2013 IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT)

Identify Emergent Trends based on the Blogosphere

Patrick Hennig
Hasso-Plattner-Institut
University of Potsdam, Germany
patrick.hennig@hpi.uni-potsdam.de

Philipp Berger
Hasso-Plattner-Institut
University of Potsdam, Germany
philipp.berger@hpi.uni-potsdam.de

Christoph Meinel
Hasso-Plattner-Institut
University of Potsdam, Germany
office-meinel@hpi.uni-potsdam.de

***Abstract*—Information about upcoming trends is a valuable knowledge for both, companies and individuals. Detecting trends for a certain topic is of special interest. According to the latest information over 200 million blogs exist in the World Wide Web. Hence, every day millions of posts are published. These blogs contain an enormous think tank of open-source intelligence. Considering the continuously growing nature of the World Wide Web a primary factor of success is the ability to include the latest data and focus on the complete data set of blogs. The structured as well as unstructured data of blogs are available offline via a single database for further analyses. This paper describes and evaluates an algorithm to detect trends based on the data published in blog posts.**

I. Introduction

An emerging trend is a topic of interest that is becoming more and more important over time. An old but often used example for an emerging trend is *Extensible Markup Language - XML* in the 1990s [1]. With the increasing amount of data that is available on the World Wide Web the need is arising to be able to detect such trends at an early stage.

Information about upcoming trends is considered to be a valuable source of knowledge for both, companies and individuals. A large number of market analysts working at monitoring a particular business field, with many employing manual methods to do so. Since the amount of available data on the internet is far too high for humans to monitor, which carries a major risk of substantial amount of information being missed, the necessity arose to detect emerging trends automatically.

It is a pretty difficult task to make all these information available offline. This is is carried out by the *intelligent crawler* [2] of the *BlogIntelligence* project.

Nevertheless, the analysis of this data is one of the key success factors. A lot of text-mining algorithms are already well known. However, in most cases they are used on a limited data set. Using a limited data set or using pre-aggregated results stands in contrast to the continuously growing nature of the World Wide Web. In former times this was not possible at all. Nowadays, in collaboration with an in-memory database the execution of these analyses becomes really fast. Therefore it is possible to provide results based on the latest data.

Since the blogosphere - all inter-connected weblogs - provides structured and unstructured data, text-mining algorithms have to be combined with the analysis of structured information in order to detect reliable trends automatically.

This paper presents an approach that combines trend-detection for structured and unstructured data. As a result, this approach fits perfectly to the semi-structured format of weblogs.

In order to demonstrate the usability of the approach presented in this work, the algorithm is evaluated by using a real blog data set gathered by the crawler running on the latest hardware at the HPI Future SOC Lab[1].

In the second section the scope of this project is discussed. Afterwards other techniques on this field of research are discussed in the related work section before the algorithm itself is presented in detail in Section IV.

In Section V the presented trend-detection algorithm gets evaluated to provide a deeper understanding of how this detection works in detail. In addition, in Section VI some future work is presented before the work is summarized in Section VII.

II. Related Work

In their article Kontostathis et al. [3] described different kinds of trend-detection systems. They divided the trend-detection systems into two main categories. The semi- automatic systems, which require user interaction for detecting emerging trends. Often these systems provide user-friendly reports and statistics for the user. In the beginning these systems supported a human expert in finding emerging trends. The second group is the group of fully-automatic trend-detection systems. The fully-automatic trend-detection approaches produce output without interaction with the user.

A trend-detection system consists of different components such as linguistic and statistical features, learning algorithms and visualization. Kontostathis and his group stated that a great deal of progress has been made towards automating the process of detecting emerging trends. Nevertheless, all systems try to present the results to the user in a user-friendly visualization.

A very sophisticated trend-detection model was introduced by Abe et al. [4]. Their system finds research keys in bibliographical data. They used a set of paper titles from two artificial intelligence conferences for their experiment. Their algorithm relies on the calculation of an importance index to find the most relevant words in the paper titles and monitor changes of this index over time. They presented really good results for their set of conference paper titles, but only focused on small paper titles for two selected conferences.

Currently, there a few approaches for trend-detection inside the World Wide Web with more or less good results. Abe et al.

[1] http://www.hpi.uni-potsdam.de/forschung/future_soc_lab.html?L=1

978-1-4799-2902-3/13 $31.00 © 2013 IEEE
DOI 10.1109/WI-IAT.2013.147

With *EnBlogue*, Alvanaki et. al [5] already attempted to use meta data provided in the World Wide Web, but they purely focused on *tags*. Therefore, this work provides an algorithm that focuses on different aspects provided by the structured and unstructured data of the blogosphere in order to detect trends.

III. Trend-Detection Prerequisites

To detect trends inside the blogosphere many different steps have to be performed first. Assuming that the necessary *structured data* has been extracted before, the initial steps for detecting emerging trends can be taken. This part is an essential factor for detecting trends later on.

A. Time Window

Due to the fact that the trend-detection should work with user input and detect the latest trends,a certain time frame has to be used. For example, the last three months or even weeks should be sufficient. It can be even meaningful for a user to play around with the size of this window

B. Importance Index

It is necessary to get the most important words or phrases from the content of a *blog post*. For text mining there are a few well-known methods available to calculate an importance value for each word of a document in a given document corpus. Assuming that the contents of all *blog posts* are the document corpus it is possible to use such an index for it.

a) TF-IDF: One of the most well-known importance index correlates the frequency of a certain term with the inverse document frequency (tf-idf). These terms are extracted beforehand by a *term extraction* method. To detect trends it is necessary to change this index slightly.

In order to detect trends, it is mandatory to know the importance of a term at a certain point in time in order to monitor changes. Therefore, the definition of the *tf-idf* has to be changed slightly. Hence, it can be defined as in the following.

$$t = t_j - t_{j-1} \tag{1}$$

$$TF{-}IDF(term_i, p_{tj}) = \frac{TF(term_i, P_{tj})}{TP(term_i, P_{tj})} \times log \frac{|P|}{PF(term_i, P)} \tag{2}$$

This describes the importance of a *term* for a certain time frame. Hereby, the specified time frame can be adjusted from days, over several hours to milliseconds. *TF* calculates the frequency of the term$_i$ in the set of posts P_{tj} inside a time frame, *TP* returns the number of terms within the set of posts P_{tj}, —*P*— is the number of posts and *PF* is the number of posts containing this term in the overall corpus. For later usage it is necessary to normalize the tf-idf measure.

C. Linear Regression

To detect trends an indicator has to be defined. For the detection of trends it is necessary to monitor changes over time; linear regression is perfectly suited for this work. Koegh et al. [6] described different time-series patterns that can be used for data mining and specified meanings for them. Therefore, a simple linear regression as described in the book *Introduction to Linear Regression Analysis* by Montgomery [7] is commonly used to monitor changes over time.

By using linear regression it is possible to calculate a trend line. For this trend line the slope and intercept is calculated using the following general definition.

$$Slope = \frac{\sum_{j=1}^{n} (y_{t_j} - \overline{y})(x_j - \overline{x})}{\sum\limits_{j=1}^{n} (x_j - \overline{x})^2} \tag{3}$$

$$Intercept = \overline{y} - Slope \times \overline{x} \tag{4}$$

The defined *slope* provides information about whether a trend is increasing or decreasing over time. In addition, the *intercept* value provides information about the baseline of the values. If the *intercept* is positive the linear regression line will not be increasing enough. Therefore, for an emerging trend the trend line has to have a negative *intercept*.

Furthermore, if the linear regression has a positive *intercept*, the trend can be popular or subsiding depending on whether the slope is positive or negative. These meanings are explained in more detail in Section IV-D3.

IV. Trend-Detection Algorithm

The different aspects monitored over time are described in more detail. An overview of all necessary steps including the preparation is shown in the following:

1) *Term extraction* based on the content of a *blog post*
2) Usage of a *time window* to get the latest trends
3) The *link* structure in the blogosphere has to be analyzed
4) *Calculation* of the *importance index* to find the most important words
5) The content of each *post* has to be analyzed
6) Usage of *tags* has to be analyzed
7) Finding *patterns* by performing a *clustering*
8) Finally assigning meanings to the clusters in order to detect emerging trends

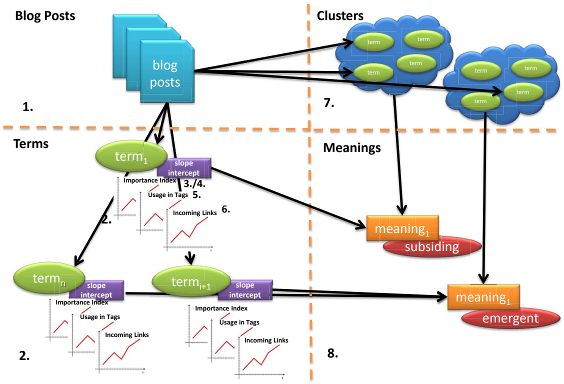

Fig. 1. Overview of the trend-detection algorithm

A. Link Analysis

One of the most powerful structures inside the blogosphere is the inter-linkage between *blogs* or *posts*. One reason for the power of the *links* inside the *blogosphere* is certainly based on the influence that incoming and outgoing links have to search engines.

This becomes particularly important for blogs, since a lot of different types of links exist between blogs. The links placed on the blogs home page are very powerful. These links often represent other blogs containing *links* to this *blog*. Since this is sometimes shown on all sub pages this can be a powerful link mechanism.

In addition to these links, links inside the content of a *post* are at least as powerful as *links* on the blogs home page. These *links* indicate *posts* even from other *bloggers* writing about a similar or related topic. For search engines these links are often very valuable regarding the given topic relevance. This is especially important, since it can trigger other discussions.

In addition there is another category of *blog links* that are used in comments. Since commenting goes along with a lot of abuse and spamming these links are almost irrelevant and have less power than other links in *posts* and *blogs*.

All of these *links* are often represented as traditional *hyperlinks* in the content of a *post*. In addition, it is often possible that some very important links are published in the *feed* of a *blog*. The *blog-crawler* is able to differentiate between these different kinds of *blog links*.

As already discussed, terms from *blog posts* and the link structure of *blog posts* are extracted. Therefore terms can be related with links.

Since a *tag* has a relation to a *post* it is possible to define the incoming links of those entities like in the following.

$$IL_p \text{ with } p \in P \tag{5}$$

$$IL_p = \{(p,x) | x \in P \land (p,x) \in RWP\} \tag{6}$$

$$P_{term,t} \text{ with } term \in TERM \tag{7}$$

$$P_{term,t} = \{p | p \in P \land (term,p) \in RTP \land TD(p) = t\} \tag{8}$$

$$il_{term,t} = \frac{\sum_{p \in P_{term,t}} |IL_p|}{|P_{term,t}|} \tag{9}$$

Hence, the indicators for the content can be defined as follows. For further combination of the indicators it is necessary to normalize this data set.

$$tspan_{i,j} = t_{i,j} - t_{i,1} \tag{10}$$

$$Slope_{link}(term_i) = \frac{\sum_{j=1}^{n}(|il_{term_i,tspan_{ij}}| - \overline{|il_{term_i}|})(tspan_{i,j} - \overline{tspan_i})}{\sum\limits_{j=1}^{n}(tspan_{i,j} - \overline{tspan_i})^2} \tag{11}$$

$$Intercept_{link}(term_i) = \overline{|il_{term_i}|} - Slope_{link}(term_i) \times \overline{tspan_i} \tag{12}$$

B. Content Analysis

The second aspect to look at is to analyze the content of each *post*. The different types of content are merged. Besides the real content of a *post* the database provides the *titles* and as well as the *short description* from the *feeds*. The shown indicators are based on the changed *importance index* defined in Section III-B in Definition 2. As stated previously, this definition measures the importance of a single word at a certain point in time. The usage of tf-idf values for detecting trends was used by Abe et el. [4] for the first time with titles of conference papers for two conferences.

First it is necessary to know the time span in between a certain time point and the first occurrence. This can be calculated by the following definition .

$$tspan_{i,j} = t_{i,j} - t_{i,1} \tag{13}$$

$$Slope_{cont}(term_i) = \frac{\sum_{j=1}^{n}(z_{term_i,t_j} - \overline{z_{term_i}})(tspan_{i,j} - \overline{tspan_i})}{\sum\limits_{j=1}^{n}(tspan_{i,j} - \overline{tspan_i})^2} \tag{14}$$

$$Intercept_{cont}(term_i) = \overline{z_{term_i}} - Slope_{cont}(term_i) \times \overline{tspan_i} \tag{15}$$

These indicators track the change of the *importance index* values over time. If there is a significant change, these indicators are one of three aspects that can classify whether a trend is coming up or not. It is important to note that these indicators have to be performed based on the *time window*, which is explained in Section III-A. If these indicators were used with the complete data set they become more and more resistant against changes.

C. Tag Analysis

Since the blogosphere consists of a semi-structured format it is not sufficient to focus on the different text sources inside the blogosphere. Furthermore, the quality of the trend-detection results can be improved by including additional meta-information. The best known parts of the blogosphere structure are *tags* and *categories*. *Tags* are keywords describing the content of a *post* in a concise form. This should make it more convenient to search inside the blogosphere.

The main problem with *tags* is that the usage highly depends on the *author* of a post. Some *authors* describe their content using too many different *tags*, while others completely forget to annotate their content with tags. Assuming that all authors use *tags* there is still the problem that *authors* can describe similar content with completely different words.

Lorelle VanFossen, a so-called *blog evangelist*, wrote a blog post especially about the problem with tagging[2]. She stated that it is very complicated for *blog authors* to use good *tags*, since they have to stop writing and start thinking like a reader would when searching for a specific topic. She also

[2] http://lorelle.wordpress.com/2005/12/12/the-problems-with-tags-and-tagging/

described how *tags* are in some way different to keywords; they are more like *categories* where the content can be grouped.

The intelligent blog-crawler [2] extracts *tags* as well as *categories*. Merging both elements and not differing between *tags* and *categories* is used for the presented approach.

Since each *post* has a time stamp as well, it is possible to analyze the usage of tags over time.

$$TU_{tg,t} \text{ with } tg \in TG \tag{16}$$

$$TU_{tg,t} = \{(tg,x) | x \in P \wedge TD(x) = t \wedge (tg,x) \in RTGP\} \tag{17}$$

$$tu_{tg,t} = |TU_{tg,t}| \tag{18}$$

Similar to the definition in Section IV-B it is possible to define a trend indicator for tags based on the data set shown above as follows.

$$tspan_{i,j} = t_{i,j} - t_{i,1} \tag{19}$$

$$Slope_{tag}(term_i) = \frac{\sum_{j=1}^{n}(tu_{term_i,t_j} - \overline{tu_{term_i}})(tspan_{i,j} - \overline{tspan_i})}{\sum\limits_{j=1}^{n}(tspan_{i,j} - \overline{tspan_i})^2} \tag{20}$$

$$Intercept_{tag}(term_i) = \overline{tu_{term_i}} - Slope_{tag}(term_i) \times \overline{tspan_i} \tag{21}$$

D. Trend-Detection

Finally, according to the indicators for each aspect from the previous Section final indicators have to be defined as well as meanings for these indicators. Three main aspects are analyzed. In order to get two final indicators for each term the indicators of the different aspects have to be combined with *term clustering* information.

1) Term Indicators: As a first step, it is necessary to define an overall indicator for the slope and intercept for each term.

$$OverallSlope(term_i) = \frac{Slope_{cont}(term_i) + Slope_{tag}(term_i) + Slope_{link}(term_i)}{numOfSetIndicators(term_i)} \tag{22}$$

$$OverallIntercept(term_i) = \frac{Int_{cont}(term_i) + Int_{tag}(term_i) + Int_{link}(term_i)}{numOfSetIndicators(term_i)} \tag{23}$$

Each slope function returns the slope for the given term. For the intercept function the behavior is similar to the one of the slope function. Therefore the *numOfSetIndicators* function returns the number of functions not returning zero.

2) Cluster Indicators: Since topic sensitive trends should be detected, it is not enough to calculate a slope and an intercept value for a single term. Since similar terms are grouped together to clusters by the *term clustering* done beforehand, it is necessary to classify whether a cluster represents an emerging trend or not. This is done by calculating the average slope and intercept within a cluster. This can be done by the following definition.

$$AvgSlope(c) = \frac{\sum_{term_i \in c} Slope(term_i)}{num(term \in c)} \tag{24}$$

$$AvgIntercept(c) = \frac{\sum_{term_i \in c} Intercept(term_i)}{num(term \in c)} \tag{25}$$

3) Meanings: Finally after calculating the indicators for each cluster, it is necessary to define meanings for these indicators. With these meanings it becomes possible to define whether a cluster contains a trend or not. As already described in Section III-C the slope and intercept is sufficient enough to classify a cluster as *emergent*, *popular* or *subsiding*. Since these meanings are well proven and used very often to classify trends, they work for the aspects of the blogosphere as well.

	AvgSlope	AvgIntercept
emergent	positive	negative
popular	positive	positive
subsiding	negative	positive

TABLE I. THE MEANING MATRIX

These meanings for the overall slope are shown in Table I.

V. EVALUATION

This Section evaluates the trend-detection algorithm itself in more detail. We want to take a deeper look whether detected trends are getting popular in the future or decreasing again.

A. Trend Prediction

Furthermore, an attempt is made to make an assumption as to how reliable the trend-detection system is using an experiment. Therefore the specified time window is divided into smaller windows of one week in order to compare the results of each window.

Therefore, a closer look at the trends that are classified as emergent inside the last week of July is necessary. In order to calculate the emergent trends for this week, the time window has to be reduced from six weeks to a single week. To make an assumption about how these trends will perform in the future, the following weeks have to be observed as well. In order to do so, it is necessary to move the time window one week ahead again and again. This experiment is focused on how many emergent trends from the first week will become popular trends in the following weeks.

Figure 2 shows the results of the experiment in more detail. In the last week of July, 3695 emergent trends can be identified. In the following weeks these emergent trends are observed. In the second week, from 3695 emergent trends, 1494 emergent

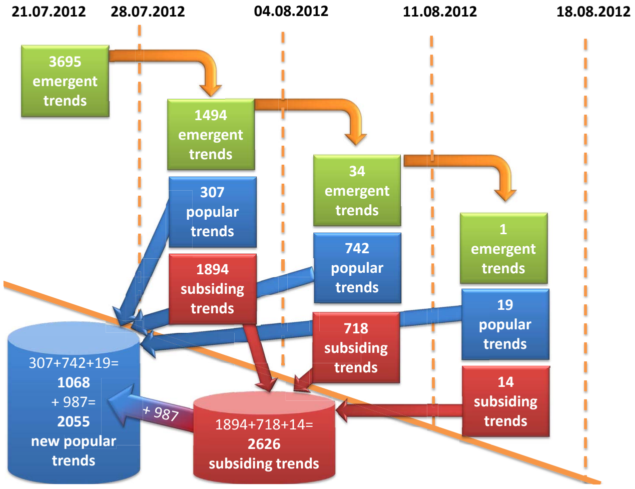

Fig. 2. Reliability of trends

trends are still classified as an emergent topic. Nevertheless, already 307 trends can be classified as popular. But 1894 of the 3695 emergent trends are again decreasing. This is not bad at all, since they have the possibility to become a popular trend again. Since the focus of the experiment is not on the subsiding trends, a trend is excluded as soon as it is classified as a subsiding trend once.

In the third week, only 34 of the 1494 as emergent trends from the week before are still emergent. This gets clearer by looking at the popular trends in the third week. From the 1494 as emergent classified trends, 742 are getting popular in this week.

Finally, in the last week there is only one emergent trend left and last but not least 19 of the 34 emergent trends are getting popular in the last week.

Furthermore from the trends that became subsiding in one week during the experiment time window, 987 of these 2626 as subsiding classified trends are getting popular as well within at least the last week of the experiment.

Finally this experiment showed that 2055 of the 3695 as emergent classified trends are really getting popular trends within the next three weeks. Of course, this is just a very limited experiment. Nevertheless, it shows that the trend-detection could probably be used for prediction of trends. Therefore, further evaluation should be conducted.

VI. Future Work

Sentiment analysis or *opinion mining* [8] makes it possible to identify whether a sentence has a positive or negative meaning. They try to analyze the words used in a sentence and provide a classification from -3 up to +3 in order to express the positive or negative meaning.

This can be useful for detecting trends as well since the classification of *emergent* or *subsiding* can be misunderstood and it goes along with positive or negative feelings. But nevertheless, the classification of *emergent* or *subsiding* has nothing in common whether it is a positive trend or a negative trend that is becoming more and more important over time.

Therefore, more classification levels for trends could be possible. By integrating a *sentiment analysis* for each trend, each *subsiding*, *emergent* or *popular* trend can be classified as well as a positive or negative trend. This would give the user a better understanding of the trends and would deliver more valuable information.

VII. Conclusion

For trend-detection based on the blogosphere three different aspects are taken into account in the presented work. As a first step it is necessary to prepare the unstructured data. Extracting terms, measuring which words are the most important ones and clustering similar terms are some key steps that have to be performed up-front. As a consequence the trend detection can take the unstructured information into account.

This work described an algorithm for the detection of trends based on the structured and unstructured data inside the blogosphere and evaluated the different parts of the algorithm by using a real data set from the blogosphere gathered by the *BlogIntelligence* tailor-made blog-crawler.

That information can help a user, on the one hand, to get more information about a certain topic and get a variety of up-to-date knowledge that is coming up in the blogosphere.

On the other hand, information about trends can be fundamental information for businesses. It is possible to comprehend sales figures and changes in sales figures based on the blogosphere. This can help to react faster in the future if the market changes and gives the possibility to adjust the direction of a company accordingly.

References

[1] A. Kontostathis, L. Galitsky, W. M. Pottenger, S. Roy, and D. J. Phelps, “A survey of emerging trend detection in textual data mining,” 2003.

[2] J. B. C. M. Patrick Hennig, Philipp Berger, “Mapping the blogosphere - towards a universal and scalable blog-crawler,” in *Proceedings of the Third IEEE International Conference on Social Computing (Social Com2011)*. MIT, Boston, USA: IEEE CS, 10 2011, pp. 672–677.

[3] A. Kontostathis, L. M. Galitsky, W. M. Pottenger, S. Roy, and D. J. Phelps, “A Survey of Emerging Trend Detection in Textual Data Mining,” *Language*, pp. 1–44, 2003.

[4] H. Abe and S. Tsumoto, “Evaluating a temporal pattern detection method for finding research keys in bibliographical data,” pp. 1–17, Jan. 2011.

[5] F. Alvanaki, S. Michel, K. Ramamritham, and G. Weikum, “En Blogue – Emergent Topic Detection in Web 2 . 0 Streams,” in *Proceedings of the 2011 international conference on Management of data - SIGMOD ’11*, New York, New York, USA, 2011, p. 1271.

[6] E. Keogh, S. Chu, D. Hart, and M. Pazzani, “Segmenting time series: A survey and novel approach,” in *In an Edited Volume, Data mining in Time Series Databases. Published by World Scientific*. Publishing Company, 1993, pp. 1–22.

[7] D. Montgomery, E. Peck, and G. Vining, *Introduction to linear regression analysis*, 3rd ed., ser. Wiley series in probability and statistics. New York, NY [u.a.]: Wiley, 2001.

[8] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” *Found. Trends Inf. Retr.*, vol. 2, no. 1-2, pp. 1–135, Jan. 2008.