Temporal-Semantic Clustering of Newspaper Articles for Event Detection

Aurora Pons-Porrata[1], Rafael Berlanga-Llavori[2] and José Ruiz-Shulcloper[3]

[2] Universitat Jaume I, Castellón (Spain)
[1] Universidad de Oriente, Santiago de Cuba (Cuba)
[3]Institute of Cybernetics, Mathematics and Physics, La Habana (Cuba)

Abstract. In this paper we introduce a new clustering algorithm for event detection in newspaper articles, which has two main features. Firstly, it makes use of the temporal references extracted from the document texts to define the document similarity function. Secondly, the algorithm works hierarchically. In the first level, documents with a high temporal-semantic similarity are grouped into individual events by applying the proposed similarity functions. In the next levels, these events are successively grouped so that more complex events and topics can be identified. The resulting hierarchy describes the structure of topics and events taking into account their temporal occurrence. These tasks cannot be currently accomplished by current Topic Detection and systems.

1. Introduction

Starting from a continuous stream of news, the event detection problem consists in determining for each incoming document whether it reports on a new event, or it belongs to some previously identified event. One of the crucial issues in this problem is to define what an *event* is. Initially, an event can be defined as something that happens at a particular place and time. However, many events can occur over several places and several time periods (e.g. a trial event). This is why many researchers of the field prefer the broader concept of *topic*, which designates an important event or activity along with all its directly related events. In fact, this research area is called "Topic Detection and Tracking" (TDT).

A TDT system is intended to discover the topics reported in the news and to organize them into these topics. Current systems mainly rely on automatic document classifiers that mainly come from the Information Retrieval field. For example, the systems proposed in [Papk99] and [Yang00] use the *Single-Pass* algorithm to detect and track news. The system presented in [Wall99] adopts a variation of the *K-means* algorithm that does not require fixing the final number of document clusters. Finally, the work in [Carb99] applies the Fractionation [Cutt92] and GAC (Group Average Clustering) algorithms to group the news that fall into a specific temporal window. All these works have in common that they use both a document clustering algorithm, preferably with near linear cost, and the publication dates of the news. This attribute is used to define classifier thresholds that depend on the temporal adjacency of the news. As proved, this property improves notably the results of these systems.

Our main interest is focused on discovering events and topics as well as how they are structured. In the context of the TDT area, we are interested in identifying not only the topics but also the possible smaller events they comprise. In our opinion, the temporal properties of news must be further exploited to achieve these purposes. Specifically, we think that regarding the time references appearing in the news texts, those documents that report about the same event can be better grouped.

In this paper we propose a hierarchical clustering algorithm for event detection that makes use of the temporal references extracted from the document texts to define a new document similarity. In the first level, documents with a high temporal-semantic similarity are grouped together by using the proposed similarity function. In the next levels, these events are successively grouped so that more complex events and topics can be identified. The resulting hierarchy describes the structure of topics and events taking into account their temporal occurrence.

2. Document Representation

The incoming stream of documents that feed our system come from some on-line newspapers available in Internet, which are automatically translated into XML (eXtended Markup Language). These XML documents preserve the original logical structure of the newspapers, so that different thematic sections can be distinguished as well as the different parts of the news (e.g. the title, authors, place, date of publication, etc.). Each news document is represented with the following feature vectors:

- A vector of weighted terms. Each term is a reduced form of a set of words appearing in the text. For instance, in Spanish all the verb tenses are represented by their infinitive (reduced form). Stop words, such as articles, prepositions and adverbs, are removed from this vector. Terms are statistically weighted using the normalized term frequency (TF). It is worth mentioning that we have not used the Inverse Document Frequency (IDF) in the weight scheme because it requires either a training corpus or an incremental adjustment.
- A vector of weighted time entities. A time entity can be either a date or a date interval expressed in the Gregorian Calendar. These dates are automatically extracted from the news texts by using the algorithm presented in [Llid01]. This algorithm applies shallow parsing techniques to detect temporal sentences and then, translates them into time entities of a time model. Time entities are statistically weighted using the frequency of their references in the text. All the time entities that are separated from the publication date more than ten days are removed from this vector. Moreover, the time entities whose frequency is smaller than a tenth part of the maximum in the vector are also removed. All these references are removed because they are considered marginal with respect to the document event.

From the above definitions, each news document is represented as follows:

- A vector of terms $T^i = (TF_1^i, \ldots, TF_n^i)$, where TF_k^i is the relative frequency of term t_k in the document d^i.
- A vector of time entities $F^i = (TF_{f_1^i}, \ldots, TF_{f_m^i})$, where $TF_{f_k^i}$ is the absolute frequency of the time entity f_k in the document d^i.

3. Document Similarity Measures

Every automatic clustering of documents, like event detection, depends on a similarity measure. Most of the document clustering algorithms presented in the literature use the cosine measure, which computes the cosine of the angle between the document vectors. In our approach, besides this measure, we need to introduce new partial similarity measures for each document component, more specially for the vector of dates. In this section we present two global similarity measures, which take into account the proposed partial similarity measures.

Let d^i and d^j be two documents. For the component of the terms we will use the cosine measure:

$$S_T(d^i, d^j) = \frac{\sum_{k=1}^{n} TF_k^i \cdot TF_k^j}{\sqrt{\sum_{k=1}^{n} TF_k^{i^2}} \cdot \sqrt{\sum_{k=1}^{n} TF_k^{j^2}}}$$

where n is the total number of terms in the document collection ζ.

For the temporal component, we define two different measures. The first one is based on the cosine measure, but it is weighted by the temporal proximity between the vectors of time entities. The main problem of this measure is the different size of these vectors and the different nature of the elements that they contain (dates and date intervals). This similarity measure is defined as follows:

$$S_F(d^i, d^j) = \frac{\sum_{k=1}^{m_i} TF_{f_k^i} \cdot TF_{s(f_k^i, d^j)} \cdot g(f_k^i, s(f_k^i, d^j)) + \sum_{k=1}^{m_j} TF_{f_k^j} \cdot TF_{s(f_k^j, d^i)} \cdot g(f_k^j, s(f_k^j, d^i))}{(2 + |m_i - m_j|) \cdot \sqrt{\sum_{k=1}^{m_i} TF_{f_k^i}^2} \cdot \sqrt{\sum_{k=1}^{m_j} TF_{f_k^j}^2}}$$

Here m_i is the number of time entities that describe the document d^i, $s(f_k^i, d^j)$ is a function that returns the most similar time entity to f_k^i (the one of minimum distance) that occurs in the document d^j. Let f_1 and f_2 be two time entities, the penalty function g is defined as follows:

$$g(f_1, f_2) = \begin{cases} 1 & if\ \ f_1 = f_2 \\ 0.8 & if\ dist(f_1, f_2) = 1 \\ \dfrac{1}{\sqrt{dist(f_1, f_2)}} & otherwise \end{cases}$$

The distance function *dist* between two time entities f_1 and f_2 depends on their nature, that is, if they are dates or date intervals. This function is defined as:

- If f_1 and f_2 are dates, then $dist(f_1, f_2)$ is the number of days between f_1 and f_2.
- If $f_1 = [a , b]$ and $f_2 = [c , d]$ are date intervals, then

  $$dist(f_1, f_2) = |f_1 \oplus f_2| - |f_1 \otimes f_2| + 0.2 \cdot (2 \cdot |f_1 \otimes f_2| - |f_1| - |f_2|),$$

  where the union interval $f_1 \oplus f_2 = [\min\{a,c\}, \max\{b,d\}]$, $f_1 \otimes f_2$ is the intersection interval and the cardinality of the interval $[l, u]$ is the number of days between the dates l and u. This metric is based on the generalized metric of Minkowski, defined in [Ichi94] for features of interval type in symbolic objects.
- If f_1 is a date and f_2 is an interval, then $dist(f_1, f_2) = dist([f_1, f_1], f_2)$.
- If f_2 is a date and f_1 is an interval, this function is defined in a similar way.

The other proposed measure for temporal component is based on the traditional distance between sets. This measure is defined as:

$$D_F(d^i, d^j) = \min_{f^i \in FR^i, f^j \in FR^j} \left\{ d(f^i, f^j) \right\}$$

where $d(f^i, f^j)$ is the distance between the dates f^i and f^j and FR^i is the set of all dates f^i that satisfy the following conditions:

- each f^i has the maximum frequency in d^i, that is, $TF_{f^i} = \max_{k=1,\ldots,m_i} \left\{ TF_{f_k^i} \right\}$
- and each f^i has the minimum distance to the publication date of d^i.

The second condition is not considered when comparing cluster representatives instead of documents. It is not difficult to see that the set FR^i is not necessarily unitary. The distance *d* is defined as follows:

- If f^i and f^j are dates, then $d(f^i, f^j)$ is the number of days between f^i and f^j.
- If $f^i = [a , b]$ and $f^j = [c , d]$ are date intervals, then

  $$d(f^i, f^j) = \min_{f_1 \in [a,b], f_2 \in [c,d]} \{ d(f_1, f_2) \}$$

- If f^i is a date and f^j is an interval, then $d(f^i, f^j) = d([f^i, f^i], f^j)$.
- If f^j is a date and f^i is an interval, this function is defined in a similar way.

Finally, to measure the global similarity between pairs of documents, we define the following two alternative functions:

1. If $S_T(d^i, d^j) \geq b_T$ and $S_F(d^i, d^j) \geq b_F$, then

   $$S^1(d^i, d^j) = W_T \cdot S_T(d^i, d^j) + W_F \cdot S_F(d^i, d^j)$$

else, $S^1(d^i, d^j) = 0$.

where W_T, $W_F \in [0,1]$ represent the relative importance of the different document components respectively. The thresholds b_T, $b_F \in [0,1]$ are the minimum partial similarities required for the semantic and temporal components, respectively.

2. If $D_F(d^i, d^j) \leq b_F$ then $S^2(d^i, d^j) = S_T(d^i, d^j)$ else, $S^2(d^i, d^j) = 0$

   where b_F is the maximum number of days required to determine whether two documents refer to the same or two distinct events .

4. Temporal-Semantic Clustering of Documents

The unsupervised pattern recognition problem over ζ consists of determining the covering set $K_1, \ldots, K_r$, r>1. Starting from this formulation, we use a clustering criterion based on topological relationships between documents. This approach is based on the following idea: given a document description set, we must find or generate a natural structure for these documents in the representation space. This structure must be carried out by the use of some similarity measure between documents based on certain property.

The clustering criteria usually have three parameters, namely: a similarity measure S, a property Π that establishes the use of S, and a *threshold* b_0. Thus, clusters are determined by imposing the fulfillment of certain properties over the similarities between documents.

According to this, we will consider the following definitions:

<u>Definition 1:</u> Two documents d^i and d^j are b_0-*similar* if $S(d^i, d^j) \geq b_0$. Similarly, d^i is a b_0-*isolated element* if $\forall d^j \in \zeta$, $S(d^i, d^j) < b_0$.

<u>Definition 2:</u> [Mart00] The set $NU \subseteq \zeta$, $NU \neq f$, is a b_0-*compact nucleus* if:

a) $\forall\, d^j \in \zeta\, [\, d^i \in NU \wedge \max\limits_{\substack{d^t \in \zeta \\ d^t \neq d^i}} \{S(d^i, d^t)\} = S(d^i, d^j) \geq \beta_0\,] \Rightarrow d^j \in NU.$

b) $[\, \max\limits_{\substack{d^i \in \zeta \\ d^i \neq d^p}} \{S(d^p, d^i)\} = S(d^p, d^t) \geq \beta_0 \wedge d^t \in NU] \Rightarrow d^p \in NU.$

c) $|NU|$ is the minimum.

d) Any β_0-isolated element is a b_0-*compact nucleus* (*degenerated*).

Notice that this criterion is equivalent to finding the connected components of the graph based on the maximum similarity. In this graph, the nodes are the documents and there is an edge from the node d^I to the node d^J if d^J is the most similar document to d^I and its similarity overcomes the threshold b_0.

4.1. b0-Compact Nucleus Algorithm

In this paper we propose a new incremental clustering algorithm of news for event detection. This algorithm is based on the β_0-compact nucleus mentioned above thereby we call it *incremental* b_0*-compact algorithm.* In this algorithm each document d^i has an *Info* field associated, which contains the document (or documents) that more closely resembles d^i along with the value of this maximum similarity. Every time a new document arrives, its similarity with all the documents of the existing clusters is calculated and their fields *Info* are updated. Then, a new cluster with the new document is built together with the documents connected to it in the graph of maximum similarity. Every time a document is added to the new cluster, it is removed from the cluster in which it was located before.

The *incremental* b_0*-compact algorithm* proposed in this paper can be described as follows:

```
Input: Similarity threshold $\beta_0$.
      Similarity measure $S$ and its parameters.
Output: Document clusters ($\beta_0$-compact nucleus) representing the identified events.
Step 1. Arrival of a document $d$.
        $MS = 0$, $MD = \emptyset$, $Info(d) = (MD, MS)$
Step 2. For each existing cluster $G'$ do
            For each document $d'$ in $G'$ do
                Calculate the similarity $S$ between $d$ and $d'$.
                If $S \geq \beta_0$ then
                    $(Max, SimilMax) = Info(d')$.
                    If $S \geq SimilMax$ then update $Info(d')$ with the document $d$ and $SimilMax$.
                    If $S \geq MS$ then update $MS$ with $S$ and $MD$ with $d'$.
Step 3. Create a new cluster $G$ with the document $d$.
Step 4. If $MS \neq 0$ then
            Add to $G$ all the documents of the remaining clusters that have in the field Info a
          document of $G$, and remove them from the clusters where they are placed.
            Add to $G$ all the documents of the remaining clusters that are included in the field Info of
          a document of $G$ and remove them from the clusters where they are placed.
```

The worst case time complexity of this algorithm is $O(n^2)$, since for each document all the documents of the existing clusters must be checked to find the most similar document.

This clustering algorithm allows the finding of clusters with arbitrary shapes, as opposed to algorithms such as *K-means* and *Single-Pass*, which requires central measurements in order to generate the clusters, and as a consequence, the shapes of these clusters are restricted to be spherical. Another advantage of this algorithm is that the generated set of clusters is unique, and it does not depend on the arrival order of the documents.

Moreover, the clustering criterion based on the b_0-compact nucleus forms disjoint, more cohesive and smaller clusters than those formed by the b_0-connected components. Hence, in this case the chaining effect is much smaller than in the b_0-connected components.

4.2. Representation of Clusters

When we apply the clustering criterion mentioned above to the document collection we obtain several clusters of news with a high temporal-semantic similarity. In this level the individual events reported by the documents are identified. In the next levels, these events are successively grouped applying the same clustering criterion so that more complex events and topics can be identified. The resulting hierarchy describes the structure of topics and events taking into account their temporal occurrence. Therefore, the cluster representatives should be determined. Once the clusters of the first level (events) have been calculated, the representatives of each cluster are determined, and they are grouped to form the clusters of the next level in the hierarchy.

The representative of a cluster *c*, denoted as $\overline{c}$, is a pair $(T^{\overline{c}}, F^{\overline{c}})$, in which $T^{\overline{c}}$ is the component of the terms, and $F^{\overline{c}}$ the temporal component. It can be calculated with both the sum and the average of the cluster's documents as follows:

1. $T^{\overline{c}} = \left(T_1^{\overline{c}}, \dots, T_n^{\overline{c}}\right)$, where $T_j^{\overline{c}} = \frac{1}{|c|} \sum_{d^k \in c} TF_j^k$, $j \in \{1,\dots,n\}$

   $F^{\overline{c}} = \left(F_{f_1}^{\overline{c}}, \dots, F_{f_s}^{\overline{c}}\right)$, where $F_{f_j}^{\overline{c}} = \frac{1}{|c|} \sum_{d^k \in c} TF_{f_j}^k$, $j \in \{1,\dots,s\}$ and s is the total number of time entities that describe the documents of this cluster.

2. $T^{\overline{c}} = \left(T_1^{\overline{c}}, \dots, T_n^{\overline{c}}\right)$, where $T_j^{\overline{c}} = \sum_{d^k \in c} TF_j^k$, $j \in \{1,\dots,n\}$

   $F^{\overline{c}} = \left(F_{f_1}^{\overline{c}}, \dots, F_{f_s}^{\overline{c}}\right)$, where $F_{f_j}^{\overline{c}} = \sum_{d^k \in c} TF_{f_j}^k$, $j \in \{1,\dots,s\}$ and s is the total number of time entities that describe the documents of this cluster.

In order to reduce the dimension of component vectors of the cluster representative we truncate these vectors. For this purpose, usually a threshold is applied to eliminate the least significant terms. So, we only keep the terms and dates whose frequency in the representative is greater than or equal to one tenth part of the maximum frequency.

5. Evaluation

The effectiveness of the clustering algorithms has been evaluated using a collection of 452 news articles published in the Spanish newspaper "El Pais" during June 1999. We have manually identified 71 non-unitary events for this collection, and the maximum size of events is 16 documents. From these events we have identified 43 topics. It is worth mentioning that this collection covers 21 events associated to the end of the

Kosovo war along with their immediate consequences. These events have a high temporal-semantic overlapping, which makes difficult their identification. Additionally, these events have an important impact over other different events.

To evaluate the clustering results we use two measures of the literature that compare the system generated clusters with the manually labelled events, namely: the F1-measure [Rijs79] and the Detection Cost [TDT89].

The F1-measure is widely applied in Information Retrieval Systems and it combines the precision and recall factors. In our case, the F1-measure of a cluster *j* with respect to an event *i* can be evaluated as follows:

$$F1(i,j) = 2 \cdot \frac{n_{ij}}{n_i + n_j}$$

where n_{ij} is the number of common members in the event *i* and the cluster *j*, n_{i} is the cardinality of the event *i*, and n_{j} is the cardinality of the cluster *j*.

To define a global measure, first each event must be mapped to the cluster that produces the maximum F1-measure:

$$\sigma(i) = \underset{j}{\arg max}\{F1(i,j)\}$$

Then the global measure can be evaluated in two ways: micro-averaging and macro-averaging. Whereas the micro-average minimizes the variance of the estimates caused by individual documents, the macro-average minimizes the variance of estimates caused by event differences. Because of the small number of events and because of its heterogeneity, the micro-average is the preferred method for event detection [TDT98]. Moreover, to take into account the disparate event sizes, it is also preferable to weight the individual F1-measures with the event size. Hence, the final global measure is calculated as follows:

$$F1 = \frac{1}{N_{docs}} \sum_{i=1}^{N_{events}} n_i F1(i, \sigma(i))$$

The detection cost is a measure that combines both the miss and false alarm errors between an event *i* and a system-generated cluster *j*:

$$C_{DET}(i,j) = P_{miss}(i,j) \cdot P_{topic} + P_{false\_alarm}(i,j) \cdot (1 - P_{topic})$$

where $P_{miss} = (n_{\mathrm{i}} - n_{\mathrm{ij}})/\, n_{\mathrm{i}}$ and $P_{false\_alarm} = (n_{\mathrm{j}} - n_{\mathrm{ij}})/(N - n_{\mathrm{i}})$, P_{topic} is the a priori probability of a document belonging to a given event, and N is the collection size.

Again, to define a global measure, each event must be mapped to the cluster that produces the minimum detection cost:

$$\sigma(i) = \underset{j}{\arg min}\{C_{DET}(i,j)\}$$

The micro-average of this measure is defined as follows:

$$C_{DET} = P_{miss} \cdot P_{topic} + P_{false\_alarm} \cdot (1 - P_{topic})$$

$$P_{miss} = \frac{1}{N_{events}} \sum_{i=1}^{N_{events}} \frac{n_i - n_{i\sigma(i)}}{n_i} \qquad P_{false\_alarm} = \frac{1}{N_{events}} \sum_{i=1}^{N_{events}} \frac{n_{\sigma(i)} - n_{i\sigma(i)}}{N - n_i}$$

In Figure 1, the F1-measure results for the two proposed similarity measures at the event level are shown. In the left graphic, the behavior curves of the similarity function S^1 are plotted against the weight of the vector of terms ($W_F = 1 - W_T$). Each of such curves corresponds to a different value of b_0. In the right graphic, the behavior curves of the similarity function S^2 are plotted against the time threshold b_F, being the last point of this graphic $b_F=\infty$. Notice that disregarding the time component at all, the best result for F1 is 0,622 with b_0=0,33. Clearly, both graphics demonstrate the usefulness of the time vector for event detection.

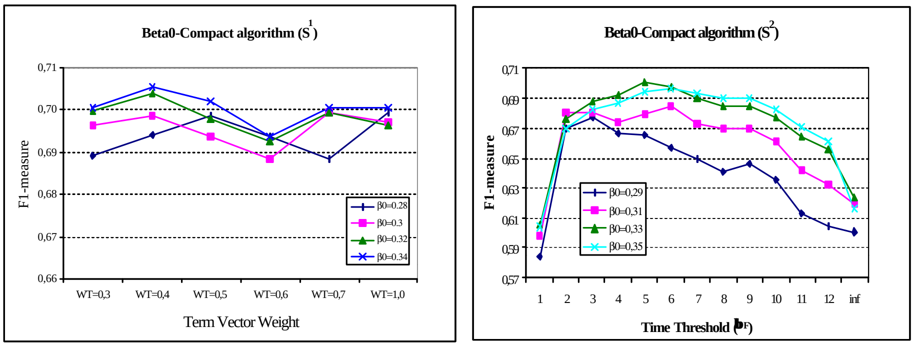

Fig. 1. F1-measure results for the β_0-Compact Algorithm (event level).

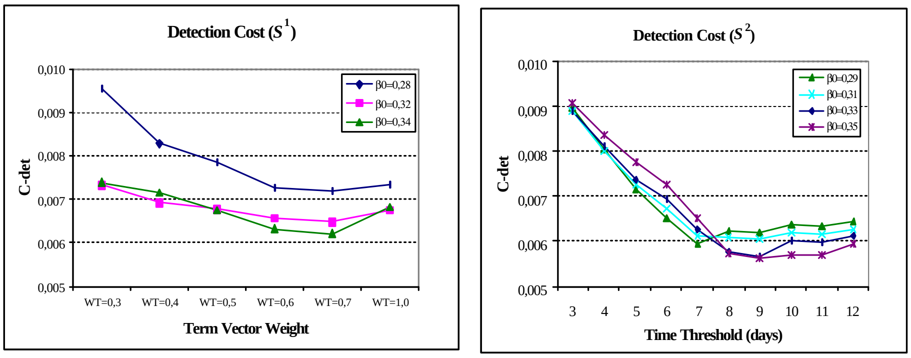

Fig. 2. Detection Cost results for the β_0-Compact Algorithm (event level).

Figure 2 shows the detection cost results at the event level. Again, regarding the time component improves the detection cost of the generated clusters. However, notice that the optimal weight and threshold for these measures (W_T=0,7 and b_F=8) differ from those obtained for the F1-measure (W_T=0,4 and b_F=5). This implies that the time vector has a smaller impact in the event detection task than in the retrieval one.

With regard to the topic level, the best result for the F1-measure is around 0,657, which has been obtained with S^1, b_0 = 0.4, W_T = 0,7, and taking the truncated average vector as the cluster representative. In general, we have observed that the impact of the time component is much smaller at the topic level.

6. Conclusions

In this paper new similarity measures between documents considering both the temporality and contents of the news articles have been introduced. Unlike other proposals, the temporal proximity is not just based on the publication date, but it is calculated using a group of dates automatically extracted from the article texts.

A new algorithm for determining a hierarchy of grouped news is also introduced. In the first level the individual events reported by the documents are identified. In the next levels, these events are successively grouped so that more complex events and topics can be identified. This algorithm is based on the incremental construction of existing β_0-compact nucleus in the collection of documents. It allows the finding of clusters with arbitrary shapes, as opposed to algorithms such as *K-means* and *Single-Pass*, which are restricted to be spherical. Another advantage of this algorithm is that the generated set of clusters is unique independently of the arrival order of documents.

Our experiments have demonstrated the positive impact of the time component in the quality of the system-generated clusters. Moreover, the obtained results for the F1-measure and the detection cost also demonstrate the validity of our algorithm for event detection tasks.

As future work, we will analyze the optimization of the β_0-compact algorithm, keeping the current cluster quality. Also we want to study other methods for calculating the cluster representatives, and the inclusion of other article attributes such as the event places in the global similarity measure.

References

[Carb99] Carbonell, J. et al. CMU: Report on TDT2: Segmentation Detection and Tracking. In *Proc. of DARPA Broadcast News Workshop*, 117-120, 1999.

[Cutt92] Cutting, D.R. et al. Scatter/Gather: a cluster-based approach to browsing large document collections. In *Proc. ACM/SIGIR 1992*, 318-329, 1992.

[Ichi94] Ichino, M.; Yagushi, H. Generalized Minkowski Metrics for Mixed Feature-Type Data Analysis. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 24(4), 1994.

[Llid01] Llido, D.; Berlanga R.; Aramburu M.J. Extracting temporal references to automatically assign document event-time periods. In *Proc. Database and Expert System Applications*, 62-71, Springer-Verlag, Munich, 2001.

[Mart00] Martínez Trinidad, J. F., Ruíz Shulcloper J., Lazo Cortés, M. Structuralization of Universes. *Fuzzy Sets and Systems*, Vol. 112 (3), pp. 485-500, 2000.

[Papk99] Papka, R. *On-line New Event Detection, Clustering and Tracking.* Ph.D. Thesis report, University of Massachusetts, Department of Computer Science, 1999.

[Rijs79] van Rijsbergen, C.J. *Information Retrieval.* Butter-Worths, London, 1979.

[TDT98] National Institute of Standards and Technology. *The Topic Detection and Tracking Phase 2 (TDT2) evaluation plan.* version 3.7, 1998.

[Wall99] Walls, F.; Jin, H.; Sista, S.; Schwartz, R. Topic Detection in Broadcast news. In *Proc. DARPA Broadcast News Workshop*, 193-198, 1999.

[Yang00] Yang, Y. et al. Improving text categorization methods for event tracking. In *Proc. ACM/SIGIR 2000*, 65-72, 2000.