Computer Standards & Interfaces 46 (2016) 79–87

Contents lists available at ScienceDirect

Computer Standards & Interfaces

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

Harvesting Big Data in social science: A methodological approach for collecting online user-generated content

M. Olmedilla [a,1], M.R. Martínez-Torres [a,*], S.L. Toral [b,2]

[a] Facultad de Turismo y Finanzas, University of Seville, Avda. San Francisco Javier s/n, 41018 Seville, Spain
[b] E. S. Ingenieros, University of Seville, Avda. Camino de los Descubrimientos s/n, 41092 Seville, Spain

ARTICLE INFO

Article history:
Received 2 December 2015
Received in revised form 2 February 2016
Accepted 2 February 2016
Available online 10 February 2016

Keywords:
Big Data
User-generated content
e-Social science
Computing
Data gathering

ABSTRACT

Online user-generated content is playing a progressively important role as information source for social scientists seeking for digging out value. Advances procedures and technologies to enable the capture, storage, management, and analysis of the data make possible to exploit increasing amounts of data generated directly by users. In that regard, Big Data is gaining meaning into social science from quantitative datasets side, which differs from traditional social science where collecting data has always been hard, time consuming, and resource intensive. Hence, the emergent field of computational social science is broadening researchers' perspectives. However, it also requires a multidisciplinary approach involving several and different knowledge areas. This paper outlines an architectural framework and methodology to collect Big Data from an electronic Word-of-Mouth (eWOM) website containing user-generated content. Although the paper is written from the social science perspective, it must be also considered together with other complementary disciplines such as data accessing and computing.

1. Introduction

A better access to information is powering the interest in Big Data [1]. Over the next years, the increasing volume of data created and collected in the Internet is expected to persist [2]. However, most of the Big Data still remains wild and unstructured. In that regard, advanced computational techniques are exploiting the potential of technology to capture and analyse such big amounts of data from the Internet in increasingly powerful ways [3]. This is offering the humanistic and social science disciplines the possibility of making many social spaces quantifiable, so they can be studied following a quantitative approach [4]. Actually, the evolution in computer aided research methods is changing the way in which social science research and data processing is done [5]. In recent years a far wider range of social scientists has become more involved about the potential of Big Data, which is creating challenges and opportunities for interdisciplinary researchers [6]. For instance, in his article in Wired magazine [1], Anderson suggested that research methodologies in social science should not only be based on building theoretical models but also on having better data and using better analytical tools. The beginning of digital convergence in the social sciences is accelerating the way phenomena are studied [7]. Besides, the recent advancements in Big Data technologies such as software tools to gather the content of interest from user-generated data facilitate the paradigm change in the so-called *modern e-Science* [4].

In general, most of the researches focus just on the analysis or modelling step of the Big Data pipeline. While that step is essential, the other phases such as data gathering are at least as important [8].

The access to massive quantities of information produced by and about people requires the application of computer science techniques [9]. Tools such as APIs (Application Programming Interface) are frequently used to get access to different subsets of content from the public stream [4]. Although APIs facilitate the automatic extraction of content, they also have some limitations when accessing to some specific data required by researchers. Actually, APIs only facilitate the information decided by the API provider. Thus, extracting meaningful information from these large-scale data repositories is still a challenging problem [10]. Whenever researchers are interested in accessing data beyond information provided by APIs, an effective in-situ processing has to be designed [8], like for example web crawlers. In that regard scientists have begun to develop web services with interfaces to collectors of Big Data sets such as Milne and Witten for Wikipedia [11], and Reips and Garaizar for Twitter [12].

In accordance with the idea developed by the aforementioned authors, who apply a methodology to collect of Big Data from different webs, this paper focuses on the computational challenges faced by social science in dealing Big Data gathering. Hence, an architectural framework and methodology to collect Big Data from a web that has user-generated content are defined. For this purpose, the paper is focused on Ciao, one of the world's largest eWOM (electronic Word-of-Mouth) communities. The rest of the paper is organized as follows.

* Corresponding author. Tel.: +34 954 55 43 10.
E-mail addresses: mariaolmedilla@hotmail.com (M. Olmedilla), rmtorres@us.es (M.R. Martínez-Torres), storal@us.es (S.L. Toral).
[1] Tel.: +34 954 55 43 10.
[2] Tel.: +34 954 48 12 93; fax: +34 954 48 73 73.

http://dx.doi.org/10.1016/j.csi.2016.02.003
0920-5489/

The next section discusses the background and provides the rationale for this study by conducting a review on the Big Data in Social Science, user-generated content and the role of the social scientists within Big Data. Then, the methodology section presents the design of the research using the web crawling approach. The case study and results' section explain the process of data gathering within the eWOM portal Ciao UK, including some experimental results in terms of time, size and database design. Afterwards, discussions and implications as well as limitations of this study and plans for future research are discussed. Finally, the last section concludes the study.

2. Research background

New perspectives in social science are now pursuing developments in Big Data [13], which is nowadays available in an abundance that was never known before. For instance, the amount of data that is produced each day already exceeds 2.5 exabytes [14] and 90% of the data in the world today was produced within the past two years [15]. Besides and according to Fan and Bifet [16], Big Data is going to continue increasing over the years to come, and each data scientist will manage a greater amount of data every year. Thus, dimension of data *volume* might be the most self-evident characteristic. Nevertheless, Big Data is also described utilizing other dimensions such as *variety* and *velocity* with which data is produced and needs to be consumed [17,18]. Those 3 dimensions form the so-called *3V model*, which are attributed to the analyst Doug Laney [19]. Other dimensions of this model comprise further aspects of Big Data such as the *veracity* the data comes with [18] and the need to turn the processed and stored data into *value* [20,21].

Nevertheless, it is important to understand that Big Data in social science is about not only the created content nor its consumption. Actually, it is also about the capture, search, discovery, and analysis tools that help gaining insights from unstructured data. In that regard, this section of the paper is focused on the Big Data collection within social sciences gathered from the literature.

2.1. Collecting Big Data in social science

With the increased automation of data collection and analysis, handling the emergence of an era of Big Data is critical [4]. Likewise, selecting the content of interest from the huge and constantly expanding universe of user-generated data exhibit one of the most fundamental challenge for applications for data collection: to explore large volumes of data and extract useful information or knowledge for future actions [22]. When using appropriate instrumentation for data collection, it is possible to take advantage of the information that comes from user-generated content such as clickstreams, tweets, user opinions, auction bids, consumer choices or social network exchanges [6]. In numerous situations, the knowledge extraction process has to be very efficient and close to real time because storing all observed data is nearly unfeasible. Hence, for an intelligent system to handle such acquisition of Big Data the essential key is to provide a processing framework, which includes considerations on data accessing and computing, as well as algorithms that can extract knowledge. In addition to providing a variety of data analysis methods, such knowledge discovery must supply a means of storing and processing the data at all stages of the pipeline, meaning from initial ingest to serving results [23]. To achieve such goal an overview of crawlers (or spiders, robots, wanderers, etc.) for collecting and indexing all accessible web documents will be introduced and then, in the methodology section of the paper, the one used to extract the data within this paper will be discussed. Nevertheless, it has to be emphasized that, all this process of gathering Big Data cannot be effectively understood from the unique disciplinary perspective of social science. Convergence among several disciplines to deal with the emergence of Big Data should be taken into account [6]. Furthermore, according to McCloskey [24], what gives accuracy to social scientists' work is not only rooted in all the way to data analysis and interpretation of the results but also in their systematic approach to data collection. To that end, a researcher can retrieve the data stored in the web through APIs provided by most social media services and largest media online retailers, which are not complicated to use. For example, the public API provided by Twitter to request specific information on the social network [25]. However, in many cases they do not provide all the data required by researchers. For instance, some additional features of users can be necessary to perform data cleaning and filtering operations, such as previous experience of users or their popularity or reputation. Such specific information is not usually available using APIs, and more computational specialized techniques are then necessary [26]. Therefore, collecting Big Data is a skill set generally restricted to those with a computational background. They use methods from the discipline of computer science such as web crawlers in order to capture the full potential of Big Data without any restriction.

The rapid growth of the web poses unprecedented scaling challenges for web crawlers, which seek out pages in order to obtain data. According to Najor [27], a web crawler is a "*programme that, given one or more seed URLs, downloads the web pages associated with these URLs, extracts any hyperlinks contained in them, and recursively continues to download the web pages identified by these hyperlinks*". Several crawling systems and architecture have been described in the literature. For instance, Chakrabarti et al. [28] and Seyfi et al. [29] describe in their papers a focused crawler and briefly outline its basic process, which seeks, acquires, indexes, and maintains pages that represent a narrow segment of the web rather than crawling the entire web. Equally, well established is the principle of operation of web crawlers stated by Cothey [30]. The author presents an experiment that examines the reliability of web crawling as a data collection technique. Prior to these authors, Pinkerton [31] describes the architecture of the web crawler and some of the trade-offs made in its design. The author specifies three actions performed by a crawler: (1) marking the document as retrieved, (2) deciphering any outbound links and (3) indexing the content of the document. Additionally, it is important to highlight that given space limitations in dealing with extremely large datasets extracted from crawling the web – especially when working with a very large and diverse information collection – there seems to be a fundamental to create a database in order to have an organized collection of data.

2.2. User-generated content in Internet

Social science has been traditionally handling collection of data in passive observation or active experiments, which aim to verify one or another scientific hypothesis [5]. On that subject, it is still a common practise in social science to develop further survey models to collect data sets directly from the users. Contrariwise, the public is increasingly choosing not to respond to surveys [32,33]. Besides, advances in data collecting technologies and data storage make it possible to obtain and preserve massive data generated directly or indirectly by users in Internet to generate valuable new insights [34]. In the same way, with the emerging capabilities to collect data sets from diverse real world contexts, Internet has become the researcher's new behavioural research lab [6]. Especially during the last years the rapid expansion of social networking applications, such as Facebook or Twitter have allowed users to generate content freely and amplify the already massive web volume of data [16]. In that regard, among the current literature there are several studies and projects in which user-generated online content have been used to carry out analysis in social sciences. For instance, Antenucci et al. [35] from the University of Michigan used Twitter data to create three job-related indexes for the US economy: job loss, job search and job posting. Likewise, in [36] the authors focused on tweets about unemployment to demonstrate how social media activity relates to the socio-economic situation across Spanish regions. In the financial field, also a growing number of papers are investigating whether the data coming from online social networks can help to improve the prediction of financial variables such as the study conducted in [37].