Towards Automatic Detection and Tracking of Topic Change

Florian Holz and Sven Teresniak

NLP Group, Department of Computer Science, University of Leipzig
`{holz,teresniak}@informatik.uni-leipzig.de`

Abstract. We present an approach for automatic detection of topic change. Our approach is based on the analysis of statistical features of topics in time-sliced corpora and their dynamics over time. Processing large amounts of time-annotated news text, we identify new facets regarding a stream of topics consisting of latest news of public interest. Adaptable as an addition to the well known task of topic detection and tracking we aim to boil down a daily news stream to its novelty. For that we examine the contextual shift of the concepts over time slices. To quantify the amount of change, we adopt the volatility measure from econometrics and propose a new algorithm for frequency-independent detection of topic drift and change of meaning. The proposed measure does not rely on plain word frequency but the mixture of the co-occurrences of words. So, the analysis is highly independent of the absolute word frequencies and works over the whole frequency spectrum, especially also well for low-frequent words. Aggregating the computed time-related data of the terms allows to build overview illustrations of the most evolving terms for a whole time span.

1 Introduction

Large collections of digital diachronic text such as the New York Times corpus and other newspaper or journal archives in many ways contain temporal information related to events, stories and topics. To detect the appearance of new topics and tracking the reappearance and evolution of them is the goal of topic detection and tracking [2,1]. For a collection of documents, relevant terms need to be identified and related to a particular time-span, or known events, and vice versa, time-spans need to be related to relevant terms. To identify relevant and new terms in a stream of text (within a predefined period of time), three main approaches have been followed. [7,8,6] measure the relevance of terms using multiple document models and thresholds based on a tf/idf comparison of text stream segments. [5] introduces the burstiness of terms during certain periods of time as an additional dimension for topic detection, and models the temporal extension of relevant terms using a weighted finite state automaton. [10] use co-occurrence patterns and their local distribution in time to detect topics over time. By their approach, every topic is represented by a co-occurrence set of terms representative of a certain period of time. Assuming topics and the terms representing them to be constant over time, topics can efficiently be related to times.

However, topics not only depict events in time, they also mirror an author's, or society's, view on the events described. And this view can change over time. In language, the relevance of things happening is constantly rated and evaluated. In our view, therefore, topics represent a conceptualization of events and stories that is not statically related to a certain period of time, but can itself change over time. Tracking these changes of topics over time is highly useful for monitoring changes of public opinion and preferences as well as tracing historical developments.

In what follows, we shall argue that

1. changing topics can be detected by looking at their change of meaning,
2. changing topics are interesting, i. e. they generally represent topics that for some period of time are “hotly discussed”, and
3. tracking the change of topics over time reveals interesting insights into a society's conceptualization of preferences and values.

If we consider the disclosure of new aspects regarding an already established and identified topic, the extraction of the meaning's change can be used to distinguish between real novelty and already known facts or recurring events.

While the public attention to a topic determines the topic's presence in media coverage, the novelty regarding this topic is somewhat independent from the amount of coverage of a story: even without worldshaking new facts, a topic can be important to society and therefore on the agenda of an editorial department. From a text mining perspective which takes the past news as given facts and aims to extract unknown and novel aspects and developments this reporting is to some degree redundant.¹ With the ability to discriminate between *novel* news on the one hand and news just referencing the *recent* (but not necessarily new) knowledge about a certain topic on the other hand, we are able to identify the novelty bearing parts in news streams.

In addition to term frequency, we consider a term's global context (see below) as a second dimension for analyzing its relevance and temporal extension and argue that the global context of a term may be taken to represent its meaning(s). Changes over time in the global context of a term indicate a change of meaning. The rate of change is indicative of how much the “opinion stakeholders” agree on the meaning of a term. Fixing the meaning of a term can thus be compared to fixing the price of a stock. Likewise the analysis of the volatility of a term's global contexts can be employed to detect topics and their change over time. We first explain the basic notions and assumptions of our approach and then present first experimental results.

2 Motivation for Our Method

Following [4], we take a term to mean the inflected type of a word, where the notion of a word is taken to mean an equivalence class of inflected forms of a

¹ For example consider the global warming. While for the last few years, this topic became more and more important, no fundamental new aspects were discovered and thus no novel facts are added to the discussion.

base form. Likewise we take the notion of a topic to mean an equivalence class of words describing an event (as computed by the global context of the topic's name), and the notion of a concept to mean an equivalence class of semantically

Table 1. The 30 most significant co-occurrences and their significance values (cut to 3 digits) in the global context of “abu ghraib” on May 10, 2004

prisoners 0.346, abuse 0.346, secretary 0.259, abu ghraib prison 0.259, iraqi 0.247, rumsfeld 0.221, military 0.218, prison 0.218, bush 0.210. prisoner 0.200, photographs 0.183, donald 0.183, secretary of defense 0.174, prisons 0.174, photos 0.174, the scandal 0.174, interrogation 0.163, naked 0.163, mistreatment 0.163, under 0.162, soldier 0.154, saddam 0.154, armed 0.154, defense 0.143, the bush 0.140, senate 0.140, videos 0.130, torture 0.130, arab 0.130, captured 0.130

Fig. 1. The co-occurrence graphs of “iraq” depict the changes in the global context of the term for three days in (a), (b) 2003 and (c) 2004

1. Built a corpus where all time slices are joined together.
2. Compute for this overall corpus all significant co-occurrences $C(t)$ for every term t .
3. Compute all significant co-occurrences $C_i(t)$ for every time slice i for every term t .
4. For every co-occurrence term $c_{t,j} \in C(t)$ compute the series of ranks $\text{rank}_{c_{t,j}}(i)$ over all time slices i . This represents the ranks of $c_{t,j}$ in the different global contexts of t for every time slice i .
5. Compute the coefficient of variation of the rank series $\text{CV}(\text{rank}_{c_{t,j}}(i))$ for every co-occurrence term in $c_{t,j} \in C(t)$.
6. Compute the average of the coefficients of variation of all co-occurrences terms $C(t)$ to obtain the volatility of term t

$$\begin{aligned} \text{Vol}(t) &= \text{avg}_j \left(\text{CV}_i (\text{rank}_{c_{t,j}}(i)) \right) \\ &= \frac{1}{|C(t)|} \sum_j \text{CV}_i (\text{rank}_{c_{t,j}}(i)) . \end{aligned}$$

Fig. 2. Computing the volatility

related words. The global context of a topic's name is the set of all its statistically significant co-occurrences within a corpus. We compute a term's set of co-occurrences on the basis of the term's joint appearance with its co-occurring terms within a predefined text window taking an appropriate measure for statistically significant co-occurrence. The significance values are computed using the log-likelihood measure following [3] and afterwards normalized according to the actual corpus size. These significance values only serve for sorting the co-occurrence terms; their absolute values are not considered at all. The position of a term in this sorted list is called the term's rank.

Table 1 exemplifies the global context computed for the term “abu ghraib” based on the New York Times corpus of May 10, 2004. The numbers in parenthesis behind a term indicate its statistical significance (normalized to the corpus size and multiplied by 10^6), which are used to rank the co-occurring terms (cf. Fig. 2).

The global context can also be displayed as a graph which contains the term and its context terms as nodes where the edges have a weight each according to the significance value of the joint appearance of the terms. Figures 1(a)–(c) illustrate the change of co-occurrences and thus the change of the global context of the word “iraq” for three different days in 2003 and 2004 based on the New York Times corpus. These Graphs show how a changing media coverage is affecting the co-occurrences of a term.

3 Method

The basis of our analysis is a set of time slice corpora. These are corpora belonging to a certain period of time, e. g. all newspaper articles of the same day.

Table 2. Volatility and global context

	Volatilität	
	low	high
No. co-occurrences constant	high-frequent terms: - stop words - “static concepts”	periodic or arbitrary concepts, e.g. “Monday”, “city center” usf.
No. co-occurrences differing	rising or declining concepts, e.g. “globalization” usf.	highly dynamic notions: - low-frequent (“weak signals”) - high-frequent

The assessment of change of meaning of a term is done by comparing the term’s global contexts of the different time slice corpora.

The measure of the change of meaning is *volatility*. It is derived from the widely used risk measure in econometrics and finance², and based on the sequence of the significant co-occurrences in the global context sorted according to their significance values and measures the change of the sequences over different time slices. This is because the change of meaning of a certain term leads to a change of the usage of this term together with other terms and therefore to a (maybe slight) change of its co-occurrences and their significance values in the time-slice-specific global context of the term. The exact algorithm to obtain the volatility of a certain term is shown in Fig. 2.

In order to reduce the time complexity of our algorithm we only take the overall most important co-occurrences into account. This is done by computing the global contexts of the terms based on an overall corpus which is the aggregation of all time slice corpora. In the case of the used New York Times corpus this means a comprehension of about 7500 days which are about 20 years. Using an overall significance threshold only the more significant terms are taken into account during the comparison of the time-slice-specific global contexts. This set of relevant co-occurring terms for a term t is named $C(t)$ in Fig. 2. Besides providing evidence for meaningful filtering the overall corpus is not used in the computation of the volatility. A co-occurring term is significant if the according co-occurrence, i.e. the pair of the original term and the co-occurring term is significant. Co-occurrences are taken as statistically significant if they a) occur at least two times in the corpus and b) their significance value computed using the log-likelihood measure exceeds a threshold. In our experiments the threshold was set that half the co-occurrences occurring at least two times passed it. Based on language statistics this means very careful filtering.

Concerning the relation between the volatility and the global context of a term the following picture can be sketched (cf. Tab. 2). One can expect that the volatility of terms like “Monday” is quite high because weekdays (and other

² But it is calculated differently and not based on widely used gain/loss measures. For an overview over miscellaneous approaches to volatility see [9].

periodic re-occurring time references) are highly ambiguous as the specification which precise day is meant is lacking. Analogously this is to be assumed for ambiguous place identifiers without specification like “city center” and “town hall” as we do no identification of the concrete referenced entity or any other semantic pre-processing.

4 Experiments

In what follows, we present results of experiments that were carried out on the basis of data based on the New York Times Annotated Corpus (NYT)³. Table 3 lists some general characteristics of this corpus. First tests were performed for German on the corpus of the project *Deutscher Wortschatz*⁴ and showed comparable results. We aim to show that our method in fact works to detect topics that were “hotly discussed” during some period of time, giving also an indication, why that has been so. For performance reasons, we only took the 50 000 most frequent terms out of the 20-year NYT corpus into account and filtered this list as follows:

- remove stop words,
- remove all terms with a frequency rank higher than 50 000 (this is equivalent to less than 850 occurrences in 20 years NYT),
- include all multi word units which are wikipedia lemmata and of a rank of at most 50 000
- remove digits, numbers and so on.

The resulting term list contains “people” as its most frequent word and “benes” (a name) as its least frequent word. As described in Sect. 3, for all of these words, its most significant co-occurrences in the overall corpus of the whole 20 years were computed. Thereafter the volatility for every term was computed.

Figure 3 shows the development of the volatility of “abu ghaib” from January 2004 to December 2006. The volatility was computed per day with a window of 30 days, i. e. for the volatility for a certain day the last 30 days before were taken into account (cf. Fig. 2). The daily frequency of “abu ghraib” is also shown in Fig. 3 as a 30-day average over the last 30 days, too. The clearly outstanding peaks of the volatility are easily connectable to certain events and their related media coverage. The first peak beginning in May 2004 is caused by the initial discussion about the torture pictures and videos taken in the prison in Abu Ghraib. This is also clearly affecting the co-occurrences of “abu ghraib” as shown in Tab. 1.

Concerning the next time span, the frequency is declining but fluctuating. These fluctuations aren’t related to a change of the topic what can be seen in the uncorrelated peaks of the volatility. The volatility peak in the end of May 2005 is caused by a widening of usage of the term “abu ghraib”, e. g.

³ <http://www ldc.upenn.edu/>

⁴ <http://wortschatz.uni-leipzig.de>

Table 3. Characteristics of the used corpus NYT

language	english
time span	Jan 87 – Jun 07
no. time slices	7 475
no. document	1.65 mil.
no. tokens	1 200 mil.
no. types	3.6 mil.
no. sig. co-occurrences	29 500 mil.
size (plain text)	5.7 GB

Fig. 3. 30-day volatility and frequency of “abu ghraib” from 2003 to 2006 based on the NYT corpus

the NYT reports about U.S. interrogators accused of misconducts and hold responsible for dead detainees in afghanistan as well as a discussion which actions are appropriate during a military interrogation and what methods are regarded as torture. Further, now “abu ghraib” is also used as the name of a district in the west of baghdad – where a huge military operation took place – which our algorithm couldn’t differentiate from “abu ghraib” as the catchphrase for the abusive military behaviour at Abu Ghraib prison.⁵

The new aspect and topic shift does not lead to an extended coverage in the New York Times but is measurable as a change of context. The peak in November and December 2005 is related to an exhibition which also was held in New York. There pictures from Abu Ghraib have been exhibited together with others from the Weimar Republic and World War II. This new aspect and its reporting is clearly affecting the global context of “abu ghraib”. Table 4 shows

⁵ Not only because of the removal of stopwords it’s not possible to extract whether the pronoun in front of “abu ghraib” was *in* or *at*.

Fig. 4. Joint representation of the 30-day volatility of “abu ghraib”, “abu ghraib prison” and “iraq” from 2003 to 2006 based on the NYT corpus

Table 4. The 30 most significant co-occurrences in the global context of “abu ghraib” on November 20, 2005

disasters, hook, grosz, international center, finalized, weighty, inkling, complement, partnerships, guggenheim museum, collaborative, the big city, easel, reaped, hudson river museum, blockbuster, enlarging, goya, weimar, art museums, eras, inconvenient, negatives, golub, poughkeepsie, grissold, big city, impressionist, staging, neuberger

this for the November, 20, when the reporting about the exhibition started. The event also does not cause a more frequent usage of “abu ghraib” in the New York Times, but is nevertheless detectable by the related change of context.

However, another interesting fact is the comparison of the volatility graphs of “abu ghraib”, “abu ghraib prison”, and “iraq” as shown in Fig. 4. Between January 2003 and May 2004 the incidents at “abu ghraib” are not an issue, whereas the invasion of Iraq and the ongoing discussions around it are clearly visible in the graph as a relatively and constantly high-volatile topic.

At the beginning of news coverage related to the abuse at the Abu Ghraib prison the peaks of “abu ghraib” and “abu ghraib prison” are strongly correlated. But at the end of 2005 the ongoing reporting was strong or long enough to establish “abu ghraib” as a synonym to the complete affair around the abuse in Abu Ghraib prison, whereupon “abu ghraib prison” is no longer a term of general interest. Now, in the western public perception, “abu ghraib” no longer stands mainly for a city with about 1 mil. inhabitants, but for the crisis in the prison of this city. From 2006 on, the torture scandal in Abu Ghraib prison – like Guantanamo Bay – is used more frequently in a symbolic way and exploited by many people for their individual needs (e. g. by Mahmoud Ahmadinejad for

Fig. 5. 30-day volatility and frequency of “iraq” from 1987 to 2006 based on the NYT corpus

propaganda reasons to appealing to Iran’s conservative base or by opponents of war to connect the war on terrorism in Iraq with cruelty, both in 2006). Once established as a symbol, the Abu Ghraib crisis is stressed controversially in many contexts and thus remains high-volatile at least until November of 2006, even though the absolute frequency of “abu ghraib” is quiet low (cf. Fig. 3).

Figure. 5 shows the development of the volatility and the frequency of “iraq” from 1987 to 2007. It clearly can be seen that volatility and frequency do not correlate. New aspects like the Gulf War in 1990 or the discussion after Sep 11, 2001 lead to an increased volatility. An increased frequency alone – like at the beginning of the invasion of Iraq in 2003 – does not affect the volatility as at this time all reported arguments have already been very well known.

5 Generating Overviews

Based on the volatility values over time, e. g. per-day values as shown in Sect. 4, it is possible to generate an overview over the collection over the whole time span or a section. This is based on the fact, that for always highly volatile terms like stop words or concepts like “Monday” the volatility remains high all the time and its variance is comparably low in comparison to less frequent but thematically evolving terms which are expected to have a low volatility in general but with peaks at moments where they are hotly discussed or new aspects show up. For examples described in detail see Sect. 4. Therefore we computed for every term the variance of the series of volatility values to get one value for each term indicating how much the topic evolved over the considered time span.

The visual overview is a 2D plot where every term’s position is given by the term’s absolute frequency and the term’s volatility. Thus the overview depicts the relation between how present a term was in the shown time span and how

6 Conclusions and Further Work

In this paper, we have presented a new approach to the analysis of topics changing over time by considering changes in the global contexts of terms as indicative of a change of meaning. First experiments, carried out using data from contemporary news corpora for German and English, indicate the validity of the approach.

In particular, it could be shown that the proposed measure of a term's volatility is highly independent from a term's frequency.

An aggregated representation allows the user to get a direct overview over the most evolving topics covered in the processed documents. In an interactive application the user can explore more and less evolving aspects of the covered time span by zooming into certain areas. If the user finds an interesting term, it's easy to provide him with the curve of the volatility of this term showing the term's development over the time span. Using the significant co-occurrences, the user can be provided the most related terms as well. Combining those overviews of subsequent time spans, it is possible to show the terms' developments as a trajectories, one for every term. So, rising or declining topics can be identified by having the according terms moving along the x -axis while they gain or loose variance of volatility in contrast to other concepts which may stay in their area over the different overview representations.

In a next step, the analysis proposed can be extended to look at individual topics changing over those time spans identified as interesting. Instead of only looking at the terms that change their meaning over time, it might also be of value to look at those terms that for some time span retain a "stable" meaning, expressing a society's unquestioned consensus on a topic, as it were. In the long run, this approach might lead to an infrastructure for easily analyzing diachronic text corpora with many useful and interesting applications in trend and technology mining, marketing, and E-Humanities.

Acknowledgments

This research has been funded in part by DFG project *Topology-based Visual Analysis of Information Spaces*⁶ as part of the Focus Project Nr. 1335 *Scalable Visual Analytics: Interactive Visual Analysis Systems of Complex Information Spaces*⁷.

References

1. Allan, J.: Introduction to topic detection and tracking, pp. 1–16. Kluwer Academic Publishers, Norwell (2002)
2. Allan, J., et al.: Topic detection and tracking pilot study final report. In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, pp. 194–218 (1998)
3. Dunning, T.E.: Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1), 61–74 (1993)
4. Heyer, G., Quasthoff, U., Wittig, T.: Text Mining: Wissensrohstoff Text – Konzepte, Algorithmen, Ergebnisse, 2nd edn. W3L-Verlag (2008)
5. Kleinberg, J.: Bursty and hierarchical structure in streams. In: KDD 2002: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 91–101. ACM Press, New York (2002)

⁶ <http://www.visualanalytics.de/node/19/>

⁷ <http://www.visualanalytics.de/>

6. Kumaran, G., Allan, J.: Text classification and named entities for new event detection. In: SIGIR 2004: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 297–304. ACM, New York (2004)
7. Swan, R., Allan, J.: Extracting significant time varying features from text. In: CIKM 1999: Proceedings of the eighth international conference on Information and knowledge management, pp. 38–45. ACM, New York (1999)
8. Swan, R., Allan, J.: Automatic generation of overview timelines. In: SIGIR 2000: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 49–56. ACM, New York (2000)
9. Taylor, S.J.: Introduction to asset price dynamics, volatility, and prediction. In: Asset Price Dynamics, Volatility, and Prediction. Introductory Chapters. Princeton University Press, Princeton (2007)
10. Wang, X., McCallum, A.: Topics over time: a non-markov continuous-time model of topical trends. In: KDD 2006: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 424–433. ACM, New York (2006)