The Journal of Systems and Software 127 (2017) 237–248

Contents lists available at ScienceDirect

The Journal of Systems and Software

journal homepage: www.elsevier.com/locate/jss

Large scale opinion mining for social, news and blog data

Nikos Tsirakis[a,*], Vasilis Poulopoulos[a], Panagiotis Tsantilas[a], Iraklis Varlamis[b]

[a] *Palo LTD, Kokkoni Corinthias P/C 20002, Greece*
[b] *Dept. of Informatics and Telematics, Harokopio University of Athens, Omirou 9, Tavros, Greece*

ARTICLE INFO

Article history:
Received 24 October 2015
Revised 9 March 2016
Accepted 4 June 2016
Available online 6 June 2016

Keywords:
Opinion mining
News streams
Social media

ABSTRACT

Companies that collect and analyze data from social media, news and other data streams are faced with several challenges that concern storage and processing of huge amounts of data. When they want to serve the processed information to their customers and moreover, when they want to cover different information needs for each customer, they need solutions that process data in near real time in order to gain insights on the data in motion. The volume and volatility of opinionated data that is published in social media, in combination with the variety of data sources has created a demanding ecosystem for stream processing. Although, there are several solutions that can handle information of static nature and small volume quite efficiently, they usually do not scale up properly because of their high complexity. Moreover, such solutions have been designed to run once or to run in a fixed dataset and they are not sufficient for processing huge volumes of streamed data. To address this problem, a platform for real-time opinion mining is proposed. Based on prior research and real application services that have been developed, a new platform called "PaloPro" is presented to cover the needs for brand monitoring.

1. Introduction

The enormous advances in social media and their power to reflect and influence public opinion made them a domain of great interest for marketeers, communication specialists and companies that want to advertise their products and services, or simply want to boost and monitor their brand name. This resulted to large amounts of data, which are created daily to various social media and the news and contain mentions to products and companies. Data can be in different formats (textual, audiovisual etc), can be written in a formal (e.g. product reviews) or informal way (e.g. comments), can be objective mentions or subjective opinions about the company or product or an aspect of it (Thet et al., 2010; Pontiki et al., 2014).

The volume and complexity of the data that can be acquired, stored and manipulated, have created a flood of data 90% of all data were generated in the last two years (SINTEF, 2013). For example, in Palo, the crawler module is able to collect a large number of data which is estimated at 1000 articles per minute during the rush hours leading to an increase of 2.5Gb per month of compressed data. The data in absolute numbers are more than 10 million records per month from all the possible data sources.

The large volume and volatility create big challenges for companies that provide social media analytics services and cope with data from multiple data streams. Big players from the Web and Databases domains invest in social media analytics with generic frameworks and platforms (e.g. IBM Social Media Analytics) that emphasize on the analytics part but do not focus on text mining, or with extensions of their existing platforms (e.g. Google news lab and Google Analytics) that incorporate content from specific social media using associate data hubs and plugins (e.g. Googles social data hub). They use Twitter, Facebook and other social media APIs to collect data in streams and provide commercial archives/feeds and associated analytics.

A big challenge for social media monitoring is to link data from various sources together, compare and integrate and bring everything in a common form for analysis and presentation. Data analysis is the next bottleneck since traditional algorithms lack of scalability and do not easily adapt to the complexity of data that needs to be analyzed. Finally, the presentation of the extracted knowledge must be carefully designed in order for the results to be self-interpreted by non-technical domain experts and assist them in getting valuable actionable knowledge.

* Corresponding author.
E-mail addresses: nt@paloservices.com (N. Tsirakis), pv@paloservices.com (V. Poulopoulos), pt@paloservices.com (P. Tsantilas), varlamis@hua.gr (I. Varlamis).

http://dx.doi.org/10.1016/j.jss.2016.06.012
0164-1212/© 2016 Elsevier Inc. All rights reserved.

Palo Ltd is a company specializing in information extraction from the web. It started by gathering content from news sites and blogs in Greece, about 5 years ago. The analysis was limited in clustering articles based on content similarity and presenting them in an aggregated form to the end users. Palo Pro is Palos social media analytics service, which was launched primarily in Greece but now expands to Serbia, Cyprus, Turkey and Romania. The service monitors and analyzes data from the web and social media, giving emphasis to entity extraction and sentiment analysis from text. In the same architecture, several modules for crawling, feed aggregation, text clustering, multi-document summarization, Named Entity Recognition, aspect extraction and opinion mining synthesize the ecosystem of Palo Services.

Palo Pro can be described as a business intelligence platform with social basis (social business intelligence) (Dinter and Lorenz, 2012) that takes advantage of the knowledge of the crowd (crowd sourcing) as expressed in social media. The benefit is both for companies, which are able to monitor the popularity of their products and for buyers who receive long-term improved services and products. The interest for such a platform is increased, for example the mobile phone industry in Greece numbers 13 million active subscribers, who are active on the internet, and comment the products of the three main competitors (packages, special offers, etc.). Although information about the popularity of each of the three partners has little value, knowledge about the course of their products in social media and the opinion formed from every new movement is valuable for any further advertisement campaign.

From the data stream management point of view, several research issues emerge, like: (a) approximate query processing techniques to evaluate slow and memory demanding queries, (b) sliding window query processing, (c) data sampling to handle an increased flow rate of the input stream etc. In this paper we present how the proposed platform handles such research issues, using real-time data filtering in the source, summarization of historical content and statistics computation over sliding windows. We also discuss additional technical challenges, that are not only data stream specific, are confronted, such as heterogeneity of data, multilingual content and scalability of the existing solution.

In the following section, we provide an overview of Palo Pro service and the infrastructure that supports them. In Section 3 we discuss the processing pipeline in more details and in Section 4 we summarize the open issues concerning the processing of big data in a real-time environment.

2. Background

2.1. Scientific background

Opinion mining is defined as the task of classifying texts into categories depending on whether they express positive or negative sentiment, or whether they enclose no emotion at all. Sentiment polarity was extracted using emotion dictionaries (comprising mostly adjectives) (Hatzivassiloglou and McKeown, 1997; Qiu et al., 2009), statistical techniques based on co-occurrence of head terms and modifiers, classification techniques such as SVM, Naïve Bayes, Maximum Entropy etc. Pang and Lee (2005) and in some cases, semantic and syntactic analyzers (Yi et al., 2003; Miyoshi and Nakagami, 2007). The topic has attracted considerable attention in recent years due to its direct applicability in real-world businesses, such as brand monitoring, marketing or prediction of election results.

The concept of ***aspect based opinion mining*** (opinion mining for different aspects of an entity) appeared in literature in 2009. Early research focused on multi-aspect entities such as movies (Thet et al., 2010) – and the opinions provided by the viewers comments for the different aspects that make up the final result (actors, director, screenplay, music, etc.) – electronic devices (Hu and Liu, 2004) and hotels (Blair-Goldensohn et al., 2008). The main principles behind these works were: (a) extracting opinions or emotions and (b) labeling entities and their aspects (head terms) and the words that convey emotion (modifiers). However, it was until 2012, that the work of Moghaddam and Ester (2012) gave a new impetus to opinion mining on individual properties of commercial products from customer overviews. A typical example of aspect based opinion mining is the evaluation of a photo camera, where users evaluate separately the ease of use, the image quality, the shutter lag and battery duration, and behind the comments attached a positive or negative score for each aspect. This approach gives a new dimension to the problem of extracting knowledge from texts adding additional granularity levels in opinion or sentiment expressed in a text. The fact that these opinions mining techniques were applied to commercial products has increased the interest of marketers and brand makers who want to handle the image of a product or company on the market and understand the preferences of potential customers. The immediate consequence to the increase of the power of comments and opinions to the commercial products is the appearance of malicious comments (spam) with positive or negative orientation that aim to alter the real image of a product (Mukherjee et al., 2012; Jindal and Liu, 2008).

Summarization is another challenging task for social media content analysts. There are many research works that focus on the summarization (Zubiaga et al., 2012) or visualization (Hao et al., 2011) of Twitter streams. The analysis focuses on a specific entity (Hao et al., 2011) or event (Zubiaga et al., 2012) of the complete Twitter stream, so assumes a filtering step in the beginning of the pipeline. Although the volume of data for an entity or event in the unit of time in Twitter is not so impressive when compared to video streams (e.g. in Hao et al., 2011 authors report about 60,000 tweets for a movie in a five days period), the processing and visualization arise several challenges for system developers. In the case of text stream analytics it is important to summarize content and export useful features in a streamed manner, but also to keep the actual data in a separate storage for further analysis (e.g. drill through analytics, post event analytics). In addition to this, the filtering step may produce several different substreams that refer to different entities or events. In the case of PaloPro for example, social media stream monitoring is available for different countries and monitors different entities in parallel. In addition to this, since PaloPro also combines social media analytics with news analytics, it collects information from data sources around the world. An aggregation module is responsible for this, and its output is also redirected to the filtering module.

2.2. Competitive systems

The big interest of businesses is reflected to a number of commercial tools that provide analysis and monitoring services of the markets. The Blogmeter[1] is one such product that has been developed by the Italian company CELI and adapted for specific markets (telephony, food, fashion, etc.). It offers tools for monitoring and reporting the image of companies, products or services to social media such as Facebook, Twitter, Google +, Pinterest etc. However, the system requires that the company has a profile in the respective social media and focuses only on analyzing the information posted on the respective websites of the companies in each medium (e.g. likes, the followers, the retweets, etc. at pins. each company). The SentiMeter[2] is a tool that gathers data from Twitter, Facebook, YouTube, Google+, Digg, Blogger, Tumblr and other

[1] http://www.blogmeter.eu.

[2] https://sentimeter.com/.