Information Credibility Analysis of Web Contents

Sadao Kurohashi[†‡] Susumu Akamine[†] Daisuke Kawahara[‡] Yoshikiyo Kato[†]
Tetsuji Nakagawa[†] Kentaro Inui[†] Yutaka Kidawara[†]

[†]National Institute of Information and Communications Technology
3-5 Hikaridai Seika-cho, Soraku-gun, Kyoto 619-0289, Japan

[‡]Graduate School of Informatics, Kyoto University
Yoshida Honmachi, Kyoto 606-8501, Japan

kuro@i.kyoto-u.ac.jp, {akamine, dk, ykato, tnaka, inui, kidawara}@nict.go.jp

Abstract

As computers and computer networks become more sophisticated, a vast amount of information and knowledge has been accumulated and circulated on the Web. They provide people with options regarding their daily lives and are starting to have a strong influence on governmental policies and business management. However, a crucial problem is that information on the Web is not necessarily credible. It is actually very difficult for human beings to judge information credibility and even more difficult for computers. However, computers can be used to develop a system that collects, organizes and relativises information and helps human beings view information from several viewpoints and judge information credibility. This paper introduces the information credibility criteria project at the National Institute of Information and Communications Technology, which aims to develop such a system, called WISDOM.

1 Introduction

The Web is a mine of information and knowledge. News articles, weather forecasts, time tables, business hours and access to shops and restaurants, school guides, facility guides, and Web services such as ticket purchasing and net banking are available on the Web. We rely heavily on the Web and find it difficult to remember out life style before the advent of the Web era.

Such information can be found by locating a Web page/site whose existence the user is already aware of or whose location he/she knows. From the viewpoint of Web search, it is called a navigational search, and the existing search engines function remarkably well in this respect.

However, information on the Web is not limited to those listed above. A much wider variety of information is available on the Web, some of which were very difficult to obtain or did not exist explicitly before the Web era; some examples of such information are opinions and the experiences of ordinary people, research papers and articles of experts, and announcements and white papers from public sectors. Such information provides people with options regarding their daily lives and is starting to have a strong influence on governmental policies and business management.

Such information is required for various purposes, including when users cannot determine exactly what he/she wants to know. It could be related to "problems in child-rearing" in general, "a decline in the children's physical strength," "measures against children's physical strength decline," or "verifying the effectiveness or safety of exercise machine X?" Such queries are known as informational search. Though the conventional search engines can provide us with some answers to such queries, the answer is not necessarily provided in a few top-ranked pages, and in such a case, the users have to arduously check many pages.

A more crucial problem in an informational search is that the information on the Web is not necessarily credible. Since the cost of sending information on the Web is very low, misleading or false information is often sent out without confirming the facts, and there are frauds and false rumors prevalent on the Web.

It is difficult even for human beings to judge information credibility, especially just from an isolated information. We usually collect related information, integrate them, and then judge their credibility on the basis of our common sense. Hence, it is far more difficult for computers to judge information credibility, but it can be developed to support human judgement by collecting, analyzing, and organizing related information automatically. In the present scenario

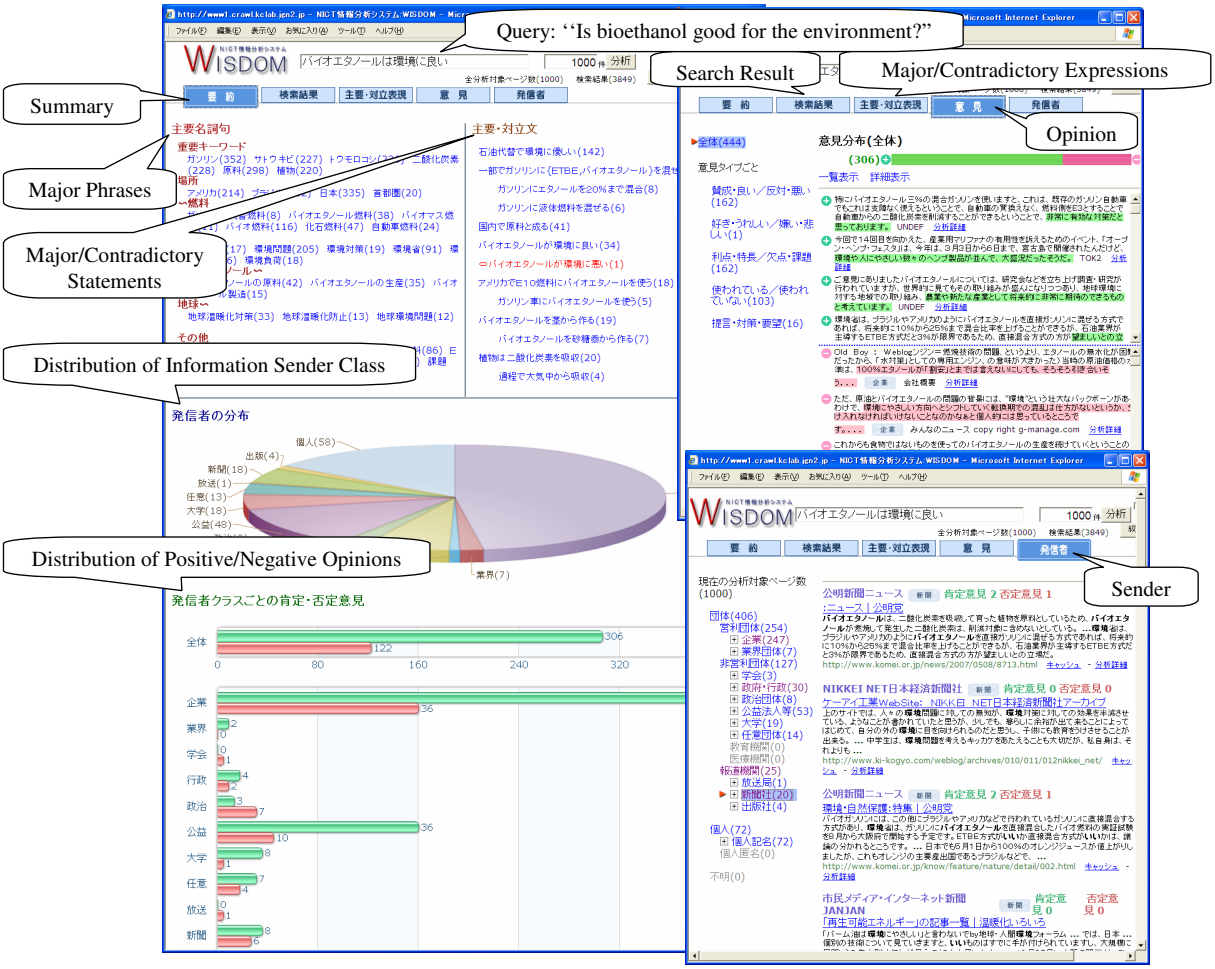


Figure 1. An analysis example of the information analysis system WISDOM.

of Web usage, very often people accept false information without confirming the facts. Information organization systems have become an indispensable technology for a well-functioning society in the Web era. On the basis of such a background, in 2006, the National Institute of Information and Communications Technology (NICT) launched a five-year project named Information Credibility Criteria Project [4].

Information organization is a promising endeavor in the area of next generation Web search. The search engine Clusty provides a search result clustering[1], and Cuil classifies a search result on the basis of query-related terms [2]. The persuasive technology research project at Stanford University discussed how Web sites can be designed to influence people’s perceptions [1]. However, as per our knowledge, no work has been carried out for supporting human judgement on information credibility and information organization systems for this purpose.

[1]http://clusty.com, http://clusty.jp
[2]http://www.cuil.com

In order to support the judgement of information credibility, it is necessary to extract the background, facts, and various opinions and their distribution, for a given topic. For this purpose, syntactic and discourse structures need to be analyzed, their types and relations extracted, and synonymous and ambiguous expressions handled properly. Furthermore, it is important to determine who the information sender is and what is the sender’s speciality as credibility criteria, which require named entity recognition and total analysis of documents.

Consequently, natural language processing (NLP) is considered to be a key technology in the project, and the Web information organization based on advanced NLP is the primary aim of the project.

2 Information Analysis System WISDOM

The project considers the following three criteria for the judgement of information credibility.

1. Credibility of information contents,

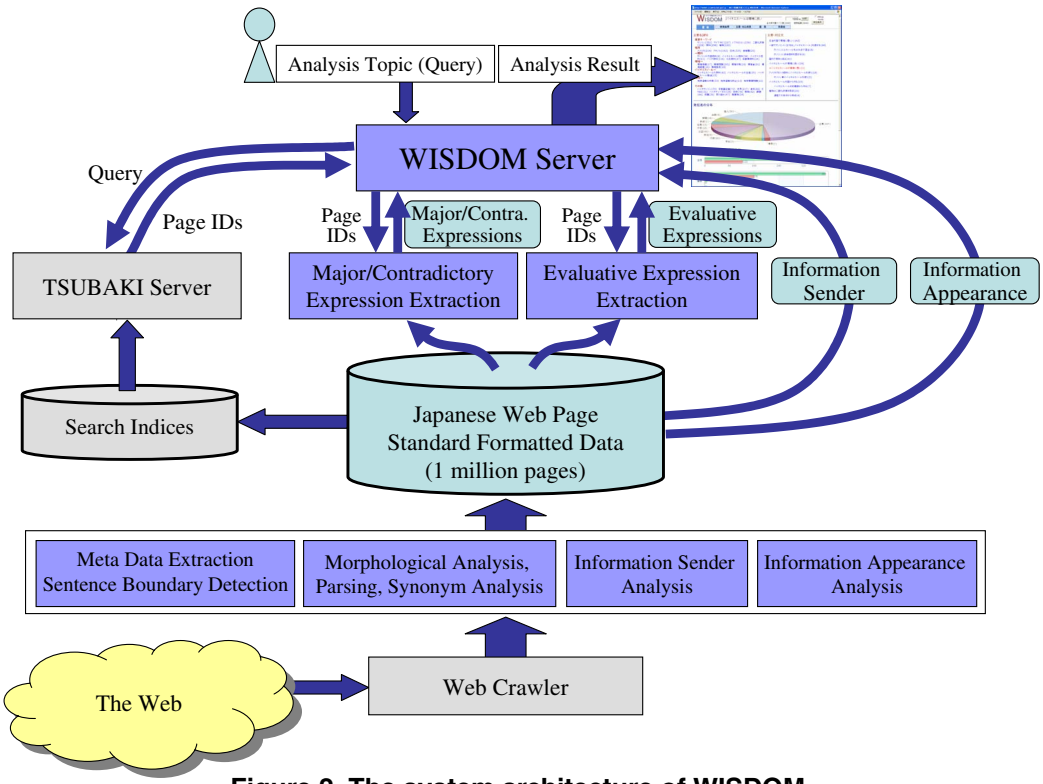


Figure 2. The system architecture of WISDOM.

2. Credibility of the information sender, and
3. Credibility estimated from the document style and superficial characteristics.

In order to help people judge information credibility from these viewpoints, we have been developing an information analysis system called WISDOM[3]. Figure 1 shows the analysis result of WISDOM on the analysis topic "Is bioethanol good for the environment?" Figure 2 shows the system architecture of WISDOM.

Given an analysis topic (query), WISDOM sends the query to the search engine TSUBAKI [8][4], and TSUBAKI returns a list of the top N relevant Web pages (N is set to 1000 usually).

Then, those pages are parallely analyzed: the extraction of major and contradictory expressions and evaluative expressions. Furthermore, the information senders of the Web pages, which were analyzed beforehand, are collected and the distribution is calculated.

The WISDOM analysis results can be seen from several viewpoints by changing tabs, using a Web browser. The leftmost tab, "Summary," shows the summary of the analysis, showing major phrases and major and contradictory statements first. By referring to these phrases and statements, a user can grasp the important issues related to the topic at a glance. The pie diagram indicates the distribution of the information sender class in the 1000 pages, such as company, industry group, and government. The names of the information senders of the class can be seen by placing the cursor on a class region. The last bar chart shows the distribution of positive and negative opinions related to the topic in the 1000 pages, for all and for each sender class. For example, with regard to "Bioethanol," we can see that the number of positive opinions is more than that of negative opinions, but it is the opposite for some sender classes. Several display units in the Summary tab are mouse sensitive, providing links to more detailed information (e.g., the page list including a major statement, the page list of a sender class, the page list containing negative opinions).

The "Search Result" tab shows the search result by TSUBAKI, ranking the relevant pages according to the TSUBAKI criteria. The "Major/Contradictory Expressions" tab shows the list of major phrases and major/contradictory statements about the given topic and the list of pages containing the specified phrase or statement. The "Opinion" tab shows the analysis result of the evaluative expressions, classifying according to for/against, like/dislike, merit/demerit, and others, and it also shows the list of pages containing the specified type of evaluative ex-

[3]Web Information Sensibly and Discreetly Ordered and Marshalled.

[4]TSUBAKI is a search engine infrastructure developed by the support of MEXT Grant-in-Aid "New IT Infrastructure for the Information-explosion Era."

pressions. The "Sender" tab classifies the pages according to the class of the information sender, for example, a user can view the pages created only by the government.

Furthermore, superficial characteristics of pages, called *information appearance*, are analyzed beforehand and can be seen in WISDOM, such as whether or not the contact address is shown in the page and the privacy policy is on the page, the volume of advertisements on the page, the number of images, and the number of in/out links.

As shown thus far, given an analysis topic, WISDOM collects and organizes relevant information on the Web and provides users with multi faceted views. We believe that such a system can considerably support the human judgement of information credibility.

3 Research Infrastructure

Since we are targeting the organization of Web information to support the judgement of information credibility, a toy-scaled research is unsuitable for that purpose. We need to obtain all Web pages beforehand, investigate their characteristics, and perform some prepossessing. An approach that obtains the necessary pages when needed using commercial search engine APIs is apparently insufficient.

Our project adopted a laborious approach based on a robust infrastructure: We prepared large-scale computation infrastructure and data infrastructure and operated a search engine by ourselves, and on the basis of these infrastructure, we constructed and evaluated an information analysis system. In practice, however, due to the limitation of these infrastructure, our project concentrates on Japanese Web pages and constructs indices for 100 million Web pages and usually utilizes them as the analysis target.

Intuitively, half of our efforts thus far has been dedicated for the development of such research infrastructure. However, we feel that the preparation of such research infrastructure, especially the common data platform, is far more important for the revitalization of information technology, and must be one of the initiatives of the national research institute.

3.1 Computation Infrastructure

We installed a cluster machine with 200 nodes (4 cores, 8GB memory and 2TB local storage for each node) and a 100TB file server for the Web crawling, indexing, and operation of a search engine and WISDOM.

The cluster nodes are allocated as follows: a few nodes for the main servers of WISDOM and TSUBAKI, 25 nodes for the search local servers of TSUBAKI, 40 nodes for the analysis of major/contradictory expressions and evaluative expressions, 30 nodes for crawling, and approximately 100 nodes for the conversion of HTML files to the standard formatted data (explained later), indexing, and other processing.

In the case of such a large-scale cluster machine, I/O concentration to the file server becomes a bottleneck. Therefore, the standard formatted Web data (6TB in total, after compression) and indices for the search engine (2TB in total) are distributed to the local disks of the nodes that utilize those data. The 100TB file server is used to store crawling data and the backup of the standard formatted Web data.

Since 2008, we also have been using a cluster machine installed at the Center for Computation, Kyoto University, for the analysis of information senders and others.

3.2 Data Infrastructure

From May to July 2007, 100 million Japanese Web pages were collected using a Web crawler developed by Taura Lab, University of Tokyo, on the NICT cluster machine. Then, the 100 million URLs were re-collected from August to September 2008, and some pages, mainly news and blog pages, were supplemented since some of the original URLs did not exist any more. The current WISDOM uses the new set of 100 million pages as an analysis target. We are planning to design a renewable crawling and to update the analysis target periodically.

The collected Web pages have been converted into the standard formatted Web data, a kind of XML format, proposed by [7]. The format includes several meta data such as URLs, crawl dates, titles and in/out links. A text in a page is automatically segmented into sentences (note that the sentence boundary is not clear in the original HTML file), and the analysis results by morphological analyzer, parser, and synonym analyzer are also stored in the standard format. Furthermore, the site operator, the page author, and information appearance (e.g., contact address, privacy policy, volume of advertisements and images) are analyzed and stored in the standard format.

3.3 Search Engine Infrastructure TSUBAKI

We use a search engine infrastructure TSUBAKI to construct indices for the 100 million Web pages and conduct a search for a given query. The indices are constructed from the standard formatted data: words, their synonyms, and dependent-governor pairs of words/synonyms.

In the NICT computation environment, the indices for one million pages are handled by a single core; that is, 25 nodes (100 cores) are used for TSUBAKI to handle 100 million pages. An analysis topic given to WISDOM is passed to TSUBAKI using the TSUBAKI API, and then the list

Table 1. Examples of obtained major predicate-argument structures and their contradictions.

Analysis topic: *reshikku syujutsu* (LASIK operation)
syujutsu-wo ukeru (undergo an operation)
↔ *syujutsu-wo uke-nai* (not undergo an operation)
shiryoku-ga kaihuku-suru (recover sight)
↔ *shiryoku-ga kaihuku-shi-nai* (not recover sight)
Analysis topic: *gousei senzai* (synthetic detergent)
gousei senzai-wo tsukau (use synthetic detergent)
↔ *gousei senzai-wo tsukawa-nai* (not use synthetic detergent)
kankyou-ni warui (bad for environment)
↔ *kankyou-ni yoi* (good for environment)

of the top 1000 relevant pages is returned to WISDOM. TSUBAKI utilizes a ranking measure based on OKAPI BM25.

4 Extraction of Major Expressions and Their Contradictions

For the organization of information contents, WISDOM extracts and presents the major expressions and their contradictions on a given analysis topic [3]. Major expressions are defined as expressions occurring with a high frequency in the set of Web pages on the analysis topic. They are classified into two classes: noun phrases and predicate-argument structures (statements). Contradictions are the predicate-argument structures that contradict the major expressions. For *yutori kyouiku* (cram-free education), for example, *tsumekomi kyouiku* (cramming education), *ikiru chikara* (life skills), and *monbu kagaku syou* (the Ministry of Education, Culture, Sports, Science and Technology) are extracted as the major noun phrases; *yutori kyouiku-wo minaosu* (reexamine cram-free education) and *gakuryoku-ga teika-suru* (scholastic ability deteriorates), as the major predicate-argument structures; and *gakuryoku-ga koujou-suru* (scholastic ability ameliorates) as the contradiction to *gakuryoku-ga teika-suru* (scholastic ability deteriorates). This kind of summarized information enables a user to grasp the facts and arguments on the analysis topic found on the Web.

We use 1000 Web pages for an analysis topic retrieved from the search engine, TSUBAKI. Our method of extracting major expressions and their contradictions consists of the following steps:

1. Extracting the candidates of major expressions from each Web page:

   The candidates of major expressions are extracted from each Web page in the search result. This process is performed easily by using the analyses of Web pages that TSUBAKI provides and by parallel computing. From the relevant sentences to the analysis topic that consist of about 15 sentences selected from each Web page, compound nouns, parenthetical expressions, and predicate-argument structures are extracted as the candidates of the major expressions.

2. Distilling the major expressions:

   Simply presenting expressions with a high frequency is not always information of high quality. This is because scattering synonymous expressions such as *karikyuramu* (curriculum) and *kyouiku katei* (course of study) and entailing expressions such as IWC and IWC *soukai* (IWC plenary session), all of which occur frequently, hamper the understanding process of users. Further, synonymous predicate-argument structures such as *gakuryoku-ga teika-suru* (scholastic ability deteriorates) and *gakuryoku-ga sagaru* (scholastic ability lowers) have the same problem.

   To overcome this problem, we distill major expressions by merging spelling variations with morphological analysis, merging synonymous expressions automatically acquired from an ordinary dictionary and the Web, and merging expressions that can be entailed by another expressions.

3. Extracting contradictory expressions:

   Predicate-argument structures that negate the predicate of major ones or that replace the predicate of major ones with its antonym are extracted as contradictions. For example, *gakuryoku-ga teika-shi-nai* (scholastic ability does not deteriorate) and *gakuryoku-ga koujou-suru* (scholastic ability ameliorates) are extracted as the contradictions to *gakuryoku-ga teika-suru* (scholastic ability deteriorates). This process is performed using an antonym lexicon, which consists of approximately 2000 pairs, extracted from an ordinary dictionary.

Table 1 lists some obtained pairs of a major predicate-argument structure and its contradiction. They are effective in grasping arguments on an analysis topic and enable a user to easily navigate to the Web pages that contain the contradictions. This is because major expressions are relatively easier to find due to their high frequencies, but contradictions mostly have low frequencies. Therefore, it is

Table 2. Evaluative Expression Types.

Evaluation(+/–): Approval/disapproval or praise/criticism
subarashii(wonderful), *doui-suru*(agree)
Emotion(+/–): Human feeling
suki(like), *kirai*(dislike)
Merit(+/–): Merit/demerit
kenko ni yoi(good for health), *urusai*(noisy)
Event(+/–): Good/bad events or situations
jyusho-suru(awarded), *kowareta*(broken)
Adoption(+/–): Adoption or promotion
saiyou-suru(adopt), "suishin-suru(promote)
Deontic: Proposal, advice, hope or request
suru-bekida(should), *nozomu*(hope)

very difficult to notice and access contradictions with a low frequency without such a support.

The current system independently displays major expressions. In the future, we will analyze and present logical relations such as cause-result and purpose-means between such expressions.

5 Extraction of Evaluative Information

The extraction and classification of evaluative information from texts are important tasks with many applications and have been actively studied recently. Most previous works on opinion extraction or sentiment analysis deal with only subjective and explicit expressions. For example, sentences such as *watashi-wa apple-ga sukida* (I like apples) and *kono seido-ni hantaida* (I oppose the system) contain evaluative expressions that are directly expressed with subjective expressions. However, sentences such as *kono shokuhin-wa kou-gan-kouka-ga aru* (This food has an anti-cancer effect) and *kono camera-wa katte 3-ka-de kowareta* (This camera was broken 3 days after I bought it) do not contain subjective expressions but evidently can contain negative evaluative expressions. From the viewpoint of information credibility, it seems important to deal with a wide variety of evaluative information including such implicit evaluative expressions [6].

A corpus annotated with evaluative information was constructed for evaluative information analysis studies. Thirty topics such as "Bioethanol" and "Pension plan" were chosen. For each topic, 200 sentences containing the topic word were collected from the Web to construct the corpus of totally 6000 sentences. For each sentence, annotators judged whether or not the sentence contained evaluative expressions. When evaluative expressions were identified, the evaluative expressions, their holders, their sentiment polarities (positive or negative), and their relevance to the topic were annotated.

We deal with a wide variety of evaluative expressions, and it is helpful to categorize them. We defined evaluation types as shown in Table 2, and this information was also annotated to the corpus. In these evaluation types, *evaluation*, *emotion*, and a part of *merit* are subjective evaluation, while *event*, *adoption*, and a part of *merit* are objective evaluation.

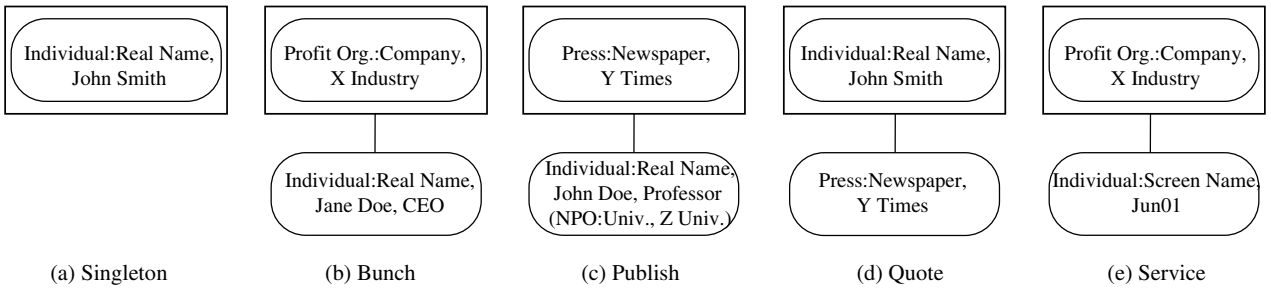
We study the automatic analysis of evaluative information using the corpus. We conducted experiments of sentiment polarity classification using Support Vector Machines. Word forms, POS tags, and sentiment polarities from an evaluative word dictionary of all the words in evaluative expressions were used as features, and an accuracy of 83% was obtained. From the error analysis, we found that it was difficult to classify domain-specific evaluative expressions; we are now planning the automatic acquisition of evaluative word dictionaries.

6 Information Sender Analysis

The source of information (or *information sender*) is one of the important elements when judging the credibility of information. For human beings, it is rather easy to identify the information sender of a Web page. When reading a Web page, whether it be deliberately or not, we attribute some characteristics to the information sender and accordingly form our attitudes toward the information. However, the state-of-the-art search engines do not provide facilities to organize a vast amount of information based on the information sender. If we can organize the information on a certain topic based on who or what type the information sender is, it would enable the user to grasp the overview of the topic or to judge the credibility of relevant information.

6.1 Information Sender Configuration of Web Pages and Information Sender Class

An *information sender* of a Web page is an entity—either an individual or an organization—responsible for the content of the Web page or its publication. The concept of information sender includes the author of the Web page, the governing body of the Web site on which the Web page is published (*site operator*), and the author of the matter quoted in the page. In order to capture the rather intricate phenomenon involved in the publication of information on the Web, we introduce the concept of *information sender configuration*, which represents the information senders of the page and the relationships among the senders. Given below are the five basic types of information sender configuration, which we have defined [2]. Figure 3 shows the examples of information sender configuration, where nodes



(a) Singleton (b) Bunch (c) Publish (d) Quote (e) Service

Figure 3. Examples of information sender configuration.

indicate information senders, and nodes within squares indicate the site operator of the page.

(a) Singleton: This type of configuration indicates that there is only one information sender of the page. The typical cases of singleton are an individual publishing materials himself/herself on a Web site operated by him/her, or a page from a corporation's Web site, which only contains information published by it as an organization.

(b) Bunch: This type of configuration indicates that the information senders of the target Web page are affiliated to the same source, particularly, when all the authors belong to the site operator. An example of bunch type is a CEO of a company conveying a message to customers through his/her company's Web site.

(c) Publish: This type of configuration indicates that the information senders of the target Web page have a *publish* relationship. A publish relationship exists when some author's writings are published by a site operator to which the author does not belong but is aware of his/her writings being published by the site operator. Through this configuration, we intend to identify cases such as a column contributed by a writer being published in some newspaper's Web site.

(d) Quote: This type of configuration indicates that the information senders of the target Web page have a *quote* relationship. When some information from an entity that is different from the site operator is quoted on a Web page, we say that there is a quote relationship between the site operator and the author of the cited information.

(e) Service: This type of configuration indicates that the information senders of the target Web page have a *service* relationship. Service type is applied to cases where the site operator provides facilities with which a third party can publish information on the Web site. Service type pages include forums where the users can post messages, or blogs where readers can post comments.

In the information sender configuration, each information sender is assigned an *information sender class*. Information sender class categorizes the information sender based on axes such as individuals vs. organizations and profit vs. nonprofit organizations. The list below shows the categories of information sender class.

1. Organization
   (a) Profit Organization
      i. Company
      ii. Industry Group
   (b) Nonprofit Organization
      i. Academic Society
      ii. Government
      iii. Political Organization
      iv. Public Service Corporation, Nonprofit Organization
      v. University
      vi. Voluntary Association
      vii. Education Institution
      viii. Medical Institution

1. Organization (cont'd)
   (c) Press
      i. Broadcasting Station
      ii. Newspaper
      iii. Publisher
2. Individual
   (a) Real Name
   (b) Anonymous, Screen Name

This classification was obtained after analyzing the dataset prepared for the evaluation of information credibility analysis [5], which contains about 2000 pages on over 20 topics. We have annotated all the Web pages in the dataset on the basis of the classification as the dataset to be used for the training of models and the evaluation of developed techniques.

6.2 Identifying the Site Operator of Web Pages

As a first step toward identifying the information sender configuration as defined above, we have developed an algorithm that identifies the site operator of a Web page as well as its sender class and have evaluated the technique against the evaluation dataset.

The identification of the site operator is performed in two steps: (1) extracting noun phrases as candidates of the site operator and (2) ranking the candidates according to features based on the NLP analysis or document structure. The identity of the information sender can be often found in the peripheral of the main content of the page, e.g., in the form

of a banner, signature, or copyright notice. Using this observation, we first identify the main part of the web page on the basis of the volume of text and extract the noun phrases from areas contained in parts of the page that were *not* recognized as the main part. Then, we rank the extracted noun phrases to select one as the site operator of the Web page.

1. Collecting relevant information: The identity of the information sender could be contained in many places other than the target page. On the basis of this observation, the proposed method collects the following three types of information other than the target page: (1) ancestor pages, (2) linked pages, and (3) WHOIS database entry.

2. Selecting extraction area: The identity of the information sender can often be found in the peripheral of the main content of the page, e.g., in the form of a banner, signature, or copyright notice. Using this observation, we identify the main part of the Web page on the basis of the volume of text and extract the texts that are contained in parts of the page that were *not* recognized as the main part.

3. Extracting candidates of site operators: The candidates of site operators are extracted from noun phrases in the selected area. On the basis of the named entity recognition, those noun phrases that can be a name of a person or organization are extracted as candidates.

4. Ranking: The extracted candidates are ranked on the basis of features such as the number of occurrences on the page, the overall occurrences on the relevant pages, the type of pages that it occurs on (target page, top page, etc.), the result of the named entity recognition, and the information of morphemes that constitutes the candidate. We used the Ranking SVM as a ranking model and trained the model using the prepared dataset.

The evaluation shows that by using the proposed method, we can identify the site operator at the precision of 60%. If we regard those cases as positive where the true site operator is ranked among the top three candidates, a precision of 70% is achieved. If we exclude blogs, considering that blogs have different properties from normal Web pages, a precision of 70% is achieved (80% among the top three). We also perform the classification of the information sender class of the site operator using machine learning techniques and achieved a precision of 74%. The result of this analysis is incorporated in WISDOM as sender information of Web pages.

As part of our future work, we aim to extend the method to identify information senders besides site operators and to integrate the analysis in order to obtain the information sender configuration. The analysis of the expertise of information senders is also within our target.

7 Conclusions

This paper described Information Credibility Criteria Project conducted at NICT. The project started in 2006, spent the first year for the system design and the evaluation data construction, the second year for the Web data crawling and the development of automatic analyses, and this year for assembling WISDOM.

As shown in this paper, WISDOM already provides quite a nice organized view for given topics and can serve as a useful tool for handling informational queries and for supporting human judgement of information credibility. We are planning to make WISDOM open to the public in the near future.

References

[1] B. J. Fogg. *Persuasive Technology: Using Computers to Change What We Think and Do (The Morgan Kaufmann Series in Interactive Technologies)*. Morgan Kaufmann, January 2003.

[2] Y. Kato, D. Kawahara, K. Inui, S. Kurohashi, and T. Shibata. Extracting the author of web pages. In *Proceedings of Second Workshop on Information Credibility on the Web (WICOW08)*, 2008.

[3] D. Kawahara, S. Kurohashi, and K. Inui. Grasping major statements and their contradictions toward information credibility analysis of web contents. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence (WI'08)*, 12 2008.

[4] S. Kurohashi. Information credibility criteria project. In *the First International Symposium on Universal Communication*, pages 49–52, 6 2007.

[5] H. Miyamori, S. Akamine, Y. Kato, K. Kaneiwa, K. Sumi, K. Inui, and S. Kurohashi. Evaluation data and prototype system wisdom for information credibility analysis. *Internet Research*, 18(2):155–164, 5 2008.

[6] T. Nakagawa, T. Kawada, K. Inui, and S. Kurohashi. Extracting subjective and objective evaluative expressions from the web. In *the Second International Symposium on Universal Communication*, 12 2008.

[7] K. Shinzato, D. Kawahara, C. Hashimoto, and S. Kurohashi. A large-scale web data collection as a natural language processing infrastructure. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC08)*, 2008.

[8] K. Shinzato, T. Shibata, D. Kawahara, C. Hashimoto, and S. Kurohashi. TSUBAKI: An open search engine infrastructure for developing new information access methodology. In *Proceedings of Third International Joint Conference on Natural Language Processing (IJCNLP2008)*, pages 189–196, 2008.