Discovering Volatile Events in Your Neighborhood: Local-Area Topic Extraction from Blog Entries

Masayuki Okamoto and Masaaki Kikuchi

Corporate R&D Center, Toshiba Corporation
1 Komukai Toshiba-cho, Saiwai-ku, Kawasaki 212-8582, Japan
masayuki4.okamoto@toshiba.co.jp

Abstract. This paper presents a method for the detection of occasional or volatile local events using topic extraction technologies. This is a new application of topic extraction technologies that has not been addressed in general location-based services. A two-level hierarchical clustering method was applied to topics and their transitions using time-series blog entries collected with search queries including place names. According to experiments using 764 events from 37 locations in Tokyo and its vicinity, our method achieved 77.0% event findability. It was found that the number of blog entries in urban areas was sufficient for the extraction of topics, and the proposed method could extract typical volatile events, such as performances of music groups, and places of interest, such as popular restaurants.

Keywords: Hot topic extraction, hierarchical clustering, locality.

1 Introduction

Along with the spread of mobile terminals, such as cellular phones and mobile personal computers, there has been growth in location-based services (LBSs) [16], such as car navigation systems, personal navigation systems, and location-based recommendation. Such LBSs mainly deal with static POI (point of interest) data and usually do not pay attention to occasional or volatile events, such as performances of singers on street corners, and new topical spots. Usually, it is difficult to find out about such events, and even if one does, it could be after the event.

The features of volatile events include the following: one or more words about an event or topic are mentioned in blog entries frequently in a short period, and these words correlate with a specific place name. For example:

- *I enjoyed XX at the new YY restaurant near the ZZ station.*
- *XX event at YY was very exciting!*

With regard to the first feature, many techniques related to hot topic extraction such as a timeline analysis-based method [1] or the burst detection method [10], which

G.G. Lee et al. (Eds.): AIRS 2009, LNCS 5839, pp. 181–192, 2009.

detects the increase in the number of documents following the occurrence of an event [7], is often applied to news documents or blog entries. However, the number of webpages that mention a particular local event is often too small for the successful extraction of topics.

With regard to the second feature, one of the issues concerns determining the type of location name that would be useful. In urban areas in Japan, many places are near one or more train stations, and station names are possible keys for accessing local events. Moreover, the locality of each event, the fact that topic terms appear only in a few specific locations, can be used as well as the topicality of each event.

As one solution, this study proposes a topic extraction method that finds new or volatile events from time-series text data obtained from blog searches carried out using a query including station names.

Although there have been studies on clustering and topic detection for webpages or blog entries, transitory location-based events have not been dealt with. The purpose of this paper is to investigate this issue. In this study, a topic extraction method based on a two-level topic clustering technique for time-series documents [9], which was applied to Japanese EPG (electronic program guide) data in the previous report, is used. Moreover, the techniques of named entity recognition and locality calculation are introduced in this paper.

The remainder of this paper is organized as follows. Section 2 shows the flow of the local-area topic extraction method. Section 3 reports on experiments for evaluating our local-area topic extraction method. Section 4 shows related works.

2 Local-Area Topic Extraction

We define the problem of local event detection as a variation of the hot topic extraction task. In this section, the extraction method is introduced.

2.1 Overview

Figure 1 shows the process flow of the proposed topic extraction method.

First, blog entries are collected in advance. Entries are collected hourly in the RSS (rich site summary) format with a common blog search engine using query words including geographic names such as ‘Tokyo’ or ‘Akihabara.’ Then, the prefiltering module removes entries including NG terms usually used for advertising or offensive to public order and morals.

Second, document vectors are generated from the collected entries using morphological analysis, named entity recognition, and IDF-based weighting function, and topics are extracted using the hierarchical clustering technique. An agglomerative approach [8] in which the Euclidean distance is used as the distance metric is employed. Then, the topic words for each topic cluster are extracted using the C-value technique [6].

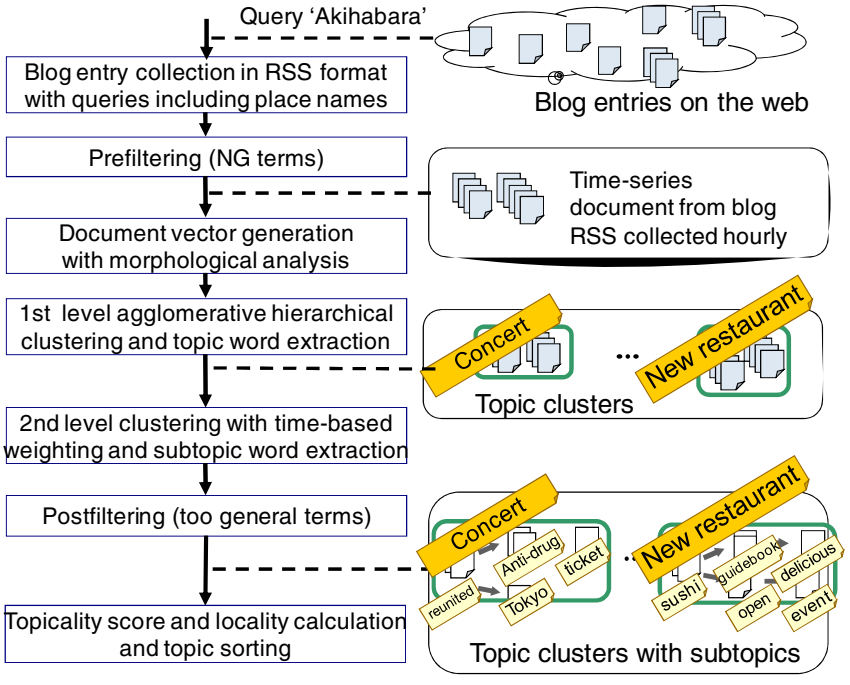


Fig. 1. Process flow of topic and event extraction in the two-level hierarchical clustering method

Third, second-level clustering is performed and subtopic clusters are extracted. A subtopic means a secondary important event or a shorter-term event in the topic. In this process, the time difference is additionally considered as a distance metric for detecting events within the same period. The subtopic words for each subtopic cluster are then extracted with the C-value technique, and the postfiltering process is performed to remove entries irrelevant to the topic, which are sometimes extracted when the topic word used is too general.

Finally, the topicality score is calculated for each topic. The topicality score is defined as the ratio of the short-term (2 days in this paper) average frequency of a topic to the long-term (3 days in this paper) average frequency of the same topic. If the Z test shows that the topicality score is significantly large for a topic, the topic is considered to be a hot topic. The locality score is also calculated for each topic. The locality score is defined as the number of locations at which a topic appears on the same day. A large locality score means that the topic appears in many locations and the topic is not a *local* event. Thus, events that share the same topic word are removed in this process.

Using the above method, the hot topics for each area are extracted. Figure 2 shows examples of extracted topics. This style is compact and applicable for a small display on a mobile terminal.

2.2 Topic Extraction

In the first-level clustering, cosine measure is used to calculate the similarity s_{ab} between two documents a and b as follows:

$$s_{ab} = \frac{\boldsymbol{d}_a \cdot \boldsymbol{d}_b}{\|\boldsymbol{d}_a\|\|\boldsymbol{d}_b\|}$$

where $\boldsymbol{d}_a$ and $\boldsymbol{d}_b$ are document vectors. Each term w in a document vector $\boldsymbol{d}$ is acquired by morphological analysis and named entity recognition (NER), and the corresponding weight is calculated with the IDF (inversed document frequency) as follows:

$$\mathrm{idf}(w) = \log_{10} \frac{N}{\mathrm{df}(w)} + 1$$

where N is the number of documents and $\mathrm{df}(w)$ is the number of documents in which w appears at least once. NER seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of time, quantities, monetary values, and percentages. NER systems have been created that use linguistic grammar-based techniques as well as statistical models [2]. We currently use linguistic rule-based techniques with over one hundred generic named entity classes covering person names, place names, organization names, numbers, and so on, originally developed for a question answering system [14].

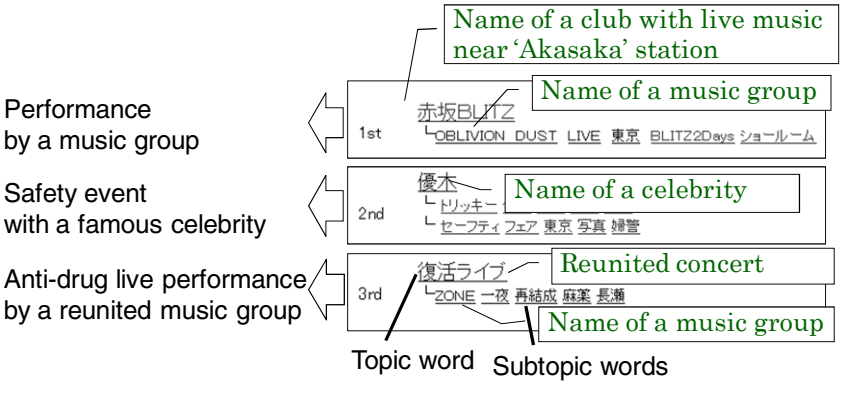


Fig. 2. Examples of extracted topics. Each topic includes a topic word and subtopic words.

Topics are extracted with the agglomerative clustering process by folding two clusters with the largest similarity score until the largest similarity score becomes lower than a threshold θ_T. When two clusters are folded, document vectors composing clusters are added as follows:

$$\boldsymbol{d}_{ab} = \boldsymbol{d}_a + \boldsymbol{d}_b \ .$$

2.3 Topic Word Extraction

A topic word is extracted from each topic cluster with a modified version of the C-value method [6]. The C-value of a candidate collocation cw is calculated as follows:

$$\text{C-value}(cw) = (\text{length}(cw) - 1)\left(\text{n}(cw) - \frac{\text{t}(cw)}{\text{c}(cw)}\right)$$

1where $\text{length}(cw)$ is the number of characters, $\text{n}(cw)$ is the number of times cw appears, $\text{t}(cw)$ is the total frequency of cw in longer candidate collocations, and $\text{c}(cw)$ is the number of those candidate collocations. In this paper, the number of characters is used for $\text{length}(cw)$ because we put higher priority on a long word than a collocation with short words, though $\text{length}(cw)$ is the number of words in the original version of C-value. Finally, a collocation with the highest C-value is selected as the topic word.

2.4 Subtopic Extraction

For the topic extraction described in Section 2.2, subtopics are extracted with the second-level clustering. In this process, a time-based similarity function is additionally introduced. A decaying function $W(a, b)$ for the similarity between clusters a and b is as follows:

$$W(a, b) = \exp(-\alpha(t_a + t_b)^2)$$

where t_a and t_b are the average times when the clusters a and b occurred respectively, and α is a constant value. Finally, similarity $\hat{s}_{ab}$ between the clusters a and b is

$$\hat{s}_{ab} = W(a, b) \cdot s_{ab} \ .$$

In addition, the IDF score for each word is recalculated on the condition that documents in the cluster are considered as the document set. Through this process, the weight for topic word decreases and the other words' weight increases. Thus, subtopic words are expected to be extracted. As described in Section 2.2, subtopics are extracted with the agglomerative clustering process by folding two clusters with the largest similarity score until the largest similarity score $\hat{s}$ becomes lower than a threshold θ_E, and then subtopic words are extracted by the same process described in Section 2.3.

2.5 Topicality and Locality Calculation

Hot topics appear more frequently in the nearest days. For example, a hot topic appears more frequently in the nearest three days than in the nearest seven days. Therefore, we need to determine whether or not the number of short-term occurrences is significantly larger than the number of long-term occurrences.

We assume that the long-term occurrence follows the rectangular distribution and determine the topicality by testing the null hypothesis 'the short-term occurrence also follows the rectangular distribution in the same way as the long-term occurrence' with Z test.

If the probability of occurrence of a document follows rectangular distribution, we can assume that the distribution of the number of documents that occurred in the latest n days v follows a binomial distribution with the occurrence probability $p = n/N$ and the trial number of times u. It is known that a probability function $Pr(v) =_u \mathrm{C}_v pv(1-p)^{u-v}$ can be approximated by a normal distribution $(up, up(1-p))$. Thus, the value of Z with the number of observed documents in the latest n days v_0 is calculated with

$$\mathrm{Z} = \frac{v_0 - up}{\sqrt{(\mathrm{up}(1-\mathrm{p}))}}.$$

We used Z score for the rank of each topic, i.e., the higher Z score a topic has, the higher the topic is ranked. In this paper, we used one day for the short term and three days for the long term.

For finding a local event, locality is also important. When a hot topic appears only in a place, it will be a local event, while a hot topic which appears in many places is not a local event but a pervasive event. In fact, events that share the same topic word are removed in this process.

Finally, topics sorted by topicality score are shown as Figure 2.

3 Experiments

We investigated to what degree extracted topics are appropriate as local events.

3.1 Topic Extraction Test Collection from Local Blogs

Before the experiment, we investigated what location is applicable. The number of relevant topics that could be extracted from local blogs was investigated. First, blogs were collected in advance for 856 train stations in Tokyo and Kanagawa in the first half of December 2008. Then, some of the stations were selected and topics were extracted to investigate the types of events that could be extracted.

Figure 3 shows the distribution of the number of blog entries. From this data it was found that there were 95 stations with 100 or more blog entries per day on average, a number considered sufficient to extract topics. Other stations had fewer entries, but it was still possible to extract small events that were referred to by a number of blog entries at around the same time. Therefore, the data in Figure 3 indicates that the number of entries is sufficient for detecting transitory events in large urban areas.

We introduce two kinds of location sets from top-rated locations: the vague location set and the non-vague location set.

The *vague location* is a location whose name is used not only for the local area but also for the wider area. For example, the word 'Tokyo' may be used to indicate the station name (Tokyo Station), the city itself, or the greater Tokyo area, which includes

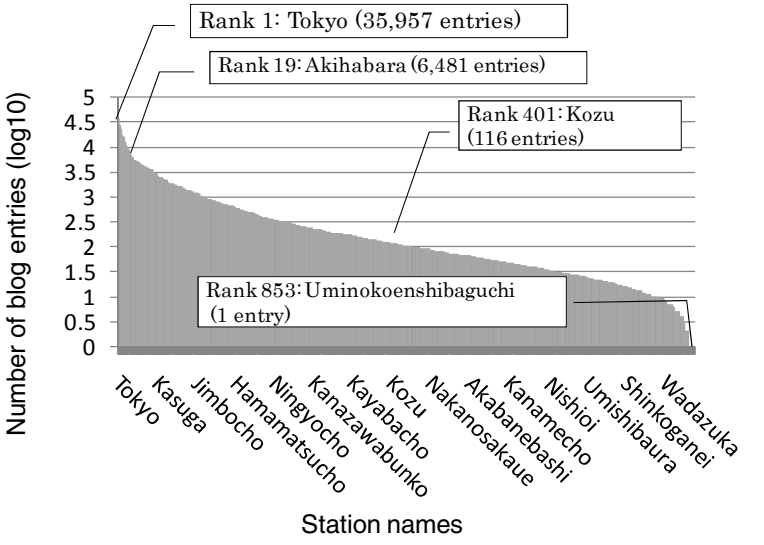


Fig. 3. Distribution of blog entries for 856 stations in Tokyo and Kanagawa in the first half of December 2008. One or more topics are extracted from 853 stations.

parts of the surrounding prefectures. Another example is the word 'Kawasaki.' This word may be used not only for the station name or the city itself but also for the last name of a person.

The *non-vague location* is a location whose name is used almost only for a local area. For example, the word 'Akihabara' may be used only to indicate Akihabara Station and its vicinity.

We chose 14 locations for the vague location set and 23 locations for the non-vague location set from locations with 10,000 or more blog entries in the first half of December 2008 (Table 1).

Table 1. Selected locations

Vague locations	Akasaka, Fuchu, Ginza, Kamakura, Kawasaki, Machida, Meguro, Nihonbashi, Ryogoku, Shibuya, Shinagawa, Tokyo, Ueno, Yokohama
Non-vague locations	Akihabara, Asakusa, Chigasaki, Daiba, Daikanyama, Ebisu, Harajuku, Ikebukuro, Jiyugaoka, Kagurazaka, Kichijoji, Korakuen, Minatomirai, Odawara, Omotesando, Roppongi, Shimokitazawa, Shinbashi, Shinjuku, Shiodome, Tsukiji, Yokosuka, Yurakucho

3.2 Extraction Accuracy

We investigated the extent to which extracted topics are recognized as local events.

Two subjects rated to what degree each topic seems to be a local event according to three levels: relevant (2); found relevant after checking the original documents (1); irrelevant (0). The relevance a user recognizes is calculated with only the rating 2, whereas the relevance of the system is calculated with the ratings 2 and 1. For each location-date pair, at most five events are rated.

Each subject determined relevance ratings based on whether or not he found both

- what kind of event or who are involved, and
- where (specific location) the event occurs or occurred.

Table 2 summarizes the results of our experiments for 764 topics from 37 locations for one week (from March 3, 2009 to March 9, 2009). In Table 2, the 'number of events' column means the number of extracted events, the 'relevance' column means the percentage of relevant topics (at least one subject rated 2 or 1), the 'relevance for user' column means the percentage of relevant topics (at least one subject rated 2).

Table 2. Relevance of events

Location	Number of evaluated events	Relevance	Relevance for user
Vague location	289	174 (60.2%)	101 (34.9%)
Non-vague location	475	335 (70.5%)	199 (41.9%)
Total	764	509 (66.6%)	300 (39.3%)

From Table 2, we found there is a large difference in relevance score between non-vague locations (70.5%) and vague locations (60.2%). It is because of the ambiguity of place names from vague locations, as we mentioned in Section 3.1. We also found there is a large difference between 'relevance' score and 'relevance for user' score. The reason is that some cue terms, such as building names or street names, are not extracted as a topic word or a subtopic word because these terms are written in various representations and it makes C-value of these terms smaller.

Table 3. Findability of events

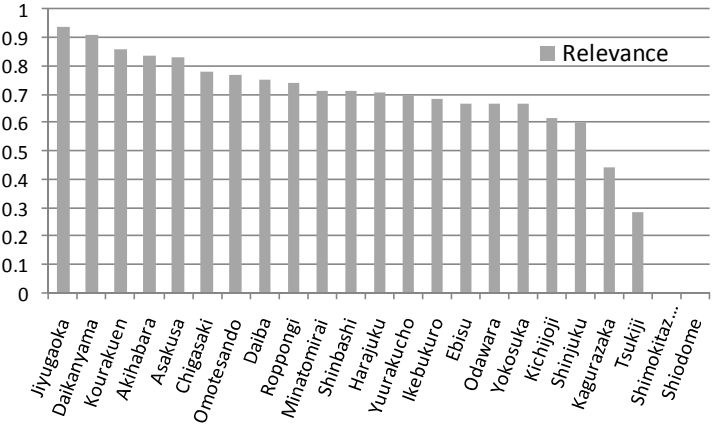
Location	Number of locations	Findability of events
Vague location	78	58 (74.4%)
Non-vague location	135	106 (78.5%)
Total	213	164 (77.0%)

Table 3 also summarizes the results of our experiments from the findability of events. In Table 3, the 'findability of events' column means the percentage for which a user recognized at least one topic as a local event (ratings 2) for each set of 5 events.

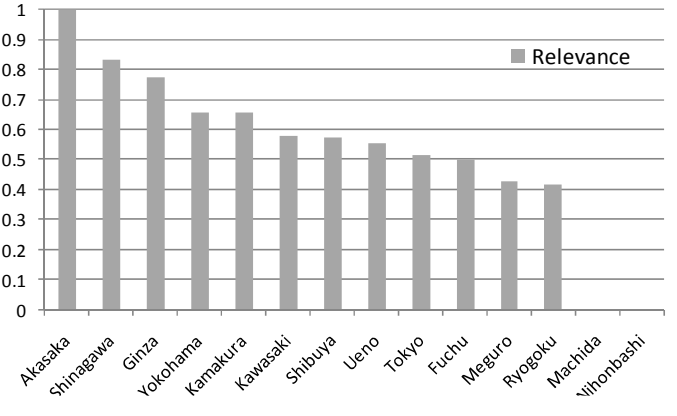
Table 3 shows that at least one subject found one or more events for 77.0% of sets of 5 events. This means that if we develop an event-recommendation application, a user can find some events by 77% for each location.

We also found the relevance score is much lower than we had expected because our previous report using EPG showed that 94.3% of topics are relevant [9]. This is because of the difference of document well-formedness between EPG documents and blog RSS documents.

Figure 4(a) shows the relevance for non-vague locations and Figure 4(b) shows the relevance for vague locations. From Figure 4, most non-vague locations achieved 0.6 or more relevance, while only 4 vague locations achieved 0.6 or more relevance.



(a) Non-vague locations



(b) Vague locations

Fig. 4. Relevance for each location

3.3 Feature of Extracted Events

Investigation of these stations revealed the following examples of typical events or spots:

- events held near a station (e.g., performances, such as a music group),
- Hokkaido souvenir festival (temporary souvenir shops),
- Ome marathon (sports event),
- Tokyo Auto Salon 2009 (a motor show), and
- Kenya Hara design event (an artist-related event in an art museum).

In addition to the above, restaurants on which bloggers wrote their impressions and incidents broadcast on news websites were also extracted.

However, problems related to place names were also revealed. One problem is the ambiguity of some place names from vague locations as mentioned in the above sections. Another problem is in the handling of the date information. The date information is useful for detecting events on a specific day, but it also functions as a hub and results in unrelated topics becoming connected. In addition, some real estate information was also found. This information, though it may sometimes be useful, is not used for extracting events. Our future research efforts will focus on solving these problems and improving the extraction results.

4 Related Works

There are many previous works on conventional topic detection and tracking (TDT) [1]. Hot topic extraction is a kind of topic detection task to find a topic that appears frequently over a period of time [3], and this technology is based on term-weighting theories, topicality calculation theories, and sentence modeling theories [5].

For term-weighting theories, TF-IDF [15] has been commonly used, and recently other methods such as TF*PDF [3] that focus on some feature of topic extraction have been proposed. For topicality calculation theories, burst detection techniques are widely used [10]. Another approach is to use the Aging Theory that models a topic's life cycle of birth, growth, decay, and death [4]. For sentence modeling theories, vector space models are widely used and achieved the best results [11].

In the topic detection task, there are some key characteristics of an event such as where the event occurred. Named entities (NEs) [2][14] play an important role and were successfully used [10].

Though we used common techniques for hot topic extraction, two new features are shown in this paper: one is the focus on local events using blog documents instead of extracting news topic from well-formed documents, and the other is the use of a two-level clustering method to extract detailed subtopics and subtopic words. For using location, a spatiotemporal theme pattern mining method is proposed [12]. This approach focuses on how the hotness of a given topic changes spatiotemporally from weblog data collected with a topic term such as 'Hurricane Katrina' with a probabilistic approach, where our approach is based on a clustering approach.

Moreover, the focus of this paper is not only on extracting local events but also on presenting the topic in a set of terms for mobile use.

There are also previous works on information presentation for mobile terminals with a small display, e.g., text summarization [13] and search result summarization [17]. We believe that the topic expression with topic words and subtopic words, described in this paper will be another successful presentation style for mobile terminals.

5 Conclusions

This paper introduced the possibility of topic extraction technologies for local-area event detection and new LBS applications. The hierarchical clustering method was used to extract topics and their transitions from blog entries collected with search queries including place names. An empirical investigation was also conducted into the types of event topics that could be extracted from blog entries collected using the names of 856 locations in the search queries. Moreover, according to an evaluation measuring relevance for 764 topics from 37 locations, we found the features of extracted topics. From these investigations, we confirmed that the proposed method was successful in extracting typical volatile events that would be difficult to extract with common web search engines or the burst detection method.

Our future work will include: improvement of filtering quality, classification of event types, and development of applications for mobile devices with geographic information systems.

References

1. Allan, J.: Topic Detection and Tracking: Event-Based Information Organization. Kluwer Academic Publication, Dordrecht (2002)
2. Borthwick, A.: A Maximum Entropy Approach to Named Entity Recognition. PhD thesis, New York Univeristy (1999)
3. Bun, K.K., Ishizuka, M.: Topic Extraction from News Archive Using TF*PDF Algorithm. In: Proceedings of International Conference on Web Information Systems Engineering (WISE 2002), pp. 73–82 (2002)
4. Chen, C.C., Chen, Y.T., Sun, Y., Chen, M.C.: Life Cycle Modeling of News Events Using Aging Theory. In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (eds.) ECML 2003. LNCS (LNAI), vol. 2837, pp. 47–59. Springer, Heidelberg (2003)
5. Chen, K.-Y., Luesukprasert, L., Chou, S.T.: Hot Topic Extraction Based on Timeline Analysis and Multidimensional Sentence Modeling. IEEE Transactions on Knowledge and Data Engineering 19(8), 1016–1025 (2007)
6. Frantsi, K., Ananiadou, S.: Extracting Nested Collocations. In: Proceedings of International Conference on Computational Linguistics (COLING 1996), pp. 41–46 (1996)
7. Fujiki, T., Nanno, T., Suzuki, M., Okumura, M.: Identification of Bursts in a Document Stream. In: Proceedings of International Workshop on Knowledge Discovery in Data Streams (2004)
8. Kamvar, S., Klein, D., Manning, C.: Interpreting and Extending Classical Agglomerative Clustering Algorithms Using a Model-Based Approach. In: Proceedings of International Conference on Machine Learning (ICML 2002), pp. 283–290 (2002)

9. Kikuchi, M., Okamoto, M., Yamasaki, T.: Extraction of Topic Transition through Time Series Document based on Hierarchical Clustering. Journal of the DBSJ 7(1), 85–90 (2008)
10. Kleinberg, J.: Bursty and Hierarchical Structure in Streams. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2002), pp. 91–101 (2002)
11. Kumaran, G., Allan, J.: Text Classification and Named Entities for New Event Detection. In: Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004), pp. 297–304 (2004)
12. Mei, Q., Liu, C., Su, H., Zhai, C.: A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs. In: Proceedings of International World Wide Web Conference (WWW 2006), pp. 533–542 (2006)
13. Otterbacher, J., Radev, D., Kareem, O.: News to Go: Hierarchical Text Summarization for Mobile Devices. In: Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006), pp. 589–596 (2006)
14. Sakai, T., Saito, Y., Ichimura, Y., Koyama, M., Kokubu, T., Manabe, T.: ASKMi: A Japanese question answering system based on semantic role analysis. In: Proceedings of Recherche d'Information Assistée par Ordinateur (RIAO 2004), pp. 215–231 (2004)
15. Salton, G., Yang, C.S.: On the Specification of Term Values in Automatic Indexing. J. Documentation, 351–372 (1973)
16. Schiller, J.H., Voisard, A.: Location-based Services. Morgan Kaufmann Publishers, San Francisco (2004)
17. Yasukawa, M., Yokoo, H.: Clustering Search Results for Mobile Terminals. In: Proceedings of Annual ACM SIGIR Conference on Information Retrieval (SIGIR 2008), p. 880 (2008)