An Approach for Topic Trend Detection

Wolfgang Gaul and Dominique Vincent

Abstract The detection of topic trends is an important issue in textual data mining. For this task textual documents collected over a certain time period are analysed by grouping them into homogeneous time window dependent clusters. We use a vector space model and a straight-forward vector cosine measure to evaluate document-document similarities in a time window and discuss how cluster-cluster similarities between subsequent windows can help to detect alterations of topic trends over time. Our method is demonstrated by using an empirical data set of about 250 pre-classified time-stamped documents. Results allow to assess which method specific parameters are valuable for further research.

1 Introduction

In order to detect emerging topics (see, e.g., Allan et al. 1998, Kontostathis et al. 2004, and Kumaran et al. 2004) or to monitor existing topics it is of interest to analyse document streams (see, e.g., Wang et al. 2007 and Wang et al. 2009) which are emitted, e.g., by news sites like spiegel.de, zeit.de, or nytimes.com. For topic detection in text mining it is common to use document clustering (see, e.g., Allan et al. 1998, Larsen and Aone 1999, and Manning et al. 2009). Let us remind SMART, the System for the Mechanical Analysis and Retrieval of Text (see, e.g., Salton 1989) as an early example for the analysis and retrieval of information by computers.

In the next Sect. 2 notation and some background information will be provided. Section 3 describes the suggested approach while in Sect. 4 an example is presented

W. Gaul (✉) · D. Vincent
Institute of Decision Theory and Management Science, Karlsruhe Institute of Technology (KIT), Kaiserstr. 12, 76128 Karlsruhe, Germany
e-mail: wolfgang.gaul@kit.edu; dominique.vincent@kit.edu

B. Lausen et al. (eds.), *Algorithms from and for Nature and Life*, Studies in Classification, Data Analysis, and Knowledge Organization, DOI 10.1007/978-3-319-00035-0_35,

to demonstrate what can be expected from our topic trend detection technique. Section 5 contains concluding remarks.

2 Notation and Background Information

We need a dictionary which is created using a corpus (see, e.g., Allan et al. 2002) composed of a set of documents $d \in D$. For every term w in the dictionary this corpus is used to compute the term frequencies tf_w as well as inverse document frequencies idf_w given by

$$idf_w = \log \frac{|D|}{|\{d : w \in d\}|}$$

where $|M|$ denotes the cardinality of a set M.

The well-known vector space model (see, e.g., Salton et al. 1975) is applied for representing the text documents that we want to analyse. Vector components in the vector space are used to reflect the importance of corresponding terms from the dictionary. The dimension z of the vector space is crucial (the smaller the dimension of z can be chosen the faster the computation).

One of the best known weighting schemes is tf-idf weighting (see, e.g., Salton and Buckley 1988 and Allan et al. 2000) which we also examined – among others – for the underlying situation.

Finally, we applied the cosine measure as (dis)similarity between documents (document – document similarities) as well as between clusters of documents (cluster-cluster similarities). If C_k^t respectively C_l^{t+1} denote clusters of documents at subsequent time windows t and $t+1$, c_k^t the centroid of C_k^t with $c_{k_w}^t$ as vector component for term w of centroid c_k^t (where a centroid is just the average vector of the documents associated with the corresponding cluster) we have

$$cos(c_k^t, c_l^{t+1}) = \frac{\sum_{w=1}^{z} c_{k_w}^t * c_{l_w}^{t+1}}{\sqrt{\sum_{w=1}^{z} (c_{k_w}^t)^2} * \sqrt{\sum_{w=1}^{z} (c_{l_w}^{t+1})^2}}$$

for the cluster-cluster measure.

With M_t as set of documents in time window t we use the cosine similarity of documents to compute a $|M_t| \times |M_t|$ matrix of dissimilarities $dis^t(i, j)$, $i, j \in M_t$, between all documents of time window t from which we get a clustering $\mathcal{K}^t = \{C_1^t, \ldots, C_k^t, \ldots, C_{|\mathcal{K}^t|}^t\}$ by application of a hierarchical cluster analysis procedure together with the number of classes $|\mathcal{K}_t|$ (which is one of the reasons to use hierachical clustering).

With the clusterings $\mathcal{K}^t$ and $\mathcal{K}^{t+1}$ from two subsequent time windows we are able to compute the dissimilarities between the corresponding sets of clusters.

Table 1 Dissimilarity matrix w.r.t. $\mathscr{K}^t$ and $\mathscr{K}^{t+1}$

	C_1^{t+1}	...	C_l^{t+1}	...	$C_{K_{t,t+1}}^{t+1}$
C_1^t					
⋮					
C_k^t			$dis^{t,t+1}(C_k^t, C_l^{t+1})$		
⋮					
$C_{K_{t,t+1}}^t$					

Table 2 Dissimilarity matrix with missing values, a vanishing cluster C_k^t, and a newly arising cluster C_l^{t+1}

	C_1^{t+1}	...	C_3^{t+1}	...	...	C_l^{t+1}	...	C_5^{t+1}	...	$C_{K_{t,t+1}}^{t+1}$
C_1^t						$\min_{C \in \mathscr{K}^{t+1}} \{dis^{t,t+1}(C, C_l^{t+1})\} >$ threshold				
C_2^t										
⋮										
C_k^t	$\min_{C \in \mathscr{K}^{t+1}} \{dis^{t,t+1}(C_k^t, C)\}$							> threshold		
⋮						$\min_{C \in \mathscr{K}^t} \{dis^{t,t+1}(C, C_l^{t+1})\}$				
$C_{\lvert\mathscr{K}^t\rvert}$										
⋮	Missing values									
$C_{K_{t,t+1}}^t$										

The matrix of dissimilarities $dis^{t,t+1}(C_k^t, C_l^{t+1})$ (determined with the help of $cos(c_k^t, c_l^{t+1})$) has size $K_{t,t+1}$ which just is the maximum of $|\mathscr{K}^t|$ and $|\mathscr{K}^{t+1}|$ (cf. Table 1).

Assume that $|\mathscr{K}^t|$ is less than $|\mathscr{K}^{t+1}|$. In this case the rows of the dissimilarity matrix from $C_{|\mathscr{K}^t|+1}^t$ to $C_{K_{t,t+1}}^t$ have missing values (Likewise, if $|\mathscr{K}^t|$ is greater than $|\mathscr{K}^{t+1}|$ the columns from $C_{|\mathscr{K}^{t+1}|+1}^{t+1}$ to $C_{K_{t,t+1}}^{t+1}$ have missing values.) which indicates that the number of clusters from different time windows don't need to coincide.

In case a value $dis^{t,t+1}(C_k^t, C_l^{t+1})$ of a pair of clusters C_k^t and C_l^{t+1} in the dissimilarity matrix is "small" cluster C_k^t corresponds to cluster C_l^{t+1}, i.e., we assume that cluster C_k^t at time window t can be assigned to cluster C_l^{t+1}.

Additionally, it can happen that the minimum of the dissimilarities of cluster C_k^t to all clusters of $\mathscr{K}^{t+1}$ is greater than a predefined threshold from which one can conclude that cluster C_k^t is not similar to any of the clusters of time window

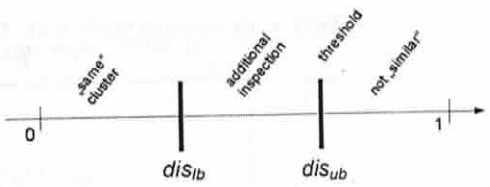

Fig. 1 *Lower* and *upper* bounds for dissimilarity checks

$t+1$, i.e., topic C_k^t has vanished (death in t) and is no longer in clustering $\mathcal{K}^{t+1}$. Another case appears if all dissimilarities in the column of C_l^{t+1} are greater than the threshold, i.e., C_l^{t+1} is a newly arising topic which was not in clustering $\mathcal{K}^t$ (birth or reappearance in $t+1$). These possibilities are depicted in Table 2.

Figure 1 tries to describe the underlying situation. When cluster-cluster dissimilarities between two clustes C_k^t and C_l^{t+1} of subsequent time windows are smaller than a problem-specific lower bound dis_{lb} it is assumed that the documents of C_k^t and C_l^{t+1} belong to the "same" cluster. However, if a problem-specific threshold dis_{ub} as an upper bound is exceeded by all cluster-cluster dissimilarities in a row (or column) of the matrix a trend has vanished (a new trend is born). In the area between dis_{lb} and dis_{ub} an additional inspection is necessary.

3 Approach

Given the explanations of the last section the following approach to support topic trend detection is suggested:

- Collect the set M_t of documents in time window t.
- Compute the $|M_t| \times |M_t|$ matrix of document-document dissimilarities $dis^t(i, j)$, $i, j \in M_t$.
- Perform hierarchical clustering to get $\mathcal{K}^t = \{C_1^t, \ldots, C_k^t, \ldots, C_{|\mathcal{K}^t|}^t\}$.
- With $K_{t,t+1} = max\{|\mathcal{K}^t|, |\mathcal{K}^{t+1}|\}$ compute the $K_{t,t+1} \times K_{t,t+1}$ matrix of cluster-cluster dissimilarities $dis^{t,t+1}(C_k^t, C_l^{t+1})$.
- Choose problem-specific dissimilarity-bounds and check for the birth (or reappearance) of topics, the death of no longer interesting topics, or the continuation of trends. In case that different lower and upper dissimilarity-bounds have to be considered additional inspection is needed to classify critical cases for which dissimilarities are situated within the bounds.

4 Example

Our test data set is a sample drawn from a set of time-stamped documents (see, e.g., Kupietz and Keibel 2009 and Kupietz et al. 2010) of the Institut für Deutsche Sprache IDS, located in Mannheim. The test documents are from newspapers

Table 3 Test configuration

$\|M_t\|$	52	65	50	35	52
time window t	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$
Cluster 1	C_1^1	C_1^2	C_1^3	C_1^4	C_1^5
Cluster 2	C_2^1	C_2^2	C_2^3	C_2^4	C_2^5
Cluster 3	C_3^1	C_3^2	C_3^3	–	C_3^5
Cluster 4	–	C_4^2		–	–

categorized by IDS into the four topics politics (P), sport (S), technique, industry, and transportation (TIT), and economy and finance (EF) which could be assigned to five time windows.

We used a dictionary with about 2 million terms and restricted our test runs to the 200, 2000, respectively 20,000 most frequent terms of that dictionary as dimension z of the vector space.

The test configuration of 254 documents is shown in Table 3. At time window $t = 1$ we had a subsample of 52 documents which could be assigned to three of the IDS topics. At time windows $t = 2$ and $t = 3$ the subsamples of 65 and 50 documents were from four respectively three topics. At time windows $t = 4$ and $t = 5$ two topics respectively three topics could be assigned. The next section will reveal which topics are hidden behind the general C_k^t-notation of Table 3.

5 Results

As writing restrictions do not allow to describe all results of the example we just explain the activities in the time windows $t = 1$ and $t = 2$ as well as the transitions between the time windows $1 \rightarrow 2$ and $3 \rightarrow 4$. We conclude with an overall view on topic trend detection situations.

5.1 *Transition Between Time Windows* $1 \rightarrow 2$

The Fig. 2(a), (b) show the dendrograms at time windows $t = 1$ and $t = 2$.

Three clusters at $t = 1$ and four clusters at $t = 2$ are marked by circles as interesting topics. In Table 4 the dissimilarity matrix between the clusterings $\mathscr{K}^1$ and $\mathscr{K}^2$ is shown. The marked cells with lowest dissimilarity values in the matrix indicate which document clusters are most similar to each other ($C_2^1 \leftrightarrow C_4^2$, $C_3^1 \leftrightarrow C_1^2$, $C_1^1 \leftrightarrow C_2^2$ although C_4^2 and C_1^1 have also a low dissimilarity). All values in the column of C_3^2 are "large", i.e., C_3^2 is a newly arising cluster, and row 4 has missing values.

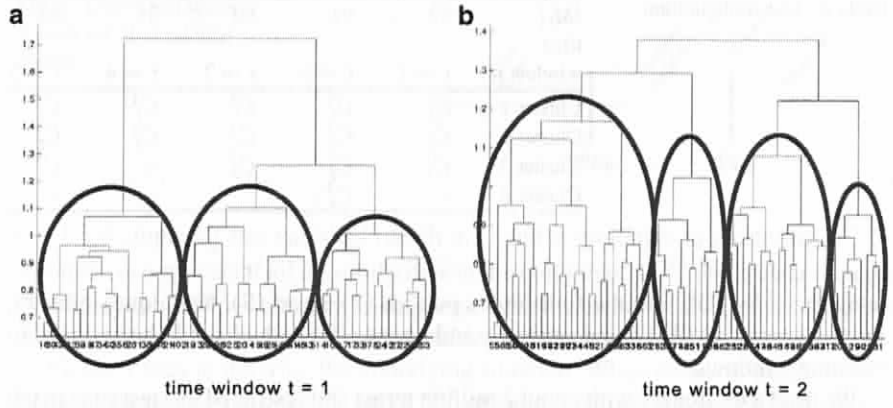

Fig. 2 Dendrograms. (**a**) Time window $t = 1$. (**b**) Time window $t = 2$

Table 4 Dissimilarity matrix w.r.t. $\mathscr{K}^1$ and $\mathscr{K}^2$

	C_1^2	C_2^2	C_3^2	C_4^2
C_1^1	0.4713	0.2807	0.3961	0.2866
C_2^1	0.5696	0.4144	0.4275	0.2210
C_3^1	0.2587	0.4170	0.5375	0.4909
C_4^1	Missing values			

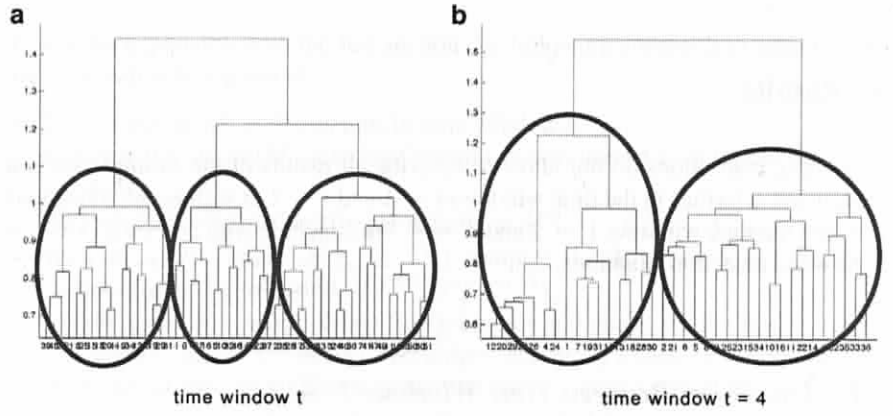

Fig. 3 Dendrograms. (**a**) Time window $t = 3$. (**b**) Time window $t = 4$

5.2 *Transition Between Time Windows* $3 \rightarrow 4$

Again, Fig. 3(a), (b) depict the dendrograms at time windows $t = 3$ and $t = 4$ together with the circles which show that a three-cluster-solution $\mathscr{K}^3$ and a two-

Table 5 Dissimilarity matrix w.r.t. $\mathscr{K}^3$ and $\mathscr{K}^4$

	C_1^4	C_2^4	C_3^4
C_1^3	0.3597	0.5690	Missing values
C_2^3	0.2634	0.5062	
C_3^3	0.3994	0.5574	

Table 6 General result

Time window t	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$
Cluster 1	P	S	EF	P	EF
Cluster 2	EF	P	P	S	P
Cluster 3	S	TIT	TIT		S
Cluster 4		EF			

cluster-solution $\mathscr{K}^4$ were chosen. This time we have $C_2^3 \leftrightarrow C_1^4$, all values in the column of C_2^4 are "large", i.e., C_2^4 is a newly arising cluster, and column 3 has missing values (because of $|\mathscr{K}^4| = 2$). Additionally, one can see that C_1^3 and C_3^3 will vanish (cf. Table 5).

5.3 *Overall View*

All in all we get the results depicted in Table 6 (see also Table 3).

Topic P (politics) exists in all time window dependent clusterings $\mathscr{K}^t$.

Topic S (sport) has vanished in time window $t = 3$, but reappeared in $t = 4$. To check whether a topic is newly arising in time window t we have to compare the centroid of that topic to the centroids of all clusters in the preceeding time windows $\tau \leq t - 2$. If we find a cluster in an earlier time window the dissimilarity of which to the actual cluster is smaller than the lower bound dis_{lb} we assume that the actual cluster is not new, if all dissimilarities are greater than the upper bound dis_{ub} we assume that a newly arising topic has been found.

The topic TIT (technique, industry, and transportation) is newly arising at time window $t = 2$ in our sample of documents but vanishes again in the time windows $t = 4$ and $t = 5$.

The chosen example was small on purpose to be able to demonstrate how the topic trend detection approach works where reappearance checks in earlier time windows are of importance in case that a topic is newly arising in a certain time window.

6 Conclusion

We described an approach for Topic Trend Detection and mentioned the problem to find an accurate threshold respectively lower and upper bounds between which an additional inspection should be performed.

The size of the vector space has an impact on the parameters mentioned. The greater the dimension of the vector space the more less frequent terms from the dictionary might have to be considered and the larger the threshold must be chosen.

References

Allan, J., Carbonell, J., Doddington, G., Yamron, J., & Yang, Y. (1998). Topic detection and tracking pilot study final report. In *Proceedings of the DARPA broadcast news transcription and understanding workshop*, Lansdowne (pp. 194–218).

Allan, J., Lavrenko, V., Frey, D. & Vikas, K. (2000). UMass at TDT 2000. In *Topic detection and tracking workshop notebook* (pp. 109–115).

Allan, J., Lavrenko, V. & Swan, R. (2002). Explorations within topic tracking and detection. In J. Allan (Ed.), *Topic detection and tracking: event-based information organization* (pp. 197–222). Norwell: Kluwer Academic.

Kontostathis, A., Galitsky, L., Pottenger, W. M., Roy, S., & Phelps, D. J. (2004). A survey of emerging trend detection in textual data mining. In M. W. Berry (Ed.), *A comprehensive survey of text mining - Clustering, classification, and retrieval*. New York: Springer.

Kumaran, G., Allan, J., & McCallum, A. (2004). Classification models for new event detection. In *International conference on information and knowledge management (CIKM2004)*. ACM.

Kupietz, M., Belica, C., Keibel, H., & Witt, A. (2010). The German reference corpus DEREKO: A primordial sample for linguistic research. In N. Calzolari, et al. (Eds.), *Proceedings of the 7th conference on international language resources and evaluation (LREC 2010)*, Valletta (pp. 1848–1854). Valletta, Malta: European language resources association (ELRA).

Kupietz, M., & Keibel, H. (2009). The Mannheim German reference corpus (DeReKo) as a basis for empirical linguistic research. In M. Minegishi, Y. Kawaguchi, (Eds.), *Working papers in corpus-based linguistics and language education*, No. 3 (pp. 53–59). Tokyo: Tokyo University of Foreign Studies (TUFS).

Larsen, B., & Aone, C. (1999). Fast and effective text mining using linear-time document clustering. In *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining KDD '99*, New York (pp. 16–22). ACM.

Manning, C.D., Raghavan, P., & Schuetze, H. (2009). *An introduction to information retrieval*. Cambridge: Cambridge University Press.

Salton, G. (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Boston: Addison-Wesley.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management, 24*(5), 513–523.

Salton, G., Wong, A., & Yang, C.S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613–620.

Wang, X., Jin, X., Zhang, K., & Shen, D. (2009). Mining common topics from multiple asynchronous text streams. In *International conference on web search and data mining*, New York (pp. 192–201). ACM.

Wang, X., Zhai, C., Hu, X., & Sproat, R. (2007). Mining correlated bursty topic patterns from coordinated text streams. In *International conference on knowledge discovery and data mining*, New York (pp. 784–793). ACM.