shows the prediction performance with the optimal number of feature words. The F-measures are improved by 10% to 20%. Figure 1 displays the F-measures with respect to the number of feature words in the vertical axis. We can see that the feature selection improved the performace for the reliability and for the subjectivity/objectivity.



(R-1) Unreliable r:1

(R-5) Reliable r:5

(S-1) Subjective s:1

(S-5) Objective s:5

(L-1) no Knowledge l:1

(L-5) Advance l:5

Figure 1. Prediction Performance

6. Conclusion and Further Work

The present paper studied the medical and the healthcare information on the Web. Asthma related Web pages are collected and analyzed. The reading level, subjectivity/objectivity and the reliability are considered and used as the viewpoints of the analysis. Manual inspection of collected pages revealed that the non-topical view is helpful. The manually assigned labels of non-topical categories are then used as the learning data for machine learning method SVM to predict the classification. It is confirmed that the optimal feature selection is effective yielding 50% to 60% prediction performance (F-measure).

The data analyzed in the present paper is very limited. Further evaluation is necessary by expanding the queries in collecting the Web pages as well as by considering other non-topical categories.

Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Number 22500220 and 24500176.

References

[1] Tokyo Metropolitan Library, Health and Medical Information Service (in Japanese), http://www.library.metro.tokyo.jp/tabid/408/Default.aspx (accessed 2013-06-12)

[2] Andrea Kenyon, Barbara Palmer Casini, Atsutake NOzoe, The public librarian's guide to providing consumer health information (Japanese translation), Japan Library Association, Tokyo, 2007

[3] Tomoko Ikari, An analysis of the content of medical services organization homepages and the choices of obstetrics service being made by mothers (in Japanese), Keie to Joho:Shizuoka Kenritsu Daigaku, Keiei Joho Gakubu gakuho 16(1), pp.27-42, 2003

[4] Tomoko Ikari, Change of Web Contents of Medical Organizations – Comparison of 2001 and 2002(in Japanese), Journal of Japanese Hospital Association 62(9), pp.780-786, 2003

[5] Eriko Hashimoto, Chihiro Wada, Tomoko Ikari, Research on Web Information Contents of Hospitals:What do Web pages of Hospitals tell to the patients?(in Japanese) 11(3), pp.69-86, 2001

[6] Atushi Toyama, Kazumi Morita, Yasuomi Toyama, Haruo Nakagaki, Report on Web Contents of Oral Health Organizations indexed in Search Engines (in Japanese), Journal of Japanese Society for Oral Health 52(4) pp.500-501, 2002

[7] Evaluation of public health center websites from the viewpoints of content, usability, and accessibility (in Japanese), Japanese Journal of Public Health 55(2), pp.93-100, 2008

[8] Kumi Sasaki, Yasuko Nagamatsu, Usefulness of Medical Information for the Foreign Residents on Homepage of Local Governments in Japan (in Japanese), Journal of St. Luke's Society for Nursing Research 12(1), pp.25-32, 2008

[9] E. Ishita, Non-topical Classification for Healthcare Information, Bulletin of IEEE Technical Committee on Digital Libraries, Vol.5, No.3, Dec.2009. http://www.ieee-tcdl.org/Bulletin/current/Ishita/ishita.html, (accessed 2013-06-12)

[10] Yukiko Sakai, Improvement and evaluation of readability of Japanese health information texts: an experiment on the ease of reading and understanding written texts on disease (in Japanese), Library and information science (65), pp.1-35, 2011

[11] Hisashi Miyamori, Susumu Akamine, Yoshikiyo Kato, Ken Kaneiwa, Kaoru Sumi, Kentaro Inui, Sadao Kurohashi, Evaluation Data and Prototype System WISDOM for Information Credibility Analysis (in Japanese), Information Processing Sciety of Japan, Technical Report on NLP 2007(76), pp.103-108, 2007

[12] Tomoko Terajima, Eri Machida, Shin-ichi Yamagata, Mayumi Mochizuki, Investigation of the Convenience of the Internet and the Reliability of Web-sites in Retrieval of Medical Information in Japan : About Medication for Elderly Patients with Myocardial Infarction(in Japanese), Japan journal of medical informatics 21(6), pp.435-443, 2002

[13] Toshihiko Sakai, Sachio Hirokawa, Feature Words that Classify Problem Sentence in Scientific Article, Proceedings of iiWAS2012, pp.360-367, 2012