A Method for Web Content Extraction and Analysis in the Tourism Domain

Ermelinda Oro[1,2] and Massimo Ruffolo[1,2]
[1]*National Research Council (CNR), Via P. Bucci 41/C, 87036, Rende (CS), Italy*
[2]*Altilia srl, Vermicelli Square, Technest University of Calabria, 87036, Rende (CS), Italy*

Keywords: Big Data, Data Integration, Web Extraction, Text Analysis, Text Processing, Social Network, Sentiment Analysis, Knowledge Extraction, Smart Tourism, Big Sport Event, Tennis Italian Open.

Abstract: Big data generated across the web is assuming growing importance in producing insights useful to understand real-world phenomena and to make smarter decisions. The tourism is one of the leading growth sectors, therefore, methods and technologies that simplify and empower web contents gathering, processing, and analysis are becoming more and more important in this application area. In this paper, we present a web content analytics method that automates and simplifies content extraction and acquisition from many different web sources, like newspapers and social networks, accelerate content cleaning, analysis, and annotation, makes faster insights generation by visual exploration of analysis results. We, also, describe an application to a real-world use case regarding the analysis of the touristic impact of the Italian Open tennis tournament. Obtained results show that our method makes the analysis of news and social media posts more easy, agile, and effective.

1 INTRODUCTION

Big data generated across the web has created numerous opportunities to bring more insights useful to make smart decisions. Therefore, technologies and approaches aimed at gathering and analyzing huge amount of contents from the Web are important in many application fields.

The tourism is one of the leading growth sectors. Many factors influence tourism growth, and one of the more considerable contribution comes from the big events, like sport tournaments. Smart tourism applications require the capability to analyze the impact of such events to understand which are weakness and strengths of the touristic offer. News published in on-line media and comments of social networks can be valuable sources that provide useful data for analyzing tourism events. The insight deduced from such an analysis is of paramount importance for smart tourism, not only for the event itself or to develop marketing strategies, but also to improve the social and economic context and, in particular to improve the knowledge for the hosting city needed to develop a smart city. For this reason, hosting a big event or more specifically big sport-event, is much appealing, but without a deep analysis a lot of opportunities can be lost. To the best of our knowledge, in literature methods for combining and analyzing heterogeneous big amounts of web contents in the tourism sector have not been defined.

This paper aims to design, propose and test a framework to facilitates the massive gathering, cleaning, and analysis of tourism related contents from different sources, such as social networks and newspapers. We describe a content analytics method based on the following main steps: (i) content extraction from web sources, (ii) content analysis and annotation, (iii) visual insight generation. The presented method is generic and applicable to other big data contexts in order to analyze unstructured data and enhance decision making capabilities. The main contribution of the paper is that we propose a method that makes content analytics: (i) easy, fluid, and applicable to many different data sources; (ii) accurate and meaningful because we adopt a sentiment analysis algorithm capable to extract sentiments and opinion targets, and make annotation and classification on the based of taxonomies and ontologies. This way, it is straightforward to get insights about specific aspects of city life and services impacting on tourism.

To show the effectiveness of the proposed method, we have applied it to the analysis of the Italian Open event, which is part of the nine major tennis tournaments in the world after those of the Grand Slam.

Oro, E. and Ruffolo, M.
A Method forWeb Content Extraction and Analysis in the Tourism Domain.
DOI: 10.5220/0006371103650370
In *Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS 2017) - Volume 1*, pages 365-370
ISBN: 978-989-758-247-9

This event has grown over the years. Most important worldwide tennis players take part to this tournament that attracts a lot of attention generating value for the city of Rome, which get many social and economic benefits from the event. We present preliminary results obtained by applying the proposed method to the evaluation of the impact of the Italian Open. In particular, we carry out an initial quantitative and qualitative analysis of texts related to the event coming from social networks and on-line media.

This article is organized as follows. Section 2 briefly describes the related work. Section 3 introduces the methodology and used tools. Section 4 presents our initial results on the Italian Open use case. Section 5 discusses and concludes the work.

2 RELATED WORK

As described in (Xiang and Fesenmaier, 2017), smart tourism is related to the capability of collecting enormous amounts of data, intelligently storing, processing, combining, and analyzing them and using big data to design tourism operations, services and business innovation. More in detail, there is a lot of work that highlights the power of analyzing sport event to support smart tourism (Marine-Roig and Clavé, 2015; Wang et al., 2013).

Different studies have been conducted to explore the impact of hosting large-scale sport tourism events focusing on mainly tangible out-comes (i.e., economic benefits) rather reputational impacts (Gibson, 1998). For example, authors in (Fourie and Santana-Gallego, 2011) have the objective to study the effect of mega-sport events on tourist arrivals by using standard datasets. Other works analyzes sentiments in social network during sport events. (Kim and Walker, 2012) examined residents' perceptions for FIFA World Cup, whereas (Thomaz et al., 2016) analyzes of Twitter content related to tourist services during the FIFA World Cup 2014. We observed that existing works lack of the capability to process and combine different types of sources by using a unique framework, and they are focused mainly on social networks. In addition, our application considers the Italian Open event, and to the best of our knowledge, there aren't existing papers analyzing the impact of tennis tournaments on tourism.

Big Data management arises with many challenges, such as difficulties in data capture (in particular, from unstructured sources), data analysis and data visualization. In (Chen and Zhang, 2014) a state-of-the-art of techniques and technologies recently adopted to deal with the Big Data problems is presented. A technology for independent reference architecture for big data systems is presented in (Pääkkönen and Pakkala, 2015). We have implemented the method described in this paper by using MANTRA Smart Data Platform described in (Oro and Ruffolo, 2015). It makes use of contextual workflows and apps to extract and manipulate data from documents. The main advantages of this platform, with respect other technologies, are the possibility to: (i) simply encapsulate our algorithms in apps of the platform and run them in a cloud based big data context, and (ii) create, in a visual way, complex applications by combining apps that embed our algorithms in contextual workflows.

To extract valuable insights from unstructured big data the capability to process natural language for recognizing interesting entities and extracting sentiments is needed. Sentiment analysis on social networks messages is attracting a lot of interest (Bifet and Frank, 2010). Walaa et al. (Medhat et al., 2014) provides a recent comprehensive overview of the sentiment analysis in text mining field. There are many tools for performing natural language processing in big data context, whether proprietary or open source (e.g., NLTK (Bird, 2006)). In this paper we use the MANTRA Language (Oro and Ruffolo, 2015). It provides a rule based approach that enables to combine knowedge representation by import and/or implement ad-hoc ontologies/taxonomies and to easily define domain specific rule sets for analyzing sentimens. Based on our knowledge, there are not other languages that have all these advantages at the same time.

3 METHOD

This section introduces a general data processing and analytics method suitable for collecting and analyzing data related to a tourism topics, like a sport event. The proposed method consists of three general stages described in detail in the following: (i) Web data extraction and preprocessing. (ii) Content analysis and annotation. (iii) Visual insight generation.

Data Collection and Preprocessing. Once the input web sites and social network has been chosen, we apply automatic extractor algorithms in order to transform unstructured data in structured form.

The *News Extractor* algorithm takes as input the home page URL of online newspapers and recognizes links of articles in the page. To recognize links of articles contained in the home page of newspaper, we use an heuristic that exploits both visual/spatial and DOM features of titles of articles abstracts, e.g. fonts, colors, header tags, anchor link attributes. They normally have similar structure and position with respect to other elements in the whole page layout. For example, the algorithm checks the existence of the right context in the expected position, i.e. images, abstract text, and other links normally are below of the title of article.

The *Article Extractor* algorithm exploits an heuristic, similar to the one use in the New Extractor algorithm, to identify and extract (in a structured format) the clean content, title, images, and date of each linked article. More in detail, the Article Extractor algorithm clean HTML tags and recognize the real content related to the news deleting noises, such as navigation menus, advertisements, other links, etc. It exploits the centrality of main text, the emphasis of the title and the existence of main images.

For each analyzed website are extracted only news that speak about the target topic. Such a selection is performed by searching terms and their synonyms that the article should contain. We design simple patterns based to regular expressions to recognize such target terms. We intent to apply deeper semantic methods in the future. For instance, in the case study related to the analysis of the Italian Open event we selected interesting news by using and combining simple terms like ”IBI”, ”Master (s)? 1000 (a)? Rome”, ”Foro Italico”, ”International (BNL)?”, ”WTA (Primier|Master)? (5)? (of)? Rome”.

In order to extract information from social network we exploit their API, such as the Twitter’s Search API. For instance, to extract posts and comments in the official Facebook page of the IBI we used the social network’s API. Different features can be extracted, e.g,: message text, shared links and images, number of received likes, number of shares of the post, relationship between messages and comments.

All extracted data are stored in a Big Data store. In our implementation we used Hadoop.

Content Analysis and Annotation. The scope of this step is to transform flat text into meaningful information and knowledge. We implemented a variety of algorithms to analyze content, opinions, and topic in social comments and in news.

Firstly, we represented a taxonomy of target entities and associated patterns to recognize such entities in the text by using the MANTRA Language (Oro and Ruffolo, 2015), which is a first order logic like language enabling to represent domain knowledge and extraction patterns by exploiting dictionaries, ontologies and built-in functions. The type and quality of information that the machine recognizes, are based on the modeled knowledge. We decided to define a custom domain taxonomy in order to have a simple knowledge representation focused on the specific applications. However, the MANTRA Language can import and exploit OWL ontologies. In the IBI case study, we represented players, sporting events, sports organizations, hashtags and emoticons (relevant for social networks). Fig. 1 shows a sample of the taxonomy applied to recognize concepts relevant for the Italian Open.

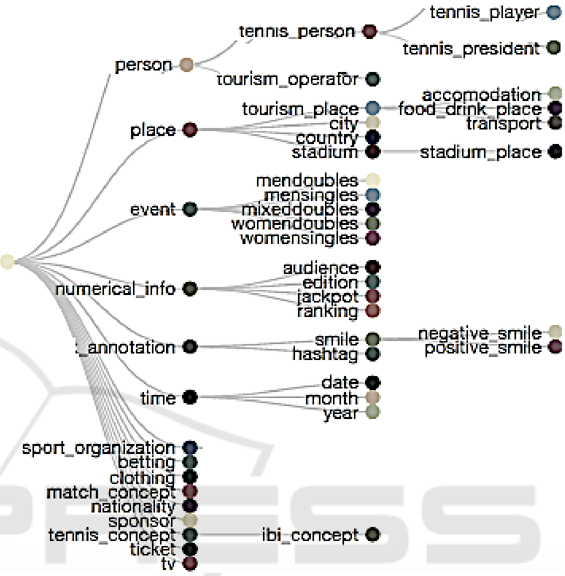

Figure 1: A sample of application taxonomy.

In addition to patterns written in MANTRA Language, in order to recognize topics in the flat text, we exploited a keyword-based approach based on the n-gram recognition and high-frequency terms (Chen et al., 2013).

Furthermore, we implemented the sentiment analyzer writing rules written in the MANTRA Language and by exploiting: (i) The sentiwordnet dictionary that label words in positive, negative or neural classes giving a certain polarity (Baccianella et al., 2010) (ii) Natural language like patterns to associate the polarity to sentences and to the whole input text. We recognize not just the positive or negative polarity. We also assign a polarity score that express the intensity of the sentiment. Furthermore, we compute the opinion-target (i.e the entity which the opinion is referred to). Polarity analysis can be helpful to get an idea of people’s reaction to various target entities, which can provide valuable insights. It is noteworthy that recognizing sentiment in social media is challenging because texts are written ignoring spelling and grammar rules, featuring lexical and syntactic problems such as slang, abbreviations, emoticons. Therefore, we found useful the possibility given by the MANTRA Language to define patterns that are not strictly based on natural language rules and that can exploit the emoticons to express mood.

Visual Insight Generation. This step consists in visual observation, evaluation, and interpretation of relationships, trends, and values get by analytical activities performed in previous steps. We adopted a data visualization approach to explore data. This enables a concise and powerful results exploration to data scientist and business users. Therefore, news and comments can be analyzed considering both quantitative information (e.g.,: number of news, various types of sources, different entities, etc.), and qualitative information to understand the type of shared contents by users, mentioned entities, opinions that derive from posts, comments and articles. Some examples of the performed analysis on IBI are shown in the following section.

To implement the proposed method, we used the MANTRA Smart Data Platform (Oro and Ruffolo, 2015), that combines semantic, big data, and cloud computing technologies. It enables the creation of Big Data Analytics applications by exploiting contextual worksflows and Apps. Apps are software modules that internally performs complex computational tasks like content extraction and analytics algorithms described above. The platform enables to combine Apps to create contextual workflow and manages the parallel, distributed excetution of workflows in the cloud. We chose MANTRA because, with respect to the tools available in literature, it consists in an holistic approach, that with his orchestrator simplify the connections of complex application encapsulated in Apps. The MANTRA Platform makes available two main GUIs that allow for implementing, in a visual way, the presented method: the MANTRA Workflow Modeler and the MANTRA Mashboard Modeler. The MANTRA Workflow Modeler (see Fig. 2) allows user to design workflows where each block-/step is a MANTRA App. The MANTRA Mashboard Modeler allows for creating dynamic dashboards in which the generated charts can be navigated and explored along different dimensions by combining and aggregating various types of collected data.

4 ANALYSIS RESULTS

In order to test proposed method, in this section we present its practical application to the extraction, acquisition, and analysis of contents about the Italian Open tennis tournament.

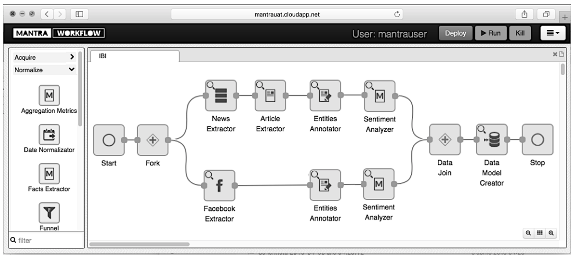

Figure 2: An Example of MANTRA Contextual Workflow. It enables to: (i) acquire clean articles from online newspapers, posts and comments from Facebook; (ii) extract entities; (iii) analyze sentiments; (iv) create the enriched dataset that can be visualized and explored by Mashboards.

Domain and Data Description. In the last decade, the Italian Open (also called IBI) recorded an extraordinary growth trend, with a higher increasing of sold tickets. But, such a measure is not the only one that should be considered. In fact, beside the significant growth in size (the public), IBI has grown also in media visibility, including social networks and online newspaper. In last years, IBI has been the main individual event for attractiveness in Rome, which annually hosts the event. Therefore, it is important to understand the reputational effects: knowledge about the high number of conversations that are generated online on IBI, and the quality in terms of content, emotions, and the different results based on the different involved media.

We considered articles of the most famous newspapers dedicated to the tennis topic, and posts/comments of the official Facebook page of IBI (Table 1) published during the years 2014 and 2015. To perform the analysis, we used the heterogeneous sources listed in Table 1 below. The dataset, created by automatically extracting contents from these sources, is publicly available[1].

Table 1: List of considered online sources.

Source	Website
FB	Internazionali-BNL-dItalia-267627913872
Internazionali	www.internazionalibnlditalia.com
Livetennis	www.livetennis.it
Ubitennis	www.ubitennis.com
Tennisitaliano	www.tennisitaliano.it
Spaziotennis	www.spaziotennis.com
Tennisbest	Tennisbest.com
Tennis	Tennis.it
Federtennis	www.federtennis.it

We apply the presented method to IBI related contents with the aim to answer the following questions: (i) When and how much do newspapers and social

[1] Datasets are available at: http://www.lindaoro.com/datasets/ibi.html

networks talk about Italian Open? (ii) What is the most discussed topic? (iii) What is the opinion about Italian Open? (iv) What are the most active users?

When and How Much do Newspapers and Social Networks Talk about Italian Open? Fig. 3 shows how the news that talk about the IBI are distributed over the time.

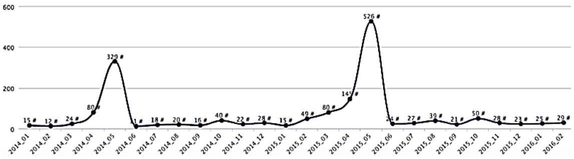

Figure 3: Distribution of IBI's news over the time.

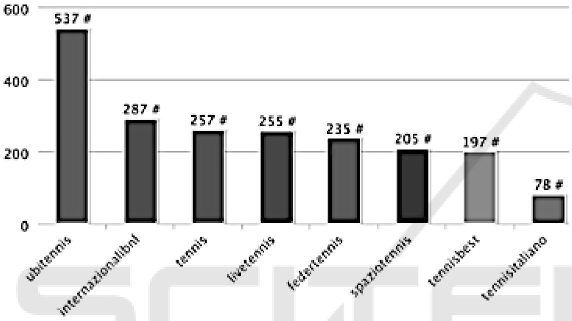

Figure 4: Number of IBI's news per newspaper.

The number of articles about IBI has been substantially increased in 2015 compared to 2014. Obviously, there is much talk of the IBI in May and in the months preceding and following the event. The histogram shown in Fig. 4 shows the websites that published the largest number of news related to the IBI domain. Likewise, analyzing Facebook fanpage we detected the increasing number of posts and comments, and also the numbers of users increased significantly over the time.

What is the Most Discussed Topic? To analyze the content of articles and posts/comments, we use the taxonomy that we have defined by using the MANTRA Language showed in Fig. 1. The news articles can be analyzed considering the title and the body of the article. As expected, the most frequent concept were event, the place where is played, and the players who disputed the matches, and in particular the most famous players in the international and national context. In Fig. 5 are shown the players that have been discussed more frequently in the analyzed years. In addition, the analysis has shown that newspapers discuss more about male events, whereas Facebook comments are often related to the female gender. Like in

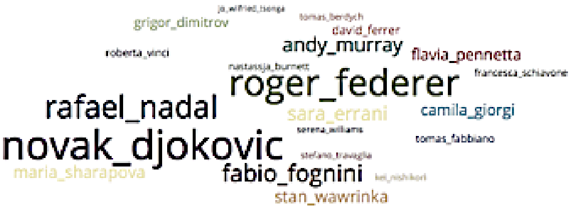

Figure 5: Tag cloud about the "tennis_player" concept.

Twitter, hashtags and smiles are often used in the social network. Most commonly used hashtags in order of frequency are: #tennis, #ibi16, #wta, #atp, #ibi14, #ibi15, #countdown, #federer, #ajdenole, #masha

What is the Opinion about Italian Open? In this subsection is shown the sentiment detected in the articles and in the post/comments. Our method is able to recognize, not just the positive or negative polarity, but also the intensity of the polarity and the opinion-target. Fig. 6 shows the percentage of articles that express positive, negative or neutral opinions. Normally, the whole news articles express positive sentiment. But, by analyzing individual sentences of news, negative statements are often identified.

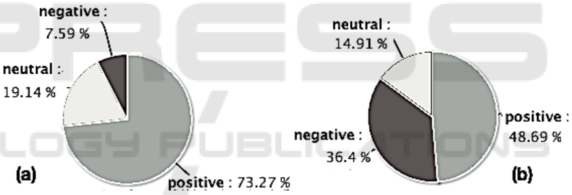

Figure 6: Percentage of negative, positive and neutral opinions expressed in the news considering (a) the whole article and (b) each single sentence.

Same analysis have been performed on Facebook posts and comments. We observed, that comments express more negative opinions than posts. This is because posts are mainly informative messages. Clicking on negative sentiment, every sentiment and related opinion-target, also associated to the modeled taxonomy of concepts, can be explored. For instance, by choosing negative sentiment, and selecting the opinion-target "Rome", we found that generally issues are about the inability to visit Rome. We deeply studied the difficulties in visiting Rome and we obtained main problems in public transports.

What are the Most Active Users? We extracted the most active users. We observed that the posts are essentially written by the International Tennis organizations, and only comments are written by Tennis fans. By graph analysis, we detected the influencers.

We used our method to collect, process, explore and monitor the topics of interest and the reputation of the Italian Open and its hosting city. Even in this early stage of the case study, our method enabled us to discover issues that touristic operators should resolve to reinforcing the ability of the city to receive tourists attracted from the tennis event.

5 CONCLUSION

In this paper we described a method for extracting and analyzing data useful for supporting strategic tourism management. The contribution of the proposed method is its ability to make content analytics processes, requiring data extraction, cleaning, and analysis, more easy and flexible. In particular, the method as been implemented by the MANTRA Smart Data Platfrom as a contextual processing workflow combining several algorithms that can be executed in parallel and disturbed way on massive contents. This makes also the method general and reusable in other areas. We described initial results deriving from the application of the method to the acquisition and analysis of contents related to the Italian Open sport event. The case study explains how tourism related content posted by users on newspaper websites and social media can be monitored and, consequently, be used as knowledge useful to improve competitiveness of tourism companies and services. Main insights obtained by analyzing contents about IBI have been the identification of difficulties in visiting Rome. Such a result, can suggest enhances that tourism organizations and destination cities can perform to promote smart tourism.

There are different suggestion for future research. About the technical point of view, we will focus on implementing and applying machine learning techniques to identify new concepts in (semi)automatic way. We will extract further information, like "annotated facts" that characterize the event (i.e. triples that describe subjects, objects, and related actions). A comparison between natural-language-based and neural-network-based methods will be performed. About the application case study, our goals will be an extension of considered sources and content extracted. We intent to perform a comparison with other events in order to identify the highest economic and reputational level that it can be reached. This comparison will allow the identification of factors that produce a better reputational impact.

REFERENCES

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204.

Bifet, A. and Frank, E. (2010). Sentiment knowledge discovery in twitter streaming data. In *Discovery Science*, pages 1–15. Springer.

Bird, S. (2006). Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics.

Chen, C. P. and Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, 275:314–347.

Chen, Y., Amiri, H., Li, Z., and Chua, T.-S. (2013). Emerging topic detection for organizations from microblogs. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 43–52. ACM.

Fourie, J. and Santana-Gallego, M. (2011). The impact of mega-sport events on tourist arrivals. *Tourism Management*, 32(6):1364–1370.

Gibson, H. J. (1998). Sport tourism: a critical analysis of research. *Sport management review*, 1(1):45–76.

Kim, W. and Walker, M. (2012). Measuring the social impacts associated with super bowl xliii: Preliminary development of a psychic income scale. *Sport Management Review*, 15(1):91–108.

Marine-Roig, E. and Clavé, S. A. (2015). Tourism analytics with massive user-generated content: A case study of barcelona. *Journal of Destination Marketing &amp; Management*, 4(3):162–172.

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113.

Oro, E. and Ruffolo, M. (2015). Using apps and rules in contextual workflows to semantically extract data from documents. In *roceedings of the 17th International Conference on Information Integration and Web-based Applications & Services*.

Pääkkönen, P. and Pakkala, D. (2015). Reference architecture and classification of technologies, products and services for big data systems. *Big Data Research*, 2(4):166–186.

Thomaz, G. M., Biz, A. A., Bettoni, E. M., Mendes-Filho, L., and Buhalis, D. (2016). Content mining framework in social media: A fifa world cup 2014 case analysis. *Information &amp; Management*.

Wang, D., Li, X. R., and Li, Y. (2013). China's "smart tourism destination" initiative: A taste of the service-dominant logic. *Journal of Destination Marketing &amp; Management*, 2(2):59–61.

Xiang, Z. and Fesenmaier, D. R. (2017). Big data analytics, tourism design and smart tourism. In *Analytics in Smart Tourism Design*, pages 299–307. Springer.