Trend template: mining trends with a semi-formal trend model

Olga Streibel, Lars Wißler, Robert Tolksdorf, Danilo Montesi
streibel@inf.fu-berlin.de, lars.wissler@googlemail.com, tolk@ag-nbi.de
Networked Information Systems Group, Freie Universität Berlin, Berlin, Germany
montesi@cs.unibo.it
University of Bologna, Bologna, Italy

Abstract

Predictions of uprising or falling trends are helpful in different scenarios in which users have to deal with huge amount of information in a timely manner,such as during financial analysis. This temporal aspect in various cases of data analysis requires novel data mining techniques. Assuming that a given set of data, e.g. web news, contains information about a potential trend, e.g. *financial crisis*, it is possible to apply statistical or probabilistic methods in order to find out more information about this trend. However, we argue that in order to understand the context, the structure, and explanation of a trend, it is necessary to take a knowledge-based approach. In our study we define trend mining and propose the application of an ontology-based trend model for mining trends from textual data. We introduce the preliminary definition of trend mining as well as two components of our trend model: the *trend template* and the *trend ontology*. Furthermore, we discuss the results of our experiments with trend ontology on the test corpus of German web news. We show that our trend mining approach is relevant for different scenarios in ubiquitous data mining.

1 Introduction

When discussing trends some of us may think about the ups and downs of NASDAQ[1], or DAX[2] curves, or changes in public opinion on politics before elections. Likewise, one can think about web trends, life style trends or daily trends, i.e. *hot topics*, in the news or on social networks. Changes in a mobile data stream also fall within the definition of a trend.
Understanding a trend as a hot topic is related to the research in *Emerging Topic Detection* (*EDT*) and *Topic Detection and Tracking* (*TDT*), the subfields of information retrieval [Allan, 2002][Kontostathis *et al.*, 2003]. A trend is defined there as *a topic that emerges in interest and utility over time*. Accordingly, common examples of trends may be the "*Arab Spring*" which emerged in political news worldwide in the beginning of 2011, as well as *the financial* and *real estate crisis* which started to emerge on business news worldwide in 2008. A graphical representation of a trend, based on GoogleTrends[3], is shown in Fig. 1.

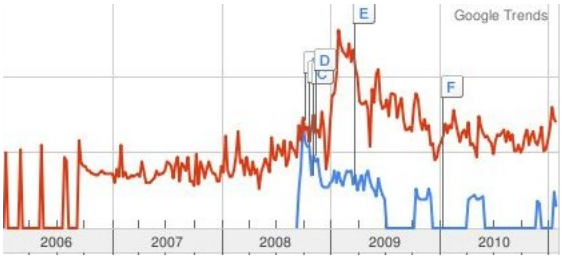

Figure 1: This graph shows a search volume index for the terms "financial crisis" (blue curve) and "insolvent" (red curve) in Germany from 2006 to 2011. Source: GoogleTrends

Several methods have been proposed for detecting trends in texts or discovering trends in the web news (see Section 3). Other works provide approaches from statistics and time series analysis that can be applied for analyzing trends in non-textual data. Our work contributes to the general understanding of *trend mining* that we see as highly relevant to ubiquitous data mining. In this paper, we explain our abstract concept of a *trend template* and go on to describe a *trend ontology* which is an instance of the trend template.

2 Ubiquitous data mining and trend mining

The Ubiquitous Data Mining (UDM) is defined as the essential part of the ubiquitous computing [Witten and Eibe, 2005]. The UDM techniques help in extracting useful knowledge from data that describes the world in movement, including the aspects of *space* and *time*. Time is the necessary dimension for trend mining– there is no trend without time. And a trend is one of the aspects of a world in movement. Before we discuss general trend characteristics, we want to mention the sociological and statistical perspectives on the trend, as well as define trend mining. This helps in understanding the trend characteristics that create the basis for the definition of our trend template later in this paper.

[1] http://www.nasdaq.com/ online accessed 04-17-2013
[2] http://dax-indices.com/EN/index.aspx?pageID=4 online accessed 04-17-2013
[3] http://www.google.com/trends/ online accessed 04-17-2013

2.1 Trend from different perspectives

Detecting trends from the sociological point of view is an analytical method for observing changes in peoples behavior over time with regard to "six attitudes towards trends" [Vejlgaard, 2008]. The definition of these six attitudes is based on eight different personality profiles of groups who participate in the trend process: trend creators, trend setters, trend followers, early mainstreamers, mainstreamers, late mainstreamers, conservatives and anti-innovators.
Detecting trends from the statistics perspective is based on trend analysis of time-series data with two goals in mind: "modeling time series (i.e. to gain insight into the mechanisms or underlying forces that generate the time series) and forecasting time series (i.e. to predict the future values of the time-series variables)" [Han and Kamber, 2006]. The trend analysis process consists of four major components: trend or long-term movements, cyclic movements or cyclic variations, seasonal movements or seasonal variations, and irregular or random movements [Han and Kamber, 2006]. A trend, in this context, is an indicator for a change in the data mean [Mitsa, 2010].

2.2 Trend mining

Since data mining can be described as "the extraction of implicit, previously unknown, and potentially useful information from data" [Witten and Eibe, 2005], we propose the use of the term *trend mining* as defined below:

DEF 2.1 **Trend mining** *is the extraction of implicit, previously unknown and potentially useful knowledge from time-ordered text or data. The trend mining techniques can be used for capturing trend in order to support user in providing previously unknown information and knowledge about the general development in users field of interests.*

3 Related Research

In general, when mining trends from textual data, at least the following three research areas should be mentioned: *emergent trend detection*, *topic detection and tracking*, and *temporal data mining*.
In [Kontostathis *et al.*, 2003] several systems that detect emerging trends in textual data are presented. These ETD systems are classified into two main categories: semi-automatic and fully-automatic. For each system there is a characterization based on the following aspects: *input data and attributes*, *learning algorithms* and *visualization*. This comparison includes an overview over the research published in [Allan *et al.*, 1998][Lent *et al.*, 1997][Agrawal *et al.*, 1995][Swan and Jensen, 2000][Swan and Allan, 1999][Watts *et al.*, 1997]. TDT research [Allan, 2002] is predominantly related to the event-based approaches. Event-based approaches for trend mining underlie the assumption that trends are always triggered by an event, which is often defined as "something happening" or "something taking place" [Lita Lundquist, 2000] in the literature. Considering a trend from the event research perspective means that trend detection has to be understood as a monitoring task. This is mostly the case for so-called short-term trends that are indeed triggered by some events and in order to detect them we have to monitor the stream in which they occur, e.g. the occurrence of "Eyjafjallajkull eruption"[4] which was reported in social networks and on the news in March 2010. However, so-called long-term trends, e.g. "financial crisis", that started to be on-topic in 2008 are not necessarily conjoined with one specific event. It is more a chain of events or even the "soft" indicators as public opinion or news. No sharp distinction has been made between the TDT and ETD research fields, which means that some research such as [Swan and Allan, 1999] or [Lavrenko *et al.*, 2000] can be in fact classified into both fields. Temporal data mining research [Mitsa, 2010] offers methods for clustering, classification, dimension reduction and processing of time-series data [Wang *et al.*, 2005]. It addresses in general the temporal data and the techniques of time series analysis on these data. One definition of temporal data is "time series data which consist of real valued sampled at regular time intervals" [Mitsa, 2010]. Temporal data mining applies the data mining methodology and deals with the same approaches for classification or clustering, that are relevant for mining trends in textual data.

4 Trend template

Based on our experiments and considerations, we outline the following assumptions about trends in the general context of this work;
A trend can be described by the following characteristics: trigger, context, amplitude, direction, time interval, and relation. Fig. 2 illustrates the trend template.
In 4.1, we more precisely define each characteristic.

4.1 Definitions

Trigger is a *thing*. They can be: an event, a person, or a topic anything that triggers the trend. A trigger can but does not have to cause a trend. A trigger makes the trend visible. An example of a trigger is *Lehman Brothers*[5] *insolvency* that can be classified as both a topic and an event.
Context is the area of the trigger. If the trigger is a topic then the context is this topic's area, e.g. *Lehman Brothers insolvency* is mentioned in the context of *real estate market*.
Amplitude is the strength of a given trend. It can be expressed by a number, the higher the number the more impact the trend has or by a qualitative value that describes the trend phase, e.g. beginning (setter), emerging (follower), mainstream, fading (conservative).
Time is necessary while spotting trend, since there can be no trend without time. It is the interval in which the trend is appearing, independent from the amplitude, e.g. the *real estate crisis* appeared between the years 2008-2011.
Relation expresses the dependency between the trigger and the context, it puts the given trigger, e.g. *Lehman Brothers insolvency* within the given context of the *real estate crisis* in a relation, e.g. *Lehman Brothers insolvency is part of the*

[4] The eruption of an Icelandic volcano in March 2010 that caused air travel chaos in Europe and revenue lost for the airlines `http://www.volcanodiscovery.com/iceland/eyjafjallajoekull.html` online accessed 04-17-2013

[5] `http://www.lehman.com/` online accessed 04-17-2013

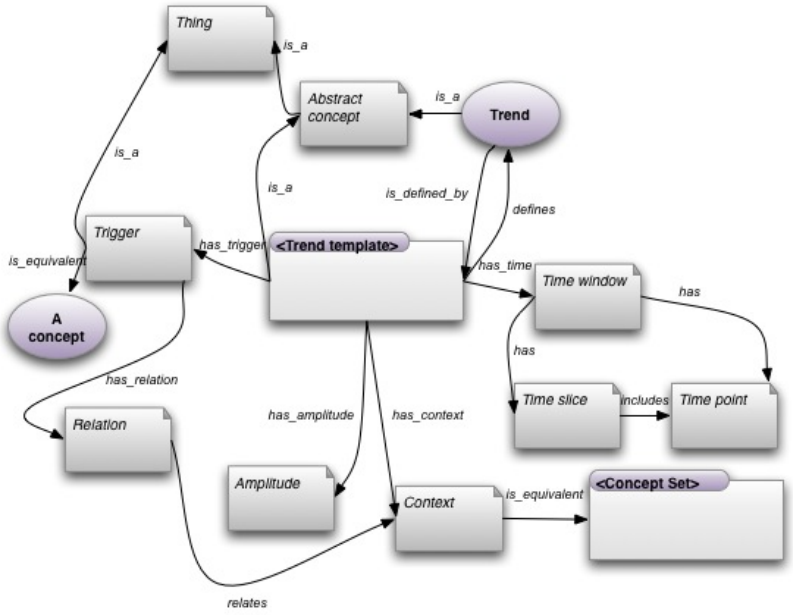

Figure 2: Trend template– an abstract conceptualization

real estate crisis.

4.2 Formal description

The trend template is an abstract model that describes the main concepts that are important and necessary for knowledge-based trend mining. In following, we more explicitly define the trend template:
DEF. 4.1: Trend template (TT) is a quintuple:

$$TT := \langle T, C, R, TW, A \rangle$$

where: T is trigger, C is context, R is relation, TW is time window, and A is amplitude.

DEF. 4.2: T- Trigger is set of concepts:

$$T := \{t_0, \ldots, t_n\}, n \in \mathbb{N} \wedge t \in T$$

so that if E, P, T_o are the sets defining:
events: $E := \{e_0, \ldots, e_n\}, n \in \mathbb{N} \wedge e \in E$
persons: $P := \{p_0, \ldots, p_n\}, n \in \mathbb{N} \wedge p \in P$
locations: $L := \{l_0, \ldots, l_n\}, n \in \mathbb{N} \wedge l \in L$
topics: $T_o := \{t_{o0}, \ldots, t_{on}\}, n \in \mathbb{N} \wedge t_o \in T_o$
then:

$$T := E \cup P \cup T_o \cup L$$

DEF. 4.3: C- Context is a union set consisting of a set of concepts and a set of relations between them where c is a context element:

$$C := C_{co} \cup R_{co}, c \in C$$

with C_{co} the set of concepts

$$C_{co} := \{c_{co0}, \ldots, c_{con}\}, n \in N \wedge c_{co} \in C_{co}$$

and R_{co} the set of relations:

$$R_{co} := \{r_{co0}, \ldots, rco_n\}, n \in N \wedge r_{co} \in R_{co} \wedge R_{co} \subseteq C_{co} \times C_{co}$$

whereas r_{co} defines a binary relation:

$$r_{co} : c_{cox}, c_{coy} \longrightarrow r_{co}(c_{cox}, c_{coy}) \wedge c_{cox} \neq c_{coy}$$

and the context element is defined by:

$$c = c_{co} \cup (c_{coi}, c_{coj})$$
$$C = C_{co} \cup C_{co} \times C_{co}$$

DEF. 4.4: *R*-Relational is a set of relations:

$$R := \{r_0, \ldots, r_n\}, n \in N \wedge r \in R \wedge R := \{T \times C\}$$

with

$$r_i : t_i, c_i \longrightarrow r_i(t_i, c_i)$$

DEF. 4.5: *TW*- Time window is a function that assigns time slice to the time points:

$$TP := \{t_{point} | t_{point} =$$
$$= ms \vee second \vee minute \vee hour \vee day \vee month \vee year\}$$
$$TS := \langle t_{point0} \ldots t_{pointn} \rangle$$
$$TW : TP \longrightarrow TS$$

DEF. 4.6: *A*- Amplitude is a function that assigns a value to the quadruple of $\langle T, C, R, TW \rangle$

$$A : T \times C \times R \times TW \longrightarrow \mathbb{N} \cup V$$

where N is the set of natural numbers and V is the set of categorical values

$$a : (t, c, r, tw) \longrightarrow n \vee v$$

5 Trend Ontology

One way of implementing the trend template is the realization of this model in the form of an ontology. We can understand the ontology as an instance of the trend template.

Based on the trend template described above, we created an applicable model, using SKOS[6] and RDFS/OWL[7] concepts and properties. Our model serves as a general model that can be extended regarding the particular application domain and applied for annotating a text corpus in order to retrieve the trend structure. The trend ontology is divided into levels meta, middle and low which correspond to three abstract layers of the model. Whereas the low level and the middle level relate to the corresponding application domain (in our case it is the German Stock Exchange, DAX), the meta level is the most interesting one. Meta ontology incorporates the general trend characteristics and can be applied to any application domain.
The central concepts of the ontology are *Trigger*, *TriggerCollection*, *Indication*, *Relational* and *ValuePartition* and have been modeled as subconcepts of *skos:Concept*, *skos:Collection* and *time:TemporalEntity*, with different semantic construction, e.g. *skos:related*, *skos:member*. The concepts mirror the composition of the trend template. *Trigger* consists of three subconcepts: event, person, location. The main goal of the meta ontology is to offer all necessary concepts and relations in order to span the trend template as a structure over a text corpus. To actually translate a specific document corpus into such a structure, meta ontology needs to be combined with a domain specific trend ontology which defines domain specific concepts, their keywords and possibly also their relations. This can either be done manually by extracting common terms as keywords and linking them to their respective concepts, or automatically by entity recognition. The pseudocode 6.1 describes the algorithm that we applied to build up the trend description on the test corpus.

Algorithm
6.1: CREATETRENDDESCRIPTION(c, o)

comment: parse $\forall$ **d**ocument $\in$ **c**orpus
comment: into **o**ntology
$parse(c, inO, outO)\{$
$model.read(inO)$
$create.reasoner(inO)$
for each $d\in$ c
do $\{$
$parse(keywords)$;
$match.model(keywords, inO)\{$
for $keyword \leftarrow 0$ **to** i
if $inO.concept.label$==$keyword$ **or**
$keyword\in$ inO.concept.label
$keyword.prefix$ **or** $keyword.postfix$==
$inO.concept.label.prefix$ **or** $.postfix$
then $matches.add(keyword)\}$
$relate.model(matches, inO)\{$
if $model.getRelation(matches).isEmpty$
then $model.createRelation(matches)$
else $model.incCounter(matches)\}\}$
$model.write(outO)$

6 Experiments

The text corpus which we call *German finance data*[8] that served as our test corpus consists of about 40,500 news articles related to the fields of business and finance, provided as XML files. The corpus is available in German and provides news articles from January 2007 to May 2008. The text was parsed in cooperation with neofonie[9] from the following sources: comdirect[10], derivatecheck[11], Handelsblatt[12], GodmodeTrader[13], Yahoo[14], Financial Times Deutschland[15], and finanzen.net[16].

In general, the content of the corpus is focused on finance and business information concerning German companies and stocks. It focuses on the situation at DAX, as well as on reviews and ratings of German companies and shares. For evaluation purposes regarding usefulness and practicability, the trend ontology has been filled with two different parts of the test corpus: stock market specific documents in Part 1 and the general business news in Part 2 (subsequently first and second part). They contain over 5,000 and 16,000 documents respectively. We specified several basic questions and respective queries as relevant for trends in general and specifically for stock market trends. Querying the ontology for the total occurrence of concepts yields the following output (shortened to some of the most relevant concepts): Germany (9,137), USA (4,808), Deutsche Telekom (442), Allianz (433), Switzerland (382), Starbucks (104). The output corresponds directly to the corpus of German stock news with a clear focus on German companies followed by the still dominant US market. A similar query for often mentioned lines of business in the context of Germany in contrast to the USA yields a major focus on the industry for Germany. 4.5% to 7.1% of the total occurrences of Germany appear in the context of different lines of industry. The USA is strong in the context of IT (9%) and services (6.9%). Moreover, we checked so-called topic structure by using our ontology. Here a general example for the concept *Germany*:

```
trendonto:#Germany (9137) has Topic
trendonto:#Financial : 1142
trendonto:#buy : 1003
trendonto:#MachineBuildingIndustry : 650
trendonto:#Share : 606
trendonto:#StockPrice : 562
trendonto:#Up : 520
trendonto:#Industry : 510
trendonto:#Investment : 468
trendonto:#Supplier : 422
trendonto:#AutomobilIndustry : 414
```

[6] http://www.w3.org/2004/02/skos/ online accessed 04-17-2013
[7] http://www.w3.org/TR/owl-features/ online accessed 04-17-2013
[8] Currently (May 2013) in the publishing process at Linguistic Data Consortium http://www.ldc.upenn.edu/
[9] http://www.neofonie.de, online accessed 04-25-2012
[10] http://www.comdirect.de/inf/index.html, online accessed 04-25-2012
[11] http://derivatecheck.de/, online accessed 04-25-2012
[12] http://www.handelsblatt.com/weblogs/, online accessed 04-25-2012
[13] http://www.godmode-trader.de/, online accessed 04-25-2012
[14] http://de.biz.yahoo.com/, online accessed 04-25-2012
[15] http://www.ftd.de/, online accessed 04-25-2012
[16] http://www.finanzen.net, online accessed 04-30-2012

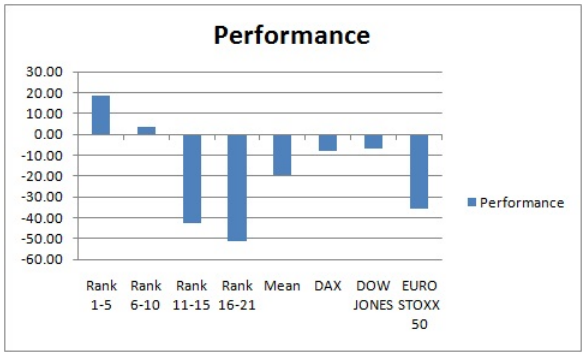

Figure 3: Performance of shares in the first corpus (5,000 documents) by ontology based ranking and comparison with share indices in the time window July 2007 to July 2011.

In Fig. 3 we show the comparison of the performance values for the stock markets as ranked by ontology (test based on time window: July 2007 to April 2008) and reported in real (time window July 2007 to July 2011). Applying the trend ontology to the test set enables to find out specific information about the certain trend that is described in the documents of the test set. Our preliminary experiments results that we partially present in this paper show that our idea of a trend template could help in harvesting knowledge from the given test data in a timely manner.

7 Conclusions and future work

This paper presents our research on knowledge-based trend mining, wherein the main contribution is our semi-formal model of a trend template. We showed that the implementation of the trend template in the form of a trend ontology allows for capturing the trend structure out of a test document set. Our experiments confirm that a knowledge-based approach for mining trends out of data allows for extended trend explanations. Currently we are comparing the trend ontology experiment results with the results from adapted K-Means clustering and LDA-based topic modeling algorithms applied on our test set.

Acknowledgments

This work has been partially supported by the "InnoProfile-Corporate Semantic Web" project funded by the German Federal Ministry of Education and Research (BMBF) and the BMBF Innovation Initiative for the New German Länder - Entrepreneurial Regions.

References

[Agrawal *et al.*, 1995] Rakesh Agrawal, Edward L. Wimmers, and Mohamed Zait. Querying shapes of histories. In *Proceedings of the 21st VLDB*, pages 502–514. Morgan Kaufmann Publishers Inc., 1995.

[Allan *et al.*, 1998] James Allan, Ron Papka, and Victor Lavrenko. On-line new event detection and tracking. In *SIGIR'98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 37–45. ACM, 1998.

[Allan, 2002] James Allan, editor. *Topic Detection and Tracking. Event-based Information Organization*. Kluwer academic publishers, 2002.

[Han and Kamber, 2006] J. Han and M. Kamber. *Data Mining Concepts and Techniques*. Morgan Kaufmann Publishers Inc., 2006.

[Kontostathis *et al.*, 2003] April Kontostathis, Leon Galitsky, William M. Pottenger, Soma Roy, and Daniel J. Phelps. *A Survey of Emerging Trend Detection in Textual Data Mining*. Springer-Verlag, 2003.

[Lavrenko *et al.*, 2000] Victor Lavrenko, Matt Schmill, Dawn Lawrie, Paul Ogilvie, David Jensen, and James Allan. Mining of concurrent text and time series. In *Proceedings of the 6 th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Workshop on Text Mining*, pages 37–44, 2000.

[Lent *et al.*, 1997] Brian Lent, Rakesh Agrawal, and Ramakrishnan Srikant. Discovering trends in text databases. In *Proceedings of the KDD'97*, pages 227–230. AAAI Press, 1997.

[Lita Lundquist, 2000] Robert J. Jarvella Lita Lundquist. *Language, Text, and Knowledge. Mental Models of Expert Communication*. De Gruyter, 2000.

[Mitsa, 2010] Theophano Mitsa, editor. *Temporal Data Mining*. Chapman Hall/CRC Press, 2010.

[Swan and Allan, 1999] Russell Swan and James Allan. Extracting significant time varying features from text. In *CIKM'99: Proceedings of the eighth international conference on Information and knowledge management*, pages 38–45. ACM, 1999.

[Swan and Jensen, 2000] Russel Swan and David Jensen. Timemines: Constructing timelines with statistical models of word usage. In *KDD-2000 Workshop on Text Mining*, 2000.

[Vejlgaard, 2008] Henrik Vejlgaard. *Anatomy of A Trend*. McGraw-Hill, 2008.

[Wang *et al.*, 2005] X. Wang, K. Smith, and R. Hyndman. Dimension reduction for clustering time series using global characteristics. In Vaidy Sunderam, Geert van Albada, Peter Sloot, and Jack Dongarra, editors, *Computational Science - ICCS 2005*, volume 3516 of *Lecture Notes in Computer Science*, pages 11–14. Springer Berlin / Heidelberg, 2005.

[Watts *et al.*, 1997] Robert J. Watts, Alan L. Porter, Scott Cunningham, and Donghua Zhu. Toas intelligence mining; analysis of natural language processing and computational linguistics. In *PKDD '97: Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 323–334. Springer-Verlag, 1997.

[Witten and Eibe, 2005] Ian. H. Witten and F. Eibe. *Data Mining Concepts and Techniques*. Morgan Kaufmann Publishers Inc, 2005.