Harvesting Big Data in Social Science: A methodological approach for collecting online user-generated content

M. Olmedilla
Facultad de Turismo y Finanzas, University of Seville
Avda. San Francisco Javier s/n, 41018 Seville, SPAIN
Telephone: +34 954 55 43 10
E-mail: mariaolmedilla@hotmail.com

M.R. Martínez-Torres*
Facultad de Turismo y Finanzas, University of Seville
Avda. San Francisco Javier s/n, 41018 Seville, SPAIN
Telephone: +34 954 55 43 10
E-mail: rmtorres@us.es
*Corresponding author

S.L. Toral
E. S. Ingenieros, University of Seville
Avda. Camino de los Descubrimientos s/n, 41092, Seville SPAIN
Telephone: +34 954 48 12 93 Fax: +34 954 48 73 73
E-mail: storal@us.es

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.

Keywords: Big Data; user-generated content; e-social science; computing; data gathering

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 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 have 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].

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 to 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 is 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. 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 explains 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 per cent 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 exhibits 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 *"program 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 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 common practice 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].

Significant challenges are present in the search to capture the full potential of user-generated content. Whereas several data collection and analysis tools have become accessible on the web (e.g. APIs and crawling methods from Twitter or Facebook) during last few years, still, they are quite limited [26]. In that respect, Jagadish et al. [8] emphasizes that a good resource to avoid those tools limitations is the development of data cleaning techniques during the gathering. Hence, as Chang et al. [6] observe, researchers should deliberate how the tools participate toward gathering and extracting maximal value from data. Web crawlers facilitate this process. While some researchers relay on APIs and other tools to retrieve user-generated content, web crawlers do a better an exhaustive harvesting as well as they are more topic-specific [38]. This is because web crawlers can also extract specific content information while browsing the target website.

2.3 The role of Social Scientists within Big Data

The phenomenon of Big Data is closely bound to the appearance of *data science* or the so-called *computational organization science*, a discipline that combines mathematics, programming and scientific instinct. Such discipline has widened researchers' perspectives on social systems, by embracing computational models that combine social science and computer science [13][8][39]. Besides, according to Manovich [26] this combination of data abundance and the appearance of computational data analysis have shaped three kinds of divisions among people, which are the so-called *"new data-classes of the big data society":* people creating data, people with skills collecting data, and people with expertise analysing data. On this line of thinking, Davenport and Patil [40] present the role of the data scientist, who basically comprehends the skills of the three aforementioned new data-classes defined by Manovich [26]. The authors emphasize that not only are important the technologies for

taming Big Data but also are the people with the skills to put those technologies to good use and to retrieve meaning from the unstructured gathered information. Additionally, one important aspect of the data scientist is the ability to write code together with the ability of analysing large quantities of data. In this regard, Boyd and Crawford [4] state that, although computational scientists have started engaging in acts of social science, it does not mean that methodological issues are no longer relevant when dealing with Big Data. As said by the authors *"understanding sample, for example, is more important now than ever"*. Following from this conceptual framework, the paper focuses attention on defining a web crawler methodology to collect Big Data from a web with user-generated content.

3. Proposed Research Methodology

Following the research background on data gathering it is clear that user-generated data can be obtained with those APIs specialized on web crawling provided by most social media services and largest media online retailers such as YouTube, Flickr or even Amazon [26]. Those APIs provide an easy technique to obtain or scrape data. For example, through Amazon Web Services, developers can access product catalogue, customer reviews, site ranking and historical pricing. Nonetheless, they offer very poor functions such as the ones for search and acquisition since the content is produced without directly involving a person [4]. In some cases, it is interesting to obtain more information than the one provided by APIs, for instance to perform some filtering over the collected data of interest. For example, there are accounts that are actually bots, or even many users that are not active and can compromise the validity and reliability of the subsequent analysis. Those users cannot be removed unless some additional information is collected such as reputation, previous experience of users, etc. Moreover,

in some cases APIs only provide a fraction of all the available information worsening one of the benefits of social Big Data, which is the possibility of collecting the whole data instead of just a sample. This might be the case of Twitter APIs, which only makes available to typical researchers a roughly 10 per cent of public tweets [4]. Besides, it is important to mention that not all websites offer an API. Consequently, web scraping is a great alternative to grabbing the required data. So, within this section of the paper, a set of commands or methods – implemented by a researcher with a computational background – are explained in order to retrieve all the data stored in a web that contains user-generated content.

The methods applied to data collection have involved two different steps: (1) data crawling from a web with user-generated content and (2) data storage in a Data Base.

Firstly, to crawl data from the web *Python* was used because it is a dynamic, portable and performing language combined with an open source web crawler framework called *Scrapy*. Although there are simpler *Python* alternatives and other open source scrapers in *Java*, *Ruby*, and *PHP*, *Scrapy* is a much better alternative because it is the most popular tool for web crawling written in *Python*, as well as it is simple and powerful, with plenty of features and possible extensions [41]. The scraping cycle went through the definition of several *items*, which are containers defined to contain the data to be collected from the page. Then, to crawl or scrape information several classes named *spiders* where programmed. *Spiders* define how a certain site will be crawled, including how to perform the crawl and how to extract structured data from their pages. To that end, the *spiders* define an initial list of URLs to download, how to follow links, and how to parse (analyse) the contents of pages to extract *items*. Obviously there is a huge amount of data in webs with user-generated content and the *spiders* provide access to useful and relevant information with the goal of browsing as many web pages as

possible. Thus, the basic algorithm programmed here was fetching the web that contains all indexed pages that link all the information to gather. To that end, the method *parse_start_url()* was called. Such method contained a list of *URLs* where the spider began to crawl from. So, the first pages downloaded were those listed there and the subsequent *URLs* were generated successively from data contained in the start *URLs*. Then, different programmed *parse()* methods where in charge of processing the response and returning scraped data and more URLs to follow. Those methods returned *item* objects. The main advantage of web crawlers is that they can also extract specific information while browsing the site. This can be done using *XPath* language, which can be easily integrated within *Scrapy*. *XPath* is a language created for doing queries in *XML* content and it is used to turn an *XML* document into a hierarchical form to better organize information into a tree structure [42]. Those selectors are applied to part of the source code of the web and perform data extractions from the *HTML* source using expressions to navigate a document and extracting information using the library *Lxml*. Using *XPath*, researchers can select whatever content they consider meaningful in the context of their ongoing research, without the limitations of APIs, restricted to the information decided by the API provider. The following Figure illustrates some examples of how were programmed the methods *parse()* and also how the language *Xpath* was used.

[FIGURE 1]

Secondly, all of these steps have involved storing information in a database. Hence, the items retrieved from the spiders were persisted to a relational database due to the data-intensive storage and in order to have an organized collection of data. A relational database consists of one or more tables that have relationships, or links, between them, either in a one-to-one or a one-to-many relationship. The relational database systems

are generally efficient since different tables from which information has to be linked and extracted can be easily manipulated by querying data in the form in which it is desired. The database was designed in *MySQL* because it is efficient, ubiquitous and has an open-source engine available for all major platforms [43]. Several tables were created containing all dataset consisting of meta-information about user-generated data. Besides and in order to look at the data and analyse it in different ways, it is possible to apply the *SQL querying language*.

Finally, because the web is constantly changing and indexing is done periodically, proper data extraction also requires solid data validation and error recovery to handle data extraction failures as well as exceptions from the data storage. Thus, crawl monitoring and diagnostics dealing with exceptions were also created.

The following diagram in Figure 1 illustrates the afore-described process and the steps performed to collect and to store the data for this paper.

[FIGURE 2]

4. Case Study and Results

As aforementioned data collection has involved accessing data from the website Ciao, which is a mass eWOM community where registered users can make reviews about any product or service. Ciao is one of the largest eWOM communities in Europe available in local-language versions, with more than 1.3 million members that have written more than 7 million reviews on 1.4 million of products [44][45][44]. Basically, the website Ciao is structured in three main sections: reviews, shopping and "My Ciao" together with their corresponding subsections as illustrated in the following Table 1. The sections *reviews* and *shopping* are organized through categories of products and

services. Principally, there are 28 main categories established by Ciao as well as subcategories created by registered users whenever they post and share reviews about any product. The section *review* also contains all the reviews, video-reviews posted by the users and the questions with the concerning a user have about a specific review. The section shopping has the top-10 list of more sell and rated products and the top-seller charts. The web also contains a member webpage section named "My Ciao" for each registered user that assembles all of the information that is relevant to Ciao members. It not only contains community and system announcements alerting members to new features and possible scheduled site downtimes, but also many useful guidelines documents offering advice on how to write reviews and comments, give ratings and use the "circle of trust".

[TABLE 1]

Ciao is available free of charge to users, who can register on the web. In order to create an account the users have to provide some data about themselves (although it is not mandatory) such as first name, gender, age or country. Likewise, they have several scopes of activity as illustrated in Figure 2: (1) create a review, (2) rate a review, a product or another user for the benefit of other consumers and (3) trust other registered users. Firstly, when creating a review some fields to fill are required such as the title, the name of the product the review belongs to, the body with the user's opinion about the product or the advantages and disadvantages about the product reviewed. The standard review must be at least 120 words long and should aim to give readers an idea of what the product in question is like. Moreover, the users can only write reviews of products that are listed in the Ciao product categories. Secondly, a user can rate either a review, or a product or another user. To rate a review, the user has to choose one of the six options listed beneath the review to describe how useful he or she found it:

exceptional, very helpful, helpful, somewhat unhelpful, not helpful, off topic. Besides, the user can score products using qualitative ratings by giving it from 1 to 5 stars depending on how satisfied he or she is with it. Thirdly, the users are able to join their own *"circle of trust"* whenever they consider another registered user's reviews are consistently interesting and helpful. A user can invite up to 100 users to join his or her *circle of trust* but an unlimited number of users can choose to trust him or her. Any user who earns another user's trust is awarded community points for having done so, which influences the weighting given to his or her review ratings, and determines how visible the user is on the website.

[FIGURE 3]

Moreover, the registered user's activity is traced including a record of all of the actions that he or she has performed as shown in the following Table 2. Some examples, among others, are the date of his/her first/last review, how many reviews has he/she received from other users, the users included in his/her circle of trust, etc.

[TABLE 2]

Considering the content of the web and taking into account the data of interest with a focus on social science research the basic algorithm programmed to crawl the web was fetching the web that contained all indexed pages that link all users, reviews, products, ratings and circle of trust belonging to a category within Ciao. For such purpose data collection has required accessing to the Member Centre page of Ciao, where all the information about registered users is presented. Firstly, a list of users was collected in order to obtain the more important pages rapidly. To that end, a *spider* or crawler that follows the hyperlink structure of the users' webpages has been developed using *Scrapy* with *Python*. Principally, the crawler browses the website of the user,

storing a list of users containing their nick, id, name, gender, location or *URLs* among other data. This has provided a fast way of maintaining an index of the web through the id and *URLs* of the users that can be queried for updates. Secondly, a link was made from the list of users in order to collect the rest of the data (reviews, products, ratings, etc.). For this purpose several spiders were programmed using *XPath* language calling the function *response.xpath()* from the *Scrapy* base library. Using this language, it is possible to navigate through elements and attributes in the *XML* document of each webpage (see Figure **1**) and extract data using selectors that can include regular expressions. In this case most of programmed *XPath* selectors have selected the link that contained the text *'Next Page'* since the majority of the pages to extract were linked to other pages through this link. Finally and once all the spiders were programed, two functions for performance evaluation measures were created. The first function was used to capture errors and when a specific *URL* generated an exception, it was stored to an error table in the database. The other function was a list of *URLs* corresponding to the pages that were already downloaded with the spider without any error. To sum up the following Figure 3 structures an example of the afore-detailed process of data gathering.

[FIGURE 4]

Gathering big data from a web site can be a time-consuming task, so programmed algorithms should be fast enough to save time. For instance, within Ciao there is a huge number of information sources as well as different levels of accessibility to its user generated-content, which presents a complex information gathering control problem. Furthermore, the scale of the dataset is very large – there are about 45 thousand registered users in Ciao UK – which has meant the crawling procedure has taken relatively long time. Nevertheless, in spite of such a time consuming and complex

process the website was completely crawled. During this process, the downloading speed has fluctuated due to exceptions and power failures as illustrated in Table 3, which shows the duration of active data downloading of each spider represented in the first column. For instance, capturing the data from the ratings of a review has been done in three steps due to the appearance of exceptions and errors from things like validating the extracted data, removing duplicated items, storing in a database, etc. which has slowed down the gathering process. Another reason of such delay was that extracting data from the ratings required accessing content from not only a webpage but from six different pages (the six options of rating: *exceptional, very helpful, helpful, somewhat helpful, not helpful, off topic*) and implement their six link extractors in order to get to the pages that contain the useful information. Conversely, the amount of time that has taken the data gathering of the user's circle of trust was insignificant in comparison with the rest of the extractions. It was because there was only one link extractor for the spider, not all the users in Ciao have a circle of trust and storing only two fields in the database was fast.

[TABLE 3]

Ciao provides a good example in the context of a statistical analysis due to the variety of information and knowledge that contains. This data of course has to be processed, stored, analysed and visualized to have any meaning. Actually, the key of translation between gathered data and posterior structured data suitable for analytics lies on well-defined data characterisations (often in tables) stored in relational databases. Therefore and as aforementioned, a relational database was designed in this paper. It is composed of several tables and ordering schemes, which gives the possibility of tracking almost everything stored inside. As can be observed, the following relational model depicted

in Figure 4 illustrates the database model based on all the gathered data with a representation in terms of tuples, grouped into relations.

[FIGURE 5]

This model organizes data into eight tables representing each item: type of users, products, reviews, circle of trust among the users, ratings of the reviews, categories of products and two tables that store errors and exceptions resulting form the data gathering. The represented relations among the tables contain a unique key for each row. For example, the table that describes a *product* with columns for id of product *(unique key)*, name, category, rating, and so forth. Another table describes a *review*: id, title, body, rating of review, date, user nick, id of product *(foreign key)*, advantages, disadvantages, and so forth. Because each row in the table *product* has its own unique key (id of product), rows in the table *product* can be linked to rows in the table *review* by storing the unique key (id of product) of the row to which it should be linked, where such unique key is known as a "foreign key". Furthermore, with the *SQL querying language* applied to the database it is possible to retrieve the data based on specific criteria. For instance, if anybody would like to know how many users have rated another user the following query should be written:

```
select r.user_nick, rr.*
from review r join review rating rr on rr.id_review=r.id_review;
```

Additionally, the next Table 4 indicates the size of the complete database, which is 760 megabytes as well as the size in rows of all the data comprised in each table of the database. As can be observed the table *review rating* is the one with the higher number of ºrows since for each review all the users can rate up to 6 types of evaluations. Otherwise the table circle of trust has only 8.356 rows, which corresponds with the

same number of users. This means that not every user has to have “trusters” thus a circle of trust.

[TABLE 4]

5. Discussion and implications

After the analysis of above-described literature and the experience in designing the case of study, indubitable it is inferred that the field of social sciences naturally marry with the enthusiasm surrounding Big Data.

The success of Big Data, and more specifically user-generated content, is indisputably related to an intelligent management of data collection and selection, which drives to data quality. To that end, new technologies (e.g. APIs for web crawling) allow collecting big quantities of data. Nevertheless, data extracted through these APIs is not tailor-made, then it is worthless if most of the of gathered data is not produced by and about people or is a fake. Thus, as demonstrated in this paper through the development of a web crawling to gather user-generated content, high data quality harvesting requires computational skills. In that regard, the above literature also reveals the role of the data scientist stated by Davenport and Patil [40], who offer both perspectives: the one of computer scientist and the social scientist. Correspondingly, Boyd and Crawford [4] reveal that there is a tendency by computational scientists to engage in acts of social science.

5.1 Research contributions

Theoretical and practical implications of this paper contribute with a new approach on gathering user-generated data of a web from a social science perspective. Following from the above-described conceptual framework, several explanations for opportunities that Big Data offers to the field of social science can be traced to several authors. Boyd

and Crawford [4] outline Big Data as a socio-technical phenomenon and explain how to handle it through the use of data collection and analysis tools. The authors describe themselves as an example of social scientists working together with computer scientists and informatics experts, who focus on Big Data in social media context. Likewise, Chang et al. [6] present a comparison of examples gathered from the literature of the importance the role Big Data plays in the so-called computational social science research. They highlight the changes among emerging collection techniques for user-generated Big Data, which they associate to the research methods and the ways they can be applied.

While those authors contribute with good ideas related to the matter of this paper, here, the specific attention resides on defining a practical framework and methodology to collect Big Data from a web that has user-generated content. To that end, a web crawler has been implemented.

When referring to a methodology of crawling user-generated data from a web there is neither a specific approach nor a common agreement in the literature. In that regard, there are some authors in the literature describing the challenges and trade-offs inherent in web crawler design. For instance, Heydon and Najork [46], who describe a scalable, extensible web crawler written in *Java* or Shkapenyuk and Suel [47], who define the design and implementation of a distributed web crawler that runs on a network of workstations, and also Cho et al. [48], who outline importance metrics, ordering schemes, and performance evaluation measures for crawling a web. Likewise, the two authors mentions within the conceptual background of this paper, Michael and Miller [34] and Reips and Garaizar [12], have also described in their papers web collectors of Big Data for Wikipedia and Twitter respectively.

Although the authors' methods explain how to implement an effective web crawler

that identify important pages early and retrieve the pages' content, the aim of this is to provide a portrait of a new model and architecture for a web crawler in which traditional data retrieval methods are challenged. Additionally, this paper contributes with a methodology that is not only understood from the perspective of the social science discipline but also includes practices on data accessing and computing.

Another important aspect is that the web crawler described within this paper permits filtering and cleaning operations – such as users distinguished by sex, only users with some experience or reputation, etc. – before the data gathering instead of a standardized way of data collection made by APIs, just as identified by Manovich [26] within the research background. Besides, due the large amount of user-generated data retrieved, this can also reveal really interesting things about human cultural behaviour in general, e.g. user interactions, participation analysis, content analysis, analysis of topics, sentiment analysis, long-tailed phenomena, etc.

5.2 Limitations and future directions

The main limitation of the paper could be that the methodology has been implemented in just one eWOM community. Nevertheless, the architectural framework can be adapted to other eWOM websites since it is possible to follow the programming code pattern as long as the website is not full of *JavaScript*. Actually, that would be achievable by rewriting something similar but for the structure of the new eWOM website. In this regard, probable future research plans on gathering information to study other eWOM communities are focused on *TripAdvisor*. Such eWOM provides reviews of travel-related content and its structure has very similar characteristics to Ciao.

Other possible methodological limitation in the analysis of this paper would be the sample of the data set. Meaning that Ciao is not representing all the population but a

particular sub-set since members using Ciao are not representative of the global population. Besides, it is important to take into account that accounts and users within Ciao are not equal since users might have multiple accounts, whereas some accounts might be used by many people. Thus, some accounts might be 'bots' producing automatic content without the involvement of a person. Furthermore, not all information within Ciao is provided by the registered users. For instance, the 95,21 % of the users are sharing their age, unlike the 11,45 % who are sharing their real name or the 10,87 % who share their location. Nonetheless, is important to remember that this limitation is just personal data from a registered user, a trace about a user's activity and interactions (reviews, score, circle of trust, etc.) is always registered.

Since data is continuing increasing in more abundance than before, further research and potential areas for future work would be reducing the crawling speed as well as the response time to gather larger collections of data, which is also a major issue. The goal would be discover another crawling strategy, i.e., a strategy for determining which URLs to download next or to have a highly optimized system architecture that could download a large number of URLs per second while being robust against errors and programming exceptions. Consequently, a better data- processing, data-gathering and data-storing technology would be necessary.

6. Conclusions

This paper outlines an architectural framework to collect Big Data from the web Ciao UK that has user-generated content from the perspective of the social science. To that end, the previous sections explore a methodology that describes the implementation of a web crawler from other disciplinary perspective: the computing science discipline.

Given the volume of the user-generated content in the web and its speed of change, the coverage of modern web crawler is relatively small. In this regard, this paper presents two important highlights. First, the implementation of an effective web crawler that can gather and identify big amounts of user-generated content. Second, the stages followed on this crawling process, which are the identification and collection of important data, and the maintenance of the gathered data in a database by employing simple selection algorithms. These highlights suggest that, in general, Big Data needs to be the work of teams of social and computer scientists. Therefore, researches themselves should overcome this distance between technology and social sciences and develop both skills. Social science needs to develop adequate methodologies to deal with huge amounts of data, such as the one outlined within this paper.

Acknowledgment

This work was supported by the Consejería de Economía, Innovación, Ciencia y Empleo under the Research Project with reference P12-SEJ-328 and by the Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad under the Research Project with reference ECO2013-43856-R.

References

[1] C. Anderson, The end of the theory: the data deluge makes the scientific method obsolete?,*Edge*(2008)http://www.edge.org/3rd_culture/anderson08/anderson08_index.html (Accessed on 15 June 2015)

[2] S. Kaisler, F. Armour, J. A. Espinosa, W. Money, Big data: Issues and challenges moving forward, System Sciences (HICSS), 46th Hawaii International Conference, (2013) pp. 995-1004.

[3] R. Eynon, The rise of BigData: what does it mean for education, technology, and media research?, Learning, Media and Technology, vol. 38 no. 3 (2013) pp. 237-240.

[4] D. Boyd , K. Crawford, Critical questions for big data: Provocations for a cultural, technological and scholarly phenomenon, Information, Communication & Society, vol. 15 no. 5 (2012) pp. 662-679.

[5] Y. Demchenko, P. Grosso, C. De Laat, P. Membrey, Addressing big data issues in scientific data infrastructure, Edited by IEEE, Collaboration Technologies and Systems (CTS) International Conference, (2013) pp. 48-55.

[6] R. M. Chang, R.J. Kauffman, YO Kwon, Understanding the paradigm shift to computational social science in the presence of big data, Decision Support Systems, vol. 63 (2014) pp. 67-80.

[7] G. King, Restructuring the Social Sciences: Reflections from Harvard's Institute for Quantitative Social Science', Political Science & Politics, vol. 47 no.1 (2014) pp. 165-172.

[8] H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, C. Shahabi, Big data and its technical challenges, Communications of the ACM, vol. 57 no.7 (2014) pp.86-94.

[9] X. Liu, Y. Wang, D. Zhao, W. Zhang, L. Shi, Patching by automatically tending to hub nodes based on social trust. Computer Standards & Interfaces, vol. 44 (2015) pp. 94-101

[10] A. Cuzzocrea, I. Y. Song, K.C. Davis, Analytics over large-scale multidimensional data: the big data revolution!, Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP , (2011) pp. 101-104.

[11] D. Milne, I.H. Witten, An open-source toolkit for mining Wikipedia, Artificial Intelligence, vol. 194, (2013) pp. 222-239.

[12] U. D. Reips, P. Garaizar, Mining twitter: A source for psychological wisdom of the crowds, Behavior research methods, vol. 43 no.3 (2011) pp. 635-642.

[13] K. M. Carley, Computational organization science: A new frontier, Proceedings of the National Academy of Sciences, vol. 99 no. 3 (2002) pp. 7257-7262.

[14] A. McAfee, E. Brynjolfsson, T.H. Davenport, D.J. Patil, D. Barton, Big data:The management revolution, Harvard Business Review, vol. 90 no. 10 (2012) pp. 61-67.

[15] SINTEF, Big Data, for better or worse: 90% of world's data generated over last twoyears,ScienceDaily,(2013)www.sciencedaily.com/releases/2013/05/130522085217.htm (Accessed on 17 June 2015).

[16] W. Fan, A. Bifet, Mining big data: current status, and forecast to the future, ACM sIGKDD Explorations Newsletter, vol. 14 no. 2 (2013) pp. 1-5.

[17] M. Schroeck, R. Smart, D. Romero-Morales, P. Tufano, Analytics: The real-world use of big data: How innovative enterprises extract value from uncertain data, IBM Institute for Business Value (2012).

[18] G. Vossen, Big data as the new enabler in business and other intelligence, Vietnam Journal of Computer Science, vol. 1 no. 1 (2014) pp. 3-14.

[19] D. Laney, 3-D Data Management: Controlling Data Volume, Velocity and Variety, META Group Original Research Note (2001)

[20] G. Geethakumari, A. Srivatsava, Big Data Analysis for Implemen- tation of Enterprise Data Security, IRACST-International Journal of Computer Science and Information Technology & Security (IJCSITS), vol. 2 no. 4 (2012) pp. 742-746.

[21] P. Russom, Big data analytics, TDWI Best Practices Report, Fourth Quarter (2011).

[22] A. Rajaraman, J.D. Ullman, Mining of massive datasets, Cambridge University Press (2012).

[23] E. Begoli, J. Horey, Design principles for effective knowledge discovery from big data, Software Architecture (WICSA) and European Conference on Software Architecture (ECSA), (2012) pp. 215-218.

[24] D.N. McCloskey, From methodology to rhetoric, The Rhetoric of Economics (University of Wisconsin Press), (1985) pp. 20–35.

[25] A.R.M. Teutle, Twitter: Network properties analysis, Electronics, Communications and Computer (CONIELECOMP) (IEEE) (2010) pp. 180-186.

[26] L. Manovich, Trending: the promises and the challenges of big social data, Debates in the digital humanities, (2011) pp. 460-475.

[27] M. Najor, Web crawler architecture, Encyclopedia of Database Systems, Springer US, (2009) pp. 3462-3465.

[28] S. Chakrabarti, M. Van den Berg, B. Dom, Focused crawling: a new approach to topic-specific Web resource discovery, Computer Networks, vol. 31 no. 11 (1999) pp. 1623-1640.

[29] A. Seyfi, A. Patel, J.C. Júnior, Empirical evaluation of the link and content-based focused Treasure-Crawler. Computer Standards & Interfaces, vol. 44, (2016) pp. 54-62. doi:10.1016/j.csi.2015.09.007

[30] V. Cothey, Web-crawling reliability, Journal of the American Society for Information Science and Technology, vol. 55 no. 14 pp. (2004) pp. 1228-1238.

[31] B. Pinkerton, Finding what people want: Experiences with the WebCrawler, Proceedings of the Second International World Wide Web Conference, vol. 94 (1994) pp. 17-20.

[32] R. Curtin, S. Presser, E. Singer, Changes in telephone survey nonresponse over the past quarter century, Public opinion quarterly, vol. 69 no. 1 (2005) pp.87-98.

[33] E. D. De Leeuw, W. D. Heer, Trends in household survey nonresponse: A longitudinal and international comparison, Survey Nonresponse, (2002) pp.41-54.

[34] K. Michael, K. Miller, Big data: New opportunities and new challenges, Computer, vol. 46 no. 6 (2013) pp. 22-24.

[35] D. Antenucci, M. Cafarella, M. Levenstein, C. Ré, M.D. Shapiro, Using social media to measure labor market flows, National Bureau of Economic Research, no. w20010 (2014).

[36] A. Llorente, M. Cebrian, E. Moro, Social media fingerprints of unemployment, arXiv:1411.3140, (2014).

[37] M. Nardo, M. Petracco Giudici, M. Naltsidis, Walking down wall street with a tablet: a survey of stock market predictions using the web, Journal of Economic Surveys, 10.1111/joes.12102 (2015).

[38] G. Pant, P. Srinivasan, F. Menczer, Crawling the web, Web Dynamics, Springer Berlin Heidelberg, (2004) pp. 153-177.

[39] A. Syed, K. Gillela, C. Venugopal, The future revolution on big data, Future, vol. 2 no. 6 (2013)

[40] T. H. Davenport, D.J. Patil, Data scientist, Harvard business review, vol. 90 (2012) 70-76.

[41] J. Wang, Y. Guo, Scrapy-Based Crawling and User-Behavior Characteristics Analysis on Taobao, Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC) (IEEE) (2012) pp. 44-52.

[42] G. Gottlob, C. Koch, R. Pichler The complexity of XPath query evaluation, Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (2003).

[43] C. Vicknair, M. Macias, Z. Zhao, X. Nan, Y. Chen, D. Wilkins, A comparison of a graph database and a relational database: a data provenance perspective, Proceedings of the 48th annual Southeast regional conference (ACM), (2010) pp. 42.

[44] F.J. Arenas Márquez, M.R. Martínez-Torres, S.L. Toral Electronic word-of-mouth communities from the perspective of social network analysis, Technology Analysis & Strategic Management, vol. 26 no. 8 (2014) pp. 927-942.

[45] M. Olmedilla, M. R. Martinez-Torres, S. Toral, Examining the Power Law Distribution among eWOM communities: A characterization approach of the Long Tail, Technology Analysis & Strategic Management, (2015) doi: 10.1080/09537325.2015.1122187.

[46] A. Heydon, M. Najork, Mercator: A scalable, extensible web crawler, World Wide Web, vol. 2 no.4 (1999) pp. 219-229.

[47] V. Shkapenyuk, T. Suel, Design and implementation of a high-performance distributed web crawler, Data Engineering. 18th International Conference on Proceedings, (2002) pp. 357-368.

[48] J. Cho, H. Garcia-Molina, L. Page, Reprint of: Efficient crawling through URL ordering, Computer Networks, vol. 56 no. 18 (2012) pp. 3849-3858.

Tables

Table 1. 3-section structure of web Ciao UK

Reviews	Shopping	My Ciao
Product main category	*Product main category*	*User's webpage*
Product subcategory	*Product subcategory*	*Member Centre*
Latest reviews	*Top 10 list of products*	*User's announcements*
Latest questions	*Ciao topseller charts*	*User's guestbook*
Latest videos	*Recent products*	*User's statistics*

Table 2. Record of all the actions a user has performed

	Information about performed activities
Registered user	*Status (online/offline)*
	Date of since when is a member
	Date of the first review
	Date of the last review
	Reviews written
	Comments written of a review
	Comments received of a review
	Ratings given
	Ratings received
	Members who trust the user
	Members the user trusts
	Community score

Table 3. Time spent on data gathering

Data gathered	Start date	End date	Duration	% of time
users	6/4/15 - 14:15	6/4/15 - 20:42	6,45 hours	1,08 %
reviews, products and categories	7/4/15 - 20:20	8/4/15 - 1:04	4,73 hours	0,79 %
user's circle of trust	8/4/15 - 8:17	8/4/15 - 8:29	0,20 hours	0,03 %
review ratings 1	22/4/15 - 17:36	16/5/15 - 23:09	24 days and 5,55 hours	97,66 %
review ratings 2	17/5/15 - 21:37	18/5/15 - 0:02	2,42 hours	0,41 %
review ratings 3	18/5/15 - 10:05	18/5/15 - 10:13	0,13 hours	0,02 %

Table 4. Size of data comprised in the database

Stored data	Size	
Table "users"	44.352 rows	760 Megabytes
Table "reviews"	105.918 rows	
Table "products"	68.650 rows	
Table "categories"	283.240 rows	
Table "review ratings"	3.444.316 rows	
Table "circle of trust"	8.356 rows	

Figures

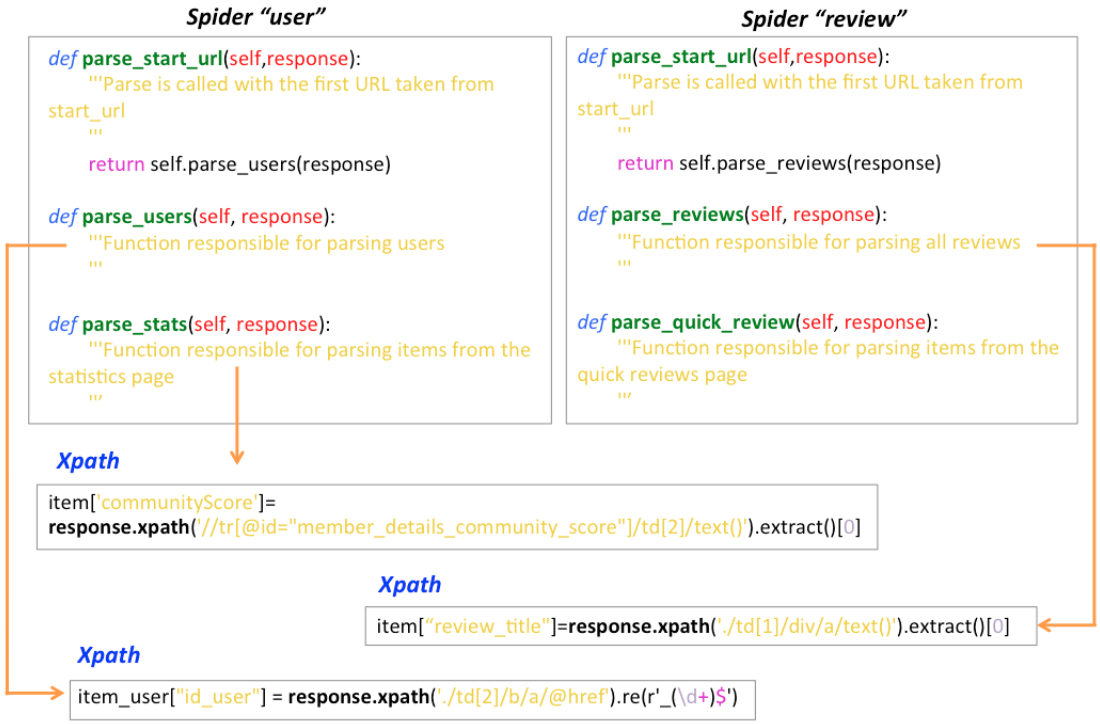

Figure 1- Programming examples of methods *parse()* and *response.xpath()*

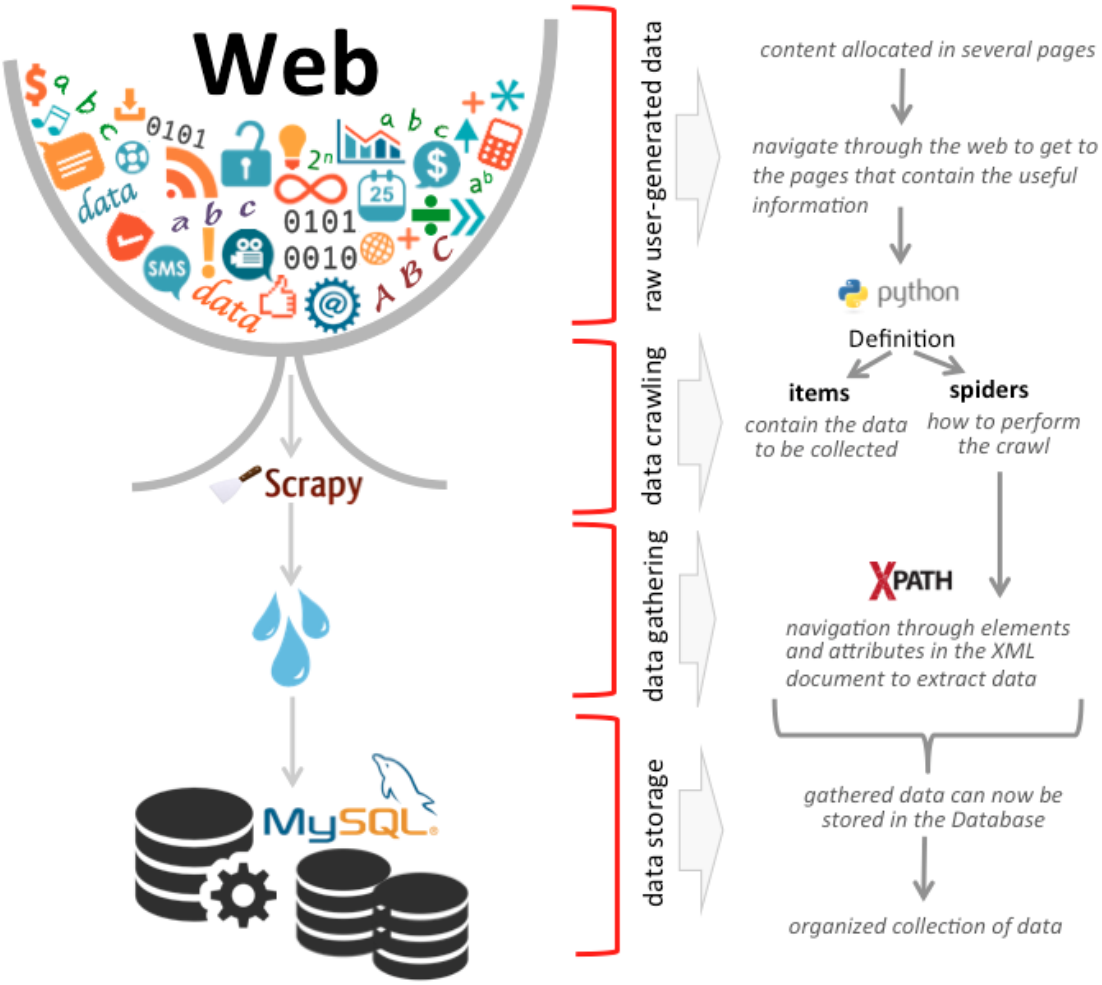

Figure 2. Process of collecting user-generated data from a web

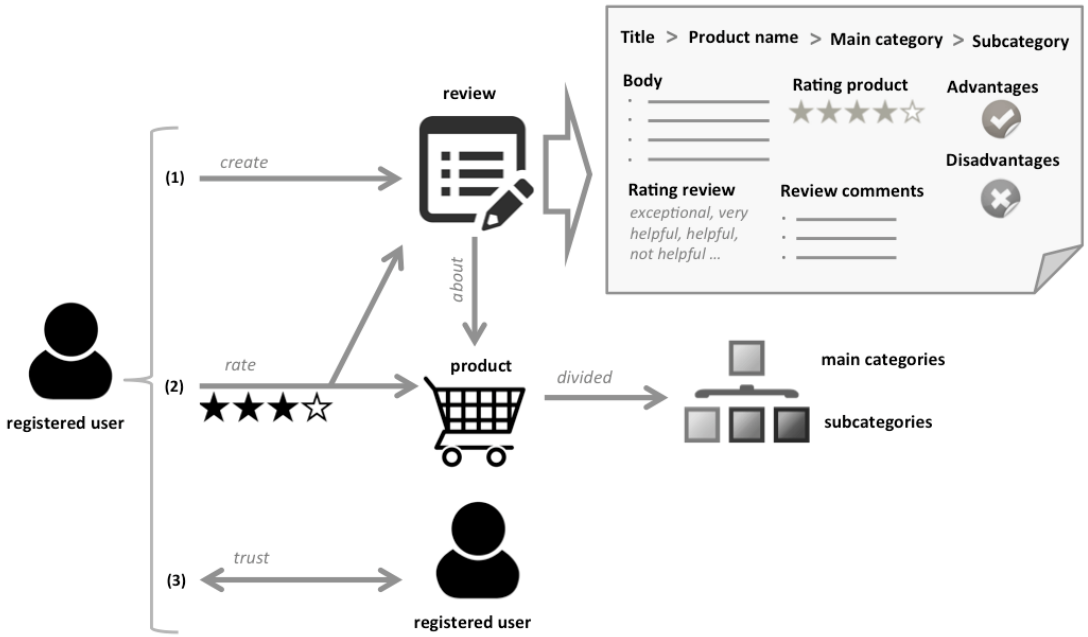

Figure 3. Scopes of activity within Ciao for a user

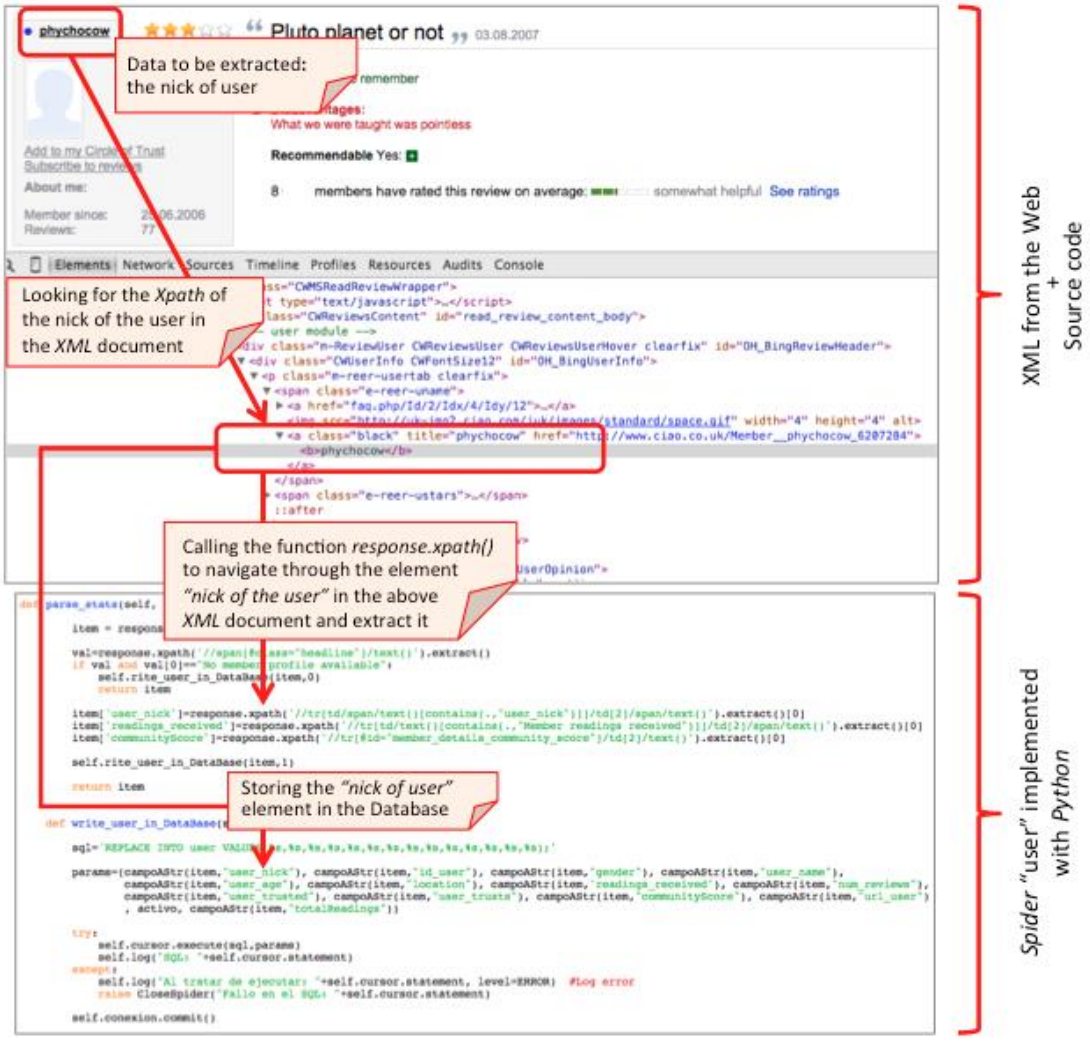

Figure 4. Structure of the process of data gathering *"nick of user"*

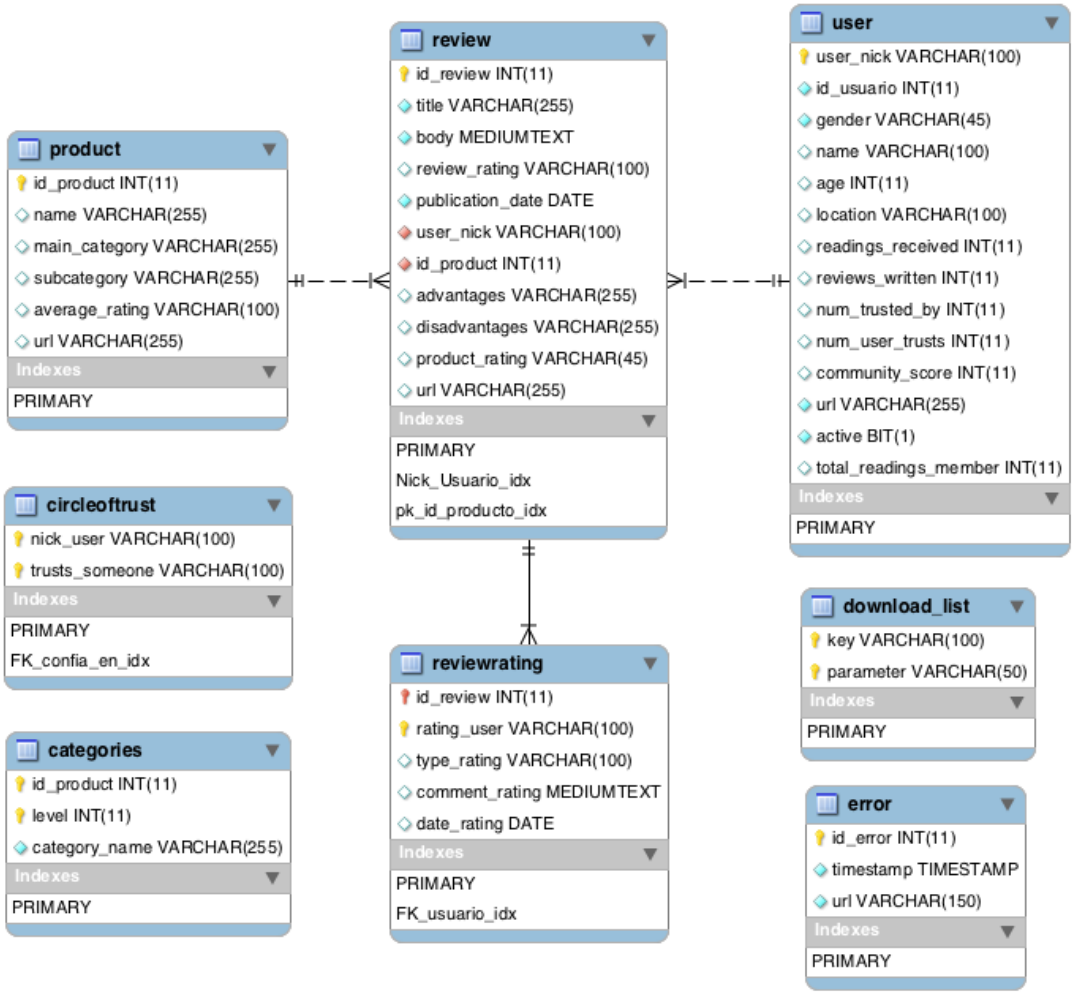

Figure 5. Relational model of the database