SIGRAD 2012
A. Kerren and S. Seipel (Editors)

Analytical Semantics Visualization for Discovering Latent Signals in Large Text Collections

C. Stab, M. Breyer, D. Burkhardt, K. Nazemi, and J. Kohlhammer

Fraunhofer Institute for Computer Graphics Research IGD, Germany

Abstract

Considering the increasing pressure of competition and high dynamics of markets, the early identification and specific handling of novel developments and trends becomes more and more important for competitive companies. Today, those signals are encoded in large amounts of textual data like competitors' web sites, news articles, scientific publications or blog entries which are freely available in the web. Processing large amounts of textual data is still a tremendous challenge for current business analysts and strategic decision makers. Although current information systems are able to process that amount of data and provide a wide range of information retrieval tools, it is almost impossible to keep track of each thread or opportunity. The presented approach combines semantic search and data mining techniques with interactive visualizations for analyzing and identifying weak signals in large text collections. Beside visual summarization tools, it includes an enhanced trend visualization that supports analysts in identifying latent topic-related relations between competitors and their temporal relevance. It includes a graph-based visualization tool for representing relations identified during semantic analysis. The interaction design allows analysts to verify their retrieved hypothesis by exploring the documents that are responsible for the current view.

Categories and Subject Descriptors (according to ACM CCS): H.5.2 [Information Interfaces and Presentations]: User Interfaces—Graphical user interfaces (GUI), Interaction styles H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing—Abstracting methods

1. Introduction

The early detection of novel developments and trends is crucially important and confers companies a significant advantage over their competitors in global competition. Concepts known from the field of strategic management like weak signals [Ans75] and environmental scanning [Agu67] constitute the theoretical foundation for detecting early warnings and systematically scanning the environment of an organization for relevant information [Sch05]. Today, different research areas focus on adopting these concepts in different kinds of software systems that aim at (semi-) automatically identifying knowledge for facilitating strategic decision making. On the one hand, business intelligence platforms consider mainly internal data, whereas competitive intelligence platforms aim at gathering semi- or unstructured data from the external environment of an organization with the aim of supporting the optimization of strategic decision making. The latter are also known as *Strategic Early Warning Systems* (SEWS).

In contrast to SEWS, common information systems usually provide keyword searches that make intensive use of information retrieval technologies. These search mechanisms are primarily focused on providing the users with easy access to information of their interest and deal with the access to information items and resources [BYRN10]. The underlying assumption of information retrieval platforms is an *initial information need* of the user that is usually expressed by a keyword query. In response to the given query the user receives a sorted list that is often supplemented by *Key-Word-In-Context* (KWIC) snippets. In contrast to typical search tasks, weak signal discovery rarely starts with an initial intention expressed as a set of keywords but rather with the exploration of key topics and trends that are latent in an underlying collection of text documents. After getting an overview, analysts dive deeper into a specific topic and achieve new hypotheses and impulses by identifying novel knowledge artifacts and recognizing relations between them.

In this paper we present an SEWS called *Signal Tracing* that

is based on analytical semantics visualizations for discovering latent signals in large text collections. The approach utilizes a backend system that provides semantic analysis and data mining tools for summarizing and querying the underlying document collection. Starting at an initial state, the interactions of the analyst are translated into different queries whose results are visualized in several visualizations. In the following section we briefly introduce the general process of a SEWS followed by a detailed description of the visualization and interaction techniques that are used in Signal Tracing for discovering latent signals in large text collections.

2. The SEWS Process

Usually a SEWS passes through three general and repeating process steps (Figure 1), namely (1) information gathering, (2) analysis & diagnosis and (3) reporting and decision making. In the first step, data is gathered and selected either fully automated or manually by detecting key issues in the environment of a company. In most cases the collected data contains mainly text documents like news feeds, company websites, newsletters, customer ratings, etc. but it may also contain pictures and video material that requires additional preprocessing (e.g. image recognition, speech2text, etc.) for extracting the main content. The gathered text collection is analyzed during the second step of the process. After some preprocessing steps like stop-word removal and stemming, different techniques like document clustering, classification, key term extraction, topic detection and tracking (TDT) and more domain specific methods like those described in [ZS06, MZ05] can be utilized to summarize the collection and to extract latent signals. Beside the described data mining methods, information visualizations are increasingly used to communicate and summarize the discovered information. In addition to common visualization techniques like pie, line and bar charts there are also a number of more sophisticated methods like topic-based visualizations [LZP*12, DWCR11], interactive maps [FMG05, PM06] or graph-based methods [ZS06, GM04] for exploring latent knowledge in text collections. Finally, in the third step the inferred signals are used to evaluate strategic possibilities, to formulate potential reactions and to assess the consequences of uncovered trends and topics.

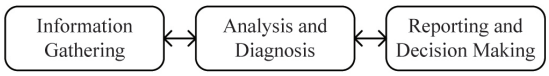

Figure 1: *Process of Strategic Early Warning Systems*

3. Visual Discovery of Latent Signals

The user interface of Signal Tracing is based on several visualization tools that cover different analytical aspects. These components are connected by brushing and linking techniques for providing multiple interactive and aspect-oriented perspectives on the underlying text collection. Altogether the user interface (Figure 2) integrates six different components: (1) Entity Explorer, (2) Trend Visualization, (3) Topic and Keyword Explorer, (4) Relation Graph, (5) Document Browser and (6) Reporting Tools.

The actual analysis and diagnosis follows a prolonged interactive loop of querying, exploring and refining using the tools provided by the Signal Tracing user interface and the analytical, semantic backend system. So the analyst is able to develop new hypotheses and to gain novel insights into the gathered sources and to infer strategic decisions. This interactive cycle includes the following tasks: (a) Entity Exploration and Selection, (b) Trend Discovery, (c) Topic Detection and Exploration, (d) Relation Analysis and (e) Validation and Verification. Thanks to the design of the user interface and its juxtaposed components, the sequence of these analytical tasks is not strictly defined but analysts are able to combine the tools in different orders. For example it is possible to insert a unknown topic identified in the Topic and Keyword Explorer as a new entity for identifying its temporal correlation with existing entities. Hence, it is possible to consider new acquired knowledge from the one perspective in another perspective to visually correlate novel insights with already existing entities in an iterative loop. A detailed explanation of how the components of the Signal Tracing user interface are related to these tasks and how the components are interactively connected is provided in the next subsections.

3.1. Entity Exploration and Selection

The Entity Explorer presents known entities and their corresponding categories that are either identified during the information gathering step of the SEWS process or emerged during the semantic analysis of the given document collection. For each entity the view includes the number of occurrences for providing an overview of the quantitative distribution. The Entity Explorer is directly connected with the adjacent Trend Visualization so that the temporal characteristics of selected entities are also visible. Thus the Entity Explorer also serves as a kind of filtering instance for selecting entities that are relevant for the further analysis. It also allows the definition of custom entities that can be incorporated in subsequent analysis steps.

3.2. Trend Discovery

The awareness of emerging and disappearing trends constitutes a decisive factor for evaluating current strategies and for accurate strategic decisions. The Signal Tracing user interface includes a Trend Visualization that represents the temporal occurrence of selected entities in an extended stacked-graph visualization. By exploring the temporal distribution, analysts are able to identify correlations between

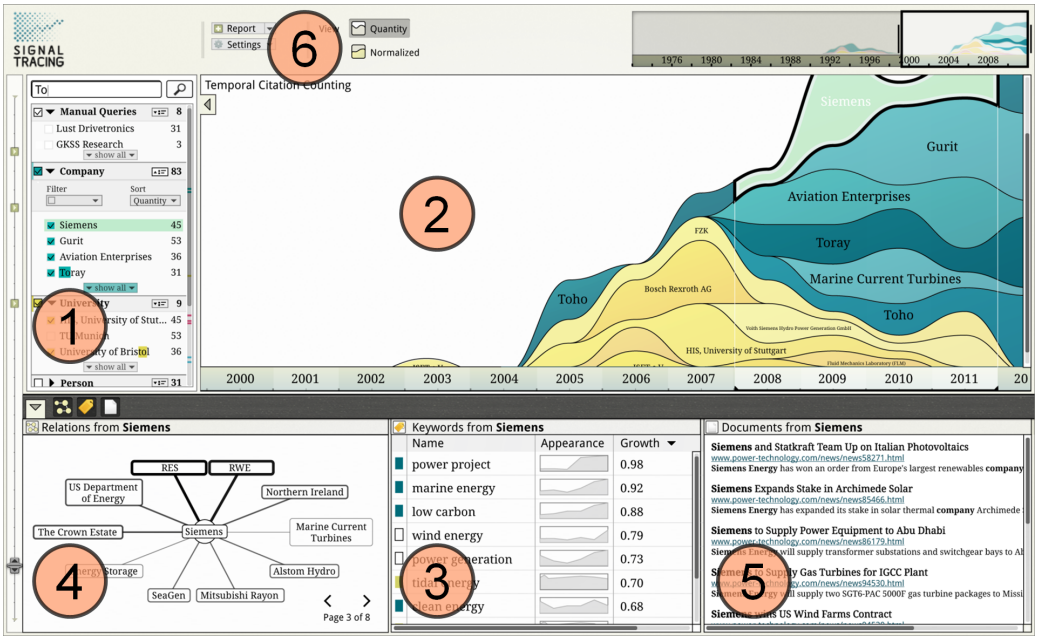

Figure 2: *The user interface of Signal Tracing incorporates the following components for discovering latent signals in large text collections: (1) Entity Explorer, (2) Trend Visualization, (3) Topic and Keyword Explorer, (4) Relation Graph, (5) Document Browser and (6) Reporting Tools.*

selected entities. The advanced overlay techniques (3.4) integrated in the stacked graph enable the exploration of topic specific and temporal co-occurrences in the given document collection.

3.3. Topic Detection and Exploration

For obtaining an overview of the content, the Topic and Keyword Explorer summarizes either the whole corpus or the current entity selection. Therefore, the background system analyses all documents that are related to the current state of the visualizations. The extracted topics and keywords are ordered according to their relevance and presented in a list view. Next to each entry, the view also includes sparklines for indicating the temporal development of each keyword and the documents that are related to the current state are included next to the Topic and Keyword Explorer in the Document Browser.

3.4. Relation Analysis

Signal Tracing provides two different types of relation discovery: Analysis of entity relations and the analysis of topic-specific and temporal co-occurrences. The first type of relations is visualized in a graph representation that shows relations between entities. The strength between the entities is indicated by the thickness and the color intensity of visible edges. So analysts are able to identify unknown relations e.g. between competitors that are often mentioned together in the documents. The graph also contains a paging feature for preventing visual clutter and switching between pages of related entities. The interactive linkage with the entity explorer allows the selection of related entities for enabling subsequent trend analysis.

The second type of relation analysis enables analysts to identify relations between topics and selected entities and to inspect their strengths over time. Visually this feature is implemented by overlays for each entity in the stacked graph. By selecting an entry in the Topic and Keyword Explorer, a query is generated for each visible entity in the Trend Visualization. The result of each query is the number of co-occurrences of an entity and the selected keyword over time that is visualized as a visual overlay in the Trend Visualization (Figure 3). The strength of the relation is indicated by the intensity of the overlay for each time interval. So it is possible to explore the evolution of a certain topic in relation to selected entities. For instance, an analyst may be interested if selected competitors are involved in a novel technology and to identify increasing or decreasing activities in this specific sector.

3.5. Validation and Verification

Finally, the Document Browser includes all documents that are responsible for the current selection and the state of the visualization respectively. So analysts are able to validate

and to verify their hypotheses by means of the sources and to collect additional data for reporting and strategic decisions.

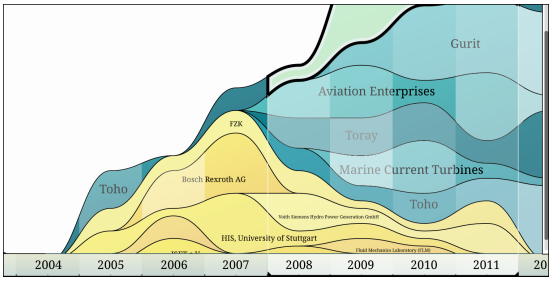

Figure 3: *The overlay technique used in the stacked graph reveals the strengths of the relations between competitors and selected topics (e.g. technologies or market sectors) over time for identifying latent trends in the underlying document collection.*

4. Related Work

Current approaches to mining strategic knowledge differ not only in the type of representation but also in the applied analysis methods. For instance Pulse [GACoR05] uses a tf-idf based clustering algorithm and sentiment analysis for extracting categories that are represented in a treemap for identifying critical issues in customer opinions. Other approaches e.g. [ZS06] [GM04] utilize co-occurrences of named-entities or keywords for providing a graph-based exploration. [ZS06] also utilizes taxonomic background knowledge for generating semantic profiles of competitors which are compared using a graph-based visualization. Tiara [LZP*12] is a visual text summarization tool that is based on Latent Dirichlet Allocation (LDA) for extracting topics whose strengths are determined at different time periods and visualized in a stacked graph but it does not contain tools to inspect co-occurrences over time. Map-based approaches [PM06, FMG05] utilize dimensionality reduction methods like Multi-dimensional Scaling (MDS), Principal Component Analysis (PCA) or Latent Semantic Indexing (LSI) for plotting high dimensional document vectors in a two-dimensional space. The resulting maps provide an overview even for very large document collections but do not reveal trends or entity relations. ParallelTopics [DWCR11] is a visual analytics system for analyzing large text corpora. It includes a document distribution view that presents the probabilistic distribution of documents across topics, a temporal view, a topic cloud and a document scatterplot. However it does not include tools for discovering latent relations in the text collection.

5. Conclusion & Future Work

In this paper we introduced Signal Tracing, a strategic early warning system for interactively discovering latent knowledge in large text collections. Besides the common process for SEWSs we presented several visualization approaches incorporated in Signal Tracing and the interactive design that fosters the identification of weak signals for strategic decisions. For the future work we plan to extend the current prototype with additional tools e.g. a visual comparison of keyword distributions might help analysts to compare different players and to apply the prototype to additional corpora.

Acknowledgements

This project (HA project no. 290/11-35) is funded in the framework of Hessen ModellProjekte, financed with funds of LOEWE - "Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz, Förderlinie 3: KMU-Verbundvorhaben" (State Offensive for the Development of Scientific and Economic Excellence). We thank Dr. Rainer Vinkemeier (C21 Consulting GmbH) and Joachim Caspar (Conweaver GmbH) for the inspiring discussions and the provision of the backend system.

References

[Agu67] AGUILAR F. J.: *Scanning the business environment*. Collier-Macmillan, 1967. 83

[Ans75] ANSOFF I. H.: Managing Strategic surprise by response to weak signals. *California Management Review 18*, 2 (1975), 21–33. 83

[BYRN10] BAEZA-YATES R., RIBEIRO-NETO B.: *Modern Information Retrieval*, 2nd ed. Addison-Wesley Publishing Company, 2010. 83

[DWCR11] DOU W., WANG X., CHANG R., RIBARSKY W.: Paralleltopics: A probabilistic approach to exploring document collections. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on* (oct. 2011), pp. 231 –240. 84, 86

[FMG05] FORTUNA B., MLADENIC D., GROBELNIK M.: Visualization of Text Document Corpus. *Informatica Journal 29*, 4 (2005), 497–502. 84, 86

[GACoR05] GAMON M., AUE A., CORSTON-OLIVER S., RINGGER E.: Pulse: Mining customer opinions from free text. In *Proc. of the 6th International Symposium on Intelligent Data Analysis* (2005), pp. 121–132. 86

[GM04] GROBELNIK M., MLADENIC D.: Visualization of news articles. In *SIKDD 2004 at multiconference IS* (2004). 84, 86

[LZP*12] LIU S., ZHOU M. X., PAN S., SONG Y., QIAN W., CAI W., LIAN X.: Tiara: Interactive, topic-based visual text summarization and analysis. *ACM Trans. Intell. Syst. Technol. 3*, 2 (Feb. 2012), 25:1–25:28. 84, 86

[MZ05] MEI Q., ZHAI C.: Discovering evolutionary theme patterns from text: An exploration of temporal text mining. In *Proc. of ACM SIGKDD* (2005), KDD '05, ACM, pp. 198–207. 84

[PM06] PAULOVICH F., MINGHIM R.: Text map explorer: a tool to create and explore document maps. In *Information Visualization, 2006. IV 2006.* (july 2006), pp. 245 –251. 84, 86

[Sch05] SCHWARZ J. O.: Pitfalls in implementing a strategic early warning system. *Foresight - The journal of future studies, strategic thinking and policy 7*, 4 (Apr. 2005), 22–30. 83

[ZS06] ZIEGLER C.-N., SKUBACZ M.: Towards automated reputation and brand monitoring on the web. In *Proc. of IEEE/WIC/ACM* (2006), WI '06, pp. 1066–1072. 84, 86