Websensors Analytics: Learning to sense the real world using web news events

Ricardo M. Marcacini
Federal University of Mato Grosso do Sul (UFMS)
Três Lagoas, MS, Brazil
ricardo.marcacini@ufms.br

Rafael G. Rossi
Federal University of Mato Grosso do Sul (UFMS)
Três Lagoas, MS, Brazil
rafael.g.rossi@ufms.br

Bruno M. Nogueira
Federal University of Mato Grosso do Sul (UFMS)
Campo Grande, MS, Brazil
bruno@facom.ufms.br

Luan V. Martins
Federal University of Mato Grosso do Sul (UFMS)
Três Lagoas, MS, Brazil
luan.martins@aluno.ufms.br

Everton A. Cherman
Onion Technology
ParqTec - Science Park
São Carlos, SP, Brazil
everton@oniontecnologia.com.br

Solange O. Rezende
Mathematical and Computer Science Institute (ICMC/USP)
São Carlos, SP, Brazil
solange@icmc.usp.br

ABSTRACT

An event is defined as "*a particular thing which happens at a specific time and place*" and can be extracted from news articles, social networks, forums, as well as any digital documents associated with metadata describing temporal and geographical information. In practice, this knowledge is a digital representation (virtual world) of various phenomena that occur in our physical world. The manual analysis of large event collections is not feasible, thereby motivating the development of intelligent data analytic tools to automate the knowledge extraction process. In this paper we present a computational tool called Websensors Analytics that uses machine learning methods for learning sensors from events to monitor and understand various phenomena in the real world. Websensors Analytics is the first initiative to analyze events in Portuguese and currently contains all the necessary features for extracting and analyzing knowledge from events: (i) web crawling to collect events in real time, (ii) statistical and natural language preprocessing techniques for event extraction (iii) machine learning methods for learning sensors, and (iv) Application Programming Interface (API) using the Websensors Analytics infrastructure. The Websensors Analytics tool is potentially useful for media analytics, opinion mining, web engineering, content filtering and recommendation systems – for both academic research and industrial applications.

KEYWORDS

Machine Learning; Web Engineering; News Events; Websensors

1 INTRODUCTION

In recent years, online content publishing platforms have allowed significant growth of digital data repositories on a wide range of topics [24]. A significant part of these data are represented by unstructured textual data, since it is a natural way to represent human knowledge. However, a manual analysis from this huge volume of data extrapolates the human capacity. Thus, intelligent data analytic tools have been very promising to automate the knowledge extraction from these repositories and support decision-making processes [16].

Among the various types of textual data published in online platforms, the analysis of events extracted from digital documents is a subject that has received great attention in the literature [9–11]. An event is defined as "a particular thing which happens at a specific time and place" and are published in news articles, social networks, and forums [3] — both by government agencies as a way of providing transparency to public management as well as by private sectors and organized society. The main motivation of these studies is the premise that the analysis of events published in *web* represents a way of mapping the digital world to the physical world. In this sense, it is possible to develop computational methods to extract knowledge from event databases to understand real-world phenomena such as epidemic monitoring and forecasting [8, 20], opinion and sentiment analysis [12], urban violence monitoring [7], school dropout prediction [25], monitoring and warning of natural disasters [21], as well as analysis of social, political and economic trends [1, 17, 23].

One of the most important steps in event analysis is the organization of related (or similar) events in clusters. The idea is that if the user is interested in an event of a particular cluster, then he will also be interested in other related events of that same cluster [5, 7]. Data clustering is a machine learning method that allows the organization of events in clusters through a criterion of proximity [2]. A good proximity measure consider various "components" of events, such as place of occurrence (where), date of publication (when), causes and effects of events (what), and related persons and organizations (who). After extracting the event components and obtaining the clustering model, we learn "websensors" to monitor events of interest, thus generating alerts and reports according to the requirements of each application domain. However, the quality of these websensors can be significantly improved if such clustering models truly reflect events and phenomena from the real world [20]. This is a recent and challenging research area that presents gaps in relation to computational tools, algorithms and evaluation in practical scenarios.

In: XVI Workshop de Ferramentas e Aplicações (WFA 2017), Gramado, Brasil. Anais do XXIII Simpósio Brasileiro de Sistemas Multimídia e Web: Workshops e Pôsteres. Porto Alegre: Sociedade Brasileira de Computação, 2017.

ISBN 978-85-7669-380-2.

In this paper we present a computational tool called Websensors Analytics that uses machine learning methods for learning sensors from events to monitor and understand various phenomena in the real world. Websensors Analytics is the first initiative to analyze events in Portuguese. The development of the tool started in 2013 and currently contains all the necessary features for extracting and analyzing knowledge from events: (i) web crawling to collect events in real time, (ii) statistical and natural language preprocessing techniques for event extraction (iii) machine learning methods for learning sensors from a event dataset, (iv) controlled access through webservices, and (v) Application Programming Interface (API) using the Websensors Analytics infrastructure. The Websensors Analytics tool is potentially useful for applications related to media analytics, opinion mining, web engineering, content filtering and recommendation systems – for both academic research and industrial applications.

2 WEBSENSORS ANALYTICS

In this section we present an overview of the Websensors Analytics tool. The architecture is divided into three main steps (Figure 1): (i) Web Crawling and Data Storage, (ii) Preprocessing and Extracting Events and (iii) Learning Sensors from Events. Each step involves a set of features available in the tool. We also present examples of applications for exploratory analysis of events and their respective mapping with a phenomenon of interest that occurred in the physical world.

2.1 Web Crawling and Data Storage

Digital documents are collected from predefined sources such as news portals, information agencies, and government websites. There is a control policy to include a new information source, which involves its area of activity, life time and the existence of an editorial team. Although the inclusion of a new information source may be requested by users, we believe that current crawlers collect a wide variety of subjects that are useful for the most application domains. Thus, there is an expectation that the dataset of collected events is representative of the phenomena occurring in the real world.

The data structure for the data crawling is XML-RSS (Really Simple Syndication)[1]. Thus, even if the data source only displays information in HTML structure, Websensors Analytics crawlers are able to parse the HTML content and obtain the corresponding XML-RSS.

Web crawling is carried out in a distributed way, in which currently two research groups (LABIC/USP and GEPIC/UFMS) collaborate in the regular execution of the crawlers. Every minute, crawler machines receive their own lists of web information sources. A central server manages the priority of crawling sources by considering an "older first" technique.

The data storage is based on a big data solution, in particular, an integration with Apache Hadoop and Mysql databases. The collected data are available to users of the Websensors Analytics through a REST webservice that responds to queries using JSON objects. Crawling status and the recent collected data can be monitored in real time at http://websensors.net.br/api/crawler.

[1]XML-RSS: https://www.w3schools.com/xml/xml_rss.asp

2.2 Preprocessing and Extracting Events

Events are extracted from digital documents, such as news articles. For this task, both statistical and linguistic techniques are used. Natural Language Processing (NLP) techniques are important in defining the event components, in general, based on the following named entities:

- **Geographic Information (where):** names of places and regions are extracted from the textual data, as well as identifying their latitude and longitude coordinates (geocoding technique). The Websensors Analytics tool uses the GeoNames[2] database to support this feature. Moreover, we have developed machine learning techniques for geographical entity disambiguation based on linguistic attributes of textual documents.
- **Temporal Information (when):** the date of publication of the news, as well as techniques for standardization of temporal expressions of the texts are used to extract temporal information. This temporal information is very useful for the alignment between a news event and a phenomenon that occurred in the real world.
- **Names of People and Organizations (who):** related events can be identified by analyzing people names and organizations involved in these events. Thus, the Websensors Analytics tool allows to identify proper names and verify the occurrence of these proper names in knowledge bases, such as Wikipedia, to obtain a list of people names and organizations.

While NLP techniques are used to identify the named entities, statistical techniques are important to identify causes and effects of the events, which are related to the **"what" component**. In this case, we propose a technique to extract related-topics called AL^{2}FIC (*Active learning for frequent itemset-based text clustering*) [14], which allows the use of domain information (*e.g.* domain expert users) or a knowledge dataset about facts of the real world (e.g. Wikipedia). Topics that co-occur in various news articles are selected and the temporal information determines when a given topic represents the cause (predecessor time information) or the effect (successor time information).

The preprocessing step allows obtaining an candidate event set $E = \{e_1, e_2, ..., e_n\}$ from a news article dataset. Each event is defined as a quadruple $e_i = \{where, when, who, what\}$. We propose an approach based on the Maximum Likelihood method to obtain a model that determines the m selected events ($m < n$) as defined in Equation 1.

$$log(p(E|\theta)) = log(\prod_{i=1}^{n} p(e_i|\theta)) = \sum_{i=1}^{n} log(\sum_{j=1}^{m} p(e_j^{sel}) p(e_i|e_j^{sel}, \theta)) \quad (1)$$

In Equation 1, θ represents the model parameters and $p(e) = p(where)p(when)p(who)p(what)$ indicates the probability of occurrence of an event from the observed data — considering that the components of the event are independent. In relation to the term $p(e_i|e_j^{sel}, \theta)$, given a selected event e_j^{sel}, the four components of the

[2]GeoNames: http://www.geonames.org/

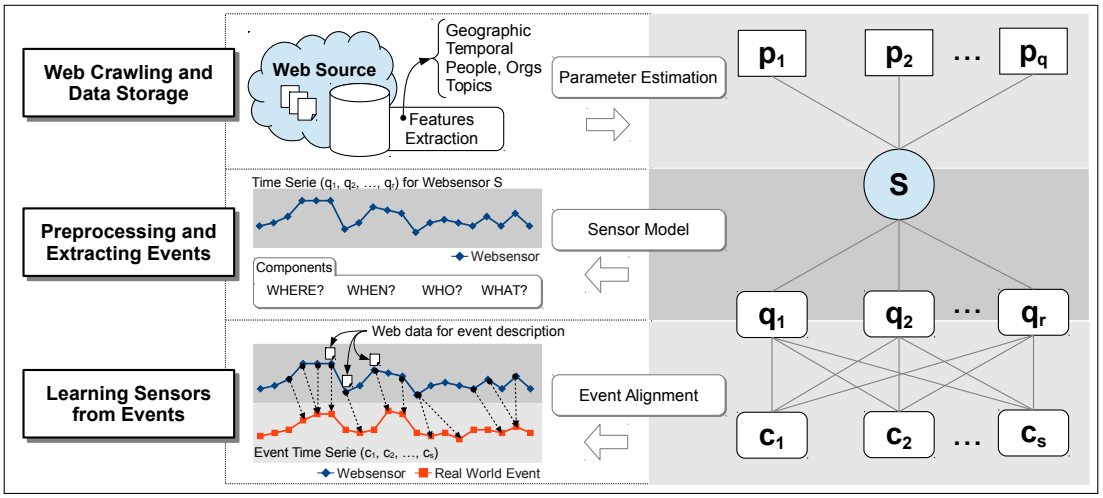

Figure 1: General architecture of the Websensors Analytics tool.

i-th candidate event are conditional independent, *i.e.*, $p(e_i|e_j^{sel}) = p(where_i|e_j)p(when_i|e_j)p(who_i|e_j)p(what_i|e_j)$.

The log-likelihood maximization problem can be solved by different methods, especially by iterative methods based on Expectation-Maximization. Users can retrieve events through our REST webservice from the Websensors Analytics tool, which returns a maximum of 5000 events per query. A real-time demonstration of the latest collected events is available at http://websensors.net.br/api/events/.

2.3 Learning Sensors from Events

We define a **websensor** as a "data model" that represents a set of related events as well as their occurrences over time. This data model can be obtained from the events using machine learning algorithms, in particular, with data clustering algorithms. However, in addition to clustering related events, we also have the requirement to identify which clusters are most related to a real-world reference event. In practice, we need to align the web events with real-world events.

Real-world events are dynamic and updated over time. Thus, these events are commonly represented by time series, such as number of accidents per week, rate of disease spread, price quotation on the stock exchange, student dropout rate, frequency that a term is used in search engines (e.g. Google Trends[3]), to cite few. Thus, the sensor learning step of the Websensors Analytics is important to answer two research questions:

(1) How to build clusters of related events for websensors learning?
(2) How to align websensors with real world events?

In relation to the first question, several clustering algorithms can be used. However, it is necessary to define a good measure of proximity between events. In general, this measure is a simple linear combination of various proximity measures — a measure for each component of the event. In previous works, we propose more advanced strategies such as the use of semi-supervised clustering with geographic and temporal constraints [7], metric learning [15, 19] and consensus clustering [6]. All of these alternatives are available in the Websensors Analytics tool.

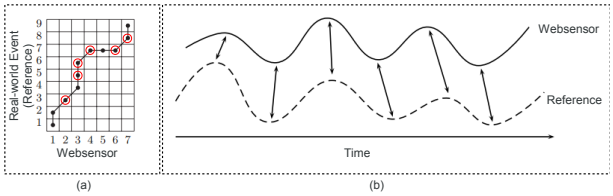

Figure 2: Alignment between the Websensor and the reference time series. Adapted from [18].

In relation to the second question, we propose a technique based on DTW (Dynamic Time Warping) to align websensors (digital world) with real-world events [13]. Figure 2 shows an example of an alignment between the time series of a websensor and the time series of a reference event. It is important to note that the time series of the websensor indicates an occurrence signal of the event-cluster over time. The temporal granularity of the series may be different (such as weeks vs. months), since the alignment performed is non-linear. DTW technique allows a distortion in relation to the time axis. This is important to analyze which are the past websensor events that tend to affect future behavior of the reference event.

The non-linear alignment between the time series is performed as follows. Let M be a distance matrix between the observations of the websensor time series Q and the reference event time series C. The cell M_{ij} indicates the distance $ws(q_i, c_j)$ between the signal c_i of the websensor at time i and the signal c_j of the reference event at time j. The non-linear alignment is represented by the least cost path in the M matrix, according to the recurrence relation described in Equation 2.

[3]Google Trends: https://trends.google.com/

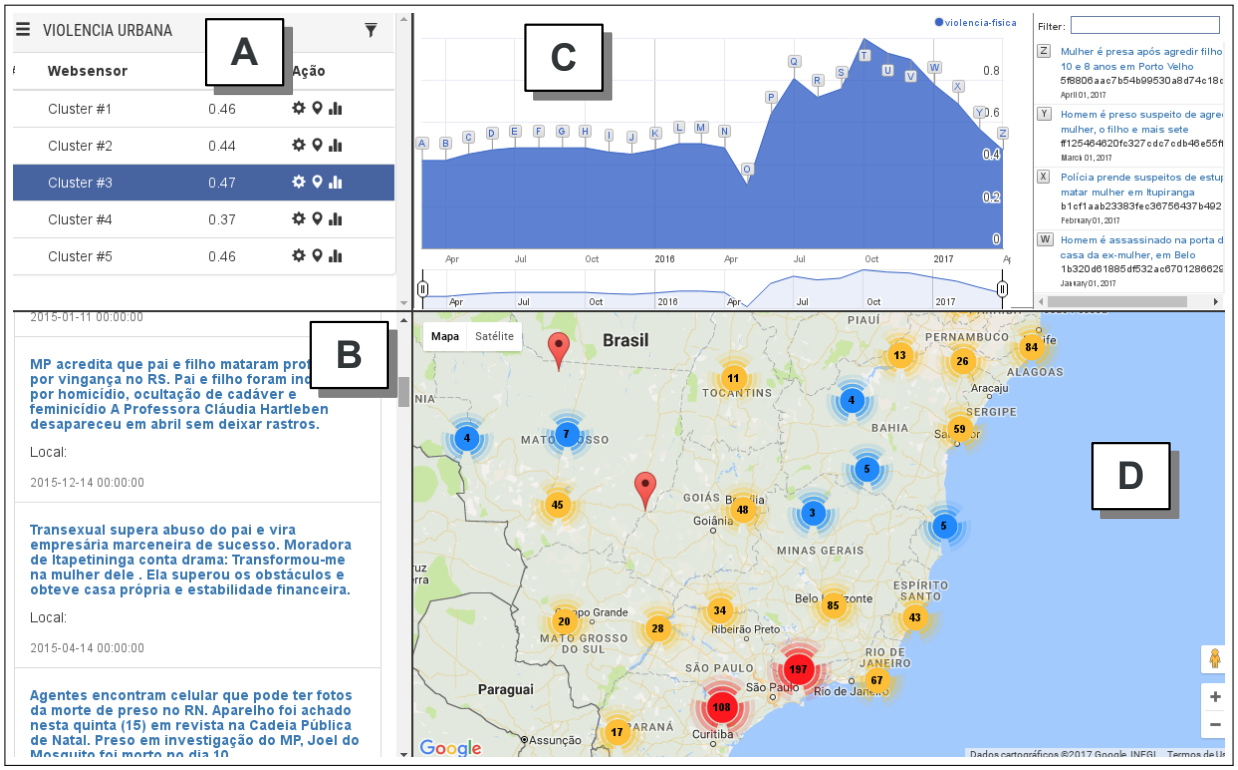

Figure 3: An example of Websensors News Analytics using the template available in the tool [7].

$$DTW(Q,C) = ws(q_i, c_j) + min\{ DTW(q_i - 1, c_j - 1), DTW(q_i - 1, c_j), DTW(q_i, c_j - 1) \} \quad (2)$$

The non-linear alignment between the websensor and the real-world event is very useful in identifying which web events were most important in each time period. Thus, users can analyze the real-world event through various web events that can explain and describe a particular phenomenon. This is one of the main contributions of Websensors Analytics tool and is best illustrated from the application examples presented in the next section.

3 EXAMPLE OF APPLICATIONS

Websensors Analytics tool have been used in some practical applications. In [13], we demonstrated that websensors can be used to improve the prediction of pulp productivity from the time series provided by BRACELPA[4], where we used an event dataset about precision agriculture, government investments, pests and the development of biotechnology on pulp domain. For example, Figure 4 illustrates the DTW alignment between a websensor on advances in forest biotechnology and Brazil's pulp export rate [22] — obtained from Websensors Analytics tool.

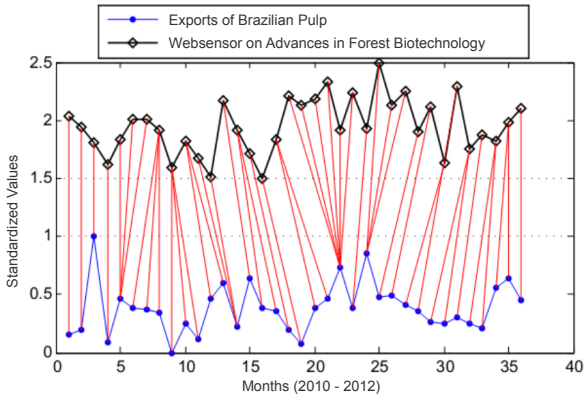

Figure 4: Alignment between the websensor time series (Advances in Forest Biotechnology) and the reference time series (Exports of Brazilian Pulp).

Another interesting application is the use of websensors for monitoring fires in agriculture — generally used for clearing cropland areas. We use events about the social impact of these fires, in particular, events related to respiratory problems caused by smoke. The reference event is the number of fire alerts identified by the INPE[5] satellite. The results provide evidence for a significant correlation

[4]BRACELPA - Brazilian Pulp and Paper Association

[5]INPE - National Institute of Space Research

between the social effects of fires, even allowing early alerts about periods in which the number of hospitalizations due to respiratory problems will increase [4].

Recently, the Websensors Analytics tool has also been exploited in industrial applications. In particular, there is the development of Websensors applications for recommendation tasks in the area of education, such as student dropout prediction, as well as the food industry by monitoring the food profile of clients via social networks, their location and menu items of restaurants[6].

The Websensors Analytics tool allows the user to develop their own interface for data visualization, in general, adapted to each application. However, our tool provides a template that can be used in several scenarios. Figure 3 shows an example of a News Analytics for events on urban violence in Brazil [7] — which was based on the template provided in the Websensors Analytics tool. Figure 3A shows the websensors that obtained the best alignment with the reference event time series. Figure 3B shows the events of each websensor selected by the user. Figure 3C shows the temporal evolution of the selected websensor as well as the main events of each period. The user can explore the websensor time evolution to understand the phenomenon of the real world. Finally, in Figure 3D the regions in which each event occurred are shown on the map.

We believe that Websensors Analytics is a useful tool for exploratory analysis, especially to map and visualize the event components (where, when, who, and what) that represent phenomena of interest that occur in the real world. Other applications have been developed and are available on the project homepage at http://www.websensors.net.br/.

4 CONCLUDING REMARKS

In this paper we presented an overview of the Websensors Analytics tool. Due to space constraints, we focus on describing the features of the tool — rather than describing in depth the methods involved. However, we present references to our previous work on each feature of the tool.

The Websensors Analytics tool provides a complete solution for Event News Analytics, with the advantage of aligning web events with real-world events. The main way to use our tool is through REST webservices, which involves exchanging JSON objects. Thus, it is possible to incorporate our tool in several applications in a transparent way. Moreover, we also presented a News Analytics template that is useful for a variety of scenarios and can be easily adapted to other applications.

We presented some urls in each section with demonstrations of the stages of our tool. In addition, an overview of the Websensors Analytics tool features is available at http://gepic.ufms.br/wfa2017/.

The Websensors Analytics tool is accessed through tokens and distributed from two proprietary licenses: academic and industrial. The academic license allows the free use of the tool for application development in research institutions, as well as collaborations between organizations and universities. The industrial license allows the use for companies with commercial purposes and can be requested at the main site of the project.

[6]Onion Menu: http://www.onionmenu.com.br/

ACKNOWLEDGMENT

The authors acknowledge the Brazilian Research Agencies FUNDECT-MS (Project 147/2016 - SIAFEM 25907), FINEP, CNPq, CAPES, and FAPESP (Process 2015/50074-6) for their support to this work. The authors also thank the NVIDIA Grant Program for the donation of the equipment to enable high-performance experiments.

REFERENCES

[1] Robert Ackland. 2013. *Web social science: Concepts, data and tools for social scientists in the digital age.* Sage.
[2] Charu C Aggarwal and ChengXiang Zhai. 2012. A survey of text clustering algorithms. In *Mining text data.* Springer, 77–128.
[3] James Allan. 2012. *Topic detection and tracking: event-based information organization.* Vol. 12. Springer Science & Business Media.
[4] T Amaral, M Nesso, and R Marcacini. 2015. Machine Learning for Websensors: Applications for monitoring fires from news (In Portuguese). In *SIICUSP-USP.*
[5] Jack G Conrad and Michael Bender. 2016. Semi-Supervised Events Clustering in News Retrieval. In *Recent Trends in News Information Retrieval Workshop.* 21–26.
[6] Geraldo N Corrêa, Ricardo M Marcacini, Eduardo R Hruschka, and Solange O Rezende. 2015. Interactive textual feature selection for consensus clustering. *Pattern Recognition Letters* 52 (2015), 25–31.
[7] Ronaldo Florence, Bruno Nogueira, and Ricardo Marcacini. 2017. Constrained Hierarchical Clustering for News Events. In *Proceedings of the 21st International Database Engineering & Applications Symposium.* ACM, 49–56.
[8] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457, 7232 (2009), 1012–1014.
[9] Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, Franciska de Jong, and Emiel Caron. 2016. A Survey of event extraction methods from text for decision support systems. *Decision Support Systems* 85 (2016), 12–22.
[10] Lei Hou and Li. 2015. Newsminer: multifaceted news analysis for event search. *Knowledge-Based Systems* 76 (2015), 17–29.
[11] Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. 2014. Event registry: learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web.* ACM, 107–110.
[12] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5, 1 (2012), 1–167.
[13] Ricardo M Marcacini, Julio C Carnevali, and João Domingos. 2016. On combining Websensors and DTW distance for kNN Time Series Forecasting. In *Pattern Recognition (ICPR), 2016 23rd International Conference on.* IEEE, 2521–2525.
[14] Ricardo M Marcacini, Geraldo N Corrêa, and Solange O Rezende. 2012. An active learning approach to frequent itemset-based text clustering. In *Pattern Recognition (ICPR), 2012 21st International Conference on.* IEEE, 3529–3532.
[15] Ricardo M. Marcacini, Marcos Aurélio Domingues, Eduardo R Hruschka, and Solange O. Rezende. 2014. Privileged information for hierarchical document clustering: a metric learning approach. In *22nd International Conference on Pattern Recognition (ICPR).* IEEE, 3636–3641.
[16] Nathan Marz and James Warren. 2015. *Big Data: Principles and best practices of scalable realtime data systems.* Manning Publications Co.
[17] Gautam Mitra and Leela Mitra. 2011. *The handbook of news analytics in finance.* Vol. 596. John Wiley & Sons.
[18] Meinard Muller. 2007. Dynamic Time Warping. In *Information Retrieval.* Springer Berlin Heidelberg, 69–84. https://doi.org/10.1007/978-3-540-74048-3_4
[19] Bruno M. Nogueira, Yuri K. Tomas, and Ricardo M. Marcacini. 2017. Integrating distance metric learning and cluster-level constraints in semi-supervised clustering. In *2017 International Joint Conference on Neural Networks (IJCNN).* IEEE, 4118–4125.
[20] Kira Radinsky and Eric Horvitz. 2013. Mining the web to predict future events. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining.* ACM, 255–264.
[21] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web.* ACM, 851–860.
[22] Diego P. Silva, Luis R. E. Jesus, Solange O. Rezende, and Ricardo M. Marcacini. 2014. Unsupervised learning of websensors for automatic extraction of descriptors on pulp production in Brazil (In Portuguese). In *SIICUSP-USP.*
[23] Hristo Tanev, Jakub Piskorski, and Martin Atkinson. 2008. Real-time news event extraction for global crisis monitoring. In *International Conference on Application of Natural Language to Information Systems.* Springer, 207–218.
[24] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. 2016. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann.
[25] Diyi Yang, Tanmay Sinha, David Adamson, and Carolyn Penstein Rosé. 2013. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the NIPS Data-driven education workshop*, Vol. 11. 14.