Black Swan: Augmenting Statistics with Event Data

Johannes Lorey[1] Felix Naumann[1] Benedikt Forchhammer[2]
Andrina Mascher[2] Peter Retzlaff[2] Armin ZamaniFarahani[2]

Hasso Plattner Institute, Potsdam, Germany
[1] firstname.lastname@hpi.uni-potsdam.de [2] firstname.lastname@student.hpi.uni-potsdam.de

ABSTRACT

A large number of statistical indicators (GDP, life expectancy, income, etc.) collected over long periods of time as well as data on historical events (wars, earthquakes, elections, etc.) are published on the World Wide Web. By augmenting statistical outliers with relevant historical occurrences, we provide a means to observe (and predict) the influence and impact of events. The vast amount and size of available data sets enable the detection of recurring connections between classes of events and statistical outliers with the help of association rule mining. The results of this analysis are published at `http://www.blackswanevents.org` and can be explored interactively.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*

General Terms

Algorithms

1. COMBINING STATISTICS AND EVENTS

Over the course of human history there have always been so-called *Black Swan* events. These historic events are defined as singular occurrences with high impact that were entirely unexpected at their time, but in retrospect appear to be the logical consequence considering the given preconditions. Examples for such events include World War I and the burst of the dot-com bubble. The term originates from the discovery of black swans in Western Australia in 1697. Until then, people had widely used the phrase Black Swan to refer to impossibilities, as it was common (Western) belief that only white swans could exist. Subsequently, the definition of the term has changed radically to the one above [7].

By combining statistical data with information on events, the Black Swan project provides a tool to aid domain experts (such as historians or econometricians) in identifying important events throughout history. The goals are (i) the automated detection of outliers in global statistics and their annotation with suitable events and (ii) the discovery of patterns of how event types and statistical developments influence one another.

In order to achieve such insight it is necessary to process statistical data, where large deviations from an expected statistical development could suggest Black Swans events. Fortunately, over the past years a reasonable amount of statistical data has been openly published by governments and other organizations. Other Internet resources and databases contain information on a large number of historical events, which may be Black Swan events.

The rest of this paper is organized as follows: Sec. 2 highlights some of the foundations of the system, such as the chosen data sources as well as the definition of event and outlier for our application. The remaining sections each explain one component of the system architecture illustrated in Fig. 1. In Sec. 3, the *extraction* component for statistical and event data is introduced. Section 4 details the outlier detection *analysis*, whereas Sec. 5 illustrates the *rule mining* process for deriving patterns between outliers and events. Finally, Sec. 6 describes the *visualization* of our results as annotated statistical graphs and provides an outline of the proposed demonstration.

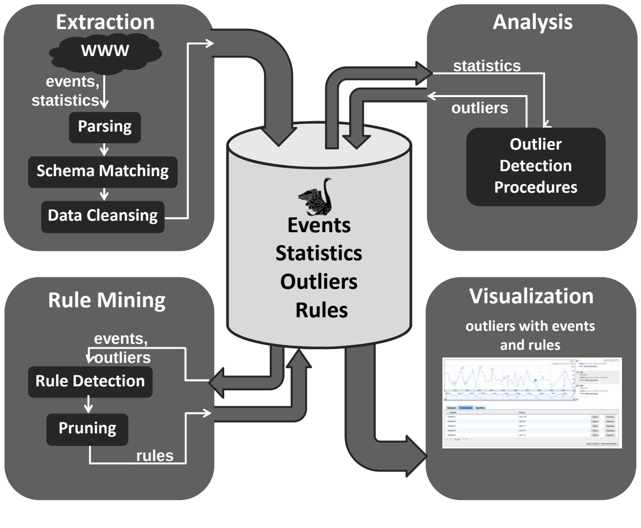


Figure 1: System architecture diagram

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

2. THE DATA

A key prerequisite for our prototype is the creation of a database containing statistical and event data. We achieve this by extracting and merging information from a number of different Internet sources as detailed in Sec. 3. Events in our system are characterized by a date, location, title, and event category. Optionally, instead of a single date, an event can also be described by a time period with start and end date. The event category provides a hierarchical taxonomy for natural disasters, political events, military conflicts, etc.

A statistical data point is characterized by the numerical value of the data point and the type, location, and year of the statistic. For instance, one value might characterize the annual average income in the United Kingdom for the year 1969 in USD. Statistical data outliers are extrema that deviate from the underlying tendency of a graph. There are several approaches to detect them as discussed in Sec. 4.

Our sources for statistical data are international organizations, such as the World Bank or IMF, but also data provided by other projects, such as Gapminder[1] or Correlates Of War[2]. For event data, we exploit DBpedia[3], EMDat[4], NOAA[5], Correlates Of War, Freebase[6], and the BBC historical timeline[7]. While EMDat and NOAA mostly contain information about natural disasters, such as droughts or earthquakes, the BBC timeline pages provide a chronology of key events in each country's history, but mostly political events. In addition, Correlates of War provides information on wars, and the DBpedia and Freebase data sets contain structured data from user-generated content.

Using the statistical outliers and events data, we detect patterns between these classes automatically. The aim of this rule mining is to decrease the number of events possibly causing an outlier in a statistic and to find probable causes for an event. Details on the rule generation are described in Sec. 5.

3. EXTRACTION

The first step in our workflow is the extraction of events and statistical data from various online sources. This process can be subdivided into three steps: parsing, schema matching, and data cleansing.

3.1 Parsing

As a first step, we retrieve and parse event data from the sources listed in Sec. 2. This goal is achieved by a number of flexible parsers, which are able to handle structured (e.g., CSV, HTML/XML, RDF) as well as unstructured formats, i.e., plain text. Depending on the source, the parsing step itself is preceded by a preprocessing step in which irrelevant parts of source data (e.g., header/footer in HTML documents) are removed. Our flexible design allows for easy adaption of existing parsers to different formats as well as integration of new parsers. We use Apache Tika[8] to parse structured sources (e.g., HTML documents) and extract relevant text, and we use the Jena Framework[9] to load and query RDF data.

With this approach we are able to extract more than 43,000 distinct events from the sources mentioned in Sec. 2. Around half the events stem from collections about natural disasters (primarily from EMDat and NOAA).

The statistics we use include around 400 specific indicators (e.g., GDP or literacy rate) collected annually for as many as 200 countries and up to 200 years. It should be noted that obviously not all statistical data is available for all countries for such a long period of time. Overall, our data comprises approx. 1,250,000 individual statistical values.

3.2 Schema Matching

The second task includes the mapping of parsed data from their source schema to the unified event and statistics schemata. Because of the different data structures in the sources used, the details for this mapping process have to be configured manually for each source. The following normalization steps are employed for schema matching:

- **Attribute de-duplication**: Semi-structured sources such as DBpedia may contain multiple attributes that essentially have the same meaning. These attributes are merged into one attribute where applicable.
- **Categorical classification** of events and statistics into a predefined hierarchical structure of categories. If a taxonomy already exists in a source, the classification can be done using a static mapping between the existing and our classification. Alternatively, classification is performed using a machine learning approach, which is especially useful for unstructured sources (text classification).
- **Geospatial classification** of events and statistics using the GeoNames database and web-service[10]. Multiple locations are allowed for events.
- **Title generation** for sources that do not allocate titles to events. Generation is performed dynamically depending on values of other attributes (like category and location).
- **Date normalization**: Date and time values of different formats are mapped to a unified schema.
- **Value normalization**: Statistical values are normalized to account for different value ranges and units. Prognostic values are removed, because we are only interested in analyzing actual historical data.

Before an entity is stored in the database a number of checks ensure that the entity is *valid* and *useful*. For example, events need to contain at least a title, year, location, and category in order to be useful for rule mining. Likewise, missing or non-numerical values in statistical time series are ignored as they might cause the data analysis algorithms to produce incorrect or no results at all.

3.3 Data Cleansing

It is likely that the same event or similar events can be found in different data sets. Using the following steps, we were able to identify and fuse around 2,500 of these duplicate events. Details on all used methods are described in [5].

First, we adapted the *Sorted Neighborhood* algorithm in a way that a window contains all events within five con-

[1] http://www.gapminder.org

[2] http://www.correlatesofwar.org

[3] http://dbpedia.org/About

[4] http://www.emdat.be

[5] http://www.ngdc.noaa.gov

[6] http://www.freebase.com

[7] for the UK, see for example http://news.bbc.co.uk/2/hi/europe/country_profiles/1038820.stm

[8] http://tika.apache.org/

[9] http://jena.sourceforge.net/

[10] http://www.geonames.org

secutive years: only the events within this time frame are compared with one another. To decide on a duplicate, all attribute values are compared using different similarity measures and weights. Here, the titles and locations have the highest impact. The weight of the start date is increased for natural disasters, as this date is typically easier to pinpoint for these occurrences than it is for, e.g., political events. Two titles are compared by using a modified *MongeElkan* distance metric that divides the similarity sum of the most similar words by the number of words from the shorter title. All words are stemmed by the *Porter Stemmer* algorithm, stop words are eliminated and the similarity of two words is defined as their *JaroWinkler* distance.

Second, these duplicates need to be clustered. We considered clustering by deriving the transitive closure of individual duplicate candidates, but found that this is computationally too expensive and will result in a low precision value. Thus, we cluster by using Nearest Neighbors and Compact Sets with varying thresholds for different groups [2].

Finally, to create a clean event we fuse an individual event cluster by fusing all contained attributes. We choose the title and start date from the most reliable source and take the latest end date if given. The values of most other attributes can be concatenated.

4. DETECTION OF OUTLIERS

To identify interesting aspects in the statistical data, we have to determine outliers in the values. We implemented several methods to achieve this goal.

Linear regression uses a linear model to describe the relationship between a scalar variable (time, in our case) and a dependent variable, which corresponds to a specific quantitative indicator, such as the gross domestic product in USD or the literacy rate in percent. We then defined an outlier to be a point on the graph that differs noticeably from the estimated linear model, i.e., it has a large residual. To find the linear approximation of a curve, we implemented three different algorithms: MM-estimation [8], least squares, and its robust variant least median squares [6]. For the latter method, the linear model is defined only by points that themselves are not regarded as outliers for this model.

As linear regression is suited only for some specific data sets, we also implemented two variants of *non-parametric regression* analyses, using the Loess-function [3] and a generalized additive model (GAM) [4], respectively.

We also experimented with methods that analyze properties of the graph itself. For instance, we implemented an algorithm that simply defines all extrema of a graph as its outliers. A more elaborate approach is to analyze not the extrema, but the slope of the graph and define those points as outliers, where the absolute change of the slope is above a certain threshold.

The last approach we implemented is an algorithm that defines a "*global statistic*" by calculating the mean of all country specific statistics and analyzing each single statistic according to its relation to the global statistic.

Each of the methods mentioned above has certain advantages and disadvantages and may be suitable only for specific data sets, e.g., indicators that are expected to increase linearly over time. While the application's 'default' setting should provide a good starting point for exploration, domain experts can also choose to apply a more appropriate analysis to the underlying data by selecting the corresponding algorithm in the web interface.

5. ASSOCIATION RULE MINING

One of our goals is to detect interesting patterns of event-outlier-combinations, such as 'in case of a major natural disaster, the annual GDP of a country declines'. In this section, we explain the process of finding these patterns using association rule mining. We start by defining the term *rule*, outline the attributes used for the machine learning process, and finally describe the rule generation process itself.

5.1 Data Preparation

A rule consists of a premise X and a consequence Y, as well as some metrics describing the quality of the rule, such as support, confidence, and conviction. Support is a measure for the frequency of attribute combinations in the data set. Confidence states how often the consequence follows from a given premise and is defined as

$$\mathrm{conf}(X \Rightarrow Y) = P(Y|X) = \frac{P(X,Y)}{P(X)}$$

Conviction is a metric on how strongly a rule holds when compared to purely random effects between premise and consequence and is defined as

$$\mathrm{conv}(X \Rightarrow Y) = \frac{P(X)P(\neg Y)}{P(X \wedge \neg Y)}$$

Association rule mining aims at discovering dependencies between variables in a single large database. The first step in data preparation was therefore to extract such a data set from our relational model. Because our interest lies in the correlation between events and outliers, we use a join of statistic outliers with events happening in the same year and location (country) as the basis for our data set. In the end, this data set contains information on the event's category, the statistic category, the tendency of an outlier, and the statistical trend leading up to the outlier. For example, in our data set we discovered that a governmental change led to a local extremum in a previously ascending consumer price index for this country in 383 cases with a conviction value of approx. 2.09. In other words: A change in government typically yielded a decline in the consumer price index.

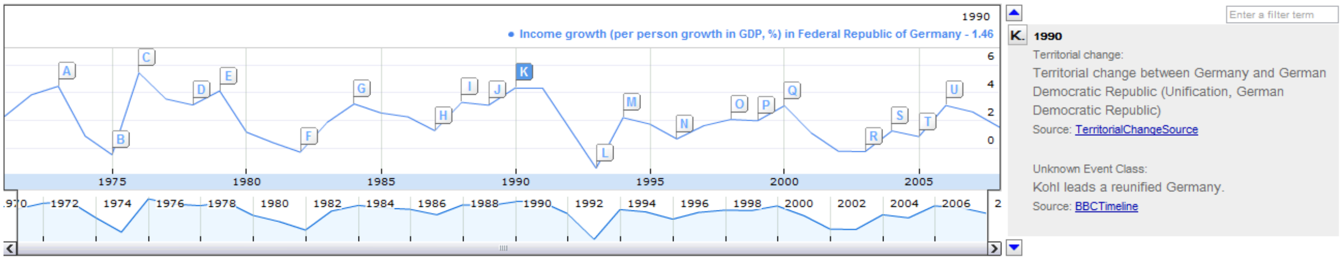
The tendency of an outlier expresses whether the outlier is a maximum or minimum in respect to the expected statistical trend. The statistical trend in the years leading up to an outlier can help to identify Black Swans for which the statistical development in general is relevant to the event, instead of just considering the outlier itself. An example for this is the widespread increase of weapons production that preceded World War I.

5.2 Rule Generation

For association rule mining we used the open source machine learning software WEKA[11]. We employed the Apriori algorithm [1] for rule mining.

One challenge in rule mining is to select the attributes one expects to occur in rules discovered during the process. For our problem we required that every rule needs to contain at least the event category and the statistical category or indi-

[11] http://www.cs.waikato.ac.nz/ml/weka/



K. 1990

Territorial change:

Territorial change between Germany and German Democratic Republic (Unification, German Democratic Republic)

Source: TerritorialChangeSource

Unknown Event Class:

Kohl leads a reunified Germany.

Source: BBCTimeline

Figure 2: The effect of the German reunification on income growth in Germany.

cator. Other attributes, e.g., outlier tendency or historical trend, are also permitted, but not required in a rule.

Since our base data set for rule generation contains about 1.1 million combinations (by joining events, statistics, and outliers for each year), we required that each determined rule must meet a minimum support of approx. 0.01% by considering only those implications with at least 100 occurrences.

As the Apriori algorithm builds all combinations of item subsets, some generated rules do not constitute a useful rule that can be visualized. This includes rules without an event or statistic attribute as well as events with an unknown event class. We prune the resulting set by removing all such rules.

In order for a user to determine the most fitting event for an outlier, each rule returns the conviction value for a given combination of an event and outlier. This value indicates to what degree a rule applies to the given combination. The more rules match for the combination and the higher the conviction, the more likely the event is for the given outlier.

6. DEMONSTRATION

We now describe how we used the generated rules to visualize statistics and annotate them with event data. The goal is to enable users to explore the correlation between statistical data and historical events interactively. Our prototype is available at `http://www.blackswanevents.org`.

On opening the prototype's website, the user may start to filter the available statistics by lists of categories, subcategories, statistic names and countries. Once a country and a statistic have been selected, the graph containing annotated outliers can be added to the screen. The user may change the displayed timespan, add more graphs by selecting other statistics or countries, and view details about the events.

6.1 Visualization of outliers

The main use case of the visualization is the interactive exploration of our data. Using the Google Web Toolkit[12], we implemented a backend that provides an asynchronous service, which can be used to query available statistics for one or more countries. The backend then retrieves this statistic from the database and scans it for outliers using one of the algorithms described in Sec. 4. The result contains the statistic's data along with a set of outliers. These outliers are mapped to rules and thus to potentially influencing events by using the rules' score method as described in Sec. 5.

Now that all outliers and corresponding events for a statistic and country have been identified, the client generates an "annotated time line" using the Google Chart Tools[13]. Each outlier is annotated with one or more corresponding event(s). By clicking on an outlier, event details are provided, including the rule that caused individual events to be associated with the outlier. For example, in Fig. 2 an outlier (a maximum) in income growth in Germany has been detected for the year 1990, which according to a mined rule is ascribed to the German reunification in the same year.

6.2 Visualization of rules

The second use case is meant to enable users to find relations between certain event classes and specific categories of statistics more easily. It does so by presenting a list of rules, which were identified by the rule mining algorithm. The user can choose one of the rules and review a selection of graphs that support this specific rule. Thus, it is easier to find and validate interesting and unexpected relations between events and statistical data.

7. ADDITIONAL AUTHORS

Soeren Discher[2], Cindy Faehnrich[2], Stefan Lemme[2], Thorsten Papenbrock[2], Robert Christoph Peschel[2], Stephan Richter[2], Thomas Stening[2], Sven Viehmeier[2]

[2] Hasso Plattner Institute, Potsdam, Germany,
firstname.lastname@student.hpi.uni-potsdam.de

8. REFERENCES

[1] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 487–499, Santiago de Chile, Chile, 1994.

[2] S. Chaudhuri, V. Ganti, and R. Motwani. Robust Identification of Fuzzy Duplicates. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 865–876, Tokyo, Japan, 2005.

[3] W. S. Cleveland. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 74(368):pp. 829–836, 1979.

[4] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall, 1st edition, 1990.

[5] F. Naumann and M. Herschel. *An Introduction to Duplicate Detection*. Morgan and Claypool Publishers, 2010.

[6] P. J. Rousseeuw. Least Median of Squares Regression. *Journal of the American Statistical Association*, 79(388):pp. 871–880, 1984.

[7] N. N. Taleb. The Black Swan: The Impact of the Highly Improbable, 2007.

[8] V. J. Yohai. High Breakdown-Point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics*, 15(2):pp. 642–656, 1987.

[12] `http://code.google.com/webtoolkit/`

[13] `http://code.google.com/apis/chart/`