Multi-lingual, Longitudinal Analysis of Future-related Information on the Web

Adam Jatowt[1], Hideki Kawai[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

***Abstract*— Future prediction is one of the crucial activities of humans and is a constantly ongoing process. In this paper, we report the results of exploratory analysis of future-related information on the Web in different languages: English, Japanese and Polish. 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. Datasets are constructed by crawling search engine indices and analyze collective views of future events discussed on the Web. We investigate multiple aspects of future-related information across different languages such as its amount, time span, typical topics, associated sentiment levels as well as the relation to the future-related content in news articles.**

Keywords— future-related information, collective predictions, opinion analysis

I. INTRODUCTION

Our life success depends to large extent on the correct prediction of the future so that we can plan our actions in the best way. Naturally, the future, in contrast to the past, is inherently uncertain and is difficult, if not impossible, to be correctly predicted. Nevertheless, forecasting future has been always a crucial and permanent activity common to many societies and cultures. One way to forecast the future is to analyze the current situation and the latest trends such as technological, economic or societal ones. Another common approach is to examine the history and to extrapolate the past, usually by finding previous situations similar to the present one [7]. However, relying solely on the history usually does not suffice as the saying goes: "History doesn't repeat itself, but it does rhyme" (Mark Twain).

Since the Web has taken a prime role in our lives and is commonly accepted as the reflection of the real world, it should become then possible to harness online content for portraying the collective views on the future. This kind of approach appears feasible as lots of future-related information is available in web pages. Such information consists of future plans, schedules, predictions, speculations, expectations and so on. As people want to organize and arrange their work and personal lives they publish information on the future course of actions. Due to its uncertain nature, the future-related information can vary among different sources and can also fluctuate over time. For example, oil peak can be predicted to occur in different time frames according to different experts, or the expected release date of a new operating system can change as the time goes by. Therefore, the analysis of future expectations of web users should focus on estimating majority opinion according to the "wisdom of crowds" concept [19] and should be possibly carried repeatedly over time. The results of this kind of analysis could be useful for individual users in decision making as well as in many domains like the futurology, sociological studies, marketing or business intelligence. For example, companies may want to know what kind of future events people are talking about or scientists could study variations in future expectations between different communities or nations.

In this paper we investigate the distribution and character of future-related information on the Web in order to analyze collective future views maintained by web users. The data collection was done at the end of 2009, thus, our datasets represent the snapshot of the future views on the Web recorded as of the end of 2009. We have collected the data by crawling a web search engine with queries containing absolute temporal expressions such as future years. Then, for each future year we have estimated the amount of information on events expected or predicted to occur in this year and created its summary. We note that this sort of future view naturally applies to the events already having associated expected dates of their occurrence. Hence, it does not cover events for which there are no concrete dates known yet. Since we analyze time-referred future events, thus, it becomes possible to portray the distribution of web authors' attention concerned with particular future years as well as to investigate the nature of collective predictions according to the temporal dimension.

Our analysis applies not only to the Web of English but is also concerned with the predictions made in Japanese and Polish languages. This allows for comparing characteristics of collective future expectations across different languages.

The remainder of this paper is structured as follows. In the next section we describe the related research, while in Section 3 we discuss the data collection. Section 4 contains the description of the main results of our analysis. In the next section we provide the discussion and conclude the paper.

II. RELATED WORK

Previous attempts of using web information for prediction often focused on the prediction of stock price movement or sales volume estimation [20,2,8]. Wuthrich et al. [20] proposed the prediction method of stock indices using historical news about companies and past information on stock indexes as training data. Their method predicts whether the stock indices in the next day will be up, down or unchanged by extracting salient keywords from past news. Choudhury et al. [2] tried to predict changes in stock prices based on blog communication patterns. By inputting features derived from communication dynamics of the blogosphere into SVM, they managed to determine and visualize the probable movement of stock prices. Gruhl et al. [8] showed that the volume of blog postings can be used to predict spikes in actual consumer purchase decisions on the example of books.

Some prediction research has been using sentiment analysis [15,16,18]. Mishne and Glance [16] and Liu et al. [15] applied sentiment analysis methods to weblog data for estimating movie success. Related to these works is [17], in which authors investigated the public mood concerned with the future on the basis of emails submitted to futureme.org, a web service that allows scheduling emails to be sent at future dates. In this paper we also study sentiment degrees of future-related information and arrange it on timeline.

Jatowt et al. [9,11] proposed generating summaries of probable future outcomes related to user-provided keywords. The proposed method was based on clustering of query-related content that refers to the future with an emphasis on the dates when this content appeared (i.e., news article timestamps). This approach was then extended by using model-based clustering in [11]. Dias et al. [3] made an exploratory study to understand how the temporal features impact upon the classification and clustering of different genres of future-related texts. In contrast to these works, we provide here the collective overview of future-related information on the Web and analyze the characteristics of such information from different viewpoints. Some of the results of this study appeared in [10] such as the forecasting curve for English language content on the Web in yearly and monthly granularities and selected representative terms related to future years in English.

Baeza-Yates [1] was the first to introduce the concept of "future retrieval" and discussed the mechanics of future search engine. Kanhabua et al. [13] proposed ranking model for predictions that takes into consideration their relevance. Future-related information which is associated with concrete future dates is however relatively rare. Kanazawa et al. [12] estimated that about 30% of predictions in news articles contain future dates based on a small scale investigation. The authors then proposed methods for retrieval and validity analysis of future-related information which is not associated with explicit future dates (i.e. time-unspecified future information). In [14] the authors proposed effective ways to automatically determine future-related information in documents using machine learning. The proposed system, called Chronoseeker, was trained with SVM to select features representing typical ways in which people refer to the future in text.

III. DATASET CONSTRUCTION

We tried to collect web content that is concerned with future events and, at the same time, is not subject to any topical bias. To this end, we queried the Bing Web search engine with special phrase queries that force the search engine returning content with future references. The queries contained temporal expressions, which explicitly refer to the time, in the form: "temp_modifier+(the)year(s)+yyyy" such as "in year yyyy", "in the year yyyy", "by the year yyyy". temp_modifier denotes a preposition that is 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 until 2050. To ensure their correctness we manually inspected the top search results returned for each pattern. We then rejected the patterns which return content unrelated to the future.

Note that we did not use relative temporal expressions such as "two years later" or "the next year". They are difficult to be mapped on the timeline due to unknown or irretrievable reference time. Note also that due to applying different temporal modifiers, the expressions denoting point in time and those referring to the beginning or the end of time period were treated equally.

We also appended stop words to the above temporal expressions in order to increase the amount of content returned by a search engine's API above the allowed limit of 1000 snippets. Each temporal expression was appended with a single stop word forming queries such as: that "till year 2032", then "till the year 2032", there "to years 2032", there "to the years 2032", although "by the year 2013". By appending the stop words we could increase the total number of queries sent to the search engine and thus obtain more documents having future-related content. This is because usually web documents do not contain all possible stop words. Note that due to negligible semantic meaning of stop words, little bias was introduced, even in the case of relatively rare stop words such as "albeit", "accordingly" or "nevertheless". Similar strategy for increasing the size of returned content was applied in Knowitall project [5].

For each issued query we have captured the hitcount value reported by the search engine. We also stored the returned snippets of up to 1000 search results. For brevity, from now on, we will call page snippets as documents.

After having finished the crawling we applied the following filtering conditions:

1) each page URL should be unique within the results collected for the same year

2) temporal expressions containing future dates should appear either in snippets or in the titles of search results

The first condition was applied to ensure duplicate URL removal within the data for the same year. The duplication rate was quite high (between 70% - 90% depending on the dataset) due to the presence of stop words in the queries. Note that if a document contained many different future dates it was

counted once for each different date. The second condition had to be applied as sometimes temporal expressions did not appear in the returned page content or they appeared only inside URL addresses.

Table 1 provides the statistics of all datasets. The datasets for Japanese and Polish languages were created in the same way using Japanese and Polish stop word lists.

As mentioned before, the datasets did not cover future-related information which lacks association with any explicit future dates. We deliberately constrain the analysis to somewhat more precise future events for which concrete occurrence dates are already known and which can be mapped on a timeline. Another reason why we have decided to consider only time-referenced, future-related information is that generating the unanchored future-related datasets from search engine indices is not trivial.

We have also constructed additional dataset by crawling English news articles with yearly granularity using the Bing's news search option. The reason for creating this dataset was to compare it with the web dataset following the assumption that news articles should contain more credible information related to the future, which should also be more grounded in time than the information in web pages. We have collected at most the top allowable 100 snippets using the Microsoft Bing API. We could not use the hitcount values as they were automatically capped to 100 by the search engine.

TABLE 1 Statistics of English Web (EN Web), Japanese Web (JP Web), Polish Web (PL Web) and English News (EN News) datasets.

Dataset	#stop words	period	#queries	#URLs
EN Web	546	2010-2050	873K	1.04mln
JP Web	310	2010-2050	495K	2.99mln
PL Web	206	2010-2050	118K	714K
EN News	546	2010-2050	873K	196K

IV. ANALYSIS

A. Time Horizon of Predictions

First, we analyze how much information there is on the Web in relation to a particular future year. We have calculated the average hitcount values in each dataset so as to plot the amount of information related to given future dates on a timeline. We call the resulting curve a forecasting curve. The points on the forecasting curve are obtained after the aggregation and normalization of hitcount values (denoted by HC) as shown in the Equation 1.

$$NS(date) = \frac{\sum_{te}\sum_{sw} HC(date, tp, sw) \Big/ HC(sw)}{|tp| * |sw|} \qquad (1)$$

tp means a given temporal pattern and sw denotes a stop word used in queries, while |tp| and |sw| are the numbers of temporal patterns and stop words, respectively. We have then normalized the curves by dividing them by the maximum value to fit the values into 0 – 1 range.

Figure 1 shows the normalized forecasting curves obtained for English, Japanese and Polish web datasets. The amount of future-related information decreases sharply over time, especially for the first 5 years. The curves appear to stabilize around 2015. The future horizon, at least considering data gathered from the Web, seems to be thus spanning about 5 years. An interesting characteristic of the curves are local peaks denoting the time points to which larger amounts of future-related content apply as also indicated in [10]. They occur in "round dates" such as 2015, 2020, 2025 or 2030 and seem to serve as sort of convenient landmarks which are easy to be referenced to. For example, many future or present plans are going to end until the round dates.

Basically, the curve shapes are similar for different languages what suggests their universal character. The most noticeable difference is a large spike at 2012 for Polish dataset related to the planned Euro Cup in Poland and Ukraine, the largest international event that Poland will host in its history. There was much discussion about the event in media as of 2009 and before as the event was expected to stimulate the country's economy and development. Another difference is the peak in 2015 year in the Japanese dataset due to the widely discussed population-related prediction in Japan at 2015, according to which, 26% of the Japanese population will be elderly people.

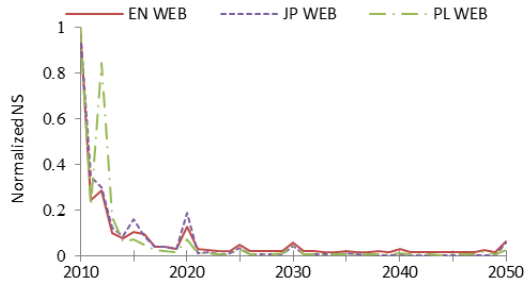

Fig. 1 Forecasting curves for the English, Japanese and Polish datasets.

B. Content of Predictions

We next investigate topics related to expectations. We extracted nouns, verbs, adjectives and adverbs from the datasets using a standard POS tagger. Next, we measured the importance of features over time using a log-likelihood ratio test. We chose this test as it makes few assumptions on underlying data distribution [4]. For each feature we constructed a 2*2 contingency table where f_i is a feature frequency inside data collected for a year i, while f_{T-i} is the frequency of the feature in the data for other years. n_i and n_{T-i} denote the numbers of documents that do not contain the feature in the year i and outside of i, respectively.

The log-likelihood ratio LL is then calculated as follows:

$$LL = 2 \times \left(f_i \times \log\left(\frac{f_i}{E_1}\right)\right) + \left(f_{T-i} \times \log\left(\frac{f_{T-i}}{E_2}\right)\right) \qquad (2)$$

where E_1 and E_2 are expected values estimated as below:

$$E_1 = \frac{(f_i + n_i) \times (f_i + f_{T-i})}{(f_i + f_{T-i} + n_i + n_{T-i})}$$
$$E_2 = \frac{(f_{T-i} + n_{T-i}) \times (f_i + f_{T-i})}{(f_i + f_{T-i} + n_i + n_{T-i})} \quad (3)$$

In Table 3 we show terms selected from within the 50 top-scored terms for some future years. The terms in the Polish and Japanese datasets have been translated into English language.

TABLE 2 Representative terms for selected years in the English Web (EW), English News (EN), Japanese (J) and Polish (P) Web datasets.

2010EW: vancouver, budget, honda, ford, toyota, car, release, japan, population, winter **2010EN**: climate, energy, copenhagen, gas **2010J**: recruit, application, new graduates, examination, test **2010P**: budget, deficit, football, republic of south africa
2011EW: chelsea, rugby, cup budget, ford, troop, cricket, contract, market, wii, myspace, fifa **2011EN**: sudan, pakistan, iraq, darfur, senate **2011J**: broadcast, digital, analogue, tv, ground wave **2011P**: parliament, tusk, euro, national, elections
2012EW: london, mayan, olympic, uefa, euro, doomsday, apocalyptic, sport, kyoto, paraolympic, obama, nostradamus, galactic, earth, nibiru **2012EN**: palin, london, olympic, doomsday, gold, population, android **2012J**: protocol, kyoto, reduction, earth, green house gas, mayan, doomsday, london **2012P**: euro, uefa, ukraine, soccer, stadium, cup, preparations, mayan, platini, prophecy, sport, end, world, highways
2013EW: eu, mobil, european, maccain, fiscal, myspace, population, iraq, troops **2013EN**: insurance, networks, earthquake, cisco, **2013J**: kyoto, protocol, climate, period, investigate, post, market **2013P**: budget, eu, grant, investments, euro, finance
2014EW: sochi, winter, xp, olympic, glasgow, ie6, microsoft, russia, brasil **2014EN**: employment, jobless, orleans **2014J**: winter, sochi, russia, xp, support, vista, home, extension **2014P**: sochi, petru, milano, olympics, winter, train, volleyball, energy
2015EW: mdg, goal, develop, hunger, global, mcfly, africa, unesco, billion, millenium **2015EN**: uranium, poverty, broadband, hungry, millenium, finland **2015J**: aging, care, silver, population, goal, develop, fuel, car **2015P**: development, strategy, mining, rocket, program, investment, eu
2016EW: olympic, bid, chicago, rio, tokyo, host, ioc, game, obama, janeiro, madrid, copenhagen, paralympic **2016EN**: rio, janeiro, olympic, brazil, chicago, golf, ioc **2016J**: olympic, tokyo, host, summer, ioc, game, rio de janeiro, sports, candidate **2016P**: culture, europe, capital, rio, olympics, chicago, gdansk, szczecin, torun, candidate, festival
2017EW: saban, alabama, ukraine, sevastopol, fleet, democracy, iraq **2017EN:** senior, terra, finance, nyse, sioux, subsidiaries, rugby **2017J**: postal services, privatization, insurance fee, schedule **2017P**: express, multibank, michalkiewicz, black sea, fleet, logistics, ireland, treaty, euro, ukraine
2018EW: cup, england, fifa, host, hockenheim, world, football, rapidshare, antivirus, jobless, expiry, moon **2018EN**: cup, fifa, beckham, england, pyeongchang **2018J**: terminator, judgment day, sky net, john connor, machine **2018**P: christian, action, saving, terminator, atomic, england, kghm, movie
2020EW: vision, energy, develop, climate, strategy, carbon, china, greenhouse, global, economy, industry, summit **2020EN**: copenhagen, greenhouse, carbon, warming, gas **2020J**: reduction, green house gas, goal, emission, hatoyama, global warming, energy, electric production, environment, democratic party, prime minister **2020P**: energy, emission, ue, gas, greenhouse, coal, strategy, poland, reduction
2025EW: energy, water, population, scenario, global, trend, vision, oil, amazon, growth, consumption **2025J**: elder, population, medical, society, care, linear, shinkansen, jr tokai **2025P**: puls, biznesu, energy, car, global, politics, euro, china, russia, strategy
2030EW: energy, oil, demand, climate, population, carbon, electric, nuclear, barrel **2030J**: energy, population, electric production, society, solar, fuel, technology, atomic power, oil **2030P**: energy, atomic, poland, trade, report, politics, increase, gay, water, alien, coal, population
2040EW: population, tokyo, citizen, climate, sea, arctic, crisis, region, scientist, demography, people, disappear, supply, polar **2040J**: north pole, melt, population, nation, vanish, future, earth **2040P**: year, euro, senior, board, ship, samsung, planet, health, number, development, earth
2050EW: population, climate, carbon, greenhouse, energy, co2, warm, temperature, gdp, g8, change, sustain, country **2050J**: green house gas, emission, toyako, goal, reduction, population, effect, earth, society, summit **2050P**: billion, population, emission, un, gas, climate, euro, strategy, energy, earth

As there are many predictions shown in Table 2, we discuss only some of them below. From 2020, the main topics of future expectations in the round years were related to the issues of environment, climate warming, energy, population, aging societies and so on. The Winter and Summer Olympics games were commonly expected international events (2012 London, 2014 Sochi, 2016 Rio de Janeiro) in year 2009 when the datasets were created. For the year 2012 internationally discussed issues centered also around Kyoto Protocol as well as surprisingly on Mayan calendar and the related "end of the world". Terms like "windows" and "xp" appearing in English and Japanese datasets are related to the expected termination of customer support for the Windows XP in 2014.

In the Japanese dataset we have noticed several topics specific to Japan such as the ones related to: new recruit period for domestic companies in 2010, the scheduled end of analogue broadcasting in 2011, the planned privatization of postal services in 2017 and, somewhat surprisingly, the "Terminator Salvation" movie whose action takes place in 2018 (this topic appears also in the Polish dataset). The movie was released in 2004.

In the Polish dataset we could observe much discussion on Euro Cup in 2012. Also, in 2016 Poland was supposed to have one of its cities chosen to become a European Culture Capital with Gdansk, Szczecin, Torun being candidate ones.

In general, we noticed that topics related to the climate, astronomy, earth, population and science were frequently discussed future issues. Apart from the actual expected events, we sometimes found words extracted from the titles of science-fiction movies, books or games. Note the difference with the information on the actually planned releases of movies or books in the future. Another interesting case is related to the present promises or schedules of future events. For example, the Japanese Prime Minister Hatoyama decided new greenhouse gases limitations for Japan to be achieved by 2020. Because of this, his name appears as one of top terms for the year 2020 in the Japanese dataset. However, most likely, he is not going to be the actual person involved in the realization of these objectives until the year 2020. In general,

in some particular cases, the information related to the future is blended with the one of the past or present. This kind of issues should be approached in any forthcoming research that attempts at constructing possible future scenarios using future-related information retrieved from text data.

C. Comparison of Predictions in News Articles vs. Web Pages

We next focused on the differences between predictions in news and web pages by synchronizing web and news datasets according to referenced years. We first constructed year vectors with term weights estimated by document frequency and also by log likelihood values in each year in both the English news and web datasets. Figure 4 shows similarities between the future-related content in news articles and web pages for each year. We have used here cosine similarity as a similarity metric.

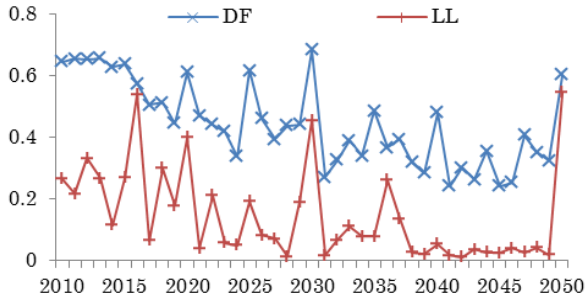

Fig. 3 Cosine similarities between document frequency (DF) and log likelihood (LL) vectors of the web and news English datasets in each year.

We can observe that the similarity decreases the more distant future dates are compared. This indicates that the further ones moves in time, the more discrepancy there is between the web-based and news-based views on future. It could be interpreted that web-based predictions for the far future are supported to lesser degree by the predictions in news articles than the web-based predictions referring to the near future. Assuming that news articles contain more probable and, in general, more trustworthy information than web pages, one cold risk saying that views on distant future expressed in the Web are, on average, less credible than the ones on near future. This seems to be due to the fact that there are fewer predictions announced in news as opposed to relatively high number of diverse expectations published on blogs or other web pages when it comes to forecasting distant future. Another observation is that, in general, there are spikes in the similarity between both datasets at round years: 2020, 2025, 2030, 2035, 2040 and 2050. This is most likely related to many schedules and plans referenced at the round dates that are common both on the Web and in news articles.

D. Sentiment Analysis

Lastly, we investigated the emotional perception of future events. Using sentiment dictionaries we measured the changes in the frequencies of positive and negative terms over time. In other words, we analyzed the levels of collective moods associated with given future time points. We followed here a simple approach used in sentiment analysis of text based on classifying moods into two broad classes: positive and negative on a statistic of the binary occurrence of a feature in a document.

The sentiment vocabularies have been created using synonyms iteratively derived from the Wordnet [6] based on two sets of seed words (positive and negative ones). In total, we have used 1637 and 4529 of positive, as well as 1756 and 6339 of negative terms for the English and Japanese languages, respectively.

Figure 4 shows the polarity changes for the English web and news datasets as well as for the Japanese web dataset with their corresponding trend lines. The sentiment value in each year has been calculated as the average of sentiment terms contained in snippets (positive and negative values in Figure 4 refer to positive and negative expressions, respectively). The trend in polarity for English news is almost flat over time. On the other hand for the English and Japanese web datasets we observe gradual decline in the polarity over time. We can conclude then that the future is, in general, portrayed more and more negatively by web users the farther ones points. For example, for the year 2043 in Japanese dataset we found a prediction made by a Brazilian futurologist about the end of the humanity. However, we could not explain the spike at 2044 in the English news dataset. The amount of data in the news articles dataset is relatively small for very distant years when compared to the web datasets (except the year 2050); therefore, we assumed this peculiar peak a noise. Actually, as we have seen in Figure 1, the amount of available data decreases drastically along the timeline, especially after the year 2020, and, thus, any text analysis should be taken keeping this in mind.

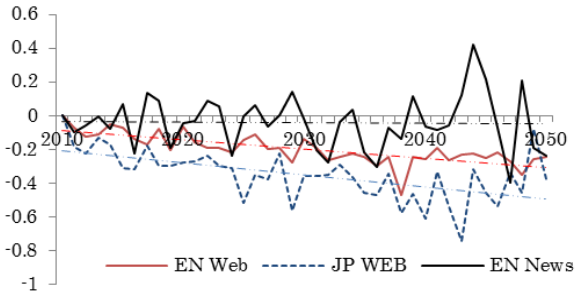

Fig. 4 Average polarity changes in the datasets and their trends.

In Figure 5 we show the distributions of positive and negative terms separately over time (positive in the upper part and negative in the lower part) normalized in relation to their respective values obtained for the year 2010. The corresponding trend lines are also displayed. For the case of English we can observe the increase in the frequency of both negative and positive expressions along with the increasing distance into the future. From this figure, it appears that sentiment volatility increases, the further future dates are considered. This could imply that the future views become more emotional. The increase in positive emotions is smaller than the one in negative emotions, hence, the average sentiment shown in Figure 6 declines. On the other hand, for the Japanese dataset we observe gradual decrease in the occurrence frequency of positive terms and the increase in the

usage of negative terms along with the time. The explanation of this result may be related to somewhat negative image of the future among Japanese due to the aging society, long-running recession and the threat of emerging economies such as BRIC countries. Quite interesting is the inverse relationship in the usage of both positive and negative sentiment expressions at round dates in English and Japanese. At these dates the number of positive terms increases in Japanese while decreases in English. The situation is reverse for the case of negative terms. This phenomenon could be explained partially by cultural differences and by the fact that round dates may contain important international and well-established events (such as the ones commonly reported in news articles in the round years as shown in Figure 4), hence their emotional descriptions tend to be more similar across different languages.

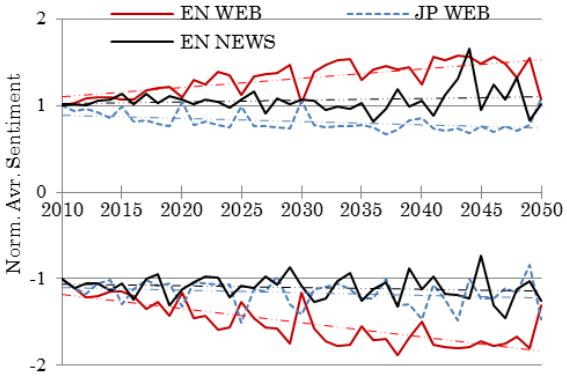

Fig. 5 Distribution of positive and negative terms in datasets with their respective trend lines.

V. CONCLUSIONS

In this paper we have analyzed collective future expectations expressed in three different languages on the Web. We think that any methods and applications that are aiming to portray the collective future images should take into account the characteristics and peculiarities of future-related information. For example, when evaluating the probability or a sentimental attitude to a certain future event that is expected to happen in a given future year, one should consider not only the absolute support level of this prediction (e.g. the number of individual predictions referring to this event and their polarity) but also relate it to the overall amount and character of information associated with this year.

We have concentrated on the subset of future-related information which concerns time-referenced events. The extension to consider also future, temporally non-referenced events would surely result in higher recall, but should necessarily involve efficient ways of filtering such information from the Web. This is however non-trivial as the future can be expressed in many different ways.

We investigated the attention amount, topics and sentiment levels of predictions as well as their comparison across different languages. We also showed that the analysis of future-related information on the Web can be achieved in a relatively simple and computationally inexpensive way through crawling search indices and light-weight text processing. We hope that in the future similar studies would bring us closer to the objective of accomplishing multi-dimensional analysis of collective future thinking of humanity.

REFERENCES

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

[2] M. D. Choudhury, H. Sundaram, A. John and D. D. Seligmann. Can Blog Communication Dynamics be Correlated with Stock Market Activity? *Proceedings of HT 2008*, pp. 55-60, 2008.

[3] G. Dias, R. Campos and A. Jorge. Future Retrieval: What Does the Future Talk About? *Proceedings of SIGIR 2011 Workshop on Enriching Information Retrieval (ENIR 2011)*, 2011.

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

[5] O. Etzioni, M. J. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, A. Yates. Web-scale Information Extraction in Knowitall: (Preliminary Results). *Proceedings of WWW 2004*, pp. 100-110, 2004.

[6] C. Fellbaum (ed.). *WordNet. An electronic lexical database.*. MIT Press, Cambridge, MA, USA, 1998.

[7] B. Flyvbjerg. From Nobel Prize to Project Management: Getting Risks Right. *Project Management Journal*, vol. 37, no. 3, pp. 5-15, 2006.

[8] D. Gruhl, R. V. Guha, R. Kumar, J. Novak, A. Tomkins. The Predictive Power of Online Chatter. *Proceedings of SIGKDD 2005*, pp. 78-87, 2005.

[9] 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.

[10] 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*, pp. 1123-1124, 2010.

[11] A. Jatowt and C. A. Yeung: Extracting Collective Expectations about the Future from Large Text Collections. *Proceedings of CIKM 2011*, pp. 1259-1264, 2011.

[12] K. Kanazawa, A. Jatowt and K. Tanaka. Improving Retrieval of Future-Related Information in Text Collections. *Proceedings of Web Intelligence 2011*, pp. 278-283, 2011

[13] N. Kanhabua R. Blanco and M. Matthews. Ranking Related News Predictions. *Proceedings of SIGIR 2011*, pp. 755-764, 2011.

[14] 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.

[15] Y. Liu, X. Huang, A. An and X. Yu. ARSA: a Sentiment-aware Model for Predicting Sales Performance Using Blogs. *Proceedings of SIGIR 2007*, pp. 607-614, 2007.

[16] G. Mishne and N. Glance. Predicting Movie Sales from Blogger Sentiment. Proceedings of the Spring Symposia on Computational Approaches to Analyzing Weblogs, 2006.

[17] A. Pepe and J. Bollen. Between Conjecture and Memento: Shaping a Collective Emotional Perception of the Future. *Proceedings of the AAAI 2008 Spring Symposium on Emotion, Personality and Social Behavior*, 2008.

[18] P. Senellart. Notes on the Timestamping of Web Pages. *Unpublished report, Télécom ParisTech, 2008* http://pierre.senellart.com/publications/senellart2008notes.pdf

[19] J. Surowiecki. The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations. Little, Brown, 2004.

[20] B. Wuthrich, D. Permunetilleke, S. Leung, V. Cho, J. Zhang and W. Lam. Daily Prediction of Major Stock Indices from textual WWW Data. *Proceedings of SIGKDD 1998*, pp.364-368, 1998.