# Improving Retrieval of Future-related Information in Text Collections

Kensuke Kanazawa, Adam Jatowt, Katsumi Tanaka
*Department of Social Informatics*
*Kyoto University*
*Kyoto, Japan*
*kanazawa, adam, tanaka@dl.kuis.kyoto-u.ac.jp*

***Abstract*—People often want to know expected future events related to given real world entities. For supporting users in the process of future scenario analysis, we propose several methods that enable to retrieve and analyze future-related opinions from large text collections. In particular, we focus on time-unreferenced predictions, which do not contain any explicit future time reference and hence are more difficult to be retrieved. As a second contribution, we propose estimating the validity of predictions by automatically searching for real-world events corresponding to the predictions. This kind of analysis aims to help detect predictions that are no longer valid as well as help estimating prediction accuracy of information sources.**

***Keywords*-future-related information retrieval; future scenario analysis;**

## I. Introduction

Lots of future-related information is currently available in text documents. Such information consists of future plans, schedules, predictions, speculations, expectations and so on. The Web or large text collections such as news articles are considered as a reflection of the society. It should be thus possible to harness them for reconstructing collaborative future image held by society. Such kind of social prediction would be generated based on opinions and expectations towards future shared by multiple users according to the "wisdom of the crowds" concept. The results of large scale analysis of future-related information could be useful for individual users in supporting decision making process as well as useful for many domains like futurology, sociological studies, marketing or business intelligence.

An effective future scenario analysis requires approaches that are characterized not only by high precision but also by high recall since widely known predictions may actually represent little value for users. Rather lesser known future opinions can offer more relative benefit. We assume in this work that the high variety of predictions is more important than their high accuracy and we thus seek approaches that result in high recall. The task of verification of predictions' probability is outside of the scope of this paper.

We propose several retrieval methods of future-related information from textual collections. We first demonstrate that future-related information which do not contain any explicit future dates is most common in news collections. Therefore we divert our efforts to design retrieval methods for time-unreferenced future-related information. This kind of predictions is more difficult to be extracted as there are numerous potential ways to refer to the future.

In addition, we categorize predictions into valid and obsolete ones. Valid predictions refer to events having some probability to happen at the reading time. On the other hand, we define obsolete predictions as the ones that refer to the already occurred events at the reading time. We propose two methods for estimating the validity of predictions according to the above definition. They could be used not only for automatically verifying past predictions but also for finding reliable predictors such as bloggers, journalists, newspapers or book authors.

The concept of using Web for forecasting future views is still new in IR and IE communities. Baeza-Yates [1] discussed the mechanics of future search engine - a service for returning relevant news articles that are directly related to particular future time periods. In our previous work [2] we demonstrated a query-dependent system for generating summaries of the most probable future outcomes concerned with particular objects. Kawai et al. [4] analyzed effective ways to automatically search for time-referenced future-related information in documents using machine learning. Time Explorer [5] is a service that allows for portraying future references in news articles and arranging them on a timeline.

The above works concern time-referenced future-related information, however, such kind of predictions account only for relatively small amount of the total future-related information. In contrast, we attempt to deal with time unspecified predictions and, in addition, we analyze their validity.

## II. Time-referenced and Time-unreferenced Predictions

Few documents concern only the future. Rather predictions appear as parts of documents. For example, they can follow reports on some latest event or analysis of current situation. Hence, we consider a sentence as a unit of analysis.

We distinguish two basic types of time related to sentences: *publication time* and *reference time*. The *publication time* is the time when a sentence is published. News and blog articles usually have the publication time directly expressed in their content. In this paper we focus on documents with known publication time.

The *reference time* is considered as the time associated with temporal expressions (e.g. “January 2013”, “next month”) contained in sentences. It is the calendar date which a temporal expression is converted to.

Time-referenced predictions can be easily retrieved by searching for sentences whose reference time is after the publication time. However the retrieval of time unspecified predictions is more difficult. Although using future referring terms such as “will” or “future” would seem to be a simple and effective method to search the predictions, the results usually have low recall and often contain noisy data. This is because such expressions may also refer to the present information or represent conditional structures. There are many predictions which do not contain fixed expressions such as “will” or “future” (e.g, “He hopes to ...”, “He plans to...”). In general one could use many ways to refer to the future.

We have analyzed the ratio of time-unreferenced and time-referenced predictions. We collected 1,000 snippets from New York Times Collection[1] for each of the following queries: “nasa”, “japan”, “apple”, “internet” and “car”. We divided the snippets into sentences and manually detected predictions. Next, temporal expressions occurring in the prediction sentences were automatically detected using GUTime[2]. GUTime is a tagger for identification and normalization of temporal expressions in text. GUTime is able to detect both absolute and relative expressions. It resolves relative temporal expressions using article timestamp and absolute dates that appear in the context of a relative expression.

Using GUTime we can count the number of time-referenced predictions, that is, sentences which contain at least one temporal expression which is newer than their publication time. In Table 1 we show the results. We can observe that less than 30% of predictions have any future reference time. Thus retrieving predictions based only on the comparison of their reference time and the query time can return, on average, about 30% of predictions, at least, in news article collections. The above analysis reveals the importance of retrieving time-unreferenced predictions for achieving high coverage of future-related opinions.

## III. Retrieval of Predictions

As mentioned before our aim is to achieve high recall prediction retrieval. time-referenced predictions can be easily retrieved by extending arbitrary query with any future dates and thus forcing search engines to return time-referenced, future-related information as shown in [2][4][3]. However as we have demonstrated in the previous section only on average 30% of total predictions can be retrieved by using this method.

Table I
The ratio of the time unspecified and specified prediction.

| Query | #Predictions | #Time specified Predictions | #Time unspecified Predictions |
|---|---|---|---|
| nasa | 329 | 134 (40.7%) | 195 (59.3%) |
| japan | 200 | 70 (35.0%) | 130 (65.0%) |
| apple | 122 | 36 (29.5%) | 86 (70.5%) |
| internet | 64 | 13 (20.3%) | 61 (79.7%) |
| car | 66 | 15 (22.7%) | 61 (77.3%) |
| Average | 156 | 29.6% | 70.4% |

In this section, we propose a method for retrieving general predictions based on calculating sentence similarities to future and past documents collections. With this approach we can retrieve both time-referenced and time-unreferenced predictions. We will discuss later how to generate the document collections. As for now let us assume that they simply provide documents containing past- and future-related information. When a target sentence to be evaluated is more similar to the future document collection than to the past one, it is regarded as a prediction. The underlying assumption is that the topics of a prediction sentence should be discussed in future documents. On the other hand, the target sentence that concerns events that already happened should be similar to topics discussed in the past documents. For calculating the similarities between the target sentence and large collections of documents we use selected terms, called characteristic terms.

Characteristic terms are grouped into positive and negative term lists depending on their relative frequencies in both the future and past collections, respectively. To generate the term lists, we calculate the bias of the term frequency between the future and past document collections. The terms which appear more frequently in one collection when compared to the other are considered to be characteristic terms.

More formally, to select a term $t$ as a characteristic term we first set null hypothesis that “the occurrence ratio of the term $t$ is same in the future and past documents collections.” For checking this null hypothesis, we test the log likelihood ratio. If the test rejected the hypothesis, the term $t$ is detected as a characteristic term.

Next, we compare the occurrence rates of the selected terms to categorize them into future and past characteristic terms. If the occurrence rate of term $t$ in the future document collection is higher than the one in the past document collection, the term $t$ is detected as a future characteristic term. In the same way, we detect past characteristic terms by finding terms that more frequently occur in the past document collection than in the future one. The weights $CV$

[1] http://developer.nytimes.com/

[2] http://www.timeml.org/site/tarsqi/index.html

of the characteristic term $t$ are as follows:

$$CV(t) = \begin{cases} LL(t) & \text{if } (LL(t) \geq \alpha \wedge \frac{\#Future_t}{\#Future} > \frac{\#Past_t}{\#Past}) \\ -1 * LL(t) & \text{if } (LL(t) \geq \alpha \wedge \frac{\#Future_t}{\#Future} \leq \frac{\#Past_t}{\#Past}) \\ 0 & \text{if } (LL(t) \leq \alpha) \end{cases}$$

$LL(t)$ is log likelihood value of the term $t$, $\#Future$ is the number of documents contained in the future document collection and $\#Future_t$ is the number of documents that contain the term $t$ in the future document collection. $\#Past$ and $\#Past_t$ are the corresponding numbers for the past document collection. $\alpha$ is a log likelihood test threshold. We set the significance level to 5%, thus $\alpha$ is 3.84. A given term is either a future characteristic term if it has a positive weight, or a past characteristic term if its weight is negative.

We rank sentences by the average weight of characteristic terms contained in these sentences. If a given sentence contains many highly-scored future characteristic terms and few highly-scored past characteristic terms, it is more probable to refer to a future event. Score $F(S)$ of a target sentence $S$ is defined as follows:

$$F(S) = \frac{\sum_{t \in Term(S)} CV(t)}{|Term(S)|}$$

$Term(S)$ is a set of characteristic terms contained in the target sentence $S$.

### A. Past and Future Document Collections

In the previous section, we assumed the existence of the past and future document collections used for detecting characteristic terms. In this section we describe the way to create such collections. We propose three approaches:

1) Topic-independent reference time method (TIRT)
2) Topic-dependent reference time method (TDRT)
3) Topic-dependent publication time method (TDPT)

The topic-independent reference time method (TIRT) creates document collections based on their reference time and independently of target sentence's topics. Documents are collected by querying news search engines with absolute temporal expressions such as general yearly queries (e.g., "2012", "2014", "1989", "1974") as explained in [3]. In particular, the query format to be issued to a search engine is defined as "$temp\_modifier$+(the)year(s)+$yyyy$" such as "in year $yyyy$", "in the year $yyyy$", "by the year $yyyy$". $temp\_modifier$ denotes a temporal preposition that is often used together with year dates, and $yyyy$ is a 4 digit number ranging from 2010 to 2050 for future and 1900 to 2009 for the past. We have prepared 39 different patterns of "$temp\_modifier$+(the)year(s)" to be used for every year. To ensure their correctness we manually inspected the top search results returned for each pattern for some selected future dates. By using the modifiers we decrease the possibility to receive content unrelated to future years (e.g., item numbers such as "page 2014").

We note that the above queries do not introduce any topical bias (topic independent method) and their results can be easily divided into future and past document collections. Documents having past reference time are grouped into the past document collection and documents having future reference time are grouped into the future document collection.

For each query we have captured the returned snippets and the titles of up to 1000 search results using Microsoft Bing search engine API[3]. After having finished the crawling we removed duplicate URLs. In total, we collected 1,044,224 unique search results.

The second approach, the topic-dependent reference time method (TDRT), creates topic-specific document collections, that are divided by their reference time. In this case documents are collected by querying news search engines with the topic represented by a short query.

We applied the same procedure as in TIRT to force search engine return sentences containing reference time. For example, for a topic "toyota" we issued queries such as: toyota "in year 2032", toyota "till the year 2017", toyota "to year 2032" etc.

In both TIRT and TDRT methods we actually use time-referenced predictions for the purpose of finding the time-unreferenced ones. The difference between TIRT and TDRT methods is that the latter is topic dependent, and thus the collections of future and past documents are assumed to be of the same topics as the one of the target sentence to be evaluated.

The third approach, the topic-dependent publication time method (TDPT) is also topic-dependent, thus the document collections are assumed to be of the same topic as the target sentence. However unlike the previous two methods it uses the publication time rather than the reference time to divide sentences into the past and future collections. The procedure is as follows. First, we issued topic query (e.g. "japan") to a news search engine to retrieve predictions. Documents that are relevant to the query were collected and divided into the future and past document collections by comparing their publication time with the publication time of the target sentence. Let $t$ be a publication time of a target sentence. When the publication time of a given document (sentence) is older than the target time $t$, the document (sentence) is grouped into the past document collection. Otherwise, it is added to the future document collection.

### B. Experiments

In this section, we describe the experimental results of the proposed prediction retrieval approach.

*1) Evaluation of Precision:* First, we calculated the precision. We collected 10,000 search results from Google News Archive[4] for each of the following queries: "apple",

[3] http://msdn.microsoft.com/en-us/library/dd900818.aspx
[4] http://news.google.com/archivesearch

Table II
PRECISION OF THE PREDICTIONS RETRIEVAL IN EACH METHOD.

| Method | Precision | Ratio of contained typical terms | Ratio of time specified pred. |
|---|---|---|---|
| TIRT | 50.7% | 73.0% | 17.9% |
| TDPT | 38.0% | 57.3% | 29.3% |
| TDRT | 28.8% | 48.1% | 52% |

“japan”, “car”, “internet”, “israel”, “nasa” and “panasonic”. We have then applied our approach and manually checked the 50 top-scored sentences. If a sentence referred to any future event, it was considered as a correct prediction. In this way we compared the precision using the three different methods for generating the future and past document collections discussed above. In addition, we calculated the ratio of sentences that contain typical future expressions: “will”, “future”, and “plan” inside the true predictions. This number shows how many correctly detected sentences could be discovered by simply using the typical future referring expressions. Note that we omited many other expressions that could be effectively used to retrieve future-related information. Also using the above terms may return false positives like “He has strong will” or “Company plan turned out to be success”. Nevertheless, the results should already be indicative of the relative extent to which the different methods can improve future-related information retrieval. Table II shows the results. We can see that TIRT approach has the highest precision. Although TDRT and TDPT methods have lower precision, the ratios of the sentences which contain the typical future terms are lower than the one for TIRT.

In addition, we also reported the ratios of the time-referenced predictions in the last column of the table. They are rather low for all the three methods. As we intend to retrieve wide range of any kinds of predictions and not only time-referenced ones, the proposed approaches fulfill our objective.

We have also calculated the duplication rates between the proposed methods within the 50 top-scored results. The duplication rate between TIRT and TDRT was 15%, the ones between TIRT and TDPT was 11.7% and between TDRT and TDPT was 17.3%. The low duplication values suggest that the combination of these approaches should result in improved performance.

*2) Evaluation of Recall:* In the second part of the experiments, we measured the coverage of the proposed method. As the sizes of our datasets are quite large, it is difficult to evaluate the recall. Thus we propose randomly choosing sentences to approximately estimate the recall. We randomly selected 50 predictions for each query. Considering these sentences as the answer set, we could approximately calculate the recall by counting the number of the answer sentences contained in the top-scored $N$ sentences. We show the evaluation results in Figure 1. TIRT approach produces the best results and can detect 80% (40 sentences) of the sample answers within almost 50% of the entire dataset. However, the remaining 20% sentences cannot be detected until almost the entire dataset is used. In these sentences, the future is not directly expressed, e.g. “NASDAQ Japan threatened to leave the Osaka Securities Exchange for another Japanese stock market.” This example sentence is mainly about the current action of NASDAQ, but it refers to the future event, that is, leaving the Osaka Securities Exchange.

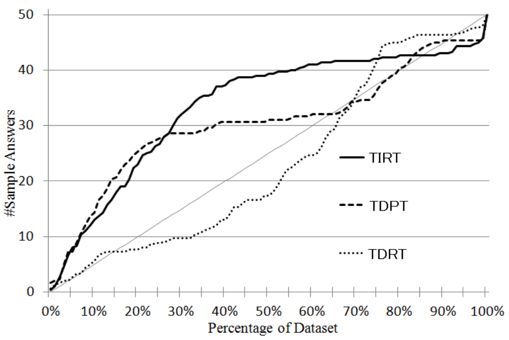

Figure 1. The number of detected sentences contained in the sample answer set.

The publication time based approach detects sentences that contain recent and often hot topics. However, not all the sample answers are about hot topics. Thus the number of sentences detected by the TDPT approach contained in the sample answer set increases only little in the 30% to 70% range of the dataset. TDRT performs the worst due to relatively low number of time-referenced predictions that could be used.

## IV. RETRIEVAL OF PREDICTED EVENTS RESULT

In this section, we introduce a new research problem of validating predictions.

When we see past predictions, we often want to know their results. Then we may ask the following questions. Have the predictions come true? Are they still valid or rather refer to the so-called “past future”? When did the predictions come true? What part of the predictions was true or false? If the predictions did not come true, what alternative events did happen? If we can find result sentences for the predictions (sentences containing evidence of predicted events), then we could have more information about the past predictions and their accuracy.

We should note that our definition of invalid predictions considers only the case when the predictions have been invalidated by the event’s occurrence (i.e. the predicted event has already occurred). It does not consider the case when the expected event was canceled or when another, newer prediction invalidated the old one. We leave these cases for future work.

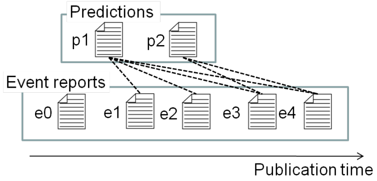

Figure 2. Calculating prediction-event's link strength.

The retrieval of prediction's result is defined as a link detection problem in a given dataset. As we consider two groups of nodes, the predictions and the reported events, the objective is to find a mapping of a node in one space to node in the other space. Our objective is to detect whether two sentences, where one is a prediction and the other is an event report, refer to the same event or not. Formally, we consider the pair $(p, e)$ where $p$ denotes the prediction and $e$ is the result candidate of the prediction $p$. For all the pairs we calculate the strength of their prediction-event report link $PRLinkScore(p, e)$ where $PRLinkScore$ is the prediction-event report link strength calculation function. We make a requirement here that considered event reports should have timestamps newer than the timestamps of the predictions as, obviously, event reports cannot be older than their predictions (see Figure 2). We propose two methods for calculating $PRLinkScore(p, e)$. Note that simple matching of predictions and event reports by calculating cosine similarity is not sufficient as predictions are often short, ambiguous and dynamically change over time. There is large syntactical difference between the expressions contained in predictions and their results. Our proposed methods are extensions of this baseline approach.

### A. Approach Based on Similar Predictions

Since predictions, due to their nature, are inherently ambiguous and dynamic, there is term appearance mismatch between the content of predictions and their results. It may be caused not only by different words or synonyms used in the both texts but also by the evolution of the expectations towards the predicted event. To fill the gap between the predictions and their results, we use similar predictions to the target prediction.

First, we detect similar predictions by calculating cosine similarities between the target prediction and all the other predictions. Then we calculate the prediction-event link strength considering the similar predictions. The equation of prediction-event link strength is as follows.

$$\begin{aligned} PRLinkScore(p, e) &= CosSim(p, e) \\ &\quad + \sum_{p' \in Similar_p} CosSim(p', e) \end{aligned}$$

$$Similar_p = \{d \mid CosSim(p, d) > \theta \wedge d \in \mathbf{P}\}$$

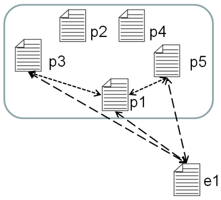

Figure 3. The approach based on similar predictions. The link strength between the target prediction p1 and event e1 is calculated using their cosine similarity and cosine similarity between e1 and other predictions p3 and p5.

$Similar_p$ is a set of similar predictions to the target prediction $p$. $\theta$ is threshold used to detect similar predictions and $\mathbf{P}$ denotes prediction collection. If similar predictions to the target prediction have the same top-scored results as the one of the target prediction, then the score of the target prediction-event report link is strengthened. Figure 3 portrays the concept behind this approach.

### B. Approach Based on Event Report Publication Time

In the previous method, we did not consider time except for the fact that we required event reports to be published after their predictions. However, intuitively, time is an important factor for analyzing predictions. The second method thus uses the publication time of event reports. Note that the reference time is not proper to be used here as few sentences contain the reference time.

To begin with, we need to set up two assumptions. The first one, which is reasonable for news, is that news about an event are reported mostly around the same time when the event occurred. The second assumption states that if sentences which are similar to the target prediction are published around the same time, then they likely describe the prediction event.

First, we calculate the similarity of every pair of the target prediction and candidate event reports. We set some threshold (0.1 in the experiments) and remove the pairs with similarities lower than the threshold. The purpose is to decrease the overall calculation cost and to avoid topic drift. For the remaining pairs, we calculate the scores of their prediction-event report links.

The equation that considers the publication time density is as follows:

$$\begin{aligned} PRLinkScore(p, e) &= CosSim(p, e) \\ &\quad + \sum_{e_i \in E} CosSim(e_i) * TimeSim(e, e_i) \end{aligned}$$

$\mathbf{E}$ denotes here the event report collection. If similar sentences to the target prediction are published around same time, the link strength values between these sentences and the prediction are increased. On the other hand, in the case when many sentences, which are different from the target prediction, are published at same time, the link values

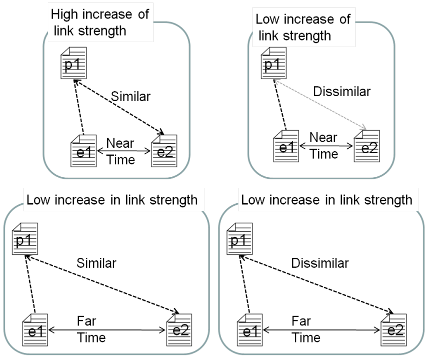
High increase of link strength
p1
Similar
e1
Near Time
e2
Low increase of link strength
p1
Dissimilar
e1
Near Time
e2
Low increase in link strength
p1
Similar
e1
Far Time
e2
Low increase in link strength
p1
Dissimilar
e1
Far Time
e2

Figure 4. The approach based on the publication time of event reports. The link strength between the target prediction p1 and event e1 is calculated using their cosine similarity and similarity between p1 and other event report e2.

between these sentences and the target prediction remain unchanged or change very little. The same applies for the case of sentences similar to the prediction which are published at different times. Figure 4 shows different cases depending on the similarities between the prediction and event reports and the time distance between the event reports.

### C. Experiments

In this section we show experiments of the prediction-event link scoring to determine whether the predictions are obsolete or valid. We used 12 queries (e.g. “international space station”, “bank of japan”, “japan whale”, “softbank”, “nasa discovery”, “hubble space telescope”, “international space station” etc.). We collected 10,000 snippets from the New York Times Article Search API and Google News Archive for each query. Predictions were manually selected from New York Times articles, while the event report candidates were selected from both the New York Times articles and Google News Archives articles. Duplicated sentences were removed. On average we selected 200 predictions for each query.

For each prediction, we applied our proposed methods and then manually checked the returned 10 top-ranked sentences. We only considered obsolete predictions in this experiment. That is, we used only predictions which have actual matching events in the datasets and which can be thus evaluated.

We evaluated the precision of the estimation of prediction validity. We empirically set a threshold of the link value (0.3 for cosine similarity method, 0.2 for both the similar prediction method and the time density method). If a target prediction had at least one pair whose link strength value was higher than the threshold, the prediction was regarded as having been already confirmed by a real time event, hence being an obsolete one. In case when there were many resulting sentences that had higher link scores than the threshold we only considered the top 10 ones. We show the results in Table III. We used cosine similarity expressed as $PRLinkScore(p, e)$ as a baseline. The similar predictions approach has higher F value than the cosine similarity approach. Time density approach is however worse than the other two approaches. On average we could confirm that extending the cosine similarity method offers good results. The similar predictions approach improves the $F$ score of the baseline by 18% and the time density approach improves it by 29%.

Table III
PRECISIONS OF THE REPORTED PREDICTIONS DETECTION.

| Method | Precision | Recall | F value |
|---|---|---|---|
| Similar Predictions | 88.9% | 88.9% | 0.44 |
| Time Density | 85.3% | 60.1% | 0.40 |
| Cosine Similarity | 69.4% | 68.7% | 0.34 |

## V. CONCLUSION

In this paper, we report on the studies towards improving retrieval and analysis of future-related information. In particular, we approach the problem of detecting time-unreferenced, future-related information and we introduce several approaches for its effective retrieval. As the second contribution we propose a novel research problem of validating future-related information and demonstrate two methods towards this objective. We test all our proposals on real datasets and show satisfactory performance.

## VI. ACKNOWLEDGMENTS

This work has been supported by Microsoft IJARC CORE6 project, “Mining and Searching Web for Future-related Information” and by MEXT Grant-in-Aid for Young Scientists B (Grant no. 22700096).

## REFERENCES

[1] R. Baeza-Yates. Searching the Future, Proceedings of ACM SIGIR Workshop MF/IR 2005, 2005.

[2] A. Jatowt, K. Kanazawa, S. Oyama, K. Tanaka. Supporting Analysis of Future-related Information in News Archives and the Web. Proceedings of JCDL 2009, pp. 115-124, 2009

[3] A. Jatowt, H. Kawai, K. Kanazawa, K. Tanaka, K. Kunieda and K. Yamada, Analyzing collective view of future, time-referenced events on the web, Proceedings of WWW 2010, 1123-1124, 2010

[4] H. Kawai, A. Jatowt, K. Tanaka, K. Kunieda and K. Yamada. “ChronoSeeker: Search Engine for Future and Past Events”, Proceedings of ICUIMC 2010, pp. 166-175, 2010

[5] M. Matthews, P. Tolchinsky, R. Blanco, J. Atserias, P. Mika and H. Zaragoza. “Searching through time in the New York Times”, Proceedings of HCIR 2010, 2010