Clustering-based Searching and Navigation in an Online News Source

Simón C. Smith and M. Andrea Rodríguez

Department of Computer Science, University of Concepción
Center for Web Research, University of Chile
Edmundo Larenas 215, 4070409 Concepción, Chile
{ssmith,andrea}@udec.cl

Abstract. The growing amount of online news posted on the WWW demands new algorithms that support topic detection, search, and navigation of news documents. This work presents an algorithm for topic detection that considers the temporal evolution of news and the structure of web documents. Then, it uses the results of the topic detection algorithm for searching and navigating in an online news source. An experimental evaluation with a collection of online news in Spanish indicates the advantages of incorporating the temporal aspect and structure of documents in the topic detection of news. In addition, topic-based clusters are well suited for guiding the search and navigation of news.

1 Introduction

The growing amount of online news posted on the WWW demands new algorithms that support topic detection, search, and navigation of news documents. This work explores the content of documents, the temporal evolution of news, and the structure of news documents to define an algorithm that creates topic-based clusters of documents (i.e., topic detection) and that uses these clusters for searching and navigating online news.

In this work, manifestations of events are seen as news. The proposed algorithm assigns news to previously detected or new events, a strategy called *single-link clustering* [16]. Basis for this work are the results from the TDT research initiative [13] that investigates new events in a stream of broadcast news stories. We consider a modification of the single-link topic detection algorithm (UMASS TDT2) that handles news as events in time [12]. Like the UMASS TDT2, the proposed algorithm assigns news to only one topic, which can also be extended to multiple topics. Such algorithm produces clusters of connected news that are then used for searching and navigating documents.

In the area of information systems, the concept of information navigation has been associated with visualization of retrieval results [11] [17] and with information access in an information space [10] [5]. This work uses a topic-based cluster as a semantic structure of connected news that can be used in a navigation process. In particular, this work has the following specific contributions:

– It implements an algorithm for Topic Detection of online news that modifies a previous algorithm (UMASS TDT2) by incorporating the structure of documents.
– It evaluates the proposed algorithm with respect to UMASS TDT 2 in a domain of news in Spanish.
– It presents a strategy for supporting the search and navigation of news. This strategy considers a cluster structure of connected news that embeds the temporal order and similarity in a stream of news.

The organization of the paper is as follows. Section 2 provides a review of related work concerning topic detection and navigation systems. Section 3 presents our algorithm for topic detection (CHILE TDT), which is compared to the UMASS TDT2 algorithm. Section 4 describes the use of clusters as a semantic structure for searching and navigation. Conclusions and further research issues are discussed in Section 5.

2 Related Work

Many advances on topic detection and tracking in online news sources are derived from the Topic Detection and Tracking (TDT) reseach initiative sponsored by DARPA [13]. Since its beginning in 1996, this research initiative has produced serveral important results. Most approaches to TDT use some sort of clustering strategy, such as single-link clustering or hierarchical group-average clustering [16]. The TDT developments started with the application of traditional clustering algorithm for topics detection [15]. Then, systems considered that topic detection and tracking of news are inherently related to the data flow in time [12]. Lately, methods considered that clusters of news may overlap; that is, a document may belong to different topics [2]. In addition, other studies have proposed algorithms that show improvements when using keyword based analysis of text documents; example of such studies are: relevance models for topic detection and tracking [15] and event tracking on domain dependency [6].

TDT2 proposes to incorporate the temporal dimension for clustering news documents with a single-link topic detection algorithm (UMASS TDT2) [12]. In this proposal the content of documents are represented as queries. If a new document triggers an existing query (i.e., the similarity between the document and the query exceeds the query's threshold), the document is considered to discuss the same topics (event) than the query; otherwise, it becomes a new event. This query's threshold is penalized by the temporal distance between the query and the new document.

A more recent work explores time and space with ontological information in the topic detection of online news [8]. This work uses semantic classes for locations, proper names, temporal expressions and general terms. Instead of representing news as a single document vector, this approach uses four vectors that reside in different spaces: spatial location, proper names, temporal dimension and general terms. It requires to extract terms with a grammar-based parser, a geographical ontology, and an automata for temporal expression pattern recognition. In this sense, this work goes beyond the simple syntactic analysis or keyword-based clustering strategy of information retrieval. An important effort goes to the grammar parsing, temporal analysis, and geographic association, but, unfortunately, no evaluation was found with respect to previous studies neither a complete description of the algorithm for its implementation and evaluation.

A topic-based cluster of news represents a collection of thematically related documents. Faced with a large collection of documents, the problem becomes to select and access the document to start with. A general idea for solving this problem is to use overviews of the information that guide users from general to more specific topics [3]. Implementations of this idea attempt to display overview information derived from the automatically extraction of most common themes in a collection. In many cases, these themes are associated with the centroids of clusters that group documents based on the similarity to one another. An innovative approach for clustering web documents uses web-snippet, which clusters the fragments of web pages that are returned by a search engine and summarizes the context of searched keywords [9] [5].

3 Topic Detection

3.1 Algorithm

This work models topics of news as events that occur in time. News has a temporal order. News that arrives at time instant t_j may be thematically related to previously arrived news at time t_i, with $t_i \leq t_j$. Essentially, the topic detection algorithm includes the same steps as the single-link algorithm described in [12]:

- News documents define queries represented by using the $n-$most frequent terms (50 terms in our case) in a document query after eliminating stopwords.
- An initial threshold for a query is defined by comparing the query with respect to the document from where it was created (Equation 4).
- When a new document arrives, it is compared to previous queries (documents) (Equation 3) and linked to the query for which the similarity value between the document and the query most exceeds the query's threshold (Equation 4). In case a news document cannot be associated with any query, it is considered as the first document of a new cluster.
- The threshold of queries are adjusted by considering the arrival time of new documents.

The derived function of similarity takes as basic elements the *terms* in the documents, which are typically used in information retrieval systems to represent text documents. Each term in a document has a weight, which is determined by the number of occurrences in the document and in the set of documents. In particular, the weight of a term k in a query q_i is defined by $w_{q_{i,k}}$(Equation 1), and the weight of a term k in a new arriving document d_i is defined by $w_{d_{i,k}}$ (Equation 2).

$$w_{q_{i,k}} = tf_{q_{i,k}} = \frac{t_{d_{i,k}}}{t_{d_{i,k}} + 0.5 + 1.5 \cdot \frac{dl_{d_i}}{avg\_dl}} \tag{1}$$

$$w_{d_{i,k}} = tf_{d_{i,k}} \cdot idf_{d_{i,k}} = tf_{d_{i,k}} \cdot \frac{log(\frac{C+0.5}{df_k})}{log(C+1)} \tag{2}$$

where

$t_{d_{i,k}}$: frequency of the term in a document

dl_{d_i} : size of the document

C : number of documents in the corpus

avg_dl : average number of terms in a document

df_k : number of documents that contain the term k

After eliminating stop-words, terms in a text document are separated in three sets: (1) terms in the title (2) terms extracted from a syntactic analysis of documents and (3) terms in the body of the news (general terms). The syntactic analysis of documents extracts all uppercase words in the body of the document, composite words in uppercase, and words between semicolon or parenthesis. By separating different components of the documents, the system aims to handle their relative importance in characterizing the information content.

A document d_j is compared to a query q_i with the Equation 3. In this definition, we assume that the comparison between a new arriving document and a query cannot exceed the similarity between a query and the document from which it was created.

$$sim(q_i, d_j) = \frac{\theta \sum_{k=1}^{N} w^T_{q_{i,k}} \cdot w^T_{d_{j,k}} + \gamma \sum_{k=1}^{N} w^S_{q_{i,k}} \cdot w^S_{d_{j,k}} + \delta \sum_{k=1}^{N} w^B_{q_{i,k}} \cdot w^B_{d_{j,k}}}{\sum_{k=1}^{N} w^T_{q_{i,k}} + \sum_{k=1}^{N} w^S_{q_{i,k}} + \sum_{k=1}^{N} w^B_{q_{i,k}}} \tag{3}$$

where

$q_{i,k}$: term k in the query i

$d_{j,k}$: term k in the document j

T, S, B : terms in the title, extracted from the syntactic analysis, and in the body of the document, respectively

θ, γ, δ : weights optimized with the training set

The threshold of a query represents the minimum possible value of similarity between a document and a query to consider the document to belong to the same topic of the query (Equation 4). The initial threshold is defined by a comparison between the query and the document from which it is created. As the temporal difference between a query and a document increases, the initial threshold also increases, making it more difficult that the document belongs to the same topic.

$$threshold(q_i, d_j) = belief(q_i, d_i) + \beta * (date_j - date_i) \tag{4}$$

$$belief(q_i, d_j) = \frac{\displaystyle\sum_{k=1}^{N} w^T_{q_{i,k}} * w^T_{d_{j,k}} + \sum_{k=1}^{N} w^S_{q_{i,k}} * w^S_{d_{j,k}} + \sum_{k=1}^{N} w^B_{q_{i,k}} * w^B_{d_{j,k}}}{\sum_{k=1}^{N} w^T_{q_{i,k}} + \sum_{k=1}^{N} w^S_{q_{i,k}} + \sum_{k=1}^{N} w^B_{q_{i,k}}}$$

Time in this equation penalizes the similarity by the time distance in days between the query and the document. This penalization reflects the behavior of news stream where the frequency of related news tends to concentrate at the beginning of an event and decrease with the time [12]. This temporal behavior refers to the time when the news were posted, and not to the temporal content in the text of news, which requires a natural language processing and may not follow the same behavior used in this model. The weight of the time distance in Equation 4 (β) is determined by an optimization process over a set of training news.

Like for the temporal dimension, we initially explored the idea of using the spatial dimension of news by taking the geographic distance between news' publications. The analysis of news document shows, however, that there is a high geographic concentration of news that makes it inappropriate to consider geographic locations as data capable of distinguishing topics. From the total number of analyzed news of a Chilean site of online news, over 40% of them are related to the Chilean capital (Santiago) and, among Chilean news, more than 80% of their geographic associations are related to Santiago. This high concentration of news is due to the fact that documents were taken from online news services with bias to report news about one specific country (Chile).

Despite the high geographic concentration of news, we did a preliminary evaluation that considers the geographic association of news as a particular component of the similarity between document and query, such as we did with time in Equation 3, but the results were negative. Therefore, we have excluded this component from the model and from the results of the experimental evaluations.

3.2 Experimental Evaluation

The experimental evaluation of the topic detection algorithm uses a set of 60,000 news obtained from a Chilean web site of online news [4] between March 2003 and October 2004. Within this time period of data collection, 30 topics were selected, and each document of this collection was manually classified into one of the 30 topics or into a null-topic (i.e., a non identified event). Each of the selected topics was characterized with a title, a description of the starting event, an id of the initial event, a summary, and principles of interpretation. The selected topics vary in the length of the time interval they were relevant and the number of documents that appeared during this time interval (Table 1). In particular, topics cover time intervals from 8 days to one year and from 6 to 530 documents.

	Topic	Time interval	# documents
0	Caso MOP-GATE	01/02/03-10/29/04	244
1	Caso Inverlink	02/03/03-10/29/04	191
2	TLC entre Chile y Estados Unidos	03/24/03-04/20/04	89
3	Votación de Chile en la ONU acerca de Cuba	04/10/03-05/05/03	11
4	Nelson Mery Acusado de violaciones a DDHH	04/16/03-09/25/04	71
5	Chile campeón Mundial de Tenis 2003	05/16/03-09/24/03	14
6	Publicación deudores crédito fiscal	06/26/03-09/24/03	12
7	Desafuero de Pinochet caso Berríos	07/11/03-10/27/04	121
8	Royalty a la gran Minería	07/22/03-09/13/04	122
9	Presos políticos en huelga de hambre	07/26/03-10/28/03	69
10	Incendios en Portugal	08/03/03-08/12/03	6
11	Caso Spinak	10/06/03-10/27/04	531
12	Asalto a consulado Argentino en Punta Arenas	11/09/03-02/10/04	22
13	Caso Matute Johns	02/13/04-05/28/04	46
14	Envío de tropas Chilenas a Haití	03/02/04-09/21/04	50
15	Atentados en Espanã del 11 de Marzo	03/11/04-10/30/04	170

Table 1. A subset of the judged events in the dataset

From the 60,000 documents, three different corpus were randomly created:

- Auxiliar Corpus. It consists of 1,500 documents used in the calculation of tf and idf.
- Training corpus. it consists of 2,000 documents randomly selected.
- Evaluation corpus. It consists of 15,000 documents; 2,200 documents belonging to one of the identified topics and 12,800 documents randomly selected from the other 57,800 documents of the corpus (i.e., from all documents minus the 2,200 documents already selected).

A preprocessing of documents identified stop-words. Stop-words were not only prepositions or articles, but also all words whose high occurrence within the whole corpus makes them less significant for characterizing the topics of news. Examples of such words are *noticia* (news), *país* (country), and *internacional* (international).

To evaluate the performance of the algorithm, we calculated cases of *miss* and *false alarm*. *Miss* is the number of documents that are not, but should be, associated with a topic. *False alarm* is the number of documents that are wrongly associated with a topic. A cost function relates both measures by a weighted sum of the probabilities of *miss* (P_{miss}) and *false alarm* (P_{false}) (Equation 5) [1]. Like TDT2, we define $cost_{miss} = 0.02$ and $cost_{false} = 0.98$, giving more weight to assigning wrong documents to a cluster.

$$cost = cost_{false} * P_{miss} + cost_{miss} * P_{false} \tag{5}$$

The evaluation compares the results of our algorithm (CHILE TDT) with the original algorithm in [12] (UMASS TDT2). For both algorithms, the experiment runs different settings with the goal of optimizing the cost function. The optimal settings with the training corpus and the final cost evaluation are presented in Table 2.

System	θ	β	γ	δ	P_{miss}	P_{false}	$Cost$
UMASS TDT2	0.0767	0.0034			0.243	0.002	0.007
CHILE TDT	14.8	0.05	12.6	13.4	0.202	0.002	0.006

Table 2. Optimal settings and final cost values

The values of the parameters for the CHILE TDT algorithm indicates that terms in the title are more relevant than terms in the text of documents. The final values indicate a small advantage of CHILE TDT due to a less number of *miss*.

4 Navigation and Search

The characteristics of the topic-based clusters derived from the previous algorithm are used for navigating and searching among online news. The idea is that clusters provide means for finding related news based on users' queries with finer granularity.

The structure of clusters derived from the TDT algorithm is a hierarchical structure, starting with the first temporal document in the cluster. Time is implicit in the hierarchy, where up level news are earlier than bottom level news. Each news document within a cluster, with the exception of the starting document, is associated with the previous and most similar document. We represent such cluster as an acyclic, directed and weighted graph.

4.1 Navigation

A simple strategy to support the navigation in the information space defined by a cluster is to highlight nodes in the cluster with a strong thematic association, or inversely, to highlight nodes that are less connected to other nodes of the cluster. In this way, users may decide to access only nodes strongly connected or to look for different information contained in the same cluster. A basic approach to determining a strong association between news is to define a threshold that allows a binary classification between associations (strong or weak associations) based on the similarity value obtained from the topic detection algorithm. This threshold may be defined/modified by the users, but also may be initially proposed by the system.

This work proposes a threshold based on the percentage of *false alarm* detected in the evaluation process, which differs from the probability of false alarm used in Equation 5. The idea is to determine the number of nodes equivalent to the percentage of *false alarm* with the lowest association similarity, having that news belonging to the set of wrong assignments should have less association similarity. For example, if the percentage of *false alarm* for the system is 30% and we have a cluster with 100 news, the thirty news with the lowest association similarity are considered to have weak association, and the threshold is defined by the thirtieth lowest similarity value.

With a threshold for the classification between weak and strong associations, the classification process in the clusters is as follows. Starting from the root, if the association similarity between a child node and its corresponding parent node is larger than the threshold, the nodes are considered strongly associated with a same highlighting color. If the association similarity is less than the threshold, in contrast, the color of the child node will be set different to the parent node.

As an example, Figure 1 shows the original cluster and the cluster with further refinement that refer to the ”Chile Campeón Mundial de Tenis 2003” (topic id 5, Table 1). In this figure, there are 4 changes of color that indicate weak associations of news. If we consider the path in the graph after the weak association, we have 6 different news that are separated from the main topic of the cluster. Based on the manual classification of news, this cluster obtained from the topic detection algorithm has 9 cases of false alarm, which include the 6 news graphically separated from the main topic of the cluster. Thus, three cases of false alarm (30% of cases of false alarm) were undetected by using the proposed threshold.

In our online implementation of the system, graphs, initially specified in Graph Description Language (GDL) [7], were translated into a Scalable Vector Graph (SVG) [14] format and visualized in a browser (Figure 2). In addition to the refinement by weak and strong associations, in this visualization the levels in the graph are associated with the temporal evolution of news, such that news that appear at the same level are concurrent. Other functionalities of the interface show the content of news as users move around the graph. We do not include a full description of the interface due to space constraints.

4.2 Search

In a search process, user queries and documents are compared for similarity. If no indexing of documents exists, the system has to do an exhaustive comparison with all documents. In this work, clusters of news provide a basic organization of documents such that it is possible to filter clusters that are not similar to the

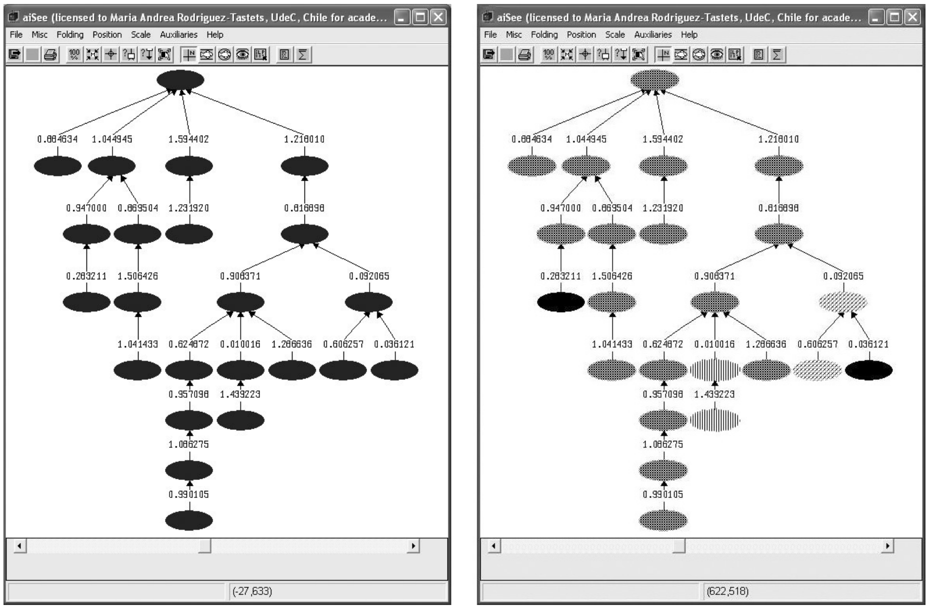
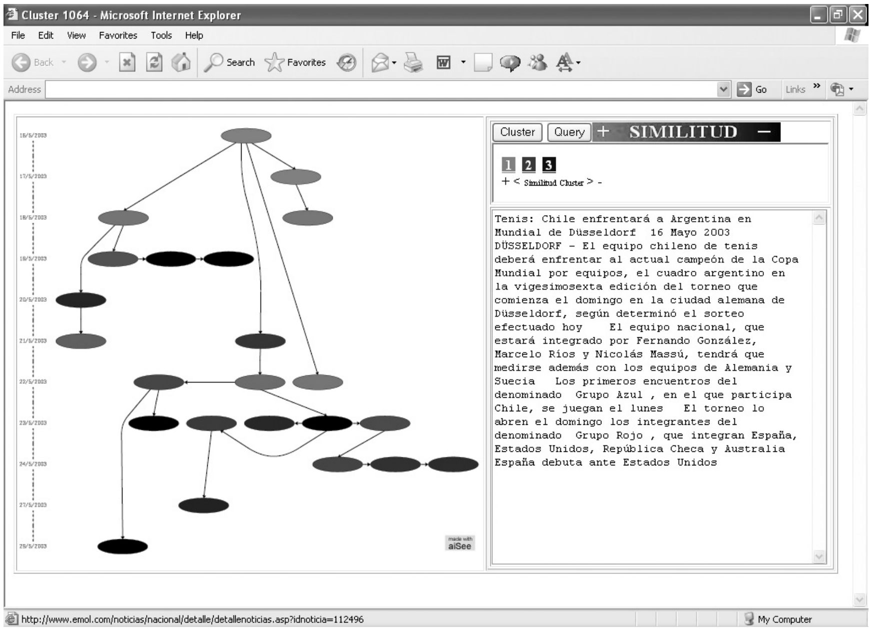


Fig. 1. Graph representation with highlighting similarity association for query "Chile Campeón Mundial de Tenis 2003"

query. Even more, since clusters associate documents based not only by keyword occurrences, but also temporal proximity, it may be possible to find thematically relevant documents that, otherwise, would be excluded from the answer with a traditional model of information retrieval.

The strategy for retrieving news documents from topic-based clusters is as follows. A user query is compared with each cluster based on a common representation. If the value of similarity is positive, and all terms in the query appear in the representation of the cluster, the cluster is selected for a second comparison between the query and each document in the cluster. The system ranks the clusters and documents such that it returns the cluster with highest similarity and, within this cluster, the documents with highest similarity values. In case that the query includes a condition expressed by a time interval, time interval of selected clusters, and then, of selected documents, must overlap the query's time interval.

A vector representation is applied to queries, clusters, and documents. The terms in the query are weighted by the same schema of the vector model. The terms in the news documents are weighted by the same schema of the topic detection algorithm. The representation of each cluster uses all terms in the documents of the cluster weighted by the *tf*idf* schema of the vector model, but with respect to the clusters. Thus, the occurrence of terms in the cluster is the number of times the term appears in the documents of the cluster. The size of



Tenis: Chile enfrentará a Argentina en
Mundial de Düsseldorf 16 Mayo 2003
DÜSSELDORF - El equipo chileno de tenis
deberá enfrentar al actual campeón de la Copa
Mundial por equipos, el cuadro argentino en
la vigesimosexta edición del torneo que
comienza el domingo en la ciudad alemana de
Düsseldorf, según determinó el sorteo
efectuado hoy   El equipo nacional, que
estará integrado por Fernando González,
Marcelo Ríos y Nicolás Massú, tendrá que
medirse además con los equipos de Alemania y
Suecia   Los primeros encuentros del
denominado Grupo Azul , en el que participa
Chile, se juegan el lunes   El torneo lo
abren el domingo los integrantes del
denominado Grupo Rojo , que integran España,
Estados Unidos, República Checa y Australia
España debuta ante Estados Unidos

Fig. 2. Online implementation of the system using the SVG format

the cluster is the total number of different terms in the cluster. Like in the topic detection algorithm, the values of *idf* are taken from the auxiliary corpus.

As an example of a search, Table 3 shows the results of the search with a query expressed by the following keywords: matute johns. This table indicates the clusters with highest similarity values and, within these clusters, the document with the highest similarity value. In addition, the fifth column indicates the number of documents in the the cluster. As additional information, the comparison with the manual classification of the topic ”caso matute johns” (topic id 13) indicates that the cluster 3090 has 4 cases of *false alarm* and 5 cases of *miss*, the latter belonging to the other clusters in the table.

As the Table 3 shows, the similarity values of the clusters may be low, since they are determined by considering all terms that are present in the documents of a cluster. Although the first cluster has the most similar value with respect to the query, the second cluster contains the document with the highest similarity value. When comparing the search over the clusters with respect to a search with the traditional vector model, the cluster with the highest similarity value gives 86% of the best ranked document obtained from the vector model. The other 14% of documents obtained from the vector model are found within the other selected clusters of the cluster-based search. When considering all found news in the selected clusters, the cluster-based search gives more results than the vector model, including cases of documents that were associated with the

Cluster ID	Similarity of cluster	Document in cluster	Similarity of document	Number of documents
3090	0.286	138604	0.378	45
730	0.228	108452	0.430	3
447	0.208	105314	0.323	3
3769	0.202	146719	0.269	2
883	0.21	110321	0.214	1
3607	0.186	144766	0.287	1
3983	0.184	148938	0.333	1
3317	0.179	140882	0.152	1

Table 3. Most similar clusters and document in these clusters for a search with keywords "matute johns"

topic, but with a lower weight for the query's keywords. Indeed, we obtained 17% of relevant news found by the cluster-based search that were not detected by the vector model.

5 Conclusions and future work

This work has presented a new topic detection algorithm for online news sources that includes text content, structure of documents and temporal content of news documents. The experimental results with the algorithm in a Spanish source of online news indicates a favorable improvement over a previous topic detection algorithm (UMASS TDT2). In addition to the topic detection algorithm, the work explored the use of this algorithm for search and navigation of news. It presents a graph-based navigation system that highlights the temporal and similarity association between news.

As future work, we expect to complete the implementation by handling updates of news in a online source. This is necessary if we expect to have good performance and avoid to compare arriving documents with all previous documents. Likewise, we expect to introduce indexing data structures that improve the search process. Within the context of future research, we want to explore the use of natural language techniques for analyzing explicit, implicit and vague temporal reference in the content of news documents. A broad domain of news with respect to geographic content would indicate whether or not the spatial content of news documents allows us to distinguish topics. Finally, we are exploring different strategies for navigating within a cluster such that a user could select, based on the information contribution of documents, what documents need to access.

Acknowledgment. This work has been funded by Nucleus Millenium Center for Web Research, Grant P04-067-F, Mideplan, Chile.

References

1. J. Allan, J. Carbonell, G. Doddington, J.Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *DARPA Broadcast News Trasncription and Understanding Workshop*, pages 194–218. [http://citeseer.ist.psu.edu/article/allan98topic.html], September 1998.
2. J. Allan, A feng, and A.Bolivar. Flexible intrinsic evaluation of hierarchical clustering for TDT. In *Twelfth International Conference on Information and Knowledge Management*, pages 263–270. ACM press, 2003.
3. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* ACM Press, 1999.
4. EMOL. El mercurio online [http://www.emol.com/].
5. P. Ferragina and A. Gulli. A perzonalized search engine based on web-snippet hierarchical clustering. In *International Conference in the World Wide Web WWW05*, pages 801–810, China, Japan, 2005. ACM Press.
6. F. Fukumoto and Y. Suzuki. Event tracking based on domain dependency. In *23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 24–28, Athens, Greece, 2000. ACM Press.
7. AbsInt Angewandte Informatik GmbH. GDL: aiSee graph visualization software: User manual unix version 2.2.07[http://www.aisee.com/manual/unix/], September 2005.
8. J. Makkonen, H. Ahonen-Myha, and M. Salmenkivi. Topic detection and tracking with spatio-temporal evidence. In Frabrizio Sebastini, editor, *Proceedings of 25th European Conference on Information Retrieval Research (ECIR 2003)*, pages 251–265. Springer-Verlag, 2003.
9. J. Mostafa. Seeking better web searches. *Scientific American Digital*, [ttp://www.sciam.com/], 2005.
10. S. Ram and S. G. Modeling and navigation of large information spaces: A semantic based approach. In *International Conference on System Science.* [http://computer.org/proceedings/hicss/0001/00016/00016020abs.htm], IEEE CS Press, 1999.
11. D. Roussinov and M. McQuaid. Information navigation by clustering and summary query results. In *International Conference on System Sciences*, page 3006. IEEE CS Press, 2000.
12. R.Papka, J. Allan, and V. Lavrenko. Umass approaches to detection and tracking at tdt2. In *Proceedings of the DARPA Broadcast News.* [http://www.nist.gov/speech/publications/darpa99/index.htm], 1999.
13. UMASS. *Topic Detection and Tracking TDT.* [http://ciir.cs.umass.edu/projects/tdt/index.html], 2005.
14. W3C. Scalable vector graphics (svg) 1.1 specification [http://www.w3.org/tr/svg/].
15. F. Walls, H. Jin, S. Sista, and R. Schwatz. Topic detection in broadcast news. In *Proceedings of the DARPA Broadcast News.* [http://www.nist.gov/speech/publications/darpa99/index.htm], 1999.
16. Y.Yang, J. Carbonelli, R. Brown, T. Pierce, B. Archibald, and X. Liu. Learning approaches for detection and tracking news events. *IEEE Intelligent Systems Special Isuue on Applications of Intelligent Information Retrieval*, 14:32–43, 1999.
17. Y. Zhang, X. Ji, C.-H Chu, and H. Zha. Correlating summarization of multi-source news with k-way graph bi-clustering. *ACM SIGKDD Explorations Newsletter*, 6(2):34–42, 2004.