A Method for the Automatic Summarization of Topic-Based Clusters of Documents

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

[1]Universidad de Oriente, Santiago de Cuba (Cuba),
aurora@app.uo.edu.cu,
[2]Institute of Cybernetics, Mathematics and Physics, La Habana (Cuba)
recpat@cidet.icmf.inf.cu
[3]Universitat Jaume I, Castellón (Spain)
berlanga@lsi.uji.es

Abstract. In this paper we propose an effective method to summarize document clusters. This method is based on the Testor Theory, and it is applied to a group of newspaper articles in order to summarize the events that they describe. This method is also applicable to either a very large document collection or a very large document, in order to identify the main themes (topics) of the collection (documents) and to summarize them. The results obtained in the experiments demonstrate the usefulness of the proposed method.

1 Introduction

Topic Detection and Tracking (TDT) is a new line of research that comprises three major sub-problems: segmenting speech-recognized TV/radio broadcasts into news stories, detecting novel events, and tracking the development of an event according to a given set of sample stories of that event [1]. An event in the TDT context is something that occurs at a specific place and time associated with some specific actions [2]. For example, the eruption of Mount Pinatubo on June 15th, 1991 is consider an event.

Starting from a continuous stream of newspaper articles, 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.

Clustering algorithms have been traditionally used in the *Event Detection* problem, such as the *K-Means*, *Single-Pass* and others [3, 4, 5]. In our approach, we use the *incremental compact algorithm* [6, 7] to solve this problem. This algorithm obtains high quality temporal-semantic clusters of documents, which represent the events of the collection, and it is independent of the document arrival order.

Another important problem that arises in the *event detection systems* is that of providing summaries for the detected events. Apart from the set of the cluster's frequent terms [8] and the relevant news titles [9], these systems do not offer any further information about the events that the generated clusters are representing. Consequently, many times users have to read the documents of the clusters to know the events they report. In the literature, the problem of summarizing a set of input documents (called *multidocument summarization*) has received much attention lately (e.g. [10, 11]).

A. Sanfeliu and J. Ruiz-Shulcloper (Eds.): CIARP 2003, LNCS 2905, pp. 596–603, 2003.

Basically, a multidocument summarization system tries to determine which sentences must be included in the summary, and then how to organise them to make the summary comprehensible. Many of these approaches are based on a sentence weight function that takes into account the position of the sentences in the documents, the length of the sentences, and the number of frequent keywords of the set of documents they include [12]. In this way, all the sentences in the document cluster must be scored to select the most appropriate for the summary. One of the main drawbacks of the current scoring procedures is that they are slow because the weight of a sentence depends on whether other sentences have been selected or not [13].

In this paper we present a novel and effective method for the multidocument summarization problem, based on the *Testor Theory* [14]. Starting from a set of document clusters, each one representing a different event or topic, our method tries to select the frequent terms of each cluster that are not included in the other clusters (testors). These terms are usually tightly related to the event of the cluster. Once these terms have been selected, the system extracts all the sentences that contain the selected terms. Finally, the system orders the extracted sentences and it produces the cluster's summary from them.

The proposed method computes very fast, and it produces good summaries for the document clusters we have analysed. Unlike the other methods in the literature, the selection of sentences is based on the discriminating frequent terms of each cluster, which can be efficiently computed.

2 Document Representation

The incoming stream of documents that feed our system comes from some on-line newspapers available in Internet, which are automatically translated into XML. This representation preserves the original logical structure of the newspapers.

From them, our detection system builds three feature vectors to represent each document d^i, namely [7]:

- *A vector of weighted terms* ($TF_1^i, \ldots, TF_n^i$), where the terms represent the lemmas of the words appearing in the content of the document, and TF_k^i is the relative frequency of the term t_k in d^i. Stop words are disregarded from this vector.
- *A vector of weighted time entities*, where time entities are either dates or date intervals. These time entities are automatically extracted from the content of the documents by using the algorithm presented in [15].
- *A vector of weighted places*. These places are automatically extracted from the content of the documents by using a thesaurus of place names.

The automatic construction of cluster summaries must take into account these three components. In [15], it was shown how the temporal entities of a document (cluster) can be summarized as a date interval called *event-time* period. Places can be easily summarized by taken a representative place from the cluster documents. Thus, in this paper we only focus on the term vector to extract the cluster summaries.

3 Basic Concepts

Before presenting our summarization method, we review the main definitions of the *Testor Theory* [14] and we define the basic concepts of this method.

In our problem, the collection of news is partitioned into clusters. Each cluster represents an event. Let ζ be a set of detected events in a news collection.

The representative of a cluster $c \in \zeta$, denoted as $\overline{c}$, is calculated as the union of the documents of that cluster, that is, $\overline{c} = \left(TF_1^{\overline{c}},...,TF_n^{\overline{c}}\right)$, where $TF_j^{\overline{c}}$ is the relative frequency of term t_j in the sum vector of the documents of that cluster.

Given a cluster c, let $T(c) = \{t_1,\ldots, t_{n_c}\}$ be the most frequent terms in the representative $\overline{c}$, i.e., the terms t_j such that $TF_j^{\overline{c}} \geq \varepsilon$, $j = 1,\ldots, n_c$ and ε is an user-defined parameter.

For each cluster c, we construct a matrix $MR(c)$, whose columns are the terms of $T(c)$ and its rows are the representatives of all clusters of ζ, described in terms of these columns. Notes that this matrix is different in each cluster.

In the Testor Theory, the set $\tau = \{x_{i_1},\ldots,x_{i_k}\}$ of features and their corresponding columns $\{i_1,...,i_k\}$ of a matrix M is called a *testor*, if after deleting from M all columns except $\{i_1,...,i_k\}$, all rows of M corresponding to distinct clusters are different. A testor is called *irreducible* (*typical*) if none of its proper subsets is a testor [14]. The *length* of the testor is the cardinal of τ.

For the calculus of the typical testors of a matrix M, the key concept is the comparison criterion of the values of each feature. One of these criteria is, for example:

$$d(v_{ik}, v_{jk}) = \begin{cases} 1 & if\ v_{ik} - v_{jk} \geq \delta \\ 0 & otherwise \end{cases}, \qquad (1)$$

where v_{ik}, v_{jk} are the values in the rows i and j in the column corresponding to the feature x_k respectively, and δ is an user-defined parameter.

In order to determine all typical testors of a matrix we use the algorithm LEX, which is described in detail in [16]. This algorithm outperforms the other algorithms to calculate the typical testors.

Let a sentence S of a document d and U be a set of terms. We call *maximal co-occurrence* of S with respect to U, and we will denoted it as $mc(S,U)$, to the set of all terms of U that also occur in S.

4 Method of Summarization

In our opinion, a summary of an event should include the terms that characterize the event, but also those that distinguish this event from the rest.

A summary of an event c consists of a set of sentences extracted from the documents in c, in which the highest quantity of terms that belong to typical testors of the maximum length of the matrix $MR(c)$ occurs. Moreover, the sentences that cover the calculated typical testor set are also added to the summary.

In order to improve the coherence and organization of the summaries, we sort the extracted sentences according to the publication date of the news and their position in the text.

In order to calculate these typical testors, we considered two classes in the matrix $MR(c)$. The first class is only formed by $\overline{c}$ and the second one is formed by all remaining cluster representatives. The comparison criterion applied to all the features is that of (1). Notice that this criterion requires that the terms frequently appear in the cluster documents but not in the other clusters.

The proposed summarization method is described as follows:

```
Algorithm Summarization of event set
Input:  $\zeta$: a set of events (clusters) of a news collection.
        $\varepsilon$: threshold of term frequencies.
        $\delta$: parameter of the comparison criterion
Output: Summary of each event.
For each event $c \in \zeta$:
    1. Construct the matrix $MR(c)$.
    2. Calculate the typical testors of the maximum length in the matrix $MR(c)$.
    3. Let $U$ be the union of all typical testors found in the step 2.
    4. For each document $d_i$ in the cluster $c$:
           For each sentence $S$ in $d_i$:
               Calculate the maximal co-occurrence $mc(S,U)$.
    5. Order decreasingly all sentences in terms of the cardinal of its maximal co-occurrence.
       Let $p_1 > p_2 > ... > p_s$ be these cardinals.
    6. $Summary (c) = \varnothing$
    7. Add to $Summary(c)$ all sentences $S$ that satisfy $|mc(S,U)| = p_1$.
    8. $W = \bigcup_{|mc(S,U)|=p_1} mc(S,U)$.
    9. $i = 2$.
    10. While $U \setminus W \neq \varnothing$ do:
            $V = \varnothing$.
            For each $mc(S,U)$ of cardinal $p_i$:
                If $mc(S,U) \cap (U \setminus W) \neq \varnothing$ then
                    Add $S$ to $Summary(c)$.
                    $V = V \cup mc(S,U)$
            $i = i + 1$
            $W = W \cup V$
    11. Sort all sentences in $Summary(c)$ according to the publication date of the news in $c$ and their
        position in the text.
```

The paragraph is a useful semantic unit for the construction of the summaries because most writers view a paragraph as a topical unit, and organize their thoughts accordingly. Therefore, if we want to obtain more extensive summaries, we can use the paragraphs instead of the sentences. Thus, we would extract the paragraphs that cover the typical testor set.

5 Experiments and Results

The effectiveness of the proposed summarization method has been evaluated using two collections. The first one contains 554 articles published in the Spanish newspaper "El País" during June 1999. We have identified 85 non-unitary events, being their maximum size of 18 documents. The collection covers 21 events associated to the end of the “Kosovo War” along with their immediate consequences, the visit of the Pope to Poland, the elections in several countries like South Africa, Indonesia and the European elections, the events related with the trials to Pinochet and the Kurdish leader Ocalan, among others.

In order to show the quality of the obtained summaries one detected event is selected. It is about the murder of the famous Mexican presenter Paco Stanley. This cluster is formed by 5 documents. Table 1 shows the titles, the publication dates and the length (number of words) of each document in this cluster.

Table 1. Documents about the murder of Paco Stanley.

Publication date	Title	Length
1999-6-8	Commotion in Mexico for the murder of a famous presenter.	531
1999-6-9	The Mexican Intelligence declares that the murdered comedian had links with the drug traffic.	748
1999-6-9	The televisions incite the indignation against Cárdenas.	298
1999-6-10	An atmosphere of collective hysteria has been created.	203
1999-6-10	The death of Stanley agitates the political atmosphere in Mexico.	540

The union of the found typical testors for this cluster is: {*murder*, *Mexican*, *death*}. The obtained summary of this event by our method is the following one:

> Paco Stanley presented several popular variety programs in the Aztec Television, and its death has caused a deep commotion in the Mexican society, in which he was a very appreciated person. Porfirio Muñoz Ledo, the mayor's competitor in the PRD internal fight for the presidential nomination of the party in the presidential elections of the 2000, annotated new causes in the murder of the showman, as he declared that 'it would be immoral' that this notorious death could affect the election process, in which Cárdenas aspires to the Republic presidency. During hours, the reporter Raúl Trejo explained, the only thing that was seen in the Mexican television screens, after murder of Paco Stanley, was a parade of laments, complaints and demands that nothing clarified, but that they became a battering ram.

As we can see, this summary offers a concise vision of the main aspects of the murder of Paco Stanley. This summary has 139 words, in contrast with the 2320 words in total that all documents of the cluster have. That is, a 94% of compression rate. It is worth mentioning that we try to maintain the original fragments in the carried out translation of this summary.

Figure 1 shows the results for the compression rate (in %) at each event with respect to its size (total number of words in the documents of the event). As we can see, there exists a tendency such that the higher event size, the greater the compression rate is.

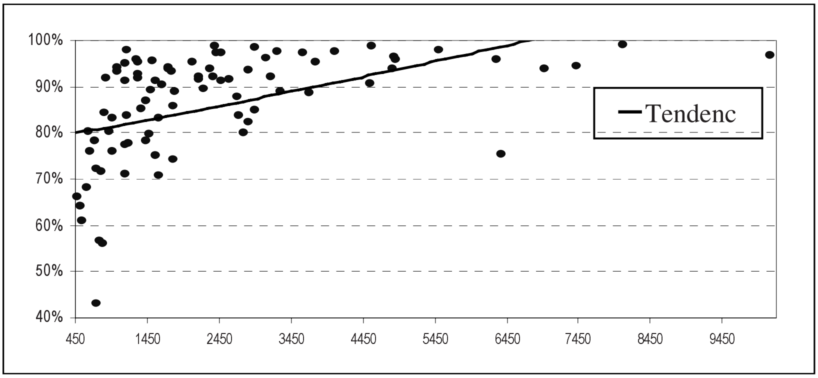

Fig. 1. Compression rate against the total number of words in each event.

In order to evaluate the effectiveness of our method, we also used the data (in Spanish) from the TREC-4 conference [17]. This collection contains 693 articles published by AFP agency during 1994. These articles are classified into 23 topics.

Table 2 shows for each topic, its size (number of documents), the size of its typical testor set (number of terms) and the obtained compression rate (in %).

Table 2. Obtained results in TREC data.

Topic	Size	TT	Rate	Topic	Size	TT	Rate	Topic	Size	TT	Rate
SP51	83	4	96.81	SP60	47	7	97.94	SP69	62	4	98.06
SP52	13	3	98.20	SP62	15	6	85.72	SP70	6	3	74.76
SP53	46	2	99.36	SP63	5	4	83.43	SP71	17	3	99.29
SP54	37	8	94.38	SP64	9	4	83.54	SP72	14	7	91.05
SP55	108	4	99.29	SP65	29	5	96.40	SP73	13	5	93.48
SP57	2	3	82.05	SP66	68	4	97.32	SP74	34	10	97.65
SP58	49	9	97.91	SP67	13	9	87.20	SP75	20	3	94.09
SP59	7	6	93.74	SP68	12	8	93.78				

Again, the obtained summaries capture the main ideas about each topic and an appreciable reduction of words is also achieved. For example, the description given by TREC to the topic SP68 is "AIDS situation in Argentine and what steps is the Argentine government taking to combat the disease". The union of the found typical testors for this topic is: {*Argentine*, *campaign*, *prevention*, *disease*, *population*, *drug*, *case*, *people*}. The summary obtained for this topic is the following one:

> The combination of the DDI drugs and hidroxamates against the AIDS disease "could be the only way to remove the virus from the seropositive people", the argentinean doctor Julio Vila, which leads a research group in France, said this Sunday. The Health authorities announced this Friday that the Argentine Government will start a educational campaign to prevent the Acquired Immune Deficiency Syndrome (AIDS). The minister of Health and Social Action, Alberto Mazza, had announced this Friday that the government will start a educational campaign, whose objective is "to instruct the population about the prevention measures that must be adopted to combat the AIDS". The campaign of preventive education

> will be started after publishing the results of a national public-opinion poll about the population knowledge on the disease, according to Mazza declarations. Although detected during the last years, the cases were just known now, pointing out that the situation has produced a prevention campaign ordered by the army chief General Martin Balza, which presumes that we are against a "serious but controllable problem". In the same way, the Army Immune Deficiency Center (CEIDE) was created within the Central Army Hospital as an institution that will be in charge of the prevention campaigns, as well as the treatment and monitoring of the AIDS cases.

Indeed, it is hard to evaluate the quality of a summarization method. In spite of this, we consider our summaries readable, coherent and excellent at capturing the main themes of the document sets. Thus, we believe that these summaries can be presented to the user as a meaningful description of the cluster contents.

6 Conclusions

In this work we presented an effective method to summarize document clusters generated by Topic Detection Systems. The proposed method employs the calculus of typical testors as its primary operation and from them, it constructs the summaries of each cluster.

The most important novelty is the use of typical testors combined with different techniques and heuristics, to produce all together better summaries.

This method enables construction of a concise representation of the focused cluster. The obtained summaries are much more descriptive than simple sets of frequent words.

The proposed method is applied to a set of newspaper articles in order to summarize the events that they describe. It is helpful to a user in order to determine at a glance whether the content of an event are of interest. The carried out experiments demonstrate the usefulness of the method. The summaries are readable, coherent and well organized. In most cases, the system successfully presents main themes, skips over minor details, and avoids redundancy. Additionally, the proposed summarization algorithm performs efficiently, taking much less time than the clustering process.

To sum up, the summarization method is robust, topic-independent and may easily be applied in other domains and other languages. Additionally, it can be applied to other document collections such as Web pages, books, and so on. For example, in a book we can consider as *clusters* some structural elements of the document (chapters, sections, etc.), being its members the different sub-structures they contain (subsections, paragraphs, etc.)

References

[1] Allan, J.; Carbonell, J.; Doddington, G.; Yamron, J. and Yang, Y.: Topic Detection and Tracking Pilot Study: Final Report. In *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, pp. 194–218, 1998.

[2] Yang, Y.; Ault, T.; Pierce, T. and Lattimer, C.W.: Improving text categorization methods for event tracking. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'00, Athens, pp. 65–72, 2000.

[3] Carbonell, J.; Yang, Y.; Lafferty, J.; Brown, R.D.; Pierce, T. and Liu, X.: CMU Report on TDT-2: Segmentation, detection and tracking. In *Proceedings of DARPA Broadcast News Workshop*, pp.117–120, 1999.

[4] Yamron, J.: Dragon's Tracking and Detection Systems for TDT2000 Evaluation. In *Proceedings of Topic Detection & Tracking Workshop*, pp.75–80, 2000.

[5] Allan, J.; Lavrenko, V.; Frey, D. and Khandelwal, V.: UMASS at TDT 2000. In *Proceedings TDT 2000 Workshop*, 2000.

[6] Pons-Porrata, A.; Berlanga-Llavori, R. and Ruiz-Shulcloper, J.: Detecting events and topics by using temporal references. *Lecture Notes in Artificial Intelligence* 2527, Springer Verlag, pp.11–20, 2002.

[7] Pons-Porrata, A.; Berlanga-Llavori, R. and Ruiz-Shulcloper, J.: Building a hierarchy of events and topics for newspaper digital libraries. *Lectures Notes on Computer Sciences 2633*, Springer-Verlag, 2003.

[8] Zamir, O.; Etzioni, O.; Madani, O. and Karp, R.M.: Fast and Intituitive Clustering of Web Documents. In *Proceedings of KDD'97*, pp. 287–290, 1997.

[9] Cutting, D.R.; Karger, D.R. and Pedersen, J.O.: Constant Interaction-Time Scatter/Gather Browsing of Very Large Document Collections. In *Proceedings of the 16th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1993.

[10] Mani, I. and Bloedorn, E.: Multi-Document Summarization by Graph Search and Matching. AAAI/IAAI 1997, pp. 622–628, 1997.

[11] Barzilay, R.; Elhadad, N. and McKeown, K.: Inferring Strategies for Sentence Ordering in Multidocument News Summarization. *Journal of Artificial Intelligence Research* 17, pp. 35–55, 2002.

[12] Mani, I.: *Automatic Summarisation*. John Benjamins Publishing Company, 2001.

[13] Marcu, D.: Discourse-based summarisation in DUC-2001, *Proceedings of Document Understanding Conference*, DUC-2001, 2001.

[14] Lazo-Cortés, M.; Ruiz-Shulcloper, J. and Alba-Cabrera, E.: An overview of the concept testor. *Pattern Recognition*, Vol. 34, Issue 4, pp.13–21, 2001.

[15] Llidó, D.; Berlanga R. and Aramburu M.J.: Extracting temporal references to automatically assign document event-time periods. In *Proceedings of Database and Expert System Applications 2001*, Springer-Verlag, Munich, pp. 62–71, 2001.

[16] Santiesteban, Y. and Pons, A.: LEX: a new algorithm for the calculus of typical testors. *Rev. Ciencias Matemáticas*, Vol. 21, No. 1, (in Spanish), 2003.

[17] http://trec.nist.gov.