Extracting Collective Expectations about the Future from Large Text Collections

Adam Jatowt
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto
606-8501, Japan
adam@dl.kuis.kyoto-u.ac.jp

Ching-man Au Yeung*
ASTRI
3/F Bio-informatics Centre
2 Science Park West Avenue, Hong Kong
albertauyeung@astri.org

ABSTRACT

News articles often contain information about the future. Given the huge volume of information available nowadays, an automatic way for extracting and summarizing future-related information is desirable. Such information will allow people to obtain a collective image of the future, to recognize possible future scenarios and be prepared for the future events. We propose a model-based clustering algorithm for detecting future events based on information extracted from a text corpus. The algorithm takes into account both textual and temporal similarity of sentences. We demonstrate that our algorithm can be used to discover future events and estimate their probabilities over time.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation

Keywords

future-related information retrieval, temporal information

1. INTRODUCTION

Future-related information can be found in textual documents such as newspapers, books and Web pages. They usually contain information such as plans, speculations, predictions and expectations and constitute an excellent source of what may happen in the future. However, the huge volume of information available makes it difficult for any person to detect all important future-related information manually, not to mention obtaining a clear picture of possible future scenarios.

In view of this challenge, we propose a framework for analyzing future-related information in a text corpus. Our goal is a system that can automatically extract future-related information from a large number of text documents, group related events and topics together and estimate the probabilities of these events based on the evidence found in the documents. We consider a query such as the name of a country, an enterprise, a celebrity or a well-known topic as the input of the system. To inform the design of our framework we first perform simple analysis of selected characteristics of future-related information (Section 3.3).

Our framework helps users understand the possible future scenarios associated with the entity given in the query based on the *collective image of the future* constructed from large textual collections. It should be noted however that our objective is not to predict the future per se. Instead, we propose a framework that can be used to provide the current view of the future that the society collectively expresses in large text collections.

Future forecasting has always been an important activity of humans and thus there should be many potential use cases for such a framework. In particular, a collective image of the future will be useful in various decision making scenarios. To name just a few, users wishing to invest in certain companies may be interested in their future perspectives and plans; fans of celebrities may want to know the schedules related to these persons; those who are considering moving to a certain city may want to know more about the development plans of the local government.

In the next section, we review related research works. Section 3 describes our dataset and presents some statistical analysis. In Section 4 we explain in detail our proposed framework. We describe several case studies and experiments in Section 5. Finally, we discuss some important issues regarding our work in Section 6, and conclude the paper in Section 7.

2. RELATED WORK

Our work is related to a number of areas including information extraction, information retrieval and data mining. While a lot of work has been done to process temporal information found in documents (e.g. [1][9]), relatively little research has been done specifically on future-related information. Baeza-Yates [2] is the first to talk about "future retrieval." He presents the idea of constructing a search engine

*Au Yeung contributed to this paper while he was at the NTT Communication Science Laboratories, Kyoto, Japan.

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

that extracts future temporal expressions from news articles and represents documents using tuples of time segments and confidence probabilities of future events.

Jatowt *et al.* [4] summarizes future-related information in both Web pages and news articles. Two methods are proposed to generate summaries of future events. Firstly, information about a certain entity is agglomerated by issuing queries containing object names and future dates (e.g. 2020, 2030) to search engines. Secondly, news articles mentioning a periodical event are analyzed in order to discover its periodicity, which can be then used to predict the future occurrence of the event. This work further develops the above concepts by taking into account the uncertainty in a piece of future-related information, and by establishing a unified and model-based approach to cluster and summarize future-related information.

Kanhabua *et al.* [6] propose ranking model for predictions that takes into consideration their relevance. *Time Explorer* [8] is a search engine that lets users search in the future and analyze future evolution of topics. Kanazawa *et al.* [5] describe methods for retrieval and validity analysis of future-related information which is not associated with explicit future dates. Our task is different from the above works as we focus on estimating the probability of events based on the aggregated evidence of future-related expressions found in text collections.

3. DATA ANALYSIS

3.1 Data Collection

We collect from Google News Archive[1] a large dataset of news articles published in the period of 1990-2010. News articles are obtained by issuing 61 queries with sub-queries specifying the time of publication to the search engine, and collecting all the search results. Our queries can be classified into the following four categories: (1) countries (e.g. Sweden, India); (2) companies (e.g. Panasonic, Toyota); (3) Persons (e.g. Johny Depp, Putin); and (4) others (e.g. Internet, science).

For each query we collect all search results with links to the original news articles and their timestamps. All news articles are downloaded and subjected to preprocessing. In some cases we are unable to collect the full text of the news articles due to subscription or other restrictions. In these cases, we collect the abstracts instead. On average each query results in 60,000 news articles or about 2.4GB text. In total, our dataset consists of about 3.6 million news articles or 145GB text.

Each news article is processed so that HTML tags and other non-content elements are removed. Next, we extract the core part of the news articles following simple heuristics based on identifying the largest chunk of text in each article. In addition, to restrict our analysis to English news articles, we perform filtering to remove articles written in other languages by using a text categorization algorithm based on n-gram matching [3].

3.2 Temporal Information Extraction

In many cases when a possible future event is mentioned in a news article, a temporal expression (e.g. "next year" and "in 2020") can be found in the same sentence. In order to understand what events would happen in the future, we need to first extract temporal expressions from the news articles. We use the GUTime tagger [7], which is the most popular, state-of-the-art temporal tagger for identifying and normalizing temporal expressions in text.

GUTime is able to detect both absolute and relative expressions. Absolute temporal expressions are defined as expressions that are unambiguously associated with a given time point or interval (e.g., 1st March 1998, Jan 2003). Relative temporal expressions, such as "next year" and "10 years later", require a reference time expression called anchor in order to be converted into absolute time expressions.

3.2.1 Temporal Expression Modeling

Temporal expressions returned by GUTime have usually single values assigned to them. In this work we modeled each temporal expression as time interval $[t_b, t_e]$ with a granularity of days. For example, "June 2012" is represented by $[01-06-2012, 30-06-2012]$. Temporal expressions that contain hours are simply converted to daily granularity. The time interval representation allows us to properly consider the meaning of temporal prepositions associated with temporal expressions, such as "after", "between", "in", "from" and "at".

For more ambiguous temporal expressions we apply simple rules to determine their boundaries. For example, seasons are roughly mapped to different periods as follows: "spring" to "March to April", "summer" to "May to August", "fall/autumn" to "September to October" and "winter" to "November to February". In addition, expressions such as "beginning", "middle" and "end" of a given period (e.g., week, month, year) are modelled as respective time spans covering one third of the period, and expressions such as "the first half of" and "the second quarter of" are converted accordingly.

In total we extract 13.1 million temporal expressions, out of which 2.7 million refer to the future, 7 million to the past and 4.2 million to the present (with respect to the publication date of the articles). Present temporal expressions are defined as expressions whose time interval, $[t_b, t_e]$, is as follows $t_b \leq t_a \leq t_e$, where t_a denotes the article timestamp. Temporal expressions related to the future and the past are determined by the following constraints, respectively: $t_a < t_b$ and $t_a > t_e$.

3.3 Distribution of Temporal Expressions

Intuitively, articles are more likely to refer to the immediate future than to the distant future, simply because there is more information about the former. Figure 1 shows the average distribution of temporal expressions in the news articles in relation to their timestamps. The horizontal axis denotes the distance of the value of temporal expressions from the timestamps of the news articles in which they appear. Negative (positive) values indicate that the corresponding temporal expression refers to the past (future) when compared to the article timestamp. The vertical axis indicates the average number of temporal expressions referring to a given time point normalized by the number of news articles published in one month.

First, we can confirm the intuitive expectation that the average number of future and past references decrease as the absolute time difference between their time reference points and the publication date of the articles increases. We also observe that the number of future references is on average

[1] `http://news.google.com/archivesearch`

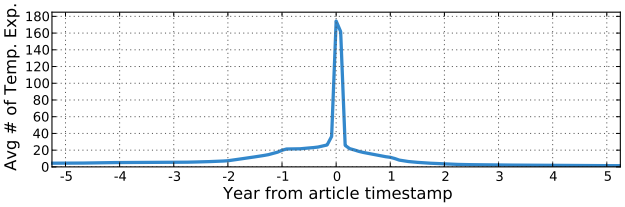

Figure 1: Distribution of temporal references.

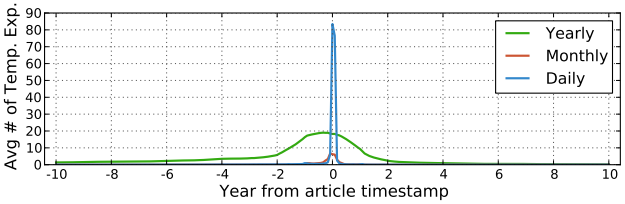

Figure 2: Distribution of temporal references according to their granularity.

smaller than that of past references (Figure 1). Another observation is that both the future and past parts of the curve decrease abruptly around the absolute values of two months, i.e. two months before and after the publication date of an article. The number of references does not change drastically beyond the 2 months from the present time.

We also investigate the usage of temporal expressions in terms of granularity (Figure 2). We categorize temporal expressions into daily, monthly and yearly depending on their minimal granularity. For example, "Monday", "tomorrow" and "17th February" are considered as daily-granularity expressions. "June" and "January 2003" are categorized into monthly expressions, while "2014" and "3 years later" are considered as yearly expressions. The graph follows our intuition that finer granularity expressions are used more often to refer to the nearer past and future. For example, it is relatively rare to use expressions of daily granularity for future or past time points that are further than 3 months from the news article publication date.

4. A FRAMEWORK FOR ANALYSIS OF FUTURE-RELATED INFORMATION

To analyze future-related information in a text corpus, we make use of a mixture model that clusters sentences based both on their *textual similarity* and *temporal similarity*. The idea is to obtain a set of clusters of sentences (presumably referring to some events in the future) such that each cluster refers to a coherent topic associated with a certain period in the future. At the same time, we also want to estimate the probability of an event (or more general a topic) happening in the future. Hence, each cluster will be associated with a probability distribution over time.

4.1 Probability Distribution of an Event over Time

When reporting events that have already happened, news articles usually provide the exact date and time of the events. However, for future events the temporal reference is usually uncertain and not all newspapers would agree on the exact time when a particular event will happen. For example, one newspaper may report that "Toyota is going to establish a new plant in Brazil in early 2012", while another may report that "Toyota's new factory in Brazil to be completed in 2011". While the two articles are probably referring to the same event, the temporal expressions are different. If we only consider the values of the temporal expressions, we will fail to cluster these two sentences together. To solve this problem, we map each temporal expression extracted from the text to a probability distribution over time. Specifically we consider the following four types of temporal expressions and suggest that they can be modeled by four different probability distributions based on their corresponding temporal modifiers.

1. A single time point (e.g. "In 2020 ...") is mapped to a Gaussian distribution.
2. An end date (e.g. "By the end of 2020 ...") is mapped to an increasing exponential distribution.
3. A start date (e.g. "from 2020 onwards, ...") is mapped to a decreasing exponential distribution.
4. A period (e.g. "... from 2015 to 2030.") is mapped to a uniform distribution.

4.2 Clustering with a Mixture Model

To group similar events together, we adopt a model-based clustering approach. We use the term "instances" to refer to the basic units in the clustering process. An instance consists of a sentence from which a temporal expression referring to the future is extracted, the probability distribution associated with the expression, and the surrounding text of that current sentence (the previous and the next sentences).

We approach this clustering task by considering a generative model of news articles. We assume that each instance extracted from the corpus is generated by first picking a topic cluster with certain probability, and then generating the terms and the temporal expression in the instance depending on the chosen topic cluster.

Formally, let the set of instances be D, and that each instance $d \in D$ is characterized by a bag of words W_d, and a probability distribution $P_d(t)$, which, as described in the previous section, reflects how likely the event mentioned in the instance would happen at different times. In addition, let Z be a set of topic clusters.

4.2.1 Basic Model-based Document Clustering

In a simple mixture model, in which documents are only characterized by a bag of words, the probability of a particular document being generated can be expressed as follows.

$$P(d) = \sum_{z \in Z} P(z) \prod_{w \in W_d} P(w|z)^{N_{w,d}} \quad (1)$$

where $N_{w,d}$ is the number of times a word w appears in document d, or it can be a score given to w with respect to d. A document is assumed to be generated by picking a particular topic cluster z and then generating terms from a probability distribution conditioned on the chosen topic.

However, we are not only interested in grouping instances that mention similar topics (e.g. putting all instances about Toyota's plans to build plants into one cluster). We are also interested in grouping instances that mention events

that are likely to happen at about the same time. In other words, having similar words is not the only criterion when clustering the instances. We also want to consider the *temporal proximity* between the instances. Hence, we extend the above mixture model as follows.

4.2.2 Considering Temporal Proximity

To perform clustering based on temporal proximity, we need to put instances with similar probability distributions in the same cluster. We let $G_d(t)$ be the probability mass function of an instance d, and let $G_z(t)$ be the probability mass function of the topic z. Also, we define $H(d|z)$ as the probability that, given a certain cluster z, d has a probability distribution defined by the function $G_d(t)$. We define $H(d|z)$ as follows.

$$H(d|z) = \frac{h(d,z)}{\sum_z h(d,z)} \tag{2}$$

where

$$h(d,z) = \frac{1}{D_{KL}(G_d||G_z)+1} \tag{3}$$

and $D_{KL}(G_d||G_z)$ represents the KL divergence between two probability mass functions. The intuition here is that $H(d|z)$ will be larger if the two probability mass functions are more similar, indicating that d has a high probability of belonging to the cluster z.

4.2.3 The Complete Model

To consider both the terms and the probability distributions of the instances, we propose the following mixture model which combines the two components described above.

$$P(d) = \sum_{z \in Z} P(z)\Big(\prod_{w \in W_d} P(w|z)^{N_{w,d}} \times H(d|z)^{\alpha}\Big) \tag{4}$$

where α is a parameter controlling the influence of temporal proximity in the clustering process.

Parameters can be estimated by using the EM algorithm. By iterating the E-step and the M-step until convergence, we obtain estimates of the parameters of the model. In the E-step, we estimate the probability $P(z|d)$ for each document.

$$P(z|d) \propto P(z) \times \prod_w P(w|z) \times H(d|z)^{\alpha} \tag{5}$$

In the M-step, we estimate the probability $P(w|z)$ and $G_z(t)$, the probability distributions over time for each topic, based on the soft clustering obtained in the E-step.

$$P(w|z) \propto 1 + \sum_w P(z|d) \times N_{w,d} \tag{6}$$

$$G_z \propto \sum_d G_d \times P(z|d) \tag{7}$$

An important parameter that needs to be tuned in our model is α. In order to take into account both textual and temporal similarities, the weight of the two components should be comparable. However, since the component of textual similarity depends on the number of words in the instance d, α should be chosen such that it is comparable to average number of words in the instances. Hence, we further define α as follows:

$$\alpha = \lambda \times \overline{N_d} \tag{8}$$

where $\overline{N_d}$ is the average number of words in the instances:

$$\overline{N_d} = \frac{1}{|D|} \sum_{d \in D} \sum_{w \in W_d} N_{w,d} \tag{9}$$

and λ is a parameter we can tune depending on how much weight we would like to put on the two components.

After the model is trained, we will be able to obtain the distribution of words for each topic ($P(w|z)$) as well as the probability distribution over time for each topic (G_z). These two pieces of information give us an idea of which topic is relevant in which period of time in the future.

4.2.4 Log-scale Timeline

In Section 3.3 we show that when a temporal expression refers to a time in the near future, it is likely to be a more specific one, while one that refers to a time in the far future is likely to be a rough indicator (e.g. only the year). To cater for this variation in the granularity of temporal expressions over time, we propose to transform the timeline from linear scale to log scale before performing the clustering process.

We assume that we are dealing with a discrete timeline, with the smallest division as a day. In a period given by $[t_1, t_2]$, we map a particular day t_x to a point t_l on the log-scale timeline spanning from 1 to 100 using the following equation:

$$t_l = \frac{\ln(t_x - t_1 + 1)}{\ln(t_2 - t_1 + 1)} \times 100 \tag{10}$$

After the transformation, days in the near future are more spread out and days in the far future will be more tightly packed together. In our experiments, we generate probability distributions for each instance after this transformation, and once we obtain the distributions of each cluster we transform the timeline back to a linear scale. In this way the events in the far future have distributions that are more spread out, indicating the higher uncertainty of these events.

5. EXPERIMENTS

Evaluating our proposed method is a challenging task because no benchmark dataset or ground truth is available. The results contain information about events or plans that were not yet realized at the time the news articles are published. Hence, it is difficult to perform quantitative evaluation. Instead, we carry out a set of case studies in order to demonstrate the effectiveness and some characteristics of our proposed method. We select three queries, namely "Germany", "Toyota" and "NASA", representing different categories of queries.

To perform the experiments, we extract news articles from our datasets that contain temporal expressions referring to a date or a period of time in or after 2011. Due to limited space here, we only present results with number of topics $|Z| = 50$ and $\lambda = 0.5$, by which we obtain reasonable results. Note that the output events are generated on the basis of 20-year old (1990-2010) collection of news articles, hence they are major events that have been usually "long-awaited" and are also scheduled in relatively far future.

5.1 Case Study: Germany

For the query "Germany", we extract 1,723 future temporal expressions. After the clustering process, quite a number of clusters are found to be associated with certain events

about Germany in the future. The three largest clusters are shown in the follow table, and the corresponding probability distributions (G_z) are shown in Figure 3.

ID	Size	Frequent Terms
1	124	energy, emissions, percent, government, power, country, climate, nuclear, renewable, million
2	39	billion, government, debt, city, economic, eastern, percent, frankfurt, euros, supply
3	27	munich, bid, year, city, new, events, world, olympic, winter, games

Table 1: Top three clusters for "Germany".

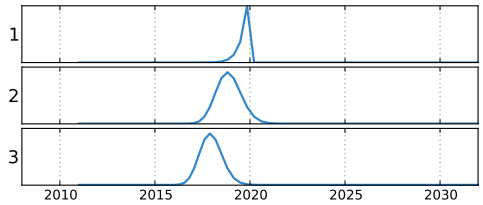

Figure 3: Probability distributions associated with the three clusters for "Germany".

Upon manual investigation of the sentences we observe that the first cluster is about energy and power consumption in Germany. In fact, this cluster is interesting because it consists of more than one topic. Firstly, there are reports about Germany's plan to cut emission to a certain level by 2020. There are also reports about Germany's plan to shut down all nuclear power plants in the country by 2020. The two topics are related and are usually mentioned together, therefore they are found in the same cluster. We also notice that the shape of the distribution of Cluster 1 is different from the other two (it is skewed to the left). This is because the year 2020 is usually mentioned in the way of "by 2020," therefore the distribution resembles an exponential rather than a Gaussian distribution.

On inspecting the instances belonging to Cluster 2, we find that they are mostly about western Germany's financial assistance to eastern Germany after unification. This long term project is scheduled to run until 2019, that is why we see a peak at 2019 in the probability distribution. On the other hand, the third cluster mostly contains sentences from news articles about Munich's bidding of the 2018 Winter Olympics.

5.2 Case Study: Toyota

For the query "Toyota", we extract 1,021 temporal expressions referring to the future.

ID	Size	Frequent Terms
1	129	new, company, car, hybrid, cars, motor, prius, vehicle, plans, electric
2	88	market, car, vehicles, plan, million, company, motor, year, share, vehicle
3	28	hybrid, vehicles, fuel, cars, company, market, global, year, million, prius

Table 2: Top three clusters for "Toyota".

We observe that the three clusters shown in Table 2 with their corresponding probability distributions in Figure 4 share

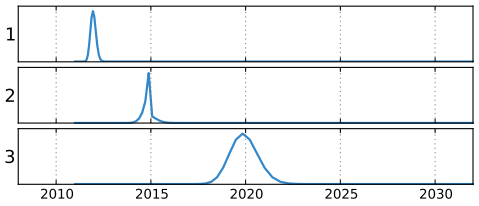

Figure 4: Probability distributions associated with the three clusters for "Toyota".

some common terms such as "hybrid", "electric" and "vehicles". However, they are actually referring to different events or plans about Toyota. The first cluster contains sentences in news articles about Toyota's plan to release both plug-in hybrid and battery electric cars in 2012, while the second cluster is about Toyota's plan to release hydrogen fuel cell vehicles by the year of 2015. Since many sentences in Cluster 2 are extracted from news articles that also discuss the projected market share of Toyota, terms such as "market" and "share" can be found in the cluster. Finally, Cluster 3 is about Toyota's plan to offer hybrid version for all of its models around 2020. This result suggests that our proposed method successfully group instances based not only on textual similarity but also the temporal expressions extracted from the text.

5.3 Case Study: NASA

Finally, we extract 2,499 temporal expressions referring to the future for the query "NASA".

ID	Size	Frequent Terms
1	276	moon, space, astronauts, return, mars, agency, president, lunar, program, new
2	139	space, launch, first, mission, shuttle, agency, flight, spacecraft, orion
3	82	earth, asteroid, space, apophis, mars, chance, hit, mission, propulsion, scientists

Table 3: Top three clusters for "NASA".

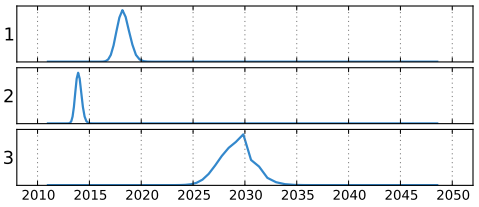

Figure 5: Probability distributions associated with the three clusters for "NASA".

This dataset mainly contains news articles about space programs and aerospace research conducted by NASA (see Table 3 and Figure 5). In the first cluster, we find news articles about NASA's plan to send manned spacecraft to the Moon again in 2018, thus we see a corresponding peak in that year in Figure 5.

Cluster 2 contains sentences from news articles about various plans of NASA to build spacecrafts. Ideally, different plans should be put into different clusters. However, since quite a lot of plans, such as a new space shuttle and the

Orion spacecraft, happen to be associated with the same period of time in the future (the year 2014), they are grouped into the same cluster. This suggests that depending on the topic we may have to give more weights to the terms instead of the temporal expressions in the clustering process.

Finally, Cluster 3 contains reports about the possibility that an asteroid called Apophis would hit the Earth in 2029. The probability distribution is much more spread out for this cluster, indicating high degree of uncertainty of this event. This is due to several reasons. Firstly, since we use a log-scale timeline during clustering, Gaussian distributions for years in the distant future as a result have a larger variance. Also, there were reports saying that the time of impact may be earlier or later. Finally, the cluster contains some sentences about the possible impacts of other asteroids in the surrounding years.

From the above three case studies, we can see that the proposed method gives reasonable and satisfactory results. In particular, we can obtain clusters that do not only contain instances that are about the same topic, but are also actually related to the same events. At the same time, we are aware of some limitations of our method. For example, in some cases when two events are highly related to each other and happen to be scheduled at about the same time, it becomes difficult to put them into different clusters, as in the case of "NASA".

6. DISCUSSIONS

The model proposed in this paper currently only takes into account some of the most important information found in the news articles. We believe several other factors can be utilized to achieve higher accuracy as shown below.

Information Freshness. Future-related information is inherently uncertain and continuously changing. Intuitively, the latest information should be more reliable than information published some time ago.

Source Credibility and Authority. Whether a piece of future-related information can be trusted is strongly related to the credibility and authority of the source. Hence, one may consider, for example, putting more emphasis on news articles from a major and reputable newspaper, and less on articles published in a blog.

Modal Expressions. It is common that modal expressions, which indicate different levels of certainty of the events, such as "may" or "is likely to" are used when news articles mention about future events. Hence, weighting instances by the modal expressions found near the temporal expressions is a potential method for improving accuracy.

Dependencies of Different Events. It is not uncommon to see sentences in the form of "A will occur if B and C happen". In other words, the probability of one event may be dependent on the probabilities of other events.

It should be noted that the above list is not exhaustive and is meant to provide a brief overview of some possible factors. In fact, our framework is flexible and can be easily extended to accommodate some of the above factors. For example, based on the credibility of the news sources and the modal expressions found in each instance, we can associate a score with each instance. We can then use these scores to weight the instances in the M-step when updating $P(w|z)$ and G_z.

While we show promising results, we are aware of certain limitations. We only focus on topics and events that are associated with temporal expressions. Therefore, a future topic/event is extracted only if it is mentioned together with some dates in the future. In fact, news articles also contain many references to future topics and events without mentioning when they are likely to happen [5]. The advantage of considering only time-referenced topics and events is that they allow us to estimate probability distributions over the timeline, and these references are more probable and credible as the time at which they will happen is more or less decided or agreed. On the other hand, we lose certain information about the complete picture of the future when we neglect topics or events that are not associated with any temporal expression.

7. CONCLUSION

Given the large amount of future-related information in text documents, it is possible to harness such data to construct a collective image of the future. We proposed a model-based framework for portraying collective images of the future, constructed based on numerous future-related information expressed in text collections. The framework supports users in understanding the currently planned or expected future events for a given query, be it a person, a company, a city or a country. We demonstrated its effectiveness on several examples of real world entities. Besides considering the additional factors mentioned in the previous section, we plan to improve our clustering algorithm and perform evaluation of larger scale in the future.

8. ACKNOWLEDGMENTS

We would like to thank Hang Li and Daxin Jiang for the discussions about this research. This research was partially supported by the MEXT Grant-in-Aid for Young Scientists B (#22700096) and by the Microsoft IJARC CORE6 Project "Mining and Searching the Web for Future-related Information".

9. REFERENCES

[1] O. Alonso, M. Gertz, and R. Baeza-Yates. Clustering and Exploring Search Results using Timeline Constructions. In *CIKM '09*, 97–106, 2009.

[2] R. Baeza-Yates. Searching the Future. In *MF/IR 2005*, 2005.

[3] W. B. Cavnar and J. M. Trenkle. N-Gram-Based Text Categorization. In *SDAIR '94*, 161–175, 1994.

[4] A. Jatowt, K. Kanazawa, S. Oyama and K. Tanaka. Supporting Analysis of Future-related Information in News Archives and the Web. In *JCDL '09*, 115–124, 2009.

[5] K. Kanazawa, A. Jatowt and K. Tanaka. Improving Retrieval of Future-related Information in Text Collections. In *WI '11*, 2011.

[6] N. Kanhabua R. Blanco and M. Matthews. Ranking Related News Predictions. In *SIGIR '11*, 755–764, 2011.

[7] I. Mani and G. Wilson. Robust Temporal Processing of News. In *ACL '00*, 69–76, 2000.

[8] M. Matthews *et al.* Searching through Time in the New York Times. In *HCIR '10*, 2010.

[9] A. Qamra, B. Tseng, and E. Y. Chang. Mining Blog Stories using Community-based and Temporal Clustering. In *CIKM '06*, 58–67, 2006.

[10] X. Zhu, Z. Ghahramani, and J. Lafferty. Time-sensitive Dirichlet Process Mixture Models. Technical report, Carnegie Mellon University, 2005.