Analyzing Collective View of Future, Time-referenced Events on the Web

Adam Jatowt[1,3], Hideki Kawai[1,2], Kensuke Kanazawa[1], Katsumi Tanaka[1], Kazuo Kunieda[2], Keiji Yamada[2]

[1]Kyoto University
Yoshida-Honmachi, Sakyo-ku
606-8501 Kyoto, Japan
{adam, kanazawa, tanaka}@dl.kuis.kyoto-u.ac.jp

[2]NEC C&C Innovation Research Laboratories
Takayama-cho, Ikoma
8916-47 Nara, Japan
{h-kawai@ab, k-kunieda@ak, kg-yamada@cp}.jp.nec.com

[3]MSR IJARC Fellow

ABSTRACT

Humans have always desired to guess the future in order to adapt their behavior and maximize chances of success. In this paper, we conduct exploratory analysis of future-related information on the web. We focus on the future-related information which is grounded in time, that is, the information on forthcoming events whose expected occurrence dates are already known. We collect data by crawling search engine index and analyze collective view of future time-referenced events discussed on the web.

Categories and Subject Descriptors

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

General Terms

Algorithms, Measurement, Experimentation

Keywords

future-related information, collective predictions, opinion analysis

1. INTRODUCTION

Our life success depends to large extent on the correct prediction of future. The usual way to forecast the future is to analyze the current situation and the latest trends such as technological, economic or societal ones. On the other hand, an indirect way for forecasting the future is to refer to the opinions of others. Since the web is commonly seen as the reflection of real world and life, it is appealing to harness online future-related content such as plans, schedules, predictions, speculations or expectations for portraying the collective views on the future.

In this paper we investigate the distribution of future-related information on the web and analyze its major topics. In order to create datasets for the analysis, we have searched for a content that is associated with future dates and, at the same time, is biased in the least possible way. We have collected the data by crawling search engines with queries composed of absolute temporal expressions. Then, for each future date we have estimated the amount of information on events expected to occur at that time. This sort of collective future view applies to the events that have assigned expected dates of their occurrence (i.e. time-referenced events). Hence, it does not cover events for which no explicit dates are commonly known.

Baeza-Yates [1] described the concept of "future retrieval" and discussed the desired mechanics of a future search engine – a service for returning relevant documents related to given future time periods. In recent work [3] we have proposed generating visual summaries of probable future events concerned with given objects by clustering web search results. In another work [4] we analyzed effective ways to automatically determine future-related information in documents using SVM.

2. DATA COLLECTION AND ANALYSIS

For data collection we have crawled web search engine index with phrase queries in English. The queries contained temporal expressions defined as "*temp_modifier*+(the)year(s)+*yyyy*" such as "in year *yyyy*", "in the year *yyyy*", "by the years *yyyy*". *temp_modifier* is a preposition often used together with year dates and *yyyy* is a 4 digit number ranging from 2010 to 2050. We have prepared 39 different patterns of "*temp_modifier*+(the)year(s)" to be used for every future year. To ensure their correctness we manually inspected the top search results returned for each pattern for some selected future dates. In order to increase the amount of content returned by a search engine API above the fixed limit we also added single stop words to the temporal expressions (e.g., however "to the year 2032", although "by the year 2013", etc.).

In this way, in total, we issued 873,054 different queries containing 546 different stop words using Microsoft Bing search engine API[1] in October 2009. Besides the *yearly dataset*, we have also created *monthly dataset* in a similar way spanning the time period from January 2010 to December 2012 with monthly granularity. In this case the temporal expressions were created according to the pattern "*month_reference*+*yyyy*" in which *month_reference* is a full or shortened name of a month (e.g., January, Jan) and *yyyy* is 2010, 2011 or 2012. The monthly dataset was created by issuing 39,312 queries in December 2009. Note that the sets of month references in the monthly dataset and prepositions in the yearly dataset are not complete.

For each query we have captured the hitcount value reported by the search engine, the returned snippets and titles of up to the top 1000 search results. Next, we filtered duplicate pages so that each

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

[1] http://msdn.microsoft.com/en-us/library/dd900818.aspx

URL was unique within the results for the same year. In total we obtained 1,044,224 and 770,715 unique search results for the yearly and monthly datasets, respectively.

First, we analyzed how much information is on the web in relation to a given future year. We have calculated the average hitcount values obtained for all the queries and normalized them by dividing by the maximum value to fit into 0 – 1 range. We could then plot the amount of information related to given future year as a *forecasting curve* (Figure 1). The amount of time-referenced future-related information decreases sharply along with the time for the first 4 years. The curve appears to stabilize around the year 2015. An interesting characteristic are local peaks at round dates such as 2015, 2020, 2025, 2030 or 2050 which probably serve as a kind of convenient temporal landmarks for referring to the future.

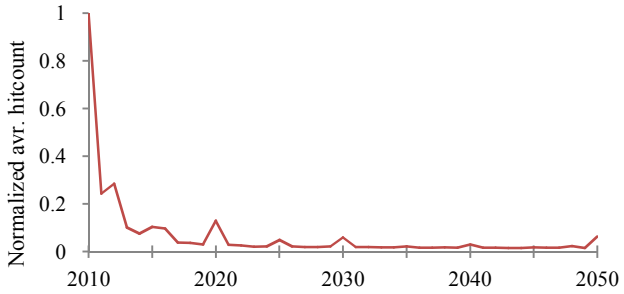

Figure 1 Forecasting curve for the yearly dataset.

In Figure 2 we show the forecasting curve for the monthly dataset. The largest drop in the hitcount values occurs from November 2010 to January 2011. This suggests that the biggest chunk of the future-related information pertains to the nearest one year period as the monthly dataset was created in December 2009. However, it could also be influenced by the change in the calendar year.

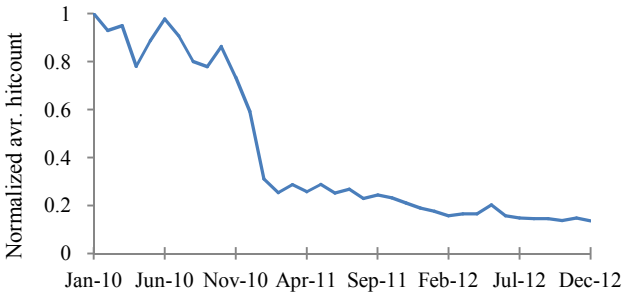

Figure 2 Forecasting curve for the monthly dataset.

We next investigated the common topics related to future years in the yearly dataset. The importance of features in each year was determined using log-likelihood ratio test [2] by comparing their occurrence frequencies in the year with the ones in other years. Table 1 shows some significant terms for the yearly dataset selected from within the 50 top-scored terms for each different year.

Table 1 Main terms for selected years in the yearly dataset.

2010	vancouver, budget, honda, ford, toyota, car, release, japan, population, winter
2011	chelsea, rugby, cup budget, ford, troop, cricket, contract, market, wii, myspace, fifa
2012	london, mayan, olympic, uefa, euro, doomsday, apocalyptic, sport, kyoto, paraolympic, obama, nostradamus, galactic, earth, nibiru
2013	eu, mobil, european, maccain, fiscal, myspace, population, iraq, troops
2014	sochi, winter, xp, olympic, glasgow, ie6, microsoft, russia, brasil
2015	mdg, goal, develop, hunger, global, mcfly, africa, unesco, billion, millenium
2016	olympic, bid, chicago, rio, tokyo, host, ioc, game, obama, janeiro, madrid, copenhagen
2020	vision, energy, develop, climate, strategy, carbon, china, greenhouse, global, economy, industry, summit
2030	energy, oil, demand, climate, population, carbon, electric, nuclear, barrel
2040	population, tokyo, citizen, climate, sea, arctic, crisis, region, scientist, demography, people, disappear, supply
2050	population, climate, carbon, greenhouse, energy, co2, warm, temperature, gdp, g8, change, sustain, country

Several popular future events can be found in Table 1. For example, the Summer and Winter Olympics games are commonly expected international events (2012 London, 2014 Sochi, 2016 Rio de Janeiro). For the year 2012, the frequently discussed issues center also around UEFA 2012 Euro Cup, Kyoto Protocol commitments and quite surprisingly on the Mayan calendar as well as the supposed prediction of "the end of the world". From the table we can also learn that the customer support for the Windows XP operating system is going to be terminated in 2014. From 2020, the main topics of future expectations and predictions in the round years seem to be related to the issues of environment, climate change, energy, population and so on.

In our future work we plan to extend the analysis to different languages and document types as well as to periodically collect future-related information on the web in order to investigate changes in its character and topics over different time spans.

3. ACKNOWLEDGEMENTS

This work has been supported by Microsoft IJARC CORE6 project, "Mining and Searching Web for Future-related Information" and by the National Institute of Information and Communications Technology, Japan.

4. REFERENCES

[1] R. Baeza-Yates. Searching the Future. *Proceedings of ACM SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval (MF/IR 2005)*, Salvador, Brasil, 2005.

[2] T. Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1), MIT Press, pp. 61-74, 1993.

[3] A. Jatowt, K. Kanazawa, S. Oyama, K. Tanaka. Supporting Analysis of Future-related Information in News Archives and the Web. *Proceedings of JCDL 2009,* ACM Press, Austin, TX, USA, pp. 115-124, 2009.

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