What is the Temporal Value of Web Snippets?

Ricardo Campos
LIAAD – INESC Porto, LA
Centre for Human Language Technology and Bioinformatics,
University of Beira Interior,
Polytechnic Institute of Tomar, Portugal
ricardo.campos@ipt.pt

Gaël Dias
Centre for Human Language Technology and Bioinformatics
University of Beira Interior, Portugal
ddg@di.ubi.pt

Alípio Mário Jorge
LIAAD – INESC Porto, LA
DCC - FCUP
University of Porto, Portugal
amjorge@fc.up.pt

ABSTRACT
The World Wide Web (WWW) is a huge information network from which retrieving and organizing quality relevant content remains an open question for mostly all implicit temporal queries, i.e., queries without any date but with an underlying temporal intent. In this research, we aim at studying the temporal nature of any given query by means of web snippets or web query logs. For that purpose, we conducted a set of experiments, which goal is to assess the percentage of web snippets or queries (in query logs) having temporal features, thus checking whether they are a valuable source of data to help on inferring the temporal intent of queries, namely implicit ones. Our results show that web snippets, as opposed to web query logs, are an important source of concentrated information, where time clues often appear. As a consequence, they can be particularly useful to identify and understand "on-the-fly" the implicit temporal nature of queries in the context of ephemeral clustering.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Search Process, Query formulation.*

General Terms
Measurement, Experiments.

Keywords
Temporal Information Retrieval, Implicit and Explicit Temporal Queries, Temporal Query Classification.

1. INTRODUCTION
Time is everywhere in the World Wide Web causing Temporal Information Retrieval (T-IR) to be increasingly involved in recent years in a systematic research process by the IR community. Notwithstanding this, few works have fully used temporal information for exploration and search purposes [3]. The lack of such an approach from major search engines, with the exception of Google[1], prevents on the one hand, users from being aware of possible historical perspectives of given subjects, and on the other hand, the modeling of user queries according to a specific period over time, potentially causing a loss in accuracy and recall. From a query point of view, this is mainly due to the fact that systems do not infer temporal intents expressed by users in a query. From the document point of view, although time clues such as dates may be found in the texts, they are usually not taken into account in the representation and indexing processes. The main reason is certainly due to the difficulties that exist to correlate temporal information to topics in documents.

Understanding the temporal intent of documents and queries is therefore of the utmost importance to produce high quality information retrieval systems. In our research, we aim at developing a methodology able to explicitly time-stamp temporal implicit queries such that temporal disambiguation can be reached "on-the-fly". This work takes place in the context of ephemeral clustering, specifically within our meta-search engine VipAccess[2] In particular, we intend to develop a language independent solution. Thus, we will specifically focus on the identification and extraction of year dates. Our framework is based on web content analysis, rather than on metadata information. As claimed by [7] this is an interesting future research direction, for which there is not a clear solution yet.

This paper is the result of part of this research. Our main purpose here is to study whether web snippets are a valuable source of data to help inferring the temporal intents of queries, either implicitly or explicitly formulated. Since results will be performed "on-the-fly", we need to adopt a web content analysis approach over the set of *k*-top web snippets retrieved, as opposed to an analysis over full web pages. In parallel with what has been done by [21], we also analyzed the temporal value of web query logs. We considered the possibility of using them for temporal query understanding purposes, but we came to the conclusion that beyond being highly dependent on the user own intents, web query logs are a particular hard temporal reference collection to access, outside the big industrial labs, and do not easily deal with query ambiguity [19].

To the best of our knowledge, this is the first work towards a comprehensive data analysis having web snippets as a data source. Indeed, [2] used them, but in another line of research, namely the construction of a time-centered snippet that highlights temporal information. Results obtained from our experiments are promising. They show that regardless of the implicit temporal nature of the query, web snippets contain a broad range of temporal information that can be used as a valuable source to help on inferring the temporal intent of queries, either implicitly or explicitly formulated.

However, one possible drawback of our research, is that web snippets are computed by search engines, which we do not control. As a consequence, basing our system upon results generated by a black box may prevent from obtaining a clear picture of the temporal values of web snippets. Nevertheless, [28] proved in the context of ephemeral clustering that web snippets are likely to provide the correct clustering of documents as they

[1] With its timeline tool incorporated within Google.com, although without any details about its architecture.

[2] http://hultig.di.ubi.pt/vipaccess [7th February, 2011].

TWAW'11, March 28, 2011, Hyderabad, India.

embody the excerpts of documents mostly related to the query terms. Given this, we think that our analysis can serve as a good approximation of the real situation. In particular, we will base our study on three different search engines: Goggle[3], Yahoo![4] and Bing[5].

The remainder of this paper is organized as follows. In Section 2, we review temporal data reference collections commonly used in the most varied activities of T-IR. We particularly emphasize data collections of web snippets and web query logs. In Section 3, we present our experiments based on a web snippet data set and a well-known web query log. In particular, we detail each data set and introduce the methodology to extract the temporal information. Then, in Section 4, we intend to assess the temporal value of web snippets and web query logs. In particular, we first analyze the temporal intents of web snippets and their use to date implicit queries. In a second experiment, we analyze explicit temporal data in web query logs. Both results are then examined in Section 5 and we conclude this paper in Section 6 with some final remarks.

2. TEMPORAL DATA COLLECTIONS

Time can be expressed in a number of different forms, depending on how the temporal intent is defined. In queries, for example, they tend to occur by means of explicit (e.g., *Football World Cup 2010*) or implicit temporal intents (e.g., *Football World Cup*). Usually, this information is stored in web query logs, which are a flat set of files that record the activity registered in a server in different ways: (1) from an infrastructural perspective (e.g., date and time of the request) or (2) from the analysis of information use (e.g., user search intentions: *CHI 2011 vs. CHI*). Due to its private nature, some collections were made publicly available for research purposes. One of the most known is the AOL collection consisting of 21,011,240 queries collected from 650.000 users over three months (01 March, 2006 - 31 May, 2006) of activity within the AOL search engine[6]. More recently, Microsoft released two large scale datasets for research purposes on learning to rank [26]. The first one consists of 30,000 queries and the later of 10,000 queries.

Temporal expressions can also be found in a number of different types of documents such as web pages, web snippets or news articles, following an explicit (e.g., *WWW 2011*), implicit (e.g., *SIGIR*) and relative temporal intent (e.g., *in the next month*). Web documents are one of the most used sources for research purposes, with some publicly available datasets such as the Clueweb09 [8]. Compared to this data source, web snippets are a simple small section of text that provides an easy way to quickly represent information, forming a very interesting set of data where dates, especially in the form of years, often appear. Both, are usually used for content analysis purposes, clearly in the opposite direction of news document collections, where the focus is not so much on content but on metadata information, namely time-stamped documents annotated with the date of creation or publication (e.g., Reuters news stories [23]; TREC corpora [20]; TDT corpus [18]; AQUAINT-2 collection [27]; TimeBank 1.2 Corpus [22]; The New York Times Annotated Corpus [24]; ACE Time Normalization (TERN) 2004 English Training Data [12]; mostly accessible through the LCD web site [17]).

3. EXPERIMENTAL SETUP

To understand the temporal value of web snippets and web query logs, we considered the use of two independent datasets. We describe each one in the following sections along with the rule-based model used to automatically identify dates within each of them.

3.1 Data Sets

For our experiments, we rely on two datasets: a series of web snippets and associated titles and URLs obtained for the execution of 465 and 450 queries executed in December 2010 and a large-scale query log dataset belonging to 2006. Both are fully available for download[7], together with all the information produced during the execution of the experiments.

3.1.1 Web Snippets Data Set

To construct our first data set, we executed a set of queries based on our meta-search engine VipAccess excluding Google from its search interface. At first sight, it may seem strange not to use Google, but there are two main reasons for that: (1) web snippets have recently began to come together with some metadata information (e.g., blog post dates), which would introduce noise in our study and (2) Google has officially deprecated its web search API in November 2010, limiting the number of requests that can be executed per day.

The selection of the queries was made from Google Insights for Search[8], which registers the hottest queries performed worldwide. We manually selected a total of 540 queries from the period between January 2010 and October 2010, as a result of an individual selection of 20 queries for each of the 27 pre-defined categories. We were left with 465 queries, including some temporally explicit ones, after removing duplicates. We then created a reduced version of 450 queries, consisting of only the implicit ones. Each query is then executed on our meta-search engine VipAccess defined to retrieve a set of 20 and 100 triple items <snippet, title, url> so as to observe any variations that may exist due to different amounts of retrieved data. In practice this represents 40 and 200 results, given our meta-search engine runs over two search engines.

Given this, we ended up by constructing three different collections[9] (*Q465R20*, *Q450R20* and *Q450R100)*, each one gathering a different number of retrieved items (see Table 1). Most of the queries belong to the categories of Internet (12.69%), Computer & Electronics (9.89%) and Entertainment (7.96%).

3.1.2 Web Query Log Data Set

Our second data set, labeled Q601, is a set of queries extracted from the AOL collection. It consists of 21,011,240 queries, 10,154,742 if we only consider single entries. From this collection, we automatically selected only those queries with explicit temporal intent. So, we ended up with 143,590 queries, from which we selected a representative sample of n=601 queries with a maximum tolerated average sampling error of E=4% for a

[3] http://www.google.com [7th February, 2011].

[4] http://www.yahoo.com [7th February, 2011].

[5] http://www.bing.com [7th February, 2011].

[6] http://search.aol.com [7th February, 2011].

[7] http://www.ccc.ipt.pt/~ricardo/software [7th February, 2011].

[8] http://www.google.com/insights/search [7th February, 2011].

[9] Where *Q* means the number of queries to run and *R* the number of results to retrieve for each query.

confidence interval of 95% (see Equation (1) based on [6]) where Z_p is the *p*-th quantile of the normal distribution, which in this case is equal to 1.96.

$$n = \frac{z_p}{4E^2}, \quad (1)$$

Each query was then manually classified into a set of 29 pre-defined categories (see [9] for a detailed description of each one). Most of the queries belong to the categories of Automotive (21.96%), Entertainment (9.48%), Sports (8.15%), Business & Economics (5.99%) and News & Events (5.16%).

3.2 Rule-based Model

The identification of dates, either within web queries or web documents, is probably the most recognized area within this recent research field, with TempEx[10], GUTime[11] and ANNIE[12] taking the lead. However, given that we only aim at detecting dates in the form of numerical years in order to meet an independent language solution, there is no need to use any of these methodologies, which are all language-dependent. As a consequence, we ended up by defining our own rule-based model. Similarly to the works abovementioned, we considered numbers in the interval [1000..2099] that satisfy some specific patterns, such as *yyyy*, *yyyy-yyyy*, *yyyy/yyyy*, *mm/dd/yyyy*, *mm.dd.yyyy*, *dd/mm/yyyy* and *dd.mm/yyyy*.

Just by using this simple rule-based method, we achieved results of almost 96% accuracy in the detection of dates within documents, namely within titles and web snippets (see Table 1). However, the application of these rules within URLs results in the return of some noisy information, which is no big surprise. Overall, false dates usually tend to occur in the response of queries belonging to the categories of Internet (e.g., *1600 YouTube Videos*), Computer & Electronics (e.g., *1024 x 768)*, Games & Toys (e.g., *1000 games*) and Food & Drink (e.g., *1001 recipes*). To reach this conclusion, we manually went through each of the items of the three collections *Q465R20, Q450R20 and Q450R100* and checked whether they were correctly labeled or not.

We performed the exact same process for the identification of dates within queries (*Q601*). Results obtained in this case show an accuracy of 86%. Again false dates tend to occur mostly within the category of Computer & Electronics (e.g., *hp 1430*).

Of course, we could improve our rule-based model by limiting the appearance of some noisy information. One way to do this would be to include some terms like *year*, or some temporal prepositions such as *in* or *between*. However, this would go in the opposite direction of our language independent goal. As such, we should try to find alternative solutions in future work. One of the possibilities is to model, with some degree of confidence, the relationship existing between the topics and the possible dates.

4. METHODOLOGY

Based on the web snippet data set, we performed a set of experiments to analyze their temporal value and their potential usefulness in the process of dating temporal implicit queries. Additionally, we also intended to study the temporal value of web query logs, particularly the relationship existing between queries and dates, i.e. temporal explicit queries. We believe this information will serve to improve our temporal knowledge with respect to the user query intents, either implicitly or explicitly defined. We detail each framework in the following sections.

4.1 Temporal Value of Web Snippets

In this first experiment, we are particularly interested in studying the existence of temporal information within web snippets, and to check if such information can be used to temporally classify implicit queries. We follow a twofold approach: (1) we collect the necessary data and (2) we classify queries with respect to their temporal nature. We summarize the overall evaluation framework in Figure 1.

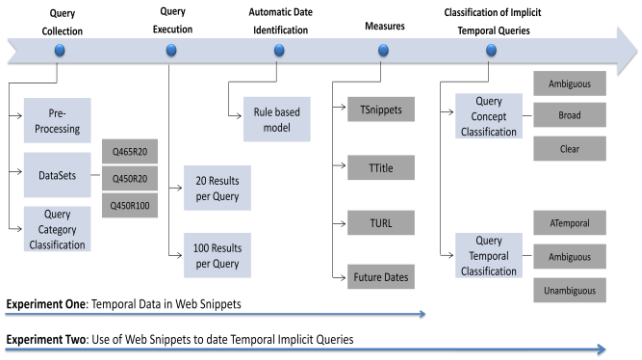

Figure 1: Web Snippets Framework.

4.1.1 Query Execution

The first step of our framework, after minor adjustments (pre processing phase and query category classification), is to process the three collections (*Q465R20*, *Q450R20* and *Q450R100*) on our meta-search engine, which defines the web snippet data set.

4.1.2 Automatic Date Identification

Upon the retrieved results, particularly over each triple item <snippet, title, url>, we ran our self-defined rule-based model in order to mark dates expressed by means of numerical patterns, particularly year dates (see Figure 2).

Figure 2: Title, Web Snippet and URL returned to the query *Alice in Wonderland*.

Each item is individually marked through an XML schema. The set of files is available for download[13].

4.1.3 Evaluation Metrics

In order to better understand and determine the temporal value of each item, we defined three basic measures taking into account the query *q*. They are represented by the functions *TSnippets(q)*, *TTitle(q)* and *TUrl(q)*. *TSnippets(q)* is computed as the ratio between the number of snippets returned with dates divided by the total number of snippets returned by our meta search engine (see Equation (2)). *TTitle(q)* and *TUrl(q)* are computed similarly and respectively defined in Equations (3) and (4).

[10] http://fofoca.mitre.com [7th February, 2011].

[11] http://www.timeml.org/site/tarsqi/modules/gutime/ [7th February,2011].

[12] http://www.aktors.org/technologies/annie/ [7th February, 2011].

[13] http://www.ccc.ipt.pt/~ricardo/software [7th February, 2011].

$$TSnippets(q) = \frac{\#\ Snippets\ Retrieved\ With\ Dates}{\#\ Snipptes\ Retrieved} \quad (2)$$

$$TTitle(q) = \frac{\#\ Titles\ Retrieved\ With\ Dates}{\#\ Titles\ Retrieved} \quad (3)$$

$$TUrl(q) = \frac{\#\ Urls\ Retrieved\ With\ Dates}{\#\ Urls\ Retrieved} \quad (4)$$

4.1.4 *Temporal Classification of Implicit Queries*

Query temporal classification is a particular hard task in the case of implicit temporal queries (e.g., *WWW*, *Scorpions*), in the sense that temporal information is not available, at least in a direct way. One possible solution to estimate the temporal value of a query is to compare it with complementary knowledge sources, such as web snippets or web queries logs. In this experiment, we aim at classifying temporal implicit queries with web snippets. A preliminary step however is involved. We need to first classify the query with regard to its conceptual ambiguity such that each different meaning[14] may have a different temporal dimension. To this purpose we follow the approach of [25], which defines three types of concept queries (see Figure 3), reducing this to a simple classification problem, with two possible classes: A (ambiguous queries) and $\overline{A}$ (broad or clear query).

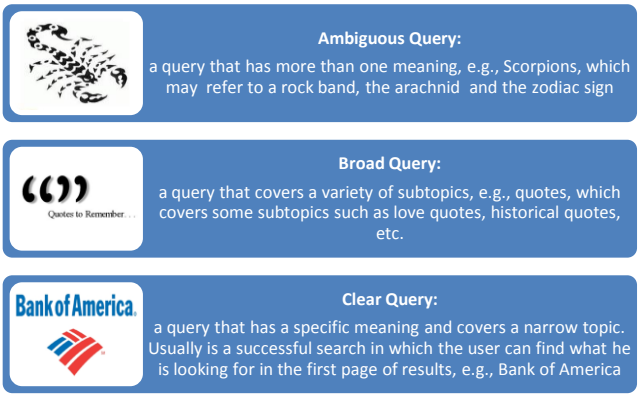

Figure 3: Taxonomy of ambiguous queries in concept. Adapted from [25].

We adopted the following methodology. First, we attempt to confirm if the query is ambiguous in terms of concept, i.e., if it belongs to class A. We use the disambiguation Wikipedia feature, to confirm, whether or not the query under study has more than one meaning. Second, if we conclude that the query has a single meaning, thus belonging to class $\overline{A}$, we attempt to confirm whether the query is broad or clear. For that purpose, we run each query on our meta-search engine, which clusters web page results through an ephemeral clustering process and retrieves its possible different sub-topics. Then, based on the assumption that only clear concept queries may be temporally classified, we introduce (see Figure 4), as in [14], three possible temporal classes: ATemporal, i.e. queries not sensitive to time (e.g. *make my trip*), Temporal Unambiguous, i.e. queries that take place in a very concrete time period (e.g. *Haiti Earthquake* or *BP Oil Spill*) and Temporal Ambiguous, i.e. queries with multiple instances over time (e.g. *World Cup* or *Oil Spill*).

[14] *WWW* may be the web conference, but also the World Wide Web itself. *Scorpions* can be associated to the rock band, the arachnid and the zodiac sign.

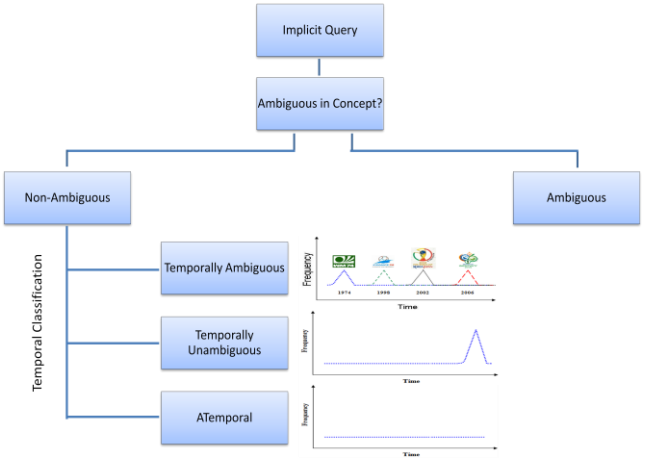

Figure 4: Temporal classification of clear concept queries.

Each query is classified into one of these three categories based on the temporal value of the triple items <snippet, title, url> retrieved. We call this temporal value temporal ambiguity. To compute it, we define (see Equation (5)) a weighted average of the *TSnippet(.)*, *TTile(.)* and *TUrl(.)* functions. Given the fact that dates occur in a different proportion in any of the items <snippet, title, url>, we value each differently through ω_f, where f is the function regarding the corresponding item. Considering the average measures obtained for *TSnippet(.)*, *TTile(.)* and *TUrl(.)*, for each of the collections (see Table 1) we define ω_f as 66.10% for *TSnippet(.)*, 20.75% for *TTile(.)* and 13.15% *TUrl(.)* for the *Q450R20* collection and respectively 50.91%, 18.14% and 30.95%, for the *Q450R100* collection.

$$TA(q) = \sum_{f \in I} \omega_f \,.\, f(q), I = \{TSnippet(.), TTitle(.), TUrl(.)\} \quad (5)$$

Based on the *TA(.)* function, we aim at classifying each query as A (ATemporal query) or $\overline{A}$ (Temporal Ambiguous or Temporal Unambiguous query). In particular, we classify each query as ATemporal if it has a *TA(.)* value below 10%. Otherwise, we conclude that the query belongs to class $\overline{A}$. In this case, we just have to decide whether the query is temporally ambiguous or not, based on the fact that it is associated to more than one year.

4.2 Temporal Value of Web Query Logs

Our purpose in this second experiment, based on the *Q601* collection, is to complement our knowledge about temporal information by understanding the explicit relationships existing between queries and dates. We are particularly interested in studying the temporal value of web query logs, but this may also prove interesting to understand two phenomena: (1) are users interested in future dates when looking for a given subject? and (2) what is the type of information they are seeking when issuing a query together with a date? In Figure 5, we summarize the overall evaluation framework. The first step is to use our rule-based model in order to automatically identify years within queries. Then, we manually identify false and future dates, and classify each of the queries in a set of 29 pre-defined categories listed in [9].

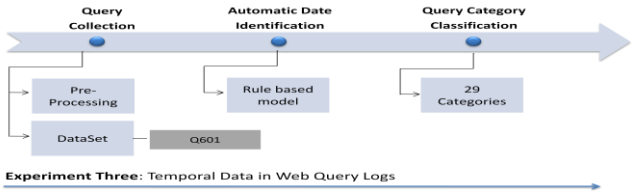

Figure 5: Web Query Logs Framework.

5. DISCUSSION

In this section, we discuss the various issues of our temporal web mining proposal over our web snippet data set and web query log. Our aim is to evaluate the extent and usefulness of temporal information in both collection. Questions discussed include, for example, to what extent temporal information in a result web snippet can indicate the temporal interest of the information needed by the user. In another strand of this research, we analyze the temporal value of web query logs and discuss to what extent the lack of a temporal value prevents the use of this kind of data source to work as knowledge base for query understanding purposes.

5.1 Temporal Value of Web Snippets

5.1.1 Temporal Data in Web Snippets

In this first experiment, we are particularly interested in studying the existence of temporal information within web snippets, namely within triple items <snippet, title, url>. We are particularly focused on extracting year dates, which are a kind of temporal information that often appears in this type of collection.

To this end, we conducted three tests in December 2010, *Q465R20*, *Q450R20* and *Q465R100*. Each corresponding query was executed on our meta-search engine VipAccess. Thus, we collected 16,648 triple items for the first test, 16,129 for the second one, and 62,842 for the final one. Then, upon these web page results, particularly over each triple item retrieved, we ran our rule-based model so as to automatically identify dates in the form of numerical years. We achieved results at about 96% of accuracy within web snippets, 98% within titles, but significantly less in the case of URLs with the worst value equals to 75%. To get these values we went through each of the items of the three collections, manually correcting each false positive occurrence.

Upon these labeled examples, we then computed the corresponding metrics *TSnippets(.)*, *TTitle(.)* and *TUrl(.)*. Obtained results (see Table 1) are according to our expectations. On average 10% of the web snippets retrieved for *Q450R20* and *Q450R100* collections have a temporal feature. This value is significantly higher for *Q465R20* collection, as it includes 15 explicit temporal queries (e.g., *hairstyles 2010*), which obviously implies the retrieval of a larger range of outcomes than usual. The occurrence of temporal features is particularly evident in the case of web snippets, but still significant in the case of titles and URLs.

We can also note that the differences between *Q450R20* and *Q450R100* collections are minimal in terms of ratio. The only exception is with *TUrl(.)*.In fact, retrieving 20 results as opposed to 100 show a decrease in the number of identified dates. This clearly shows that the first web snippets retrieved by search engines do not tend so much to include years in their web snippets. A detailed analysis within the whole set of results led us to conclude that this last collection will obviously retrieve a large range of different dates, which may be useful for a full understanding of the temporal value of a given query.

Another important remark is that the occurrence of dates within triple items <snippet, title, url> often occurs with more than one date, with values close to 23% in the case of web snippets, but in a much smaller scale for the remaining dimensions (see Figure 6).

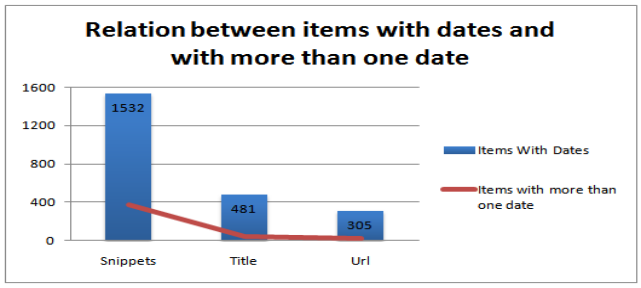

Figure 6: *Q450R20* items marked with more than one date.

It also becomes clear when observing Figure 7, that from 2003 onwards the existence of dates occur in a more intense manner, with a high peak in the period of 2008-2010 and that future dates, first introduced by [5] in 2005, also occur to a considerable extent, particularly in the case of titles, where 20% of the results retrieved have a future temporal intent. More recently [13] investigated the distribution of such information in documents and in historical document collections.

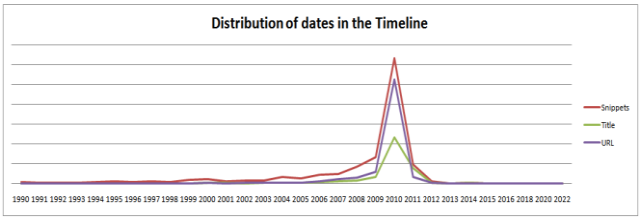

Figure 7: Distribution of dates in the timeline for *Q450R100* experiment.

Overall, we can conclude that dates occur more frequently in response to queries belonging to the categories of Dates (e.g. *calendar*), Sports (e.g., *football*), Automotive (e.g., *dacia duster*), Society (e.g., *baby*) and Politics (e.g., *Barack Obama*).

An individual analysis of one or two queries may also show us some interesting results. For instance, the implicit query *Tour de France* clearly formulated with an inherent temporal nature has a value of 77.78% for *TSnippet(.)* in *Q450R20*, with dates ranging at an annual basis, from the far distant year of 1903 to the current year 2011. Another example is the query *Toyota Recall* (see Figure 8), which despite the occurrence of other events over time may benefit of a clear temporal positioning because of a more distinct peak of occurrences between 2010-2011.

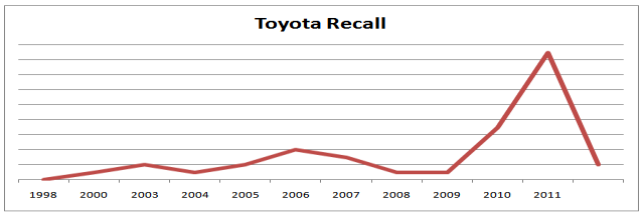

Figure 8: Graphical representation of the query *Toyota Recall* in the period of 1998-2011.

In Table 1 we can observe a summary of the results for the different experiments.

Table 1. Summary of Results for *Q465R20, Q450R20* and *Q450R100*

Tests	Web Page Results	Automatic Date Identification Accuracy		Average Measures	Future Dates
Q465R20	16,648	**WSp**	95.8%	12.4%	18.6%
		Title	97.9%	5.69%	13.8%
		URL	85.1%	4.26%	9.64%
Q450R20	16,129	**WSp**	94.3%	9.50%	6.5%
		Title	95.8%	2.98%	18.9%
		URL	75.0%	1.89%	8.8%
Q450R100	62,842	**WSp**	93.1%	9.19%	7.9%
		Title	95.3%	3.27%	19.7%
		URL	87.4%	5.59%	5.7%

Both tests show that documents in the web, particularly web snippets, are full of temporal expressions. However, they are not always exploited to increase the quality of the retrieval process.

5.1.2 Dating Implicit Queries using Web Snippets

Given the temporal value of web snippets, we aim at understanding if this temporal information can be used to automatically disambiguate query terms, namely implicit temporal queries. We rely on the *Q450R20* and *Q450R100* collections. Our approach is twofold. First, we classify each query with regard to its concept ambiguity. Final results (see Table 2) show that most of the queries are ambiguous in concept, although there is a large set of clear queries, which do not offer any doubt in terms of their meaning and a small set of broad queries.

Table 2. Classification of Queries both in Concept and Temporal meaning for the *Q450R20* collection

Conceptual Classification	Number Queries
Ambiguous	220
Clear	176
Broad	54

Temporal Classification	Number Queries	%
ATemporal	132	75%
Ambiguous	40	23%
Unambiguous	4	2%

We then temporally classified each of the 176 clear concept queries according to its temporal ambiguity *TA(.)* value. We classified each query as ATemporal if it had a *TA(.)* value below 10% (see Figure 9) and as temporally ambiguous or unambiguous, otherwise.

	Results for **20 Results per Query**						
2	Query	AmbQuery	TempQuery	tsnippets	ttitle	turl	Temporal Ambiguity
56	tattoos for girls	Clear Query	ATemporal	2,86%	0,00%	0,0%	1,89%
190	bed bugs	Clear Query	ATemporal	0,00%	0,00%	0,0%	0,00%
201	oil spill	Clear Query	TempAmbiguous	26,47%	17,65%	0,0%	21,16%
204	bp oil spill	Clear Query	TempUnambiguous	15,15%	9,09%	0,0%	11,90%
340	love quotes	Clear Query	ATemporal	0,00%	0,00%	0,0%	0,00%
452	make my trip	Clear Query	ATemporal	0,00%	0,00%	0,0%	0,00%
453							
1	Results for **100 Results per Query**						
2	Query	AmbQuery	TempQuery	tsnippets	ttitle	turl	Temporal Ambiguity
56	tattoos for girls	Clear Query	ATemporal	4,3%	0,00%	0,72%	1,89%
190	bed bugs	Clear Query	ATemporal	2,4%	0,00%	1,57%	0,00%
201	oil spill	Clear Query	TempAmbiguous	19,8%	6,87%	9,16%	14,19%
204	bp oil spill	Clear Query	TempUnambiguous	15,3%	8,03%	18,98%	15,13%
340	love quotes	Clear Query	ATemporal	0,8%	0,00%	0,76%	0,00%
452	make my trip	Clear Query	ATemporal	3,0%	0,75%	0,75%	0,00%

Figure 9: Temporal Ambiguity results for *tattoos for girls, bed bugs, oil spill, bp oil spill, love quotes* and *make my trip*.

The final analysis (see Table 2) based on the *Q450R20* collection show that of the total number of clear concept queries, 25% have implicit temporal intent, of which 23% are temporal ambiguous queries and 2% unambiguous. The remaining 75% are ATemporal queries i.e. queries for which no temporal intent is inherent. Results are very similar for *Q450R100* collection.

These values contrast with those presented in [19] who, based on web query logs, estimated that about 7% of all queries have an implicit temporal nature.

5.2 Temporal Value of Web Query Logs

In this experiment, we intend to analyze the temporal value of web query logs. In particular, we are interested in seeking for explicit temporal queries e.g. *Iraq War 1991* or *World Cup 2000*. For this purpose, we ran our rule-based model over the overall AOL queries log data set. We ended up with 143,590 queries. This represents only 1.41% of the entire collection, which is in line with the value of 1.5% presented by [21]. However, if we consider there are some false positives, due to the set of dates wrongly marked by our rule-based model, we may end up with an even lower value.

To better estimate these results, we decided to analyze the *Q601*collection. Each query was manually classified according to one of two categories: real date or false date. A further analysis of the results allowed us to conclude that 14.14% of the sample i.e. 87 queries were false positives. If we generalize these results to the whole temporal explicit queries set, then we can conclude that instead of having 143,590 implicit temporal queries we should only have 123,286 queries, causing, albeit not significantly, the reduction of the initial value of 1.41% to 1.21%.

Based on this study, we can conclude that dates are seldom used by the user to express his intents. Moreover, web query logs do not provide a data source adapted to concept disambiguation, although some studies tend to prove the contrary [19]. Besides, they are extremely hard to access outside the big industrial labs and highly dependent on the user own intents. We can particularly think in the two following illustrative cases: (1) queries that have never been typed, thus not existing in the web search log e.g. *Blaise Pascal 1623* (his year birth date) or (2) less year qualified queries that may be as relevant as the most frequents ones.

To better understand this problem, we can observe Figure 10, which clearly shows that despite significant peaks along the timeline, there is a lack of queries outside the period 2000-2006.

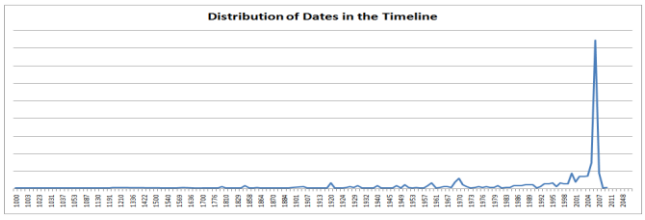

Figure 10: Distribution of dates in the timeline for *Q601* experiment.

Another interesting remark to note is that, notwithstanding a decrease from 2006 onwards (we recall that this collection is from 2006), future dates still represent 3.49% of the sample collection and that users are more interested in queries belonging to the categories of Automotive, Entertainment and Sports when queries are explicit in its temporal intent.

6. Conclusion

In recent years, time has been gaining an increasing importance in Information Retrieval in a large number of sub-areas. However, and despite the fact that the documents are full of temporal expressions, existing IR systems still do not exploit this information. On the one hand, this prevents systems to model the search process according to specific periods of time. On the other hand, it prevents the user to have an historical perspective of the results.

Inferring the user intentions and the period the user has in mind may therefore play an extremely important role in the possible improvements of IR systems by adding the temporal dimension, practically nonexistent until now. Aware of this, some works have been emerging in the last few years, yet most only use temporal information extracted within metadata, particularly from web news documents [7] [10] [14], [15], and only a few, [1] [3] [4] and [19] have attempted to extract temporal information within the contents of web documents, although the latter with the help of web query logs. Simultaneously other studies have appeared related with recency ranking, as opposed to user intentions understanding. In this regard [11], [16], [29]suggest to rank documents by relevance taking into account its freshness.

In this paper, we showed that query understanding is possible through the use of web snippets. Our experiments, strictly dependent on the value of the temporal information retrieved for a given query, show that web snippets are a very rich data source, where dates, especially years, often appear, and that they can embody a very important data source in query understanding, thus allowing "on-the-fly" temporal disambiguation for real-time IR systems. Moreover, working with web snippets may afford faster processing.

Our approach, however, may be negatively influenced by the production of the results of existing search engines considered in our meta-search engine, which we do not control. Such can cause for example, some documents to be discriminated from the very outset, preventing a broader temporal analysis. An elucidative example of this drawback is given by the query *Iraq War*, which if issued in any of the major search engines, will mostly retrieve results with regard to the conflict of 2003, as opposed to 1991 despite the fact the web is full of results concerning *Iraq War 1991*. Given this, we need to evaluate the feasibility of developing a search engine, albeit of a small scale, which will also enable us to compare a full text analysis approach with a web snippet based one.

7. ACKNOWLEDGMENTS

This research was partially funded by a grant (Reference: SFRH/BD/63646/2009) under the Portuguese Foundation for Science and Technology. A special thank to the four reviewers for the quality of their comments, which were extremely important to significantly improve the quality of this article.

8. REFERENCES

[1] Alonso, O., & Gertz, M. (2006). Clustering of Search Results using Temporal Attributes. *In SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 597 - 598). Seattle, Washington, USA. August 6 - 11: ACM Press.

[2] Alonso, O., Baeza-Yates, R., & Gertz, M. (2009). Effectiveness of Temporal Snippets. *In WSSP2009: Proceedings of the Workshop on Web Search Result Summarization and Presentation associated to WWW2009: 18th International World Wide Web Conference.* Madrid, Spain. April 20 - 24: ACM Press.

[3] Alonso, O., Gertz, M., & Baeza-Yates, R. (2009). Clustering and Exploring Search Results using Timeline Constructions. *In CIKM 2009: Proceedings of the 18th International ACM Conference on Information and Knowledge Management.* Hong Kong, China. November 2 - 6: ACM Press.

[4] Arikan, I., Bedathur, S., & Berberich, K. (2009). Time Will Tell: Leveraging Temporal Expressions in IR. *In WSDM 2009: Proceedings of the 2nd ACM International Conference on Web Search and Data Mining.* Barcelona, Spain. February 09 - 12: ACM Press.

[5] Baeza-Yates, R. (2005). Searching the Future. In S. Dominich, I. Ounis, & J.-Y. Nie (Ed.), *MFIR2005: Proceedings of the Mathematical/Formal Methods in Information Retrieval Workshop associated to SIGIR 2005: 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Salvador, Brazil. August 15 - 19: ACM Press.

[6] Barbetta, P. A., Reis, M. M., & Bornia, A. C. (2004). *Estatística para Cursos de Engenharia e Informática.* Lisboa: Atlas.

[7] Berberich, K., Bedathur, S., Alonso, O., & Weikum, G. (2010). A Language Modeling Approach for Temporal Information Needs. In C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, et al. (Eds.), *In Lecture Notes in Computer Science - Research and Advanced Technology for Digital Libraries, ECIR 2010: 32nd European Conference on Information Retrieval* (Vol. 5993/2010, pp. 13 - 25). Milton Keynes, UK. March 28 - 31: Springer Berlin / Heidelberg.

[8] Callan, J., Hoy, M., Yoo, C., & Zhao, L. (2009, 02). *ClueWeb09 Dataset*. Retrieved 12 23, 2010, from Carnegie Mellon University: http://boston.lti.cs.cmu.edu/Data/clueweb09/

[9] Campos, R. (2011). *Analysis of Temporal Data in Explicit Temporal Queries. AOL dataset.* Available at: http://www.ccc.ipt.pt/~ricardo/software

[10] Dakka, W., Gravano, L., & Ipeirotis, P. G. (2008). Answering General Time Sensitive Queries. *In CIKM 2008: Proceedings of the 17th International ACM Conference on Information and Knowledge Management* (pp. 1437 - 1438). Napa Valley, California, USA. October 26 - 30: ACM Press.

[11] Dong, A., Chang, Y., Zheng, Z., Mishne, G., Bai, J., Zhang, R., et al. (2010). Towards Recency Ranking in Web Search. In WSDM2010: In Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (pp. 11 - 20). New York, USA. February 3 - 6: ACM Press.

[12] Ferro, L., Gerber, L., Hitzeman, J., Lima, E., & Sundheim, B. (2005). ACE Time Normalization (TERN) 2004 English Training Data v 1.0. (L. D. Consortium, Ed.) Philadelphia, USA.

[13] Jatowt, A., Kawai, H., Kanazawa, K., Tanaka, K., & Kunieda, K. (2010). Analyzing Collective View of Future, Time-referenced Events on the Web. In WWW2010: Proceedings of the 19th International World Wide Web

Conference (pp. 1123 - 1124). Raleigh, USA. April 26 - 30: ACM Press.

[14] Jones, R., & Diaz, F. (2007). Temporal Profiles of Queries. *In TOIS: ACM Transactions on Information Systems , 25(3). Article No.: 14*.

[15] Kanhabua, N., & Nørvåg, K. (2010). Determining Time of Queries for Re-Ranking Search Results. *In ECDL2010: Proceedings of The European Conference on Research and Advanced Technology for Digital Libraries.* Glasgow, Scotland. September 6 - 10: Springer Berlin / Heidelberg.

[16] Li, X., & Croft, B. W. (2003). Time-Based Language Models. *In CIKM 2003: Proceedings of the 12th International ACM Conference on Information and Knowledge Management* (pp. 469 - 475). New Orleans, Louisiana, USA. November 2 - 8: ACM Press.

[17] Linguistic Data Consortium. (1992). *Linguistic Data Consortium*. Retrieved 12 23, 23, from Linguistic Data Consortium: http://www.ldc.upenn.edu/

[18] Linguistic Data Consortium. (n.d.). *Topic Detection and Tracking*. Retrieved 12 23, 2010, from Linguistic Data Consortium: http://projects.ldc.upenn.edu/TDT/

[19] Metzler, D., Jones, R., Peng, F., & Zhang, R. (2009). Improving Search Relevance for Implicitly Temporal Queries. *In SIGIR 2009: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 700 - 701). Boston, MA, USA. July 19 - 23: ACM Press.

[20] NIST. (2000, 08 01). *Text REtrieval Conference (TREC) Home Page*. Retrieved 08 07, 2009, from National Institute of Standards and Technologies: http://trec.nist.gov/

[21] Nunes, S., Ribeiro, C., & David, G. (2008). Use of Temporal Expressions in Web Search. In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, & R. White (Eds.), *Lecture Notes in Computer Science - Advances in Information Retrieval, ECIR 2008: European Conference on IR Research* (Vol. 4956/2008, pp. 580 - 584). Glasgow, Scotland. 30th March - 3rd April: Springer Berlin / Heidelberg.

[22] Pustejovsky, J., Verhagen, M., Sauri, R., Littman, J., Gaizauskas, R., Katz, G., et al. (2006). TimeBank 1.2. (L. D. Consortium, Ed.) Philadelphia, USA.

[23] *Reuters Corpora @ NIST*. (2004, 11 2). Retrieved 12 23, 2010, from Text Retrieval Conference: http://trec.nist.gov/data/reuters/reuters.html

[24] Sandhaus, E. (2008). *The New York Times Annotated Corpus*. Retrieved 12 23, 2010, from Linguistic Data Consortium: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T19

[25] Song, R., Luo, Z., Nie, J.-Y., Yu, Y., & Hon, H.-W. (2009). Identification of Ambiguous Queries in Web Search. *In Information Processing & Management: An International Journal , 45(2)*, 216 - 229.

[26] Tao, Q., Tie-Yan, L., Wenkui, D., Jun, X., & Hang, L. (2010, 06 16). *Microsoft Learning to Rank Datasets*. Retrieved 12 23, 2010, from Microsoft Research: http://research.microsoft.com/en-us/projects/mslr/default.aspx

[27] Vorhees, E., & Graff, D. (2008). *AQUAINT-2 Information-Retrieval Text Research Collection*. Retrieved 12 23, 2010, from Linguistic Data Consortium: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T25

[28] Zamir, O., & Etzioni, O. (1998). Web Document Clustering: A Feasibility Demonstration. In SIGIR 1998: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 46 - 54). Melbourne, Australia. August 24 - 28: ACM Press.

[29] Zhang, R., Chang, Y., Zheng, Z., Metzler, D., & Nie, J.-y. (2009). Search Result Re-ranking by Feedback Control Adjustment for Time-sensitive Query. In NAACL2009: Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies, (pp. 165 - 168). Boulder, Colorado, USA. May 31 - June 5.