Future Retrieval: What Does the Future Talk About? *

Gaël Dias
University of Beira Interior
Covilhã, Portugal
DLU/GREYC
Caen, France
ddg@di.ubi.pt

Ricardo Campos
Polytechnic Institute of Tomar
Tomar, Portugal
LIAAD - INESC Porto
Porto, Portugal
ricardo.campos@ipt.pt

Alípio Jorge
University of Porto
Porto, Portugal
LIAAD - INESC Porto
Porto, Portugal
amjorge@fc.up.pt

ABSTRACT

Predicting the future has always been one of the main aims of human beings in order to adapt their behavior and maximize their chances of success. With the advent of the Web, which indexes a wealth of temporal information, a great number of research have been proposed in the area of Temporal Information Retrieval, but Future Retrieval has remained a difficult problem to handle. In this paper, we propose to understand what the future is about. In particular, we present an exploratory study to understand how the temporal features impact upon the classification and clustering of different "genres" of future-related texts.

Categories and Subject Descriptors

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

General Terms

Algorithms, Measurements, Experimentation

1. INTRODUCTION

Predicting the future has always been the Holy Grail of Mankind. Although we cannot know the future, a lot can be inferred from it by mining huge collections of texts such as the Web and Microblogs (e.g. Twitter, Facebook). Indeed, future events, prophecies or forecast analyses have traditionally been edited on written documents. As a consequence, retrieving texts with their future intent is likely to benefit a lot of Web and Business Intelligence applications. For example, from the query *Dacia*, retrieving information such as *Dacia plans the release of no less than 8 new models and facelifts by 2015* is of the utmost importance for allowing companies and individuals to adapt their behavior and maximize their chances of success. This research area is commonly known as Future Retrieval (FR).

The foundations of Future Retrieval have been settled by Ricardo Baeza-Yates [4]. The overall idea is to seek for future temporal references within the Web from text/time queries. To answer to such a challenge, a FR system should be composed of (1) an information extraction module that recognizes temporal expressions, (2) an Information Retrieval system that indexes articles together with time segments and (3) a Text Mining system that given a time query, finds the most important topics associated with that time segment.

The challenges posed by this definition are difficult to handle. As a consequence, simpler tasks have been tackled [1] [2]. In particular, Adam Jatowt et al. [2] propose to summarize future events based on a partitioning clustering algorithm. Their idea is to explicitly introduce the time feature into the measure of similarity between documents. For that purpose, they propose the linear interpolation defined in Equation 1 between two documents d_i and d_j as a new time-related similarity measure, where $TermDist(d_i, d_j)$ (resp. $TimeDist(d_i, d_j)$) corresponds to the Euclidean distance between the vectors of time features (resp. tf.idf features).

$$Dist(d_i, d_j) = (1 - \beta).TermDist(d_i, d_j) + \beta.TimeDist(d_i, d_j) \quad (1)$$

The important result of their study is the fact that best results are obtained for $\beta = 0.2$. It is clear that the impact of the introduction of the time feature in the similarity measure is low. In this paper, we propose to go further in the analysis of future-related documents and try to understand the effects of temporal features both for classification and clustering based on three genres of future-related Web snippets[1]: informative (the future date can not be predicted in advance), scheduled (the future date can be predicted in advance) and rumors. For this purpose, we use six different learning algorithms (i.e. Naive Bayes, Multinomial Naive Bayes, K-NN, Weighted K-NN, Multi-Class SVM and K-means) having in mind that our main objective is not to reach high accuracy results but instead understand the impact of temporal features over different learning paradigms.

[1]The scope of our research is Temporal Ephemeral Clustering for Meta Search Engines. As a consequence, we focus on Web snippets instead of Web documents.

2. DATA ANALYSIS AND COLLECTION

Web news stories have mostly been the source of analysis for Future Retrieval as their temporal intent is usually easier to define [4] [2] compared to Web snippets or Web documents on general. Usually, Web news stories are informative or scheduled events and rumors or prophecies are unlikely to appear. However, our analysis of future-related Web snippets shows that many rumors can be identified. We illustrate these three text "genres" in the following three sentences.

1. *Sony Ericsson release postponed for February 2011 due to Software issues.* (Informative - not predictable)

2. *It has been announced that Qatar will host the 2022 FIFA World Cup.* (Scheduled - predictable)

3. *Avatar 2? in 2013? Cameron intends to complete his next film in 3 to 4 years.* (Rumor)

Unlike existing works [4] [2], which deal with Web news stories, we propose to use the entire Web to retrieve a large spectrum of future-related Web documents. Moreover, [4] [2] base their analysis on a small number of manually defined queries (3 for [4] and 20 for [2]). In order to propose a more robust set of experiments, we searched the Web based on a set of 450 queries[2] over 27 categories extracted from Google Insights for Search[3], which registers the hottest and rising searches performed worldwide. In particular, queries were extracted between January and October 2010.

To build our dataset, we used Yahoo! and Bing APIs to retrieve 200 Web results for each query, thus giving rise to a set of unique 62,842 Web snippets. In order to extract the temporal information from the set of Web snippets, we performed a pattern matching methodology as proposed in [3], which focuses exclusively on years as in [1][4]. From this labeled data set, we extracted 508 future Web snippets from a set of 5,777, which contained year dates. This means that 9.19% of the Web snippets contain year dates and only 0.81% contain future dates. Moreover, a further analysis showed that 82.48% of the future dates were related to a near future (i.e. a few months after the query time) and only 17.52% were related to a further future (i.e. at least one year after the query time).

Based on our previous study about the "genres" of future-related documents, we analyzed the distribution of the 508 extracted Web snippets according to these three categories in Table 1. These results are particularly interesting, as many future-related documents are scheduled documents as shown in [2] and [4], although most of future-related texts deal with informative statements, but also evidence gossips, comments or even prophecies (e.g. *Maya predictions of the end of the world in 2012 have seriously scared some people*). As a consequence, we also studied the distribution of Web snippets by text "genre" over their focus dates (i.e. near or far future). The results are presented in Table 2 and show that while informative texts tend to occur mostly within near future dates, scheduled events and rumor texts happen at distant years, especially for the gossip/rumour category.

[2]The queries do not contain time features.
[3]http://www.google.com/insights/search
[4]Our research is based on language-independent ephemeral clustering. For that purpose, we only use year dates.

The final data set consists of 508 future-related Web snippets, manually labeled as informative, scheduled or rumor as well as if they occur in a near or far future. This data set will be used to infer the impact of temporal features over classification and clustering as shown in the next section.

3. EXPERIMENTS AND RESULTS

Our exploratory analysis tries to define whether the temporal features play any role in future-related Web Snippet classification and clustering or not. The basic idea is to understand whether the category of future-related documents can be discovered by using only specific linguistic features or it can be improved by including temporal features.

The first experiments, which we conducted are based on whether a unigram model, combined or not with temporal features, is able to classify future-related Web snippets by their "genre" (i.e. informative, scheduled or rumors). For that purpose, we built four different balanced data sets: (1) one with all unigrams present in the Web snippets together with their year dates (D1), (2) one with unigrams present in the Web snippets withdrawing their year dates (D2), (3) one with unigrams present in the Web snippets with their year dates plus the mention of their belonging to a near or far future (D3) and finally (4) one with unigrams present in the Web snippets without their year dates plus the mention of their belonging to a near or far future (D4). The experiments were run on the basis of a 5-fold cross-validation for boolean and tf.idf unigram features for five different classifiers: the Naive Bayes algorithm (boolean), the Multinomial Naive Bayes algorithm (tf.idf), the K-NN[5] (boolean), the Weighted K-NN[6] (tf.idf) and the Multi-Class SVM (boolean and tf.idf). Results are presented in Tables 3 and 4 and show that the importance of the temporal features is heterogeneous for the classification task.

For the boolean case, both the Naive Bayes and the K-NN show improved results with the use of explicit year dates (i.e. D1 vs. D2 and D3 vs. D4). Moreover, the Naive Bayes largely outperforms the K-NN in accuracy. However, the K-NN shows the highest differences with and without explicit time features (i.e. year dates). As a consequence, we can confirm the results of [2] as the K-NN has been used with the Euclidean Distance. However, with the SVM, which reaches the best results overall, the impact of explicit time features is negative. Indeed, best results are obtained without any mention of time features. So, it would seem that the language used in each text "genre" is enough to classify future information. Interestingly, for the tf.idf representation, all results (except one) are worst than for the boolean case. Moreover, the behavior of each algorithm changes. In particular, all algorithms provide best results for D4 compared to D3, which would mean that the explicit mention of time feature would not benefit the classification task. However, the introduction of the near/far future date feature improves results overall, except for the Weighted K-NN. Another interesting conclusion is the fact that all algorithms show improved results with D1 compared to D2 (even the Multi-Class SVM), which would mean that the year dates are important in this case, when the near/far future date feature is missing.

[5]K=10.
[6]K=10 and the $1/distance$ weight.

To complete our study, we also proposed a set of experiments based on the well-known K-means clustering algorithm to understand the impact of temporal features within this process. The idea is to automatically retrieve three different clusters (informative, scheduled and rumors) based on the previous representations of Web snippets i.e. D1, D2, D3 and D4. As in the classification task, we provide experiments for the boolean and tf.idf cases. Results are presented in Table 5 and also show different results depending on the representation of Web snippets. For the boolean case, the introduction of explicit year dates only improves the results when combined with the near/far future date feature. For the tf.idf case, the use of the explicit time stamps improves the results when the near/far future date feature is missing. However, when the near/far future date feature is present, the use of explicit year dates has a negative effect on the results. In fact, this is not a surprise as the results of the K-means are similar to the (Weighted) K-NN results, which are all based on the Euclidean Distance.

The results obtained from our analysis are subject to discussion. Indeed, depending on the representation of Web snippets and on the algorithm family, the temporal issue may or may not have any influence. For the classification task, the SVM gives the overall best results without any temporal information with 79.22% accuracy for the boolean case, although the same Multi-Class SVM shows improved results for the tf.idf case when the near/far future date is introduced, reaching 79.20% accuracy. Moreover, the probabilistic learning and the lazy learning families always evidence best results when any time feature is used, to the exception of the Multinomial Naive Bayes for D3. As such, we can conclude that in most of the experiments, the time feature improves the results but this conclusion may not always stand and the time feature must definitely be treated in a special way depending on the learning algorithm and on the Web snippet representation. For the clustering task, and in particular for the for the K-means algorithm, the same conclusions can be drawn as for the lazy learning paradigm[7]. However, further experiments should be proposed with different (e.g. probabilistic) clustering algorithms to assess new exhaustive results.

4. CONCLUSION

In this paper, we proposed to analyze the impact of temporal features upon classification and clustering of future-related Web snippets. Our motivation for this study was the inconclusive results presented by [2], who show that temporal features slightly help to cluster future-related Web snippets. As a consequence, we proposed a set of exhaustive classification and clustering experiments based on three different future-related text "genres": informative, scheduled and rumors. Results show interesting issues as they depend on the learning algorithms and the Web snippet representation. It is true that fine-grained or coarse grained temporal features usually improve future-related text classification and clustering, but this may not be always the case. As such, further experiments must be carried out with different representations of time-related features in the learning process to reach final conclusions.

[7]This can be easily explained as the K-means is also based on the Euclidean Distance.

Acknowledgments

This research is funded by the Portuguese Foundation for Science and Technology through the VIPACCESS project with Reference PTDC/PLP/72142/2006 and the PhD scholarship with Reference SFRH/BD/63646/2009. Other funds also came from the ACM Special Interest Group on Information Retrieval (SIGIR, the organization).

5. REFERENCES

[1] J. Adam, K. Hideki, K. Kensuke, T. Katsumi, K. Kazuo, and Y. Keiji. Analyzing collective view of future, time-referenced events on the web. In *19th International World Wide Web Conference (WWW 2010)*, April 2010.

[2] J. Adam, K. Kensuke, S. Oyama, and T. Katsumi. Supporting analysis of future-related information in news archives and the web. In *Joint Conference on Digital Libraries (JCDL 2009)*, June 2009.

[3] R. Campos, G. Dias, and A. Jorge. What is the temporal value of web snippets? In *1st International Temporal Web Analytics Workshop of the 20th International World Wide Web Conference (TWAW 2011)*, March 2011.

[4] B.-Y. Ricardo. Searching the future. In *ACM SIGIR Workshop on Mathematical/Formal Methods for Information Retrieval (MF/IR 2005)*, August 2005.

APPENDIX

Table 1: Classification According to Text Genre.

Informative	Schedule	Rumors
255 (50.20%)	136 (26.77%)	117 (23.03%)

Table 2: Classification According to Focus Time.

	Informative	Schedule	Rumors
Near Future	55.85%	25.78%	18.37%
Far Future	23.60%	31.46%	44.94%

Table 3: Accuracy Results for the Boolean Case.

Algorithm	D1	D2	D3	D4
Naive Bayes	78.06%	77.21%	78.60%	78.06%
K-NN	58.11%	56.98%	62.67%	57.55%
Multi-Class SVM	79.20%	79.77%	78.63%	79.20%

Table 4: Accuracy Results for the tf.idf Case.

Algorithm	D1	D2	D3	D4
Multinomial N.B.	76.35%	75.78%	75.49%	76.53%
Weighted K-NN	59.25%	50.99%	56.41%	57.54%
Multi-Class SVM	75.21%	74.36%	74.92%	79.20%

Table 5: K-means Results.

Case	D1	D2	D3	D4
boolean	43.59%	43.59%	45.02%	41.88%
tf.idf	39.04%	35.90%	40.74%	51.00%