COMPLEX EVENT PROCESSING AND DATA MINING FOR SMART CITIES

Alexandra Moraru, Dunja Mladenić
Artificial Intelligence Laboratory
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
e-mail: alexandra.moraru@ijs.si, dunja.mladenic@ijs.si

ABSTRACT

Complex Event Processing (CEP) is emerging as a new paradigm for continuous processing of streaming data in order to detect relevant information and provide support for timely reactions. The main role of a CEP engine is to detect the occurrence of event patterns on the incoming streaming data. However, the problem of discovering the event patterns, although strongly related to the data mining field, has not been studied from the perspective of CEP applications.

This paper presents the first steps towards defining a framework that would allow seamless integration of CEP and data mining method. We present the smart cities scenarios as a good working-field for experimentation. A concrete use case is discussed and preliminary results are presented for real-live data that has been collected.

1 INTRODUCTION

The avalanche of data which information systems had to face in the last years influenced their evolution and characteristics. Continuous, on-time processing of incoming data streams imposed particular requirements [1], which traditional Database Management Systems (DBMS) were not able to fulfil. Consequently, due to the market needs, new tools have been developed, able to process multiple data sources, often streams, in a timely fashion in order to extract relevant information. Grouped under the domain of *event processing* (or, according to [2] information flow processing domain), two main types of such systems have emerged: Data Stream Management Systems (DSMS) and Complex Event Processing (CEP) systems.

The term *event processing* here refers to a broad study area. In [3] the term of *event processing* is coined to "any form of computing that performs operations on events". The key concept is that of an **event** which can represent anything that happens or is observed as happening (e.g. a mouse click, a sensor reading, water level increase, a river flood, spring coming, etc.). A common characteristic of event processing applications is to continuously receive such events from different **event sources** (e.g. sensors, software modules, blogs, etc.). The central module processing the events, called the CEP **engine**, detects **event patterns** from the incoming data streams and outputs the detected or predicted complex events which can be further used by other **event consumers**, or it can return as an input to the CEP engine. The event patter's role is to specify how the incoming events should be processed in order to extract relevant information. The language used to define these patterns should have the ability of specifying complex relationships among events flowing into the CEP engine.

The typical approach in defining patterns of events is to manually specify them. This is done either by domain experts, capable of providing the definition of event patters or by using other tools externally of the CEP systems in order to discover these patterns and then encode them in the event processing language (EPL). However, we see the integration of machine learning algorithms with the CEP system, as a solution for direct support in definition of event patterns. Although massive amount of research has been conducted in the areas such as pattern recognition and multisensor data fusion, the systems developed for many of the CEP applications do not provide a seamless integration with such techniques, but rather consider the human component responsible for defining the complex events patters that should be monitored and detected. Therefore, an important improvement for applying machine learning algorithms in event-based application is to develop a framework that would allow easy integration of existing algorithms with event processing techniques.

A first step in achieving such integration is choosing a scenario for running experiments. One example is the smart cities scenarios, as there can identified many data sources and use cases for data mining and CEP. In this paper we are proposing such a scenario, identify the data sources and run preliminary experiments for analysing the data. Future steps are discussed in the direction of using CEP engines with the patterns discovered and defining a framework for an easier integration of data mining and CEP.

The rest of the paper is structured as follows: Section 2 describes the smart cities scenario and introduces one use case considering the city of London. Data integration and preprocessing is presented in Section 3, while Section 4 discusses the result of data mining. Finally we conclude the paper and identify future directions.

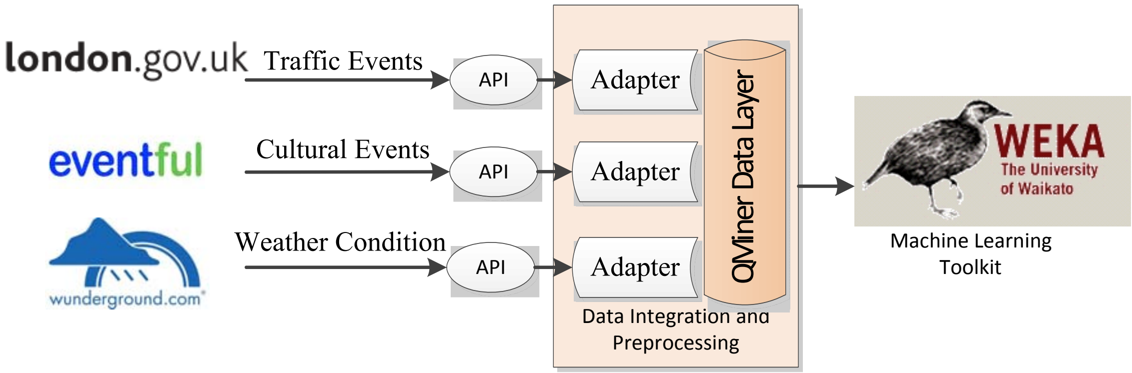

Figure 1. *Data Sources.*

2 SMART CITIES SCENARIO

The high level requirements for making a city smarter, as envisioned by IMB in the larger Smarter Planet[1] program, refer to collaboration and coordination between city agencies managing different domains (e.g. water management, transportation, buildings, etc.) in order to be able to optimize the limited resources and to efficiently and effectively deliver city services. Moreover, different technologies may enable smarter cities, such as: communication channels (e-mail, instant messaging, etc.), business rules, data sharing (data models, accessibility) and integration of different sources of data [4].
In another study [5], the classification of cities as smart is made based on 6 criteria: economy, people, governance, mobility, environment and living. Out of these, we focus on smart mobility, which refers to transport (accessibility, modern transport systems) and availability of ICT infrastructure.

2.1 London Use Case

The final goals for our experimental scenario will be to (1) find patters for appearance of traffic disruptions that could be then applied by a CEP engine for sending different alarms and (2) discover interesting correlations between cultural events happening in a city, social media and their influence on traffic.
The first step toward our goals is to identify data sources of potential useful information. Some specific sources are listed below:

- Traffic data (bus schedules and delays, congested roads, etc.). Sources: Bing Maps[2], Traffic for London[3] (Tfl).
- Weather conditions. Sources: Weather Underground[4], Yahoo! Weather[5], AccuWeather[6], etc.
- Events happening in the city: Live music, conferences, festivals, galleries, sports, etc. Sources: Eventful.com, upcoming.org, last.fm, zvents.com, socialevents.com.
- Social media about the events (microblogging and news). Sources: Twitter[7], IJS newsfeed[8].

2.2 Description of the Data Sources Used

After receiving the data through the data sources APIs, custom built adapters are used for storing data in a uniform data structure, which allows us to integrate all the sources for generating the input dataset for data mining. As illustrated in Figure 1, for the data storage functionality we have used the QMiner infrastructure which is based on tightly integrated and scalable custom software modules.
The data mining algorithm applied is for learning association-rules. The Weka toolkit [7] was used for running the experiments.
For our preliminary results we have used data only from the sources depicted in Figure 1, which have been crawled through several API made available by the source providers. Depending on the how often the sources were updated, different time intervals were used for crawling data as can be observed in **Error! Reference source not found.**; data was collected for a period of one month, between 16th of July to 16th of August 2012.

Table 1: *Time intervals for data collection*

Source	Update time interval
Tfl Road Disruptions	5 minutes
Current Weather Conditions[4]	30 minutes
Events (from Eventful.com)	Once per week

The road disruption events are identified with an unique id and have the following properties: start and end time, location details, time of last update, type, severity and category. The category property is described in **Table 2** , as it has predefined values which are used in the analysis of the results, while for the rest of properties more details can be found in [6]. The total number of road events registered is 3090. The type of the events indicates if the event has

[1] http://www.ibm.com/smarterplanet/us/en/
[2] http://www.bing.com/maps/
[3] http://www.tfl.gov.uk/
[4] http://www.wunderground.com/
[5] http://weather.yahoo.com
[6] http://www.accuweather.com/
[7] http://twitter.com/
[8] http://newsfeed.ijs.si/

Table 2: *Categories of road events and their frequency*

Category	**Interpretation**	**Percentage**
Works	planned and emergency road works	7%
Accident	road traffic accidents	30%
Signal Failure	automatic traffic signals failure	6%
Breakdow n	vehicle breakdown	19%
Incident	emergency incident	4%
Event	cultural events	<1%
Hazard	dangerous structures, fire, flooding, ice and spillages	6%
Other Cause	abnormal loads, unexplained congestion, etc.	27%

happened during the period for which the data was crawled, or is it scheduled to happen in the future (caused by road works or other planned cultural events). Therefore, the number of road events analysed is 2305, and the percentage for each category is also presented in **Table 2**.

The weather conditions represent a report for the whole area of London city, and presents properties such as: temperature, wind speed and a short text description (e.g.: clear, partly cloudy, rainy, etc.). Unfortunately the precipitation information, which could presumably have an influence on traffic disruptions, was not available from this data source.

The cultural events are described by time, location, performers and can be of more categories (e.g. music, concert, arts, sports, etc.). The total number of events happening collected is 5931, in 1707 different locations. The top 10 most frequent categories of events are presented in **Table 3** (one event can have more categories).

3 DATA PREPROCESSING

The data preprocessing consisted of two main steps: (1) data integration and (2) generation of the set of instances. In the data integration step, the traffic events were correlated with cultural events and weather data based on time and location. We have also defined nearby events, as follows: if an event is situated at predefined maxim distance from the current event and starts with a predefined maximum time interval before or after the current event, than it is considered as nearby.

Table 3: *Frequent categories of cultural events*

Category	Freq.	Category	Freq.
Music	2652	Sales	215
Performing Arts	1138	Festivals-Parades	200
Other	887	Family-fun-kids	157
Singles social	279	Outdoors-recreation	124
Sports	351	attractions	100

Table 4: *Instance Attributes*

Attribute	Type	Distinct / [min-max]	Missing Values
category	Nominal	8	No
severity	Nominal	3	No
day	Nominal	7	No
time	Nominal	4	No
duration	Nominal	3	27%
traffic events nearby	Nominal	2	No
weather	Nominal	12	2%
Temperature (˚C)	Numeric	[10-30]	2%
wind speed (kph)	Numeric	[0-32]	2%
events nearby	Nominal	2	
cultural events categories (25)	Nominal	2	No

The maximum distance and time interval have been tested with different values, as it will be explained in the next section.

In the second step, we have created a set of instances around the road events, where each instance has 37 attributes, described in **Table 4**. As there can be more categories of cultural events, each is represented as an attribute which has value *t* if category is present and *f* otherwise. An example of an instance is illustrated in **Figure 2**. The total number of instances is equal to the number of road events analysed, which is 2305.

Hazard, Moderate, medium, Monday, Afternoon, 0, Mostly Cloudy, 16, 26, 1, f, f, t, f, t, f, f, f, f, f, f, f, f, f, f, f, f, f, f, f, f, f, f, f

Figure 2: *Example of an instance from the input dataset*

4 PRELIMINARY RESULTS

First, analysis was done on each attribute. The analysis with respect to day of the week and time of the day is illustrated in **Figure 3**, and presents the expected correlation.

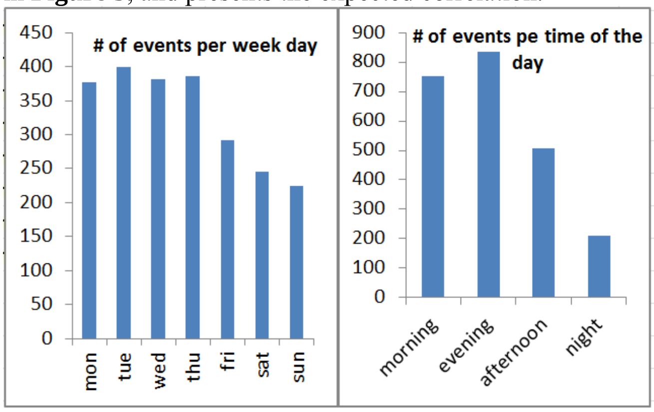

Figure 3: *Number of events per week day (left) and time of the day (right)*

Another aspect of importance is the presence of nearby events (road or cultural) for an instance of road event. This has been calculated for three sets of parameters, and the results in **Table 5** show the number of instances where nearby traffic events (#tfl), respectively cultural events (#evt) exist.

Table 5: *Number of instances that have nearby events for different constraints on distance and time difference*

Distance (m)	Time diff. (mins)	#tfl	#evt
500	60	9	104
1000	60	31	213
1000	240	96	487

Possible correlations between the attributes of our dataset have been studied using association rules. The algorithm used for discovering such rules is the Apriori algorithm, implemented in Weka. We choose the dataset with the constraints of 1000 meter in distance and 60 minutes time difference. As Weka crashed when running the algorithm on all the attributes we first try removing all the categories of cultural events from the attributes, reducing the number of attributes to 10. However no relevant rules were found.

As our interest was in the relation of different traffic events categories (listed in **Table 2**) with nearby cultural events, we have reduced the dataset to 213 instances (for which the constraints on distance and time where 1000 meters, respectively 60 minutes), which had at least one nearby cultural event. Although the rules obtained are not necessarily related to traffic events, they do illustrated normal relations, such as: cultural events are more often in the evening (rule 1) or that some cultural events categories are related (rule 2)

Rule 1: Weather = Clear, music = t, performing_arts = t (23) ==> Time=Evening (21) [conf:(0.91)]

Rule 2: singles_social = t. performing_arts = t (33) ==> music = t (28) [conf:(0.85)]

5 CONCLUSIONS AND FUTURE WORK

A number of conclusions can be drawn after our preliminary experiments as follows. Data must be collected for a longer time period, allowing thus generation of cleaner and more precise datasets. Although the 27% of missing values for the duration attribute (see **Table 4**) is not very high, it does influence the calculation of the nearby events. In our experiments we used an approximation of average duration, for being able to obtain other nearby events. A larger dataset would allow for extraction of data of better quality, which can then be used more successfully for association-rule learning.

Determining the nearby events may be done differently for different categories of cultural events; (e.g. a big music concert affects a larger area then a sales event) a more thoroughly study for determining the parameters involved is needed.

There is no clear evidence of the influence of cultural events over road events. However, common sense tells us that there should be. Therefore we shall continue our study once we gather more data.

As future steps we consider integration of more data sources, and first we will focus our attention on social media. Next we will also consider visualisation techniques that can provide a faster insight into the data, and then proceed with data mining methods.

Finally, once the event pattern obtained, we will research into connecting them to a CEP engine.

Acknoledgements

This work was supported by the Slovenian Research Agency and the ICT Programme of the EC under PlanetData (ICT-NoE-257641), PASCAL2 (IST-NoE-216886) and ENVISION (IST-2009-249120).

References

[1] Stonebraker, M., Çetintemel, U., Zdonik, S.: The 8 requirements of real-time stream processing. ACM SIGMOD Record. 34, 42-47 (2005).

[2] Cugola, G., Margara, A.: Processing Flows of Information : From Data Stream to Complex Event Processing. ACM Computing Surveys.

[3] Niblett, P.: Event Processing In Action. (2010).

[4] Wang, Q., Meegan, J., Freund, T., Li, F.T., Cosgrove, M.: Smarter City: The Event Driven Realization of City-Wide Collaboration. 2010 International Conference on Management of e-Commerce and e-Government. 195-199 (2010).

[5] Giffinger, R., Fertner, C., Kramar, H., Kalasek, R., Pichler-Milanovic, N., Meijers, E.: Smart cities Ranking of European medium-sized cities. , Vienna, Austria (2007).

[6] Transport for London, Live Traffic Disruptions – Data Dictionary, (last accessed 1st September 2012), http://www.tfl.gov.uk/assets/downloads/businessandpartners/data-dictionary-live-traffic-disruptions.pdf

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten. The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, Issue 1, 2009.