Geo-Temporal retrieval filtering versus answer resolution using Wikipedia

Jorge Machado
INESC-ID, Lisbon
Rua Alves Redol 9 Apartado
130691000-029 Lisboa
+351213100300
jorge.machado@estgp.pt

José Borbinha
INESC-ID, Lisbon
Rua Alves Redol 9 Apartado 13069
1000-029 Lisboa
+351213100300
jlb@ist.utl.pt

Bruno Martins
INESC-ID, Lisbon
Rua Alves Redol 9 Apartado 13069
1000-029 Lisboa
+351213100300
bruno.martins@ist.utl.pt

ABSTRACT

We describe an evaluation experiment on GeoTemporal Document Retrieval created for the GeoTime evaluation task of NTCIR 2011. This work describes the retrieval techniques developed to accomplish this task. We describe the collections used in the workshop, detailing the composition of the collections in terms of geographic and temporal expressions. The first contribution of this work is the collections' statistics, which by itself reveals the relevance of this subject. Our parsing techniques found millions of references related with the dimensions of relevance time and space. Those references were used to index the documents in order to score them in those dimensions. We also introduce a technique to find extra references in Wikipedia using Google Search Service and the same parsers used in the collections. Those references were used in four different scenarios depending on the queries: first we used the references found in topics to filter documents without geographic or temporal expressions and used pseudo relevance feedback to expand topics with no references using the indexes created for places and dates; in other approach we used the Wikipedia references to filter documents from the result set, in a last approach we expanded all topics with the Wikipedia references. Finally we used another technique based on metric distances calculated through coordinates (latitudes and longitudes) and dates in order to create a scope for documents and topics, and rank them according to the distance between each other.

Keywords

Geographic and Temporal Information Retrieval, Probabilistic Models, Multidimensional Retrieval Models, Wikipedia.

1. INTRODUCTION

This work was motivated by the GeoTime evaluation task that was part of NTCIR 9 workshop[1]. The GeoTime task aims to evaluate retrieval techniques focusing in geographic and temporal retrieval.

Nowadays, evidences representing semantic dimensions of relevance, such as temporal and geographic evidences, which are extracted from text, are hot topics in information retrieval. Temporal and Geographic dimensions of relevance are two important sources of evidences very common in documents. Extracting those evidences and including them in retrieval models are considered a major step to understand human language and contribute with better searching systems.

To use these evidences it is necessary to extract them from text in documents and, if possible, normalize them in order to make possible a proper use. However it stills very difficult to understand the real meaning of human sentences, and many times the results are not as good as expected in theory. Like in previous year at NTCIR, sources of news articles were used to perform the experiments. This year three new collections were added to the corpus, in a total of almost one million documents, which is already a relevant source for statistical analysis. Our objective for this year's participation was to understand the difference between document filtering using evidences (which we used last year), and the document filtering using features extracted from outer sources (in this case we used Wikipedia). We tried to do that by filtering in two distinct ways. At first place we used probabilistic calculation with BM25 and we filtered documents without the geographic and temporal terms obtained in the topic or in the Wikipedia. In second place we recurred to a common technique where we assigned to each document a bounding box in space and another one in time, using metrical signatures (coordinates in space and dates' intervals in time). To create the bounding boxes we used the terms extracted from Wikipedia, as explained in the follow sections. This way, Wikipedia was used as a pseudo question answer system. The technique was to try to get the best dates and places from Wikipedia documents and use them to either filter the document's result set or to expand the original queries, depending on the run. We assumed that this technique could introduce high noise in the results, but that problem was left for future resolution. Our main objective was to try to understand if the terms extracted from Wikipedia by this technique include or not the relevant dates and places for each topic. We didn't consider natural language processing; we only tried to score Wikipedia sentences with a score function developed for this purpose and then extract the terms. In parallel, we apply the document filtering using the places and the dates automatically found in topics, and also applied filtering/expand queries using the pseudo-relevance feedback technique proposed by Rochio [18]. Like in previous year we used authomatic parsers to extract the geographic and temporal terms from documents.

In this paper we describe an experiment to set some directions in Geo-Temporal retrieval research. In section 2 we describe the entire experiment phase, including the collection processing step of the articles from the test collections New York Times (2002-2005), Xinhua English (1998-2000), Korea Times (1998-2001) and Maniachi Daily (1998-2001). In that same section we also present our first contribution, which is the statistical information about geographic and temporal expressions extraction, and we also detail the documents processing and, finally, the topic processing. In section 3 we discuss the results and in section 4 we conclude and set some future directions.

[1] http://research.nii.ac.jp/ntcir/

2. EXPERIMENTAL HISTORY

Next section 2.1 details the characteristics of the collection, and in section 2.2 we detail the collection processing. The experiment consisted in 5 runs over 25 GeoTemporal topics. The topics are detailed in the overview document for the GeoTime task [15]. Topics consisted in questions mostly using the adverbs *when* and *where,* or providing some geographic references of temporal expressions in form of restrictions for the retrieved documents. In section 2.3 we detail our topic processing step.

We used 3 systems for our experiment. First, we extracted geographic entities using the online service Yahoo PlaceMaker[2]. For temporal expressions extraction we used the TIMEXTAG[3] tagger, developed at University if Amsterdam. The indexes were created with our tool LGTE[4], based on the Lucene text indexer with extensions for probabilistic models, geographic retrieval and hierarchical indexes. For this year we also built a new extractor to obtain the best ranked geographic and temporal terms from Wikipedia's documents.

2.1 Collection Extraction Statistics

The 2011 GeoTime task of NTCIR-9 used a broader corpus of news articles from four collections of 797.216 documents written in English covering dates from January 1998 to December 2005. More details about the collection can be found in the overview documents for the GeoTime task [12] [15].

Geographic and temporal expressions were extracted using, respectively, PlaceMaker and TIMEXTAG. This section is dedicated to report the extracted data in order to validate the relevance of the geo-temporal information used in the experiment. In Tables 2 to 4 we report the extraction with Yahoo PlaceMaker, while in tables 5 to 9 we report the TIMEXTAG extraction of temporal expressions. In Table 1 we summarize the totals of documents with geographic places extracted. In Table 2 we show the place types distribution over the collection, as extracted by PlaceMaker.

Table 1 – Geo-Parsing General Statistics.

	Documents	%
Documents with Places	766.420	96,14%
Documents with no Places found	30.796	0,37%
Total Documents	*797.216*	*100,00%*
Place References	8.315.446	
Average Number of References / Document	10,43060601	

Table 2 – Place types distribution over documents.

Woeid Types	Doc Frequency	References	%References
Town	672.747	3.070.257	36,92%
Country	545.654	3.162.293	38,03%
State	257.234	823.738	9,91%
POI	152.914	358.741	4,31%
Suburb	97.459	184.424	2,22%
County	83.637	165.160	1,99%
Colloquial	51.069	86.397	1,04%
Continent	69.301	128.520	1,55%
Supername	67.306	113.548	1,37%
ZIP	22.623	31.069	0,37%
LandFeature	10.435	17.935	0,22%
Airport	18.454	30.426	0,37%
Island	13.996	22.287	0,27%
HistoricalTown	6.617	11.296	0,14%
Ocean	14.402	19.142	0,23%
Sea	13.042	17.595	0,21%
Drainage	6.454	9.400	0,11%
LocalAdmin	21.210	54.782	0,66%
Miscellaneous	439	720	0,01%
HistoricalState	844	1.288	0,02%
Estate	481	636	0,01%
HistoricalCounty	3.134	4.881	0,06%
DMA	511	859	0,01%
Market	42	44	0,00%
Zone	8	8	0,00%
Total	**2.130.013**	**8.315.446**	*100,00%*

In Table 3 we present the Yahoo confidence degrees. The results show that more than 80% of the extracted places have a degree of confidence higher or equal to 7, in a scale of 1 to 10. This is a good indicator to use geographic entities in this collection.

Table 3 – Yahoo Place Maker confidence degree.

Yahoo Confidence Degree	References	% References
9	3.684.449	44,31%
8	1.372.099	16,50%
10	1.237.413	14,88%
7	698.750	8,40%
6	718.240	8,64%
5	247.207	2,97%
4	144.810	1,74%
3	95.947	1,15%
2	73.144	0,88%
1	43.387	0,52%
Total	*8315446*	*100,00%*

The totals for normalized WOEID (Where on Earth IDentifier) [5]identifiers are found in Table 4. *BelongTos* are the places belonging to the tree of administrative parent regions for one given place (starting in a parent defined as the smaller administrative region containing the given place, following by the smaller administrative region containing the parent of the parent, and so on).

[2] http://developer.yahoo.com/geo/placemaker/

[3] http://ilps.science.uva.nl/resources/timextag

[4] http://code.google.com/p/digmap/wiki/LuceneGeoTemporal

[5] http://developer.yahoo.com/geo/geoplanet/guide/concepts.html

Tables 5 to 9 summarize the collection characteristics in terms of temporal expressions. Table 5 counts the number of temporal expressions (*timexes*) found in collection with TIMEXTAG system.

In Table 6 we display the formats of expressions normalized by TIMEXTAG versus unknown or expressions impossible to normalize.

Time periods are represented in TIDES schema, using the structure PnK, where n is the number of time periods that have passed and K represent days (D), months (M), years (Y) or weeks (W). This kind of expression is followed by an anchor which is a normalized date, and finally a direction that define if the anchor marks the start or the end of the period. As an example consider the duration P2W with anchor 201001 and direction STARTING. This means that the period is the first 2 weeks of January 2010. In Table 7 we present all duration expressions that we were able to expand and index. Expressions "Week of the Year (YYYY-Wn)" are not included in this table. As we can see in the Table 7, this kind of expression is very usual, so we believe we still need new techniques to make a good use of that to improve the retrieval.

Table 4 - Normalized WOEID's.

	Indexed Expressions	References
	Geo	
Place Woeids	104015	8315446
Administrative Scopes Woeid	5974	766420
Geographic Scopes Woeid	7635	766420
BelongTos	88382	97909301
All Woeids	*206006*	*107757587*

Table 5 – Temporal Expressions general statistics.

	Documents	%
Docs with Timexes	783643	98,30%
Docs with no Timexes found	13573	1,70%
Docs with Indexable Time Exprs	770526	96,65%
Docs with no Indexable Time Exprs	26690	3,35%
Docs Failed Anotation	0	0,00%
Docs	*797216*	*100,00%*
Time References	*7079966*	
Total Mapped Temporal Expressions	*9759226*	
Invalid Temporal Expressions	*52*	

Table 6 - Normalized formats statistics.

Temporal Expression Formats Found	Total	%
Y	1.045	0,02%
YY	22.645	0,35%
YYY	74.022	1,13%
YYYY	1.159.229	17,77%
YYYY-MM	430.313	6,60%
YYYY-MM-DD	2.528.222	38,76%
YYYY-Wn	128.008	1,96%
UNKNOWN	2.178.625	33,40%
Found References	*6.522.109*	*100,00%*

Table 7 – Duration expressions expanded and indexed.

Expanded Timexes	Direction	Anchor Format	Timexes
PnD (Starting)	*STARTING*	YYYY-MM-DD	1.399
PnD (Ending)	*ENDING*	YYYY-MM-DD	3.701
PnW (Starting)	*STARTING*	YYYY-Wn	1.462
PnW (Ending)	*ENDING*	YYYY-Wn	5.508
PnM (Starting)	*STARTING*	YYYY-MM	2.979
PnM (Ending)	*ENDING*	YYYY-MM	10.505
PnY (Starting)	*STARTING*	YYYY	13.421
PnY (Ending)	*ENDING*	YYYY	72.869
PtH (Starting)	*STARTING*	hh	1.613
PtH (Ending)	*ENDING*	hh	801
PtM (Starting)	*STARTING*	mm	203
PtM (Ending)	*ENDING*	mm	601
PnDecades (Starting)	*STARTING*	YYY	1.283
PnDecades (Ending)	*ENDING*	YYY	6.881
PnCenturies (Starting)	*STARTING*	YY	236
PnCenturies (Ending)	*ENDING*	YY	435
Distinct Timex2 Durations Found			*9668*
References			*123.897*

Table 8 - Duration expressions not used.

Not Used Timexes	Direction	YYYY	YYYY-MM
TimeExpr/Anchor		Anchor Format	
PnD (BEFORE)	*BEFORE*	YYYY-MM-DD	77.169
PnD (AFTER)	*AFTER*	YYYY-MM-DD	1
PnD (NULL)	*NULL*	UKNOWN	398
PnW (BEFORE)	*BEFORE*	YYYY-Wn	33.019
PnW (NULL)	*NULL*	UNKNOWN	346
PnW (AFTER)	*AFTER*	YYYY-Wn	1
PnM (BEFORE)	*BEFORE*	YYYY-MM	49.916
PnM (AFTER)	*AFTER*	YYYY-MM	1
PnM (NULL)	*NULL*	UNKNOWN	891
PnY (BEFORE)	*BEFORE*	YYYY	191.525
PnY (AFTER)	*AFTER*	YYYY	8
PnY (NULL)	*NULL*	UNKOWN	1.817
PtH (BEFORE)	*BEFORE*	-	27.669
PtH (AFTER)	*AFTER*	-	0
PtH (NULL)	*NULL*		193
PtM (BEFORE)	*BEFORE*	YYYY-MM-DD	33.261
PtM (AFTER)	*AFTER*	YYYY-MM-DD	0
PtM (NULL)	*NULL*	UNKOWN	613
PnDecades (BEFORE)	*BEFORE*	YYY	13.502
PnDecades (AFTER)	*AFTER*	YYY	5
PnDecades (NULL)	*NULL*	UNKOWN	5
PnCenturies (BEFORE)	*BEFORE*	YY	3.620
PnCenturies (AFTER)	*AFTER*	YY	0
PnCenturies (NULL)	*NULL*	UNKOWN	0
Total References			*647.484*

In Table 8 we present the expressions of durations that were not used. Mainly are expressions representing time periods with not well defined limits. For example the expression "before 2010" is true for events that happened in 2008 but also for events that happened in 2000. Probably it is not a good idea to index this kind of expressions with temporal tokens expressing a date because the periods could easily cover very big time intervals. Index the document with temporal limits is probably a better choice but the topic processing must consider that fact and understand the user needs to map those needs to the equivalent query. In this case *a greater than* or *smaller than* query is a possibility. Anyway, this kind of expression was not addressed in our experiment, but is targeted for future research.

Table 9 summarizes the indexed expressions as key points (which are expressions that could be directly normalized to a date using only the document date), generated expressions resulted from event ordering techniques, duration expressions, and finally the documents publishing dates.

Table 9 – Indexed Temporal Expressions

	Temporal		% References
Key Points	19.371	4.343.484	54,76%
GenPoints	8	139	0,00%
Expanded Indexable Time Exprs from Durations	5.628	2.791.092	35,19%
T1 - Indexable Time Expressions	***25.007***	***7.134.715***	***89,95%***
Document DateTime	2.824	797.216	10,05%
T2 - Total Indexable Expressions (include DateTime)	***27.831***	***7.931.931***	***100,00%***

2.2 Collection Processing

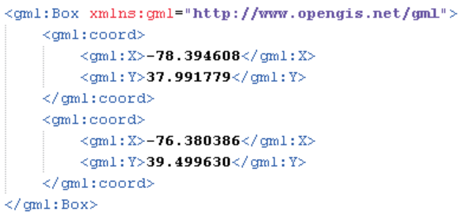
Our experiment aimed to compare the filtering and query expansion approaches using geographic and temporal expressions extracted from documents in three different contexts: first, places and dates in the topic description used to filter documents with the same expressions; second, places and dates obtained from Wikipedia processing after search the topic in Google and restricting results to Wikipedia; finally, use the same Wikipedia references and/or the references parsed from topics description to generate one spatial signature and one temporal signature for the topic. This signature was used to calculate a merged score based in the document signature. This last approach used coordinates and dates distance, as explained in the following sections.

Last year we split the documents in sentences, assuming that each sentence was an independent document. This was done in order to improve the filtering of places and temporal references using a smaller context assuming that relevant reference should only be considered if found inside relevant statements. However this technique did not achieve good results. Many documents had the relevant information spared along different sentences and the score of the sentences containing good places and dates was overlapped with a poor score obtained with BM25 calculated from the text of those the sentences. For that reason, in this year participation, we used the documents as a whole in all runs. We left the sentence splitting because we concluded that the context is very relevant to obtain good documents.

We started our processing by extracting places' references and places' types from the documents (these results are available online, for community[6]) as well as the geographic signature represented as a bounding box with north, south, east, west coordinates represented by latitudes and longitudes on earth surface. We parsed all the documents by the Yahoo Placemaker service available for general use. The representation used for places was the WOEID, for example the WOEID for Lisbon is 2346573, so we created the token "WOEID-2346573" to index it. In a second step of this phase we also extracted the places' parents represented by WOEID's and usually called *belongTos*. This could be important to understand a place's context in the earth surface in order to obtain not only the specific place but also the broader region where it belongs. In a second phase we extracted all the temporal references and structured them in the TIDES schema using the TIMEXTAG system. The temporal references were used when they could be normalized, meanwhile we indexed all temporal references in order to classify a document as having or not temporal references and this was used as a base filter in all runs. We addressed the following types of timexes: time points defined as dates expressed in natural language; generated points defined as relative dates expressed in natural language and normalized using an anchor – that is also a generated or is a time point itself; and weeks defined with the number of the week in year and normalized as a set of days. Finally we also used anchored durations defined as time intervals that were able to be normalized and expanded using an anchor that is the time point assigned by TIMEXTTAG for the start date and the end date. We used day granularity to index all the normalized dates. We used the following format to index dates: YYYY[MM[DD]] (where Y is a digit for year, M for months and D for days; the parenthesis means not mandatory parts). This means that we indexed years using tokens such as "2005" for the year of 2005, or "200511" for the month of November 2005. Period durations were normalized to the equivalent set of dates which could be days, months or years depending on the duration scope. For example a duration referring to the period from 1 January 2010 to 31 January 2010 was normalized to January 2001 using the token "201001". A duration referring to the Week 4 of 2010 was normalized to the equivalent set of days from 18 January to 24 January using the tokens "20100118", "20100119", and so on until "20100124".

Tables 4 and 9 summarize the number of expressions indexed.

The temporal signatures and the geographic signatures were generated in the following way. Geographic signatures were used directly from the bounding box assigned by PlaceMaker. Figure 1 illustrates an example of such a bounding.



```
<gml:Box xmlns:gml="http://www.opengis.net/gml">
    <gml:coord>
        <gml:X>-78.394608</gml:X>
        <gml:Y>37.991779</gml:Y>
    </gml:coord>
    <gml:coord>
        <gml:X>-76.380386</gml:X>
        <gml:Y>39.499630</gml:Y>
    </gml:coord>
</gml:Box>
```

Figure 1 - Example of a bounding box assigned by PlaceMaker.

[6] http://deptal.estgp.pt:9090/collectionsir/GeoTime2011/

To create the temporal bounding boxes we created our own technique. Our signature is an interval of two dates. To find the interval we calculate the standard deviation measure of all dates. The start date is calculated subtracting the standard deviation to the centroide obtained from all dates, the end date is the centroide plus the standard deviation. The following formula defines this signature:

$$\sigma(d) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu(d))^2} \quad \mu(d) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Figure 2 – Standard Deviation equation.

In the equation represented by Figure 2 x_i is one date extracted from a document and is represented in milliseconds, N is the total number of dates found for that document and µ is the centroide of the document obtained as the average of all normalized dates. The interval is defined by the following formula:

$$DateInterval(d) = [\ \mu(d) - \sigma(d)\ ;\ \mu(d) - \sigma(d)]$$

Figure 3 - Date interval assigned to the document.

We also assume that this technique must be improved in the future. A good alternative might be to try to cluster the dates using techniques such as kMeans and try to group spared dates in the document assigning several intervals instead of only a broader one.

We created 4 groups of indexes: Contents, WOEID's, Timexes and Signatures. The coverage of geographic hierarchies was done at index level using the *belongTos* every time a topic asks for some place inside another (e.g all cities in USA, we want documents where USA must be indexed in belongTos index). The coverage of hierarchic dates was done at query level using a wildcard * (e.g. earthquakes in 2002 results in the temporal filter 2002* that will retrieve all documents indexed with temporal expressions started by 2002). The signatures were indexed in three inverted lists: the start date; the end date; and the centroide date. The geographic bounding boxes used 6 indexes: the centroide; the radium in miles; and the geographic coordinates of north, south, east and west limits. The purpose of these indexes is explained in the section 2.4, where we define the formula to calculate geographic distances and temporal distances as a score.

2.3 Topic Processing

Each of the 25 topics of GeoTime had an identifier, a description and a narrative. This processing is structured in four distinct phases. First we parsed the topic with PlaceMaker and TIMEXTAG to extract places an dates. Second we parsed the descriptions with a very simple grammar developed, last year and based in rules to understand if a place or a date were restrictions or information to the query. Third, for those topics without places or dates extracted we used the description to search in the Wikipedia website (using Google) in order to find places and dates to extend the topics. Finally we used the dates to assign a temporal scope to the document using the same technique of standard deviation introduced in the previous section. These phases are detailed in the rest of this section.

The **first phase** is trivial and we gone jump over it. In the **second phase** we parsed the topics with a semi-automatic processor supervised by ourselves. Like in last year participation we aimed to split the topics automatically in three dimensions of relevance: terms, places and times. The terms were obtained removing stopwords and reducing words with a stemming step implemented in the package Lucene SandBox[7] tool that uses the Porter Stemming technique. We extracted places, place types and temporal expressions and we removed them from the keywords query. We also aimed to filter the user needs using restrictions in time and places dimensions. We used the grammar specified in the last year participation to preprocess the 25 topics and create a representation for each one. Generally, the representation of the topic consisted in one filter chain of logic filters and a query part consisting in text, space and time terms. The filter chain aims to represent topic restrictions captured from the text of the topic. This year we used only the topic description to capture restrictions and query terms. We add a new feature to the schema to identify topics where the user asks for the last occurrence of an event. This brought us 5 types of features, as illustrated in Table 10.

Table 10 - Indexed tokens used in filters.

Features	Found values
woeidType	country, city, province
timeType	year, year-month, exact-date, any
place	Yahoo PlaceMaker WOEID references
time	Normalized Expressions found with TIMEXTAG
lastTime	Identification of a user intention of obtain the last event of a subject (*NEW*)

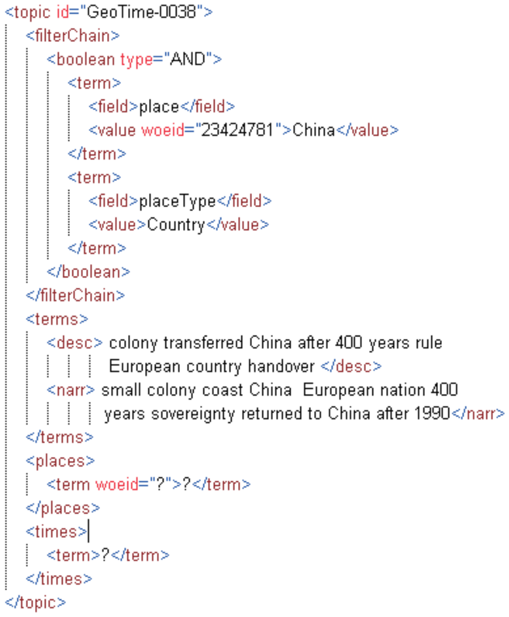
Like in previous participation we captured place names, temporal expressions, place types and temporal expressions types. We refer for the previous paper to detail the objective of these features [16].

All terms found using the previous technique, including adverbs in questions, user references, places, times, places properties and time properties, were removed from the text fields description and placed in filters as geographic or temporal terms filters. Places' names and normalized dates references not considered by the previous set of rules were removed from the terms fields description and placed in their own dimensions of relevance queries. We also removed the stopwords and punctuation characters. The follow example illustrates the topic GeoTime-0038 (Note that the question mark in places and dates means that the user wants to find those features; Those fields only have places instead of question mark if the user specifies a place in topic that is considered a restriction but instead is a complement for the topic itself for example we refer topic 49):

"The International Library of Children's Literature is a part of the National Diet Library of Japan."

In the sentence "Japan" is not near an expression of the kind "the user wants events occurred in Japan" or "Where in Japan …" or "When … in Japan". In this case the place "Japan" is considered by our parser as a complement to the topic, so we used as a place query instead of a place filter.

[7] http://lucene.apache.org/java/2_3_2/lucene-sandbox/index.html



```
<topic id="GeoTime-0038">
    <filterChain>
        <boolean type="AND">
            <term>
                <field>place</field>
                <value woeid="23424781">China</value>
            </term>
            <term>
                <field>placeType</field>
                <value>Country</value>
            </term>
        </boolean>
    </filterChain>
    <terms>
        <desc> colony transferred China after 400 years rule
                European country handover </desc>
        <narr> small colony coast China  European nation 400
               years sovereignty returned to China after 1990</narr>
    </terms>
    <places>
        <term woeid="?">?</term>
    </places>
    <times>
        <term>?</term>
    </times>
</topic>
```

Figure 4 - Topic parsed and structured.

For such topics we created a set of filters including a base filter to remove documents without temporal and geographic references, which is represented with the question mark.

In the **third phase** of the processing we used an algorithm developed to found dates and places relevant to the subject of the topic using Wikipedia. The package is available in the LGTE repository for general use[8]. The algorithm is structured in the following steps:

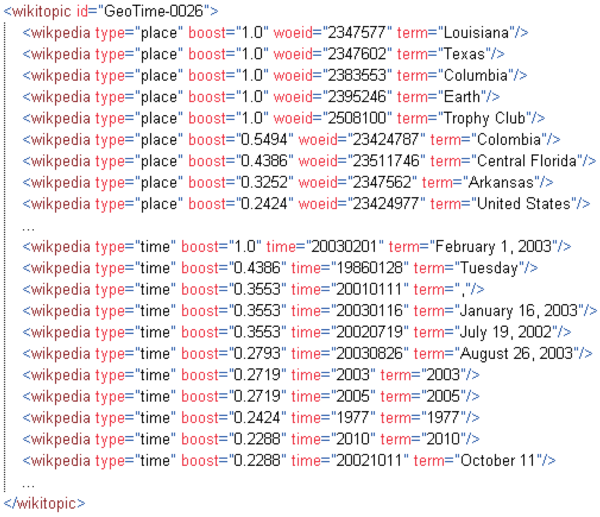
1. Use the topic to search in Google restricting results to *site:en.wikipedia.org;*
2. Obtain the first N documents from result set;
3. Geoparse the document with PlaceMaker and extract temporal expressions with TIMEXTAG;
4. Obtain the first P paragraphs from each document;
5. Create a temporary collection of paragraphs and index them in order to score the paragraphs with BM25. Each paragraph includes also the title of the Wikipedia document in order to give additional context to the paragraph.
6. Take out paragraphs with zero places or dates;
7. Score the paragraphs;
8. Create a list of place references and date references together with the score assigned to the paragraph;

Formalizing the paragraph scoring is given by the following equation:

$$Score\begin{pmatrix} w, \\ p, \\ q, \\ posWG, \\ posP \end{pmatrix} = bm25\begin{pmatrix} q = terms\ in\ topic\ desc, \\ p = title(w) \cup text(p), \\ C = \sum_{p=1}^{P} p_i) \end{pmatrix} . \alpha^{posWG} . \beta^{posP}$$

In this equation *w* is the Wikipedia document, *p* is the paragraph being ranked, *q* is the description of the topic, *posWG* is the position of the Wikipedia document in Google results list, and *posP* is the position of *p* in Wikipedia document starting in 0 for the first paragraph. The *bm25* is calculated considering each paragraph *p* as a document and the virtual collection C is the entire set of paragraphs *p* in the same document. In the formula the *bm25* is calculated using the terms in query *q* scored considering the paragraph *p* and the entire collection C of P paragraphs. The parameters α and β are decay factors, in that sense the score decreases as big is the position in Google and as big is the position of the paragraph list in the Wikipedia document. We used α = β = 0.8. We used a maximum of 5 documents and the 5 top paragraphs with expressions (dates and/or places) in each document. We also used a score threshold of 0.2 for paragraphs.

To illustrate the output of this phase we show Figure 5 an example of found terms for topic 26. The topic was "*Where and when did the space shuttle Columbia disaster take place?*". As we show, the expansion returns very good keywords in space and time. The disasters took place over Texas and were found parts of the shuttle also in Louisiana. The exact date was 1 February 2003 and that was the top scored expression. Most topics returned good expressions. But we also found much noise in the middle.

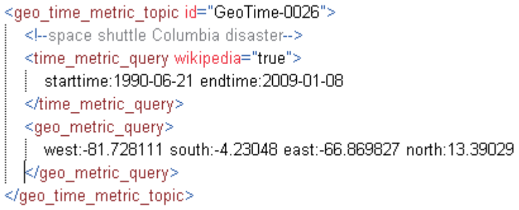


```
<wikitopic id="GeoTime-0026">
    <wikpedia type="place" boost="1.0" woeid="2347577" term="Louisiana"/>
    <wikipedia type="place" boost="1.0" woeid="2347602" term="Texas"/>
    <wikipedia type="place" boost="1.0" woeid="2383553" term="Columbia"/>
    <wikipedia type="place" boost="1.0" woeid="2395246" term="Earth"/>
    <wikipedia type="place" boost="1.0" woeid="2508100" term="Trophy Club"/>
    <wikipedia type="place" boost="0.5494" woeid="23424787" term="Colombia"/>
    <wikipedia type="place" boost="0.4386" woeid="23511746" term="Central Florida"/>
    <wikipedia type="place" boost="0.3252" woeid="2347562" term="Arkansas"/>
    <wikipedia type="place" boost="0.2424" woeid="23424977" term="United States"/>
    ...
    <wikipedia type="time" boost="1.0" time="20030201" term="February 1, 2003"/>
    <wikipedia type="time" boost="0.4386" time="19860128" term="Tuesday"/>
    <wikipedia type="time" boost="0.3553" time="20010111" term=","/>
    <wikipedia type="time" boost="0.3553" time="20030116" term="January 16, 2003"/>
    <wikipedia type="time" boost="0.3553" time="20020719" term="July 19, 2002"/>
    <wikipedia type="time" boost="0.2793" time="20030826" term="August 26, 2003"/>
    <wikipedia type="time" boost="0.2719" time="2003" term="2003"/>
    <wikipedia type="time" boost="0.2719" time="2005" term="2005"/>
    <wikipedia type="time" boost="0.2424" time="1977" term="1977"/>
    <wikipedia type="time" boost="0.2288" time="2010" term="2010"/>
    <wikipedia type="time" boost="0.2288" time="20021011" term="October 11"/>
    ...
</wikitopic>
```

Figure 5 - Topic expansion using wikipedia for topic 26.

In **phase four** we assigned a temporal scope to the topics in order to use it in our run 5. The scope was assigned using the standard deviation already explained in previous sections, using the dates found in Wikipedia and in the topic. In Figure 6 we illustrate the temporal scope assigned to topic 26, as well as the geographic bounding box assigned by Yahoo PlaceMaker.

[8] http://code.google.com/p/digmap/wiki/LuceneGeoTemporal



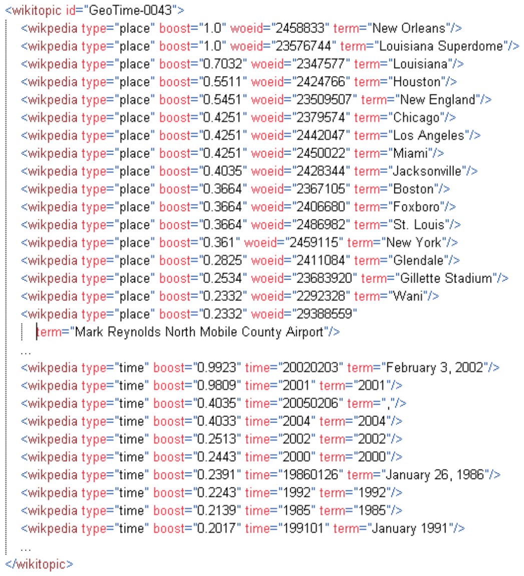
```
<geo_time_metric_topic id="GeoTime-0026">
   <!--space shuttle Columbia disaster-->
   <time_metric_query wikipedia="true">
      starttime:1990-06-21 endtime:2009-01-08
   </time_metric_query>
   <geo_metric_query>
      west:-81.728111 south:-4.23048 east:-66.869827 north:13.39029
   </geo_metric_query>
</geo_time_metric_topic>
```

Figure 6 – Temporal generated for topic 26 using standard deviation.

Analyzing this topic reveals one source of problems. As we can see the deviation is very big, from 1990 until 2009, which was due to the noise obtained from Wikipedia. Besides the standard deviation purpose, some dates were very distant of the real one and that leads to very big standard deviations. In this case the dates introducing the noise were 1986 and 1977. This will need to be addressed in future. The geographic scope is returned by the Yahoo Placemaker service using as text the top 5 scored paragraphs of Wikipedia document.

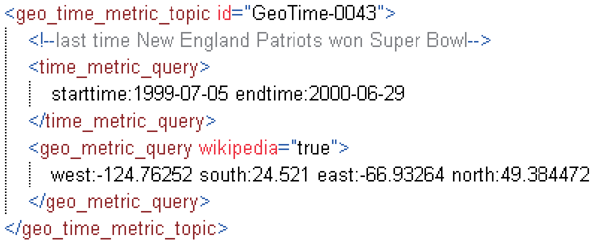
Another example is the Topic 43, "*When was the last time the New England Patriots won the Super Bowl*". The Patriots won the SuperBowl in: 2002 in New Orleans (Louisiana), 2004 in Huston (Texas) and 2005 in JaksonVille (Florida). All of them were returned by the system as shown in Figure 7.

Once more the temporal scope was not very good as we show in Figure 8.



```
<wikitopic id="GeoTime-0043">
   <wikpedia type="place" boost="1.0" woeid="2458833" term="New Orleans"/>
   <wikpedia type="place" boost="1.0" woeid="23576744" term="Louisiana Superdome"/>
   <wikpedia type="place" boost="0.7032" woeid="2347577" term="Louisiana"/>
   <wikpedia type="place" boost="0.5511" woeid="2424766" term="Houston"/>
   <wikpedia type="place" boost="0.5451" woeid="23509507" term="New England"/>
   <wikpedia type="place" boost="0.4251" woeid="2379574" term="Chicago"/>
   <wikpedia type="place" boost="0.4251" woeid="2442047" term="Los Angeles"/>
   <wikpedia type="place" boost="0.4251" woeid="2450022" term="Miami"/>
   <wikpedia type="place" boost="0.4035" woeid="2428344" term="Jacksonville"/>
   <wikpedia type="place" boost="0.3664" woeid="2367105" term="Boston"/>
   <wikpedia type="place" boost="0.3664" woeid="2406680" term="Foxboro"/>
   <wikpedia type="place" boost="0.3664" woeid="2486982" term="St. Louis"/>
   <wikpedia type="place" boost="0.361" woeid="2459115" term="New York"/>
   <wikpedia type="place" boost="0.2825" woeid="2411084" term="Glendale"/>
   <wikpedia type="place" boost="0.2534" woeid="23683920" term="Gillette Stadium"/>
   <wikpedia type="place" boost="0.2332" woeid="2292328" term="Wani"/>
   <wikpedia type="place" boost="0.2332" woeid="29388559"
      term="Mark Reynolds North Mobile County Airport"/>
   ...
   <wikpedia type="time" boost="0.9923" time="20020203" term="February 3, 2002"/>
   <wikpedia type="time" boost="0.9809" time="2001" term="2001"/>
   <wikpedia type="time" boost="0.4035" time="20050206" term=","/>
   <wikpedia type="time" boost="0.4033" time="2004" term="2004"/>
   <wikpedia type="time" boost="0.2513" time="2002" term="2002"/>
   <wikpedia type="time" boost="0.2443" time="2000" term="2000"/>
   <wikpedia type="time" boost="0.2391" time="19860126" term="January 26, 1986"/>
   <wikpedia type="time" boost="0.2243" time="1992" term="1992"/>
   <wikpedia type="time" boost="0.2139" time="1985" term="1985"/>
   <wikpedia type="time" boost="0.2017" time="199101" term="January 1991"/>
   ...
</wikitopic>
```

Figure 7 - Topic expansion using wikipedia for topic 43.



```
<geo_time_metric_topic id="GeoTime-0043">
   <!--last time New England Patriots won Super Bowl-->
   <time_metric_query>
      starttime:1999-07-05 endtime:2000-06-29
   </time_metric_query>
   <geo_metric_query wikipedia="true">
      west:-124.76252 south:24.521 east:-66.93264 north:49.384472
   </geo_metric_query>
</geo_time_metric_topic>
```

Figure 8 - Temporal generated for topic 43 using standard deviation.

2.4 Runs Description

We participated in GeoTime with 5 distinct strategies, all of them based on BM25. We used filters or query expansion when no filters were defined. If the topic requested places and/or dates, which was allays true, we also used a base filter to remove documents without geographic or temporal expressions. If the extracted expressions were not considered as filters, then they were used as query terms. The keywords component of the query was built using twice the description, in order to increase its discriminatory power, while the narrative was not used. In this sense, our term queries were composed by keywords, places and times using for that purpose several indexes in order to obtain a unique score. The BM25 model was used considering independent indexes. In first place we calculated the score of each term in its index, and in second we summed all term scores to obtain the document score. We also used boost factors in each dimension of relevance like in last year participation, but this year we used a normalized score from 0 to 1 in each index (places, dates and text). This was done for all runs and bellow we detail each particular strategy.

- INESC-EN-EN-01-D - Our first run used documents description index and the base filter to remove documents without geographic or temporal expressions, depending on the topic restrictions.
- INESC-EN-EN-02-D - Our second run used document description index, the base filter and the filters defined in topic processing using places and dates or query expansion if no filters were defined.
- INESC-EN-EN-03-D - Our third run used the places and dates found in Wikipedia to expand the query.
- INESC-EN-EN-04-D - Our third run was similar to the second one but instead of using only the places and dates in the topic we also used the terms found in Wikipedia to filter results without those terms.
- INESC-EN-EN-05-DN – Our last run used a score based on distances obtained from the document scope and from the query scopes.

2.4.1 Filters and Query Expansion in RUN2

This subsection summarizes the technique used in run 2 already used in the last year. The base filter made use of one index created to mark geo-temporal documents.

The geographic restrictions filters used three geographic filters: *places, belongTos* and *placeTypes* containing the types of places detailed in Table 2.

For temporal restrictions we used two indexes, *timeExpressionsFormat* and *timePoints*. We also used the index *expandedTimeDurations* for query expansion together with all the others. However we didn't use this last index for filtering purposed once this filter was made from generated dates, an interval of two dates, a start and an end date, a week, a decade or a century will result in several dates representing it. The first one indexed the formats presented in Table 6, while the second indexed the key points of Table 9 and the third one indexed expanded time expressions that we generated from time periods, including weeks and expressions described in Table 7.

For topics with zero filters and zero geographic and temporal expressions we used blind relevance feedback query expansion. We based our method in Rochio algorithm, with modifications to use multiple indexes. This technique is detailed in [14]. We considered the geographic indexes *belongTos* and *places* with relative weights of 0.3 and 0.7 respectively. This was done because we think that *belongTos* is good for filtering but not so good for expansion because it is an extension that could not be in the original text (we are only supposing that a document talking about Texas could be talking about USA, and thus put USA in the expansion means that we are trying to retrieve every documents talking about USA). We used temporal indexes *timePoints* and *expandedTimeDurationswith* relative weights of 0.7 and 0.3 respectively. We used a maximum number of 15 terms in the first 5 documents with a decay factor of 0.15. The topics where we used the query expansion were 32, 33, 37, 40, 43, 46, 48.

2.4.2 Wikipedia Terms

The terms found in Wikipedia were used as expansion terms in run3 using as boost factor the score obtained from the extraction which was the score of the paragraph. In run4 we used the terms as filters. Was created a sequential OR filter with all terms obtained from Wikipedia. Note that we only used dates and places. An example of an expanded query is given below(the queries were very big, so we exemplify with a small and incomplete query):

```
t_point:(2010*^0.5)^0.7 t_duration:(2010*0.6)^0.3
(g_place:(WOEID-2347577)^1.0)^0.7 g_belongTos:(WOEID-2346573^1.0)^0.3
```

In the query the first factor is the score from Wikipedia and the 0.7 and 0.3 are the broader factors assigned to those indexes for the reasons explained in previous subsection. *Durations* are less important than *time points* and *belongsTos* are less important than *places*. The internal factor of 0.5 assigned to 2010 is illustrative of a Wikipedia term of a paragraph with score 0.5 assigned by our algorithm introduced in section 2.3, the same is valid for the WOEID's.

We used another factor to increase the power of recent dates when the query contains expressions like "When was **the last time** the New England Patriots". We explain this factor in the next subsection, since the technique used was the same that we used for metric queries. The two topics where this technique was applied were topics 27 and 43.

2.4.3 Metric

Finally, we conclude by introducing the technique to score the temporal and geographic scopes of the queries and documents. This technique is similar to the one used by our group in previous works, but we used now a smoother approach. The following equation defines an S curve to score documents based on the document centroide of the bounding box, the centroide given by the query and the diagonal distance between the edges of the query.

$$f(d,r) = \frac{1}{1+(\frac{\alpha d}{r}\,e)^{\beta}}$$

In the equation d is the distance between centroides (query and document) and the radium is given by twice the diagonal of the query. In geographic queries the diagonal was the distance between the northwest point and the southeast point of the query. To calculate temporal score we used the end date minus the start date of the query as radium, in this sense we are admitting documents placed at twice the radium from the centroide of the query (in other words we are admitting documents out of the start date).

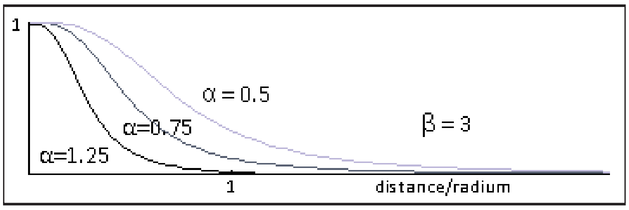


Figure 9 - S curve to calculate geographic and temporal scores.

Our smooth parameters α and β were used to tune the curve. We used β = 3 and α = 0.5.

We must go now back to section 2.4.2 to explain a situation where this curve was used to change the boost factor assigned to dates in run 3. This curve was also used in other means to improve results where the topic asked for the last occurrence of an event. In those cases we used the curve to calculate an extra boost factor for dates. The last date found in expansion was assigned an extra boost factor of 1 and for older dates we used this curve to smooth the other factor by multiplying them by this one. The radium was calculated as the difference of years from the last date that is smaller than the query date and the older date. With this we could increase the boost factor of recent dates found in Wikipedia.

3. RESULTS

We present the formal results means in Table 11 and the analysis per topic in Figure 10. The best run was again BM25 using the description of the document and filtering those documents without temporal expressions or geographic expressions depending on the query. Meanwhile Wikipedia filtering run 4 and Wikipedia Expansion run3 returned very close results. This make us believe that using an optimized tuning and probably improving the technique of choosing the geo and temporal expressions could perform better.

Table 11 – Formal Metrics Means

RunName	MAP	Q	nDCG@10	nDCG@100	nDCG@1000
INESCID-EN-EN-01-D	0,3260	0,3497	0,4591	0,4563	0,5791
INESCID-EN-EN-02-D	0,2006	0,2093	0,3805	0,3284	0,3684
INESCID-EN-EN-03-D	0,3200	0,3362	0,4499	0,4550	0,5224
INESCID-EN-EN-04-D	0,3027	0,3183	0,4457	0,4378	0,5110
INESCID-EN-EN-05-DN	0,1392	0,1474	0,3137	0,2493	0,3195

Our run5 does not perform well as we expected in the first place because of the technique used to assign a bounding box in time.

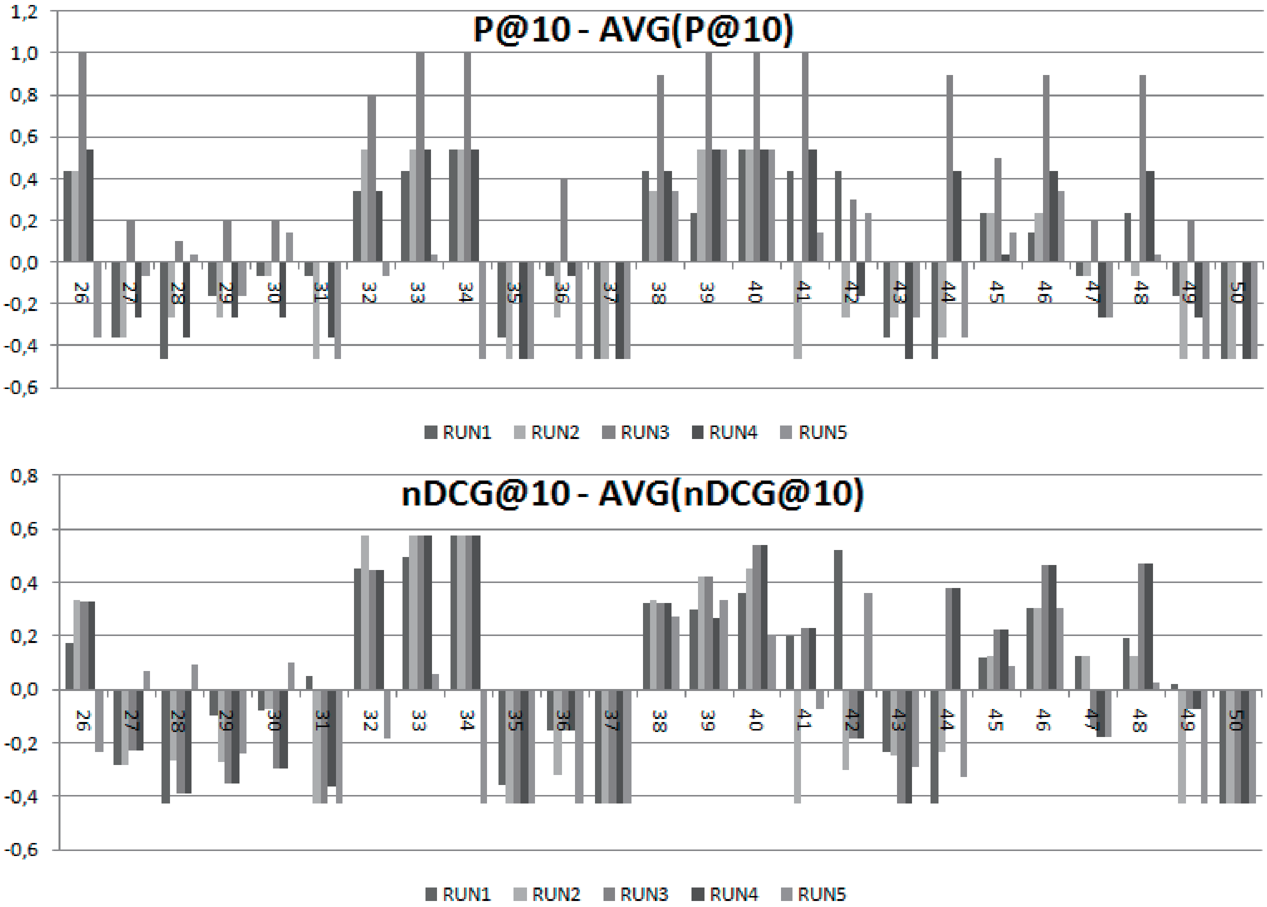


Figure 10 - Results per topic.

We need to review this technique and make it better as we explained in the previous section. Considering the noise introduced by Wikipedia, we could now compare the set of topics expanded with Wikipedia and with Rochio pseudo relevance feedback. In run 2 only the topics without filters assigned (32, 33, 37, 40, 43, 46, 48) used Rochio, in run 3 all topics used Wikipedia query expansion. So these set 7 topics are the eligible to compare the two approaches. As we can see, run 3 is much better than run 2. Rochio performs with a better nDCG in 32 but a worst P@10. The same was true for topic 33. In 37 they are similar, but in 40, 43, 46 and 48 the Wikipedia expansion is much better. We shall consider that Rochio is a generally accepted and good method, but the use of Wikipedia has a big potential but in order to take advantage of it, we must reduce its weaknesses, such as the noise.

Comparing the same set of topics in runs 3 and 4, we note that the results are also better for expansion (3) than for filtering (4), meanwhile the filtering using Wikipedia is not worst that the expansion with Rochio, this means that the filter and expansion are probably candidates for a merged approach. The true is that expansion and filtering probably will grow together with future improvements.

Considering now run 5 we must refer topic 42 "*Middle East King died 1999*". This topic has a reference date. In cases like this we didn't use Wikipedia terms to generate the temporal scope, we only used the query terms because it makes sense that if we already have the time, we don't need to introduce noise. The point is that the temporal scope assigned by our method was 1998-2000 and only the run 1 and 5 obtained good results considering the nDCG measure. This means that the metric distance between the scopes assigned to documents and the temporal scope assigned to the topic really work in some cases. The problem in the other topics was the scope assigned to the query using Wikipedia terms. The results made us to analyze several documents to find if the temporal scope assigned by standard deviation was valid. We choose randomly 30 documents and we found that 25 of them had very good temporal scopes assigned. Note that we did not use the document date, only the speech references, some of them of course anchored in document date. We think that this happens because news articles are probably a much more focused source of temporal information than Wikipedia which in other hand tries to relate many events in the same document. We cannot prove, because is only one topic, but we believe that metric scores calculated with this kind of S curves based on distance could perform very well when the date is given in the topic and the documents are news articles.

4. CONCLUSIONS

This work aimed to contribute with another step in the quest of GeoTemporal retrieval. We introduced three new approaches to score documents using geo-temporal information: document filtering, query expansion and metric distances converted to scores We used several methods already known and used by the retrieval community such as Wikipedia query expansion, S-curves based on distances, boost factors, geographic hierarchical information. We found that all of them are good when correctly used: Metric distances are good when queries define the wanted date; topics without references to places or dates could be expanded by scoring Wikipedia paragraphs and extract the references inside; and finally query expansion using those references is better than filter the documents without them, probably due to boost factors that are not used in filtering.

We believe that the main challenge in future is not to find new techniques but instead find ways to choose and combine the technique given the topic where user describes its needs.

5. ACKNOWLEDGEMENTS

This work was partially supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds. We would also like to thanks to Professor Pável Calado and the students Pedro Barros and Irina Lopes, from the INESC-ID / Technical University of Lisbon at IST, for their valuable contribution in the topic development, and also to students Duarte Santos and Bruno Crespo, from the ESTG / Polytechnic Institute of Portalegre for their valuable contribution in the assessments for the GeoTime English Task.

6. REFERENCES

[1] D. Ahn, J. van Rantwijk, and M. de Rijke. A Cascaded Machine Learning Approach to Interpreting Temporal Expressions. In: Proceedings NAACL-HLT 2007.

[2] E. Saquete, R. Muñoz , P. Martínez-Barco. Event ordering using TERSEO system. Special issue: Application of natural language to information systems (NLDB04). Pages: 70 – 89, 2006.

[3] Omar Alonso, Michael Gertz, Ricardo Baeza-Yates. On the value of temporal information in information retrieval. ACM SIGIR Forum. Volume 41, Issue 2 (December 2007).

[4] Jorge Machado, Bruno Martins e José Borbinha, "LGTE: Lucene Extensions for Geo-Temporal Information Retrieval". GIIW, ECIR, Toulouse, 2009.

[5] T. Mandl et al. "An evaluation resource for Geographical Information Retrieval". In Proceedings of the 6th International Conference on Language Resources and Evaluation, 2008.

[6] F. Gey, at al. "GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview". In A. Nardi, C. Peters, and J. L. Vicedo, editors, Cross Language Evaluation Forum: Working Notes for the CLEF 2006 Workshop

[7] Ray R. Larson. "Cheshire at GeoCLEF 2008: Text and Fusion Approaches for GIR", GeoCLEF 2008, CLEF 2008, Aarhus, Denmark, September 17-19, 2008.

[8] D. Ferres, H. Rodríguez. "TALP at GeoCLEF 2007: Results of a Geographical Knowledge Filtering Approach with Terrier". CLEF, Budapest, Hungary, 2007

[9] Nuno Cardoso, Patrícia Sousa and Mário J. Silva, "The University of Lisbon at GeoCLEF 2008" GeoCLEF 2008, CLEF 2008, Aarhus, Denmark, September 17-19, 2008.

[10] Ferro Lisa, Inderjeet Mani, Beth Sundheim and George Wilson. TIDES Temporal Annotation Guidelines. Version 1.0.2 MITRE Technical Report, MTR 01W0000041. 2001.

[11] Roser Saur´ý, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky. TimeML Annotation Guidelines. Version 1.2.1. January 31, 2006

[12] Fredric Gey†, Ray Larson, Noriko Kando, Jorge Machado, Tetsuya Sakai. NTCIR-GeoTime Overview: Evaluating Geographic and Temporal Search. NTCIR, Tokyo, Japan, 2010.

[13] James F. Allen. Towards a general theory of action and time. Artificial Intelligence 23, 2, July 1984.

[14] Jorge Machado, Bruno Martins and José Borbinha. Experiments with N-Gram Prefixes on a Multinomial Language Model versus Lucene's off-the-shelf ranking scheme and Rocchio Query Expansion (TEL@CLEF Monolingual Task). ECDL/CLEF, Corfu, Greece, 2009.

[15] Fredric Gey†, Ray Larson, Noriko Kando, Jorge Machado, Masaharu Yoshioka. NTCIR9-GeoTime Overview - Evaluating Geographic and Temporal Search: Round 2. GeoTime, NTCIR-9, Tokyo, Japan, 2010.

[16] Jorge Machado, Bruno Martins and José Borbinha. GEOTIME: Experiments with Geo-Temporal Expressions Filtering and Query Expansion at Document and Phrase Context Resolution. Proceedings of NTCIR-8 Workshop Meeting, June 15–18, 2010, Tokyo, Japan.

[17] Porter, M. F. (1980). "An algorithm for suffix stripping". In: Sparck Jones, K. & Willett, P. (eds.), Readings in Information Retrieval., pp. 313 - 316. San Francisco: Morgan Kaufmann.

[18] Rocchio, J. J.: Relevance Feedback in Information Retrieval: In: The SMART Retrieval System. Experiments in Automatic Document Processing: pp 313 - 323. Prentice Hall. (1971)