Application of the Concept-Based Similarity Measure in Topic Detection

SATYA P KUMAR SOMAYAJULA[1] , DEEPTHI BODDHANA[2]
[1]*Asst. Professor, Department of CSE, Avanthi College of Engg & Tech, Tamaram, Visakhapatnam, A.P., India.*
[2]*Department of CSE, Avanthi College of Engg & Tech, Tamaram, Visakhapatnam, A.P.,India*

Abstract - **This paper introduces the Concept-based similarity measure into the Temporal -Semantic Clustering model for event detection in newspaper articles. The document similarity function is defined in terms of two similarity measures. Initially a context-based similarity measure that uses the vector of weighted terms is used to determine the similarity between the documents. Later another partial similarity measure that uses the vector of weighted time entities along with the previously determined concept-based measure to determine the combined similarity measure. Hierarchical approach is used to cluster documents based on this similarity measure.**

Keywords - **Temporal-Semantic Clustering, Concept-based clustering, Text Mining**

I. INTRODUCTION

The rapid rise in the number of news articles has made the task of discovering, tracking and organizing the topics associated with them difficult. For each incoming document detecting an event consists of determining whether it reports a new event or an older event that has occurred before. All the currently used Topic Detection and Tracking (TDT) systems use a document clustering algorithm along with the date of publication of the news articles. Identifying the topics available in the documents and grouping the events based on the temporal information available in them is the main aim of the TDT Systems. In this paper the hierarchical clustering algorithm presented in [1] is used . Documents that have high temporal – semantic similarity are grouped together based on the modified similarity function. This similarity function is based on the temporal - semantic similarity function of [1] which in turn uses the concept-based similarity function presented in [2] instead of using the cosine measure. In this paper the hierarchical clustering algorithm presented in [1] is used. At the initial level documents with high temporal –semantic similarity are grouped together using the modified document similarity function to detect the occurrence of an event. Grouping of documents at later levels will result in more complex events and topics being identified. This results in the documents being clustered basing on the content as well as temporal occurrences available in the documents.

II. DOCUMENT REPRESENTATION

The newspaper articles available in the documents are translated into XML documents which preserve the original logical structure of the documents so that the different thematic sections can be distinguished as well as different parts of the news. Each news document is represented using the following feature vectors:

- A vector of weighted terms. In the given document the sentences are separated, the stop words are removed and the stemming is also done. The contribution of each term to the semantics of the document at the sentence, document and corpus levels is determined by computing the conceptual frequency ctf, term frequency tf and document frequency df measures using the concept-based analysis algorithm as presented in [2].
- A vector of weighted time entities. A time entitiy may be either a date or a time interval expressed in Gregorian calendar. The dates are automatically extracted from the news texts by using algorithm presented in [Llid01]. This algorithm detects temporal sentences and translates them into time entities of a time model. Time entities are statistically weighted according to their frequency of references in the text.

Using the above each document is represented as follows:

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

III. CONCEPT-BASED SIMILARITY MEASURE

For the purpose of clustering the documents based on the term frequency values the cosine measure was used in [1]. However considering the advantages of concept-based similarity on the quality of clustering as presented in [2], the use of concept-based similarity measure is being proposed instead.

The concept based similarity between two documents d_1 and d_2 is calculated by

$$sim_c(d_1,d_2)= \sum_{i=1}^{m} \max\left(\frac{li_1}{Lvi_1},\frac{li_2}{Lvi_2}\right) \times weighti_1 \times weighti_2$$

The concept based weight of concept i in document d is calculated by

$$weight_i= (tfweight_i + ctfweight_i) \times \log\left(\frac{N}{dfi}\right)$$

The *tfweight*$_i$, *ctfweight*$_i$ values represent the weight of concept i in document d at document level and sentence level respectively.

The $\log\left(\frac{N}{df_i}\right)$ value rewards the weight of the concept i on the corpus level when concept i occurs in small no. of documents. The tf_{ij} and ctf_{ij} values are normalized by the length of the document vector of term frequency and conceptual term frequency of document d as

$$tfweight_i = \frac{tf_{ij}}{\sqrt{\sum_{j=1}^{cn}(tf_{ij})^2}}$$

$$ctfweight_i = \frac{ctf_{ij}}{\sqrt{\sum_{j=1}^{cn}(ctf_{ij})^2}}.$$

IV. TEMPORAL-SEMANTIC SIMILARITY MEASURE

The Temporal – Semantic Similarity measure is defined in [2] as

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

Here m_i is the number of time entities that describe the document d^i, $s\left(f_k^i, d^j\right)$ is a function that returns the most similar time entity to f_k^i 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:

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

The distance function *dist* between two time entities $f_1 \ and \ f_2$ depend on their nature, that is , whether they are dates or date intervals as defined in [1].
The other proposed measure for temporal component is based on the traditional distance between sets. This measure is defined as :

$$D_f(d^i,d^j) = \min_{f^i \epsilon FR^i, f^j \epsilon FR^j} \{d(f^i,f^j)\}$$

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 conditions given in [2].
In order to measure the global similarity between pairs of documents, two functions as defined in [1] are used:

1. If $S_T(d^i,d^j) \geq \beta_T$ and $S_F(d^i,d^j) \geq \beta_F$, then $S^1(d^i,d^j) = W_T . S_T(d^i,d^j) + W_F . S_F(d^i,d^j)$
else, $S^1(d^i,d^j) = 0$ where $W_T, W_F \ \epsilon \ [0,1]$ represent the relative importance of the different document components respectively. The thresholds $\beta_T, \beta_F \ \epsilon \ [0,1]$ are the minimum similarities required for the semantic and temporal components respectively.

2. If $D_F(d^i,d^j) \leq \beta_F$ then $S^2(d^i,d^j) = S_T(d^i,d^j)$ else, $S^2(d^i,d^j) = 0$
where β_F is the maximum number of days required to determine whether two documents refer to the same or two distinct events.

V. CLUSTERING

Incremental β_0-Compact Nucleus algorithm for event detection:
The clustering criteria is based on the following approach: Given a document description set, we must find or generate a natural structure for these documents in the representation space. These structure must be carried out by the use of some similarity measure between the documents based on certain property.
The clustering criteria has three parameters: a similarity measure S, a property π that establishes the use of S, and a *threshold* β_0. Thus, clusters are determined by imposing the fulfillment of certain properties over the similarities between documents.
The following definitions as presented in [1] are used:
Definition: Two documents d^i and d^j are β_0-similar if $S(d^i,d^j) \geq \beta_0$. Similarly, d^i is a β_0-isolated element if $\forall \ d^j \ \epsilon \ \zeta$, $S(d^i,d^j) < \beta_0$.
Definition : [Mart00]The set $NU \subseteq \zeta$, $NU \neq \phi$, is a β_0-compact nucleus if :

a) $\forall \ d^j \ \epsilon \ \zeta \ [d^i \epsilon \ NU \ \wedge \max_{d^j \ \epsilon \ \zeta, d^t \neq d^i} \{S(d^i,d^t)\} = S(d^i,d^j) \geq \beta_0 \ ] \Rightarrow d^j \ \epsilon \ NU$

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

c) $|NU|$ is the minimum.

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

Thus 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 β_0.

The Incremental β_0-Compact Nucleus algorithm presented in [1] is being used for incremental clustering of the documents.In this algorithm each document d^i has an *Info* field associated with it which contains the document or documents that more closely resemble d^i along with the value of the 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 along 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 algorithm can be described as follows:

```
Input:  Similarity threshold    , Similarity measure S and its parameters.
Output: Document clusters(                    )representing the identified events.
Step1.  Arrival of document d.
        MS=0, MD= Ø, Info(d)=(MD,MS)
Step2.  For each existing cluster G' do
              For each document d' in G' do
                    calculate the similarity S between d and d'.
                    If S     then
                    (Max,SimilMax)=Info(d').
                    If S   SimilMax then update Info(d') with the document d and Similmax
                    If S   MS then update MS with S and MD with d'.
Step3.  Create a new cluster G with the document d.
Step4.  If MS   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 clustering algorithm allows the finding of clusters with arbitrary shapes and is also independent of the arrival order of the documents.

Representation of Clusters:

When we apply the clustering criterion we obtain several clusters of news with 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.

VI. EVALUATION

A collection of 452 news articles published in the Spanish newspaper "El Pais " during June 1999 was used to evaluate the effectiveness of the clustering algorithm presented in [1]. For this purpose the system generated clustering results were compared with manually labeled events using the F1-measure [Rijs79] and Detection Cost [TDT89].The results had demonstrated the positive impact of using the time component in the quality of sytem-generated clusters.

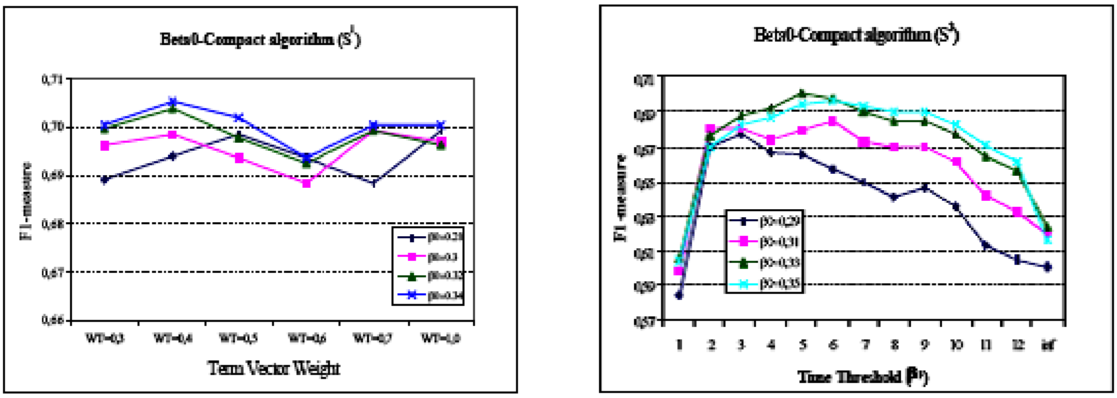


Fig1. F-1 measure results for the   –compact algorithm.

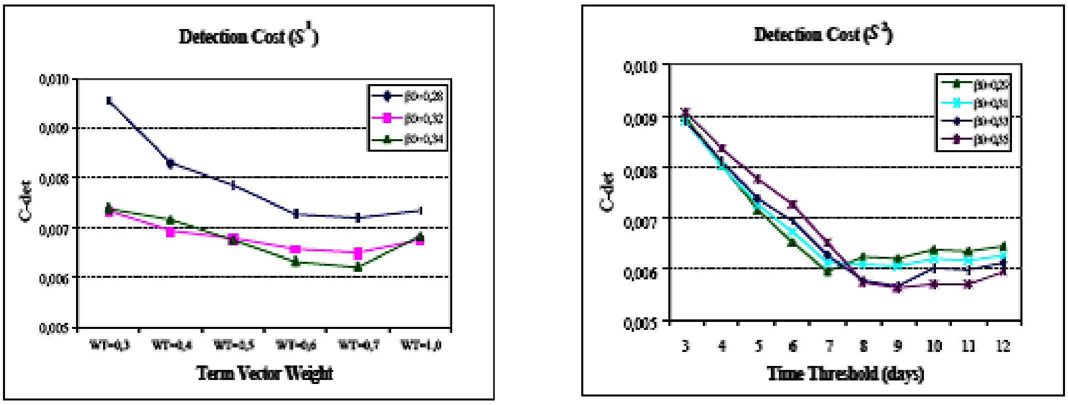
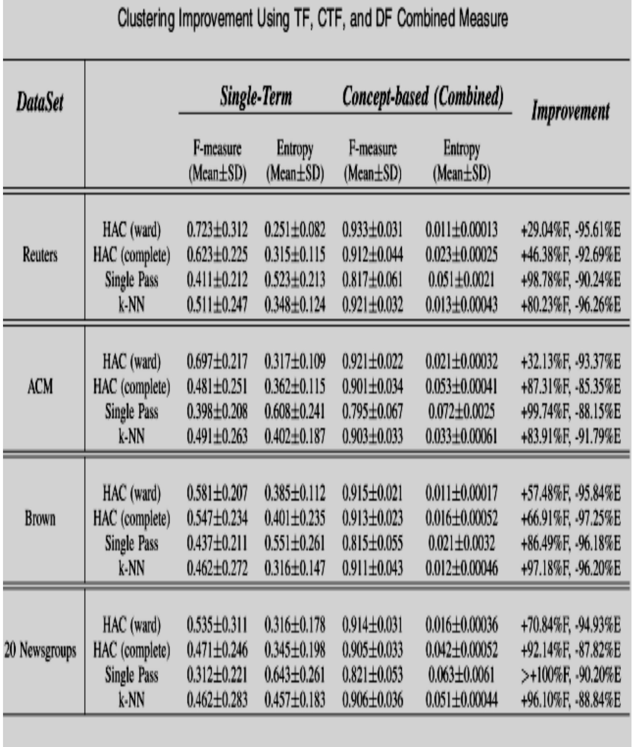


Fig2.Detection cost results for the   –compact algorithm.

The Concept-based mining model had significantly improved the quality of clustering as presented in [2].For this purpose four data sets were used. The first data set contained 23,115 ACM articles from the ACM library. The second data set contained 12,902 documents from the Reuters data set. The third data set contained 361 samples from the Brown corpus. The fourth data set contained 20,000 messages collected from 20 Usenet groups. The similarities calculated using the sentence-based, document-based, corpus-based and the combined approach concept analysis are used to compute four similarity matrices among the documents. Three standard document clustering techniques were chosen for testing the effect of concept-based similarity on clustering i.e., Hierarchical Agglomerative Clustering, Single-Pass Clustering and k-Nearest Neighbour. The F-measure and Entropy measures were used to evaluate the quality of clustering. The results indicate significant improvement in the quality of clustering as indicated by these tables below:



Clustering Improvement Using TF, CTF, and DF Combined Measure

DataSet		Single-Term		Concept-based (Combined)		Improvement
		F-measure (Mean±SD)	Entropy (Mean±SD)	F-measure (Mean±SD)	Entropy (Mean±SD)	
Reuters	HAC (ward)	0.723±0.312	0.251±0.082	0.933±0.031	0.011±0.00013	+29.04%F, -95.61%E
	HAC (complete)	0.623±0.225	0.315±0.115	0.912±0.044	0.023±0.00025	+46.38%F, -92.69%E
	Single Pass	0.411±0.212	0.523±0.213	0.817±0.061	0.051±0.0021	+98.78%F, -90.24%E
	k-NN	0.511±0.247	0.348±0.124	0.921±0.032	0.013±0.00043	+80.23%F, -96.26%E
ACM	HAC (ward)	0.697±0.217	0.317±0.109	0.921±0.022	0.021±0.00032	+32.13%F, -93.37%E
	HAC (complete)	0.481±0.251	0.362±0.115	0.901±0.034	0.053±0.00041	+87.31%F, -85.35%E
	Single Pass	0.398±0.208	0.608±0.241	0.795±0.067	0.072±0.0025	+99.74%F, -88.15%E
	k-NN	0.491±0.263	0.402±0.187	0.903±0.033	0.033±0.00061	+83.91%F, -91.79%E
Brown	HAC (ward)	0.581±0.207	0.385±0.112	0.915±0.021	0.011±0.00017	+57.48%F, -95.84%E
	HAC (complete)	0.547±0.234	0.401±0.235	0.913±0.023	0.016±0.00052	+66.91%F, -97.25%E
	Single Pass	0.437±0.211	0.551±0.261	0.815±0.055	0.021±0.0032	+86.49%F, -96.18%E
	k-NN	0.462±0.272	0.316±0.147	0.911±0.043	0.012±0.00046	+97.18%F, -96.20%E
20 Newsgroups	HAC (ward)	0.535±0.311	0.316±0.178	0.914±0.031	0.016±0.00036	+70.84%F, -94.93%E
	HAC (complete)	0.471±0.246	0.345±0.198	0.905±0.033	0.042±0.00052	+92.14%F, -87.82%E
	Single Pass	0.312±0.221	0.643±0.261	0.821±0.053	0.063±0.0061	>+100%F, -90.20%E
	k-NN	0.462±0.283	0.457±0.183	0.906±0.036	0.051±0.00044	+96.10%F, -88.84%E

Fig3. Clustering Improvement using tf, ctf and df combined measure.

As the above results indicate the temporal-semantic clustering has given better results used to cluster newspaper articles using the cosine measure for the term vectors. The concept based similarity measure also has improved the quality of clustering when the importance of the concept to the sentence semantics was considered at the sentence, document and corpus levels. We have tried to bring about the advantages of both the temporal aspect as well as the concept-based approach so that the quality of clustering will be enhanced furthermore if the concept-based similarity measure is used instead of the cosine measure for determining term vector similarity.

VII. CONCLUSIONS

We have tried to improve the quality of Temporal-Semantic clustering by introducing the concept-based similarity measure for determining the term vector similarity instead of the currently used cosine measure for document similarity. By bringing together the hierarchical clustering algorithm for Temporal-Semantic clustering and the Concept-based similarity measure for term vector similarity between documents the quality of clustering will be enhanced to a large extent which in turn enhances the quality of topic detection.

REFERENCES

[1] Temporal-Semantic Clustering of Newspaper Articles for Event Detection, Aurora Pons-Porrata, Rafael Berlanga-Llavori and Jose Ruiz-Shulcloper,"*Proceedings of Pattern Recognition In Information Systems*,2002" pg 104-113.

[2] An Efficient Concept-based Mining Model for enhancing Text Clustering, S.Shehata, F.Karray and M.Kamel, "*IEEE Trans. On Knowledge and Data Mining*",vol 22.no.10.,October 2010.

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

[4][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.

[5][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.

[6][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.

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

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

[9][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.

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

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

[12]S. Shehata, F. Karray, and M. Kamel, "*Enhancing Text ClusteringUsing Concept-Based Mining Model*," Proc. Sixth IEEE Int'l Conf.Data Mining (ICDM), 2006.