Adv Data Anal Classif
DOI 10.1007/s11634-016-0241-2

REGULAR ARTICLE

Evaluation of the evolution of relationships between topics over time

Wolfgang Gaul[1] · **Dominique Vincent**[1]

Received: 13 January 2015 / Revised: 20 January 2016 / Accepted: 20 February 2016

Abstract Topics that attract public attention can originate from current events or developments, might be influenced by situations in the past, and often continue to be of interest in the future. When respective information is made available textually, one possibility of detecting such topics of public importance consists in scrutinizing, e.g., appropriate press articles using—given the continual growth of information—text processing techniques enriched by computer routines which examine present-day textual material, check historical publications, find newly emerging topics, and are able to track topic trends over time. Information clustering based on content-(dis)similarity of the underlying textual material and graph-theoretical considerations to deal with the network of relationships between content-similar topics are described and combined in a new approach. Explanatory examples of topic detection and tracking in online news articles illustrate the usefulness of the approach in different situations.

Keywords Topic relationships · Topic trend detection · Text processing · Content-(dis)similarity · Information clustering

Mathematics Subject Classification 01-08 · 62H30 · 68M11 · 68P10 · 68U15 · 68W27 · 90C35 · 91C20

✉ Wolfgang Gaul
wolfgang.gaul@kit.edu

Dominique Vincent
dominique.vincent@kit.edu

[1] Institute of Information Systems and Marketing, Karlsruhe Institute of Technology, Zirkel 2, 76131 Karlsruhe, Germany

Published online: 11 March 2016

1 Introduction

Information overload has led to activities to install processing devices that help to find, structure and analyze important subject areas within the flow of information as well as track whether and how such subject areas may vary over time.

When information is made available textually together with time tags, the name document(s) will be used below. In such situations one can divide an interesting inspection period into time intervals, assign the (time-stamped) documents to time intervals, group content-similar documents of the same time interval into subsets, use the size of these subsets in order to measure the importance of the topics addressed in the assigned documents, and check how topics evolve over time.

To do this, considerations such as how the (dis)similarity of documents can be assessed, how content-similar document subsets can be determined and assigned to topics, and how relationships between content-similar document subsets in different time intervals can be traced (to mention just the important ones), have to be made accessible to computer processing.

Here, knowledge concerning science directions as text processing (e.g., how to determine weight vectors which represent the contents of documents), cluster analysis (e.g., how to group documents into content-similar clusters), and graph theory (e.g., how to visualize the network of the relationships between content-similar clusters) has to be combined for which short descriptions of salient aspects are given in the next Sect. 2 together with basic notations. Section 3 explains how one can deal with the network of relationships between content-similar topics in different time intervals via graph-theoretical tools, how topics emerge and wear off given the document clusterings in the underlying inspection periods, whether and how subgraphs in which content-similar topics are connected indicate that topics, seemingly unrelated in single time intervals, have similarities, and describes a new approach how to evaluate the evolution of relationships between topics over time. SPON (SPiegel ONline) documents are selected in Sect. 4 to illustrate the usefulness of the theoretical considerations and to show how the suggested approach works. In Sect. 5 concluding remarks together with a selection of literature concerning as well the 'topic detection and tracking' area as the related science directions of (in alphabetical order) clustering, graph theory, and text processing are presented.

2 Underlying situation and basic notations

2.1 Reference corpus and dictionary as prerequisites for linguistic research

With $|M|$ as notation for the cardinality of a set M consider as the starting point for this article a set of documents called reference corpus, $\mathcal{R} = \{d_1, \ldots, d_s, \ldots, d_{|\mathcal{R}|}\}$, in which a document $d_s = (w_{s_1}, \ldots, w_{s_b}, \ldots, w_{s_{R_s}})$ is designated as a list of R_s words or terms $w_{s_1}, \ldots, w_{s_{R_s}}$ contained in it. $\mathcal{R}$ has to be 'large' to allow linguistic research. For given $\mathcal{R}$ it is possible to derive a so-called local dictionary $\mathcal{L} = \{x \mid \exists\, d_s \in \mathcal{R} : x = w_{s_b}\}$ which contains all words or terms x of the reference corpus $\mathcal{R}$. Earliest and most recent documents in $\mathcal{R}$ describe a time frame for the interpretation of text processing results. When new documents appear, which are of importance for inclusion in the

reference corpus, new terms might appear as well, thus, reference corpus and local dictionary may have to be adjusted over time in order to remain up-to-date.

2.2 Aspects of text processing

Well-known from the text processing literature is that for every term x and every document d_s one can determine characteristics, e.g., the term frequency tf_{x,d_s} of term x based on document d_s, a normalization $\overline{tf}_{x,d_s}$ of tf_{x,d_s}, a general term frequency tf_x, an inverse document frequency idf_x (in mathematical notation: $tf_{x,d_s} = \sum_{b=1}^{R_s} \delta_{\{w_{s_b}=x\}}$, $\overline{tf}_{x,d_s} = tf_{x,d_s}/\max_x\{tf_{x,d_s}\}$, $tf_x = \sum_{s=1}^{|\mathcal{R}|} tf_{x,d_s}/\sum_{s=1}^{|\mathcal{R}|} R_s$, $idf_x = log(|\mathcal{R}|/|\{d_s|\exists\, b : w_{s_b} = x\}|)$ with δ as Kronecker delta), and remove certain terms (e.g., less frequent and less important ones) to reduce $\mathcal{L}$ to a smaller dictionary $\mathcal{L}' \subset \mathcal{L}$. Of course, interpretability / recovery of the contents of the documents and the size $Z = |\mathcal{L}'|$ of the reduced dictionary (which influences computational speed) have to be balanced when a document d is represented by a weight vector $v^d = (v_1^d, \ldots, v_z^d, \ldots, v_Z^d)'$ where in this article $v_z^d = idf_z \cdot \overline{tf}_{z,d}$ has been used as weight of term z.

The computation of term weights based on a suitable $\mathcal{R}$ (and $\mathcal{L}'$) is necessary to allow a comparison of any documents d_i, d_j for which the cosine measure $\cos(v^{d_i}, v^{d_j})$ as one of the most often used tools of describing document-document similarity can be applied.

2.3 Information clustering within a time interval *t*

Since huge interesting document sets possibly have to be examined the application of cluster analysis can help to find subsets of content-similar documents. Let D be an underlying set of documents and $D^t \subset D$ the set of documents assigned to time interval t. In order to find content-similar subsets of documents in time interval t, standard practice, well-known from the cluster analysis literature, is as follows: compute the $|D^t| \times |D^t|$ matrix of document-document dissimilarities $dis^t(i, j) = 1 - \cos(v^{d_i}, v^{d_j})$, $d_i, d_j \in D^t$, apply hierarchical clustering to the documents of D^t to determine a family of clusterings hierarchically ordered according to the number of clusters, use the elbow-criterion to achieve a reasonable estimate for the 'right' number of clusters, and select the corresponding clustering $\mathcal{K}^t = \{C_1^t, \ldots, C_k^t, \ldots, C_{|\mathcal{K}^t|}^t\}$, where the subset of documents $C_k^t \subset D^t$ describes cluster k in time interval t. As textual material is clustered the labels 'information clustering' and 'document clustering' are used interchangeably. Let $c_k^t = (\ldots, c_{k_z}^t, \ldots)'$ denote the centroid of the weight vectors of the documents of C_k^t.

2.4 Dissimilaritiy matrix determination between clusters in different time intervals

For different time intervals t, T

$$\cos(c_k^t, c_l^T) = \frac{\sum_{z=1}^{Z} c_{k_z}^t \cdot c_{l_z}^T}{\sqrt{\sum_{z=1}^{Z} (c_{k_z}^t)^2} \cdot \sqrt{\sum_{z=1}^{Z} (c_{l_z}^T)^2}}$$

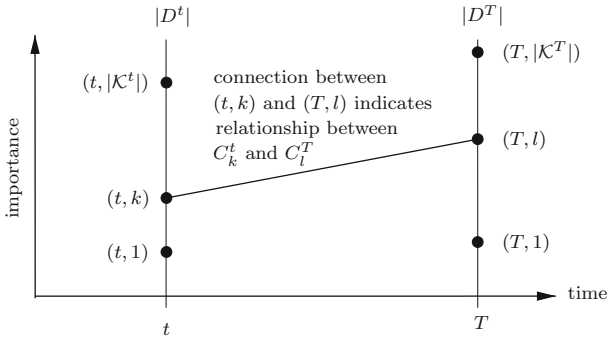

Fig. 1 Time-importance diagram for cluster relationship description between two time intervals t, T

is suggested to describe cluster-cluster similarities of clusters C_k^t, C_l^T. Notice that $dis^t(i, j)$ used for document clustering in time interval t and

$$dis((t, k), (T, l)) = 1 - \cos(c_k^t, c_l^T)$$

which depicts the relationship between clusters C_k^t, C_l^T in different time intervals both rely on the cosine measure but explain kinds of dissimilarity that have to be distinguished. When the cluster-cluster dissimilarity $dis((t, k), (T, l))$ for two clusters C_k^t, C_l^T is smaller than a problem-specific lower bound dis_{lb}, one assumes that the contents of the clusters are related, when a problem-specific upper bound dis_{ub} is exceeded the corresponding clusters are assumed to have no relationship. For $dis((t, k), (T, l))$ in the interval between dis_{lb} and dis_{ub} an additional inspection of the documents of the clusters involved may be necessary to decide whether a relationship exists.

2.5 Graph-theoretical time-importance description of topic relationships and term clouds for topic illustration

With $imp(t, k) = |C_k^t|/|D^t|$ the importance of the documents of C_k^t within the set D^t of all documents in time interval t can be assessed and time-importance diagrams can be drawn to describe the relationships between clusters in different time intervals where re-numberings of the clusters according to size in the single time intervals may be necessary. Figure 1 is an example for such a time-importance diagram and depicts how the relationships between document clusters in two time intervals t, T can be visualized using a bipartite graph.

In order to support the assignment of a topic label $Top(C_k^t)$ to a cluster C_k^t of documents, a term cloud (tag cloud in computer science language) featuring the hundred highest weighted terms of the centroid of the cluster, is created to provide an intuitive, illustrative, and rapid inspection of the subject area tackled in C_k^t, see, e.g., Figs. 4, 6, and 8 in the application Sect. 4.

As a bipartite graph allows only a description of relationships of topics between two time intervals, for a more general visualization and evaluation of topic trends let

$G(N, A)$ denote a graph with $N = N(G)$ as set of nodes and $A = A(G)$ as set of arcs of graph G where nodes (t, k) correspond to topics $Top(C_k^t)$/ clusters C_k^t and arcs $((t, k), (T, l))$ indicate that the corresponding topics are related.

3 Methodology

3.1 Preliminaries

For two time intervals t, T with $t < T$ a bipartite graph as depicted in Fig. 1 is used for the time-importance diagram description. However, if one wants to check an inspection period $[t_1, t_3]$ which includes more than two time intervals, a possibly 'much larger' graph $G_{[t_1,t_3]}$ which describes the topic relationships with respect to $[t_1, t_3]$ is needed with nodes (t, k_t) for all time intervals $t \in [t_1, t_3]$ and clusters $k_t \in \{1, \ldots, |\mathcal{K}^t|\}$. Since in addition to the document clusterings in the single time intervals cluster-cluster dissimilarity matrices

$$(dis((t_3 - \gamma - \chi, k_{t_3-\gamma-\chi}), (t_3 - \gamma, k_{t_3-\gamma}))),$$
$$\chi = 1, \ldots, t_3 - t_1 - \gamma, \quad \gamma = 0, \ldots, t_3 - t_1 - 1,$$

describing the relationships between the clusters of all pairs of the involved time intervals now have to be determined, the length of the inspection period can become crucial. An intuitive idea is to use graphs $G_{[t_1+\tau,t_2+\tau]}$, $t_1 < t_2 < t_3$, with shorter inspection period length $\lambda = t_2 - t_1 + 1$ for all $\tau = 0, \ldots, t_3 - t_1 - \lambda + 1$, (see, e.g., Fig. 2 in which a graph $G_{[t_1,t_3]}$ is described by subgraphs $G_{[t_1+\tau,t_2+\tau]}$ where $t_2 = \frac{t_1+t_3}{2}$ was selected for convenience) and try to combine 'smaller' subgraphs to construct 'larger' graphs in order to explain phenomena which occur in topic relationships based on inspection periods covering a considerably longer time span. With the notation $\cup$ for the union and $\oplus$ for the concatenation of graphs one can check 'smaller' graphs and use their combinations for relationship evaluation in 'larger' inspection periods where concatenation describes the special situation that the node sets of the graphs overlap only in one time interval. Note, however, that in Fig. 2, e.g., $G_{[t_1,t_3]}$ and the concatenation $G_{[t_1,t_2]} \oplus G_{[t_2,t_3]}$ are different, as arc $((t_1 + 1, 3), (t_2 + 1, 2)) \in A(G_{[t_1+1,t_2+1]})$ (see Fig. 2b) is not contained in the arc set $A(G_{[t_1,t_2]} \oplus G_{[t_2,t_3]})$ (see Fig. 2a, e), but in $A(G_{[t_1,t_3]})$. An example in which the union of graphs with overlapping node sets in more than one time interval is used for topic relationship evaluation is described in Sect. 4.2.2.

Additional aspects of interest for a graph-theoretical description of topic relationships are: A path $P((t, k), (\rho, l))$, $t, \rho \in [t_1, t_3]$, is a sequence of arcs $(((t, k), (t_\alpha, k_{t_\alpha})), ((t_\alpha, k_{t_\alpha}), (t_\beta, k_{t_\beta})), \ldots, ((t_\mu, k_{t_\mu}), (t_\nu, k_{t_\nu})), ((t_\nu, k_{t_\nu}), (\rho, l)))$. Topics $Top(C_k^t)$ and $Top(C_l^\rho)$ can be characterized as content-similar if they are connected via a path $P((t, k), (\rho, l))$.

Although the underlying graphs are undirected (the relationship between two nodes is symmetric), the position of the time coefficient t in the description of a node (t, k) can be important.

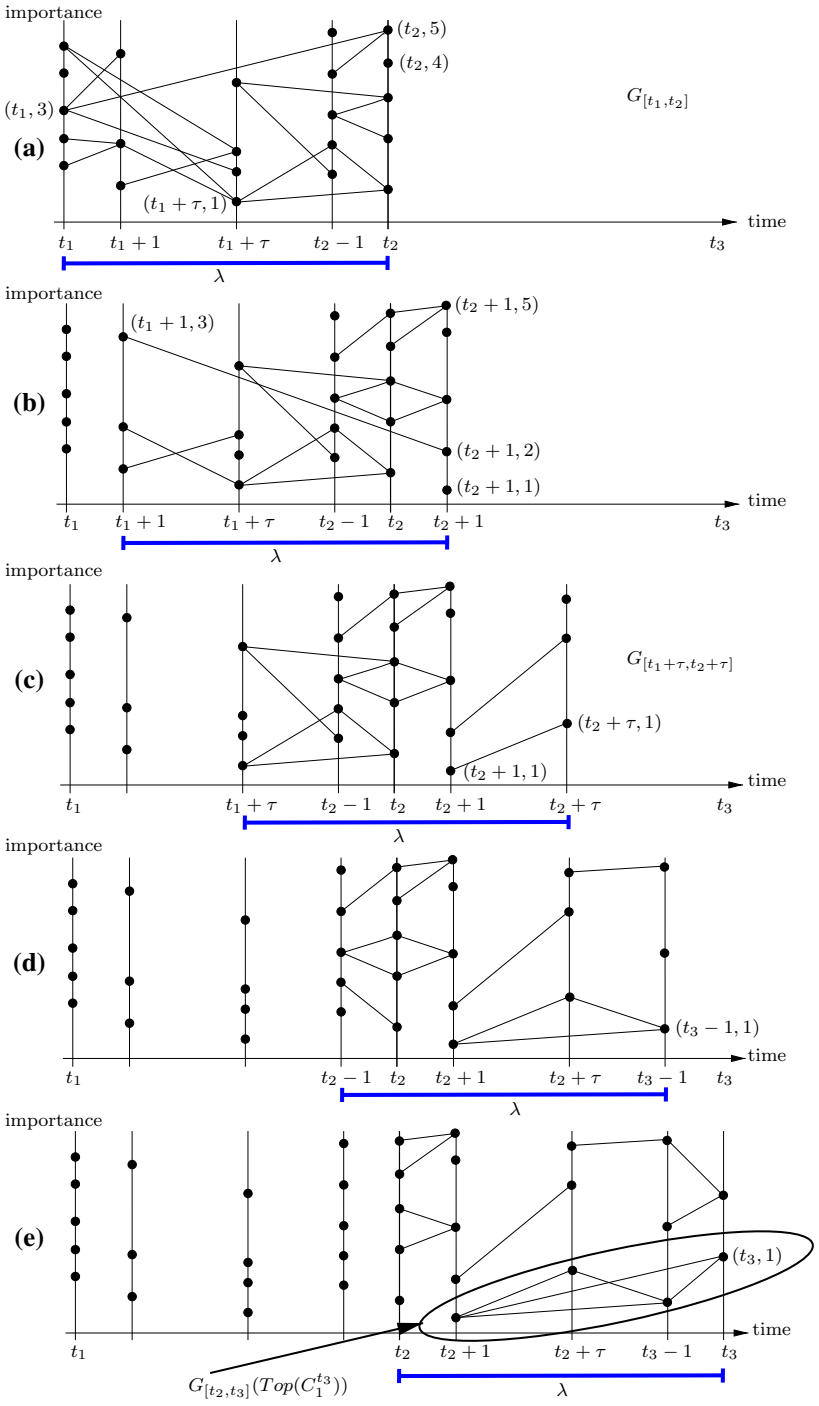

Fig. 2 Evolution of topic relationships over time

A topic $Top(C_k^t)$ is called an *origin* in $G_{[t_1+\tau,t_2+\tau]}$ if there is no path in $G_{[t_1+\tau,t_2+\tau]}$ from a content-similar topic in the past of t to $Top(C_k^t)$ ($\{P((\rho,l),(t,k))|t_1+\tau \leq \rho < t\} = \emptyset$).

A topic $Top(C_k^t)$ is called a *close* in $G_{[t_1+\tau,t_2+\tau]}$ if there is no path in $G_{[t_1+\tau,t_2+\tau]}$ from a content-similar topic in the future of t to $Top(C_k^t)$ ($\{P((t,k),(\rho,l))|t < \rho \leq t_2+\tau\} = \emptyset$).

A topic $Top(C_k^t)$ is called a *merger* in $G_{[t_1+\tau,t_2+\tau]}$ if it is directly related to more than one of the topics of $G_{[t_1+\tau,t_2+\tau]}$ in the past of t ($|\{((\rho,l),(t,k))|t_1+\tau \leq \rho < t\}| > 1$).

A topic $Top(C_k^t)$ is called a *split* in $G_{[t_1+\tau,t_2+\tau]}$ if it is directly related to more than one of the topics of $G_{[t_1+\tau,t_2+\tau]}$ in the future of t ($|\{((t,k),(\rho,l))|t < \rho \leq t_2+\tau\}| > 1$).

When checking relationships between content-similar topics one wants, e.g., to track the evolution of topic trends over time and, here, it is of interest to find those topics in which news appear for the first time (origins) or in which topic trends end (closes). Topics which are directly related to several topics in the past (mergers) or in the future (splits) have bundling properties which might also be helpful in the evaluation of the evolution of topics over time with the help of graph-theoretical tools.

Normally, $G_{[t_1+\tau,t_2+\tau]}$ contains several subgraphs of content-similar topics which are not connected to each other. These subgraphs are marked as $G_{[t_1+\tau,t_2+\tau]}(Top(\cdot))$ as they can be distinguished by a topic $Top(\cdot)$ to which all the topics assigned to their nodes are content-similar (Note that any of these topics can be selected to mark the corresponding subgraph. We use a close in such a connected subgraph of content-similar topics for the marking.). Subject areas of interest as well as single topics are constrained by the corresponding connected subgraphs of content-similar topics and the chosen inspection periods $[t_1+\tau, t_2+\tau]$ and can vary when τ changes.

A topic can simultaneously be a merger and a split as $Top(C_1^{t_1+\tau})$ in $G_{[t_1,t_2]}$ (see Fig. 2a). However, properties can appear and vanish due to the limited and changing inspection periods, e.g., $Top(C_1^{t_1+\tau})$ is still a split in $G_{[t_1+1,t_2+1]}$ (see Fig. 2b) but changes to an origin in $G_{[t_1+\tau,t_2+\tau]}$ (see Fig. 2c). $Top(C_1^{t_2+1})$ and $Top(C_1^{t_2+\tau})$ are content-similar in $G_{[t_1+\tau,t_2+\tau]}$ (see Fig. 2c) and evolve into members of a larger connected subgraph $G_{[t_2,t_3]}(Top(C_1^{t_3}))$ with additional topics $Top(C_1^{t_3-1})$ and $Top(C_1^{t_3})$ in which $Top(C_1^{t_2+1})$ is an origin (see, e.g., Fig. 2e). Even in $G_{[t_1,t_2]} \oplus G_{[t_2,t_3]}$ this subgraph $G_{[t_2,t_3]}(Top(C_1^{t_3}))$ is not connected to other subgraphs and is an example for which the marking $G_{[t_2,t_3]}(Top(C_1^{t_3}))$ or even $G_{[t_1,t_2]} \oplus G_{[t_2,t_3]}(Top(C_1^{t_3}))$ makes sense. It describes a set of content-similar topics as connected subgraph separated from all other topics depicted in Fig. 2, and shows how—when evaluating the evolution of topic relationships over time—one can and has to concentrate (because of the possible magnitude of the set of all relationships) on connected subgraphs of content-similar topics of current interest.

Another important aspect is that topics seemingly unrelated in a single time interval as, e.g., $Top(C_4^{t_2})$ and $Top(C_5^{t_2})$ (see Fig. 2a), can be content-similar because they are connected by a path via $Top(C_5^{t_2+1})$ (see Fig. 2b). Such a situation can arise when topics in time intervals different from the time interval t, in which the elbow criterion of hierarchical clustering was the reason for the selection of $\mathcal{K}^t = \{C_1^t, \ldots, C_{|\mathcal{K}^t|}^t\}$, are content-similar with respect to different clusters in $\mathcal{K}^t$ and connected by a path. Of

course, content-similarity has to be relativized, e.g., the length of a path connecting two topics can play an important role.

In Fig. 2, as an example how topic relationship evaluation with the help of graph-theoretical tools is possible, a 'look-forward' description (from t_1 into the future of t_1 up to t_3) is depicted. Conversely, one can start from a topic $Top(\cdot)$ of interest and apply a 'look-back' strategy to find content-similar topics in the past of $Top(\cdot)$. It should be noted, however, that depending on the chosen inspection period length λ, arcs $((t, k), (\rho, l))$ with $|t - \rho| > \lambda$ are not visible in Fig. 2.

3.2 Approach

Given the explanations so far the following approach to support the evaluation of the evolution of relationships between topics over time is suggested (This approach describes only the steps which have to be carried out. Algorithms/computer routines selected in single steps are not mentioned as they are not needed in order to understand the approach):

Let D denote an underlying set of documents in the starting situation, T the present time interval, and λ the length of the selected inspection period. Assume that $D^T \subset D$ and that document clusterings $\mathcal{K}^t = \{C_1^t, \ldots, C_{|\mathcal{K}^t|}^t\}$, $t \in [T - \lambda, T - 1]$, are already available.

[1] Select the set D^T of documents assigned to time interval T.
[2] Compute the $|D^T| \times |D^T|$ matrix of document-document dissimilarities $(dis^T(i, j),\ d_i, d_j \in D^T)$.
[3] Perform hierarchical document clustering to get $\mathcal{K}^T = \{C_1^T, \ldots, C_{k_T}^T, \ldots, C_{|\mathcal{K}^T|}^T\}$.
[4] Determine the matrices of cluster-cluster dissimilarities $(dis((T-\gamma-\chi, k_{T-\gamma-\chi}), (T-\gamma, k_{T-\gamma})))$, $\chi = 1, \ldots, \lambda-\gamma-1,\ \gamma = 0, \ldots, \lambda-2$, where matrices known from the past of T do not have to be recalculated.
[5] Construct $G_{[T-\lambda+1,T]}$ and check the topic relationships for content-similar subgraphs $G_{[T-\lambda+1,T]}(Top(C_{k_T}^T))$.
[6] Set $T := T + 1$, go to [1] (note that when the underlying set D contains only documents from the time interval T and the past of T, one has to wait until the new document set D^{T+1} is available).

As the just mentioned approach constitutes an essential part of the underlying paper, a verbalized explanation is added: The set D^T of documents of time interval T is provided (step [1]) for the computation of dissimilarities between the documents of this time interval (step [2]), which is standard in text processing.

Based on these document-document dissimilarities, hierarchical clustering together with the application of the elbow-criterion is used to determine a 'reasonable' number of clusters in time interval T (step [3] in which also the centroids of the obtained clusters have to be calculated).

As clusterings for the time intervals $t \in [T - \lambda, T - 1]$ are assumed to be available (i.e., steps [1], [2], and [3] have already been applied to the time intervals of the preceding inspection period $[T - \lambda, T - 1]$), now, the cluster-cluster dissimilarities between the clusters of all pairs of the relevant time intervals (remember Fig. 1 which explains

how the cluster-cluster relationships between two time intervals can be visualized) are determined (step [4]), again, by applying a standard procedure from text processing. This step can become cumbersome depending on the chosen inspection period length.

Based on the matrices of cluster-cluster dissimilarities provided for the inspection period $[T-\lambda+1, T]$, the graph $G_{[T-\lambda+1,T]}$ is constructed and, selecting interesting topics, e.g., closes $Top(C_{kT}^{T})$ of this graph, the subgraphs $G_{[T-\lambda+1,T]}(Top(C_{kT}^{T}))$ of content-similar topics are checked (step [5]). It should be kept in mind that the same connected subgraph can be denoted in different ways depending on the fact which $Top(\cdot)$ of the content-similar topics of this subgraph was selected for marking. Finally, it is assumed that the approach proceeds from time interval T to time interval $T+1$ (step [6]).

In steps [4] and [5] 'look-back'-strategies (from the current time interval T) are applied. When the approach moves on to $T+1, T+2, \ldots, T+\tau, \ldots$, relationships to older topics are increasingly no longer in the focus of the evaluation unless the inspection period length is extended. Note, however, that within a 'look-back' strategy interesting topics or subject areas or phenomena may be tracked down for which one wants to check in the past whether content-similar topics have already arisen earlier. In this case, one can just reset the approach to a time interval $T'(< T)$ which, then, serves as starting point to search for roots of these phenomena. Here, it may be necessary to extend the inspection periods to more than λ time intervals. In such a situation the concatenation and/or union of adequate subgraphs, which may cover different inspection period lengths, come into play.

The formulation of the approach possesses generality. Starting with whatever set D of documents and subset $D^{T} \subset D$ of time interval T, the approach finds an interesting subgraph of content-similar topics not connected to other subgraphs either in one inspection period (see Sect. 4.2.1) or the union of interesting subgraphs in more than one inspection period, when restarts are performed to check longer time spans in the past (see Sect. 4.2.2). When the search for interesting content-similar topics in the past is unsuccessful with inspection periods of length λ, the approach can be repeated with longer inspection periods (see Sect. 4.2.3).

These possibilities are explained in more detail in the next section in which examples depicting known political events are described on the basis of online news documents.

4 Topic relationship evaluation

4.1 Reference corpus and SPON (SPiegel ONline) articles

DeReKo, the reference corpus of the Institut für Deutsche Sprache, IDS, located in Mannheim (see, e.g., Kupietz 2009; Kupietz et al. 2010) together with its local dictionary (about 2 million terms) and the weights belonging to it, have been used as the basis.

As documents for the description of demonstrative examples, which show how the evaluation of topic relationships can be tackled, different sets of SPON (SPiegelONline) articles were available. 'Der Spiegel' as one of the best known German political magazines, also internationally, launched 'Spiegel Online' with separate and inde-

pendent editorial staff in 1994. For this article 179,313 crawled SPON documents were considered. First, we used samples to check how results concerning clustering and topic relationship evaluation react to time interval lengths (14 days, 10 days, 7 days), and, finally, selected for convenience calender weeks as time intervals because differences in the overall comparison of the samples were not substantial. For the inspection period length, normally 12 time intervals (about 3 months) were selected. For the document clusterings in the single time intervals the number of clusters used in this evaluation was restricted to a maximum of 15. Additionally, the choice of an appropriate dimension of the reduced dictionary $\mathcal{L}'$ was examined. The following examples are based on $Z = |\mathcal{L}'| = 20{,}000$.

4.2 Examples

Out of the different categories of SPON documents, 'politics' (34,222 articles between December 29, 2008 and July 3, 2011) was selected for two reasons: (1) More than 100 contributions per single calender week were available with exceptions only between Christmas and the New Year. (2) Examples based on political events are probably known to larger audiences.

Notice that for a single inspection period of λ time intervals (calender weeks) in the 'politics' category more than $\lambda \cdot 100$ documents had to be checked. Besides the document clusterings in the time intervals $\frac{\lambda(\lambda-1)}{2}$ cluster-cluster dissimilarity matrices (with dimensions of less than or equal to 15×15) had to be computed per inspection period of length λ.

From the many content-similar subgraphs which were evaluated the following three examples are described in more detail: Example 4.2.1 depicts an extremely small subgraph within an inspection period of 12 time intervals for which the topic assignment task is easy. In example 4.2.2 it is shown how the union of two content-similar subgraphs from different but overlapping inspection periods of 12 time intervals each helps to ascertain a topic trend of overriding importance. Finally, example 4.2.3 describes a situation in which about a year has to be examined to reveal relationships between content-similar topics which establish an overriding topic trend. Here, either an inspection period with the considerable length of nearly a year (with the relevant computational burden) has to be used or a sequence of shorter inspection periods of different lengths togeher with the concatenation and/or union of adequate subgraphs.

4.2.1 Resignation of a Minister of Defense

This is an example from early 2011 in which content-similar topics in a subject area of considerable interest at that time can be found in only three time intervals 112, 113, 114 (7. KW 2011, 8. KW 2011, 9. KW 2011; KW $\widehat{=}$ calendar week), see Fig. 3 which is based on an inspection period of $\lambda = 12$ time intervals. $Top(C_{13}^{112})$ describes the origin and $Top(C_4^{114})$ the close in the small subgraph $G_{[106,117]}(Top(C_4^{114}))$. Term clouds of these topics are depicted in Fig. 4 and show that the contents of the documents assigned to the selected clusters are similar although the viewpoints of the author(s)

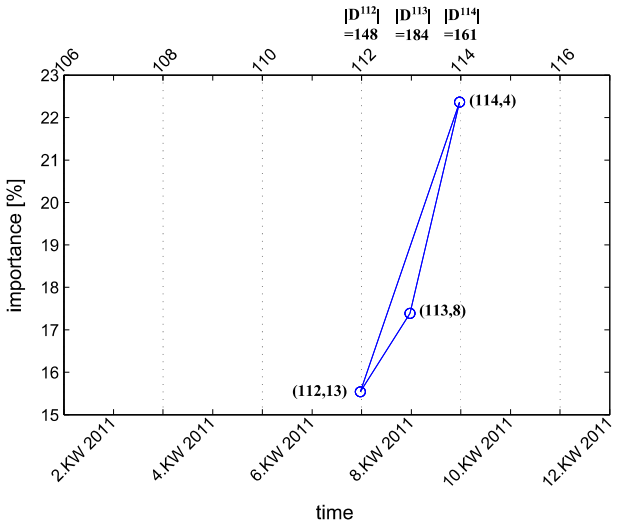

Fig. 3 Content-similar subgraph $G_{[106,117]}(Top(C_4^{114}))$ concerning the resignation of a Minister of Defense

Fig. 4 Term clouds for **a** the origin $Top(C_{13}^{112})$ and **b** the close $Top(C_4^{114})$ of $G_{[106,117]}(Top(C_4^{114}))$

of each document may differ from document to document. With $|D^{112}| = 148$, $|D^{113}| = 184$, $|D^{114}| = 161$ the importance of the topics in this subject area is increasing ($imp(112, 13) = 15,54 \times 10^{-2}$, $imp(113, 8) = 17,39 \times 10^{-2}$, $imp(114, 4) = 22,36 \times 10^{-2}$). Although 66 cluster-cluster matrices with corresponding dissimilari-

ties between the document clusters of the involved time intervals had to be computed and checked, only three relationships remained to create $G_{[106,117]}(Top(C_4^{114}))$ showing that topic trend detection can be simple. Within three time intervals the Minister of Defense (Verteidigungsminister) resigned (Vorwürfe $\widehat{=}$ accusations in Fig. 4a; Rücktritt $\widehat{=}$ resignation in Fig. 4b). When checking time intervals before $t = 112$ and after $t = 114$ no content-similar topics were found in the underlying SPON articles.

4.2.2 Arab spring

This example shows that the evolution of relationships between topics over time can involve the task of establishing a topic trend of overriding importance based on inspection periods covering a longer time span.

$G_{[120,131]}(Top(C_{11}^{131}))$, see Fig. 5, with $Top(C_5^{120})$ as origin and $Top(C_{11}^{131})$ as close describes a subgraph of content-similar topics concerning the situation in Libya and the role of NATO within the underlying inspection period [120, 131] of $\lambda = 12$ time intervals, see the term clouds in Fig. 6a, b.

Two topics, $Top(C_{12}^{121})$ and $Top(C_{13}^{121})$, which have the property to be splits in $G_{[120,131]}(Top(C_{11}^{131}))$, are closes in the content-similar subgraph depicted in Fig. 7 for the inspection period [110, 121] of $\lambda = 12$ time intervals. This subgraph can be denoted as $G_{[110,121]}(Top(C_{12}^{121}))$ or as $G_{[110,121]}(Top(C_{13}^{121}))$, see the remarks in Sect. 3.1 and step [5] of the approach described in Sect. 3.2. We will use the notation $G_{[110,121]}(Top(C_{12}^{121}))$ below. The subgraphs $G_{[110,121]}(Top(C_{12}^{121}))$ and $G_{[120,131]}(Top(C_{11}^{131}))$ overlap, see the nodes (120, 5), (121, 12), (121, 13) $\in N(G_{[110,121]}(Top(C_{12}^{121})) \cup G_{[120,131]}(Top(C_{11}^{131})))$ and reveal that the subject area of topics content-similar to $Top(C_{11}^{131})$ dates back to the inspection period [110, 121] or even earlier time intervals. However, e.g., the term clouds of the two origins $Top(C_9^{110})$ and $Top(C_{10}^{110})$ of $G_{[110,121]}(Top(C_{12}^{121}))$ show, see Fig. 8, that, at least in the begin-

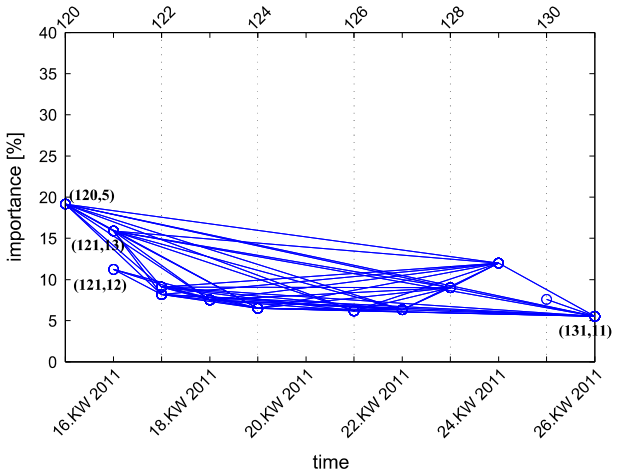

Fig. 5 $G_{[120,131]}(Top(C_{11}^{131}))$ as content-similar subgraph concerning the situation in Libya and the role of NATO

Fig. 6 Term clouds for **a** the origin $Top(C_5^{120})$ and **b** the close $Top(C_{11}^{131})$ of $G_{[120,131]}(Top(C_{11}^{131}))$

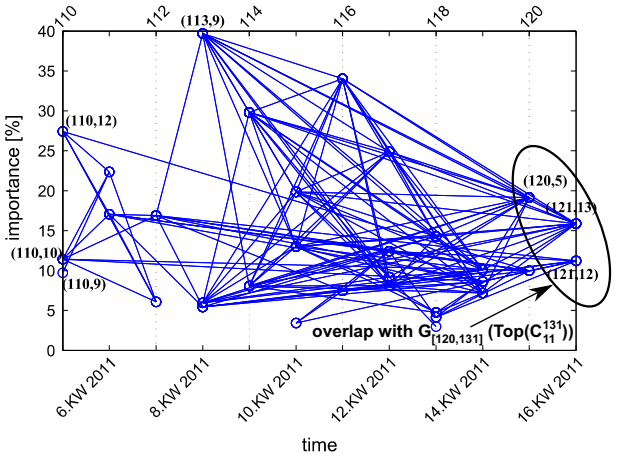

Fig. 7 $G_{[110,121]}(Top(C_{12}^{121}))$ which overlaps with $G_{[120,131]}(Top(C_{11}^{131}))$

ning of this subgraph, the situations in Egypt and Tunisia are emphasized in documents concerning these topics.

This illustrates that content-similarity via path connectivity can broaden the horizon with respect to topic trend selection of overriding importance. $G_{[110,121]}(Top(C_{12}^{121}))\cup G_{[120,131]}(Top(C_{11}^{131}))$, as a connected subgraph in the inspection period [110, 131]

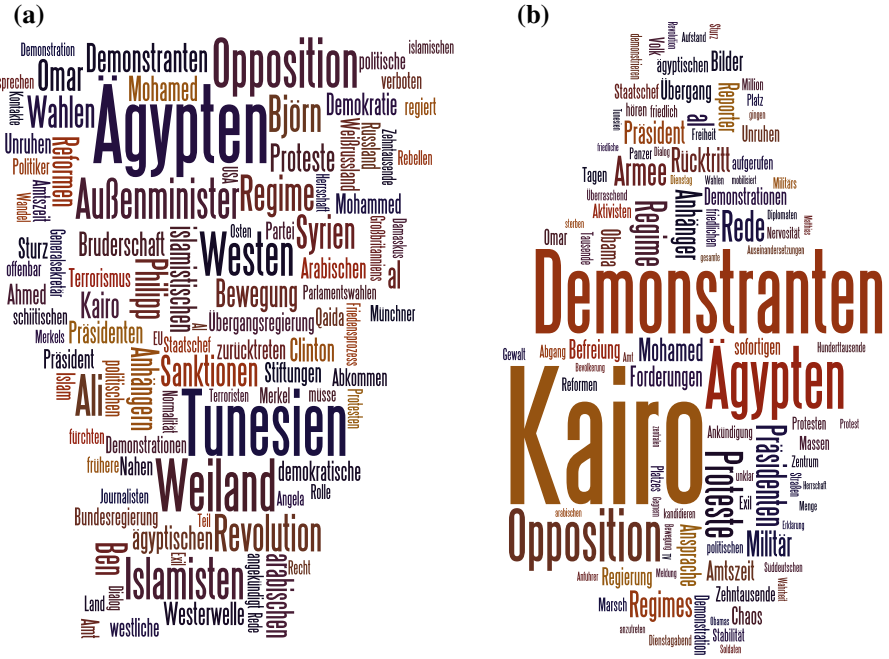

Fig. 8 Term clouds of two origins. **a** $Top(C_9^{110})$ and **b** $Top(C_{10}^{110})$ of $G_{[110,121]}(Top(C_{12}^{121}))$

of length $\lambda = 22$, is devoted to a topic trend which could be called the 'Arab spring'. It should be noted that 231 cluster-cluster dissimilarity matrices would have to be computed if one wanted to check all relationships in [110, 131] but only $2 \cdot 66 - 1 = 131$ matrices if the overlapping inspection periods [110, 121] and [120, 131] are used. For an interested person who applies a 'look-back' strategy, starting with $Top(C_{11}^{131})$, the inspection of $G_{[110,121]}(Top(C_{12}^{121})) \cup G_{[120,131]}(Top(C_{11}^{131}))$ would be of assistance although it will not necessarily contain all content-similar relationships possible in the inspection period [110, 131]. Note, also, that the importance of the topics is decreasing over time according to the outcomes of our approach. The topic with highest importance in $G_{[110,121]}(Top(C_{12}^{121})) \cup G_{[120,131]}(Top(C_{11}^{131}))$ is the upper topic $Top(C_9^{113})$ in the time interval $t = 113$ (8. KW 2011) which combines and compares aspects in Tunisia, Egypt, and Libya, is a split, and has a lot of relationships to topics which are tackled in time intervals in the future of $t = 113$. The graph-theoretical aids to explain the evolution of relationships between topics over time within the frame of time-importance diagrams are obviously very helpful.

4.2.3 Fukushima effects on German nuclear power policy

In Fig. 9 an inspection period of $\lambda = 53$ time intervals is shown together with a subgraph $G_{[75,127]}(Top(C_{10}^{127}))$ of content-similar topics the structure of which reveals that three parts of particular noteworthy subject areas of content-similar topics can be separated and that based on the application of 'look-back' strategies, as explained in

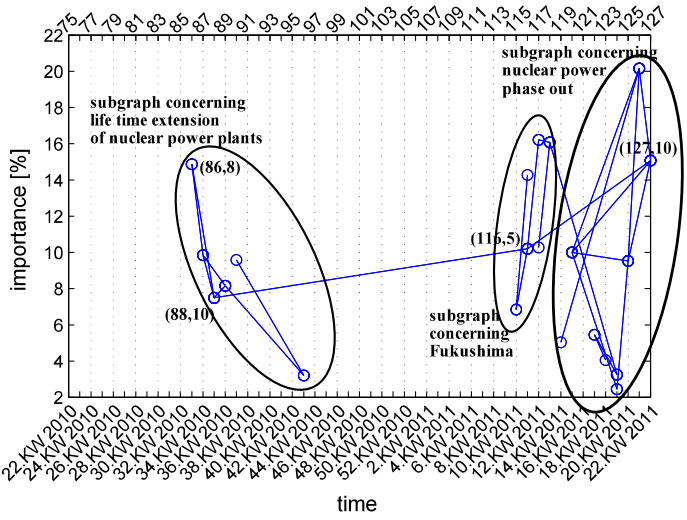

Fig. 9 $G_{[75,127]}(Top(C_{10}^{127}))$ visualizes changes concerning German nuclear power policy

Sect. 3.2, an inspection period of at least 29 time intervals would be needed to recognize that topics $Top(C_{10}^{88})$ and $Top(C_5^{116})$ are content-similar within the considered SPON articles.

This time, term clouds of selected topics to illustrate how the meanings of topics can change, will not be shown. An interested person who starts at $T = 127$ with $Top(C_{10}^{127})$ and applies 'look-back' strategies, would find content-similar topics related to a trend concerning 'nuclear power phase out' in the inspection period part [119, 127] of the 12 weeks inspection period [116, 127] and recognize that—shortly before in the past in the inspection period part [116, 118]—topics with respect to the 'Fukushima tragedy' are content-similar to the starting topic $Top(C_{10}^{127})$. Here, something like 'nuclear catastrophe effects' could be selected as the overriding topic trend. If (s)he were to continue (her) his topic mining activities, (s)he would have to go back another 29 time intervals (i.e. 306 cluster-cluster dissimilarity matrices would have to be computed and checked based on the SPON articles under consideration) starting from $Top(C_5^{116})$ until topics related to discussions about 'life time extension of nuclear power plants' take shape which are content-similar to $Top(C_{10}^{127})$. Inquisitive as to whether further content-similar topics might occur in the past of $Top(C_{10}^{88})$, the application of 'look-back' strategies can be continued. Figure 9 shows the situation within the inspection period [75, 127] and illustrates that content-similar topics before $T = 86$ were not found. Such an evaluation reveals that a remarkable change in German nuclear power policy has taken place for which 'Fukushima effects on German nuclear power policy' could be selected as topic description of overriding importance.

An application of the approach under Sect. 3.2 with a constant inspection period length of $\lambda = 12$ would not be able to find the graph depicted in Fig. 9, as arc ((88, 10), (116, 5)) needs an inspection period length of at least $\lambda = 29$ time intervals. However, when the approach has checked the past of $Top(C_{10}^{127})$, e.g., the concatenation of the 12 weeks inspection periods [116, 127], [105, 116], [94, 105] and, finally,

found the subgraph of content-similar topics with respect to the topic trend 'life time extension of nuclear power plants' in the 12 weeks inspection period [85, 96], one can apply the approach with a larger λ to examine whether paths exist which are able to bridge the gap between [85, 96] and [116, 127].

4.3 Discussion

The selected examples show only subgraphs of content-similar topics within the much larger network of relationships between all topics, which will appear in such situations, and indicate that graph-theoretical considerations using content-similarity via path connectivity are able to support topic detection and to trace topic trends.

The examples also show that the approach can handle arbitrary situations provided that a starting time interval T and a set of documents D with $D^T \subset D$ are available.

Of course, the length of inspection periods plays an important role as it influences computational speed and—if extremely long—can lead to graphs in which topic trend evaluation can become more difficult and enhance the burden to concentrate on content-similar subgraphs of particular interest. The choice of dis_{lb} as bound for the decision whether topics are related, as well as the dimension $Z = |\mathcal{L}'|$ of the reduced dictionary and the number of clusters in the single time intervals, selected via application of the elbow-criterion and constrained to a maximum of 15 for the document sets in this paper, are further parameters affecting the evaluation of topic relationships with the help of the presented approach.

The above-mentioned examples illustrate that the suggested methodology, which combines text processing, information clustering, and the visualization of topic relationships over time with the help of graph-theoretical tools in time-importance diagrams, works well and can reveal overriding topic trends.

The 'little' example of Sect. 4.2.1 was included on purpose to show that—when a subject area covers only a few time intervals in a longer inspection period and is related to a person of public interest—topic relationship evaluation and trend detection can be uncomplicated.

For the example of Sect. 4.2.2 an inspection period of nearly half a year (the union of two overlapping inspection periods composed of 12 time intervals each) was checked. Based on the 'look-back'-strategies mentioned in the approach described in Sect. 3.2, it is interesting to see how the meaning of topics along paths of content-similar topics may evolve. While $Top(C_{11}^{131})$ and $Top(C_5^{120})$, the term clouds of which are depicted in Fig. 6a, b, describe the situation in Libya and the role of NATO at the different time intervals and suggest that $G_{[120,131]}(Top(C_{11}^{131}))$ is devoted to the political change in that country, e.g., $Top(C_9^{110})$ and $Top(C_{10}^{110})$ as two of the origins of $G_{[110,121]}(Top(C_{12}^{121}))$ have Egypt and Tunisia as weight-important terms in the corresponding term clouds, see Fig. 8a, b, and indicate that changing views concerning political developments in Arabic countries seem to constitute the overriding topic trend in $G_{[110,121]}(Top(C_{12}^{121})) \cup G_{[120,131]}(Top(C_{11}^{131}))$. Such a trend could be summarized as the 'Arab Spring' although this may not be a combination of terms which have weights relatively high enough (compared to the $Z = 20{,}000$ terms of the reduced dictionary $\mathcal{L}'$) to be considered in the term clouds of the relevant topics. Remember

that it has been pointed out that the shape of the graph(s) includes information about the importance of the topics involved and adds further possibilities to scrutinize the contents of single topics.

The example of Sect. 4.2.3 is based on an inspection period of about a year (53 time intervals) and the subgraph shown in Fig. 9 depicts different subject areas of German policy development. Within less than a year there was a change in the corresponding topics from 'life time extension of nuclear power plants' to 'nuclear power phase out' caused by the 'Fukushima tragedy'. Here, the term Fukushima, although contained in the local dictionary of IDS at weight position 316,322 far behind $Z = 20{,}000$ and thus not considered in the terms used for text processing, can be utilized to formulate an overriding topic trend description. This is an example indicating the flexibility possible within the application of the approach. Restarts in time intervals in the past, for which different inspection period lengths can be tried to check whether content-similar topics occurred already earlier, can be performed.

The three examples selected illustrate the generality of the approach and reveal the range for applications:

From easy topic trend detection within one inspection period to situations in which the examination of gaps of considerable length between different inspection periods might be of interest for trend analysis.

5 Concluding remarks

In order to support the search for additional explanations concerning topics related to a subject area of current interest, tools are needed that, apart from comparing contents of other adequate actual media, check the flow of information in available documents to find aspects that are content-similar to those topics of the present subject area, which have attracted attention in such a way that further evaluation activities seemed appropriate.

To the best of our knowledge there is no contribution in the literature which uses only documents in corresponding time intervals together with text processing and information clustering to allow a graph-theoretical treatment of the network of topic relationships of the kind described in this paper. As our approach does not need probabilistic assumptions as presented in some other references in the literature, we do not reference work in which such assumptions are used. Thus, in order to embed this contribution into the underlying research field it is sufficient to mention only the following publications:

Early references (e.g., Allan et al. 1998; Wayne 1998; Oard 1999; Walls et al. 1999; Rajaraman and Tan 2001; Allan 2002a, b; Allan and Lavrenko 2002; Allan 2002c as an edited volume of several contributions, and Pons and Berlanga 2002) used labels as 'topic detection, tracking, trend analysis' for the problem, which can be abbreviated by the acronym TDT (topic detection and tracking).

Already Wayne (1998), who worked for the NSA (National Security Agency), mentioned the interest of DARPA (Defense Advanced Research Projects Agency), NIST (National Institute of Standards and Technology), and NSF (National Science Foundation) in the TDT research field.

Hints regarding internet usage can be found in, e.g., Wei and Lee (2004), Khy et al. (2008) (online documents) as well as Bun and Ishizuka (2006) and Mei et al. (2006), in Benhardus (2010), Mathioudakis and Koudas (2010) (Twitter), or in Steiner et al. (2013) (Wikipedia).

Of course, authors looked for additional information to tackle the TDT research field [e.g., Kim and Myaeng 2004; Li et al. 2006; Tu and Seng 2012 (importance of different temporal information), Jin et al. 2007 (place information)].

Graph-theoretical considerations were used, e.g., in Yang et al. (2009) (event evolution graphs) or in Oliveira and Gama (2010) (bipartite graphs with definitions of mergers, splits, etc., for this special case).

One paper, in which—although with a different research direction which is, however, connected to topic detection—internet usage and additional information were brought together, is Gaul (2011) about Web page importance ranking. Given the references of this area, Web page importance ranking appears to be unrelated to TDT research but also combines content-similarity of Web information (checked via text processing) with the link structure of Web pages (modeled with the help of graph-theoretical considerations) as additional information and even provides page ranking values for assessing which pages seem to be the most important ones.

Further references concerning TDT could have been cited but the above-mentioned listing has provided enough aspects to allow an embedding of our problem description into the line of research concerning the evaluation of the evolution of relationships between topics over time. As one of our aims was to simplify the handling of the problem and to avoid assumptions whose relevance for practical applications needs to be checked, standard knowledge from text processing, cluster analysis, and graph theory was used. The determination of document-document and cluster-cluster dissimilarities is well-known in text processing. From cluster analysis only hierarchical document clustering in the single time intervals was mentioned (because then the elbow-criterion to estimate the 'right' number of clusters/topics can be applied). For the comparisons between clusters/topics of different time intervals, networks of relationships between topics were constructed and visualized in time-importance diagrams by graphs from which adequate subgraphs of content-similar topics could be chosen.

To sum up, documents are the items that have to be clustered, weight vectors assigned to the documents allow to determine dissimilarities between documents in a single time interval, centroids of the weight vectors of the documents assigned to clusters in different time intervals are used for the determination of relationships between topics, and graph-theoretical tools are applied for the visualization of the evolution of the relationships between topics over time in time-importance diagrams. Finally, examples are selected to help to recognize the potential of the suggested approach.

As clustering, graph theory and text processing (in alphabetical order) have proved to be salient research fields in the situation we described, the following early references—for foundations as well as further research directions that could be of interest for the problem addressed—are mentioned, e.g., Bock (1974, 1980) for publications concerning clustering, Salton (1989) for text processing and Brandes and Erlebach (2005) as more recent volume for a collection of contributions concerning graph-theoretical applications.

References

Allan J (2002a) Detection as multi-topic tracking. Inf Retr 5(2–3):139–157

Allan J (2002b) Introduction to topic detection and tracking. In: Allan J (ed) Topic detection and tracking. Kluwer Academic Publishers, Norwell, pp 1–16

Allan J (ed) (2002c) Topic detection and tracking: event-based information organization. Kluwer Academic Publishers, Norwell

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, VA, USA, pp 194–218

Allan J, Lavrenko V, Swan R (2002) Exploration within topic tracking and detection. In: Allan J (ed) Topic detection and tracking. Kluwer Academic Publishers, Norwell, pp 197–224

Benhardus J (2010) Streaming trend detection in Twitter. In: UCCS REU for Artificial Intelligence, Natural Language Processing and Information Retrieval, Final Report

Bock HH (1974) Automatische Klassifikation. Theoretische und praktische Methoden zur Gruppierung und Strukturierung von Daten (Cluster-Analyse). Vandenhoeck & Ruprecht, Göttingen

Bock HH (1980) Clusteranalyse—Überblick und neuere Entwicklungen. Oper Res Spektrum 1(4):211–232

Brandes U, Erlebach T (eds) (2005) Network analysis: methodological foundations, vol 3418. Lecture Notes in Computer Science. Springer-Verlag New York Inc, Secaucus

Bun KK, Ishizuka M (2006) Emerging topic tracking system in WWW. Knowl Based Syst 19(3):164–171

Gaul W (2011) Web page importance ranking. Adv Data Anal Classif 5:113–128

Jin Y, Myaeng SH, Jung Y (2007) Use of place information for improved event tracking. Inf Process Manage 43(2):365–378

Khy S, Ishikawa Y, Kitagawa H (2008) A novelty-based clustering method for on-line documents. World Wide Web 11(1):1–37

Kim P, Myaeng SH (2004) Usefulness of temporal information automatically extracted from news articles for topic tracking. ACM Trans Asian Lang Inf Process 3(4):227–242

Kupietz M, Keibel H (2009) The Mannheim German reference corpus (DeReKo) as a basis for empirical linguistic research. In: Minegishi M, Kawaguchi Y (eds) Working Papers in Corpus-Based Linguistics and Language Education, Tokyo University of Foreign Studies (TUFS), 3, pp 53–59

Kupietz M, Belica C, Keibel H, Witt A (2010) The German reference corpus DeReKo: A primordial sample for linguistic research. In: Calzolari N, Choukri K, Maegaard B, Mariani J, Odijk J, Piperidis S, Rosner M, Tapias D (eds) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta

Li B, Li W, Lu Q (2006) Topic tracking with time granularity reasoning. ACM Trans Asian Lang Inf Process 5(4):388–412

Mathioudakis M, Koudas N (2010) Twittermonitor: trend detection over the twitter stream. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, ACM, New York, NY, USA, SIGMOD '10, pp 1155–1158

Mei Q, Liu C, Su H, Zhai C (2006) A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In: Proceedings of the 15th International Conference on World Wide Web, ACM, New York, NY, USA, WWW '06, pp 533–542

Oard DW (1999) Topic tracking with the prise information retrieval system. In: Proceedings of the DARPA Broadcast News Workshop, pp 209–211

Oliveira M, Gama J (2010) Bipartite graphs for monitoring clusters transitions. In: Cohen P, Adams N, Berthold M (eds) Advances in intelligent data analysis IX, vol 6065., Lecture Notes in Computer Science. Springer, Berlin Heidelberg, pp 114–124

Pons-Porrata A, Berlanga-Llavori R, Ruiz-Shulcloper J (2002) On-line event and topic detection by using the compact sets clustering algorithm. J Intell Fuzzy Syst 12(3,4):185–194

Rajaraman K, Tan AH (2001) Topic detection, tracking, and trend analysis using self-organizing neural networks. In: Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer-Verlag, London, UK, UK, PAKDD '01, pp 102–107

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

Steiner T, van Hooland S, Summers E (2013) MJ no more: using concurrent wikipedia edit spikes with social network plausibility checks for breaking news detection. Computing Research Repository. arXiV:1303.4702

Tu YN, Seng JL (2012) Indices of novelty for emerging topic detection. Inf Process Manage 48(2):303–325
Walls F, Jin H, Sista S, Schwartz R (1999) Topic detection in broadcast news. In: Proceedings of the DARPA Broadcast News Workshop, Morgan Kaufmann Publishers, Inc, pp 193–198
Wayne CL (1998) Topic detection and tracking (tdt)—overview and perspective. In: DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne Conference Resort, Lansdowne Virginia
Wei CP, Lee YH (2004) Event detection from online news documents for supporting environmental scanning. Decis Support Syst 36(4):385–401
Yang C, Shi X, Wei CP (2009) Discovering event evolution graphs from news corpora. IEEE Trans Syst Man Cybern Part A Syst Hum 39(4):850–863