New Ways of Accessing Web Information by Extracting Knowledge from Text

Genichiro Kikui[†] *and Yoshihiro Matsuo*

Abstract

This article introduces the field of automatic extraction of semantic information from Web texts and describes our current efforts centered on *rich indexing technology*. If semantic information can be extracted from the huge amount of text available on the World Wide Web and converted into a format that can be understood by computers, it should be possible to access Web information in completely new ways.

1. Importance of accessing text-based information

The World Wide Web contains a huge amount of content and a huge number of services. One of the most important functions of a portal service is to organize these sources of information and help users access the information they need. There is a great demand for this function. According to a survey conducted in 2006, 92% of the more than 80 million Internet users in Japan make use of a search service at least once a day [1].

Text processing occupies a key role in supporting information access on the Web. Needless to say, there is a massive amount of text on the Web, covering a wide array of content even if we consider only Japanese-language text. In particular, the spread of consumer generated media (CGM) like weblogs (blogs) and online forums in recent years has helped to expand the amount of information related to ideas, opinions, and impressions of general users—a type of information that hardly appears in traditional media. Being able to identify and extract such information from the huge amount of text on the Web is essential in order to satisfy user needs. This is true even for non-text content like pictures and music because text-based information attached to that content in the form of captions and links can also be extracted to give users more extensive support.

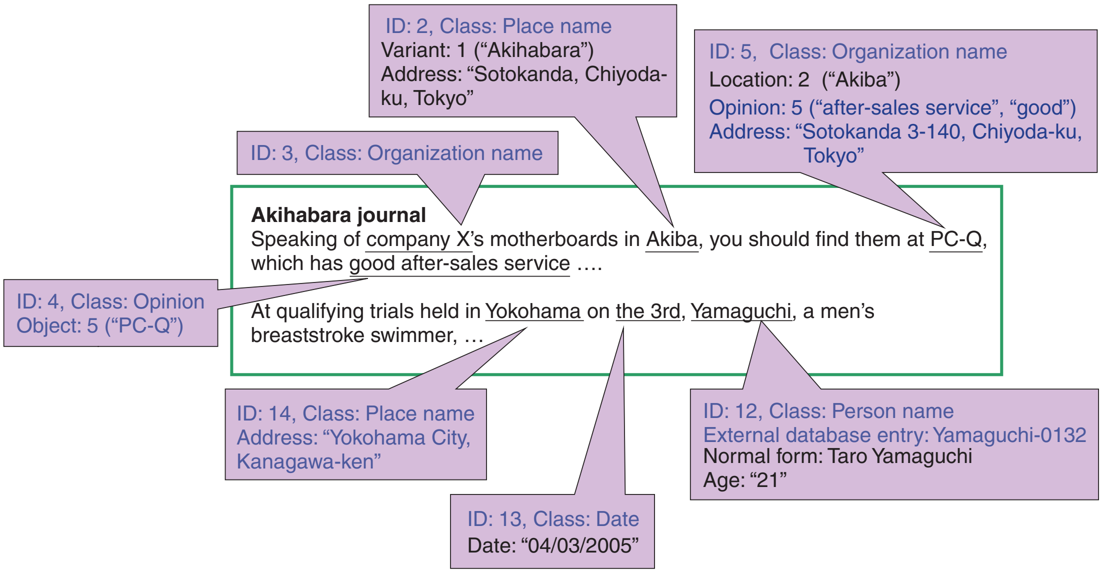
Against the above background, this special feature focuses on Web-based text and introduces natural language processing technology to support access to information contained in text. A comprehensive view of research and development trends in portal technologies is given in Ref. [2].

2. Issues in accessing text-based information

When people set out to search for text-based information on the Web, they generally use a search engine and enter keywords. The search engine will locate Web pages that include the input keywords and will list about ten items at a time using a ranking system established by each search-engine company. This method has been widely adopted because of its simplicity and general versatility. However, it suffers from two key problems.

The first problem is *misses* and *junk* in search results. For example, if a user wishes to get some information about a certain athlete, then entering only the person's surname will unfortunately result in a search for all people with the same surname, including unintended results (i.e., junk). The user might then try entering the athlete's full name. This will reduce the amount of junk, but will still find many other people with the same name unless the combination is uncommon. However, misses can still occur. Pages mentioning the athlete in question may fail to

[†] NTT Cyberspace Laboratories
Yokosuka-shi, 239-0847 Japan
Email: kikui.genichiro@lab.ntt.co.jp


ID: 2, Class: Place name
Variant: 1 (“Akihabara”)
Address: “Sotokanda, Chiyoda-ku, Tokyo”
ID: 5, Class: Organization name
Location: 2 (“Akiba”)
Opinion: 5 (“after-sales service”, “good”)
Address: “Sotokanda 3-140, Chiyoda-ku, Tokyo”
ID: 3, Class: Organization name
ID: 4, Class: Opinion
Object: 5 (“PC-Q”)
ID: 14, Class: Place name
Address: “Yokohama City, Kanagawa-ken”
ID: 13, Class: Date
Date: “04/03/2005”
ID: 12, Class: Person name
External database entry: Yamaguchi-0132
Normal form: Taro Yamaguchi
Age: “21”

Akihabara journal
Speaking of company X’s motherboards in Akiba, you should find them at PC-Q, which has good after-sales service ….

At qualifying trials held in Yokohama on the 3rd, Yamaguchi, a men’s breaststroke swimmer, …

Fig. 1. Examples of information assigned by rich indexing technology.

be found if he or she is usually referred to on the Web by a nickname or by surname only.

The second problem is that the target of a keyword search is not information but text. Of course, this is fine when searching for text itself, for example, a novel, but when the real target of one’s search is information related to something such as a profile or opinions about a particular person, store, product, etc., then the only option is to perform a keyword search using the name of the person or thing and to read each of the results returned until the desired information is found. Thus, using existing text search systems, it is impossible to automatically extract what many people are saying about something in blogs.

3. From text to semantic information

To solve the above problems, we must, in the end, analyze *the meaning that individual linguistic expressions have in text* and extract that meaning in a form that can be easily processed by computers. For example, linguistic expressions having the same meaning should be converted to the same data and stored in a database. However, conducting such an analysis for all possible linguistic expressions would require long-term research from a basic level. In our current study, we narrow down our search target to *named entity expressions* or, simply, *named entities* (described below) that have high practical worth with the aim of achieving early implementation of technology for extracting semantic information.

3.1 Rich indexing technology

Rich indexing technology takes individual named entities that appear in text and assigns rich information to each to indicate (1) what kind of thing that entity corresponds to in the real world and (2) how that entity is referenced in text. Here, named entities correspond to the names of people, places, organizations, and artifacts (e.g., products and services) and time expressions. They are language entities that hold the key to deciphering the meaning of text.

An example of information assigned by rich indexing technology is shown in **Fig. 1**. Input text is shown inside the large rectangle and automatically assigned information is given in the pink callouts. For the first line of input text, assigned information includes the descriptions that Company X is an organization name (company name), that PC-Q is also an organization name (company name), that the latter, in this text, refers to a store’s branch in Sotokanda, and that the writer evaluates it as having good after-sales service.

For the second line, assigned information includes the descriptions that Yamaguchi is not a place name but rather part of the name Taro Yamaguchi, a competitive swimmer; that this name is connected to Yamaguchi-0132 in the personal name database; and that “the 3rd” refers to April 3, 2005.

Assigning additional information in this way makes it possible to search for information about a certain item regardless of the form (character string) in which

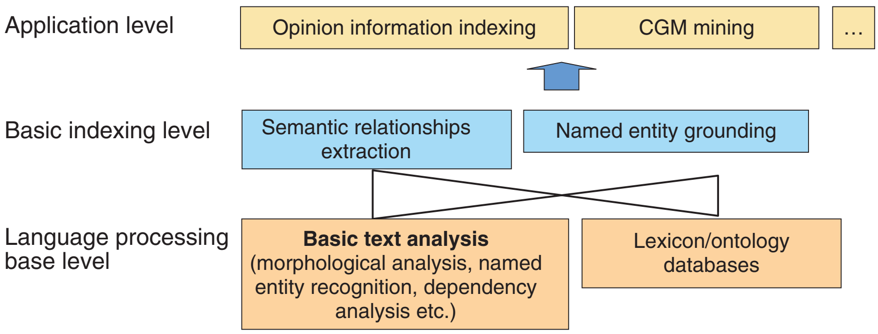

Application level
Opinion information indexing
CGM mining
...
Basic indexing level
Semantic relationships extraction
Named entity grounding
Language processing base level
Basic text analysis (morphological analysis, named entity recognition, dependency analysis etc.)
Lexicon/ontology databases

Fig. 2. Constituent technologies of rich indexing.

it appears in text. For example, to learn about the swimmer Taro Yamaguchi, a user need only search for any text connected to Yamaguchi-0132 and that information will be retrieved without excessive searching. If it is known how a language entity in text is associated with something in the real world, then a broad range of applications can be considered. For example, the name of a singer appearing in text can be linked to a product database for CD sales, enabling effective tie-ups with advertising and online sales.

Furthermore, collecting assigned information in a table facilitates the tabulation, sorting, and searching of information, the same as in (relational) databases. To give an example, retail establishments that are written about in blogs and the opinions given about them can be collected in table form. In this way, the evaluation of a particular store can be determined, and if address information is combined with each store, a list of stores receiving high praise in a certain district can be retrieved.

3.2 Constituents of rich indexing

Rich indexing technology can be broadly divided into three levels, as shown in **Fig. 2**. To begin with, the bottom layer corresponds to base technology for processing the Japanese language. It is divided, in turn, into basic Japanese text analysis and vocabulary/ontology databases. The basic text analysis breaks down an input Japanese sentence into words and analyzes the syntactic relationships among those words (e.g., subject-predicate relationship). This process includes the extraction of named entities that play a central role in rich indexing. A vocabulary/ontology database includes dictionaries that treat the meaning of words and the relationships between them.

Next, the middle layer consists of elemental technologies characteristic of rich indexing, namely semantic relationship extraction and named entity grounding. Semantic relationship extraction aims at analyzing semantic relationships between named entities in sentences and storing them in a database. Named entity grounding relates a named entity to an object in the real world to which it refers.

Finally, the top layer extracts service-oriented knowledge based on information gathered from the lower layers. It includes subjective information indexing that extracts opinions about people, organizations, things, and services from word-of-mouth text, as in blogs and CGM mining, which can retrieve information about things discussed in CGM in the manner of a database.

Some of these technologies are described in more detail in other articles of this special feature.

3.3 Services enabled by rich indexing

Extracting information such as profiles or evaluations of named entities from a huge amount of Web text enables us to develop various new applications, as shown in **Fig. 3**. These might include opinion searching that analyzes and presents Web-based opinions about products and services and company-relationship mining that analyzes Web-based documents about company activities and extracts relationships between companies, as in “they jointly developed a certain product”. Such services are not restricted to portal services—they can also contribute to system development related to text-based information access and to businesses like application service providers.

4. Towards general-purpose knowledge extraction technology

Rich indexing technology shows one direction in the use of Web text as a knowledge source, but it is still limited compared with the human ability to

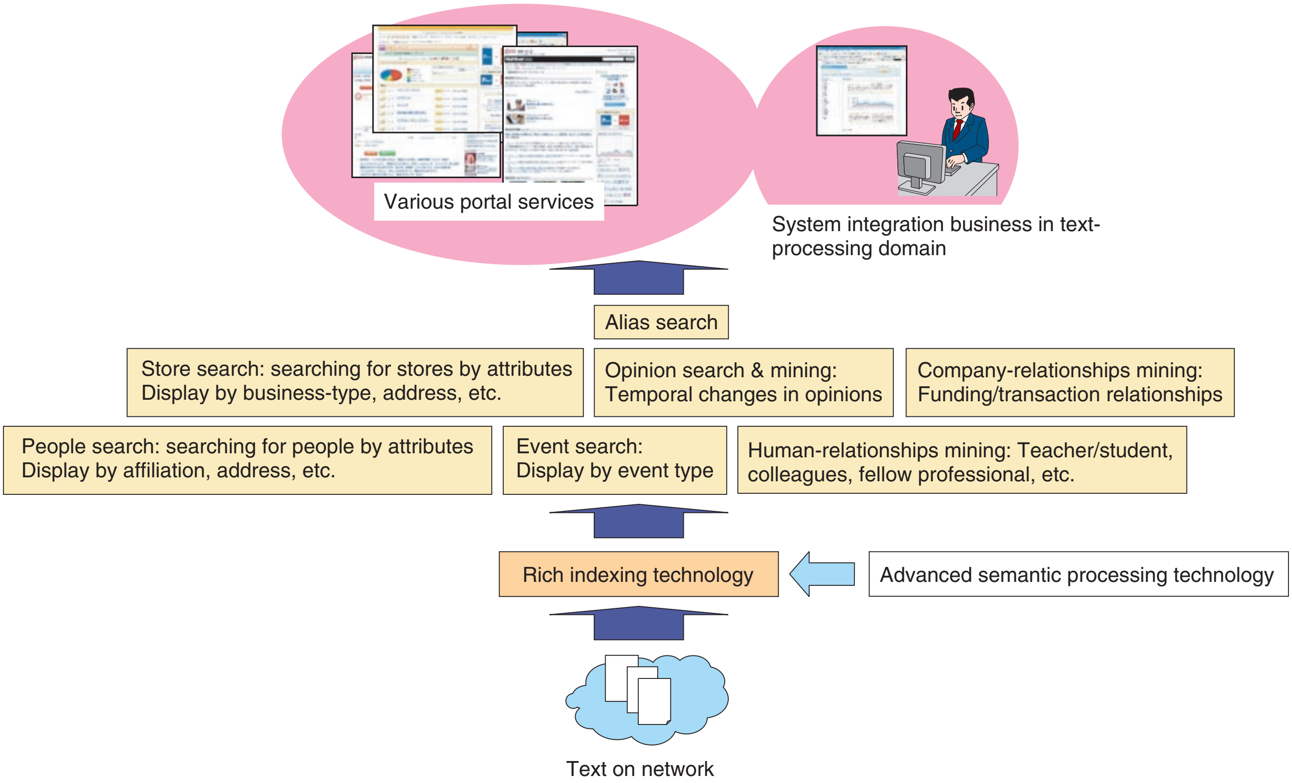

Various portal services
System integration business in text-processing domain
Alias search
Store search: searching for stores by attributes
Display by business-type, address, etc.
Opinion search & mining:
Temporal changes in opinions
Company-relationships mining:
Funding/transaction relationships
People search: searching for people by attributes
Display by affiliation, address, etc.
Event search:
Display by event type
Human-relationships mining: Teacher/student, colleagues, fellow professional, etc.
Rich indexing technology
Advanced semantic processing technology
Text on network

Fig. 3. Services enabled by rich indexing.

extract knowledge from text. With the aim of producing general-purpose technology for extracting semantic information, a number of trials are being conducted centered on NTT Communication Science Laboratories.

To begin with, research is being performed on extracting semantic information as in "What happened?" with regard to general language entities in addition to named entities. And as an extension of this, there is research on inferring background meaning that is not specifically stated. For example, on reading the statement "A defeated B", we also assume that A played a game. This is something that we do automatically. The above research is introduced in the fifth article in this special feature entitled "Challenges for General-purpose Semantic Analysis Technologies".

In addition, we human beings have the ability to learn new language entities and usage patterns and make use of them immediately. Research is also active in giving computers this ability, and some results of that research are being incorporated into rich indexing technology.

5. Concluding remarks

This article introduced our on-going research on extracting knowledge from text, focusing on rich indexing technology. This technology has two main functions: associating named entity expressions with their external referents and extracting semantically related pairs of named entities. These functions enable users to efficiently search for information.

The topic of *extracting semantic information (or knowledge) from text* that we have been discussing up to now is an issue that was also taken up in the field of artificial intelligence about twenty years ago. Some readers might feel that such a function is just a fantasy that is never coming true. To be sure, it is still extremely difficult to deal with background meaning because that requires an ability that approaches human intelligence. There are, however, some key differences between now and then. First of all, as explained in the following articles, today's systems are completely different from earlier ones. They make full use of modern computer power and large-scale language databases, so they can handle Web text and blogs in a very accurate manner. Secondly, the amount of text on the network already far exceeds the amount for which semantic information could be extracted manually, so automatic extraction of meaning by computer—even if not perfectly accurate—can be expected to provide a storehouse of new and useful information.

References

[1] "Internet White Paper 2007," Impress R&D, 2007.

[2] Special Feature: "Navigational Technologies for Next-generation Portal Services," NTT Technical Review, Vol. 4, No. 8, 2006.

Genichiro Kikui

Group Leader, Senior Research Engineer, Supervisor, Speech, Acoustics and Language Laboratory, NTT Cyber Space Laboratories.

He received the B.E., M.E., and Ph.D. degrees from Kyoto University, Kyoto, in 1984, 1986, and 2007, respectively. He joined NTT Communications and Information Processing Laboratories in 1986. He was with ATR Spoken Language Communication Research Laboratories from 2001 to 2006. His research focus is on mono- and cross-language information navigation and machine translation of text and speech. He is a member of the Information Processing Society of Japan (IPSJ), the Association for Natural Language Processing (NLP), the Japanese Society for Artificial Intelligence, and the Association for Computational Linguistics.

Yoshihiro Matsuo

Senior Research Engineer, Supervisor, Speech, Acoustics and Language Laboratory, NTT Cyber Space Laboratories.

He received the B.S. and M.S. degrees in physics from Osaka University, Osaka, in 1988 and 1990, respectively. He joined NTT Communications and Information Processing Laboratories in 1990. His research interests include multimedia indexing, information extraction, and opinion analysis. He is a member of IPSJ and NLP.