Mining Future Spatiotemporal Events and their Sentiment from Online News Articles for Location-Aware Recommendation System

Shen-Shyang Ho
School of Computer Engineering
Nanyang Technological University
Singapore, 639798
ssho@ntu.edu.sg

Mike Lieberman
Center for Automation Research
Institute for Advanced Computer Studies
University of Maryland
College Park, MD, 20742
codepoet@cs.umd.edu

Pu Wang
Google, Inc
Mountain View, CA, 94043
puwang@google.com

Hanan Samet
Center for Automation Research
Institute for Advanced Computer Studies
University of Maryland
College Park, MD, 20742
hjs@cs.umd.edu

ABSTRACT

The future-related information mining task for online web resources such as news articles and blogs has been getting more attention due to its potential usefulness in supporting individual's decision making in a world where massive new data are generated daily. Instead of building a data-driven model to predict the future, one extracts future events from these massive data with high probability that they occur at a future time and a specific geographic location. Such spatiotemporal future events can be utilized by a recommender system on a location-aware device to provide localized future event suggestions.

In this paper, we describe a systematic approach for mining future spatiotemporal events from web; in particular, news articles. In our application context, *a valid event is defined both spatially and temporally*. The mining procedure consists of two main steps: recognition and matching. For the recognition step, we identify and resolve toponyms (geographic location) and future temporal patterns. In the matching step, we perform spatiotemporal disambiguation, de-duplication, and pairing. To provide more useful future event guidance, we attach to each event a sentiment linguistic variable: positive, negative, or neutral, so that one may use these extracted event information for recommendation purposes in the form of "avoid Event A" or "avoid geographic location L at time T" or "attend Event B" based on the event sentiment. The identified future event consists of its geographic location, temporal pattern, sentiment variable, news title, key phrase, and news article URL. Experimental results on 3652 news articles from 21 online new sources collected over a 2-week period in the Greater Washington area are used to illustrate some of the critical steps in our mining procedure.

ACM SIGSPATIAL MobiGIS'12, November 6, 2012. Redondo Beach, CA, USA

Categories and Subject Descriptors

H.4.m [**Information Systems Applications**]: Miscellaneous; I.2.6 [**Computing Methodologies**]: Artificial Intelligence—*Learning*; I.2.7 [**Computing Methodologies**]: Natural Language Processing—*Text Analysis*

General Terms

Design

Keywords

Event Extraction, Toponym Recognition and Resolution, Temporal Pattern, Sentiment Classification, Supervised Latent Dirichlet Allocation, Support Vector Machine

1. INTRODUCTION

Online web sources provide rich information that can be used as input for recommendation system. In particular, one can find substantial amounts of information related to future spatiotemporal events in news articles that are useful for future event recommendations. The task of "searching the future" is first discussed in [3, 4] when the temporal patterns in a news article are considered formal attributes in a text document . Generally, this task can be extended to any web content. Recently, this future-related information extraction task for web content has been getting more attention [11, 12] [and references therein] due to its potential usefulness in supporting decision making by an individual in a world where massive amount of new data are generated daily. Instead of building a

prediction model based on historical data to predict the future, one extracts future-related information from these data with high occurrence probability. To enable temporal annotation and tagging, Mani and Wilson [19] introduced a temporal annotation specification and an approach for resolving a class of time expression based on handcrafted and machine learned rules. Verhagen et al. [36] extended Mani and Wilson's work by developing a software tool TARSQI (Temporal Awareness and Reasoning Systems for Question Interpretation) that automated temporal annotation. Recently, Wang et al. [38] proposed a spatiotemporal knowledge harvesting framework to construct trajectory of individuals from spatiotemporal information extracted from news archives.

In this paper, we describe an information extraction framework to mine spatiotemporal future events from news articles. The application objective for this extraction framework is to provide a location-aware recommendation system with information on future events. In our context, a future event will be extracted only when *its location and time can be identified or deduced* from a news article. For example,

E1: Washington Post (27 September 2010): a possible tornado occurring at Clarksville, Maryland in an hour after a news alert (at noon).

E2: Baltimore Sun Blog (23 September 2010): possible traffic congestion near Merriweather Post Pavillion in Columbia, Maryland on 24 September 2010 due to an outdoor concert performance.

E3: Baltimore Sun (27 September 2010): a casino opening a few days ahead of schedule on 27 September 2010 at Perryville, Maryland.

To enhance the utility for the mined future events for the recommendation application, one needs to further identify the event sentiment. In other words, one needs to identify whether the event is a positive, neutral, or negative event based on either the event type or the sentiment expressed in the text. For example, E1 and E2 are labeled as negative events as they are events related to bad weather and bad traffic, respectively. For E3, it can be labeled as positive if one considers gambling as a recreational activity. Or, it can be labeled as negative if gambling is considered as a vice. The event sentiment of E3 can also depend on the news article content.

For application purposes, we include "near-past" event in our mining task due to its relevance to current and future. For instance, a traffic accident occurring one half hour ago on a highway may still be affecting the traffic flow. It may be a precautionary information for a user who drives past a recent crime scene where the culprit is still at large. These pieces of information are useful for personalized decision making and event recommendation based on user location.

The information sources that we use are feeds from news sources in a real-time setting. Towards this end, a record in the event database for the recommender system consists of six attributes: spatial (name, latitude, longitude), temporal (day, month, year, time [interval] (if available)), key phrase (text before and after a temporal pattern), sentiment, information source (URL), news article title, all of which are extracted from a news article. The record will be removed from the event database when it is a past event occurring at a fixed time interval from the current time.

A scenario of the application of the extracted information for a location-aware recommender system on a mobile device is as follows. John stays in the Baltimore suburban area. He carries a GPS-enabled mobile device with the recommendation application that can get future event recommendation based on his GPS location. On 27 September 2010, if John was driving near Clarksville, Maryland, the recommendation application would advise him to drive away (due to negative sentiment, see Example E1) and it would provide the news article title, key phrase, and URL as evidence to support the recommendation. If John was driving near Perryville, Maryland, he would be encouraged to visit the casino (if the sentiment is positive, see Example E3) with specific details available from the news article when he clicked on the URL.

Figure 1 shows three different events extracted from news articles from local news sources around the Greater Washington region with respect to the user geographic location (green car icon) determined using the user's network routing addresses. One "near past" event is an accident happened in the morning near to the user's current location. A "near past" event that *continues to have an effect to the present* is the shooting incident at Johns Hopkins Hospital in Baltimore (although no marker is shown in Figure 1) which resulted in a "temporarily restricted access to the [hospital] main building". A future event occurs two days from the present related to an electronic devices re-cycling event. The latitude and longitude values for an event is used to mark the event on the map while the geographic name is shown in the information box. Date, event sentiment, key phrase, and the news article title provide a brief description for the event. A URL is included for users who require more details.

The main contributions of the paper are (i) the introduction of a new future event information type extracted from news articles that is useful for location-aware recommender systems, (ii) a clear description and explanation of how such information is mined from news articles, and (iii) a description of how the extracted information can be used for location-aware recommendation purposes. In particular, we describe (i) temporal pattern recognition, (ii) toponym recognition and resolution, (iii) de-duplication, disambiguation, and matching for the spatial and temporal patterns, and (iv) event sentiment classification using statistical supervised learning. Data consists of 3652 news articles from 21 online new sources collected over a 2-week period (16-29 September 2010) in the Greater Washington area are used to illustrate some of the critical steps in our mining procedure. The main motivation for data collection in the Greater Washington area is to develop a location-aware future event recommender on mobile devices for the Greater Washington area. A key premise for localized data collection is that one can obtain more relevant and region-specific future events for the local users.

The rest of the paper is organized as follows. First, we briefly describe some related work on future event mining and recommender systems, temporal and spatial pattern extraction and disambiguation, and sentiment analysis. Then, we describe our event mining procedure in detail, highlighting some of the main subtasks such as near-past and future temporal pattern recognition, rule-based toponym recognition and resolution, rule-based spatiotemporal deduplication and disambiguation, and sentiment classification using supervised learning. Instead of presenting experimental results in a later section, we discuss our experimental results as we describe the subtasks within the mining procedure.

2. RELATED WORK

Currently, web-based recommender systems for news articles have been implemented, both commercially (e.g. Google news) and as research prototypes, and they have been extensively studied [29, 30, 33] (and reference therein). The news article recommendations are usually based on personal preference (e.g. news topic) and the general public interest in a social context (e.g., Twitterstand [31]). These types of recommendation rely on high-level text analysis

Figure 1: Examples of extracted near past and future events near to a user geographic location (green square car icon) on Google Map on September 16, 2010.

such as topic modeling and classification, and news article ranking based on social information sources. Minkov et al. [21] introduced a low rank collaborative approach for future event recommendation without previous feedback that was applied to scientific seminar recommendation. Ye et al. [41] proposed an approach to recommend locations for location-based social network. Levandoski et al. [13] proposed a generic location-aware recommendation system that uses location-based ratings to produce recommendations. The approach was tested on movie and location recommendations. Venetis et al. [35] proposed a framework that takes a user location and a collection of near-by places to rank places for recommendation purposes. Tekeuchi and Sugimoto [32] proposed a real-world recommendation system that makes recommendations of shops based on users' past location history. The system uses a place learning algorithm that efficiently find users' frequented places and assign the proper names. The discovered users' frequented shops are used as input to the item-based collaborative filtering algorithm for the recommendation system.

Jatowt et al. [11] proposed two methods for future information detection and summarization from news archives and the web. One method is based on future temporal expressions analysis; the other one depends on periodic pattern detection in historical data. Jatowt et al. [12] investigated the distribution of future event information on the web and analyze its major topics. Ling and Weld [18] described an information-extraction system that extracts temporal relations between times and historical events. Brants et al. [6] described a method to detect previously unseen events (news stories) using an incremental TF-IDF model.

A multi-level generative model that establishes relationships between latent topics and geographical regions was proposed and applied to geotagged microblogs [9]. Two graphical models designed specifically for text data to address textual interactions between named entities (e.g., persons, organizations, locations) and the topics were proposed and applied to news articles [22]. Similarly, [20] extended the probabilistic latent semantic analysis model to extract topics from text data with context information such as time and location. Wang et al. [37] also proposed a probabilistic graphical model to explicitly model the relationship between locations and topics.

Recently, Gao et al. [10] proposed a system for extracting events and their corresponding spatiotemporal context from photo images. Ye et al. [40] investigated the use of textual and geographic features of locations to identify relevant location to the main them of a travelogue. Popescu et al. [26] defined an event as an "activity or action with a clear, finite duration" in which a particular entity or object is the main focus of the events. An approach was proposed to extract such an event, and its corresponding set of actions, and audience opinion. Yan et al. [39] interpreted a news article as consisting of event-centered "atomic text snippets". They investigated the event snippet extraction problem and described a fine-grained news digestion framework for the extraction problem using semantic, syntactic, and visual features.

Sentiment analysis (or classification) deals with the "computational treatment of opinion, sentiment, and subjectivity in text" [24]. Previous work in sentiment classification focused on applications such as movie reviews [25] and product reviews [7]. Machine learning techniques such as naive Bayes, maximum entropy method, and support vector machines, were used for sentiment classification on various features such as unigram, bigram, word frequencies and presence, and parts of speech. It was suggested that sentiment classification is much more difficult than topic classification [25]. Cui et al. [7] also demonstrated that discriminative models such as support vector machines outperform generative models. Focusing on customer reviews of products, Ding et al. [8] proposed a holistic lexicon-based approach to handle opinion mining problem by exploiting external evidences and linguistic conventions of natural language expressions. This approach allowed the system to handle context dependent opinion words and hence handled major difficulties in existing algorithms. Recently, sentiment analysis has been extended to social networks and micro-blogging [23].

Geotagging is the process of identifying and disambiguating references to geographic locations. Amitay et al. [1] employ a hierarchical gazetteer approach to develop the "Web-a-Where" system with several enhancements for geo/non-geo and geo/geo disambiguations that improved on existing gazetteer approaches. Lieberman et al. [17] introduced the concept of local lexicon into geotagging news article to eliminate incorrect tagging of a less prominent geographic location appearing in a local new source by a prominent geographic location. In other words, a local lexicon set related to a news source has priority over a general lexicon set during geographic location resolution process.

3. EVENT MINING APPROACH

This section contains an overview of the future event mining task for news articles. We describe the four main subtasks, namely: future temporal pattern recognition, toponym recognition and resolution, spatiotemporal disambiguation and matching, and event sentiment classification.

3.1 Overview

Figure 2 is an overview of the future event mining task and its subtasks. The input data are news articles from online news source. The mining procedure consists of two main steps: recognition and matching. For the recognition step, (i) future temporal patterns, both absolute future times (e.g., October 16 2011) with respect to news article publication times and relative times (e.g., this Thursday, next week, tomorrow), are recognized, (ii) toponyms are recognized and resolved based on the application of seven heuristic rules [17], and (iii) news article title and URL are also identified as attributes to describe and provide additional information for a future event. For the matching step, (i) spatiotemporal disambiguation and de-duplication are needed to pair up toponyms and future temporal patterns to identify the future events, (ii) key phrases are extracted based on the position of the temporal pattern in the article, and (iii) sentiment for each event is derived using statistical supervised learning.

The output is a record in the event database consisting of six attributes: spatial (name, latitude, longitude), temporal (day, month, year, time [interval] (if available)), key phrase (text before and after a temporal pattern), sentiment, information source (URL), news article title.

3.2 Future (and Near Past) Temporal Pattern Recognition

Similar to the GUTime tagger in the TARSQI Project [36] and the TempEx tagger [19], we handle both absolute times and relative times. However, we only consider those temporal patterns that are *near past* or *future* with respect to a reference time, which is the publication timestamp of the news article. All identified temporal patterns are converted to a standardized format for comparison. We use the local time (e.g., Eastern Standard Time), ordinal date (Day 1 to 365/366 and year) format. A formal definition of a *near past* and *future* event is as follows.

DEFINITION 1. *Let t_p be the publication timestamp for a document d. A* **event** *is an incident described in d from which one can obtain from d its geographic location, and t the time information. A* **future** *event satisfies the property that $t > t_p$. A* **near-past** *event satisfies the property that $t + \epsilon = t_p$ such that ϵ is a small value depending on the context.*

In this paper, t is the ordinal day and ϵ is set to 0. In other words, our approach returns all events occurring on the same day as the publication timestamp as well as subsequent days.

Examples of the absolute and relative temporal patterns used for matching are shown in Tables 1 and 2, respectively. Absolute temporal patterns are identified and then compared to the publication timestamp to decide whether it is a future (and near-past) event or not. For relative future temporal patterns, the temporal pattern may include "this" which is used as an *adjective* to indicate the nearness in time such as "this Saturday" which is a future temporal expression or "this morning" which is ambiguous as to on whether it is past or future, but it is clear to be on the current day. Another adjective used in a future temporal pattern is "next" that indicates an "immediate following in time" such as "next week" and "next day". Again, we use the publication timestamp to compute the future date. For "next week", if we cannot identify the exact day, then we use the date for the Monday of that week.

	Temporal Pattern	Examples
1	<day> <month>	31 August; 31 Aug.
2	<month> <day>	February 13; Feb. 13
3	<year>	*current or next year*
4	<time><:/.><am/pm>	8:30am; 10.30pm

Table 1: Absolute Temporal Patterns

	Temporal Pattern	Examples
1	this <day_variable>	this morning; this Saturday
2	next <day/hour/week>	next week; next hour
3	in <integer> <hours/days/weeks_variable>	in three hours; in two days
4	tomorrow <with/without variables>	tomorrow afternoon

Table 2: Relative Future Temporal Patterns

After identifying absolute and future temporal patterns, we then identify interval patterns, if they exist. Examples of interval patterns are shown in Table 3.

	Temporal patterns	Examples
1	starting <at/.> <time> (<to/-> <time>)	starting 11.30am; starting 11.30am to 1.30pm
2	from <time> <to/-> <time>	from 11.30am - 1.30pm
3	<time> <to/-> <time> <am/pm>	3-7pm

Table 3: Temporal Interval Patterns

Based on the above matching rules used for extracting future temporal patterns, at least 282 news articles (7.72%), out of the 3652 news articles from 21 online new sources collected over a 2-week period (16-29 September 2010) in the Greater Washington area, contain at least [1] one near past (same day, but earlier than the publication time) or one future temporal pattern. At least 28 documents contain information on more than one near past or future events. One notes that by using the matching rules, one can achieve very high precision, but since the matching rules are not exhaustive and do not infer implicit future temporal patterns, the recall performance varies. By manual analysis, 150 of the 282 news documents (53.19%) contain information about future events. One can extract future event information from at least 4.11% of the news documents we collected. While web news from the main national/local

[1]We use "at least" to remind the readers that our approach is not an exhaustive search and that the result can be taken as an empirical lower bound for the number of future temporal pattern.

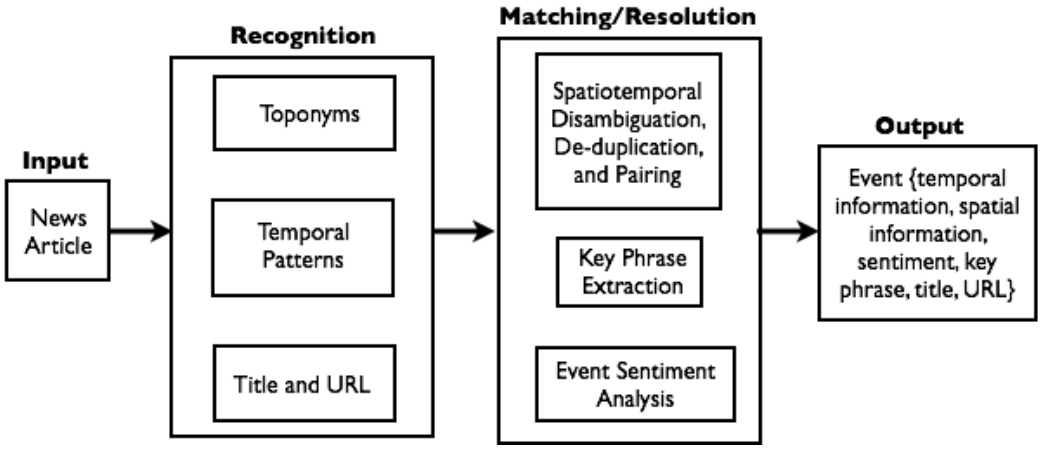

Figure 2: An overview of mining spatiotemporal future event from a news article.

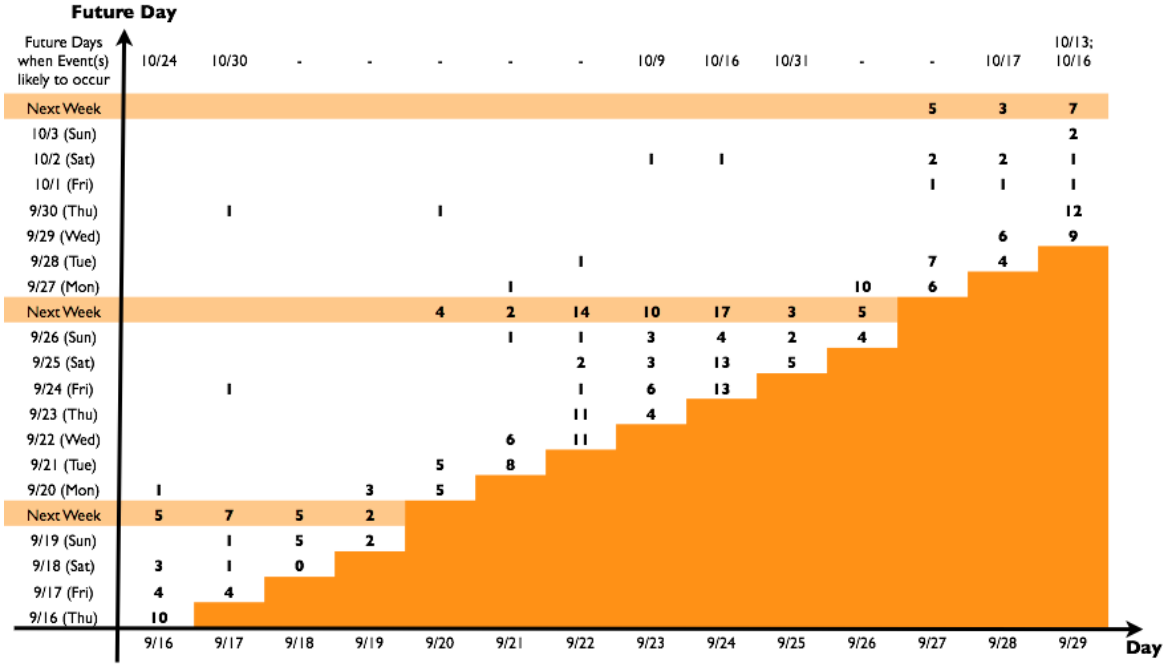

Figure 3: Number of extracted unique near past and future temporal patterns over 14 days of news documents.

broadsheet (Washington Post) make up more than a quarter of the future temporal patterns extracted, one interesting observation is that online local newspapers (Wonkette and Baltimore Star) contribute to another quarter of the extracted patterns.

Figure 3 shows the number of unique future temporal patterns that we extracted based on our future temporal pattern recognition approach for each day from September 16 to 29. Note the regular appearance of ambiguous "next week" information for each day. Figure 3 also shows that one can retrieve sufficient amount of near-future information from news sources for real-world applications. Also observes that in Figure 3 we do not show whether the temporal pattern extracted on the same day as the publication date is a near past or future temporal pattern with respect to the publication time. One can assume that a near past event is as relevant as a future event.

3.3 Toponym Recognition and Resolution

We now briefly describe the spatial information mining process (different from [2]) which closely follows the procedures in [17] and consists of two steps: toponym recognition [14] and toponym resolution [15]. The main idea is the definition of a local spatial lexicon, consists of a set of toponyms of close proximity, attaching to a news source, especially the local one. These local spatial lexicons related to their corresponding news sources are different from a global lexicon, consisting of prominent places that everyone knows, used by all news sources during toponym resolution. In most cases where the news sources and the news categories are local, the local lexicons supersede the global lexicon.

We use a hybrid toponym recognition technique consisting of Part-Of-Speech (POS) tagging, Named-Entity Recognition (NER), and rule-based heuristics recognition, followed by matching phrases that can be found in a gazetteer (database of geographic locations with geographic coordinates and associated metadata). The GeoNames gazetteer, an open gazetteer built from multiple gazetteers, is used for the toponym recognition.

The toponym recognition step consists of seven heuristic rules according to the following order.

1. Dateline toponyms: toponyms that appear at the beginning of the news article and provide the general geographic location for the news article.
2. Relative geography: phrases that define an imprecise geographic region based on proximity to another geographic location.
3. Comma group: toponyms that frequently occur together, separated by commas.
4. Location/container: toponyms occurring together that satisfy with a hierarchical containment relationship.
5. Local lexicon: a set of toponyms that is associated to a news source.
6. Global lexicon: a set of toponyms that is found in a curated set of well-known places.
7. One sense: all instances of a specific toponym in an article will have the same resolution.

Rule 7 is applied after each of the Rules 1 to 6. This propagates a resolved toponym to all later identical toponyms. If none of the rules can be used to resolve a toponym, then the toponym is not resolved instead of giving it a default sense such as the most common resolution or interpretation.

For a detailed evaluation of the toponym recognition and resolution procedures, see [14, 15, 16, 17, 28].

3.4 Spatiotemporal Disambiguation and Matching

After toponym recognition/resolution and future temporal pattern recognition, one needs to pair-up a toponym and a future temporal pattern to establish the existence of a future event. This matching process is defined by a function $f : X \rightarrow Y$, where X is the future temporal pattern set and Y is the toponym set. A future temporal pattern has to pair-up with a toponym, but not the other way. In other words, f (or the matching process) is injective and non-surjective.

There are five possible cases where one needs to consider when one performs toponym-temporal disambiguation and matching, namely:

1. $|X| = 0$ or/and $|Y| = 0$,
2. $|X| = |Y| = 1$,
3. $|Y| > 1, |X| = 1$,
4. $|X| > 1, |Y| = 1$,
5. $|X| > 1, |Y| > 1$.

For the first case, no future event is identified. For the second case, it is a direct one-to-one match.

For Case 3-5, one needs to de-duplicate the set of toponyms and the set of temporal patterns since the earlier future temporal pattern recognition and toponym recognition/resolution only tagged the phrases and identified their positions in the text. For the future temporal patterns, de-duplication is more complex as, for example, "afternoon" and "morning" on the same day are considered different, but "3 p.m to 5 p.m" and "afternoon" may be similar. In this paper, "morning", "afternoon", and "evening" are defined to be non-overlapping time intervals.

When one has a single temporal pattern and multiple toponyms (Case 3), there are three heuristic rules to consider in the list order, namely:

H1: Matching toponyms in the title.

H2: Matching the leaf node in a hierarchical containment relationship among multiple toponyms.

H3: Matching a toponym based on proximity of its occurrence to the temporal pattern in the text.

Note that any dateline toponym in the news article is ignored as it represents the news geographic source and usually not the event's geographic location. H2 comes before H3 as we place emphasis on the existence of related toponyms in a news article. When there are multiple leaf nodes, then H3 is used to decide the toponym match.

EXAMPLE 1. **Spatiotemporal Disambiguation/ Matching**
Key phrase with temporal pattern: "Hamlin will probably contribute on special teams in **this Sunday***'s game against the Pittsburgh Steelers."*
Toponyms: "Owing's Mill", "Baltimore", "Maryland", "Pittsburgh".
Matching: "this Sunday" and "Baltimore".

In the above example, "Owing's Mill" is a dateline toponym in the news article. Hence, it is ignored. The matching toponym is "Baltimore" and not "Pittsburgh" as H2 is preferred over H3.

When there is a single toponym and multiple temporal patterns (Case 4), all the temporal patterns will match to the single toponym due to the injective nature of our matching process. From our analysis result, we note that this case is very rare.

For Case 5, when a news article contains multiple toponyms and temporal patterns, we repeat the heuristic rulesets in Case 3 for each temporal pattern. One observation from our analysis result is that if there are a large number of toponyms or/and temporal patterns, then the text is likely to be a descriptive, non-factual article. Another observation is that a news article containing a weather forecast usually consists of a larger number of temporal patterns than toponyms.

After we have a list of toponym-temporal pattern pairs, we extract a corresponding key phrase containing the temporal pattern for each toponym-temporal pattern pair.

3.5 Event Sentiment Analysis

The event sentiment classification task involves the determination of the *anticipated user attitude or feeling towards an identified event*. A person has a *positive* feeling for a festival, an entertainment event, a concert or a sport event. On the other hand, an accident, slow traffic, poor weather or a crime induces *negative* feeling. An event such as an "electronic recycling event" (see Figure 1), which may not affect the user attitude or feeling towards it, will be *neutral*. For this classification task, we use the bag-of-word representation for a news article. First, we split all text into terms, and remove terms of pure digits. Then we use a porter stemmer [27] to stem every word. Next, we perform feature selection on words that removes words with document frequency less than 3. Each document is represented by a vector with each dimension corresponding to a unique word, with the entry of that dimension as the term frequency (TF), the number of times the word appeared in this document. *Classification is based on the simple and intuitive assumption that the sentiment of the news article corresponds to the sentiment of the events found in the news article*.

We applied two classification approaches, namely the supervised Latent Dirichlet Allocation (sLDA) [5] and the Support Vector Machine (SVM) [34], to the event sentiment classification task. LDA (a generative graphical model) and its variants have been popular topic modeling approaches for text data. SVM (a discriminative classifier) has strong theoretical justification and competitive practical performance.

We conducted experiments to compare the two approaches. First, we manually labeled and grouped 214 news articles (out of the 3652 news articles) into three sentiment categories: neutral, positive, and negative. Positive news articles contain news related to topics such as festival, entertainment, and sports. Negative news articles contain news related to topics such as crime, accident, poor weather, and traffic. The rest are included into the neutral category. Then, we performed the leave-one-out cross-validation (LOOCV) for each approach. For the LOOCV, we constructed a learning model or a classifier using 213 news articles and tested the sentiment classification accuracy for a single news article. We repeat this procedure 214 times leaving a different news article out for testing.

For sLDA, we fix the number of topic at 25. The number of iteration for the E-step and the M-step are 50 and 20, respectively. For the SVMs, we use the linear SVM, polynomial SVM, and the Gaussian SVM. We set $C = 10000$ and due to the unbalance nature of the data set, we weigh the C based on the ratio of the number of news articles in the different categories Table 4 shows the LOOCV performance of the approaches we used for the classification. Table 4 (right column) shows that the linear SVM and the Gaussian SVM are better than the sLDA. The third degree polynomial performs worst among the four approaches.

Classifier	Accuracy
sLDA	68.69%
Linear SVM	73.83%
Polynomial (3rd) SVM	42.52%
Gaussian SVM	73.36%

Table 4: News Articles Sentiment Classification.

One observes from [25] (Figure 1 and 2) that the human-based classifier (using positive and negative word lists) achieve an accuracy of 58% to 69% for the sentiment classification problem (in the movie reviews domain). If one assumes that this human baseline performance is applicable to our problem domain, then the sLDA performs almost as good as human, and the Linear SVM and the Gaussian SVM perform better than a human.

The identified events from a news article are assigned the news article sentiment. A recommendation system can then advise a user either to avoid a geographic location or to attend a future event based on this event sentiment. A challenging issue for our event sentiment task is the assignment of event sentiment when there are multiple future events in a single news article when the events have different sentiment.

4. FUTURE WORK AND CONCLUSIONS

In this paper, we describe a systematic approach for future event mining from web; in particular, news articles. The mining procedure consists of two main steps: recognition and matching. For the recognition step, we identify and resolve toponyms and future temporal patterns. In the matching step, we perform spatiotemporal disambiguation, de-duplication, pairing, and sentiment classification. Example 2 below shows an example of an identified future event that consists of its geographic location, temporal pattern, sentiment value, key phrase from text, news title, and news article URL.

EXAMPLE 2. **A Future Event Output Record.**
Spatial: (Lat, Lon) = (39.0993, -76.8483); Laurel
Temporal: Date = 2010 09 18; this Saturday, Sept 18.
Sentiment: Neutral
Key Phrase: "The Woman's Club of Laurel's annual yard sale will be held **this Saturday, Sept. 18** *from 8 a.m.-noon on Bond Mill Road between Brooklyn Bridge Road and Orem Drive.'*
Title: "West **Laurel***: Graduate needs instruments to fulfill his musical mission"*
URL: http://www.explorehoward.com/community/74907/graduate-needs-instruments-fulfill-his-musical-mission/

With the geographic location, date/time of occurrence, and the event sentiment, the identified future events from news articles are useful for location-aware future event recommendation system (see Figure 1).

There are many issues that require further investigations such as better spatiotemporal event sentiment analysis and temporal pattern extraction, and robust performance evaluation. In particular, we perform document (news article) level sentiment classification on the extracted spatiotemporal events. In other words, we assume that (i) the sentiment for the spatiotemporal event(s) extracted from a news article is reflected by the whole document content and (ii) if there are multiple events in a news article, they have identical sentiment. From our manually labeled data used for performance evaluation, Assumption (ii) seems to hold. To eliminate the two assumptions from spatiotemporal event sentiment classification task, a better approach would be to apply sentence level or text segment level sentiment classification. One would evaluate the sentiment of the sentence containing a spatiotemporal event. One difficulty for this approach would be the limited amount of information from one sentence. Text segment level approach could overcome this limitation by providing more information relevant to the spatiotemporal event.

A key part of recommendation systems, not discussed in this paper and require further investigation, is personalized recommendations based on some innate interest or relevancy metric. For instance, traditional collaborative filtering makes use of user rating history (e.g., movie ratings) to provide personalized movie recommendations. Similar idea can be applied here, for instance, based on user travel histories, or user event attendance histories (e.g., FourSquare check-ins), or online news viewing histories. This will avoid flooding all nearby users with a future event that some users may not be interested.

5. ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under Grants IIS-07-13501, IIS-08-12377, CCF-08-30618, IIS-09-48548, IIS-10-18475, and IIS-12-19023.

6. REFERENCES

[1] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In *SIGIR*, pages 273–280. ACM, 2004.

[2] W. G. Aref and H. Samet. Efficient processing of window queries in the pyramid data structure. In *Proceedings of the 9th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, pages 265–272, Nashville, TN, 1990.

[3] R. Baeza-Yates. Challenges in the interaction of natural language processing and information retrieval. In *CICLING 2004*, volume 2945 of *LNCS*, pages 445–456. Springer, 2004.

[4] R. Baeza-Yates. Searching the future. In *ACM SIGIR Workshop MF/IR*, 2005.

[5] D. M. Blei and J. D. McAuliffe. Supervised topic models. In *NIPS*. MIT Press, 2007.

[6] T. Brants, F. Chen, and A. Farahat. A system for new event detection. In *SIGIR*, pages 330–337. ACM, 2003.

[7] H. Cui, V. O. Mittal, and M. Datar. Comparative experiments on sentiment classification for online product reviews. In *AAAI*. AAAI Press, 2006.

[8] X. Ding, B. Liu, and P. S. Yu. A holistic lexicon-based approach to opinion mining. In *WSDM*, pages 231–240, 2008.

[9] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. In *EMNLP*, pages 1277–1287. ACL, 2010.

[10] M. Gao, X.-S. Hua, and R. Jain. Wonderwhat: real-time event determination from photos. In *WWW (Companion Volume)*, pages 37–38, 2011.

[11] A. Jatowt, K. Kanazawa, S. Oyama, and K. Tanaka. Supporting analysis of future-related information in news archives and the web. In *JCDL*, pages 115–124. ACM, 2009.

[12] A. Jatowt, H. Kawai, K. Kanazawa, K. Tanaka, K. Kunieda, and K. Yamada. Analyzing collective view of future, time-referenced events on the web. In *WWW*, pages 1123–1124. ACM, 2010.

[13] J. J. Levandoski, M. Sarwat, A. Eldawy, and M. F. Mokbel. Lars: A location-aware recommender system. In A. Kementsietsidis and M. A. V. Salles, editors, *ICDE*, pages 450–461. IEEE Computer Society, 2012.

[14] M. D. Lieberman and H. Samet. Multifaceted toponym recognition for streaming news. In *Proceedings of the 34th International Conference on Research and Development in Information Retrieval (SIGIR'11)*, pages 843–852, 2011.

[15] M. D. Lieberman and H. Samet. Adaptive context features for toponym resolution in streaming news. In *Proceedings of the 35th International Conference on Research and Development in Information Retrieval (SIGIR'12)*, pages 731–740, 2012.

[16] M. D. Lieberman, H. Samet, and J. Sankaranarayanan. Geotagging: Using proximity, sibling, and prominence clues to understand comma groups. In *Proceedings of 6th Workshop on Geographic Information Retrieval*, 2010.

[17] M. D. Lieberman, H. Samet, and J. Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *ICDE*, pages 201–212. IEEE, 2010.

[18] X. Ling and D. S. Weld. Temporal information extraction. In *AAAI*. AAAI Press, 2010.

[19] I. Mani and D. G. Wilson. Robust temporal processing of news. In *ACL*, 2000.

[20] Q. Mei and C. Zhai. A mixture model for contextual text mining. In *KDD*, pages 649–655, 2006.

[21] E. Minkov, B. Charrow, J. Ledlie, S. J. Teller, and T. Jaakkola. Collaborative future event recommendation. In *CIKM*, pages 819–828. ACM, 2010.

[22] D. Newman, C. Chemudugunta, and P. Smyth. Statistical entity-topic models. In *KDD*, pages 680–686, 2006.

[23] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*. European Language Resources Association, 2010.

[24] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2007.

[25] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *EMNLP*, 2002.

[26] A.-M. Popescu, M. Pennacchiotti, and D. Paranjpe. Extracting events and event descriptions from twitter. In *WWW (Companion Volume)*, pages 105–106, 2011.

[27] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[28] G. Quercini, H. Samet, J. Sankaranarayanan, and M. D. Lieberman. Determining the spatial reader scopes of news sources using local lexicons. In *Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 43–52, 2010.

[29] H. Samet, M. D. Adelfio, B. C. Fruin, M. D. Lieberman, and B. E. Teitler. Porting a web-based mapping application to a smartphone app. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 525–528, 2011.

[30] H. Samet, B. E. Teitler, M. D. Adelfio, and M. D. Lieberman. Adapting a map query interface for a gesturing touch screen interface. In *Proceedings of the Twentieth International Word Wide Web Conference (Companion Volume)*, pages 257–260, 2011.

[31] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *GIS*, pages 42–51. ACM, 2009.

[32] Y. Takeuchi and M. Sugimoto. Cityvoyager: An outdoor recommendation system based on user location history. In *UIC*, volume 4159 of *Lecture Notes in Computer Science*, pages 625–636. Springer, 2006.

[33] B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. Newsstand: a new view on news. In *GIS*, page 18. ACM, 2008.

[34] V. N. Vapnik. *The nature of statistical learning theory*. Springer, 2nd edition, 2000.

[35] P. Venetis, H. Gonzalez, C. S. Jensen, and A. Y. Halevy. Hyper-local, directions-based ranking of places. *PVLDB*, 4(5):290–301, 2011.

[36] M. Verhagen, I. Mani, R. Sauri, J. Littman, R. Knippen, S. B. Jang, A. Rumshisky, J. Phillips, and J. Pustejovsky. Automating temporal annotation with tarsqi. In *ACL*. The Association for Computer Linguistics, 2005.

[37] C. Wang, J. Wang, X. Xie, and W.-Y. Ma. Mining geographic knowledge using location aware topic model. In *GIR*, pages 65–70. ACM, 2007.

[38] Y. Wang, B. Yang, S. Zoupanos, M. Spaniol, and G. Weikum. Scalable spatio-temporal knowledge harvesting. In *WWW (Companion Volume)*, pages 143–144, 2011.

[39] R. Yan, L. Kong, Y. Li, Y. Zhang, and X. Li. A fine-grained digestion of news webpages through event snippet extraction. In *WWW (Companion Volume)*, pages 157–158, 2011.

[40] M. Ye, R. Xiao, W.-C. Lee, and X. Xie. Location relevance classification for travelogue digests. In *WWW (Companion Volume)*, pages 163–164, 2011.

[41] M. Ye, P. Yin, and W.-C. Lee. Location recommendation for location-based social networks. In *GIS*, pages 458–461. ACM, 2010.