Searching the Future*

Ricardo Baeza-Yates

Center for Web Research
Dept. of Computer Science, University of Chile
Blanco Encalada 2120, Santiago, Chile
E-mail: rbaeza@dcc.uchile.cl

ABSTRACT

In this paper we define a new retrieval problem: *future retrieval*. The idea is to use news information to obtain future possible events and then search events related to our current (or future) information needs. In another words, we include time as a formal attribute for a document. We present a simple ranking model based on time segments, a prototype for it and some specific examples. This work also poses new challenges for natural language processing, information extraction, and answer evaluation.

Keywords: information extraction, formal models.

1. INTRODUCTION

Humans have always wanted to know their future, resorting from religious texts and astrology, to fortune tellers. Although we cannot know the future, a lot can be predicted about it because many things are planned several years in advance. The main sources for knowledge about future events are news. In fact, just looking at future years in Google News,[1] it is possible to find more than 50 thousand articles, with more than 10% having the year in the headline, and with references at least up to year 2100.

In this paper we define a new problem that we call *future retrieval* (FR) that consists in extracting temporal information from news and combine it with standard full-text retrieval to answer queries that mix text and time. In this paper we prove the feasibility of a FR system by using news data already processed for temporal entities, a simple probability model for future events, a model based on a set of time segments and a generic simple ranking extension to any IR model. We also include some results from a prototype of a FR system, which serves as a proof-of-concept of the feasibility and usefulness of future retrieval. Hence, our main contribution is this new problem and the challenges that poses.

The expression "searching the future" has also been used recently for what is called prospective search. For example, in a publisher/subscribe model, people can subscribe to topics of their interest (sometimes expressed as queries), which are matched to any new published information[2]. Our goal is different and in more ambitious.

The outline of the paper is as follows. In section 2 we define the problem. Section 3 presents related work. Section 4 presents our retrieval model. Section 5 presents our prototype, to end with some concluding remarks.

2. PROBLEM STATEMENT

Based in the Google News service, in December 1st, 2003, we found more than 100 thousand references to years 2004 and on. Considering that, according to our sample data shown later, about 80% of the news about the future refer to the immediate future (days, weeks, a few months), and on average more than once per article, we estimate that at least there are half a million references to future events in Google News. Assuming that there is a ten-fold repetition redundancy (similar articles in different newspapers), a reasonable estimation is about 50 thousand unique articles about thr future. A similar analysis only on headlines gives around 10% of that number. In July 15th, 2005, the references to years 2006 or more were over 250 thousand.

For example, in 2034, we found the following among almost 100 news:

1. The license of nuclear electric plants in Arkansas and Michigan will end.
2. The ownership of Dolphin Square in London must revert to an insurance company.
3. Voyager 2 should run out of fuel.
4. Long-term care facilities may have to house 2.1 million people in the USA.
5. A human base in the moon would be in operation.

*ACM SIGIR Workshop MF/IR 2005
Partially funded by Millennium Nucleus Center for Web Research, Mideplan, Chile.

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

[2] For example, pubsub.com offers this service in over 13 million sources.

So, when searching for *energy or health* in the future, we would like to retrieve 1 and 4, classified by year. If Searching for *2034 and space*, we would want to obtain 3 and 5. As mentioned before, we call this new retrieval problem, future retrieval, extending in some sense the concept of temporal databases to the field of IR (although temporal databases worry about the past, not the future).

A FR system should have the following components [3]:

- An information extraction (IE) module that recognizes temporal expressions as times, dates, and durations (a particular kind of named entity recognition) and quantifies the likelihood of the future event (for example, news 1 and 2 will most probably happen, while 3 to 5 might have lower probability of happening).
- An IR system that indexes articles together with time segments and allows text queries, and optionally a time segment, such as the second query posed before. Any IR ranking can be then extended in the time dimension, projected to a time segment, and sorted according to ranking or time.
- A text mining system that given a time query (a time segment for example), finds the most important topics associated with that segment. For example, *space travel* or *NASA* for 2034.

Here we face several challenges, some related directly to IR, and some related to natural language processing (NLP). The IE module has the following challenges: (1) temporal expression recognition, that can be done fairly well in at least 90% of the cases (dates are more important than times or durations in our case); and (2) finding the tense and mood of the main verb, which is a harder problem, to determine a confidence level for the occurrence of each event. A simple first approach is just to distinguish *will* and *must* from *should, could* and *would*.

From the IR side, the challenges are: (3) to add time segments to documents based in (1), and (4) define how to combine any IR ranking scheme with time segments. The text mining part is the most complex, as relies in news topic extraction, which is left aside for future work.

In the sequel we use input data already processed for temporal entities to solve (1), a simple probability model for future events to solve (2), a model based on a set of time segments to solve (3), and a generic simple ranking extension to any IR model to solve (4) based in (3).

3. RELATED WORK

Our work is related to several areas: temporal databases, temporal entity recognition, text mining (in particular information extraction from news), and stock prediction. None of those areas has work on searching the future based in news text, but we borrow concepts from all of them. We present those areas in the order used before.

Temporal databases is a well-established field in the database community that deals with the problem of storing data that has time information [18, 5]. This includes manipulating and reasoning about time based data. We use a simple set of time segments model based, based on several of almost a dozen temporally enhanced entity-relationship models available [6].

Entity recognition is a well defined task in NLP [4]. Temporal entities are one of the most important [1], and they are difficult to extract completely. Using state-of-the-art techniques, more than 90% of them can be extracted (see for example [16]). This problem is also important for tagging news itself [17]. To the best of our knowledge we have not found any work that tags only future temporal entities and extracts the correct time segment. A related problem is information triage, which refers to the task of monitoring a variety of information sources (e.g. news stories, email, stock-prices), and providing users with well-filtered, prioritized and ranked set of information items [11]. However, information triage focus on relating different sources and rank news based on its present importance, while learning from the data. On the other hand, it mixes numerical and textual information, as we do.

Text mining and information extraction have developed in the recent years [13, 14], including discovering trends [10, 12] as well as event extraction [2]. There is specific work on language models for news recommendation [8], as well as mining at the same time text and time-series [9].

There is some work on predicting the future, in particular related to the stock market. In this case the input is numerical part data, and the prediction should help to decide what actions to take (sell, buy, etc.). Recent work has included the use of text as well as time-series for exchange rate forecasting from news headlines [15, 7] or predicting daily stock indices using Web data [19].

4. RETRIEVAL MODEL

This section deals with the inclusion of temporal events in documents and how to mix traditional ranking with temporal events.

Any temporal entity e can be traduced to a time segment $S = [t_1, t_2]$. The time segment could be open to the future (that is, $t_2 = \infty$), but t_1 is at least the present, defined as $t_1 = today$. We define a temporal event E_e as a tuple (S_e, C) where S_e is the time segment associated to the temporal entity e, and C is a confidence probability (or level) that the event will actually happen ($0 \leq C \leq 1$).

A document has several temporal entities. Hence, we associate to a document all temporal events associated to temporal entities in the document. That is, for a document d, we have

$$E_d = \cup_{e \in d} E_e \ .$$

Let define $M(V)$ as the maximum confidence level of a set of events V.

Notice that we can plot time versus confidence of the set E_d by taking the maximum confidence at any time (which is equivalent to using the most probable event if more than one will happen). Other ways to combine overlapped events are possible, but we use this one because it has a reasonable meaning. We will call the resulting curve in the graph, the *trace* of E_d. In figure 1 we show an example.

We now extend any ranking model to include a set of temporal events, in a very simple way, by considering only word queries. Let $r = r(d, q_W)$ be the normalized ranking (weight) of a given full-text retrieval system for document d when the query is q_W (that is $0 \leq r \leq 1$). Then, the new time-based ranking is computed using

$$tr(d, q_W) = r(d, q_W) \times M(E_d)$$

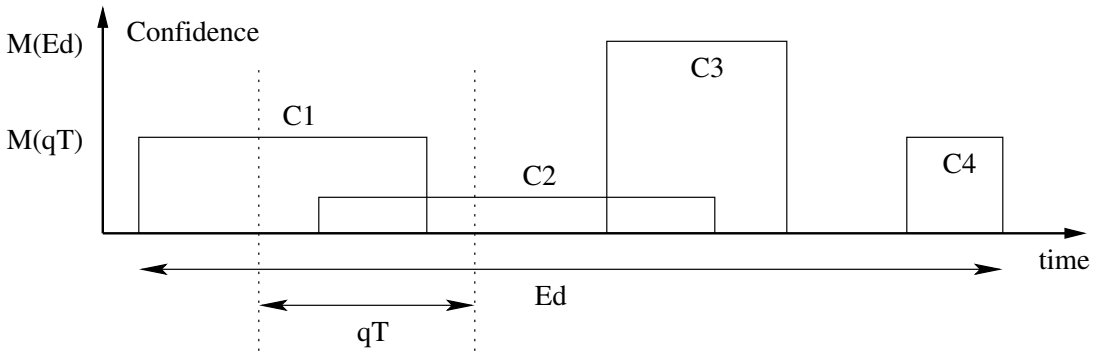

Figure 1: Example of a document temporal trace.

In simpler terms, the best document will depend on the content as well as the most probable event in it, in an independent fashion.

We define a *time query* as a time segment[3]. In this case the ranking is simply the maximal confidence restricted to the time segment of the query, that is $M(E_d \cap q_T)$ (see example in figure 1 which uses $M(q_T)$ to indicate this case). Finally, a complete query combining words and time will be ranked by using $tr(d,q) = r(d,q_W) \times M(E_d \cap q_T)$, as before.

4.1 Variants

If q_T is a finite time segment, an alternative ranking can be defined as follows. Let $A = A(E_d \cap q_T)$ be the area under the trace of E_d restricted to the time segment defined by q_T. Then, we can use $tr(d,q) = r(d,q_W) \times A(E_d \cap q_T)$.

Another alternative would be to divide the document in text passages, each one of them having just one temporal event. However, this poses new problems, such as delimiting the text of a temporal event, how to handle text overlaps among events, how to handle time overlaps among events and their confidence levels, etc. Several other variations are possible, but we chose to start with a simple model, to later make it more complex.

5. EXPERIMENTAL RESULTS

5.1 Model Choices

Without loss of generality, we used years as time unit for our time segments. Although a finer granularity is possible, most future references were for years, with the exception of references to following days.

As underlying word ranking model we use the Boolean model, as this allows to study just the effect of the event confidence model. Our confidence model is also simple. We have two levels, one for events that almost surely will happen, and another for possible events. We decided arbitrarily to use 100% and 50% confidence for those two levels, although other numbers are equally valid.

5.2 News Data

As news input we use the NIST 1999 information extraction entity recognition evaluation data sample[4], that contains 94 Associated Press and New York Times articles from 1998, with a total size of 540Kb. Each article contains tags for a date, headline, and text context. In addition, this data is already tagged for two temporal entities: dates and durations. We manually filtered this data to extract only the articles that referenced the future, removing other tagged entities. Table 1 shows the result of the filtering, that reduced the database to 45 articles and 187Kb. To each article we added the time segments and confidence levels.

	Articles	Dates	Durations
Total	94	538	246
Future	45	88	21
Percent	48%	16%	9%

Table 1: News filtered for future temporal entities.

About 87% of the news having references to the future were in the same year (and many to the following days). However, 44% of them had some reference to the next two years and 31% to three or more years up to 2020. The estimation for Google News gives around 62% and 36% for the last two percentages (the rest was after 2020).

Regarding the confidence levels, 66% were scheduled events which almost surely would happen (all but one of them were for the same year), while the possible events referenced in half of the cases to two or more years ahead. So, as expected, there is a high correlation between confidence level and event proximity.

5.3 Prototype

For the prototype we used the Amberfish XML search toolkit[5], which allows Boolean full-text search in XML data. This software is freely available and runs in a PC under Linux, which was our development environment. Each news article was an XML segment having four tagged fields: title, body, E_d (only years), and $M(E_d)$. Time segments were treated expanding them to all years in the data as well as in the queries. A query was evaluated by intersecting a full-text search in the title and body tags content with a full-text search for any of the years in the E_d tag content. Finally, the ranking was obtained by sorting the answers using the $M(E_d)$ tag con-

[3]It is possible to define it as a set of segments, but this can be done by simply extending the normal Boolean query operations to operate also in time segments.

[4]Available at http://www.itl.nist.gov/iad/894.01/tests/ie-er/er_99/er_99.htm

[5]Distributed by Etymon Systems, Inc.

Query:

Year: 2002

Ranking	Headline	Confidence
1.	MUSICIANS ON BROADWAY TO VOTE ON A CONTRACT	100%
1.	NEW YORK PLEASED BY HOUSE BILL ON TRANSPORT	100%
3.	ANALYSIS: TAXING INTERNET SALES - GOVERNORS VS. TAX FREEDOM ACT	50%
3.	BUOYANT CLINTON TAKES ON GOP SENATORS, BIG TOBACCO	50%
3.	SMITHSONIAN FIRES ARCHITECT OF NATIONAL INDIAN MUSEUM	50%

Query: billion

Year: 1999-2000

Ranking	Headline	Confidence
1.	NEW YORK PLEASED BY HOUSE BILL ON TRANSPORT	100%
2.	ANALYSIS: TAXING INTERNET SALES _ GOVERNORS VS. TAX FREEDOM ACT	50%
2.	BUOYANT CLINTON TAKES ON GOP SENATORS, BIG TOBACCO	50%
2.	CREDIT WARNING BY MOODY'S ON JAPANESE BONDS	50%

Query: Kenya OR Malaysia

Year: 1998

Ranking	Headline	Confidence
1.	MALAYSIA	100%
1.	MALAYSIA	100%
1.	SECURITY COUNCIL MEMBERS EXPRESS NEED FOR POLITICAL WILL TO HELP	100%
4.	KENYANS PROTEST TAX HIKES	50%

Figure 2: Examples from our prototype.

tent and the title tag content as secondary key (that is, confidence ties were given in alphabetical order).

A simple form allows to input the word query (optional) and the year segment. The output was a simple HTML page showing the numerical ranking, headline, and confidence level. As the data available is small, did not make sense to do a formal evaluation of the confidence level, as it is not clear how to choose meaningful queries of a few days in 1998 and then check if the events happened or not. Nevertheless, to have a feeling on the type of answers, figure 2 shows three query examples. In the last example the first two headlines are the same although the news are different: one about Malaysian economy, and the other about the Commonwealth Games.

Notice that the problem of assessing the quality of the confidence level of the answers, after they have happened, it is an interesting research problem in its own right.

6. CONCLUDING REMARKS

We have shown that at least the first two modules of a FR system are feasible and that the results could be used in commercial or political decision systems. We are currently evaluating devising better ranking techniques for our system by using larger news collections, where we can use a first part as news base, and a second part to evaluate whether things happened or not (another NLP problem). The text mining module needs future collaborative research from the NLP side.

Our system can also be used for advanced forensic search, extending the search to only past temporal entities. However, in forensic retrieval (that is, searching the past), we can argue that a probabilistic model does not make that much sense, as events in the past did (almost always) happened.

We believe that our main contribution is the problem itself, along with a proof-of-concept of its feasibility. Further work includes:

- Estimation of confidence levels from time segments and their associated text.
- Evaluation of the answer in a FR system. For example, detecting which events did or not happen in a given news collection and comparing that with the confidence level estimated *a priori*.
- Efficient search. We used a very simple scheme by using only years, but if complete time segments are used (e.g. day granularity), a different type of index is needed.

We believe that future research on this problem will spark collaboration with other areas such as NLP and text mining.

Acknowledgements

We are grateful to the excellent comments and helpful pointers provided by Steven Bird.

7. REFERENCES

[1] K. Ahmad, P. de Oliveira, P. Manomaisupat, M. Casey, T. Taskaya. Description of Events: An Analysis of Keywords and Indexical Names. In *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002: Workshop on Event Modelling for Multilingual Document Linking*, pp. 29-35, 2002.

[2] J. Allan, editor. Topic Detection and Tracking: Event-based Information Organization, Kluwer, 2002.

[3] R. Baeza-Yates. Searching the Future, Technical Report, CS Dept., University of Chile, January 2004.

[4] R. Dale, H. Moisl, and H. Somers, editors. Handbook of Natural Language Processing, Marcel Dekker, NY, 2000.

[5] C.J. Date, H. Darwen, and N. Lorentzos. Temporal Data & the Relational Model, Morgan Kaufmann, 2002.

[6] H. Gregersen, C.S. Jensen. Temporal Entity-Relationship Models-A Survey. *IEEE Transactions on Knowledge and Data Engineering* pp. 464-497, 1999.

[7] P. Kroha, R. Baeza-Yates. A Case Study: News Classification Based on Term Frequency. *Sixth International Workshop on Theory and Applications of Knowledge MAnagement* (TAKMA 2005), Copenhagen, Denmark, August 2005.

[8] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan. Language models for financial news recommendation. In *Proceedings of the ninth international conference on Information and knowledge management*, VA, pp. 389-396, 2000.

[9] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan. Mining of Concurrent Text and Time-Series, In *KDD-2000 Workshop on Text Mining*, Boston, MA, 2000.

[10] B. Lent, R. Agrawal, R. Srikant. Discovering Trends in Text Databases, In *Proc. 3rd Int. Conf. Knowledge Discovery and Data Mining*, D. Heckerman, H. Mannila, and D. Pregibon, and R. Uthurusamy, editors. AAAI Press, pp. 227–230, 1997.

[11] S. Macskassy, and F. Provost. Intelligent information triage. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, New Orleans, LA, pp. 318–326, 2001.

[12] U.Y. Nahm, and R.J. Mooney. Text Mining with Information Extraction. In *Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases*, pp. 60-67, Stanford, CA, March 2002.

[13] M.T. Pazienza, J.G. Carbonell, and J. Siekmann, editors. Information Extraction: Towards Scalable, Adaptable Systems, LNAI, Springer, 1999.

[14] M.T. Pazienza, editor. Information Extraction in the Web Era: Natural Language Communication for Knowledge Acquisition and Intelligent Information Agents, LNAI 1714, Springer, 2003.

[15] D. Peramunetilleke, and R.K. Wong. Currency exchange rate forecasting from news headlines. In *Proceedings of the thirteenth Australasian conference on Database technologies*, Vol. 5, pp. 131-139, 2002.

[16] F. Schilder, and Chr. Habel. Temporal information extraction for temporal question answering. In *Proceedings of the 2003 AAAI Spring Symposium in New Directions in Question Answering*, Stanford University, CA, 2003.

[17] F. Schilder, and Chr. Habel. From Temporal Expressions to Temporal Information: Semantic Tagging of News Messages. *In Proceedings of ACL'01 workshop on temporal and spatial information processing*, pp. 65 -72, Toulouse, France, 2001.

[18] A. Tansel, J. Clifford, S. Gadia, S. Jajodia, A. Segev, and R. Snodgrass, editors. *Temporal databases: theory, design, and implementation*, Benjamin-Cummings, CA, USA, 1993.

[19] B. Wuthrich, D. Permunetilleke, S. Leung, W. Lam, V. Cho, and J. Zhang. Daily Prediction of Major Stock Indices from textual WWW Data, *HKIE Transactions*, Vol. 5, No. 3, pp. 151-156, 1998.