Disambiguating Implicit Temporal Queries by Clustering Top Relevant Dates in Web Snippets

Ricardo Campos[1,2,6], Alípio Mário Jorge[1,3], Gaël Dias[4,6], Célia Nunes [5,6]

[1] LIAAD – INESC TEC
[2] Polytechnic Institute of Tomar, Portugal
[3] DCC – FCUP, University of Porto, Portugal
[4] HULTECH/GREYC, University of Caen Basse-Normandie, France
[5] Department of Mathematics, University of Beira Interior, Covilhã, Portugal
[6] Center of Mathematics, University of Beira Interior, Covilhã, Portugal

ricardo.campos@ipt.pt, amjorge@fc.up.pt, gael.dias@unicaen.fr, celian@ubi.pt

ABSTRACT

With the growing popularity of research in Temporal Information Retrieval (T-IR), a large amount of temporal data is ready to be exploited. The ability to exploit this information can be potentially useful for several tasks. For example, when querying "*Football World Cup Germany*", it would be interesting to have two separate clusters $\{1974, 2006\}$ corresponding to each of the two temporal instances. However, clustering of search results by time is a non-trivial task that involves determining the most relevant dates associated to a query. In this paper, we propose a first approach to flat temporal clustering of search results. We rely on a second order co-occurrence similarity measure approach which first identifies top relevant dates. Documents are grouped at the year level, forming the temporal instances of the query. Experimental tests were performed using real-world text queries. We used several measures for evaluating the performance of the system and compared our approach with Carrot Web-snippet clustering engine. Both experiments were complemented with a user survey.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Query Formulation*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software – *Performance evaluation*

Keywords

Temporal Information Retrieval, Temporal Clustering, Dating Implicit Temporal Queries, Temporal Query Understanding.

1. INTRODUCTION

With so much information available on the Web, clustering of search results appears as a valid alternative to help users in their process of seeking information. One of the advantages of this alternative interface is to offer users with a quick overview of a topic, without the need to go through an extensive list of results. Within this context, Web-snippet clustering appears as an interesting approach to group similar results on the basis of the retrieved result set. The resulting data is a set of flat or hierarchical clusters generated on the fly, which can be instantly used for interactive browsing purposes. Over the years some clustering engines have been proposed, including *iBoogie*[1], *Yippy*[2] and *Carrot*[3]. While all these systems present a large number of topic clusters, they seldom include a temporal feature as part of the cluster description as we will show in this paper. The lack of such a time-oriented analysis, makes it difficult for clustering search engines to return results with a temporal perspective. Moreover, it prevents users to be aware of the possible temporal structure of a given topic.

In this paper, we focus on disambiguating a text query with respect to its temporal purpose and propose an approach that temporally clusters the results of a text query. Whereas most of the temporal queries issued by users are implicit by nature [4], detecting its underlying temporal intent may thus be particularly useful for a large number of tasks, such as user query understanding or temporal clustering.

In essence, our method is a two stage process, combining the identification of relevant temporal expressions extracted from Web snippets with a clustering methodology, where documents are grouped into the same cluster if they share a common year. The resulting clusters directly reflect groups of individual years that consistently show a high connectivity to the text query.

The advantage of our approach is that instead of considering all the temporal expressions as equally relevant, as currently common in most of the T-IR tasks, we determine which ones are more relevant to the user text query. One consequence of this, is a direct impact on the quality of the retrieved clusters, as non-relevant or wrong dates will be discarded. In summary, the main contributions of this work are: (1) we present a temporal document representation model, based on the most important topics of the Web snippets and their respective dates; (2) we introduce a novel approach to identify temporal expressions relevant to a particular query by relying on a content-based approach and a language-independent methodology; (3) we propose a soft flat overlapping temporal clustering algorithm, where documents are highly related if they share a common year; (4) we publicly provide a set of queries and ground-truth results to the research community, hence our evaluation results can be compared to future approaches; (5) we present an evaluation of our approach using several performance measures and a comparison against a well known Web-snippet clustering engine; and (6) we conduct an user study to validate our approach.

The remainder of this paper is structured as follows. In section 2 we present the related work. In section 3 we introduce our algorithm. Experimental setups and results are discussed in Section 4. Finally, we conclude this paper in Section 5 with some final remarks and future research directions.

2. RELATED WORK

Temporal clustering is a relatively new subfield of T-IR. However, despite its importance, little work has been carried out

[1] http://www.iboogie.com [7th September, 2012]
[2] http://search.yippy.com/ [7th September, 2012]
[3] http://search.carrot2.org/stable/search [7th September, 2012]

978-0-7695-4880-7/12 $26.00 © 2012 IEEE
DOI 10.1109/WI-IAT.2012.158

over the years. Within the context of detecting and tracking events by time, Mori et. al. [13], and Shaparenko et. al. [14] were the first to consider temporal clusters. In another line of work, Jatowt et. al. [9] proposed a clustering approach to summarize future-related information and a model-based clustering algorithm [8] for detecting future events based on information extracted from a text corpus. The task of clustering search results by time, which is the focus of our research, was first introduced by Alonso et.al [1][2]. In his first work [1] the authors assume two different clustering views: *topics* and *time*. Clustering by topics is based on traditional clustering approaches, supported on features extracted from the title and the text snippet, whereas clustering by time relies on temporal attributes extracted from the metadata of the document and from its contents. This paper, was later extended [2] by introducing a clustering algorithm called *TCluster*, where each cluster is formed by a set of documents sharing a temporal expression. The organization of the clusters along a timeline $T = \{T_d, T_w, T_m, T_y\}$, allows the exploration of the documents at different levels of granularity, namely days, weeks, months and years. Unfortunately, none of these works measure whether the temporal expressions found are indeed relevant or query-related. The closer to this respect, is proposed by Alonso et. al [2] however for purposes of document ranking in clusters. As such, in a perspective of the document and not of the relevancy of the date. The lack of a solution that ensures an effective relationship between the text query and the dates found in the documents, causes both systems to be highly dependent on the ability of the temporal tagger to determine correct temporal expressions, which may thus compromise the quality of the clusters.

In this paper we focus on adding top relevant temporal features to post-retrieval clustering based on a web content analysis that extracts dates within Web contents given a particular text query. This method involves the formation of clusters showing a high connectivity to snippets sharing a common year, based on a second-order co-occurrence measure that filters out irrelevant dates. To the best of our knowledge only three works [10] [12] [16] have been proposed with regard to the identification of top relevant expressions given a user implicit temporal query. In detail, Metzler et. al. [12] mine query logs to identify implicit temporal information needs. Kawai et. al [10], on the other hand, developed a chronological events search engine for the Japanese language based on Web snippets analysis, where noisy temporal patterns are removed through machine learning techniques trained over a set of text features. Finally, Strötgen et. al. [16] extend this idea by proposing an enriched temporal document profile for each document, where each temporal expression found is represented by a larger number of different features. Our approach differs from previous works in relevant temporal expressions identification [10] [12] [16] in several aspects. First, we do not make use of query logs, nor train a classifier. We also do not use a set of heuristics extracted from the contents of a document. Instead, in our approach, relevant temporal expressions are detected based on corpus statistics and a general similarity measure that makes use of co-occurrences of words and years extracted from the contents of the Web snippets. Second, our methodology is language-independent as we do not use any linguistic-based techniques. Instead, we use a rule-based model solution supported by language-independent regular expressions. Third, besides estimating the degree of relevance of a temporal expression, we present an appropriate threshold-based classification strategy to determine whether a date is or is not relevant to the query. Finally, instead of suggesting a ranking of the documents or a timeline-based visualization of the temporal expressions, we propose to apply a clustering algorithm in the scope of temporal clustering. While we already achieved an initial stage of flat clustering by time, our proposal still lacks an approach focused on topics. These should be addressed in future work as part of a global project.

3. THE ALGORITHM

In this section, we describe our method of disambiguating text queries with temporal dimensions. We divide this method into the following five subtasks: (1) Web search, (2) Web snippet processing, (3) Query-Date relevance identification, (4) Relevant date classification and (5) Temporal clustering of Web snippets. Each one will be described in the following sections.

3.1 Web Search

In what follows, we assume a query to be either explicit, that is, a combination of both text and time, denoted q_{time}, or implicit, i.e., just text, denoted q_{text}. In this paper, we deal with the latter ones since handling explicit temporal queries is a trivial process. For purposes of better readability, we denote a query simply as q. Similarly to Kawai et. al [10], we use a Web search API to access an up-to-date index search engine. Given a text query q, we obtain a collection of n Web snippets $S = \{S_1, S_2, \dots, S_n\}$.

3.2 Web Snippet Representation

Each S_i, for $i = 1, \dots, n$, denotes the concatenation of two texts, i.e. $\{Title_i, Snippet_i\}$ and is represented by a bag-of-relevant-words and a set of candidate temporal expressions. In what follows, we assume S_i to be represented by two different sets denoted W_{S_i} and D_{S_i}

$$S_i \rightarrow (W_{S_i}, D_{S_i}) \tag{1}$$

where

$$W_{S_i} = \{w_{1,i}, w_{2,i}, \dots, w_{k,i}\} \tag{2}$$

is the set of the k most relevant words/multi-words associated with a Web snippet S_i, and

$$D_{S_i} = \{d_{1,i}, d_{2,i}, \dots, d_{t,i}\} \tag{3}$$

is the set of the t candidate years associated to a Web snippet S_i. Moreover,

$$W_S = \bigcup_{i=1}^{n} W_{S_i} \tag{4}$$

defines the set of distinct relevant words extracted for a query q, within the set of Web snippets, S i.e. the relevant vocabulary. Similarly,

$$D_S = \bigcup_{i=1}^{n} D_{S_i} \tag{5}$$

is defined as the set of distinct candidate years extracted from the set of all Web snippets S. In this work, relevant words are identified using the methodology proposed by Machado et. al. [11], who define a numeric heuristic based on word left and right contexts distribution analysis. This metric is specifically tuned towards the tokenization process of Web snippets in order to overcome the problems faced by usual tokenizers, sentence splitters or part-of-speech taggers, which due to the specific structure of Web snippets, fail to correctly process this type of collection. Moreover, Machado et. al. [11] show that standard collocation extraction strategies also fail compared to longest frequent substrings identification. As a consequence, multiword

unit identification is done as in Zamir et. al. [17]. Due to space limitations, we do not detail this pre-processing step as it can easily be reproduced from [11] and [17], and it is commonly used in Web snippet processing. Furthermore, a simple rule-based model supported on regular expressions is used to extract explicit temporal dates satisfying certain specific explicit patterns (e.g., *yyyy*, *yyyy-yyyy*, *yyyy/yyyy*, *mm/dd/yyyy*, *mm.dd.yyyy*, *dd/mm/yyyy* and *dd.mm/yyyy*). Although it is possible to extract temporal expressions with finer granularities, such as month and day, we are particularly interested in working at the year granularity level in order to keep language-independency and allow longer timelines for visualization. As such, all the temporal expressions detected according to the aforementioned patterns end up normalized to the year granularity level. An example is given as follows: $norm(29/10/2012) = 2012$.

Finally,

$$W^* = \mathrm{W_S} \cap \mathrm{W_{S_{d_i}}} \tag{6}$$

is defined as the set of distinct words that result from the intersection between the set of words $\mathrm{W_S}$ and the set $\mathrm{W_{S_{d_i}}}$ which contains the words that co-occur with date d_i, in any Web snippet, S_i, from S.

3.3 Temporal Similarity Measure

In this section we introduce our temporal similarity measure *GenTempEval* (*GTE*) with the purpose of identifying top relevant dates. Given a query q and a date $d_i \in D_S$ we can measure their relatedness $GTE(q, d_i)$ by using the scoring function defined in Equation 7. This score is computed by *GTE* taking into account the co-occurrence of the date d_i with respect to each word $x \in W^*$, under the following principle.

P1: The more closely a given date is correlated to the set of corresponding distinct most relevant words associated to the query (i.e., the result of the intersection between the set of words co-occurring with the query and the set of words co-occurring with the date), the more closely the query will be associated to the date.

GTE is defined in Equation 7, where sim is a similarity measure and F an aggregation function of the several $sim(W^*, d_i)$ that combines the different similarity values produced for the date d_i in a single value capable of representing its relevance:

$$GenTempEval(q, d_i) = F(sim(W^*, d_i)). \tag{7}$$

A wide range of combinations with different $F's$ and $sim's$ have been proposed in [3]. In this paper we assume that F is the Median function and sim is *InfoSimba* (IS) [6] a semantic vector space model supported by corpus-based word correlations (see Equation 8).

$$IS(V_x, V_y) = \frac{\sum_{i \in V_x} \sum_{j \in V_y} S(i,j)}{\left(\sum_{i \in V_x} \sum_{j \in V_x} S(i,j) + \sum_{i \in V_y} \sum_{j \in V_y} S(i,j) - \sum_{i \in V_x} \sum_{j \in V_y} S(i,j)\right)} \tag{8}$$

In detail, *IS* calculates the correlation between all pairs of two context vectors V_x and V_y. Without loss of generality, V_x and V_y can be seen as the context vector representations of each of the two items of a (x, d_i) pair, respectively. Each vector is represented by a combination of words and dates. The similarity between each pair is determined by any first order similarity measure $S(.,.)$ relating items i and j. In this paper, we use the DICE coefficient since it has shown better results compared to other measures [3]. Each of these similarity values is stored on a global conceptual temporal correlation matrix denoted M_{ct}. In detail:

$$M_{ct} = \begin{bmatrix} A_{k \times k} & B_{k \times t} \\ B^T_{t \times k} & C_{t \times t} \end{bmatrix}_{(k+t) \times (k+t)} \tag{3}$$

where $[A]_{k \times k}$ is the $k \times k$ matrix which represents the similarity between k words, $C_{t \times t}$ is the $t \times t$ matrix which represents the similarity between t candidate dates, $B_{k \times t}$ is the $k \times t$ matrix which represents the similarity between k words and t candidate dates, and $B^T_{t \times k}$ is the transpose of this matrix. A more thorough discussion of this issue, along with many more experiments, can be found in [3].

3.4 Relevant Date Classification

In order to determine whether a date is or is not relevant we use a classical threshold-based strategy, where a date is considered to be: (1) relevant, if $GTE(q, d_i) \geq \lambda$, and (2) irrelevant or wrong date, if $GTE(q, d_i) < \lambda$.

The final set of m relevant dates for the query q is derived from the decomposition of D_S into D_S^{Rel}, as follows:

$$D_S^{Rel} = \{d_1^{Rel}, d_2^{Rel}, \dots, d_m^{Rel}\}, \tag{9}$$

where $d_1^{Rel} < d_2^{Rel} < \cdots < d_m^{Rel}$.

Note that d_1^{Rel} and d_m^{Rel} represent the lower and the upper temporal bound of the query q respectively. Similarly D_{S_i} is decomposed into

$$D_{S_i}^{Rel} = \{d_{1,i}^{Rel}, d_{2,i}^{Rel}, \dots, d_{u,i}^{Rel}\}, \tag{10}$$

meaning the set of u relevant dates d_i for the query q associated to the Web snippet S_i.

3.5 Web Snippets Clustering

The first step of our clustering approach is to choose an appropriate measure that calculates the similarity between each of the snippets. In this work, instead of using a usual similarity measure, we cluster each snippet according to its associated years. Our assumption, defined in Principle 2, is that:

P2: Two snippets are temporally similar if they are highly related to the same set of dates.

One of the advantages of our clustering model is that instead of considering all the temporal expressions as equally relevant, we determine which ones are more relevant to the user text query. Based on this, each snippet S_i is no longer represented by a set of candidate temporal expressions, but by a set of relevant temporal ones. We redefine S_i as follows:

$$S_i \rightarrow (W_{S_i}, D_{S_i}^{Rel}) \tag{11}$$

where

$$W_{S_i} = \{w_{1,i}, w_{2,i}, \dots, w_{k,i}\} \tag{12}$$

remains as the set of the k most relevant-words associated to the snippet S_i and

$$D_{S_i}^{Rel} = \{d_{1,i}^{Rel}, d_{2,i}^{Rel}, \dots, d_{u,i}^{Rel}\} \tag{13}$$

is the set of u relevant dates that replace the prior set D_{S_i} consisting of candidate temporal expressions.

Each Web snippet S_i can be assigned to possible many clusters $C = \{C_1, C_2, \dots, C_m\}$ since its text can contain several different

$D_{S_i}^{Rel}$ relevant temporal features. According to Campos et. al [4] 23% of the Web snippets have on average more than one date[4]. A single cluster C_j, for $j = 1, \dots, m$ can be seen as a container including documents sharing the same year. The final set of clusters is ranked on the basis of the timeline, and consists of m clusters, where m is the number of relevant dates found within D_S^{Rel}. As such, each C_j cluster is labeled directly by D_S^{Rel}. Note that one of the challenges in our problem is to disambiguate the query through a balanced number of clusters, so that the search for information through the list of results is not replaced for a search within the set of clusters. This will be denoted as requirement 1 (**R1)**.

The final step of our clustering algorithm is to rank the documents inside each cluster C_j. For this purpose, we use a soft clustering strategy that estimates a level of membership for each snippet S_i found within each cluster C_j. In order to leverage all the information we have, documents are ranked to reflect the relevance of the snippet $S_i$within the cluster C_j according to the query q, both in the conceptual and in the temporal dimensions. This membership is provided by the value $rank(S_i, C_j)$ computed by *GTE* and *IS* as follows:

$$rank(S_i, C_j) = \alpha * \sum_{l=1}^{u} GTE(q, d_{l,i}^{Rel}) + (1-\alpha) * \sum_{h=1}^{k} IS(q, w_{h,i}), \quad \alpha \in [0,1] \qquad (14)$$

Central to this ranking is the similarity computed by *GTE* between the query and each of the relevant dates found in the snippet and also the similarity between the query and each of the relevant concepts found in the snippet, computed by *IS*. Without loss of generality, q and $w_{h,i}$ (arguments of *IS*) can be seen as the context vector representation for each of the two items defined in (8) as V_x and V_y. These are formed by the set of the best relevant words and dates related to q and $w_{h,i}$, respectively. Similarly to what happens in our temporal similarity measure, *IS* is combined with the DICE coefficient.

In the following we formalize two requirements that the ranking function should fulfill:

R2: S_i is more relevant for C_i than S_i', if $rank(S_i, C_j) > rank(S_i', C_j)$.

R3: S_i is more relevant for C_j than for C_j', if $rank(S_i, C_j) > rank(S_i, C_j')$.

In the next section we will experimentally evaluate our approach.

4. EVALUATION AND RESULTS

The experimental validation of our approach is twofold. First, we aim to evaluate the ability of our clustering algorithm to correctly identify relevant temporal clusters C_j and snippets S_i for the query q. Second, we aim to compare our clustering proposal with current Web-snippet clustering engines. To these respect we have used the clustering engine *Carrot*, since both *iBoogie* and *Yippy* do not allow tests over an external data source. In order to test our approach, we rely on our own, publicly available, dataset [5], named *WC_DS*. The fact that there is no standard text collection for temporal clustering purposes lead us to develop a ground truth collection. *WC_DS* consists of 42 text queries (see Table 6 or Table 7) selected from the 27 categories of Google Insights for Search[5] 2010 and 2011 Webpage trends, after removing duplicates, atemporal queries and queries with multiple meanings. Each query was issued in Bing[6] search engine on December 2011, collecting the top best 50 relevant web results, using for this purpose the Bing Web search API, parameterized with the *en-US* market language parameter. Of the 2100 web snippets retrieved, only those annotated with at least one candidate year term were selected. The final set consists of 582 web snippets.

The ground truth was then obtained over this dataset by conducting two relevance human judgments: (1) evaluation of the quality of the snippets with respect to the cluster label, and (2) evaluation of the quality of the clusters with respect to the set of top relevant dates identified.

The former judgment was performed on top of 656 distinct $(S_i, d_{h,i})$ pairs, where S_i is the set of 582 Web snippets annotated with at least one year candidate, and $d_{h,i} \in D_{S_i}$, $h = 1, .., t$, is the set of candidate dates for the snippet S_i. Since each candidate date found in a snippet can potentially originate a cluster, the task of evaluating the temporal relevance of the snippets is the task of evaluating the proper identification and significance of its dates in the context of the query. Based on this, each $(S_i, d_{h,i})$ pair was assigned a relevance label on a 2-level scale: not a date or temporally irrelevant to the query (score 0) and temporal relevant to the query (score 1). An example of this task given in Table 1.

Table 1. Year candidates for the query *Haiti Earthquake* extracted from the Web snippet with id 39.

Title	2011 Haiti Earthquake Anniversary
Snippet	As of 2010 (see 1500 photos), the following major earthquakes have been recorded in Haiti. The 1st one occurred in 1564. 2010 has been a tragic date, however in 2012 Haiti will organize the Carnival…

While there are a few year candidates, only clusters "*1564"* and "*2010"* are relevant to the query. "*2012"* is not query-related, "*1500"* is not even a date and "*2011"* may be considered not very relevant. As the task did not show to be prone to different judgments, we did not apply a multi-annotator scheme. The final list of judgments consists of 119 $(S_i, d_{h,i})$ pairs labeled with score 0, and 537 with score 1.

The second human judgment, which supports the quality assessment of the clusters, consists of 235 distinct (q, d_i) pairs[7], where q is the query and $d_i \in D_S$ the set of distinct candidate dates, potentially clusters, extracted from the set of all Web snippets S. Defining whether a cluster is or not relevant depends then on the number of corresponding relevant and irrelevant $(S_i, d_{h,i})$ classifications. Consider for example the pair (*avatar movie*, *2009*), where "*avatar movie*" is the query and "*2009*" is a candidate date. In this example we assume that "*2009*" was found within seven Web snippets and that six out of seven $(S_i, 2009), i = 1, .., 7$ pairs where classified by the human evaluator as relevant. As relevant classifications comprise the big majority, "*2009*" is automatically determined as a relevant cluster. The ground truth for this query thus consists of a

[4] Although we adopt an overlapping methodology that enables a document to be in more than one cluster, a Web snippet can be simply placed in a single main cluster. Such a cluster could be easily determined by $GTE(q, d_i)$ based on the similarity value computed for the query with respect to each of the dates found in the snippet.

[5] http://www.google.com/insights/search [7th September, 2012]

[6] http://www.bing.com [7th September, 2012]

[7] 86 (q, d_i) pairs labeled with score 0, and 149 with score 1

hypothetical temporal cluster termed "*2009*" and seven Web snippets, six classified as relevant, both to the query and to the cluster label, and one classified as irrelevant. This task is formalized in Equation (15), where $\#Rel$ represents the number of $(S_i, d_{h,i})$ whose relevance judgment equals 1, and $\#IRel$ represents the number of $(S_i, d_{h,i})$ whose relevance judgment equals 0.

$$(q, d_i) = \begin{cases} 1, if\ \#Rel \geq \#IRel \\ 0, if\ \#Rel < \#IRel \end{cases} \quad (15)$$

In the following, we describe each one of the three experiments performed. The first set of experiments uses *WC_DS* to evaluate our clustering algorithm. The second set of experiments uses *WC_DS* to compare our temporal clustering approach with *Carrot* Web-snippet clustering engine. Finally, the last set of experiments test the performance of our approach on real web user environment by conducting a user study over the same dataset.

4.1 Clustering Algorithm Evaluation

Our primary goal is to evaluate the clustering accuracy of our proposal. In the following two sections, we evaluate the quality of both clusters and snippets, by using the 235 distinct (q, d_i) pairs and the 656 distinct $(S_i, d_{h,i})$ pairs, respectively.

4.1.1 Clustering Evaluation

Our clustering algorithm consists of a two stage process that combines the identification of relevant temporal expressions extracted from Web snippets with a clustering methodology, where documents are grouped into the same cluster if they share a common year. In order to evaluate our methodology we compare the final list of temporal clusters generated by our *relevant date classification* model, against the *WC_DS* ground truth dataset.

Since we are interested in the potential agreement between the clusters and the identification of the relevant dates our objective is to search for the best cut-off λ value. For this purpose, we use common IR measures to reach a decision. In detail, we calculate *Precision* (Equation 16), *Accuracy* (Equation 17) *Recall* or *Sensitivity* (Equation 18), *F1-Measure* (Equation 19), *Balanced Accuracy* or *Efficiency* (Equation 20) and Negative Predictive value (Equation 21)

$$Precision = \frac{TP}{TP+FP} \quad (16)$$

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (17)$$

$$Recall\ (Sensitivity) = \frac{TP}{TP+FN} \quad (18)$$

$$F1 - Measure = \frac{2*Precision*Recall}{Precision+Recall} \quad (19)$$

$$Balanced\ Accuracy\ (Efficiency) = \frac{0.5*TP}{TP+FN} + \frac{0.5*TN}{TN+FP} \quad (20)$$

$$Negative\ Predictive\ Value = \frac{TN}{TN+FN} \quad (21)$$

where True Positives (TP) is the number of years (thus clusters) correctly identified as relevant, True Negatives (TN) is the number of years correctly identified as irrelevant or incorrect, False Positive (FP) is the number of years wrongly identified as irrelevant and False Negative (FN) is the number of years wrongly identified as relevant. Different tests have been performed over the 235 distinct (q, d_i) pairs, following a 5-fold cross validation approach with 80% of learning instances for training and 20% for testing. Best results pointed to 94.3% F1 performance, 93.2% Accuracy, 92.6% Balanced Accuracy, 94.2% Recall and 94.5% Precision in identifying relevant dates and thus forming relevant temporal clusters using $GTE(q, d_i) \geq \lambda$ where $\lambda = 0.35$ (see Figure 1).

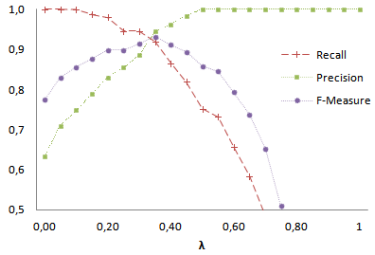

Figure 1: Performance Results vs λ.

Applying this λ to any retrieved results will enable to filter out irrelevant clusters and to select relevant ones. Table 6 shows the complete set of clusters identified for each query. Cluster labels whose dates were classified as wrong or irrelevant for the set of snippets they refer to, appear identified with a single strikethrough. While there is a query "*george bush iraq war*" with 11 clusters, the average number does not exceed the value of 3.40. This is in line with requirement **R1**. Indeed, while topic clustering systems present an excessive number of clusters this does not seem to be the case of our temporal clustering proposal. This is mostly due to two reasons. On the one hand, there is a clear reduced number of dates occurring in snippets when compared to words. On the other hand, our clustering algorithm is built upon the identification of top relevant dates, thus filtering out some wrong or irrelevant years. Indeed, 78 out of 90 candidate years were correctly filtered out by our system resulting in a negative predictive value of 86,7%. In order to quantify the effect of the application of the GTE in the clustering evaluation we performed a further experiment (denoted *Non-GTE*) which measures effectiveness when all dates are used. Results are summarized in Table 2 and show a huge difference between using GTE or not.

Table 2. Clustering Evaluation of *GTE* and Non-*GTE* over *WC_DS*

System	*F1*	*Precision*	*Recall*
Non-GTE	0.776	0.634	**1**
GTE	**0.943**	**0.945**	0.942
Improv.	0.167	0.311	-0.058

4.1.1 Snippets Evaluation

In this section we evaluate the quality of the snippets with respect to the cluster label. We rely on the *WC_DS* ground truth collection, in particular in the set of 656 distinct $(S_i, d_{h,i})$ pairs. To evaluate our proposal we use common IR measures already defined in the previous section, where TP is the number of the retrieved snippets that are relevant to the cluster label, TN is the number of snippets that were correctly classified as irrelevant with respect to the date, and thus do not appear in the final list of the results, FP is the number of the retrieved snippets that are irrelevant to the cluster label and FN is the number of relevant snippets missed by the system. Results obtained point to 95.9% F1, 92.9% Accuracy, 84.9% Balanced Accuracy, 94.6% Precision and 97.1% Recall, suggesting the appropriateness of our solution in correctly position the snippets with regard to the temporal cluster. In order to quantify the effect of the application of the GTE in the snippet evaluation we performed a further experiment (denoted *Non-GTE*) which measures

effectiveness when all dates are used. Results are summarized in Table 3 and show, again, a considerable difference between using GTE or not.

Table 3. Snippet Evaluation of *GTE* and Non-*GTE* over *WC_DS*

System	*F1*	*Precision*	*Recall*
Non-GTE	0.908	0.832	**1**
GTE	**0.959**	**0.946**	0.971
Improv.	0.051	0.114	-0.029

In order to better understand both drawbacks and strengths of our proposal we present an example for the query "*true grit*". Representations of the three clusters obtained are pictured in Figure 2.

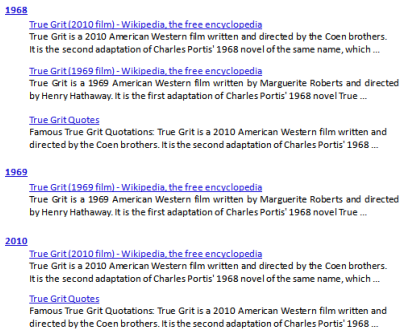

Figure 2: Temporal Disambiguation of the query "*True Grit*".

The snapshot shows the potential of our approach in disambiguating implicit temporal queries. By looking at the figure, we can quickly identify three main temporal clusters, $\{1968, 1969, 2010\}$ showing similarity with the query. Of the six candidate years initially identified by our pattern recognition system, three of them $\{1982, 2006, 2011\}$ were simply filtered out by *GTE*. It is also noteworthy that the use of *GTE* in our clustering algorithm, apparently seems to be sufficient to achieve a good quality of the snippets results as the system achieved 95.9% F1 performance. Yet, a new similarity measure that focuses on the individual temporal processing of each snippet, in line with what has been proposed by [10] and [16], can be further studied in the future, so that the snippets selection process does not depend on the *GTE* application similarity measure.

The figure also illustrates our overlapping clustering methodology as cluster "*1968*" overlaps with cluster "*1969*" and "*2010*". We believe this will certainly help the user to identify relations between each of the different dates. While overlapping could be an interesting characteristic of a temporal system, it may pose some problems in case of snippets containing a large number of dates, for which there is no further associated snippets. A clear example is given for the query "*Fernando Alonso*" and the snippet "*1988 - 1990 Karting Infant Category. Asturias Champion (won all 8 races), winner Galicia's Championship, winner Asturias Championship. 1990 - 1991 Karting Cadet Category"*. This snippet by itself will form four temporal clusters simply related to the same snippet, thus hindering the user navigation task. In such cases, snippets will likely be better fitted into a single main cluster as described in section 3.5. A more sophisticated temporal cluster presentation would also help to improve these problems.

A further relevant aspect is the language-independent characteristic of the system, making it possible to return relevant snippets from different languages. An example is the spanish text "*Natural de Tuilla (Asturias). Nacido en 1981, jugador profesional de futbol*" retrieved for the query "*David Villa*".

Note that although in this example the date "*1981*" is in the year granularity level, our system is also capable of detecting finer-grained expressions from different languages. Some examples are "*January 20 1987*" expressed in English, *20 de enero de 1987*" in Spanish or "*20 janvier 1987*" in French. Both will be normalized to the year "*1987*" according to the function *norm* defined in section 3.2.

It is also noteworthy that the final list of snippets consists of texts having at least one year annotation. As such, Web snippets not containing any identifiable year are not represented in the final list of results. While this cannot be seen as a problem, given the temporal purpose of the system, it can be improved in the future by applying a measure of similarity between the words found in the snippet and each of the relevant years retrieved for the query. This is a rather simple process as similarity values are already recorded in M_{CT} conceptual temporal correlation matrix. Despite our current good results, the ranking algorithm also leaves much room for improvement. For example, in case we have more than one date, we can weigh their similarity differently, or we can take into account the number of different dates found, or even its position. Another possibility is to consider more features, such as the distance between the query and the date/word within the snippet, in line with what has been proposed by [2]. These should be studied in future work.

4.2 Comparison against *Carrot* Web-Snippet Clustering Engine

In the second set of experiments we compare our proposal with Web-snippet clustering engines. In particular, we tested our data against *Carrot*. We have also considered the possibility to compare our approach against *iBoogie* and *Yippy* clustering engines. Yet, given the impossibility to test them against *WC_DS*, we decided not to do it, as results were likely to be influenced. As such, we simply rely on *Carrot* for this experiment. Our aim is twofold. First, we aim to show that our clustering algorithm is able to determine a wider number of temporal clusters when compared to *Carrot*. Second, we aim to assess the behavior of *Carrot* in correctly identifying relevant temporal clusters and snippets, so as to compare their results with the ones obtained by our temporal approach.

For this purpose, we used the *Carrot Document Clustering Workbench*[8] which enables to test *Carrot* with our WC_DS dataset. In order to obtain *Carrot* results, we run each of the 42 text queries on the Workbench over the *WC_DS* dataset. We used Lingo [14] an overlapping clustering algorithm, which is also used for *Carrot* live demos, and defined the *cluster count base* parameter to 100 with the purpose of obtaining the highest possible number of temporal clusters. This parameter was combined with the *allow numeric labels*, in order to enable labels to contain numbers. Note that *Carrot* does not apply a date filtering schema and as such the entire set of snippets, consisting of either relevant, wrong or irrelevant dates, will be retrieved and placed across a different set of clusters. Yet, as we intend to assess its temporal nature we will only rely on the set of clusters (and its corresponding snippets) labeled with a year, either a single numeric value "*2009*", or a combination between years and text, e.g., "*1955 October*" or "*Susan Magdalene Boyle Born 1 April 1961*".

[8] http://project.carrot2.org/download.html [7th September, 2012]

The final set of results undergoes an evaluation process to assess the performance of *Carrot* in terms of forming both relevant temporal clusters and snippets. For this purpose, results are matched against the *WC_DS* ground truth dataset and compared by means of common IR measures. Results pointed to 62.9% F1 performance, 63.4% Accuracy, 68.7% Balanced Accuracy, 49.0% Recall and 88.0% Precision in identifying relevant temporal clusters and 67.9% F1 performance, 57.5% Accuracy, 64.6% Balanced Accuracy, 54.0% Recall and 91.5% Precision in terms of snippet performance. Table 4 and Table 5 summarize both dimensions for the *GTE* and *Carrot* methodologies, showing that *GTE* improves F1 in 0.299 and 0.279, in terms of both clustering and snippets performance compared to *Carrot*.

Table 4. Clustering Evaluation of *Carrot* and *GTE* over *WC_DS*

System	*F1*	*Precision*	*Recall*	*Accuracy*	*BAccuracy*
Carrot	0.629	0.879	0.489	0.634	0.686
GTE	**0.943**	**0.945**	**0.942**	**0.932**	**0.926**
Improv.	0.314	0.066	0.453	0.298	0.240

Table 5. Snippet Evaluation of *Carrot* and *GTE* over *WC_DS*

System	*F1*	*Precision*	*Recall*	*Accuracy*	*BAccuracy*
Carrot	0.678	0.915	0.539	0.575	0.645
GTE	**0.959**	**0.946**	**0.970**	**0.929**	**0.849**
Improv.	0.281	0.031	0.431	0.354	0.204

We are aware that we are comparing two different types of approaches with different purposes, and that we expect *Carrot* to perform worse when compared to our temporal approach. Yet, the idea is precisely to show that a specific clustering temporal approach, based on the identification of relevant temporal expressions, is likely to benefit a wide range of implicit temporal queries, for which search engines continue to fail.

Table 6 and Table 7 summarize the set of temporal clusters retrieved for each of the 42 text queries. Note that the apparent lack of years in queries such as "*tour de france*" or "*football world cup*" (see Table 6), does not rely on some problem of date identification, but rather on the lack of temporal features retrieved by the Web search API for each of the queries.

Anecdotal evidence of the clusters presented in both tables, show that *GTE* is capable of retrieving a large number of temporal clusters. Two illustrative examples are the cases of "*slumdog millionaire*" and "*waka waka*" whose temporal instances were correctly identified by *GTE* but ignored by *Carrot*. Another interesting case is the query "*avatar movie*", which in addition to "*2009*" was also tagged by Carrot with an irrelevant date, in the case "*2011*". A final example, is given for the query "*osama bin laden*" for which *GTE* was able to identify a further relevant date "*2001*" when compared to Carrot. We believe applying a dedicate temporal similarity measure with the purpose of identifying relevant temporal expressions will improve the quality of the results retrieved and thus will help the user in his process of temporally disambiguate the query. In order to prove our assumption we conducted a user survey. Results are shown in the following section.

4.3 User Study

In order to test our clustering approach in real web user environments, we conducted a user-survey. Our objective was to evaluate the ability of our clustering algorithm in correctly identifying relevant clusters and snippets and in filtering out irrelevant ones. For this experiment, we used the set of results comprising the *WC_DS* dataset (without the human annotations) showing the users the set of temporal clusters (and corresponding snippets) retrieved by our approach together with those that were filtered out (similarly to what is shown in Table 6). Each query was evaluated by 6 workers using the following scale, in line with what has been proposed by [2]:

- Excellent. All irrelevant items were filtered out and all the remaining ones are relevant.
- Good. The search results are very relevant but there might be better results. Most irrelevant items were filtered out and most remaining ones are relevant.
- Fair. Somewhat relevant. There are many items that are inaccurate, either remained or were filtered out incorrectly.
- Not Relevant. The search result is not good because it contains too many wrong decisions.
- I don't know. I can't evaluate the quality of the search results.

The most frequent response was "*Excellent*" (see Figure 3) with an average of 4.30. Overall, the annotators obtained about 0.46 of agreement level by applying the Fleiss Kappa statistics [7]. Although this represents a low agreement between the annotators it does not compromise the validity of the results, as disagreements mostly concern to differences between classifying a query as "Excellent" and "Good", and not between "Excellent" and "Fair" classifications. This can be easily proved, as Kappa agreement gets improved to 0.81 if we simply divide the set of results into the class of relevant quality assessments (Excellent + Good) and the class of irrelevant quality ones (Fair + Not Relevant + I don't know).

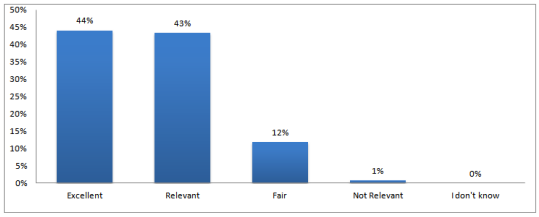

Figure 3: Evaluation of the quality of the results produced by our clustering algorithm for the entire set of 42 queries.

5. CONCLUSIONS

In this paper, we propose a strategy for the temporal clustering of search engine query results, where snippets are clustered by year. We rely on a novel temporal similarity measure named *GTE* which enables to detect top relevant years and filter out irrelevant ones. Results obtained show that the introduction of *GTE* benefits the quality of the clusters generated, by retrieving a high number of precise relevant dates. Comparative experiments have also been performed over Carrot Web-snippet clustering engine. Results showed that our clustering approach is more effective than the approach of Carrot in temporally disambiguating a query. These results were complemented with a user survey showing that users mostly agree with the set of temporal clusters retrieved by our system.

Although our current approach already enables a simple form of temporal clustering, which can guarantee that the detected temporal expression is related to the query, we cannot claim that the Web snippets inside the same temporal cluster are topic related. As such, the aim of our larger project to which these results contribute to, is to provide an effective clustering algorithm that ranks snippets, both based on their temporal and conceptual proximities.

Table 6. List of Queries. Temporal Clusters obtained by *GTE*

george bush iraq war *1946, 1990, 1991, 1995, 2000, 2001, 2002, 2003, ~~2004~~, 2005, 2009*	**tour de france** *1903, 2009, 2010, 2011, 2012*	**ryan dunn** *1977, 2002, 2003, 2006, 2010, 2011*	**steve jobs** *1955, 1970, 1998, 2005, 2011*
slumdog millionaire *2008*	**britney spears** *1981, 2008*	**troy davis** *1969, 1989, 1991, 2011*	**david villa** *1981, 2008, 2011, ~~2012~~*
football world cup *1930, 2006, 2010, ~~2011~~, ~~2012~~, 2014, 2018, 2022*	**justin bieber** *1994, 2011*	**adele** *1988, 2006, 2008, 2009, 2011*	**dan wheldon** *1978, 2005, 2011*
walt disney company *1920, 1923*	**rebecca black** *1997, 2011*	**lady gaga** *1986, ~~2004~~, 2008*	**dacia duster** *~~1180~~, 2009, 2011*
lena meyer-landrut *1991, 2010, 2011*	**kate middleton** *1982, 2010, 2011*	**swine flu** *2009, 2011, 2012*	**waka waka** *2010*
fernando Alonso *1981, 1988, 1990, 1991, 2005, 2006, 2011*	**david beckham** *1975, 2006, 2007, 2011*	**fiat 500** *1936, 1955, 1957, 1975, 2012*	**obama** *1961, 1964, 2008, ~~2011~~, 2012*
sherlock holmes *1887, 2009, 2011*	**volcano iceland** *1918, 2004, 2010*	**kate nash** *1987, 2006, 2007, 2008, 2009*	**katy perry** *1984, 2008, 2009, 2010, 2012*
california king bed *2010, 2011*	**bp oil spill** *2010, 2011*	**tour eiffel** *1989, 1959*	**haiti** *1953, 1956, 2010*
osama bin laden *1957, 2001, 2011*	**little fockers** *2000, 2010*	**fukushima** *2001, 2011*	**nissan juke** *2011, 2012*
amy winehouse *1983, 2000, 2011*	**marco simoncelli** *1987, 2011*	**true grit** *1968, 1969, 2010*	**susan boyle** *1961, 2009*
haiti earthquake *2010*	**avatar movie** *2009*		

Table 7. List of Queries. Temporal Clusters obtained by *Carrot*

george bush iraq war *1991, 2001, 2002, 2003, ~~2004~~*	**tour de france** *2010, 2011, 2012*	**ryan dunn** *1977*	**steve jobs** *1955, 2005, 2011*
slumdog millionaire	**britney spears**	**troy davis** *1991, 2011*	**david villa** *1981, 2008*
football world cup *2010, 2014, 2018*	**justin bieber** *1994, 2011*	**adele** *2011*	**dan wheldon** *1978, 2005, 2011*
walt disney company *1923*	**rebecca black** *2011*	**lady gaga** *~~2011~~*	**dacia duster** *2009, ~~2011~~*
lena meyer-landrut *1991, 2010, 2011*	**kate middleton** *2011*	**swine flu** *2009*	**waka waka**
fernando Alonso *2005, 2006, 2011*	**david beckham** *2006, 2007, 2011*	**fiat 500** *1936, 1955, 2007, ~~2011~~, 2012*	**obama** *2008, ~~2009~~, ~~2010~~, ~~2011~~, 2012*
sherlock holmes *2009, 2011*	**volcano iceland** *1918, 2004, 2010*	**kate nash** *1987, 2007, ~~2011~~*	**katy perry** *2008, 2010*
california king bed *2010*	**bp oil spill** *2010*	**tour eiffel** *~~1989~~*	**haiti**
osama bin laden *1957, 2011*	**little fockers** *2010*	**fukushima** *2011*	**nissan juke** *2011, 2012*
amy winehouse *1983, 2000, ~~2008~~, 2011*	**marco simoncelli** *1987*	**true grit** *1968, 1969, 2010, ~~2011~~*	**susan boyle** *1961, 2009*
haiti earthquake *2010*	**avatar movie** *2009, ~~2011~~*		

6. ACKNOWLEDGMENTS

This work is funded by the ERDF through the Programme COMPETE and by the Portuguese Government through FCT - Foundation for Science and Technology and grant (Reference: SFRH/BD/63646/2009). It is also supported by the Center of Mathematics, University of Beira Interior, project PEst-OE/MAT/UI0212/2011. We would also like to thanks the human annotators for their hard work.

7. REFERENCES

[1] Alonso, O., & Gertz, M. (2006). Clustering of Search Results using Temporal Attributes. *In SIGIR'06,* pp. 597 - 598. Seattle, Washington, USA. August 6 - 11: ACM Press.

[2] Alonso, O., Gertz, M., & Baeza-Yates, R. (2009). Clustering and Exploring Search Results using Timeline Constructions. *In CIKM'09.* Hong Kong, China. November 2 - 6: ACM Press.

[3] Campos, R., Dias, G., Jorge, A., & Nunes, C. (2012). GTE: A Distributional Second-Order Co-Occurrence Approach to Improve the Identification of Top Relevant Dates in Web Snippets. *In CIKM'12*. Maui, Hawaii, USA. October 29–November 2.

[4] Campos, R., Jorge, A., & Dias, G. (2011). Using Web Snippets and Query-logs to Measure Implicit Temporal Intents in Queries. *In SIGIR-QRU'11*, pp. 13 - 16. Beijing, China. July 28.

[5] Campos, R. (2011). http://www.ccc.ipt.pt/~ricardo/software

[6] Dias, G., Alves, E., and Lopes, J. (2007). Topic Segmentation Algorithms for Text Summarization and Passage Retrieval: An Exhaustive Evaluation. In *AAAI'07,* 1334-1340. Canada. July 22-26.

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

[8] Jatowt, A., & Yeung, C. M. (2011). Extracting Collective Expectations about the Future from Large Text Collections. *In CIKM'11,* pp. 1259 - 1264. Glasgow, Scotland, UK. October.

[9] Jatowt, A., Kawai, H., Kanazawa, K., Tanaka, K., & Kunieda, K. (2009). Supporting Analysis of Future-Related Information in News Archives and the Web. *In JCDL'09,* pp. 115 - 124. Austin, USA. June 15 - 19.: ACM Press.

[10] Kawai, H., Jatowt, A., Tanaka, K., Kunieda, K., & Yamada, K. (2010). ChronoSeeker: Search Engine for Future and Past Events. *In ICUIMC'10,* pp. 166 - 175. Republic of Korea. January 14 - 15.

[11] Machado, D., Barbosa, T., Pais, S., Martins, B and Dias, G. (2009). Universal Mobile Information Retrieval. In HCII'09, USA.

[12] Metzler, D., Jones, R., Peng, F., & Zhang, R. (2009). Improving Search Relevance for Implicitly Temporal Queries. *In SIGIR'09,* pp. 700 - 701. Boston, MA, USA. July 19 - 23: ACM Press.

[13] Mori, M., Miura, T., & Shioya, I. (2006). Topic Detection and Tracking for News Web Pages. *In WIC'06,* pp. 338 - 342. Hong Kong, China. December 18 - 22: IEEE Computer Society Press.

[14] Osinski, S., & Weiss, D. (2005). A Concept-Driven Algorithm for Clustering Search Results. In *IEEE Intelligent Systems*, 20(3),48-54.

[15] Shaparenko, B., Caruana, R., Gehrke, J., & Joachims, T. (2005). Identifying Temporal Paterns and Key Players in Document Collections. *In TDM'05,* pp. 165 - 174. USA. November 27 - 30.

[16] Strötgen, J., Alonso, O., & Gertz, M. (2012). Identification of Top Relevant Temporal Expressions in Documents. *In WWW-TWAW'12,* pp. 33 - 40. Lyon, France. April 17: ACM - DL.

[17] Zamir, O., and Etzioni, O. (1998). Web Document Clustering: A Feasibility Demonstration. In SIGIR'98, 46-54. Australia