GTE: A Distributional Second-Order Co-Occurrence Approach to Improve the Identification of Top Relevant Dates in Web Snippets

Ricardo Campos[1,2,6], Gaël Dias[4,6], Alípio Mário Jorge[1,3], 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, gael.dias@unicaen.fr, amjorge@fc.up.pt, celian@ubi.pt

ABSTRACT
In this paper, we present an approach to identify top relevant dates in Web snippets with respect to a given implicit temporal query. Our approach is two-fold. First, we propose a generic temporal similarity measure called *GTE*, which evaluates the temporal similarity between a query and a date. Second, we propose a classification model to accurately relate relevant dates to their corresponding query terms and withdraw irrelevant ones. We suggest two different solutions: a threshold-based classification strategy and a supervised classifier based on a combination of multiple similarity measures. We evaluate both strategies over a set of real-world text queries and compare the performance of our Web snippet approach with a query log approach over the same set of queries. Experiments show that determining the most relevant dates of any given implicit temporal query can be improved with *GTE* combined with the second order similarity measure InfoSimba, the Dice coefficient and the threshold-based strategy compared to (1) first-order similarity measures and (2) the query log based approach.

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, Implicit Temporal Queries, Temporal Query Understanding, Query Log Analysis.

1. INTRODUCTION
Recent years have seen significant progress in Temporal Information Retrieval (T-IR), which aims to exploit temporal information in order to improve the Web search process. Despite the fact that great improvements have been achieved, existing solutions still face challenges that are not easy to deal with such as a deep understanding of the temporal intents of user text queries. While in the case of explicit temporal queries (e.g. "*Fukushima 2011*") the retrieval task can be relatively straightforward as the temporal purpose is explicitly defined by the user, in the case of implicit temporal ones (e.g. "*Iraq War*") it is much more complex as it involves estimating the temporal part of the query. Given that most of the temporal queries issued by users are implicit by nature [5], detecting its underlying temporal intent turns out to be a very interesting problem and a real need to improve the performance of search systems. In this context, most state-of-the-art methodologies consider any occurrence of temporal expressions in Web snippets and other Web data as equally relevant to an implicit temporal query. This is obviously not true for most query results.

In this work, we aim to define the temporal intents of implicit temporal queries in order to further improve the Web search process. As referred by Berberich et. al. [3] this is a challenging problem for which there is no clear solution yet. For that, we propose a language-independent strategy to associate top relevant years to any text query by analyzing its corresponding Web snippets. As shown by Alonso et. al. [1] [2], snippets are an interesting alternative for the representation of Web documents, where dates, especially in the form of years, often appear. However, this diversity of temporal expressions poses some challenges since only a few of them are actually relevant to the query. Hence, our goal is twofold: (1) select the most relevant dates for a given query and (2) discard all irrelevant or incorrect ones. The contributions of this work can be summarized as follows: (1) we propose a novel approach to properly tag text queries with relevant temporal expressions by relying on a content-based approach and a language-independent methodology; (2) our measure, outperforms well-known first order similarity measures improving precision in correctly tagging dates against a query log based approach and (4) we publicly provide a set of queries and ground-truth results to the research community. The remainder of this paper is organized as follows. In Section 2, we describe related work. In Section 3, we propose the *GTE* (*Generic Temp Eval*) similarity measure and present the two strategies for date classification. Experimental setups and results are discussed in Section 4 and Section 5, respectively. Finally, we conclude this paper in Section 6 with some final remarks.

2. RELATED WORK
To the best of our knowledge only three works [9] [11] [13] have been proposed with regard to the identification of top relevant expressions given a user implicit temporal query. Metzler et. al. [11] mine query logs to identify implicit temporal information needs. They propose a weighted measure that considers the number of times a query q is pre- and post-qualified with a given year y. A query is then implicitly year qualified if it is qualified by at least two unique years. A relevance value is then given for each year found in the document. This work proposes an interesting solution as it introduces the notion of correlation between a query and a year but lacks in query coverage as it depends on query log analysis.

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

Kawai et. al [9], on the other hand, developed a chronological events search engine for the Japanese language based on Web snippets analysis. In order to collect a large number of temporal expressions, the authors expand the query with language dependent expressions related to event information such as year expressions, temporal modifiers and context terms. Then, noisy temporal patterns are removed using machine learning techniques trained over a set of text features. Finally, Strötgen et. al. [13] 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. Temporal expressions are extracted by applying the HeidelTime tagger and all the features are combined into a single relevance function based on a set of pre-defined heuristics. Although being an interesting approach, this research lacks a further evaluation in terms of IR metrics. Moreover both proposals reveal a dependency in terms of language.

3. OUR APPROACH

In this section, we describe our method to identify top relevant dates related to text queries with temporal dimensions. We divide this method into the following three main subtasks: (1) Web snippet processing, (2) Date-Query relevance identification and (3) Relevant date classifications. In particular, step (1) has been well studied as there are many successful methods for the extraction of text and temporal information. However, to the best of our knowledge there are only three works [9] [11] [13] with regard to step (2) and only one [9] related to step (3). Although this process is easier in the case of explicit temporal queries (e.g., "*War 1945*"), it turns out to be very difficult in the case of implicit temporal ones (e.g. "*Avatar Movie*"). For that purpose, we first describe the Web search and Web snippet module. Then, we propose *GTE*, a generic temporal similarity measure to assess the similarity between text queries and temporal expressions. And finally, we propose two different classification models in order to retrieve the top relevant temporal expressions and filter out irrelevant ones: (1) a threshold-based classification strategy that defines the boundary between relevant and irrelevant temporal associations and (2) a SVM classifier based on a combination of multiple similarity measures, which act as different features for classification.

3.1 Web Snippet Processing

A query can either be 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 explicit queries are easier to deal with. For better readability, we denote a query simply as q. Although we have focused on Web snippets in our experiments, our temporal similarity measure is also applicable to any document collection embodying temporal information, such as Wikipedia pages or Twitter posts. Similarly to Kawai et. al [9], 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 $S = \{S_1, S_2, \dots, S_n\}$. Each S_i, for $i = 1, \dots, n$, consists of its title and its text, i.e. $\{Title_i, Snippet_i\}$ and is represented by a bag-of-words and a set of candidate temporal expressions. Specifically, $W_{S_i} = \{w_{1,i}, w_{2,i}, \dots, w_{k,i}\}$ is defined as the set of k most relevant words/multi-words associated with a Web snippet S_i and $D_{S_i} = \{d_{1,i}, d_{2,i}, \dots, d_{t,i}\}$ as the set of t candidate years associated to a Web snippet S_i. Moreover $W_S = \bigcup_{i=1}^{n} W_{S_i}$ 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, $D_S = \bigcup_{i=1}^{n} D_{S_i}$ 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. [10], 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. Indeed, due to the specific structure of Web snippets, these tools usually fail to correctly process this type of collection. Due to space limitations, we do not detail this pre-processing step as it can easily be reproduced from [10], 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-independence 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. Finally, $W^* = W_S \cap W_{S_{d_i}}$ is defined as the set of distinct words that results from the intersection between the set of words W_S and the set $W_{S_{d_i}}$ which contains the words that appear together with date d_i, in any Web snippet S_i, from S.

3.2 GTE: Temporal Similarity Measure

We formally define the problem of query temporal tagging as follows.

<u>Problem Definition</u>: given a query q and a date $d_i \in D_S$ assign a degree of relevance to each (q, d_i) pair. To model this relevance, a temporal similarity value v is defined by a similarity measure $sim(q, d_i)$, $v \in [0,1]$.

The proposed formulation attempts to identify relevant dates d_i for q and minimize any errors that might arise from considering irrelevant or wrong dates. As we will demonstrate in this paper, the relevance between a (q, d_i) pair is better defined if, instead of just focusing on the self-similarity between the query q and the date d_i, all the information existing between W^* and d_i is considered. As such, we will not only define the similarity between the query word and the candidate date, but also between each of the most important topics extracted from the Web snippets and their respective candidate dates. Based on this principle, the *GTE* measure is formalized in Equation 1, 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 representing its relevance:

$$GenTempEval(q, d_i) = F(sim(W^*, d_i)) \quad (1)$$

We consider three different F functions, specifically (1) the Max/Min, (2) the Arithmetic Mean and (3) the Median. Extensive experiments have been performed to assess the different aggregation functions [4]. Overall, the median gave more satisfactory results. As such, we will only focus on this approach in the remainder of this paper. The overall strategy of our query time tagging relevance model is shown in Algorithm 1.

Algorithm 1: Assign a degree of relevance to each (q, d_i) pair

Input: query q
1: S ← RequestSearchEngine(q)
2: For each $S_i \in S$, i = 1,..,n
3: Apply Text Processing
4: W_{S_i} ← Select best relevant words/multi-words in S_i
5: D_{S_i} ← Select all temporal patterns in S_i
6: $W_S \leftarrow \bigcup_{i=1}^{n} W_{S_i}$
7: $D_S \leftarrow \bigcup_{i=1}^{n} D_{S_i}$
8: For each $d_i \in D$
9: Compute $GTE(q, d_i)$
Output: (q, d_i) relevance

The algorithm receives a query from the user, fetches related Web snippets from a given search engine and applies text processing to all Web snippets. This processing task involves selecting the most relevant words/multi-words and collecting the candidate years. Words and dates are then associated to a list of distinct terms. Finally, each date is given a temporal similarity value $GTE(q, d_i)$ which is computed with any $sim(.,.)$ similarity measure. While $sim(.,.)$ can be any similarity measure, either of first or second order, we believe it can be modeled more effectively if it is based on a second order similarity measure. Our hypothesis, which will be supported in the experiments section, is that second order similarity measures carry valuable additional relations in both the word $X \in W^*$ and the date d_i context vectors, which cannot be induced if a direct co-occurrence approach is used. The use of a second-order co-occurrence measure requires however the definition of a context vector for each of the two items, such that the word X and the date d_i are similar if their context vectors are also similar. In this context, most of the works apply the cosine similarity measure, in order to assess the similarity between the two context vectors. However, as most of them rely on exact matches of context words, their accuracy is low since language is creative and ambiguous [8]. This is particularly evident in the case of relations between words and dates, where the cosine similarity measure may not even be applied. In order to overcome these drawbacks we apply the *InfoSimba* (*IS*) similarity measure [7] (see Equation (2)), which can be seen as a semantic vector space model supported by corpus-based word correlations.

$$IS(V_x, V_y) = \frac{\sum_{i \in V_x} \sum_{j \in V_y} S(i,j)}{\left(\sum_{i \in V_x} \sum_{j \in V_x} S(i,j) + \sum_{i \in V_y} \sum_{j \in V_y} S(i,j) - \sum_{i \in V_x} \sum_{j \in V_y} S(i,j)\right)} \qquad (2)$$

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. For this purpose, five possible representations such as (*W*;*W*) (*D*;*D*), (*W*;*D*), (*D*;*W*) and (*WD*;*WD*), have been defined where *W* stands for a word-only context vector, *D* for a date-only one and *WD* for a word and date context vector. Overall, the (*WD*;*WD*) representation gave more satisfactory results, meaning that each of the two context vectors can be better defined if represented by a set of words and a set of temporal patterns. The similarity between each pair of the two context vectors is determined by any first order similarity measure $S(.,.)$ (e.g., *PMI, EI* or *DICE*) relating items *i* and *j*. For this purpose, we build a global conceptual temporal correlation matrix M_{ct}, which will serve to store the similarity value obtained between the most important words and the candidate dates. In detail,

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

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 the matrix.

3.3 Relevant Date Classification

Based on the similarity value obtained from the *GTE* measure, we need an appropriate classification strategy to determine whether the candidate temporal expressions are actually relevant or not. For that, we propose two approaches. The first one is to use a classical threshold-based strategy. Given a (q, d_i) pair, the system automatically classifies a date based on the following expression: (1) relevant, if $GTE(q, d_i) \geq \lambda$, and (2) irrelevant or wrong date, if $GTE(q, d_i) < \lambda$, where λ has to be tuned to at least a local optimum. The second strategy uses a SVM learning model. For this purpose, a set of different first order and second order similarity measures are defined for each (q, d_i) pair, in line with what has been proposed by Pecina et. al. [12]. As such, each (q, d_i) pair can be seen as a learning instance associated to the set of different characteristics, thus defining a classical learning problem. In the following section, we define the experimental set-ups.

4. EXPERIMENTAL SETUPS

Since no benchmark for (q, d_i) pairs exists, we built two new data sets, consisting of 42 text queries, one based on Web snippets and another one based on query logs, both publicly available for research purposes [6]. For the WC_DS, we queried Bing search engine for each of the 42 queries, collecting the top best 50 relevant Web results, which resulted in a set of 582 relevant Web snippets with years and 235 distinct (q, d_i) pairs. The ground truth was then obtained by manually labeling each one of the 235 distinct (q, d_i) pairs. Each pair was assigned a relevance label by a human judge on a 2-level scale: not a date or irrelevant (score 0) and relevant date (score 1). An overall example of this task is given in Table 2. The second data set QLOG_DS, which will be used to compare with the Web snippet approach, was constructed based on Google and Yahoo! auto-completion engines, which suggest a set of ten expanded queries for any given query. So, as to enable a fair comparison, we retrieved the highest number of possible date results for each of the 42 text queries, by using three different query combinations: (a) "*query*", (b) "*query + 1*" and (c) "*query + 2*", which enable to capture the query together with dates starting at 1 and 2 respectively. Like for the previous approach, candidate dates were extracted based on the set of regular expressions introduced in section 3.1. Each (q, d_i) pair was then manually labeled in the same way as for the first data set.

In order to better understand the different experiments, we propose the following notations. The different versions of the *GTE* combined with *IS* are represented as *IS_(X;Y)_SM_F(q,d$_i$)*, where (*X*;*Y*) means the type of the context vectors, *SM* is any similarity measure used in *IS* and *F* is the aggregator function. More, the first order measures applied to the *GTE* are noted *SM_F(q,d$_i$)* where *SM* is any measure (i.e. *PMI*, *DICE, Jaccard*, *EI*, *NGD*, *WebJaccard*, *WebOverlap*, *WebDICE*, *WebPMI*). Outside *GTE* (i.e. without aggregation function), we propose a set of different second order similarity measures based on *IS* and noted *IS_(X;Y)_SM(q,d$_i$)*, which evaluate the temporal similarity by <u>exclusively</u> taking into account query q and date d_i and not their correlated words W^*. Similarly, we directly apply first order similarity measures without the aggregator function, denoted *SM*.

In order to evaluate all strategies, we propose classical evaluation metrics in IR based on a confusion matrix with True Positives (TP) being the number of years correctly identified as relevant, True Negatives (TN) being the number of years correctly identified as irrelevant or incorrect, False Positive (FP) being the number of years wrongly identified as irrelevant and False Negative (FN) being the number of years wrongly identified as relevant. As such, we calculate *Precision* (P), *Recall* (R), *F*1-*Measure* (F1) and *Balanced Accuracy* (BA).

$$P = \frac{TP}{TP+FP} \quad R = \frac{TP}{TP+FN} \quad F1 = \frac{2*P*R}{P+R} \quad BA = \frac{0.5*TP}{TP+FN} + \frac{0.5*TN}{TN+FP} \qquad (4)$$

5. RESULTS AND DISCUSSION

In this section, we describe the set of experiments for both the data sets WC_DS and QLOG_DS.

5.1 Experiments on WC_DS

We performed a set of experiments with different sizes of N for the context vector and different threshold values T in order to decide whether we should consider as input for the context vector, all the terms or just those having a similarity value higher than T. For this purpose, we limited the parameters within the ranges of $5 \leq N \leq +\infty$ and $0 \leq T \leq 0.9$ and combined them as: {*T0.0N5, T0.0N10, T0.0N20, T0.0N+∞,..., T0.9N5, T0.9N10, T0.9N20, T0.9N+∞*}. Results showed that the best combination was obtained for *T0.05N+∞* (i.e. the selection of all (*N=+∞*) terms related to the (q, d_i) pair having a similarity value $T \geq 0.05$). In detail, the best biserial correlation value (i.e. 0.80 – see Table 2) is given for the Median aggregator function for *IS_(WD;WD)_DICE_M*, denoted *BGTE* (i.e. Best *GTE*) in the remainder of this paper. Table 2 lists the similarity scores between a sub-set of (q, d_i) pairs to compare the *BGTE* with baseline methods. The highest biserial correlation coefficient is reported by the proposed *BGTE* with a notable improvement compared to all measures, in particular to Web-based ones. The next step of our approach is to define an appropriate classification strategy to determine whether a date is or is not relevant to a query.

5.1.1 Threshold Classification on WC_DS

Our first approach is to use a classical threshold-based strategy as described in section 3.3 where λ has to be tuned to at least a local optimum. To avoid over-fitting and understand the generalization of the results, we followed a 5-fold cross validation approach for all the proposed measures with 80% of learning instances for training and 20% for testing. A summary of the experimental results can be found in Table 3 and Table 4, for the non-aggregated approach and for *GTE* with the Median function which obtained the best results. We can observe that *BGTE* can achieve 94.3% F1 performance, 92.6% (BA), 94.5% (P) and 94.2% (R) matching a cutoff of λ = 0.35. These results were complemented with a ROC curve, which indicates an almost perfect classifier with an Area Under Curve (AUC) of 0.953. When compared to non-*GTE* and non-*IS* similarity measures (see Table 3), the *BGTE* can produce 19.9% *F*1 improvements over the best performing measure i.e. *WebPMI* with 74.4% *F*1. Further experiments show that by simply adding the Median aggregator function to the simple *IS_(WD;WD)_DICE* results in an improvement of 11.3% in terms of *F*1. Indeed, all similarity measures with *GTE* outperform their baselines in terms of *F*1, indicating that using the Median as part of the model improves the performance of the system. In our next experiment, we compare the results of *BGTE* with the baseline rule-based model, which selects all of the temporal patterns found as correct dates within a given data set. As a consequence, for a fair comparison, we fixed a Recall of 1 for the *BGTE*. Results are presented in Table 1 (in the last two columns). While the *BGTE* threshold strategy is forced to have a recall equal to one, it still significantly outperforms the baseline model. To assess if the difference between using the *BGTE* or the baseline rule-based model for the correct classification of a (q, d_i) pair is significant, we performed the McNemar's test, a non-parametric method particularly suitable for non-independent dichotomous variables. The test resulted in a Chi-squared statistic value equal to 126.130 with a p-value < 2.2e-16. This indicates that the difference of the correct date classifications is significantly different. Based on this result, we also built a confidence interval for the difference of means for paired samples between the number of misclassified dates given by the rule-based method and by the *BGTE*. The interval obtained [1.42; 2.30] clearly shows that the rule-based model retrieves, on average, more irrelevant or incorrect dates than the *BGTE* measure, with a 95% confidence level.

5.1.2 SVM Classification on WC_DS

In alternative to the threshold-based strategy, which uses a single similarity measure to classify any query date pair, we propose to train a SVM model based on a combination of similarity measures. For this purpose, we defined a set of different first order and second order similarity measures for each (q, d_i) pair, in line with Pecina et. al. [12]. As such, each (q, d_i) pair can be seen as a learning instance described by 24 different similarity measures and its manually defined class label (relevant or not relevant). Experiments were run over the implementation of the sequential minimal optimization algorithm to train a support vector classifier using a polynomial kernel with the default parameters of Weka software. A 5-fold cross validation was performed before and after a feature selection process based on principal component analysis. After feature selection, only 14 similarity measures remained for the learning process. Results are illustrated in Table 5 for both situations, with and without feature selection with (1) respective accuracies of 88.6% and 90.3%, (2) respective *F*1-measures of 88.5% and 90.2% and (3) respective *AUC*s of 87.6% and 89.4%. The first conclusion of these results confirms the experiments of Pecina et al. [12], although in a different domain, that most similarity measures, when combined, can lead to improved results as they behave differently. As a consequence, feature selection may not lead to improved results. The second conclusion is that a unique adapted similarity measure in a threshold-based classification strategy can improve results over a classical learning process. Indeed, compared to the threshold-based classification strategy, the results obtained by the SVM classification are worse than only using *BGTE*. In the same experimental conditions, the threshold-based strategy shows performances of 92.6% accuracy (improvement of 2.3%), 94.3% *F*1-measure (improvement of 4.1%) and 95.3% AUC (improvement of 5.3%).

5.2 Experiments on QLOG_DS

In this section, we compare the *BGTE* measure with the threshold strategy over the WC_DS against QLOG_DS and the baseline, which corresponds to take into account all the retrieved dates. Table 1 presents the overall results for both approaches.

Table 1. Performance Results for both Approaches.

	Google_QLogs	Yahoo_QLogs	*BGTE*	Baseline
Precision	0.653	0.647	**0.748**	0.634
Recall	1	1	1	1
*F*1-Meas.	0.790	0.786	**0.856**	0.776

Once again, it is important to note that for a fair evaluation, we base the comparison on a Recall equal to 1. The results are shown individually for each auto-completion tool of Google (Google_QLogs) and Yahoo! (Yahoo_QLogs). The results show that *BGTE* achieves 85.6% of *F*1 performance and 74.8% of Precision, which is significantly higher than those achieved by each of the two completion engines. As in the previous experiment, we built a confidence interval for the difference of means, for paired samples, between the number of misclassified dates given by each of the two query log approaches and *BGTE*. The interval obtained for Google_QLogs is given by [1.32, 3.20] and for Yahoo_QLogs it is [1.44, 3.47]. These intervals show that both approaches retrieve on average a significant number of irrelevant or incorrect dates when compared to *BGTE*, with 95% of confidence. Not surprisingly, results show that query logs are able to return a great number of potential query related year dates, when compared to Web snippets. But, interestingly, we found that a large number of these temporally explicit queries consist of misleading temporal relations i.e. users may execute incorrect temporal queries as they may not know the exact date related to their query.

Table 2. List of (q, d_i) examples with the BGTE for the Median aggregator function compared to baseline methods.

(q, d_i) Pair	Class	*BGTE*	*NGD*	*WebJaccard*	*WebDICE*	*WebPMI*	*PMI*	*DICE*	*Jaccard*	*EI*
True grit – 1969	1	0.896	0.360	0.290	0.012	0.325	0.378	0.255	0.194	0.217
True grit – 2010	1	0.812	0.327	0.336	0.201	0.414	0.378	0.750	0.679	0.759
Avatar movie – 2009	1	0.670	0.325	0.516	0.621	0.455	0.261	0.412	0.330	0.214
Avatar movie – 2011	0	0.346	0.330	0.454	0.515	0.432	0.261	0.102	0.074	0.043
California king bed – 2010	1	0.893	0.334	0.398	0.388	0.417	0.518	0.329	0.257	0.287
Slumdog millionaire – 2009	0	0.000	0.311	0.350	0.251	0.461	0.388	0.069	0.049	0.055
Tour Eiffel – 1512	0	0.286	0.331	0.288	0.001	0.267	0.432	0.075	0.054	0.060
Lady gaga – 1416	0	0.336	0.337	0.289	0.003	0.275	0.368	0.066	0.047	0.053
Haiti earthquake – 2010	1	0.605	0.328	0.339	0.210	0.426	0.449	1.000	1.000	1.000
Sherlock Holmes – 1887	1	0.839	0.342	0.292	0.020	0.330	0.388	0.135	0.099	0.111
Dacia duster – 1466	0	0.096	0.323	0.288	0.000	0.206	0.378	0.067	0.048	0.054
Waka waka – 1328	0	0.246	0.321	0.288	0.000	0.102	0.492	0.084	0.061	0.068
Waka waka – 2010	1	0.944	0.328	0.332	0.188	0.420	0.492	0.742	0.670	0.749
Bp oil spill – 2006	0	0.277	0.300	0.350	0.248	0.454	0.545	0.094	0.068	0.076
Bp oil spill – 2010	1	0.838	0.328	0.323	0.154	0.426	0.254	0.384	0.304	0.211
Volcano Iceland – 2010	1	0.749	0.000	0.288	0.000	0.290	0.368	0.000	0.000	0.000
Point Biserial Correlation	-	0.800	-0.065	-0.110	-0.002	-0.081	-0.031	0.385	0.366	0.358

Table 3. Evaluation results on WC_DS for $sim(q, d_i)$.

Measure	λ	Recall	Prec.	BAcc.	F1	AUC
IS_(WD;WD)_ EI	0.15	0.638	0.953	0.786	0.763	0.795
IS_(WD;WD)_DICE	0.15	0.754	0.924	**0.823**	**0.830**	**0.803**
IS_(WD;WD)_PMI	0.24	0.738	0.709	0.598	0.720	0.597
EI	0.05	0.473	**0.986**	0.730	0.639	0.537
PMI	0.05	0.376	0.648	0.521	0.473	0.561
DICE	0.05	0.598	0.817	0.712	0.687	0.728
Jaccard	0.05	0.526	0.885	0.703	0.659	0.696
WebPMI	0.91	0.768	0.725	0.576	0.744	0.600
WebDICE	0.11	0.497	0.593	0.464	0.538	0.565
WebJaccard	0.05	0.489	0.583	0.322	0.530	0.616
WebOverlap	0.15	0.704	0.616	0.489	0.650	0.605
NGD	0.75	**0.852**	0.580	0.502	0.690	0.529

Table 4. Evaluation results on WC_DS for $M(sim(W^*, d_i))$.

Measure	λ	Recall	Prec.	BAcc.	F1	AUC
IS_(WD;WD)_ EI_M	0.25	0.932	0.896	0.846	0.898	0.891
IS_(WD;WD)_DICE_M	0.35	0.942	**0.945**	**0.926**	**0.943**	**0.953**
IS_(WD;WD)_PMI_M	0.16	0.980	0.727	0.682	0.833	0.714
EI_M	0.05	0.890	0.652	0.614	0.748	0.578
PMI_M	0.10	**1**	0.684	0.579	0.812	0.575
DICE_M	0.15	0.958	0.723	0.669	0.823	0.656
Jaccard_M	0.10	0.881	0.792	0.729	0.833	0.769
WebPMI_M	0.42	0.949	0.612	0.517	0.743	0.526
WebDICE_M	0.79	0.377	0.630	0.519	0.462	0.536
WebJaccard_M	0.04	0.701	0.586	0.468	0.617	0.648
WebOverlap_M	0.90	0.630	0.640	0.483	0.619	0.551
NGD	0.75	**1**	0.693	0.547	0.817	0.547

Table 5. Best Overall Classification for each group of measures.

Attribute Set	Balanced Accuracy	Average *F*1-Measure	Average AUC	Correct Date			Incorrect or Irrelevant Date		
				Precision	Recall	*F*1-Measure	Precision	Recall	*F*1-Measure
All Measures	0.903	0.902	0.894	0.920	0.926	0.923	0.872	0.862	0.867
All Measures after Feature Selection	0.886	0.885	0.876	0.907	0.913	0.910	0.849	0.839	0.844

6. CONCLUSION AND FUTURE WORK

In this work, we proposed a new temporal similarity measure, *GTE*, which allows different combinations of first order and second order similarity measures to compute the temporal intents of query dates (q, d_i) pairs. In particular, we showed that the combination of the second order similarity measure *IS* combined with the *DICE* coefficient shows improved results over all other combinations based on the threshold classification strategy. Comparative experiments have also been performed on two different data sets (*WC_DS* and *QLOG_DS*).

Results showed that the Web snippets approach is more effective than the query log based one. The results indicate that the introduction of an additional layer of knowledge, with a second order similarity measure may affect the efficiency of a broad set of T-IR systems, by retrieving a high number of precise relevant dates. As a consequence, we plan to use this new classifier in the field of Temporal Clustering. Indeed, as the methodology is language-independent and does not depend on lists of stop-words, it can be applied to real-world search scenarios.

7. ACKNOWLEDGMENTS

This work is funded by the ERDF through the Program COMPETE and by the Portuguese Government through FCT - Foundation for Science and Technology within the project FCOMP-01-0124-FEDER-022701 and the PhD grant with reference SFRH/BD/63646/2009. It is also supported by the Center of Mathematics of the University of Beira Interior, with the project PEST-OE/MAT/UI0212/2011.

8. REFERENCES

[1] Alonso, O., Baeza-Yates, R., and Gertz, M. (2009). Effectiveness of Temporal Snippets. In WSSP'09 - WWW'09. Madrid, Spain.

[2] Alonso, O., Gertz, M., and Baeza-Yates, R. (2011). Enhancing Document Snippets Using Temporal Information.In SPIRE'11. Italy

[3] Berberich, K., Bedathur, S., Alonso, O., and Weikum, G. (2010). A Language Modeling Approach for Temporal Information Needs. In ECIR'10. UK.

[4] Campos, R., Dias, G., Jorge, A. M & Nunes, C. (2012). Enriching Temporal Query Understanding through Date Identification: How to Tag Implicit Temporal Queries? In WWW-TWAW'12, 41-48. France.

[5] Campos, R., Jorge, A., & Dias, G. (2011). Using Web Snippets and Query-logs to Measure Implicit Temporal Intents in Queries. In SIGIR-QRU11, pp. 13 - 16. Beijing, China. July 28.

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

[7] 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.

[8] Freitag, D., Blume, M., Byrnes, J., Chow, E., Kapadia, S., Rohwer, R., et al. (2005). New Experiments in Distributional Representations of Synonymy. In CoNLL'05. Michigan, USA.

[9] Kawai, H., Jatowt, A., Tanaka, K., Kunieda, K., & Yamada, K. (2010). ChronoSeeker: Search Engine for Future and Past Events. In ICUIMC'10.

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

[11] Metzler, D., Jones, R., Peng, F., and Zhang, R. (2009). Improving Search Relevance for Implicitly Temporal Queries. In SIGIR'09. USA.

[12] Pecina, P., and Schlesinger, P. (2006). Combining Association Measures for Collocation Extraction.In COLING/ACL'06. Australia.

[13] Strötgen, J., Alonso, O., & Gertz, M. (2012). Identification of Top Relevant Temporal Expressions in Documents. In WWW-TWAW'12.