An Exploratory Study on How Temporal Information Impact Classification and Clustering of Future-Related Web Documents

Ricardo Campos, Gaël Dias, Alípio Jorge

An Exploratory Study on the Impact of Temporal Features on the Classification and Clustering of Future-Related Web Documents

Ricardo Campos[1,2,4], Gaël Dias[1,3], and Alípio Jorge[4]

[1] HULTIG, University of Beira Interior, Covilhã, Portugal
[2] Polytechnic Institute of Tomar, Tomar, Portugal
[3] DLU/GREYC, Univeristy of Caen Basse-Normandie, Caen, France
[4] LIAAD – INESC Porto LA, University of Porto, Porto, Portugal
ricardo.campos@ipt.pt, ddg@di.ubi.pt, amjorge@fc.up.pt

Abstract. In the last few years, a huge amount of temporal written information has become widely available on the Internet with the advent of forums, blogs and social networks. This gave rise to a new challenging problem called future retrieval, which consists of extracting future temporal information, that is known in advance, from web sources in order to answer queries that combine text of a future temporal nature. This paper aims to confirm whether web snippets can be used to form an intelligent web that can detect future expected events when their dates are already known. Moreover, the objective is to identify the nature of future texts and understand how these temporal features affect the classification and clustering of the different types of future-related texts: informative texts, scheduled texts and rumor texts. We have conducted a set of comprehensive experiments and the results show that web documents are a valuable source of future data that can be particularly useful in identifying and understanding the future temporal nature of a given implicit temporal query.

Keywords: Temporal Information Retrieval, Prospective Search, Temporal Web Mining, Temporal Classification, Temporal Clustering.

1 Introduction

With the advent of the World Wide Web, a huge amount of temporal data became available on the Internet ready to be exploited. This gave rise to the emergence of a new research area called Temporal Information Retrieval (T-IR). The purpose of this research area is to detect temporal data in documents and rank Internet search results based on temporal information. Alongside this, a new topic called future retrieval (FR) was introduced by Baeza-Yates [2] in 2005, with the specific goal of searching for future temporal references within web documents in order to answer queries that combine text and time. Such a system should include three components: (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. An example of this type of system would be a

system that would return information, such as *Dacia Coming in 2012* or *Dacia plans 8 new models by 2015,* for the query *Dacia*. This would benefit a number of users who are looking for future-related contents. Despite the relevance of this topic, little research has been conducted on using temporal information features for future search purposes, and the only known temporal analytics engine is Recorded Future [13].

1.1 Motivation

Although we cannot know the future, a lot can be deduced about it by mining huge collections of texts such as weblogs and microblogs (e.g. *Twitter*, *Facebook*). It is possible to look for events that are planned in advance, based on existing information. The following sentences show examples of three types of texts: informative texts, texts about scheduled events and rumors.

1. *Sony Ericsson Yendo Release Postponed for February 2011 Due to Software Issues.* (Informative – not predictable)
2. *The 2022 FIFA World Cup will be the 22nd FIFA World Cup, an international football tournament that is scheduled to take place in 2022 in Qatar.* (Schedule - predictable)
3. *Avatar 2? Arriving in 2013? James Cameron intends to complete his next film, another 3D epic, within three to four years.* (Rumor)

Based on this information we could, for example, decide whether or not to buy a property given the presumed tax increase, or to re-direct business negotiations based on economic predictions. Understanding the future temporal intent of web documents is therefore of the utmost importance. This is a particularly difficult task that has been mostly supported by a reliable collection of web news articles, annotated with a timestamp and that mainly consists of informative and scheduled texts. However this can also be based on several other types of sources such as web documents. In contrast to web news articles, web documents, especially those from social networks, suffer however from the problem of containing a large number of comments, predictions or plans, all expressed by means of rumors. This has led some authors, such as Adam Jatowt et al. [9], to question its credibility. However, what apparently seems to be a drawback, can actually constitute a great opportunity to infer the user's interests. For example, James Cameron may discover that people are interested in another 3D Avatar movie; mobile companies may redirect their core business to the development of mobile applications due to the growth of this industry that is expected to reach an impressive $35 billion by 2014; environmentalists may be interested to know that the easyJet airline plans to cut CO2 emissions by 50% by 2015.

Despite the relevance of this information, none of the proposed works to date have focused on this type of future-related documents, either of an informative nature, or rumors or scheduled texts. Therefore, developing an effective model to classify web documents with regard to their future intent, based on the temporal features incorporated in the text, is extremely important. Consequently, two challenging issues need to be considered: (1) Do web documents have enough temporal information for future analysis? (2) Can text classification and clustering be improved based on the existing future-related information in web documents? The aim is to answer these questions in this paper. We are particularly interested in considering a specific type of

query, the so-called implicit temporal query, which as stated by Campos et. al. [5], constitutes approximately 35% of all queries. This work takes place within the context of ephemeral clustering. The goal is to develop a language independent solution. As a consequence, this research focuses on the identification and extraction of numerical dates with years. Moreover, the study is based on the analysis of web content, rather than on a metadata-based approach. As noted by Klaus et al. [3], this is an interesting direction for future research, for which there is not yet a clear solution. Our study is based on web snippets which can be a powerful data source for future prediction as they reflect the views of society, as this paper will demonstrate.

1.2 Overview

The motivation behind this research is that: (1) to the best of our knowledge, this is the first work based on comprehensive future data analysis with web documents as a data source and implicit temporal queries (2) it is also the first work that aims to understand the impact of temporal features on both classification and clustering based on three genres of future-related texts (informative texts, scheduled texts and rumors).

Extensive experiments have been conducted to perform both types of analysis. In particular, the distribution of year dates present in snippets, in titles and their respective URLs was studied. Five different classification algorithms (Naive Bayes, Multinomial Naive Bayes, K-NN, Weighted K-NN, Multi-Class SVM) and one clustering algorithm (K-means) were then used to explore the main ideas. It must be noted that the main objective is not to attain a higher level of accuracy in the results, but instead to understand the impact of temporal features on different learning paradigms. The results of this paper show that web snippets are a valuable source of future data that can be particularly useful in identifying and understanding the future temporal nature of a given implicit temporal query. This exploratory study also shows that to some extent, the use of temporal features has an impact on the classification and clustering of future-related texts.

2 Related Work

Little research has been conducted so far in this area. However, there are some studies that do focus on this domain. Kira Radinsky et al. [12] use patterns in web search queries to predict whether an event will appear in tomorrow's news. Gilad Mishne et al. [11] predict movie sales through blogger sentiment analysis. Yang Liu et al. [10], focus on the same line of research and attempt to predict sales performance.

Indeed, it seems that only Baeza-Yates [2] and Jatowt et al. [8] [9] are concerned with Future Retrieval in a more general way. For example, Baeza-Yates [2] was the first to define this problem and to introduce a basic method for searching, indexing and ranking documents according to their future features. Each document is represented by a tuple consisting of a time segment and a confidence probability that measures whether the event will actually happen or not in this time segment. On the other hand, Jatowt et al. [9] propose the generation of visual summaries of future-related information on user queries using two methods. The first method is based on calculating the probability of the next instances of periodical events appearing in the future, through the analysis of past data, such as the statistics on document creation over time from the Google News Archive search engine. The second method relies on

the analysis of explicit future-related information contained in documents. Future events are clustered using a partitioning clustering algorithm in order to answer queries on named entities, such as the names of people or places. For that purpose, they propose the linear interpolation between two documents d_i and d_j as a new time-related similarity measure. This is illustrated in Equation 1.

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

The best results in terms of precision occur for $\beta = 0.2$. Consequently, it is clear that the impact of future-related features is relatively reduced. Finally, Adam Jatowt et al. [8] conducted an exploratory analysis of future-related information on the Internet. For that purpose, they gathered a set of queries in English, composed of temporal expressions. The queries (873.054) with a year reference ranging from 2010 to 2050 belong to the yearly dataset and the queries (39.312) with a month reference and a year ranging from 2010 to 2011 belong to the monthly dataset. Each query is then executed on Bing which is set to retrieve up to 1000 results, resulting in two sets of 1.044.224 and 770.715 unique Internet search results. Their analysis relies on the average number of hits obtained from the search engine for all of the queries and they show that (1) future-related information clearly decreases after a few years, with some occasional peaks, and that (2) most of the near future-related contents are related to expected international events. However, distant years are mostly linked to predictions and expectations that relate to issues such as the environment and climate change.

3 Measuring the Future Temporal Nature of Web Documents

This section assesses whether web snippets are a valuable source of data that can help deduce the future temporal intent of queries that do not specify a year. Unlike Jatowt et al. [8], the analysis is not based on the execution of future temporal explicit queries (queries with temporal expressions), but it is based on implicit queries. Subsequently, restrictions have not been placed on the language, type and topic of the query. Furthermore, this analysis is not based on the number of hits reported by the search engine, but on the detection and manual analysis of future dates that occur within the set of results retrieved. Moreover, in accordance with the work produced by Jatowt et al. [9], the impact of introducing future features on the process of classifying and clustering future-related web contents will be studied. However, unlike this work, more than 20 queries are used, and each text is classified according to three possible genres: informative web snippets, scheduled texts and rumors. This framework consists of four steps. Fig. 1. outlines the overall evaluation framework.

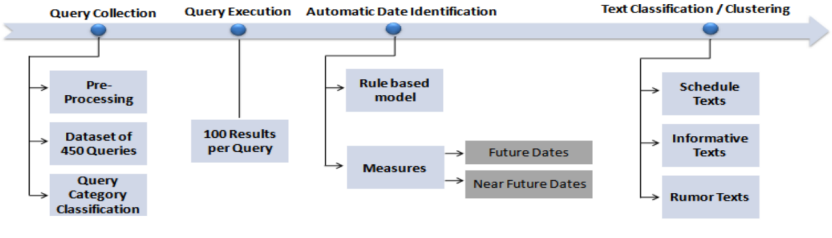

Fig. 1. Overview of the framework

3.1 Query Collection and Query Execution

Our dataset was collected by crawling the web in response to a set of diverse queries (see Table 1) selected from Google Insights for Search [7], which registers the most common searches performed worldwide. These queries are from the period January-October 2010. Twenty queries were manually selected for each of the 27 categories, which resulted in a total number of 540 queries. Since this research is looking at how web snippets temporally behave towards implicit temporal queries, explicit queries have not been included. Therefore, the final query collection consists of 450 queries (without duplicates) mostly belonging to the categories of the Internet.

Table 1. Example of queries

dacia duster	toyota recall	hairstyles	unemployment
avatar	lady gaga	tour de france	bank of america

To build the dataset, Yahoo! and Bing APIs were used and defined to retrieve a total of 200 web results per query resulting in a unique set of 62.842 web snippets. Each web snippet includes a title, a text (also known as a snippet), and a link, known as the URL.

3.2 Automatic Date Identification

In order to extract the temporal information from these data, a pattern matching methodology was performed, as proposed in [5]. Since the aim is to detect dates in the form of numerical years, to make the search efficient, and to keep the system language-independent, a custom built rule-based model was developed, which achieves results of almost 96% accuracy in the detection of dates within documents, particularly within titles and web snippets [5]. From this labeled data set, 5.777 web snippets, titles and URLs containing year dates were extracted. Each text was then manually labeled as indicative of a *near* or *distant future* purpose depending on the dates found in the text. If the date identified was from 2011, the text was labeled as a *near future* intention. If the date was later than 2011, or if the text had both a near and a distant future date, the text was labeled as having a *distant future* nature. The function *NearFuture*, is computed to all the queries, as the ratio between the number of dates retrieved labeled with the year 2011, divided by the total number of dates retrieved (see Equation 3).

$$NearFuture(q) = \frac{\#\ 2011\ Dates\ Retrieved}{\#\ Future\ Dates\ Retrieved} \tag{3}$$

In order to understand the future temporal value of each item more clearly and determine its value more easily, a basic measure represented by the function *FutureDates(q)* was defined, which is computed as the ratio between the number of dates retrieved with a future nature, for a given query *q,* divided by the number of dates retrieved for the same query (see Equation 2).

$$FutureDates(q) = \frac{\#\ Future\ Dates\ Retrieved}{\#\ Dates\ Retrieved} \quad (2)$$

A date in a document is considered of a future nature, if, regardless of the document timestamp, the focus time (the time of the content) is superior to the reading time (the time of the query). In this paper, the reading time is December 2010. As such, all years later than 2010 are considered future dates in the web snippets (e.g. *in 2014 the World Cup…*), with regard to the execution of an implicit query (e.g., *World Cup*). The final dataset consists of 508 web snippets, 419 titles and 195 URLs containing future dates. This data set will make it possible to perform classification and clustering tasks in order to understand the impact of temporal features.

3.3 Text Classification

Each of these texts (508 web snippets, 419 titles, 195 URLs) was then manually classified by three annotators who were asked to place them in three future temporal classes: informative texts, schedule texts or rumor texts. Fleiss' Kappa statistic [6] was used in order to measure the consistency between the different annotators. Results show Kappa was found to be 0.93, meaning an almost perfect agreement between the raters. Most of the differences occur in the classification of rumor and scheduled texts.

3.4 Data Analysis and Discussion of Results

This section outlines a number of issues on future temporal web mining analysis. The issues discussed include for example, the temporal value of future dates with regard to a given future year, the frequency of occurrence in a near future temporal window, related categories and text genres. Unlike conventional T-IR systems, where the amount of temporal information available is relatively significant, in a future retrieval system, values are naturally lower. That is perfectly clear in Table 2, where from a total number of 62.842 web snippets retrieved, 5.777 have temporal features and only 508 are of a future nature. This means that 9.2% of the web snippets contain year dates, but only 0.81% contain future dates. This is due to the fact that people talk about the past more than the future, which hampers the extraction of information in large quantities from a future retrieval system. Overall, these results would clearly be higher if the execution of explicit temporal queries had also been included (e.g. *hairstyles 2010* and subsequently *hairstyles 2011*). A recent work [5] has shown that 3.49% of the queries in a web query log collection have a future temporal intent, and that this value is inherently linked to the occurrence of higher results.

Nevertheless, it must be noted that the nature of a search in a conventional system is naturally different from a search in a future retrieval system, in which a person does not need much information to meet their objectives. Subsequently, it is important to note that albeit in a reduced scale, 149 queries, from the total number of 450 queries issued, retrieved at least one future date within the web snippet item (see Table 3), of which 32 had more than one future date. This means that of the 33.1% queries that retrieved a future date in a web snippet, 21.4% had more than one future date.

Table 2. Future dates analysis. *A* represents Absolute values, *R* represents relative values.

Item	Dates		Future Dates			Near Future Dates			Categories	
			#	A	R	#	A	R		
Web Snippet	5777	9.2%	508	0.8%	8.7%	419	0.6%	82.4%	Automotive Society Finance	33.1% 15.5% 11.0%
Title	2058	3.2%	419	0.6%	20.3%	373	0.5%	88.7%	Automotive Beauty Finance	49.3% 28.5% 23.8%
URL	3512	5.5%	195	0.3%	5.5%	167	0.2%	85.6%	Automotive Computer Sports	23.6% 9.2% 8.0%

Table 3. Classification of queries in terms of the occurance of future dates

Item	One Future Date		> One Future Date	
Web Snippet	149	33.11%	32	21.47%
Title	113	25.11%	14	12.38%
URL	75	16.67%	10	13.33%

Two of these cases are illustrated in the two following sentences: (1) *Japan plans to establish a robot moon base by 2020 with a landing by 2015* and (2) *FIFA denied that the process for the 2018-2022 World Cup was corrupt*. A closer study also shows that most of the future dates with relative (R) values occur in titles. Indeed, from a total number of 2.058 items tagged with dates, 20.3% (see Table 2) have a future temporal intent. This constitutes a rich set of information that could be considered when trying to infer the temporal nature of implicit queries. Regardless of a continuous shortage of future dates as we move forward in the calendar, a great number of references to far distant years are still found. The occurrence of dates is largely predominant in 2011, but consistent until 2013. Thereafter, there are some quite small peaks in 2014 and 2022 that mostly relate to the Football World Cup, which coincides with the results of [8]. Overall, the occurrence of future dates is very common in items retrieved in response to queries belonging to the categories of Automotive (e.g., *dacia duster*), Finance & Insurance (e.g., *Bank of America*), Beauty & Personal Care (e.g., *Hairstyles*), Sports (e.g., *football*) and Computer & Electronics (e.g., *hp*). A more detailed analysis of each of the three items: titles, snippets and URLs will now be presented.

Titles. On average, more than 90% of the future dates are related to the near future. This information is mostly related to economic forecasts, such as the expected growth of India, or the prediction that 2011 will be a good year to buy property. Some other examples are related to IT companies, for example the release date for electronic devices, or sport events, as illustrated in these titles: (1) *2011 will be best year to buy a home, says BSA*, (2) *Experts bet on India growth story in 2011*, (3) *Tour de France organizers unveil climb-heavy 2011 route* and (4) *Nokia to launch tablet in Q3 2011*. As we move forward in the calendar, reference years become more scarce such as with scheduled events, including the Football World Cup or rumors relating to environmental issues or company previews: (1) *Mobile App Revenue Estimated at $35*

Billion by 2014, (2) *Octopus Paul joins England's 2018 World Cup bid* and (3) *Qatar Plans 'Island Stadium' For 2022 World Cup*.

Snippets. Unlike in the titles, the occurrence of future dates in web snippets is not very common. Still, they occur in 8.79% of cases and they mostly include a short temporal window, such as 2011. Once again, we can note that most of the texts are related to economic forecasts concerning the worldwide crisis we are currently facing. References to upcoming events can also be seen, such as the Detroit Auto Show and an interesting political text on a visa agreement between Turkey and Azerbaijan: (1) *Honda is planning a major jump in hybrid sales in Japan in 2011*, (2) *Next-generation Ford 2012 Escape unveiled at the 2011 Detroit Auto Show* and (3) *Visa agreement expected to be signed between Turkey and Azerbaijan in 2011*. As with titles, business plans prevail in far distant years. References to PayPal accounts can be seen as well as sales of mobile applications or Adidas plans. Even those related to scheduled events have an economic nature, such as the Qatar Football World Cup reference. In addition, there are other quite interesting examples, one related to the translation of the Bible, another to the environment and another with the calendar of holidays until 2070. Some examples include: (1) *Avatar 2? in 2013? Cameron intends to complete his next film in 3 to 4 years*, (2) *IDC predicts sales of mobile apps will be a $35 billion industry by 2014*, (3) *Wycliffe's mission is to see a Bible translation in every language by 2025* and (4) *Calendar of all legal Public and Bank Holidays worldwide, until 2070*.

URLs. As expected, the occurrence of future dates in URLs is scarce when compared to web snippets or even titles. Indeed, only 5.6% of the links have a future temporal nature. Regardless of the fact that future dates are very uncommon in URLs, they can still be very useful in some specific cases. A careful observation of the list below leads to the conclusion that future dates in URLs are as descriptive as in titles or even in web snippets. Predictions are mostly related to IT companies, economic forecasts, and automotives, as this example shows (1) *http://www.grist.org/article/2010-11-15-fords-first-electric-car-to-be-sold-in-20-cities-in-2011*. Finally, references to far distant dates also appear in URLs such as (1) *http://www.london2012.com/* and (2) *http://msn.foxsports.com/foxsoccer/worldcup/story/world-cup-bid-usa-loses-2022-world-cup-bid-to-qatar*.

Web Snippets Genre. The distribution of items was also analyzed according to the three categories: informative texts, scheduled texts and rumor texts. On average (see Table 4), almost 77% of the texts have either an informative nature or concern a scheduled event which has a very high probability of taking place. The remaining 23% relate to rumor texts, which lack confirmation in the future. Some examples are listed here: (1) *WebOS tablet will arrive in March 2011. Details are not officially* (Rumor), (2) *Tickets for Lady Gaga 2011 Tour* (Scheduled Event) and (3) *Latest Hairstyles 2011* (Informative).

Table 4. Classification of texts according to genre

Item	#Items	Scheduled Events		Informative		Rumor	
web snippet	508	136	26.77%	255	50.20%	117	23.03%
Title	419	85	20.29%	248	59.19%	86	20.53%
URL	195	38	19.49%	101	51.79%	56	28.72%

Table 5. Classification of texts according to genre for near and distant future dates

Item	Near Future			Distant Future		
	Schedule	Informative	Rumor	Schedule	Informative	Rumor
Web Snippet	25.7%	55.8%	18.3%	31.4%	23.6%	44.9%
Title	15.0%	65.4%	19.5%	63.0%	8.7%	28.2%
URL	13.7%	56.8%	29.3%	53.5%	21.4%	25.0%

While informative texts mostly occur with near future dates, schedule events and rumor texts occur more frequently with far distant years (see Table 5). Words such as *latest*, *new*, *review*, *information*, *schedule*, *announce*, *official* and *early* are usually used to describe the near future in informative texts, such as information on product releases (e.g., *Dacia Duster*, *Audi*, *Toyota*, *Ford*, *Honda*, *Nissan*, *Nokia*, *Microsoft*) and upcoming scheduled events (e.g. *Auto Show*).

As we move forward in the calendar, it is more common for texts to be related to events planned in advance and to also be of a rumor nature. These are associated with events that require confirmation in the future, as shown in Table 5. Long term schedule events such as the Olympic Games in London or the FIFA Football World Cup in Brazil and also in Qatar, and rumor words such as *planning*, *report*, *preview*, *coming*, *expecting*, *rumor*, *scenarios*, *reveal* and *around* often replace words with a near future nature, such as *early* or *new*. Another interesting issue to note is the fact that future dates are mostly year related and fewer are related to months or days. This becomes more apparent further into the future. Exceptions only occur with scheduled events. The following sentence is an illustrative example: *Tour de France: from Saturday July 2nd to Sunday July 24th 2011, the 98th*.

4 Classification and Clustering of Future-Related Texts

This paper aims to understand whether data features influence the classification and clustering of future-related texts according to their nature: informative, scheduled or rumor. It is important to note that the goal is not to achieve high accuracy results but to understand if these three genres can be discovered by simply using specific linguistic features, thus avoiding the importance of time for these tasks, or if instead, the inclusion of temporal features plays an important role.

4.1 Classification of Future-Related Texts

This study includes cross-domain experiments by selecting and issuing queries for the set of 27 categories available. The Aue et al. [1] and Boey et al. [4] model that suggests training a classifier on a domain-mixed set, in order to tackle cross-domain learning, was used. Experiments are based on two collections[1]: one consisting of 508 snippets and another consisting of 419 text titles, both tagged with future dates. URL texts were not included in this experiment. A selection of 117 text snippets of Informative nature, 117 of Scheduled intent and 117 of Rumor purpose were collected, together with 86 text titles of Informative nature, 86 of Scheduled intent and 86 of Rumor purpose. The result is a set of 351 balanced texts snippets and 258

[1] Available at `http://www.ccc.ipt.pt/~ricardo/software` [17th June, 2011].

Table 6. Datasets structure

Dataset	Web Snippet		Near/Distant Future	Class
	Unigram	Year Dates		
D1	x	x		x
D2	x			x
D3	x	x	x	x
D4	x		x	x

Table 7. Web Snippet classification results for the boolean and tf.idf cases

Algorithm	Case	Dataset	Accuracy	Scheduled		Informative		Rumor	
				Precis.	F-Mea	Precis.	F-Mea	Precis.	F-Mea
Naïve Bayes	Boolean	D1	78.1%	84.2%	75.4%	77.8%	77.8%	74.1%	80.5%
	Boolean	D2	77.2%	80.8%	74.1%	78.6%	78.6%	73.3%	78.6%
K-NN	Boolean	D1	58.1%	52.0%	58.1%	56.7%	51.4%	67.9%	64.6%
	Boolean	D2	57.0%	48.2%	60.3%	67.3%	43.0%	68.3%	63.3%
Multi-Class SVM	Boolean	D1	79.2%	87,3%	81,3%	75,2%	77,7%	76,6%	78,8%
	Boolean	D2	79.8%	87,0%	80,2%	75,6%	78,7%	78,2%	80,5%
Multi-Class SVM	TF.IDF	D1	75.2%	83,0%	76,5%	69,5%	72,7%	74,8%	76,7%
	TF.IDF	D2	74.4%	85,6%	77,6%	66,9%	71,2%	73,6%	74,8%
M. Naïve Bayes	TF.IDF	D1	76.4%	78.6%	78.6%	79.4%	72.0%	72.3%	78.0%
	TF.IDF	D2	75.8%	76.0%	77.3%	79.6%	72.6%	72.7%	77.1%
Weighted K-NN	TF.IDF	D1	59.3%	87.5%	61.9%	65.3%	49.7%	48.8%	63.3%
	TF.IDF	D2	51.0%	51.5%	55.0%	66.7%	35.2%	46.9%	56.2%
Naïve Bayes	Boolean	D3	78.6%	84.4%	76.1%	77.8%	77.8%	73.9%	80.0%
	Boolean	D4	78.1%	83.5%	75.7%	79.1%	78.4%	73.4%	79.7%
K-NN	Boolean	D3	62.7%	59.1%	62.7%	57.1%	57.6%	74.0%	68.2%
	Boolean	D4	57.6%	50.0%	59.5%	60.7%	50.0%	59.7%	57.2%
Multi-Class SVM	Boolean	D3	78.6%	86,3%	80,4%	73,8%	76,5%	77,2%	79,2%
	Boolean	D4	79.2%	87,1%	80,7%	74,2%	77,6%	77,9%	79,5%
Multi-Class SVM	TF.IDF	D3	74.9%	83.7%	76.3%	67.7%	72,0%	75,8%	76,8%
	TF.IDF	D4	79.2%	87.1%	80.7%	74.2%	77,6%	77,9%	79,5%
M. Naïve Bayes	TF.IDF	D3	75.5%	78.3%	77.6%	78.4%	71.0%	71.2%	77.3%
	TF.IDF	D4	76.5%	75.2%	75.2%	82.8%	73.3%	77.0%	76.1%
Weighted K-NN	TF.IDF	D3	56.4%	86.8%	54.1%	66.7%	49.5%	46.3%	61.3%
	TF.IDF	D4	57.5%	50.0%	59.5%	60.7%	50.7%	68.4%	61.3%

balanced text titles, from which four datasets D1, D2, D3 and D4 (see Table 6) were built, respectively. Each dataset is labeled with the respective text genre/class. In particular, (D1) consists of texts with their year dates, (D2) consists of texts withdrawing their year dates, (D3) consists of texts with their year dates plus the mention of their belonging to a near or distant future and (D4) consists of texts without their year dates plus the mention of their belonging to a near/distant future.

Experiments are 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 K-NN (k = 10, boolean), the Multinomial Naive Bayes algorithm (tf.idf), the Weighted K-NN (K = 10 and weight=1/distance, tf.idf) and the Multi-Class SVM (boolean and tf.idf). Results are presented in Table 7 and show that the importance of

temporal features in the classification task is heterogeneous, as it depends on the learning algorithm and on text representation.

In general, all of the algorithms (see Fig. 2), with the exception of SVM (boolean) show improved results in terms of accuracy with the simple use of explicit year dates. The greatest difference is in the Weighted K-NN algorithm. However, both Naïve Bayes and SVM (boolean) largely outperform the Weighted K-NN in terms of accuracy. In contrast, the dates do not have a great impact if combined with near/distant future knowledge. Indeed, Multi-Class SVM (boolean and tf.idf), Multinomial Naïve Bayes and Weighted K-NN provide better results for D4 than D3. Equally, in the comparison between D1 and D2, the greatest difference in accuracy occurs with the K-NN algorithm. Once again, the Naïve Bayes and SVM (boolean) achieve the best results.

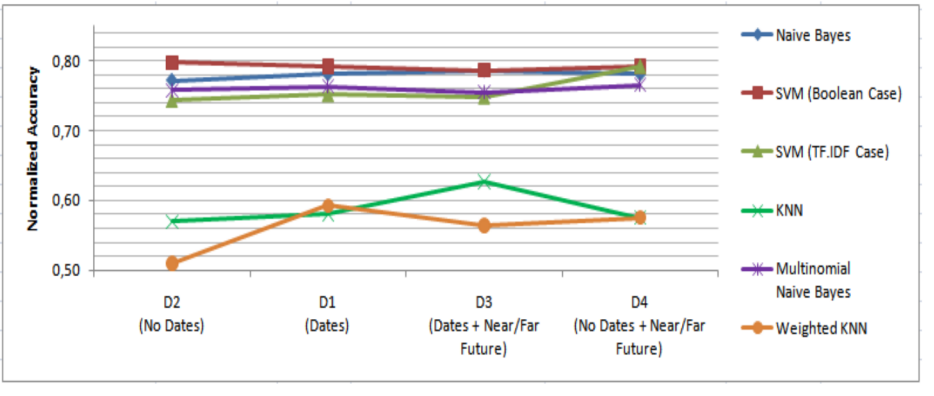

Fig. 2. Overall analysis of global accuracy for Web Snippets texts

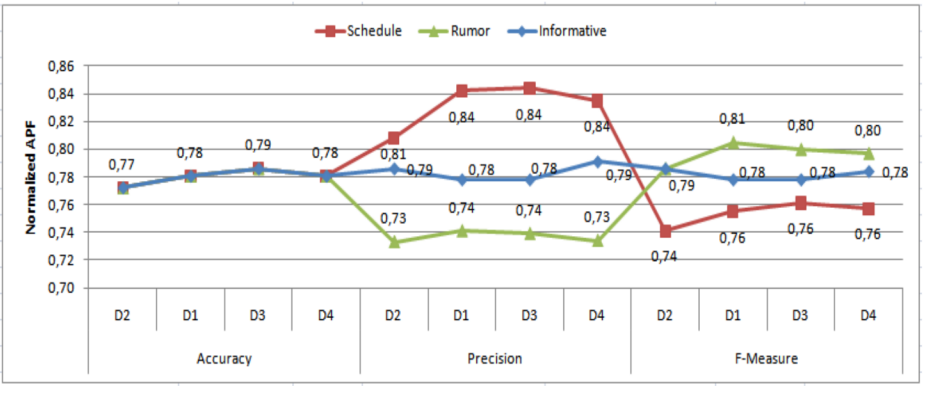

Fig. 3. Text genre analysis for Naïve Bayes (D1,D2) and (D3,D4) comparison

An individual analysis of each text genre (informative, scheduled, rumor) also led to the conclusion that the introduction of temporal features has an overall positive impact on precision in the classification of scheduled texts. In contrast, the

classification of informative texts is more accurate without dates and this is uncertain in the case of rumor texts. Overall these conclusions are confirmed by F-Measure for scheduled and informative texts, but interestingly, not for rumor texts, which show an overall positive impact with F-Measure with the introduction of time features. The best results, however, occur for the SVM algorithm (boolean) without the use of any temporal features. Fig. 3 shows the results for the specific case of Naïve Bayes.

The same experiments performed on the web snippets were then performed on the set of 258 balanced text titles. The results are shown in Table 8.

Table 8. Title classification results for the boolean and tf.idf cases

Algorithm	Case	Dataset	Accuracy	Scheduled		Informative		Rumor	
				Precis.	F-Mea	Precis.	F-Mea	Precis.	F-Mea
Naïve Bayes	Boolean	D1	78.1%	83.5%	75.7%	79.1%	78.4%	73.4%	79.7%
	Boolean	D2	79.9%	77.8%	83.2%	74.1%	80.0%	96.4%	75.2%
K-NN	Boolean	D1	54.3%	71.6%	62.7%	44.7%	60.4%	100%	24.5%
	Boolean	D2	55.4%	56.9%	67.0%	51.2%	61.0%	100%	17,0%
Multi-Class SVM	Boolean	D1	74,4%	75,0%	75,9%	66,7%	72,3%	85,3%	75,3%
	Boolean	D2	76,4%	74,7%	78,5%	70,5%	74,0%	86,8%	76,6%
Multi-Class SVM	TF.IDF	D1	72.9%	71.4%	73.4%	66.7%	70.3%	83.1%	75.2%
	TF.IDF	D2	76.4%	73.5%	78.3%	71.3%	74.4%	87.9%	76.3%
M. Naïve Bayes	TF.IDF	D1	77.9%	78.9%	80.7%	70.4%	78.4%	90.0%	74.0%
	TF.IDF	D2	76.4%	76.5%	81.5%	69.3%	74.9%	88.1%	71.7%
Weighted K-NN	TF.IDF	D1	53.1%	70.0%	62.8%	43.8%	59.1%	100%	20,8%
	TF.IDF	D2	53.1%	53.5%	64.2%	50.8%	60.0%	100%	11,0%
Naïve Bayes	Boolean	D3	72.9%	71,8%	71,3%	63,6%	74,4%	96,2%	72,5%
	Boolean	D4	77.9%	75,3%	79,8%	71,0%	78,8%	96,3%	74,3%
K-NN	Boolean	D3	53.9%	71,9%	61,3%	44,5%	60,4%	100%	24,5%
	Boolean	D4	52.7%	70,4%	63,7%	43,6%	58,9%	100%	17,0%
Multi-Class SVM	Boolean	D3	75,2%	75,9%	76,3%	67,3%	73,7%	86,6%	75,8%
	Boolean	D4	75,6%	76,4%	77,7%	66,7%	73,3%	89,1%	76,0%
Multi-Class SVM	TF.IDF	D3	73,6%	73.0%	74.3%	67.0%	73.0%	84.8%	73.7%
	TF.IDF	D4	74.4%	75.0%	75.9%	65.7%	73.2%	88.7%	74.3%
M. Naïve Bayes	TF.IDF	D3	77.1%	77.8%	79.5%	70.0%	78.6%	89.7%	72.2%
	TF.IDF	D4	77.1%	76.5%	81.5%	71.3%	77.0%	88.1%	71.1%
Weighted K-NN	TF.IDF	D3	52.3%	69.7%	60,5%	43,4%	59,0%	100%	46,8%
	TF.IDF	D4	51.1%	62.7%	61,5%	44,1%	58,6%	100%	43,7%

Overall, it is clear that most of the algorithms (see Fig. 4) perform worst in terms of accuracy with the introduction of temporal features, meaning that time characteristics do not have a great impact on the classification task. This does not happen with the Multinomial Naïve Bayes, which has one of the best overall results, only supplanted by the Naïve Bayes algorithm.

This is confirmed by a detailed analysis of all three types of text genres, where the Multinomial Naïve Bayes algorithm shows successful results. Overall, for almost all of the algorithms, scheduled texts benefit with the introduction of temporal features, which is not as clear in the case of informative texts. Another interesting result is that

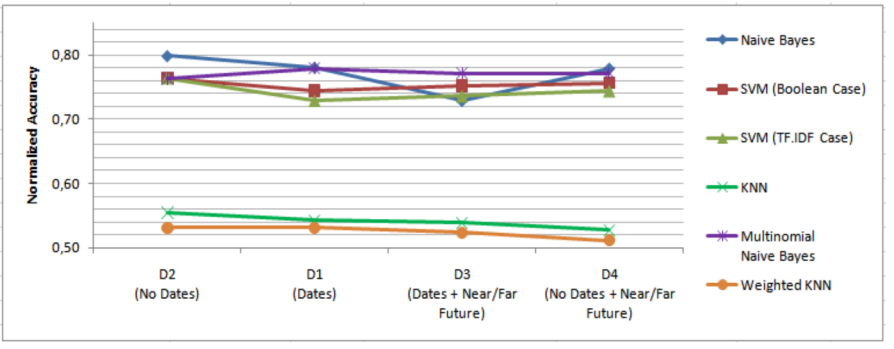

Fig. 4. Overall analysis of global accuracy for Titles texts

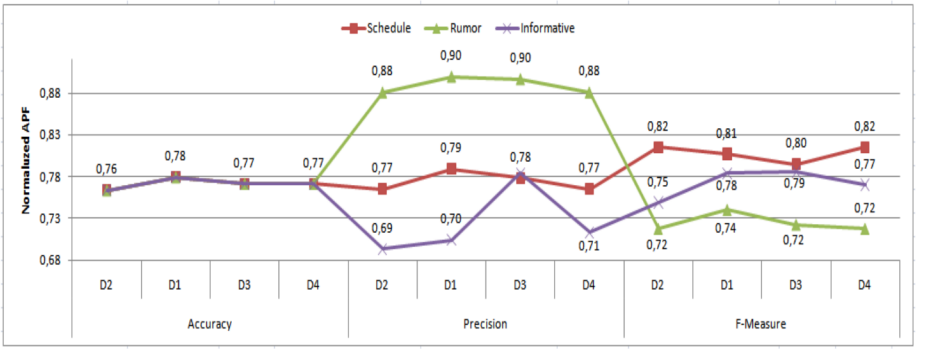

Fig. 5. Text genre analysis for Multinomial Naïve Bayes (D1,D2) and (D3,D4) comparison

precision in rumor texts is very high. However, with the exception of the Multinomial Naïve Bayes algorithm, time features do not have an overall impact on the classification task. The following figure (see Fig. 5) shows these results for the specific case of the Multinomial Naïve Bayes algorithm.

4.2 Clustering of Future-Related Texts

Finally, a set of experiments based on the well known K-means clustering algorithm was proposed in order 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 same representations of web snippets, the D1, D2, D3 and D4. As in the classification case, experiments for the boolean and tf.idf cases, and for snippets and text titles are shown.

Results for text snippets are presented in Table 9 and show that they are more sensitive to the near/distant future feature, as the best results, for the Boolean case, are obtained for D3. However, the best overall results are obtained by using the K-means over D4, which only takes into account a coarse-grained temporal feature. It must also be noted that scheduled texts have a very high precision rate of almost 85% with a positive impact on the use of temporal features.

Table 9. Web snippet clustering results for the K-means in the boolean and tf.idf cases

Algorithm	Case	Dataset	Correctly Clustered	Scheduled Precision	Informative Precision	Rumor Precision
K-Means	Boolean	D1	43.59%	34.7%	59.5%	41.1%
		D2	43.59%	34.7%	59.5%	41.1%
		D3	45.02%	36.0%	55.8%	50.0%
		D4	41.88%	33.9%	46.6%	43.6%
	tf.idf	D1	39.04%	84.6%	35.6%	20.0%
		D2	35.90%	83.3%	34.4%	29.4%
		D3	40.74%	25.0%	38.0%	50.6%
		D4	51.00%	43.4%	50.5%	58.4%

This is a clear contrast to text titles clustering, as the best results occur for D3 in the tf.idf representation, with nearly a 13% impact when compared to D4 (Table 10). Moreover, the use of temporal features, either alone or combined with near/distant future knowledge, show a positive impact in the clustering task, but for rumor texts they reach an impressive value of almost 85% in terms of precision. The results obtained were not conclusive for D1 and D3 (Boolean case), in that more than two clusters were not found. A more detailed analysis led to the conclusion that this is mostly because the system appears to have some difficulties in splitting schedule texts from those of a rumor nature.

Table 10. Title clustering results for the K-means in the boolean and tf.idf cases

Algorithm	Case	Dataset	Correctly Clustered	Scheduled Precision	Informative Precision	Rumor Precision
K-Means	Boolean	D1	39,54%			
		D2	42,25%	34.9%	47.5%	84.5%
		D3	39,54%			
		D4	42,25%	34.9%	47.5%	84.5%
	tf.idf	D1	41,87%	34.7%	37.6%	82.4%
		D2	41,87%	37.1%	37.0%	79.3%
		D3	53,49%	68.0%	45.0%	82.8%
		D4	41,87%	37.5%	35.8%	79.3%

5 Conclusion

In this paper, we conducted an exploratory analysis of future information on the Internet. Results show that titles, particularly in the near future, contain a broad range of temporal information, which is still significant in the case of text snippets and URLs. In addition, we conclude that texts are more often of a scheduled and rumor nature as we move forward in the calendar, contrary to what happens with informative texts, which are unlikely to appear. The high precision of these results and the work presented by Adam Jatowt et al. [9], who has shown that temporal features can help cluster future-related web snippets, led to our final experiments. We performed a set

of exhaustive classification and clustering tests based on the three different future-related text genres (informative, scheduled and rumors). The results obtained from our analysis are subject to discussion. Indeed, depending on the representation of the text and on the algorithm family, the temporal issue may or may not have any influence.

For the classification task, the SVM and the Naïve Bayes provide the best overall results for text snippets and text titles respectively. However, none of these results was obtained using temporal features. Moreover, the probabilistic learning and the lazy learning families always show the best results for the classification of text snippets when any time feature is used, with the exception of the Multinomial Naive Bayes and the Weighted K-NN for D3. This is the opposite of what happens with the classification of text titles, where most of the algorithms perform better without temporal features. Furthermore, we can also conclude that in general, the introduction of temporal features has an overall positive impact on the classification of scheduled texts, both in snippets as well as in text titles. Interestingly we can also note that the detection of rumor texts benefits from the introduction of temporal features, particularly in the probabilistic algorithms. For the clustering task, and in particular for the K-means algorithm, the impact of temporal features is more apparent in D1 for snippets and in D3 for text titles. Moreover, the identification of schedule texts is particularly easy in text snippets, while rumor texts are easily identified in text titles.

We believe that this information will serve to improve temporal knowledge in terms of the aims of the user's query, and is a step towards the formation of a future search engine, where the returned documents relate to future periods of time. As such, time features must definitely be treated in a special way and further experiments must be carried out with different representations of time-related features in the learning process, to reach final conclusions and to assess new exhaustive results in the clustering process.

Acknowledgments. This research was part-funded by the PhD grant with reference SFRH/BD/63646/2009 from the Portuguese Foundation for Science and Technology (FCT). This work was also supported by the VIPACCESS project with reference PTDC/PLP/72142/2006 funded by the FCT.

References

1. Aue, A., Gamon, M.: Customizing Sentiment Classiers to New Domains: a Case Study. In: RANLP 2005, Borovets, Bulgaria, September 21-23 (2005)
2. Baeza-Yates, R.: Searching the Future. In: MFIR 2005 associated to SIGIR 2005, Salvador, Brazil, August 15-19 (2005)
3. Berberich, K., Bedathur, S., Alonso, O., Weikum, G.: A language modeling approach for temporal information needs. In: Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., van Rijsbergen, K. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 13–25. Springer, Heidelberg (2010)
4. Boey, E., Hens, K., Deschacht, K., Moens, M.-F.: Automatic Sentiment Analysis of On-Line Text. In: ELPUB 2007, Vienna, Austria, June 13-15 (2007)
5. Campos, R., Dias, G., Jorge, A.M.: What is the Temporal Value of Web Snippets? In: TWAW2011 Associated to WWW 2011, Hyderabad, India, March 28 (2011)

6. Fleiss, J.L.: Measuring Nominal Scale Agreement Among many Raters. Psychological Bulletin 76(5), 378–382 (1971)
7. Google Insights for Search, http://www.google.com/insights/search
8. Jatowt, A., Kawai, H., Kanazawa, K., Tanaka, K., Kunieda, K.: Analyzing Collective View of Future, Time-referenced Events on the Web. In: WWW 2010, Raleigh, USA, April 26 - 30, pp. 1123–1124 (2010)
9. Jatowt, A., Kawai, H., Kanazawa, K., Tanaka, K., Kunieda, K.: Supporting Analysis of Future-Related Information in News Archives and the Web. In: JCDL 2009, Austin, USA, June 15-19, pp. 115–124 (2009)
10. Liu, Y., Huang, X., An, A., Yu, X.: ARSA: A Sentiment-Aware Model for Predicting Sales Performance Using Blogs. In: SIGIR 2007, Amsterdam, Netherlands, pp. 607–614 (July 2007)
11. Mishne, G., Glance, N.: Predicting Movie Sales from Blogger Sentiment. In: CAAW 2006 Associated to AAAI 2006, Boston, USA, July 16-20 (2006)
12. Radinsky, K., Davidovich, S., Markovitch, S.: Predicting the News of Tomorrow Using Patterns inWeb Search Queries. In: WIC 2008, Sydney, Australia, pp. 363–367 (2008)
13. Recorded Future, http://www.recordedfuture.com/