Using Web Snippets and Query-logs to Measure Implicit Temporal Intents in Queries

Ricardo Campos
LIAAD – INESC Porto, LA
HULTIG, University of Beira Interior,
Polytechnic Institute of Tomar, Portugal
ricardo.campos@ipt.pt

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

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

ABSTRACT
Understanding the user's temporal intent by means of query formulation is a particular hard task that can become even more difficult if the user is not clear in his purpose. For example, a user who issues the query *Lady Gaga* may wish to find the official web site of this popular singer or other information such as informative or even rumor texts. But, he may also wish to explore biographic data, temporal information on discography release and expected tour dates. Finding this information, however, may prove to be particularly difficult, if the user does not specify the query in terms of temporal intent. Thus, having access to this data, will allow search mechanisms to improve search results especially for time-implicit queries. In this paper, we study different approaches to automatically determine the temporal nature of queries. On the one hand, we exploit web snippets, a content-related resource. On the other hand, we exploit Google and Yahoo! completion engines, which provide query-log resources. From these resources, we propose different measures to understand the temporal nature of queries. We compare these measures by analyzing their correlation. Finally, we conduct a user study to temporally label queries.

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

General Terms
Algorithms, Experimentation.

Keywords
Temporal Information Retrieval, Implicit Temporal Queries, Query Classification, Temporal Query Understanding.

1. INTRODUCTION
The temporal intent of queries may be explicit or implicit. Explicit temporal queries are the most obvious ones, carrying explicit temporal evidence stated by the user. Some examples are *SIGIR 2011*, *Iraq War 2003* or even future temporal queries such as *Football World Cup 2014*. Despite an apparent timeless nature, implicit temporal queries embody inherent temporal evidence. They consist of a set of keywords implicitly related to a particular time interval, which is not explicitly specified by the user. Some examples are *Tour de France, Miss Universe* or *Haiti earthquake*.

Understanding the temporal nature of a query, namely of implicit ones, is one of the most interesting challenges [3] in Temporal Information Retrieval (T-IR). However, few studies have attempted to answer questions like "How many queries have a temporal intent?", "How many of those are temporally ambiguous?" and "Do they belong to some prevalent category?". If we are able to answer to these questions, we may estimate how many queries will be influenced by a prospective temporal approach. A further automatic identification of temporal queries would enable to apply specific strategies to improve web search results retrieval. However, inferring this information is a hard challenge. First, different semantic concepts can be related to a query. An example is the query *Scorpions* that may be the rock band, the arachnid or the zodiac sign, each one with a different temporal meaning. Second, it is difficult to define the boundaries between what is temporal and what is not and so is the definition of temporal ambiguity. Third, even if temporal intents can be inferred by human annotators, the question is how to transpose this to an automatic process. One possible solution to date time-implicit queries is to seek for related temporal references over complementary web resources, such as document collections (e.g., web pages, web snippets, news articles, web archives) or web usage data (e.g., web query logs). In this paper we first summarize the three types of temporal queries. We then propose two different studies for the classification of implicit temporal queries based on the temporal value of web snippets and the temporal value of web query logs.

2. TYPES OF TEMPORAL QUERIES
Unlike explicit temporal queries, where the temporal nature is clearly defined by the user, implicit temporal queries can have a variety of inherent temporal nature. Similarly to [8], we classify time-implicit queries into one of the three following categories:

Type A - ATemporal: those not sensitive to time, i.e., queries not temporally related, like for instance *make my trip*.

Type B - Temporal Unambiguous: queries that take place in a very concrete time period like *Haiti Earthquake* which occurred in 2010.

Type C - Temporal Ambiguous: queries with multiple instances over time, such as *Football World Cup*, which occurs every four years. But also, queries such as *Pascal*, in the sense of the philosopher, where the user can be interested in different time segments of his life (e.g. birth date, date of death).

3. IMPLICIT TEMPORAL QUERIES
The extraction of temporal information is usually based on a metadata-based approach upon time-tagged controlled collections, such as news articles, which are informative texts typically annotated with a timestamp [6] [8] [9] [10]. This information can be particularly useful to date relative temporal expressions found in a document (e.g., *today*) with a concrete date (e.g., document creation time). However, it can be a tricky process if used to date implicit temporal queries as referred in [12]. Indeed, the time of the document can differ significantly from the actual content of the document. An example is a document published in *2009* but which contents concern *2011*. For instance, if a document is

related to the time-implicit query *Miss Universe* (see Figure 1), then there exists a high probability to associate the document temporal information to the document timestamp i.e. *2009* instead of *2011*.

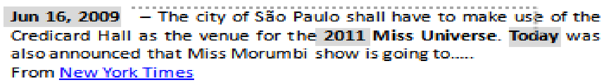
Jun 16, 2009 – The city of São Paulo shall have to make use of the Credicard Hall as the venue for the 2011 Miss Universe. Today was also announced that Miss Morumbi show is going to.....
From New York Times

Figure 1: Use of Timestamp to date Miss Universe Query.

In this case, a content-based approach, which would extract temporal information from the content of the document, would obviously be the most appropriate solution to determine whether a query is temporally implicit or not. Another approach is to timestamp the queries based on similar year-qualified queries (e.g., *Miss Universe 2009*, *Miss Universe 2010*) stored in web query logs. Both differ in how they deal with query-dependency: while a web content approach simply requires the set of web search results, an approach based on web query logs implies that some versions of the query have already been issued.

3.1 Web Snippets

One of the most interesting approaches to date implicit temporal queries is to rely on the exploration of temporal evidence within web pages. As claimed in [3], this is an interesting future research direction, for which there is not yet a clear solution [1] [2]. Our main purpose in this section is to check if web snippets can be used to date time-implicit queries based on the temporal information existing in the title, in the text (generally known as snippet) and in the link (URL) of the web snippet. To this end, we conducted an experiment [5] where we studied the temporal value of web snippets[1]. We executed a set of 450 implicit queries extracted from Google Insights for Search Collection[2], which registers the hottest and rising queries performed worldwide, defining a retrieval of 200 results per query. For each of the queries we computed three measures, *TSnippets(.)*, *TTitle(.)* and *TUrl(.)*, which assess how strong a query is temporally related. All represent the ratio between the number of respective items retrieved with dates divided by the total number of retrieved results. Results show that on average 10% of the web snippets retrieved for a given implicit query, contain year dates, of which 23% have more than one date. In this paper, we introduce a new insight to this approach. We propose to calculate the Pearson correlation coefficient between each of the dimensions (*TSnippets(.)*, *TTitle(.)* and *TUrl(.)*) and conclude that the correlation between web snippets and titles is very high (0.83), meaning that there is a strong correlation between the occurrence of dates within both dimensions (see Figure 2).

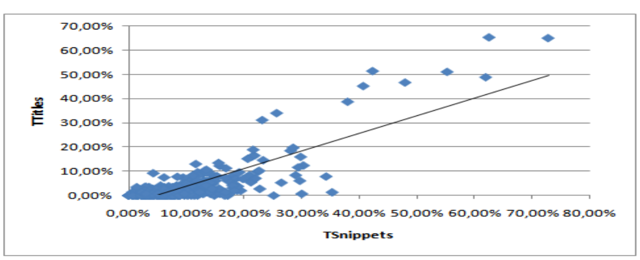

Figure 2: Snippets vs Titles Scatter Plot.

[1] Available at http://www.ccc.ipt.pt/~ricardo/software [17th June, 2011]
[2] http://www.google.com/insights/search [17th June, 2011]

Our next step is to manually classify each query with regard to its temporal intent based on the values of *TSnippets(.)*, *TTitle(.)* and *TUrl(.)*. We first start by classifying the query in accordance to its concept ambiguity following the approach of [13] who defines three types of concept queries: ambiguous, broad and clear. Results in Table I show that most of the queries are ambiguous in concept, followed very closely by clear queries, which do not offer any doubt in terms of their meaning and by a small set of broad queries.

Table I: Concept and Temporal Classification of Queries.

Conceptual Classification	Number Queries	Temporal Classification	Number Queries	%
Ambiguous	220			
Clear	176	ATemporal	133	75%
		Temporal	43	25%
Broad	54			

Second, we aim at classifying the queries with regard to its temporal value. However, given that each concept of a query can have a different temporal dimension, we only focus on the temporal classification of clear concept queries. To this end, we computed, for each of the 176 clear queries (e.g., *Toyota recall*, *lady gaga*, *Dacia duster*, *hairstyles*), a temporal ambiguity value. Given the fact that dates occur in different proportions in the three items i.e. titles, snippets and urls, we value each feature differently through ω_f (18.14% for *TTitles(.)*, 50.91% for *TSnippets(.)* and 30.95% for *TUrl(.)*), where f is the function regarding the corresponding item based on equation (1).

$$TA(q) = \sum_{f \in I} \omega_f \, . \, f(q) \,, \; I = \{TSnippet(.), TTitle(.), TUrl(.)\} \quad (1)$$

With this measure we can define a simple model for the automatic temporal classification of queries. A query is ATemporal if *TA(.)* is below 10%, otherwise the query is defined as Temporal. Our experiments, strictly depending on the value of the year dates found in the web snippets, show that of all clear concept queries, 25% have implicit temporal intent. The remaining 75% are ATemporal queries. Moreover, we showed that year dates occur more frequently in response to queries belonging to the categories of sports, automotive, society and politics.

In order to evaluate our simple classification model, we conducted a user study. Using the same 176 clear concept queries, we asked three human annotators to judge their temporality. Human annotators were asked to consider each query, to look at web search results and to classify them as ATemporal or Temporal. Judgments were made assuming that the human annotator did not have any kind of specific temporal purpose in the execution of the query. The basic idea was to check if human annotations were correlated to our simple classification methodology just by looking at the set of web search results, even if there was a total absence of temporal intent. An inter-rater reliability analysis using the Fleiss Kappa statistics [7] was performed to determine consistency among annotators. Results showed a value of 0.89, meaning an almost perfect agreement between the raters. Overall, results pointed at 35% of implicit temporal queries from human annotators, while only 25% were given by our methodology. Then, we used the same test (Fleiss Kappa) to determine the consistency among each of individual annotators and the system decision. Results showed an average Kappa of 0.40 with a percentage of overall agreement of approximately 75%. A more thorough analysis also showed that most of the differences in the automatic classification of the system and the human annotator rating were related to electronic device queries, which introduce noise in the automatic definition of year dates (e.g., *Nikon d3000*). Overall, we may conclude that the occurrence of year dates in web snippets by itself is not sufficient to temporally classify these kinds of specific queries and that complementary information, such as the number of instances or the number of different dates should be considered in future approaches. Furthermore, this information can also be very useful in order to improve the results set returned to the user, either by adjusting the score of a document in response to an implicit temporal query, or by proposing alternative temporal search results exploration, through timelines or temporal clusters. We propose to do this by clustering the *k* top web snippets retrieved from the execution of any given query in [4]. The use of web documents to date queries not entailing any temporal information can be however a tricky process. The main problem is related to the difficulties underlying the association of the year date found in the document and the query. An elucidative example is shown in Figure 3 where the dates appearing in the text and in the URL are not related with the query *Miss Universe*.

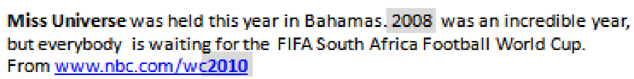
Miss Universe was held this year in Bahamas. 2008 was an incredible year, but everybody is waiting for the FIFA South Africa Football World Cup. From www.nbc.com/wc**2010**

Figure 3: Use of web contents to date Miss Universe Query.

3.2 Web Query Logs

Another approach to date implicit temporal queries is to use web query logs. The temporal activity of a search engine can be recorded in two different ways: (1) from an infrastructural perspective, i.e., date and time of the request and (2) from the user activity dimension, i.e., user search query such as *Football World Cup 2010*. This can be seen as a Web Usage Mining approach. This latter information can particularly be useful to infer temporal intents in queries not containing a specific year such as *Tour de France*, based on similar year-qualified queries such as *Tour de France 2010*. However, one of the problems of this approach lies in the fact that the number of year-qualified queries is too reduced. Indeed in [5], we showed that explicit temporal queries represent 1.21% of the overall set of a well-known query log collection (AOL Collection). Most of them belong to the categories automotive (21.96%), entertainment (9.48%) and Sports (8.15%). An additional problem is that we may have to deal with queries that have never been typed e.g. *Blaise Pascal 1623* (his birthday year date). This tends to get even worse in specific applications that lack massive user interaction. Another problem is that query logs are hard to access outside big industrial laboratories. Moreover, web query logs are not adapted to concept disambiguation. For most part of the queries this is a real problem that would result in inconclusive information to the user. Consider for example the query *Euro* and suppose a web query log that has some related explicit temporal queries like *Euro 2008* or *Euro 2012*. Having access to this information makes it possible to assign the query with the dates 2008 or with the future date 2014. Yet, this information is insufficient to disambiguate the query in its several meanings, causing the system to date the query *Euro* with temporal information regarding the European football world cup, when in fact it could be related to the European currency. One of the first works in this line was proposed by [11] who presented an interesting approach based on the access to a query log with frequency information. In order to temporally qualify a query, the authors introduced a weighted measure that considers the number of times a query q is pre- and post-qualified with a given year *y* as shown in Equation 2.

$$WA(q,y) = \#(q.y) + \#(y.q) \qquad (2)$$

A query is then considered implicitly year-qualified if it is qualified by at least two different years. Moreover, the authors introduce a confidence measure to confirm that the query has an implicit temporal intent. This value is defined in Equation 3 where the sums on the denominator of the equation go all over pre- and post-qualifications of the query q.

$$\alpha(q) = \frac{\sum_y WA(q,y)}{\sum_x \#(q.x) + \sum_x \#(x.q)} \qquad (3)$$

If the query is always qualified with a single year then $\alpha(q) = 1$. Overall results show that 7% of the queries have implicit temporal intent. These values contrast with the 25% that we present in our study based on web content analysis.

3.2.1 Yahoo! and Google Temporal Value

In this section, we aim at quantifying the temporal value of a query in Yahoo! and Google web logs and compare it with web snippets by means of a Pearson correlation coefficient with a confidence interval for paired samples. We already showed that the number of explicit temporal queries existing in web logs can be very small but we must also take into account that the simple fact that a query is year-qualified does not necessarily mean that it has a temporal intent. An illustrative example is the query *make my trip* which is substantially more qualified with words than with temporal features. In order to measure this value, we introduce two measures *TLogYahoo(.)* and *TLogGoogle(.)* similarly to *TTitle(.)*, *TSnippets(.)* and *TUrl(.)*. To compute these values, we rely on Google and Yahoo! auto-complete query search feature, which constitutes a powerful mechanism to understand the temporality of a given query based on the fact that it displays user queries supported by real user search activities. *TLogGoogle(.)* and *TLogYahoo(.)* are defined in Equation 4 and 5 respectively as the ratio between the number of queries suggested with year dates divided by the total number of suggested retrieved queries.

$$TLogGoogle(q) = \frac{\#\, Suggested\ Queries\ Retrieved\ With\ Dates}{\#\, Suggested\ Queries\ Retrieved} \qquad (4)$$

$$TLogYahoo(q) = \frac{\#\, Suggested\ Queries\ Retrieved\ With\ Dates}{\#\, Suggested\ Queries\ Retrieved} \qquad (5)$$

For example, if we type in the query *bp oil spill*, both Yahoo! and Google search engines suggest a set of 10 queries, of which only one, in this case in Yahoo! search interface (see left hand part of Figure 4) includes a year date. This means that for the query *bp oil spill TLogYahoo(.)* would be 10% and *TLogGoogle(.)* 0% (see right hand part of Figure 4).

Figure 4: Auto-complete Query Suggestion for the query Bp Oil Spill in Yahoo! (on the left) and Google (on the right).

Based on the values obtained for each of the 176 clear concept queries, we calculated the Pearson correlation coefficient (see Table II) to compare the temporal value of web snippets by means of *TSnippets(.)*, *TTitle(.)* and *TUrl(.)* values with the temporal value of web logs by means of *TLogGoogle(.)* and *TLogYahoo(.)*.

Table II: Pearson Correlation Coefficient.

	TLogGoogle	TTitle	TSnippet	TUrl
TLogYahoo	0.63	0.61	0.52	0.48
TLogGoogle		0.69	0.63	0.44

Final results show that best correlation values occur between *TTitle(.)* and *TLogGoogle(.)* with a value of 0.69 and between *TSnippet(.)* and *TLogGoogle(.)* with 0.63. This means that as dates appear in the titles and snippets, they also tend to appear, albeit in a more reduced form, in the auto-complete query suggestion of Google.

An additional analysis led us to conclude that the temporal information is more frequent in web snippets than in any of the query logs of Google and Yahoo! (see Figure 5 and Figure 6). Overall, while most of the queries have a *TSnippet(.)* value around 20%, *TLogYahoo(.)* and *TLogGoogle(.)* are mostly near to 0%.

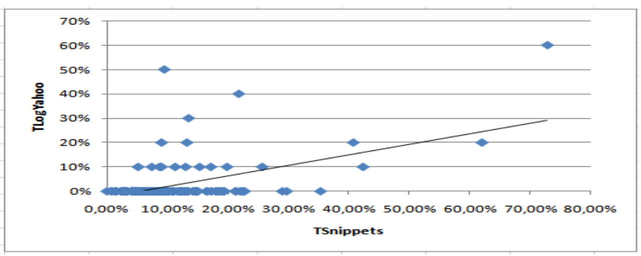

Figure 5: *Tsnippets(.)* vs *TLogYahoo(.)* Scatter Plot.

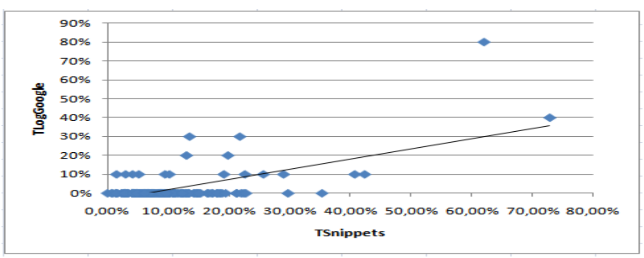

Figure 6: *Tsnippets(.)* vs *TLogGoogle(.)* Scatter Plot.

Finally, we also studied, both in web snippets and in web query logs, how strongly a given query is associated to a set of different dates. For that purpose, we built a confidence interval for the difference of means, for paired samples, between the number of times that the dates appear in the web snippets and in the web query logs. We obtained the intervals [5.10; 6.38] and [5.12; 6.43] for *TLogYahoo(.)* and *TLogGoogle(.)* with 95% confidence. These results show that the number of different dates that appear in web snippets is significantly higher than in either one of the two web query logs.

4. CONCLUSION

In this paper, we showed that web snippets are a very rich data resource, where dates, especially years in snippets and titles, often appear correlated. Overall, 10% of the web snippets retrieved for a query with no stated temporal information has dates. Moreover, we conducted a user study that estimates that 35% of the queries are temporal in their intent. Our experiments showed that temporal query understanding is possible based on web snippets search results. However, the simple occurrence of dates may not be entirely sufficient for some types of queries. This can be improved in future work by considering complementary temporal information, such as the number of instances or the number of different dates. We also showed that the explicit temporal value of web query logs is nearly 1.5%. This value appears to be very small. Moreover, the simple fact that a query is year-qualified does not necessarily mean that it has a temporal intent. For that purpose, we compared web snippets collections to Google and Yahoo! auto-complete query suggestions (i.e. web query logs). We concluded that web snippets are statistically more relevant in terms of temporal intent than web query logs.

5. ACKNOWLEDGMENTS

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

6. REFERENCES

[1] Alonso, O., & Gertz, M. (2006). Clustering of Search Results using Temporal Attributes. In SIGIR 2006 (pp. 597 - 598). August 6 - 11.

[2] Alonso, O., Gertz, M., & Baeza-Yates, R. (2009). Clustering and Exploring Search Results using Timeline Constructions. In CIKM 2009. November 2 - 6: ACM Press.

[3] Berberich, K., Bedathur, S., Alonso, O., & Weikum, G. (2010). A Language Modeling Approach for Temporal Information Needs. In ECIR 2010. UK. March 28 - 31: Springer Berlin / Heidelberg.

[4] Campos, R. (2011). Using k-top Retrieved Web Snippets to Date Temporal Implicit Queries based on Web Content Analysis. In SIGIR 2011. Beijing, China. July 24 - 28.: ACM Press.

[5] Campos, R., Dias, G., & Jorge, A. M. (2011). What is the Temporal Value of Web Snippets? In TWAW 2011 associated to WWW2011. Hyderabad, India. March 28.: CEUR Workshop Proceedings.

[6] Dakka, W., Gravano, L., & Ipeirotis, P. G. (2008). Answering General Time Sensitive Queries. In CIKM 2008 (pp. 1437 - 1438). California, USA. October 26 - 30: ACM Press.

[7] Fleiss, J. L. (1971). Measuring Nominal Scale Agreement Among many Raters. *In Psychological Bulletin , 76(5)*, 378 – 382.

[8] Jones, R., & Diaz, F. (2007). Temporal Profiles of Queries. In TOIS, 25(3). Article No.: 14.

[9] Kanhabua, N., & Nørvåg, K. (2010). Determining Time of Queries for Re-Ranking Search Results. In ECDL2010. September 6 - 10.

[10] Li, X., & Croft, B. W. (2003). Time-Based Language Models. In CIKM 2003. (pp. 469 - 475). November 2 - 8: ACM Press.

[11] Metzler, D., Jones, R., Peng, F., & Zhang, R. (2009). Improving Search Relevance for Implicitly Temporal Queries. In SIGIR 2009. (pp. 700 - 701). Boston, MA, USA. July 19 - 23: ACM Press.

[12] Nunes, S., Ribeiro, C., & David, G. (2007). Using Neighbors to Date Web Documents. In WIDM2007 associated to CIKM2007. (pp. 129 - 136). Lisboa, Portugal. November 9: ACM Press.

[13] Song, R., Luo, Z., Nie, J.-Y., Yu, Y., & Hon, H.-W. (2009). Identification of Ambiguous Queries in Web Search. In IPM, 45(2), 216 - 229.