# 5-3 Development of the Information Analysis System WISDOM

**KIDAWARA Yutaka**

NICT Knowledge Clustered Group researched and developed the information analysis system "WISDOM" as a research result of the second medium-term plan. WISDOM has functions that users find high-credible information from huge amount of Web pages. WISDOM is the comprehensive and integrated system based on Natural Language Processing (NLP), Information Retrieval (IR), Machine Learning (ML), Database (DB) and High Performance Computing (HPC) Technology. The system has processing capability of Web information analysis, publisher detection, reputation information extraction, display all the processing result within proper category. The paper describes overview of WISDOM.

***Keywords***
Natural language processing, Information analysis, Information retrieval, Huge data management, Big data

## 1 Introduction

The Internet has been established and broadband has progressed. Information distributed via the Internet greatly affects our life. Going through a ubiquitous era in which information was utilized by not only the PC but also various terminals, information distributed on the Internet, for the most part, has largely changed. In the beginning, information could not be sent without a certain level of expertise and the general user was a consumer of information browsing information. But, as the Internet advances, and broadband becomes ubiquitous, even a user with no expert knowledge is able to easily send many different kinds of information. Looking back at years of 2006 through to 2011 (in the second medium-term plan), it was an era in which a general consumer (till then called as Consumer Generated Media (CGM)) could readily send information via a PC or mobile device. This environmental change has been called Web 2.0 and it was the beginning of and explosion in information in which the amount of information distributed and stored on the Internet is undergoing rapid growth. Also, it was the beginning of an era in which the term [cloud computing] was newly coined and that a large amount of data is processed by the network. It was the cusp of a new era resulting from the broadband Internet and the ubiquitous use of mobile terminals.

This has brought great change to overall quality of information. The environment in which the general user can readily send leads to a diversification of information, but this environment does not necessarily have advantages and at the same time produces a large amount of unreliable information lacking credibility. So it becomes quite difficult to find high quality information. Even in the typical search engines, it is not rare that a search result exceeds over several million returns. Although the general user can never comprehend the entirety of the information, he or she is held responsible for judging the quality of information and therefore, confusion can be caused by information (especially in the case of frequently incorrect) content.

We foresaw such an era in 2005 (which was the final year of the first medium-term plan) and discussed how to address this problem (information credibility) with a Media Interaction Group. Thus, the problem was set as an important theme when planning the second medium-term plan. This problem was succeeded by a Knowledge Clustered Group of the second medium-term plan and we tackled the development of the information analysis engine WISDOM.

This paper describes the structure of the information analysis engine WISDOM (Section **2**), presents technology supporting the WISDOM in Section **3**. In Section **4**, a usage example of the WISDOM is shown and associated technologies are introduced in Section **5**. Section **6** is conclusion.

## 2 Information analysis engine WISDOM

### 2.1 Information credibility analysis support

In many cases, the credibility of information stored on the Internet is different depending upon a viewpoint of the user. Therefore, it is not easy to automatically judge the credibility of information. The WISDOM delegates judgment (of the credibility) to the user with the aim of precisely providing background knowledge, facts, points of issue, contrastive points and opinion distribution and so forth about a problem targeted for analysis (to aid in judgment). To that end, it is necessary to analyze a structure of a sentence or text, its nature or relation and analyze/display different expressions having the same meaning and/or ambiguous expressions. Further, as an important clue to judge credibility, a sender of information and the expertise of an organization to which the sender belongs must be displayed. In addition to named entity recognition, such as personal names and/or organization names, comprehensive analysis of texts becomes imperative. Figure 1 shows a decision-making procedure using information credibility support that the WISDOM strives to accomplish.

In order to achieve decision-making support, development of the WISDOM, positions a natural language processing technology as a core technology. It then sets up the following evaluation axes (with a view to a user credibility support including a link analysis technology).

1. Information content credibility
2. Information sender credibility
3. Information appearance credibility

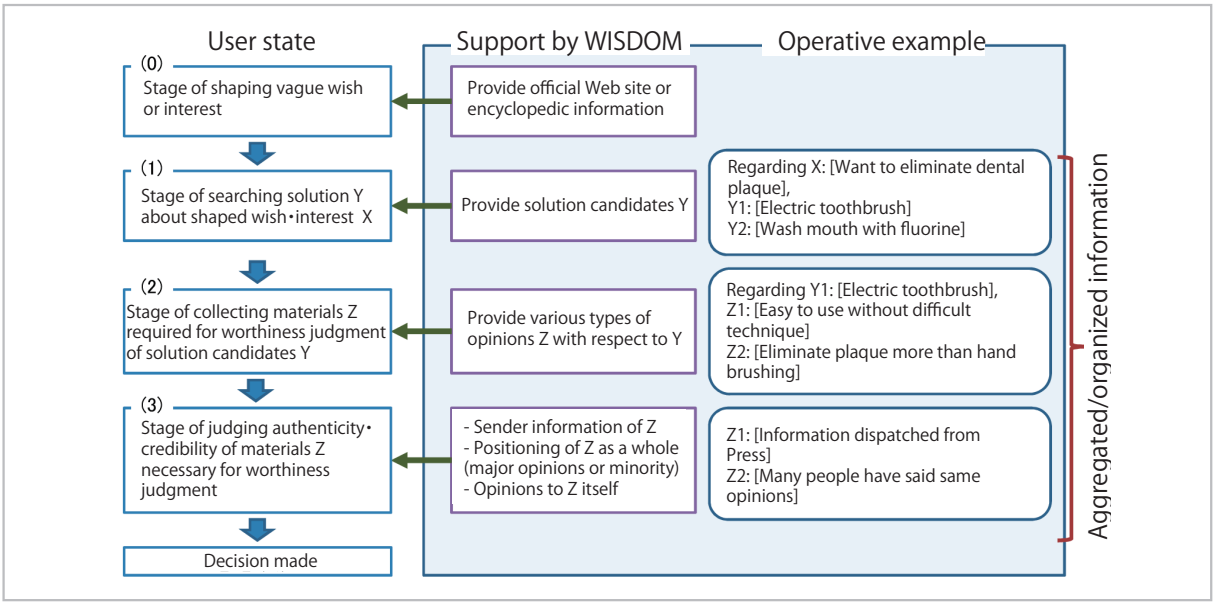


**Fig.1** *Decision-making process and support by WISDOM*

The WISDOM is designed to analyze and provide information based on these viewpoints. To achieve this, collection, accumulation and management of a large volume of web information, sentence information contained there in, analysis of the structure of information, and analysis of web link information, by use of cutting-edge technology is required.

### 2.2 Structure of WISDOM

The information analysis engine WISDOM is classified, as shown in Fig. 2, largely into an Information Analysis Platform Unit, an information analysis engine unit and a front end unit. These units' details will be explained in the following chapter.

## 3 Technologies supporting WISDOM

### 3.1 Technology structuring information analysis base

For an information analysis base, we achieve a management mechanism to precisely and quickly collect, accumulate and then access a huge amount of information.

#### *3.1.1 Crawler*

Crawlers have a tendency to only collect web information and do not have any technological element. However, they must collect web information (so as not to give an excessive load of information. Also, the frequency of updating information varies depending upon the website; Thus consideration must be paid to scheduling. A WISDOM crawler includes a typical crawler and another crawler, or a deep crawler, that visits a specific URL as a base and collects non-acquired pages while tracking back links of the same domain and a RSS feed crawler that scrapes RSS feeds and collects non-acquired pages. The WISDOM crawler collects 10 million Web pages per day. A ratio of these web pages are such that updated pages are about 72%, new pages are 27%, the deep crawler collects 0.5% and 0.5% is acquired by the RSS feed crawler. Upon operation, in consideration of bandwidth (100 Mbps) available for crawler, pages are collected in 4 parallels.

#### *3.1.2 Data pool*

**Crawl data pool**

Crawled Web pages are processed by a URL string filter, a robot.txt filter, a content-type filter, a language filter, a dictionary filter and so forth. Various kinds of information (for crawling and analysis as post-processing) are output. A compressed page file with various

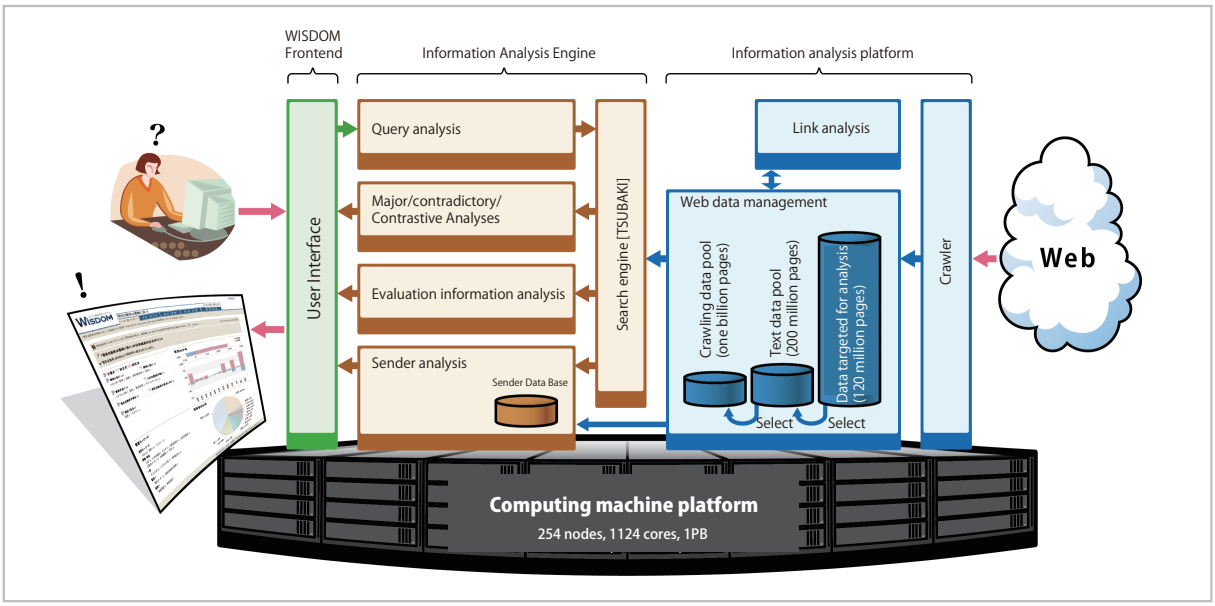


**Fig.2** *Entire structure of information analysis system WISDOM*

information and page data, information links are registered as a crawl data pool. This data is used as information during the next crawl and utilized as original data for information analysis.

**Text data pool**

From acquired Web pages stored in the crawl data pool, pages with high likelihood of being SPAM are eliminated by analyzing text information contents and/or link information and so forth. Further analysis, such as appearance analysis information as well as 200 million pages are selected as pages targeted for analysis. Furthermore, these pages are analyzed as a target and Web standard format (XML) representing characteristics of the page is created. In the web standard format, there are results of syntactic analysis performed to link information, text ID information and text string extracted as sentence, and the results are utilized as information for a method of analyzing various types of information and search engines (to be explained later).

### *3.1.3 Off-line information analysis*

**Link analysis**

When selecting Web pages targeted for analysis, it is important to discover and eliminate useless pages (SPAM). SPAM is grouped into three types: 1) content spam, 2) a link farm, and 3) impersonation. Content SPAM hides useless information such as hidden texts, tiny texts, numeration of words and a title different from its content and thus, affects rankings of search engines.

In search engine ranking, link farms use a technique utilizing link information. The link farm intends to increase the value of the link, creating a great amount of linked pages. The impersonation provides different content in accordance with crawlers and agents of web browsers and thus, crawling cache pages are different from real pages. Using WISDOM, a web spam extraction is implemented based upon the link structure. This technique represents the web link structure as a large graph, extracts strong connecting components and thereby obtains a bow tie structure as shown in

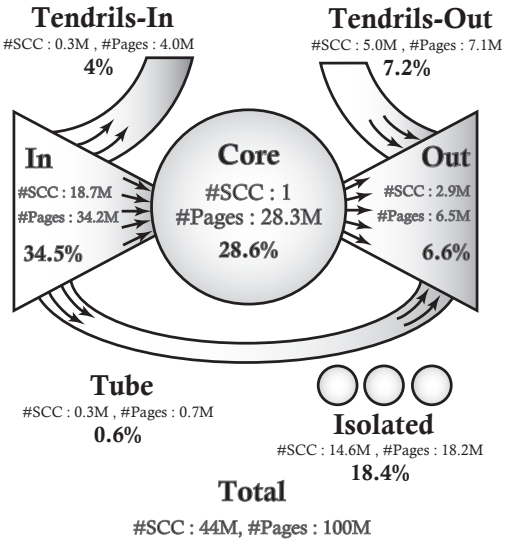


**Fig.3** *Bow tie structure of the Web link*

Fig. 3. After extracting a high-density subgraph by use of this algorithm, SPAM judgment of the sub graph is conducted by SVM and estimation results of a host unit are aggregated. Further, WISDOM develops an algorithm detects SPAM by deflected page rank linking trust and anti-trust in a host graph.

**Appearance analysis**

Web pages have structures and information to be sent is sorted and written based on the structure. On the other hand, on SPAM pages, the meaning of the structure sometimes does not reflect the content of the information. Thus, there are many inconsistencies as characteristics in appearance. Moreover, on Web pages run by businesses, pages that require a site policy and/or contact information might give rise to doubts over whether content is closely checked or not. In this way, the matching of ideal information or its structure to content from appearance of the page greatly contributes to the credibility of the Web page. For this reason, after a structural analysis of a Web page, WISDOM analyzes what kind of information is written and then analyzes for ideal information and the locations to be written.

**Sender analysis**

To determine the credibility of informa-

tion, information about sender becomes a very important element. As to information sent by experts and information obviously sent by amateurs, grounds of the information are largely different. In a case of Web information, whether the sender is clear or anonymous can be seen once the content of the information is understood. Thus, identification of senders and/or information providers can be regarded as a task of information extraction. In the development of the WISDOM, we focus attention on a relationship between major portions on Web pages and appearance locations of the information sender names thereon, and develop a method for identifying senders by use of this relationship.

In the WISDOM, we define reality, including persons and/or organizations responsible for content of information on a Web page and its disclosure as a sender and classify senders into a site operator and author. Further, as an information sender class, the senders are classified into six types, and analysis results of sender information of Web pages are sorted.

As to the identification of the sender, after selection of a page region targeted for extraction a sentence targeted for extraction is selected and a candidate for an information sender. To achieve these operations, after analyzing the Web structure, for information about a site operator, (since the information about the site operator tends to appear in a banner at a top of a page or in a copyright notice at a bottom thereof), where information senders frequently appear, intensive analysis is performed therein.

Further, since the sender information is also contained in the text, there is a low ratio that particles other than [no] in Japanese among particles contained in the sentence (representing the sender) is used in the text. And, as a result of morphological analysis, in consideration of high provability that personal names, organization names, end of organization name and non-defined terms are contained in the text, the WISDOM extracts sentences with a high likelihood as a target sentence. As to sentences extracted in this way, they are classified by machine learning as a composition such as (1) appearance frequency in an entire information source, (2) page frequency where a candidate appears, (3) type of text where a candidate appears, (4) part-of-speech property of component words, (5) front morpheme/end morpheme, (6) number of morphemes, (7) location within a page and (8) whether the extraction stems from the copyright notice or not.

### 3.2 Information analysis engine

**Query analysis**

Each of information analysis functions has the necessary information given by query analysis input by the WISDOM. It is assumed that a query is input by a noun string or word string, or a natural sentence, and to determine what WISDOM analysis performed by the query analysis, so the information analysis function becomes one of important processing in the WISDOM. Query input is classified into a topic and a sub topic, and the topic and the sub topic are given to an evaluation expression analysis. The topic is further passed to an analysis of major/contradictory/contrastive relations. For extraction of this topic and sub topic, a syntactic analyzer KNP* is used.

**Analysis of major/contradictory/contrastive relations**

Related keywords and major/contradictory/contrastive sentences about topics extracted by the query analysis are extracted from a targeted Web page set. The related keywords and major sentences are a linguistic expression that frequently appears on the targeted Web page set. Moreover, respective noun phrases and predicate argument structures (sentences) are targeted for extraction. The contradictory sentence (meaning a sentence that contradicts with and is inconsistent with the major sentence), and the contrastive sentence denotes a sentence that is in contrast with the major sentence. To achieve these analyses and extrac-

* http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html

tions, the predicate argument structure is extracted. The predicate argument structure is comprised of one predicate and argument extracting one or more categorematic strings relating to this predicate. After extracting these predicate argument structures, a synonymous predicate argument structure and a containment relationship are analyzed and then, the analyzed predicate argument structures are aggregated. Further, to classify the major, contradictory and contrastive sentences, after reversing the negative flag and reversing the predicate to antonym, a predicate argument structure set is sought out again and then, the WISDOM finds out the information.

**Evaluation information analysis**

This analysis automatically extracts and classifies affirming/opposing opinions and/or evaluations about topics obtained by a query analysis from Web pages, and outputs opinions and evaluations. The WISDOM analyzes the evaluated information, classifying the opinion and/or evaluation into seven types such as [emotion], [criticism], [merit or advantage], [decision to accept or not], [incident], [ought to do or be] and [demand]. For these classifications, 100 topics are selected and with respect 200 sentences per one topic from collected Web information. The information is evaluated by a human being and then it is given as tag information. A corpus of 2,000 sentences tagged with the evaluated information is created and the corpus is used as (teacher) data for machine learning. Using the corpus, the evaluated information is classified by SVM expanding to multi-class classification, using a method of Pairwise. First, after learning a binary classifier, determining whether or not an evaluation expression given by use of the SVM is related to the topic, examples of obtained evaluated expressions are classified by

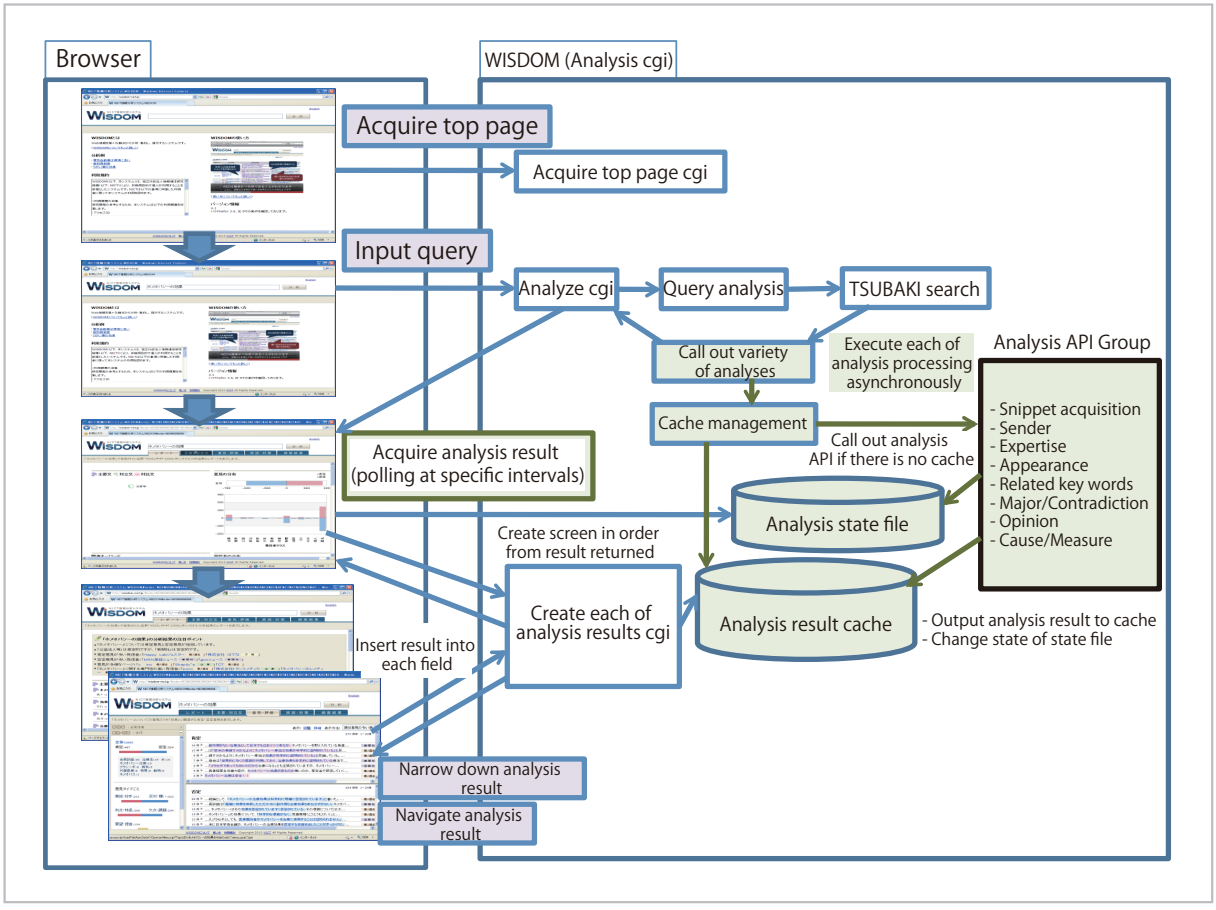


**Fig.4** *Process flow of WISDOM*

a binary classifier. The distance from a separation plane is output as relevance, and a high relevance is output as evaluation information.

### 3.3 WISDOM frontend

**User interface**

The WISDOM is used via browsers. Aggregation of the large amount of information and analysis thereof is conducted by a server. Input of the query and/or changing of analysis functions is achieved by a click in the browser or change of tab. A processing flow is shown in Fig. 4.

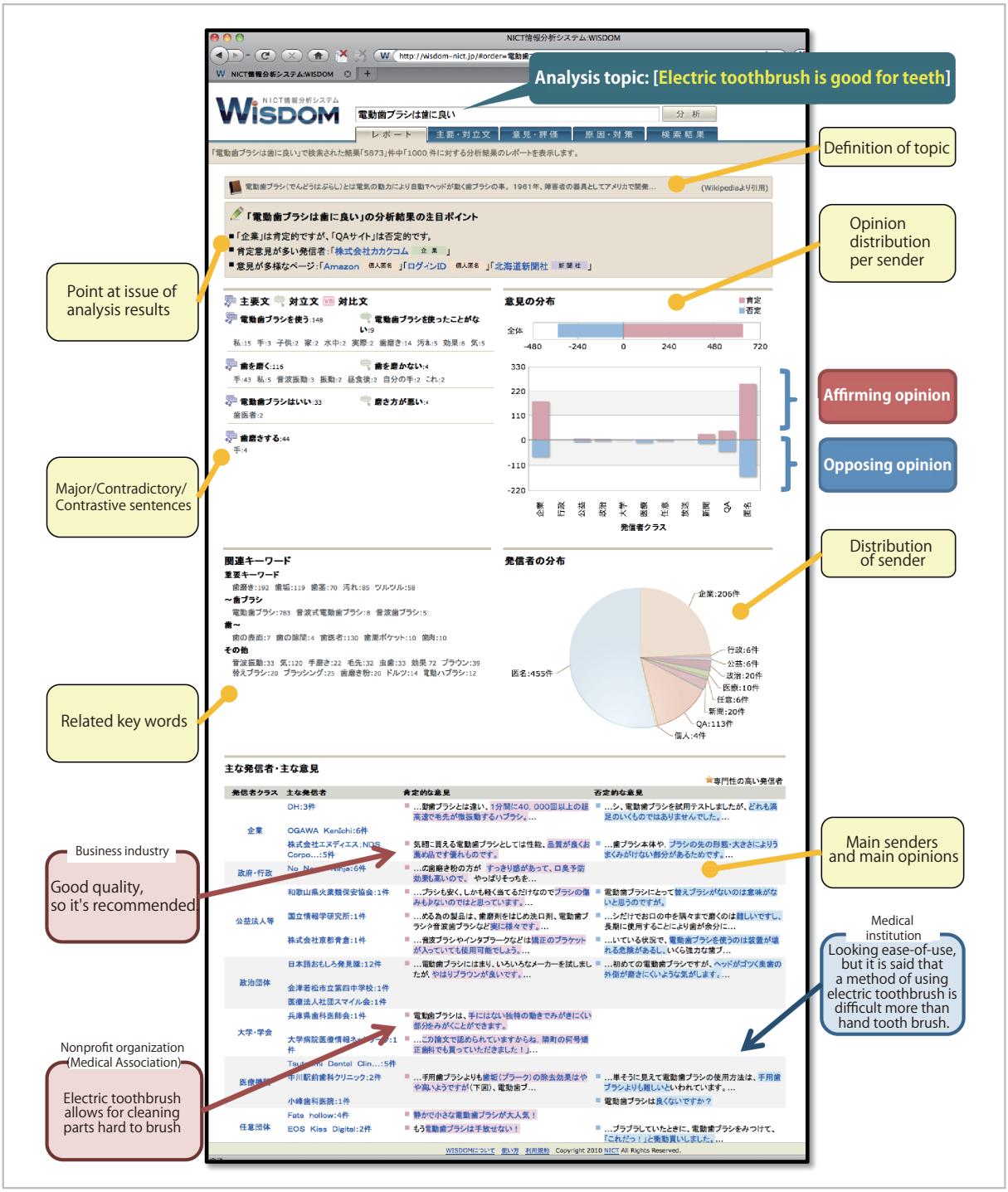


**Fig.5** *Usage example of WISDOM*

## 4 Analysis by WISDOM

In the WISDOM, each function is changed by the tab on a Web interface, and by displaying several types of analysis results described herein (thus far), results can be evaluated. The most distinctive interface of the WISDOM is a report screen. An output result (report) with respect to a sentence targeted for analysis like [an electric toothbrush is good for teeth] is shown in Fig. 5. In this page, noticeable points of the analysis result and/or related keywords thereof, the sender distribution is displayed together, and an overview as to the analysis target can be done.

## 5 Associated study

The present study aims at addressing a very difficult theme, which is the credibility analysis of information. How a user takes an output of a computing machine can be classified into four types: presumed credibility, reputed credibility, surface credibility and experience credibility (by studies of Foggs et al. [1]. Foggs et al. [2] further organize these concepts and that information credibility is judged mainly on a basis of [trustworthiness] and [expertise]. Rich et al. [3] discussed a cognitive process and strategy of credibility judgment targeting university students. According to this discussion, a process for a human being to judge information credibility has two types, that is, predictive judgment and evaluation judgment. It is said that credibility judgment is a process that repeatedly judges predictive judgment and evaluation judgment, and that credibility judgment of information is a complicated cognitive practice including critical thinking. Suggestion by all of these studies and analyses is such that the information credibility is a multiple issue comprised of combinations of various factors, and it can be said that information credibility is not equal to the authenticity of information or correctness thereof. Based upon these reports, the WISDOM is designed to analyze from the viewpoint of (1) credibility based upon the sender, (2) credibility based upon appearance characteristics of information, (3) credibility judging from evaluation of information and (4) credibility on a basis of the meaning and content of the information. As an expert search targeting the Web, there is a proposal from the expert search that Castillo et al [6] use Wikipedia and a proposal of a semantic Web approach [7]. Nakajima et al. [8] proposed a method of ranking blogs based on the familiarity in a specific field, but a method of the present study targets typical Web pages and is superior in that there is no need to prepare dictionaries. As to extraction of reputation information, Kobayashi et al. [9] and Miyazaki et al. [10] propose a method of extracting the reputation information (about products) from review articles and/or blog articles. This method is different in that the method of the present study has an eye for extracting the evaluated information contained in objective descriptions like [product X broke three days after purchased] and so forth.

## 6 Conclusion

This paper gives an overview of the WISDOM. With the WISDOM, highly advanced information processing technologies such as the natural language processing technology, information search technology, machine learning technology and moreover, the huge information management technology and parallel processing technology functions in a fusion manner. At universities and/or laboratory institutes in Japan, the WISDOM system enables the functioning of massive and exclusive systems that are different from other systems. Even in terms of quickly starting the study with attention paid to a view of the information credibility, this system is unique worldwide. When compared with the year 2006, in which the second medium-term plan started, presently, as information is increasing at an explosive rate, it becomes important to be able to find valuable information.

When the Great East Japan Earthquake occurred on March 11, 2011, worthiness of in-

formation published by SNS (like Facebook, Twitter and so forth) was extensively recognized. Amid the progress of such diversification of information and/or large scaled information more than more, this study has been implemented as [Research and Development of Key Technology about Information Credibility Criteria] Project (so-called Information Credibility Project) by a Knowledge Clustered Group in the second medium-term plan. It was undertaken by a Media Interaction Group in the first medium-term plan, and in the third medium-term Plan, Information Analysis Laboratory and Information Services Platform Laboratory were established within Universal Communication Research Institute. Then, much fruit has been yielded while accelerating the study in this field. In such a framework, we work diligently (on research and development) to make the debut as WISDOM2015 a reality. With its functions refurbished at the end year of the Third Medium-Term Plan. Due to space limitation, this paper only touches on a part of functions of the WISDOM. For more detail, please see reference documents [11].

## Acknowledgment

Regarding the development of the WISDOM, we would like to extend our heartfelt gratitude to Professor Sadao Kurohashi, Kyoto University and Professor Kentaro Inui, Tohoku University who both participated in the project as a visiting researcher of the Knowledge Clustered Group, members of the Information Credibility Project and all staff members of the former Knowledge Clustered Group.

### *References*

1 Fogg, B. J. and Tseng, H., "The Elements of Computer Credibility," Proceedings of the SIGCHI conference on human factors in computing systems, ACM Press, pp. 80–87, 1999.

2 Fogg, B., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., Paul, J., Rangnekar, A., Shon, J., Swani, P. et al., "What makes Web sites credible?: a report on a large quantitative study," Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 61–68, 2001.

3 Rieh, S. and Hilligoss, B., "College Students' Credibility Judgments in the Information-Seeking Process," The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning, pp. 49–71, 2007.

4 Demartini, G., "Finding Experts Using Wikipedia," Proceedings of the 2nd International Workshop on Finding Experts on the Web with Semantics (FEWS'07), pp. 33–41, 2007.

5 Jung, H., Lee, M., Kang, I.-S., Lee, S.-W., and Sung, W.-K., "Finding Topic-centric Identified Experts based on Full Text Analysis," Proceedings of the 2nd International Workshop on Finding Experts on the Web with Semantics (FEWS'07), 2007.

6 C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, and S. Vigna, "A reference collection for web spam," SIGIR Forum, 40(2): pp. 11–24, December 2006.

7 C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri, "Know your neighbors: web spam detection using the web topology," In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 423–430, 2007.

8 Shinsuke NAKAJIMA, Yoichi INAGAKI, and Tomoaki KUSANO, "Blog Ranking Method Based on Bloggers' Knowledge Level for Providing Trustable Information," Journal of the Database Society of Japan (in Japanese), Vol. 7, No. 1, pp. 257–262, 2008.

9 Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto, "Designing the Task of Opinion Extraction and Structurization," Information Processing Society of Japan Technical Report (in Japanese), Vol. 2006, No. 1,

pp. 111–118, 2006.
10 Rintaro MIYAZAKI and Tatsuomi MORI, "Creation of a Corpas for Sentiment Analysis based on Product Reviews and Analysis of its Features," Information Processing Society of Japan Technical Report (in Japanese), 2008-NL-187, Vol. 15, pp. 99–106, 2008.
11 "The Information System WISDOM," ISBN 978-4-904020-01-2. (in Japanese)

(Accepted June 14, 2012)

***KIDAWARA Yutaka**, Ph.D.*
*Director General, Universal Communication Research Institute*
*Digital Content Management, Ubiquitous Computing, Information Retrieval, Information Analysis*