LonMaps: An Architecture of a Crime and Accident Mapping System based on News Articles

Hideaki Ito
School of Engineering, Chukyo University
101 Tokodachi, Kaizu-cho, Toyota, Aichi, Japan
Email: itoh@sist.chukyo-u.ac.jp

***Abstract*—LonMaps is an information system for a crime and accident mapping system based on news articles, which enables extraction of certain information items from the news article. To realize the system, four types of information items are extracted, which are crimes/accidents (incidents), places, dates, and personal names. For capturing incidents, a thesaurus consisting of two types of terms, daily and legal terms, is used. Daily terms are used in daily life, while legal terms are used in legal situations. Places, dates, and personal names are extracted on the basis of typical news report patterns. Finally, experiments are performed using LonMaps to evaluate its effectiveness of processing queries and of extracting information items.**

***Keywords*—crime and accident map; news article; thesaurus; sentence pattern; extraction.**

I. INTRODUCTION

Recently, several types of systems for information presentation and management have been developed by integrating maps [6]. When locations play an important role in such information and are presented on maps, a system that deals with such information becomes more useful. A collection of local news about crimes and/or accidents (incidents) is one such type of content that is published in newspapers. In order to annotate such news articles, it is required to extract key items from articles. Key items are elements such as the types of incidents, places, dates, and personal names. Extracted items are useful not only for mapping incidents on maps but also for managing and retrieving news articles.

To integrate local news articles and maps and to manage news articles, LonMaps (Local News Map System) is being developed [14]. Maps indicate places where incidents occur. The system analyzes a single news article to extract information items consisting of an incident, a place, a date, and a personal name. Local news articles have similar structures, so patterns for representing news are found relatively easily. News articles are analyzed by such patterns. After the analysis, the places of incidents are displayed on maps and their positions on the maps are obtained using geo-coding. LonMaps is implemented using GoogleMaps [11].

The type of incident is extracted from the news article. News articles consist of legal terms and daily terms. Legal terms are used in laws, courts, and police departments, i.e., legal situations, while daily terms are used in daily life. Therefore, it is necessary to reduce the gap between the two types of terms, and to annotate a news article using legal terms is needed for preventing ambiguous representation of incidents. Even if the article is described in only daily terms, suitable legal terms are captured from them. To reduce this gap and to correlate the two types of terms, a thesaurus is constructed. This thesaurus is used not only for annotating news articles but also for retrieving them.

Several types of location-centered and geographic information system are being developed with integrated maps [6]. Ilarriet al. [13] discussed that location-aware information is useful in our daily life. When a disaster happens, a crisis map becomes a social tool [10]. Some systems that indicate crimes on maps are developed and they are referred to as crime maps [2], [8], [17], [18]. Locations involving local news articles are indicated on maps available online [1]. A system that allows end users to note local information on maps has been proposed in [12]. A news article analysis systems have been developed in [9], [16]. Moreover, information extraction systems have been developed in [19], [20], where sentence analysis, grammatical knowledge, and templates are used. Some systems extract places from news articles, and they indicates the extracted places on maps [4], [21].

Many existing thesauri are built depending on the application domain. They are utilized for retrieval and annotation [3], [5]. The thesaurus of LonMaps is designed for crimes and accidents, and is used to correlate daily terms and legal terms collected from actual news articles. There are some types of links used to connect two terms. For retrieval and annotation, traversing of the thesaurus is controlled depending on types of terms and links.

Development of LonMaps is currently underway, and its overview, the details about the procedures of using the thesaurus, as well as place and date extraction are discussed in [14]. This paper shows some considerations about this system, description of news articles, name extraction patterns, and some experimental results of query processing and of extracting information items.

This paper is organized as follows. Section II describes design of components. Some mechanisms using a thesaurus are shown in Section III. Section IV shows sentence patterns for extracting information items. Some experimental results are shown in Section V. Finally, conclusion and future work are described in Section VI.

II. AN OVERVIEW OF LONMAPS

A. *Design of LonMaps*

When a news article is read, it is well-known that the '5W1H' (what, when, where, who, why, and how) of an

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-319-3

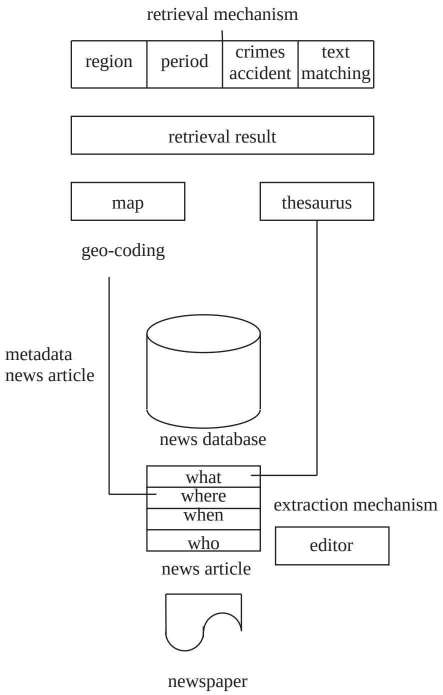


Figure 1. An overview of LonMaps.

```
<articles>
     <article id>  ; news id
          <issue year month day />  ; issued date
          <happen year month day />; occurrence date
          <newspaper which> </newspaper> ; 'which' is a distinction
     of morning or evening editions of the newspaper. The value is the name of a
     newspaper.
          <accident>  ; annotation by incidents.
               <man-assigned> </man-assigned>   ; man assigned
     incident list, if required.
               <captured> </captured>; a list of captured incidents in
     terms of legal terms.
          </accident>
          <address lng lat> </address>; 'lng' and 'lat' are
     'longitude' and 'latitude', respectively.  The value is the places of an address
     and/or location .
          <headline> </headline>;  the headline of a news article.
          <text> </text>  ; the body of a news article.
     </article>
          . . .
</articles>
```

Figure 2. A structure of a news article in XML.

incident are key items [15], [21]. We deal with four of these when using LonMaps: '4W' (what, when, where, and who).

The requirements of the system are as follows: (1) extraction of places where incidents occurred and indication of places on maps, (2) annotation of news articles in legal terms, and (3) clarification of the relationships between daily terms and legal terms using a thesaurus that consists of these two types of terms.

The features of the system are as follows:

- Four information items are captured from a news article: incidents, places, occurrence dates, and persons. These items are captured using the thesaurus and patterns that are defined by us. The dates, the places, and personal names are captured on the basis of patterns of the sentences. In many news articles, conventional patterns are typically used to describe the article.
- The thesaurus is provided for determining the relationship between legal terms and daily terms. This thesaurus includes not only terms representing incidents but also terms directly and/or indirectly related to incidents.

The thesaurus of the system is used to reduce the gap between daily terms and legal terms. For example, in retrieval of news articles, when a user specifies a query in terms that are a conventional representation of crimes or a usual representation of incidents, i.e., daily terms, some news articles cannot be retrieved, because they do not include recognizable daily terms. Daily terms are not always used for representing incidents, so legal terms may be required to retrieve the news article. In addition, it is necessary to annotate news articles in legal terms to describe incidents through formal and uniform representations. Since a news article may not include any legal terms, it is necessary to find suitable legal terms from a collection of daily terms that appear in the article.

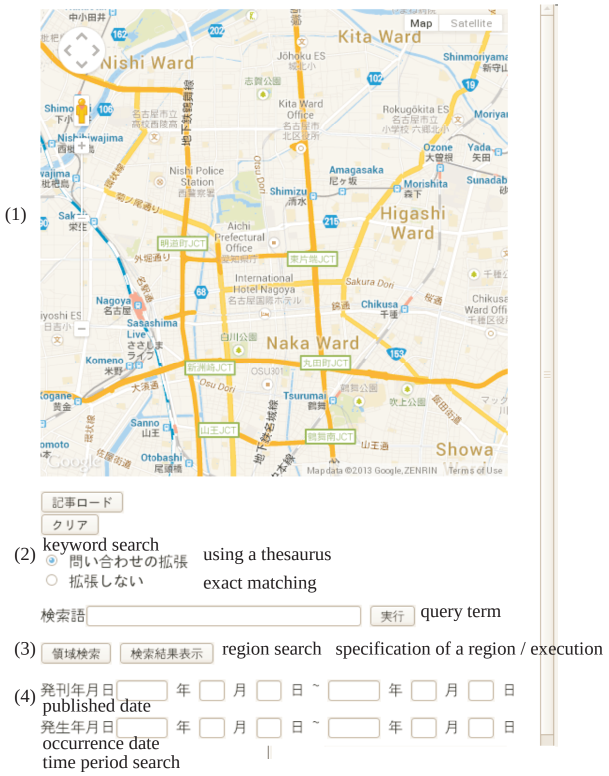
Figure 1 shows the overview the architecture of LonMaps. The system consists of a retrieval mechanism, an extraction mechanism, an editor, and a news database. The retrieval mechanism retrieves news articles. The extraction mechanism analyzes news articles. Moreover, an editor is used to edit a news database, e.g., defining a news article and modifying elements of news articles. The extraction mechanism is a main component of this editor.

B. Description of News Articles

A news database is a collection of news articles described in XML. Figure 2 shows tags for describing a news article. In this description, metadata of the news article are included. The metadata include elements such as publication date and the newspaper name. An entire set of news articles is indicated by `<articles>`. Each news article is indicated by `<article>`. The `id` attribute is an article identifier. `<newspaper>` indicates the name of a published newspaper. `<issue>` specifies a publication date. `<happen>` specifies a date when an incident occurred. `<accident>` specifies a type of incident. `<man-assigned>` denotes a list of incidents assigned by a person. `<captured>` is a list of incidents captured by the system. `<address>` specifies a place, and its longitude and latitude are obtained using geocoding provided by GoogleMaps. `<headline>` specifies the headline. `<text>` specifies the main text. The system tries to capture the values of `<happen>`, `<incident>`, and `<address>`.

C. Description of Queries

Three types of queries are available in LonMaps: a keyword query, a time period query, and a region query. Figure 3 shows the structure of a screen for specifying



● 問い合わせの拡張
○ 拡張しない

Figure 3. Structure of a screen for specifying a query in LonMaps.

queries. The first part of Figure 3 is a map. The second part is for a keyword query and a form to input query terms. If an expansion using the thesaurus is selected, given terms are expanded. Otherwise, when terms are given, exact matching is applied. A query is specified in query terms and logical connectives, i.e., and, or, and not. The third part of the screen is for a region query. A region is specified first, and then news articles within a specified region are sought. The fourth part is for a time period query. The occurrence date and/or a published date are specified as the query. After queries are specified, retrieval results are obtained. The retrieval mechanism shows the results on the map. Although markers are not shown in Figure 3, the retrieved articles are indicated by markers.

III. UTILIZATION OF THE THESAURUS

A. *Structure of the Thesaurus*

Daily terms are usually used and easy to understand in daily life, while legal terms are defined explicitly. To retrieve and annotate news articles, it is necessary to correlate these two types of terms. The thesaurus is constructed from a collection of news articles and the Japanese compendium of laws, whose conceptual structure is shown in Figure 4. There are two types of legal terms. One is the formal names of laws. Laws are defined in a hierarchical structure. The other is the legal representation of particular crimes and accidents. In contrast, daily terms are informal names for incidents, relevant terms, verbs, and conjugations. Relevant terms do not directly represent incidents but frequently co-occur with other daily terms. Moreover, the original form of a verb and its conjugations are considered daily terms.

Four types of links are defined: “is-a”, “general-term”, “associated-with”, and “conjugation-form”. Two legal terms are connected by “is-a”. A collection of legal terms is organized into a hierarchical structure. Two related terms are connected by the “associated-with” link to each other. This link is used for connecting not only two daily terms but also a daily term and a legal term. A verb and its conjugations are connected by “conjugation-form.” Moreover, if an inverse link of a link is required, it is set explicitly. Furthermore, “general-term” and “associated-with” are used for connecting legal terms and daily terms. The former is used for converting the legal term to corresponding daily terms. The latter connects two terms that often co-occur and are strongly associated with each other.

A part of the thesaurus is shown in Figure 5. Its root is the node “root”. At the second level, the root node has two children: “law” for defining laws and “accident” for defining accidents.

B. *Query Processing using the Thesaurus*

When a term defined in the thesaurus is given as a query term, the given term is expanded by traversing connected links from the given term to others in turn. When a given term is the name of a law, at first, descendants of the given term are obtained by following “is-a” links recursively. Next, by following “general-term” links connecting legal terms, daily terms are obtained. Finally, relevant terms and conjugations are obtained by “associated-with” and “conjugation-form” links. A set of these obtained terms is treated as the response to the query terms.

When a daily term is given as a query term, a legal term corresponding to the given term is first obtained. Next, the terms that are connected to the given term by “associated-with” and “conjugation-form” links are obtained. For each obtained daily term, the procedure for processing a daily term is then applied recursively.

C. *Annotation using the Thesaurus*

To find an incident of a news article, it is examined whether terms defined in the thesaurus appear, or not. Consequently, some daily terms are founded. Legal terms related to these daily terms are sought. For example, assume a daily term appearing in a news article, as shown in Figure 6. The related legal terms are captured by the inverse link of “general-term”. Moreover, certain terms connected to the given term with “associated-with” links are also captured. Consequently, legal terms are selected as candidate terms for annotation.

In this example, some candidate legal terms are found. If a term is positioned lower than others in the hierarchical structure of the thesaurus, the term is considered as an incident, since a lower term tends to represent a specific incident more precisely than higher terms. For example, assume “murder” and “robbery” are obtained from the

ISBN: 978-1-61208-319-3

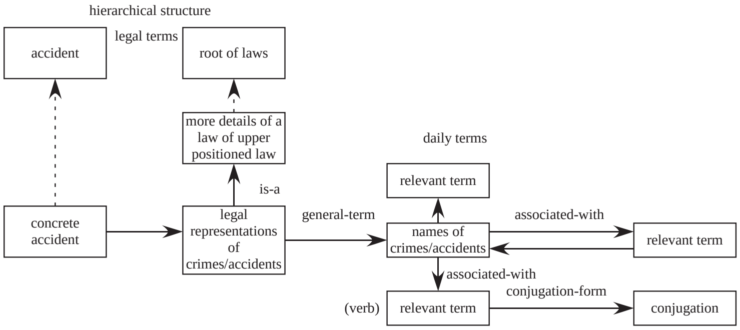


Figure 4. Components of the thesaurus.

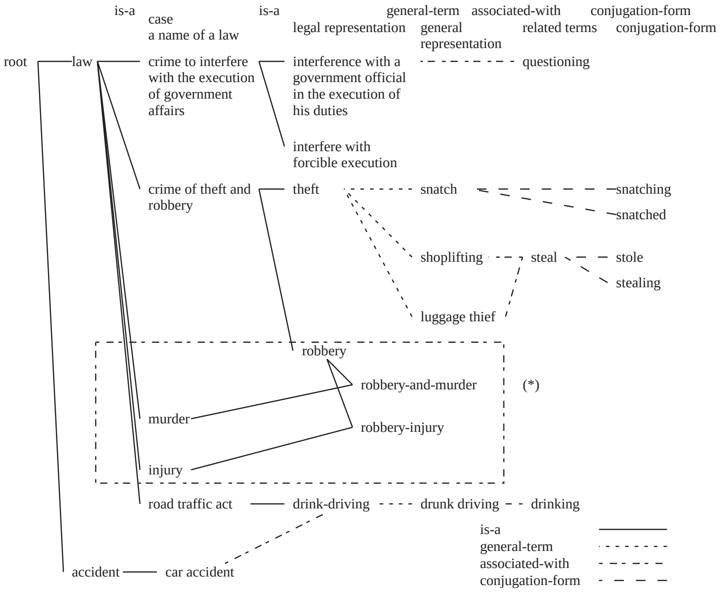


Figure 5. A part of the thesaurus.

original news article. The annotation mechanism tries to find more specific terms. Then, “robbery-and-murder” is selected rather than both “murder” and “robbery”, as shown in the part of the thesaurus marked by “(*)” in Figure 5. More details of the procedures used for traversing the thesaurus are described in [14].

IV. Extraction of Information Items

A. Extraction of Places

To extract places where an incident occurred, the following procedures are applied: (1) a part of a sentence that includes a place is extracted using sentence patterns, (2) area names are extracted by applying morphological analysis to the part of the extracted sentence, (3) when certain area names are omitted, they are complemented, and finally, (4) resolution of anaphora references is applied.

The patterns are shown in Figure 7. These patterns consist of three parts. The first part is a term representing a time period, the second a place, and the third a postpositional particle in Japanese. Here, words are separated by spaces in English but not in Japanese. For example, let a sentence including the place be “14日午前4時ごろ、名古屋市熱田区1番3の市道交差点で(at the municipal road crossing in 3, Ichiban, Atsuta, Nagoya around 4 a.m. in the morning on the 14th)”. The string between ごろ、(around) and で (at) is extracted. ChaSen [7], the morphological analysis system for Japanese is applied to the string for obtaining nouns included in this string. ChaSen divides a string into words by filling spaces between words, and searches for nouns. Some nouns from the beginning of this are obtained, as long as possible. The identified nouns are then concatenated. The resulting sequence is seemed as a place. For example, “名古屋市熱田区1番3” is extracted. Then, the complement of omitted area names and anaphora resolution are applied, if needed.

B. Extraction of Dates

Analogous to extraction of a place, extraction of an occurrence date is achieved on the basis of patterns and the published day. Figure 8 shows patterns for describing an occurrence date. First, a sequence of numbers and “日 (day)” are found. A modifier for representing time appears in the same sentence, e.g.,“午前 (in the morning)”. For example, let “14日午前 (in the morning on the 14th)”

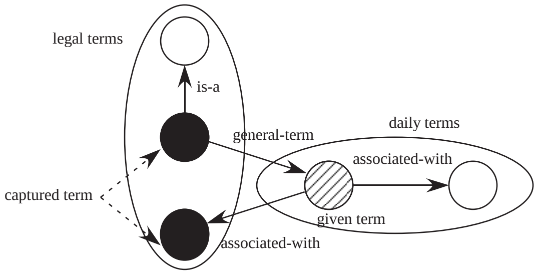


Figure 6. An annotation procedure when a daily term appears in an article.

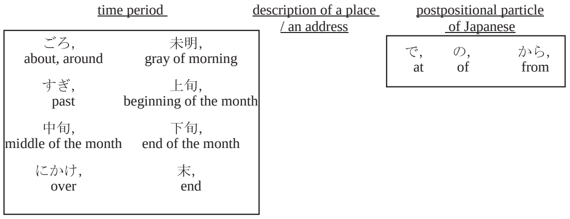


Figure 7. Patterns for extracting an occurrence place.

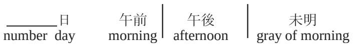


Figure 8. Patterns for extracting an occurrence date.

appear. The number 14 is extracted as an occurrence day. A month and a year are omitted in many cases. They are complemented on the basis of the published date of the news article, since the month of the occurrence day and the month of the published day are usually the same. The candidate day of the occurrence date and the published day of a published date are compared. If the candidate day is less than the published day, the month of the occurrence date and the month of the published date are the same. If the candidate day is greater than the published day, the month of the occurrence date is determined as the last month of the month of the published date in many cases.

C. *Extraction of Personal Names*

Personal names appear in a news article as victims, suspects, and other parties. Although such personal names are important in news, it is considered that presenting personal names in older news articles is unimportant over the long term. So, the system does not print personal names. To achieve this, personal names are extracted using patterns.

Figure 9 shows patterns that describe names of a suspect and/or a victim associated with crimes. Here, word order is different in English and Japanese. A personal name is a full name when the name is presented the first time. The full name is a sequence of his/her last name and first name in Japanese. A suspect and a victim are distinguished by the modifiers used with the names. Modifiers for a suspect and for a victim are, for example, 'suspect' and 'title', respectively. In Figure 9, path (a) is the most popular pattern. The sequence of this pattern is 'occupation', 'name', 'modifier (suspect or title)' and '(age)'. In (b), a comma between an occupation and a name is noted. In (c), a case particle is used. In (d), an address is shown. In cases where a name appears several times, the full name is omitted, afterwards. A personal name is referred by the last name and his/her modifier. Then, such references are treated as unprintable words. Moreover, as for occupation, some representations such as a worker, an office worker, or a therapist, are used.

TABLE I PRECISION OF QUERY PROCESSING USING A THESAURUS.

	precision
legal terms for crime names	0.92
daily terms for crime names	0.48
verbs in daily terms	0.84
nouns in daily terms	0.55

V. EXPERIMENTAL RESULTS

Query processing with the thesaurus was examined. As a query term, four types of terms were used: legal terms for crime names, daily terms for crime names, verbs in daily terms, and nouns in daily terms. Precision of their retrieval results were measured, as shown in Table I. The precision for legal terms representing crimes were better than the precision for other types of terms, since the information needs representing crimes in legal terms were more precise. The precision for verbs was better than the precision for nouns and crimes in daily terms. It appears that verbs were related to crimes more directly than nouns. Precision was worse for nouns in daily terms and crimes represented in daily terms. These terms were ambiguous and were not directly related to precise crimes. Recalls were high because the thesaurus was used to expand a given query term and many related terms were obtained.

Next, validation of annotation was measured. The results are shown in Table II. About one hundred news articles were examined for annotation. The average number of legal terms obtained using the thesaurus was 4.5 for one article. Among these terms, 2.8 suitable legal terms were included. The ratio of suitable terms to obtained terms was 0.21. This ratio was computed as (the number of common terms assigned by a person and by the system)/(the number of legal terms captured by the system). Person assigned terms included both legal and daily terms. Moreover, the ratio of suitable legal terms among obtained legal terms was computed to be 0.78. This was computed as (the number suitable legal terms captured)/(the number of legal terms captured). As described above, news articles were annotated by a person before extraction of legal terms using the system. Then terms that represented concrete incidents and appeared frequently in news articles were selected for annotation in many cases. It was difficult to annotate in legal terms since daily terms need to be converted to legal terms. The results of these two annotation experiments by a person and by the system indicate that annotation often involved daily terms when done by a person, whereas by the system, suitable legal terms were found from daily terms using the thesaurus.

The validity of places obtained from the appearance of written words was evaluated. Places where an incident occurred were extracted using description patterns, and places were obtained by reading a news article. When the

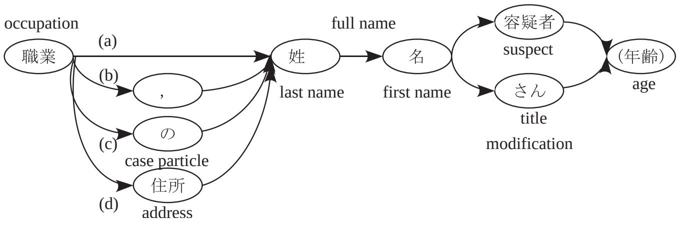


Figure 9. Patterns for extracting a personal name.

TABLE II Validity of annotation.

mean value of the number of captured legal terms	4.5
mean value of valid number of legal terms	2.8
ratio of suitable terms among captured terms	0.21
ratio of valid legal terms	0.78

TABLE III Validity of occurrence place extraction.

	validity
pattern matching and concatenating nouns	0.78
pattern matching, ChaSen, complementing place names and anaphora reference resolution	0.91

place extracted by the system was the same as that extracted by a person, the system's response was assumed as correct. The validity is computed as (the number of news articles where the places are captured correctly)/(the number of treated news articles). The result of this experiment is shown in Table III. The validity of extracted place was 0.78 when only patterns were used. By applying morphological analysis, omission complement, and anaphora resolution, the validity improved to 0.91.

Personal name extraction was also examined. In our experiment, about 96% of appearances of personal names were covered by described patterns. When a personal name represented in a full name was obtained, the personal name described by only her/his last name was captured. However, there were cases where several names of suspects appeared in one sentence. LonMaps does not currently have the capability to classify such patterns.

VI. Conclusion and Future Work

LonMaps is being developed as the first step toward building local news maps. This system provides mechanisms for retrieval of news articles and extraction of information items from them using a thesaurus and sentence patterns.

There are many cases that do not specify specific places in a news article. A place is specified as a division or an area of a town in many cases and this presents a problem in how to display a general area within a town. Moreover, the system captures incidents and retrieves articles on the basis of the thesaurus. When the thesaurus is used, the plausibility of relationships between terms is not introduced. When queries are processed and when incidents are extracted, the relationships between terms are treated as strict relationships. However, it is necessary to reflect plausibility of such relationships for introducing certainty of obtained news articles in retrieval and obtained terms in annotation.

References

[1] 47NEWS, http://www.47news.jp/, 12, 2013.

[2] Aichi Prefectural Police Department: http://www.pref.aichi.jp/police/gaitou/map/index.html, 12, 2013.

[3] J. Bhogal, A. Macfarlan, and P. Smith, "A Review of Ontology based Query", Information Processing and Management, Vol.43, pp. 866-886, 2007.

[4] D. Buscaldi, B. Magnini, "Grounding Toponyms in an Italian Local News Corpus", Proc. GIR, Article No. 15, 2010, doi:10.1145/1722080.1722099.

[5] C. Carpineto, G. Romano, "A Survey of Automatic Query Expansion in Information Retrieval", ACM Computing Survey, Vol. 44, No, 1, Article No. 1, 2012, doi:10.1145/2071389.2071390.

[6] K. Chang, Introduction to Geographic Information Systems, McGraw-Hill, 2010.

[7] ChaSen, NAIST Computational Linguistics Lab, http://chasen-legacy.sourceforge.jp/, 12, 2013.

[8] CrimeReports, http://www.crimereports.com/, 12, 2013.

[9] T. Furugori, R. Lin, T. Ito, and D. Han, "Information Extraction and Summarization for Newspaper Articles on Sasso-Jiken", IEICE Tran. Inf. & Syst., Vol. E86-D, No.9, pp. 1728-1735 2003.

[10] R. Goolsby, "Social Media as Crisis Platform: The Future of Community Maps/Crisis Maps", ACM Tran. on Intelligent Systems and Technology, Vol.1, No.1, Article No. 7, 2010, doi:10.1145/1858948.1858955.

[11] Google, http://www.google.co.jp/maps. 12, 2013.

[12] M. Haklay and P. Weber, "OpenStreetMap: User-Generated Street Maps", IEEE Pervasive Computing, pp. 12-18, Oct.-Dec., 2008.

[13] S. Ilarri, E. Mena, and A. Illarramendi, "Location-dependant Query Processing: Where We Are and Where We Are Heading", ACM Computing Serveys, Vol.42, No.3, Article No. 12, 2010, doi:10.1145/1670679.1670682.

[14] H. Ito, "An Overview of a News Map System for Local News in Newspapers", Frontiers in Artificial Intelligence and Applications, Vol. 243, pp.1031 - 1040, 2012.

[15] Z. Li, M. Wang, J. Liu, C. Xu, and H. Lu, "News Contextualization with Geographic and Visual Information", Proc. MM, pp. 133-142, 2011.

[16] A. Mehler, Y. Bao, X. Li, Y. Wang, and S. Skiena, "Spatial Analysis of News Sources", IEEE Tran. on Visualization and Computer Graphics, Vol.12, No.5, pp. 765-772, 2006.

[17] Metropolitan Police Department, http://www.keishicho.metro.tokyo.jp/toukei. 12, 2013.

[18] R. Paynich and B. Hill, Fundamentals of Crime Mapping, Jones and Bartlett Publishers, 2010.

[19] S. Sarawagi, "Information Extraction", Foundations and Trends? in Databases, Vol.1, No.3, pp 261-377, 2007.

[20] J. Strötgen, M. Gertz, and P. Popov, "Extraction and Exploration of Spatio-Temporal Information in Documents", Proc. GIR, Article No. 16, 2010, doi:10.1145/1722080.1722101.

[21] B. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling, "NewsStand: A New View on News", Proc. ACM GIS, Article No. 18, 2008, doi:10.1145/1463434.1463458.