Lifelong Knowledge Acquisition with Topic Monitoring

Myra Spiliopoulou Rene Schult
Faculty of Computer Science, Otto-von-Guericke Univ. Magdeburg
{myra,schult}@iti.cs.uni-magdeburg.de

Abstract: Industry workers and their R&D colleagues rely on knowledge stored in patents, reports, workplans, guidelines and norms. With time, the subjects considered but also the terminology used in such documents change. This impeds the categorisation of old documents and reduces the effectiveness of keyword-based search. We present a document stream mining solution that organises document archives into topics and monitors the changes of the topics and their characteristic keywords over time.

1 Introduction

A considerable part of knowledge in industry is documented in process descriptions, product specifications, regulations and norms. Hence, the *Knowledge Acquisition* task of the *knowledge management value chain* [LLS06, Ch. 6] gains in importance and demands for powerful tools for categorisation and regular keyword-based information acquisition. However, advances in text mining and keyword-based information filtering rely on a fixed terminology and thus come short in the analysis and retrieval of information that accumulates over a long time. In this study, we propose a method that assists in the lifelong *adaptive* categorisation of information stored in document archives.

A document archive is a *growing* collection. This growth implies that topics which were very important in the past may become obsolete, while new topics emerge. Research advances provide several examples to this end; topics like “ambient intelligence” or “galaxy dynamics” are rather young, while “alchemy” is not as popular as it used to be during the middle ages. As a collection of documents grows, the *terminology* may change as well. For example, made-up words like “bubblesort” and acronyms like SQL and XML have emerged in computer science, while words like PASCAL or COBOL have recently lost some of their earlier popularity. The effective acquisition of information from the whole of a growing archive requires the discovery and *monitoring* of topics that describe the archive’s contents at different timepoints. To this purpose, we observe an archive as a document stream and perform *topic monitoring* upon this stream, taking account of emerging and disappearing words that describe the topics.

The paper is organised as follows. We first discuss advances on topic monitoring. Section 3 contains a concise description of our method for topic monitoring upon document clusters [RM06, SS06]. In Section 4 we describe our experiments on *backward and forward* topic monitoring. The last Section concludes our study.

2 Related Work

Topic discovery over a fixed document collection is a well-studied issue. It is an unsupervised learning problem, since topics are not a priori defined by librarians; they are derived by analysing document content. When the collection accumulates though, topic discovery must be accompanied by *topic monitoring*, which includes the detection of emerging topics and the decay of obsolete ones.

In his seminal work of 2002, Allan introduced the subject of *Topic Detection and Tracking* (TDT) [Al02]. TDT refers to detecting and tracing *stories* (a "topic" is a "story") and encompasses the tasks of story segmentation, first story detection, cluster detection, tracking and story link detection. Despite the conceptual similarity between TDT and the identification of emerging topics in a stream, the two subjects are not identical. Although some stories may be perceived as topics (e.g. the "tsunami" story that started in December 2005), a topic is not necessarily a story (e.g. the documents of a track in this conference adhere to the same topic but are not parts of a story). The task of *topic trend discovery* with the new definition of *topic* has been first discussed in the survey of Kontostathis et al [KGP$^+$03].

Topic monitoring over a stream of documents has then been studied with methods that discover clusters (as in [MY04, AY06]) and methods that discover latent models over the data (e.g. [MZ05, BL06]). In the former case, the documents are clustered on content similarity and a *topic* is a cluster *label*, i.e. a list of words that are characteristic of the cluster, whereby definitions of "characteristic word" vary. In the latter case, a latent model is discovered with Latent Semantic Indexing or Probabilistic LSI as in [MZ05] or with Latent Dirichlet Allocation as in [BL06]: It consists of word-topic associations as individual distributions over the documents, whereby all topics are present in each document to different extends.

Topic monitoring methods observe the document collection as a stream: The model is built and then adapted dynamically. During adaptation, topics may emerge, mutate, get merged, get split or die out. Aggarwal and Yu adapt the clusters by assigning each new document to the cluster with the most similar *droplet* [AY06], i.e. to a summary consisting of two vectors of words characteristic to the cluster. A cluster dies out if no new documents join it. A cluster emerges if there are documents that do not fit to existing clusters. Moringa and Yamanishi build and adapt *soft clusters* and place emphasis on the detection of emerging topics on assessing the importance of existing topics [MY04]. Mei and Zhai use PLSA to discover and adapt topics against a *globally valid* background model [MZ05].

Common to all these methods is the assumption of a predefined feature space, i.e. of knowing all possible words in advance. We alleviate this assumption in our method MONIC for cluster monitoring over arbitrary data [SNTS06] and our approach for topic monitoring [RM06, SS06] that we describe hereafter.

3 Lifelong Topic Evolution Mining

We study a stream of arriving documents on a variety of *implicit*, not a priori known topics. The terminology of new documents may change, both with respect to emerging and disappearing words and with respect to the distribution of words that were and remain present. We observe this stream in a sequence of discrete timepoints. Our goals are (a) to discover the topics that characterize the documents of each timepoint, (b) to detect topics that survive from one timepoint to the next, those that emerge and those that decline. In this way, documents associated with the same topic can be identified as such, even if they appear at different timepoints and adhere to different terminologies.

3.1 Building Topics at each Timepoint

We observe a stream of documents and study it at different timepoints $t_1, t_2 \ldots$. At each t_i, we perform document clustering upon an accummulated set of documents D_i to build topics. Document clustering involves first transforming documents into vectors over a feature space composed of the words that appear in the documents. At timepoint t_i, the feature space consists of the M most frequent words in D_i, selected after stemming and stop-word removal [RM06, SS06]. For vector clustering, we have studied different algorithms and found good results with bisecting K-Means, as reported in [SNTS06], and with the density-based clustering algorithm DBSCAN [EKSX96], as discussed in [SS08].

Once the clusters are built, each cluster acquires a *label*, defined as the set of *characteristic* words for the cluster's members. These can be the most frequent words or the words with the highiest information gain within the cluster. In [RM06, SS06], we specify that a label is based on word frequency, i.e. it consists of all words that are more frequent within the cluster than a threshold τ. Essentially, a label is a *topic*, although we use the latter term for labels that appear in more than one timepoints.

3.2 Topic monitoring over time

For topic monitoring, we observe the stream of arriving documents at the timepoints $t_1, \ldots, t_n$. The dataset D_i upon which we build the topics for timepoint t_i can consist of all documents seen from t_1 on or of only the documents that arrived during the last w timepoints. In the former case, we would build topics that describe *all* texts that ever arrived. In the latter case, we would forget all but the recent texts and concentrate in finding topics that are characteristic for them only. The amount of documents "remembered" at each timepoint t_i is thus governed by the value of the variable w, which is essentially a *window*, through which the stream is sliding.

The clusters derived at timepoint t_i constitute a set ξ_i. A topic is then the label of a cluster $c \in \xi_i$. Then, topic monitoring refers to the question "How did the topics found in t_i differ from those found in t_{i-1} and in the earlier past?" With this information, we can distinguish between *persistent topics* that encompass labels of many adjacent timepoints [RM06] and *emerging topics* that essentially correspond to previously unseen labels [SS06].

To solve this task, our algorithm compares the topics of each timepoint to those of the previous one. In particular, for the current timepoint $t_i, i > 1$ we consider each topic x

found in t_{i-1} and find its best match in the set of topics T_i of t_i. The notion of *matching* is based on the common words between x and the topics in T_i: If there is a topic $y \in T_i$ that contains the same words as x, then x has *survived* in y. If there is no such topic, we find all topics in T_i that have at least one word in common with x and identify for each such topic y the words that have a similar frequency in x and y. These candidate are sorted on the number of such shared words. In case of a tie, that topic is chosen for which the shared words have higher frequency. Details can be found in [RM06, SS06].

The adaptation to changes in the terminology is central to our approach. As we have explained in the introduction, retaining the old terminology is inappropriate, because emerging topics that use previously unseen words will be overlooked. At the other extreme, one might decide to always use the most recent terminology at each t_i . However, this requires the (expensive) re-vectorisation of all documents in D_i. Hence, we retain the old terminology as long as a minimal quality of the clusters is guaranteed. We measure *quality* on the number of topics built at each timepoint. If the clusters are so heterogeneous that no topics can be derived from them, then the old feature space is discarded and re-built as the set of the most frequent words in D_i. Then, the documents in D_i are vectorized again and topic discovery is performed as explained in subsection 3.1.

The topic monitoring process is sensitive to the size of the dataset D_i. As explained at the beginning of this subsection, D_i consists of the documents accummulated in $(t_{i-1}, t_i]$ as well as past documents, subject to a sliding window w. Intuitively, the more past documents are remembered, the larger is the impact of the old terminology and of old topics: Temporary hypes are difficult to detect, while topics that enjoy a constant popularity over time are easier to identify. In [RM06], we have set the sliding window to be as large as the whole time axis, essentially forcing the topic monitoring algorithm to remember all old documents. With this approach, we could highlight *persistent topics* that appeared at many timepoints. In [SS06], we have rather considered a small window w that allowed us to detect emerging topics and study temporary changes of the terminology. We consider this latter approach more appropriate for lifelong topic monitoring: For a given timepoint, the scholar is determining the direction of monitoring (forward, backward or both) and the window size, beyond which data and terminology should be forgotten. In the next section, we use the small-window option to monitor a large document archive in a forward and a backward fashion and study the topic evolution in it.

3.3 Underlying Components for Topic Monitoring

Our algorithm relies on open source technology for the non-core functionalities. For document preprocessing we use our earlier developed tool DIAsDEM [GSW01], which is currently available under `SourceForge`. DIAsDEM has been designed for fine-grain categorization and annotation of document sentences [GSW01, WS02]. DIAsDEM also incorporates K-Means and Bisecting-K-Means for text clustering. Hence, we have applied both algorithms for topic monitoring and opted for the latter, more robust method. However, our topic monitor can be coupled with any clustering algorithm. In [Sc07], we have also used DBScan [EKSX96], an algorithm that is robust to noise.

As pointed out at the beginning of this Section, a topic consists of the most characteristic words in a cluster according to some ranking. Our core algorithm ranks words on frequency, but can import more sophisticated ranking functions. We import such functions, as well as further clustering algorithms from the ‘Text Clustering Toolkit” (TCT) [1]. We also consider the TCT cluster quality indices, since the re-computation of the feature space is triggered by a decrease in cluster quality. TCT offers among else Normalized Mutual Information (NMI) index [SG02] and Rand index [Ra71]. In [SS08] we use these indices to study our method over DBScan versus Bisecting-K-Means.

4 Topic Monitoring in a Digital Library

The ACM Digital Library subarchive H2.8 on “Database Applications” has been already used in our past experiments for topic monitoring [RM06, SS06, Sc07]. Here, we juxtapose the evolution of topics in forward monitoring from 1996 towards 2004 and in backward monitoring from 2004 in the past.

For text mining, we considered only the document titles and keywords, not the contents. For preprocessing, we applied TCT for document stemming and stopword removal. We defined the feature space as the set of words appearing in at least 3 documents of the current timepoint. We have specified $K = 5$ for the number of clusters at each timepoint. A topic was set to contain the $N = 5$ most frequent words in the cluster.

To study the impact of terminology evolution, we built the feature space anew at each timepoint. To identify differences between backward and forward inspection, we have set the window size equal to the whole time horizon. We have thus maximised the influence of remembered documents upon the newly built topics.

In Fig. 1, we see the evolution of the 5 topics in forward inspection from 1996 to 2004 (Left) and in backward inspection from 2004 to 1996 (Right). The lines are the *lifelines* of the topics. A line that continues over adjacent timepoints indicates that the topic has survived, i.e. there are still relevant documents to be returned by a search. If the lifelines of topics cross (e.g. Left subfigure, topic 2 of 1998 towards topic 5 in 1999), then the topics are mixed and have acquired new semantics. Line discontinuation means that the topic has died out and cannot be found beyond the line’s end.

The survival of a topic does not imply that the keywords originally associated to it are still representative throughout its lifeline. An inspection of each topic’s constituent keywords provides clues as to which keywords should be used for search at each timepoint: For example, consider topic 5 in the left subfigure. This is the data mining topic. From 1996 to 2000, it is on `association rules` and mutates gradually to `knowledge discovery` from 2000 on (two most frequent words of the top-5). It is further associated with keywords like `cluster` (1998, 2001) and `spatial` (1997, 2001, 2002).

In the backward inspection (right subfigure), data mining is topic 1. Its dominant words from 2004 till 2002 are `knowledge discovery`. A further frequent word is `visual`

[1] University of Dublin, http://mlg.ucd.ie/content/view/18/

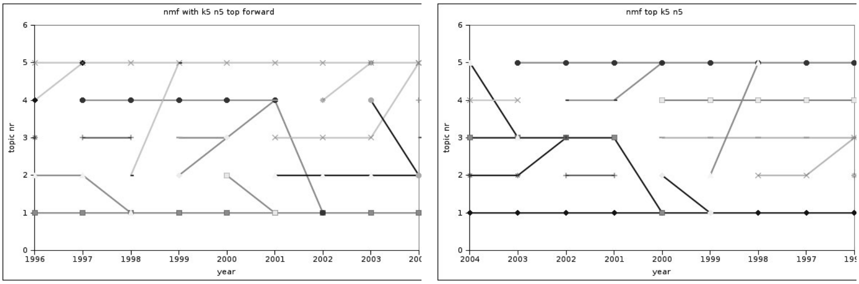

Abbildung 1: Forward and backward topic monitoring on the H2.8 document collection

(2003 till 1999). However, topic 1 shifts towards visualisation and clustering for multidimensional data, and the early-year keywords `association rules` do not show up. An explanation is the growing dominance of the class *data mining* in H2.8: After 2000, this class is larger than all the others together and grows faster. In the forward inspection, its impact on the feature space is supressed until 2004, but in the backward inspection, the feature space of 2004 dominates until 1998. This influence could be supressed more effectively by a small sliding window.

5 Conclusions

We have presented a tool for the support of lifelong knowledge acquisition. Its objective is the monitoring of topics upon a document collection, while allowing for changes in the terminology of the documents. Topic monitoring can be performed in both a forward fashion, in which information is regularly acquired from accummulating document archives, as well as in a backward fashion, where topics in an old document collection are traced starting with a contemporary terminology.

One of the lessons learned from our forward and backward inspection of topics in an archive is the need to account for skew among topics in the same timepoint and in changes in their popularity among timepoints. We intend to deal with this need by (a) considering decay functions at the price of forgetting faster the terminology and documents that are contemporary to the user and (b) by studying the potential of topic-word associations as in probabilistic latent semantic indexing [BL06] rather than working with document clusters.

Our tools contribute to the general issue of dealing with a changing world. Next to the volatility of the terminology, we intend to deal with the validity of the information provided by these sources with respect to error likelihood and up-to-date content.

Literatur

[Al02] Allan, J.: *Introduction to Topic Detection and Tracking*. Kluwer Academic Publishers. 2002.

[AY06] Aggarwal, C. C. und Yu, P. S.: A Framework for Clustering Massive Text and Categorical Data Streams. In: *Proceedings of the SIAM conference on Data Mining 2006*. April 2006.

[BL06] Blei, D. M. und Lafferty, J. D.: Dynamic topic models. In: *Proc. of 23rd Int. Conf. on Machine Learning*. Pittsburgh, PA. 2006.

[EKSX96] Ester, M., Kriegel, H.-P., Sander, J., und Xu, X.: A Density-Based Algortihm for Discovering Clusters in Large Spatial Database with Noise. In: *KDD'96: Proc. of 2nd Int. Conf. on Knowledge Discovery in Databases and Data Mining*. 1996.

[GSW01] Graubitz, H., Spiliopoulou, M., und Winkler, K.: The DIAsDEM framework for converting domain-specific texts into XML documents with data mining techniques. In: *Proc. of the 1st IEEE Intl. Conf. on Data Mining*. pp. 171–178. San Jose, CA. Nov. 2001. IEEE.

[KGP$^+$03] Kontostathis, A., Galitsky, L., Pottenger, W., Roy, S., und Phelps, D.: *A Survey of Emerging Trend Detection in Textual Data Mining*. Springer Verlag. 2003.

[LLS06] Laudon, K., Laudon, J., und Schoder, D.: *Wirtschaftsinformatik - Eine Einführung*. Pearson Studium. 2006.

[MY04] Moringa, S. und Yamanichi, K.: Tracking Dynamics of Topic Trends Using a Finite Mixture Model. In: *Proc.of 10th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'04)*. pp. 811–816. Seattle, Washington. Aug. 2004. ACM Press.

[MZ05] Mei, Q. und Zhai, C.: Discovering Evolutionary Theme Patterns from Text - An Exploration of Temporal Text Mining. In: *Proc. of 11th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'05)*. pp. 198–207. Chicago, IL. Aug. 2005. ACM Press.

[Ra71] Rand, W.: Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*. (66):846–850. 1971.

[RM06] Rene Schult und Myra Spiliopoulou: Expanding the Taxonomies of Bibliographic Archives with Persistent Long-Term Themes. In: *Proceedings of the 21st Annual ACM Symposium on Applied Computing (SAC'06)*. April 2006.

[Sc07] Schult, R.: Comparing Clustering Algorithms and their Influence on the Evolution of Labeled Clusters. In: *Procedings of DEXA 2007*. pp. 650–659. Sep 2007.

[SG02] Strehl, A. und Gosh, J.: Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*. (3):583–617. 2002.

[SNTS06] Spiliopoulou, M., Ntoutsi, I., Theodoridis, Y., und Schult, R.: Monic – modeling and monitoring cluster transitions. In: *Proc. of 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'06)*. pp. 706–711. Philadelphia, USA. Aug. 2006. ACM.

[SS06] Schult, R. und Spiliopoulou, M.: Discovering emerging topics in unlabelled text collections. In: *Proc. of ADBIS'2006*. Thessaloniki, Greece. Sept. 2006. Springer.

[SS08] Schult, R. und Spiliopoulou, M.: Contributing to market monitoring with topic evolution monitoring. *Datenbank-Spektrum, Schwerpunkt Business Intelligence*. 2008. accepted in April 2008.

[WS02] Winkler, K. und Spiliopoulou, M.: Structuring domain-specific text archives by deriving a probabilistic XML DTD. In: *6th European Conf. on Principles and Practice of Knowledge Discovery in Databases, PKDD'02*. pp. 461–474. Helsinki, Finland. Aug. 2002. Springer Verlag.