# GENERATING A VISUAL OVERVIEW OF LARGE DIACHRONIC DOCUMENT COLLECTIONS BASED ON THE DETECTION OF TOPIC CHANGE

Florian Holz[1], Sven Teresniak[1], Gerhard Heyer[1] and Gerik Scheuermann[2]
[1]*Natural Language Processing Group,* [2]*Image and Signal Processing Group*
*Institute of Computer Science, University of Leipzig, Leipzig, Germany*

Keywords: Visual analytics, Document collection summarization, Browsing document collections, Topic detection, Change of meaning, Volatility, Time-sliced corpora.

Abstract: Large digital diachronic document collections are a central source of information in science, business, and for the general public. One challenge for the efficient visualization of these collections is the automatic calculation and visualization of the main topics. These topics can then serve as the basis for an overview of the content and any subsequent interactive visual analysis. We introduce the new language processing concept of volatility of terms measured as the change of the context of terms. We demonstrate that volatility can serve as an excellent basis for the visual overview of large collections using two different examples.

## 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. Therefore, a good starting point for any visual analysis of such collections is a visual representation of contained topics. To detect the appearance of new topics and tracking the reappearance and evolution of them is the goal of topic detection and tracking (Allan et al., 1998; Allan, 2002). 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.

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 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). We use these two dimensions as the basis of our visual overview of the topics in the collection.

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. Therefore, we offer the user to study frequency and volatility of a selected topic over time to let him find time spans of interest with respect to the topic. We first explain the basic notions and assumptions of our approach and then present first experimental results.

## 2 BASIC NOTIONS

Following (Heyer et al., 2008), 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 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 con-

This research has been funded in part by DFG Focus Project Nr. 1335 Scalable Visual Analytics under DFG number SCHE 663/4-1.

Holz F., Teresniak S., Heyer G. and Scheuermann G. (2010).
GENERATING A VISUAL OVERVIEW OF LARGE DIACHRONIC DOCUMENT COLLECTIONS BASED ON THE DETECTION OF TOPIC CHANGE.
In *Proceedings of the International Conference on Imaging Theory and Applications and International Conference on Information Visualization Theory and Applications*, pages 153-156
DOI: 10.5220/0002836701530156

Table 1: The 30 most significant co-occurrences and its significance value (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

text of the topic's name), and the notion of a concept to mean an equivalence class of semantically 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 (Dunning, 1993) 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. 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. 1).

## 3 THE SETTING

The processing of large and very large document collections has several difficulties which make it hard to provide substantial help for an user who wants to access certain documents, especially when the exact item or its position is unknown to the user. The state-of-the-art interfaces for accessing large document collections are indeces like google and other search engines, which rely mainly on indexing all or statistically relevant terms, and structured catalogues like (web) opacs, which need annotated metadata for each document and use these for filtering.

The most hampering aspect is the large amount of data itself and the complexity of its analysis. For instance computing the global contexts of terms in a corpus has a time and space complexity of $O(n^2)$, where $n =$ number of types is about 1,000,000 to 10,000,000. Therefore it is difficult to compute and even to define appropriate and use- and meaningful measures describing terms and their relations. Thus most analyses rely on term frequency, which is efficiently computeable, and e. g. often relevance measures comparing local term frequency in a document to the total frequency in a reference corpus.

We aim for a new paradigm in interacting with large time-related corpora. Obviously, it is impossible to present information about every document in a large collection at once, because if there are for instance 1.6 Mio documents like in the New York Time corpus (cf. Sect. 6), there are only about 0.82 pixels per document for visualization, aasuming a standard screen with 1 280 × 1 024 pixels. So an aggregated view on the content is necessary, and this view should enable a visualization-based interactive exploration of the collection which is driven by the users attention and intent by providing him details on demand.

Therefore we want to identify the most relevant terms in the sense that these terms are related to the most considerable developments over the time span of the corpus. We establish the measure of volatility of a term (see next section) to cover the change of its global context which indicates a change of usage of the term. So we can provide an overview over the most evolving topics as an entrance into the whole collection.

## 4 VOLATILITY COMPUTATION

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. 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 finance[1], and based on the sequence of the significant co-occurrences in the global context sorted according to their significance values (see Sect. 2) 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. 1. For the detailed natural language processing background see (Holz and Teresniak, 2010).

[1] But it is calculated differently and not based on widely used gain/loss measures. For an overview over miscellaneous approaches to volatility see (Taylor, 2007).

1. Build 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 $\mathrm{t}_i$ for every term $t$.
4. For every co-occurrence term $c_{t,j} \in C(t)$ compute the series of ranks $\mathrm{rank}_{c_{t,j}}(i)$ variing $i$ which represents the ranks of $c_{t,j}$ in the different global contexts of $t$ for every time slice $\mathrm{t}_i$.
5. Compute the coefficient of variation of the rank series $\mathrm{CV}(\mathrm{rank}_{c_{t,j}}(i))$ for every co-occurrence term in $c_{t,j} \in C(t)$.
6. Compute the average of the coefficients of variation of all co-occurences terms $C(t)$ to obtain the volatility of term $t$

$$\mathrm{Vol}(t) = \underset{j}{\mathrm{avg}} \left( \underset{i}{\mathrm{CV}} \left( \mathrm{rank}_{c_{t,j}}(i) \right) \right) .$$

Figure 1: Computing the volatility.

## 5 VISUAL OVERVIEW

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 variance computed as the variance all per-day volatilities (cf. Sect. 6). The *x*-axis comprises of the rank of the term in the frequency-sorted term list while the *y*-axis indicates the volatility of the term. Thus the overview depicts the relation between how present a term was in the shown time span and how much the related topic evolved over it. The overview provides a simple and intuitive aggregation of the document collection. Figure 2 shows such an (zoomed) overview computed for all articles of the New York Times corpus in 2004. The high-frequent terms are on the right side, the low-frequent ones on the left. The *x*-axis is displayed logarithmic. According to the power law distribution of term frequencies in natural language (cf. Zipf's law), the logarithmic view leads to a concentration of the most terms in the middle of the *x*-axis which would in a linear view mostly to be found indistinguishably right next to the *y*-axis.

This represenation allows the user to get an 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 the curve of the volatility of this term showing the term's development over the time span like shown in Fig. 3 (cf. Sect. 6). Using the significant co-occurrences the user can be provided the most related terms as well.

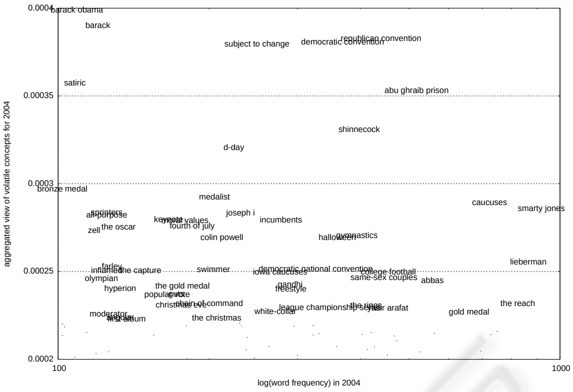

Figure 2: Variance of volatility according to word frequency for 2004, zoomed.

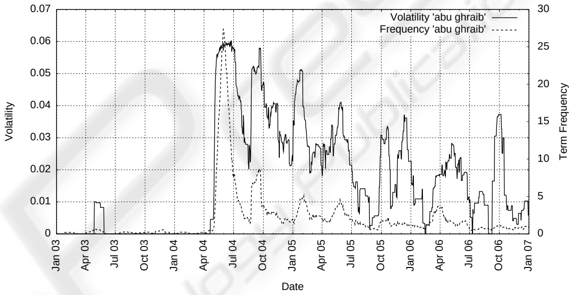

Figure 3: 30-day volatility of "abu ghraib" from 2003 to 2006 based on the NYT corpus.

## 6 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)[2]. Table 2 lists some general characteristics of this corpus.

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. 1). 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 (cf. Tab. 1).

The volatility does not generally correlate with the word frequency as e. g. the volatility peak in April 2005 shows. It's caused by the news coverage of a suicide attack at the prison on April 5. The new as-

[2] http://www.ldc.upenn.edu/

pect and topic shift does not lead to an extended coverage in the New York Times but is measureable as a change of context. The peak in November and December 2005 is related to an exhibition where pictures from Abu Ghraib have been shown together with others from the Weimarer Republic and World War II. 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. Table 3 shows this for the November 20, when the reporting about the exhibition started.

Once established as a symbol, the Abu Ghraib crisis is stressed controversely 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). For previous experiments and more detailed examples see (Holz and Teresniak, 2010).

Table 3: 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, griswold, big city, impressionist, staging, neuberger

## 7 CONCLUSIONS

In this paper, we have presented a new approach to the analysis of large diachronic document collections. We proposed to analyze the change of topics over time by considering changes in the gobal contexts of terms as indicative of a change of meaning. Measuring this change it is possible to visualize a substantial amount of a large time-related data volume concentrating on the most evolving topics and to provide a simple and intuitive overview over the wohle document collection. 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.

This overview is planned to be the basis of an advanced interactive exploration application. So in a next step we plan to combine the the overview and term-specific analysis within one user interface which provides also zooming and filtering options. Therefore the user is planned to be able to select a certain term or set of terms to get the volatility development within an selectable time frame and to have access to the most prominent documents of this time frame which have high impact on the volatility of the term(s). It is also intended to provide access to such terms which are most heavily changing the global context of the previously chosen term indicating into which direction its meaning and usage changes within the selected time frame. Combining those overviews of subsequent time spans, it is possible to show the terms' developments as a trajectories 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.

Table 2: 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 |

## REFERENCES

Allan, J. (2002). *Introduction to topic detection and tracking*, pages 1–16. Kluwer Academic Publishers, Norwell, MA, USA.

Allan, J. et al. (1998). Topic detection and tracking pilot study final report. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218.

Dunning, T. E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Heyer, G., Quasthoff, U., and Wittig, T. (2008). *Text Mining: Wissensrohstoff Text – Konzepte, Algorithmen, Ergebnisse*. W3L-Verlag, 2nd edition.

Holz, F. and Teresniak, S. (2010). Towards automatic detection and tracking of topic change. In Gelbukh, A., editor, *Proc. CICLing 2010: Conference on Intelligent Text Processing and Computational Linguistics, LNCS 6008*. Springer LNCS.

Taylor, S. J. (2007). Introduction to asset price dynamics, volatility, and prediction. In *Asset Price Dynamics, Volatility, and Prediction*, Introductory Chapters. Princeton University Press.