DANIEL JORGE RIBEIRO NUNES MARTINS

DATA EXTRACTION IN E-COMMERCE

UNIVERSIDADE DO ALGARVE
Instituto Superior de Engenharia
2016

DANIEL JORGE RIBEIRO NUNES MARTINS

DATA EXTRACTION IN E-COMMERCE

Mestrado em

ENGENHARIA ELÉTRICA E ELETRÓNICA

Área de Especialização em Tecnologias de Informação e Telecomunicações

Trabalho efetuado sobre a orientação de:

Prof. Doutor Pedro Cardoso
Prof. Roberto Lam

UNIVERSIDADE DO ALGARVE
Instituto Superior de Engenharia
2016

DATA EXTRACTION IN E-COMMERCE

DECLARAÇÃO DE AUTORIA DE TRABALHO

Declaro ser o autor deste trabalho, que é original e inédito. Autores e trabalhos consultados estão devidamente citados no texto e constam da listagem de referências incluída.

Resumo

A simplicidade do protocolo HTTP [19] e a extrema flexibilidade dos navegadores web (clientes HTTP) potenciaram o crescimento do número de sites e por sua vez o comércio eletrónico.

O comércio eletrónico, também conhecido como *e-commerce*, é um sistema que consiste na compra e venda de produtos ou serviços através da internet [22]. Sendo a internet um meio de comunicação utilizado por milhões de pessoas, a gestão da informação que é disponibilizada e a análise do mercado concorrente torna-se uma tarefa bastante árdua para quem gere um negócio de *e-commerce*. Para que os gestores se possam posicionar melhor perante os concorrentes surge a necessidade de criar mecanismos automáticos capazes de extrair informação das várias fontes web (websites).

A hotelaria é um mercado em que o *e-commerce* é imprescindível fazendo da internet o seu maior ponto de venda, seja através de canais de venda ou através dos seus próprios websites. Em simultâneo, os referidos canais apresentam informações importantes sobre a forma de comentários dos hóspedes, relativamente à reputação do hotel e seus concorrentes.

Existem dois métodos principais para a procura de informação na web [93], sendo esses: (a) a extração manual através de cópia e colagem e a (b) extração automática através de web robots.

Relativamente à extração manual, algumas empresas contratam pessoas para efe-

tuar a extração manual dos dados. Este método consiste em procurar pela web e copiar/colar ficheiros, reformatar texto, imagens, documentos, ficheiros multimédia e outros dados. Este método de extração de dados torna-se dispendioso, pois exige bastante tempo e mão de obra.

Por outro lado, para efetuar a extração de dados da web automaticamente, é necessário um crawler (web robot) para visitar as várias páginas web existentes, partindo de uma URL semente. À medida que estas URLs vão sendo visitadas pelo crawler, extraiem-se os dados da página HTML correspondente. Posteriormente por norma esses dados são armazenados numa base de dados, de forma a tornar o acesso aos dados mais eficiente.

Nesta dissertação é apresentada uma solução para alguns problemas apresentados, em que o principal foco é a extração automática de informação de quatro canais de venda de reservas de alojamento, sendo esses *Booking.com*, *Tripadvisor*, *Expedia* e *Bestday*. A informação que se pretende extrair tem como função auxiliar os gestores hoteleiros a analisar a disponibilidade de quartos, os preços praticados e a opinião dos hóspedes relativamente aos hotéis concorrentes. Essa informação será extraída com recurso a *web robots*, capazes de analisar *HTML* e interagir com as páginas web simulando o comportamento humano. Esta simulação de comportamento tira partido dos canais de venda seguirem um padrão de navegação de modo a que o utilizador siga facilmente os passos até efetuar a compra. Por cada um dos canais de venda que se pretende extrair informação foi criado um web robot diferente, pois as páginas web estão estruturadas de maneira diferente.

Descrevendo sucintamente o processo global, cada *web robot* começa por efetuar a pesquisa no formulário do respetivo website com um conjunto de parâmetros que são configuráveis. Após efetuar a pesquisa, são percorridos todos os hotéis que satisfizeram os critérios previamente definidos e de seguida é extraída a informação presente nos canais de venda, como sejam: os preços, as ofertas, os comentários e a localização do hotel. Esses dados são agrupados e armazenados numa base de dados não

relacional. Nesta fase os dados armazenados estão em bruto, i.e., sem qualquer tratamento.

Posteriormente, num processo independente (assíncrono), esses dados serão consolidados através de algumas regras previamente definidas de modo a eliminar redundância e a aumentar a consistência dos mesmos. Neste processo de consolidação existem várias preocupações, sendo possivelmente a principal a associação dos dados extraídos das diferentes páginas. Esta problemática surge devido à discrepância dos nomes dos hotéis nos diferentes canais de vendas. Além disso existem muitas outras discrepâncias entre os canais sendo as mais importantes: o número de estrelas das unidades hoteleiras, o nome dos quartos e a escala de pontuação dos hóspedes. Após concluído todo este processo de tratamento da informação, os dados são armazenados numa base de dados final. Ao contrário da base de dados usada na primeira fase, esta é uma base de dados relacional, o que significa que os dados estão devidamente estruturados possibilitando assim o uso por vários tipos de aplicações.

Depois de recolhidos e consolidados, a finalidade dos dados é serem: (a) Utilizados por modelos de previsão matemáticos que analisam os preços praticados pelos hotéis nos últimos anos e geram uma previsão de preços que os hotéis irão praticar no futuro, e (b) utilizados para verificar a reputação dos hotéis tendo em conta os comentários dos hóspedes.

Este trabalho não só apresenta a implementação dos web robots e da construção dos dados, como também uma vertente de análise da reputação dos hotéis através da análise dos comentários e pontuação dos hóspedes. A análise desses comentários e pontuações consiste em aplicar algumas regras de semântica e algumas métricas de modo a entender quais são os índices de satisfação dos hóspedes dos hotéis. Através destes indíces é possível verificar a importância de um hotel no mercado, pois num negócio são os clientes que definem o seu sucesso.

Esta dissertação apresenta um conjunto de quatro artigos resultantes em parte do trabalho desenvolvido pelo autor no projeto "SRM: Smart Revenue Management" fi-

nanciado pelo QREN I&DT, n.º 38962, promotor VISUALFORMA - Tecnologias de Informação, SA e co-promotor Universidade do Algarve. Abaixo segue-se a listagem dos artigos que compoem este trabalho:

- **Martins, D.**, Lam, R., Rodrigues, J.M.F., Cardoso, P.J.S., Serra, F. (2015) **A Web Crawler Framework for Revenue Management**, In Proc. 14th Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases (AIKED '15), in Advances in Electrical and Computer Engineering, Tenerife, Canary Islands, Spain, 10-12 Jan, pp. 88-97. ISBN: 978-1-61804-279-8.

- Ramos, C.M.Q., Correia, M.B., Rodrigues, J.M.F., **Martins, D.**, Serra, F. (2015) **Big Data Warehouse Framework for Smart Revenue Management**. In Proc. 3rd NAUN Int. Conf. on Management, Marketing, Tourism, Retail, Finance and Computer Applications (MATREFC '15), in Advances in Environmental Science and Energy Planning, Tenerife, Canary Islands, Spain, 10-12 Jan., pp. 13-22. ISBN: 978-1-61804-280-4.

- **Martins, D.**, Ramos, C.M.Q, Rodrigues, J.M.F., Cardoso, P.J.S., Lam, R., Serra, F. (2015) **Challenges in Building a Big Data Warehouse Applied to the Hotel Business Intelligence**, In Proc. 6th Int. Conf. on Applied Informatics and Computing Theory (AICT'15), in Recent Research in Applied Informatics, Salerno, Italy, 27-29 June, pp. 110-117. ISBN: 978-1-61804-313-9.

- Choupina, R., Correia, M.B., Ramos, C.M.Q, **Martins, D.**, Serra, F. (2015) **Guest Reputation Indexes to Analyze the Hotel's Online Reputation Using Data Extracted from OTAs**, in Proc. 6th Int. Conf. on Applied Informatics and Computing Theory (AICT'15), in Recent Research in Applied Informatics, Salerno, Italy, 27-29 June, pp. 50-59 ISBN: 978-1-61804-313-9.

Palavras Chave: Rastreador Web; Comércio Eletrónico; Reputação Online; Gestão de Receitas; Grande Volume de Dados; Armazém de Dados.

Abstract

Eletronic commerce, know as *e-commerce*, is a system that consists in buying and selling produtcs/services over the internet. The internet is used by millions of people, making the management of the available information (e.g. competitor analysis market) a very difficult task for those operationg an e-commerce business. So that the managers can better position their companies against competitors, comes the need to create automatic mechanisms to extract information from various web sources (websites).

The hotel business is a market where *e-commerce* is essential since the internet is their biggest selling point, either through sales channels or through their own websites. At the same time, these channels have important information, regarding the reputation of the hotel and their competitors, for instance in the form of guest comments.

In this thesis a solution to some of those problems is presented, in which the main focus is the automatic extraction of information from sales channels, such as *Booking.com*. The extracted information is used to help the hoteliers in the analysis of the prices and opinions of hotel's guests. That information will be extracted using *web robots*, able to analyze and interact with web pages, by simulating human behavior. This behavior simulation takes advantage of the navigation patterns present on most sales channels, so that users can easily follow the steps to the final purchase.

Briefly describing the overall process, the *web robot* begins by filling the web site search form with a set of configurable parameters. For each hotel that met the search criteria the most relevant information is extracted, such as: prices, offers, comments and location of the hotel. The collected data is grouped and stored in an intermediate database. Once collected, the data is: (a) used by mathematical prediction models that analyze the prices of the hotels in recent years and generate a forecast of prices that hotels will practice in the future and, (b) used to check the hotel's reputation taking into account the comments of the guests.

This thesis presents a set of four papers resulting in past from the author's work in project "SRM: Smart Revenue Management" financed by QREN I&DT, no. 38962, with promotor VISUALFORMA - Tecnologias de Informação, SA and co-promoter University of the Algarve.

Keywords: Web Crawler/Robot; E-commerce; Online Reputation; Revenue Management; Big Data; Data Warehouse.

Acknowledgements

Firstly, I would like to thank to my thesis advisors, Dr. Pedro Cardoso and Dr. Roberto Lam for all the support, availability and valuable critical thinking that only thus made possible the completion of this work.

Thanks to the QREN for financing the SRM I&DT project, n.º 38962, which from this work resulted, to project leader VisualForma - Tecnologias de Informação S.A. and all the colleagues who participated in the SRM project.

I also have to thank to my parents, brother and gradparents who have always supported me and endured the emotional and economic level, making able the completion of this work.

Last but not least, I want to thank my friends, specially João Silva and Ana Lúcia, my classmates and my lab colleagues for all the support and all the determination they gave me to complete this project.

Contents

Acronyms and Abbreviations

AGRI	*Aggregated Guest Reputation Index*
AJAX	*Asynchronous Javascript and XML*
API	*Application Programming Interface*
B2C	*Business To Consumer*
BDW	*Big Data Warehouse*
BI	*Business Intelligence*
BSON	*Binary Structured Object Notation*
Crawler	*Same as web crawler, web robot or bot*
CRM	*Customer Relationship Management*
DB1	*Primary Database*
DB2	*Secondary Database*
DB	*Database*
DM	*Data Marts*
DOM	*Document Object Model*
DW	*Data Warehouse*
ERM	*Entity-Relationship Model*
ETL	*Extraction, Transformation and Load*
GDS	*Global Distribution Systems*
GPS	*Global Positioning System*

HTML	*HyperText Markup Language*
HTTP	*Hypertext Transfer Protocol*
JSON	*JavaScript Object Notation*
KPI	*Key Performance Indicators*
NLP	*Natural Language Processing*
NoSQL	*Not only Structure Query Language*
OLAP	*Online Analytical Processing*
OR	*Online Reputation*
OTA	*Online Travel Agents*
OWL	*Web Ontology Language*
OWL DL	*Web Ontology Language Description Logic*
PMS	*Property Management System*
POS	*Part-Of-Speech*
RDB	*Relational Database*
RDBM	*Relational Database Model*
RM	*Revenue Management*
SGRI	*Semantic Guest Reputation Index*
SQL	*Structure Query Language*
SRM	*Smart Revenue Management*
URL	*Uniform Resource Locator*
ViDE	*Vision-Based Data Extractor*
ViDIE	*Vision-Based Item Data Extractor*
ViDRE	*Vision-Based Data Record Extractor*
VINTs	*Visual Information and Tag structure based wrapper generator*
W3C	*World Wide Web Consortium*
WWW	*World Wide Web*
XHTML	*Extensible Hypertext Markup Language*
XML	*Extensible Markup Language*

XPATH	*XML Path Language*
YM	*Yield Management*
YTD	*YearTo-Date*

1

Introduction

For a number of years, the hospitality industry and its partners, e.g., Global Distribution Systems (GDS) and Online Travel Agents (OTA), have been promoting their transactional services on the web [42]. These websites provided hotels with certain types of business information on a free partnership base, mainly because they had the need to promote their services, thus making them willing to promote alliances and to facilitate information, as a means to achieve higher growth rates. This was typically a win-win situation but, more recently, this scenario is changing in what concerns business relations between Hotel and OTAs.

With the amount of data that daily circulates through the web and a number of users estimated at 3 billion in 2015 [49], there is a lot of information about competitors,

the hospitality industry and about consumers trends which is within hand's reach. This information is increasingly more accessible to organizations at lower costs, presenting a new challenge on creating platforms that are able to deal with this huge amount of data that organizations have at their disposal.

Due to the intense competition, hotel managers are trying to promote their services at OTAs' sites that hold the highest market shares. These predominant OTAs, taking advantage of their privileged position, started to demand extra fees for promotion and for the facilitation of business intelligence data, besides higher booking commissions.

Other hand, hotel managers strive to achieve the best possible revenues but, in order to do that, they need to be in possession of updated and reliable information about their competitive set, e.g.: hotels with similar location, facilities, class of service, number of rooms or Guest's Reputation Index.

Simultaneously, many travelers consult different websites before booking online or to contact a hotel booking service, which reinforces the idea of the increasingly important role that the OTAs have in choosing/promoting a particular hotel.

Further more, common travelers plan their vacations or travels using the internet to search for information about tourism products that they intend to consume, such as accommodation, transportation and entertainment. They search for information about other travelers opinions, to know if they had a good experience in the destinations, they intend to visit. Therefore, it is common to have travelers making their decisions, about what they want to experiment in their holydays, using the information they have access to in the internet. In resume, they will buy according to their preferences and the opinions of others travelers [55].

To ensure the quality that guests seek, the hotels follow the Revenue Management (RM) concept or Yield Management (YM) that is a sophisticated form of supply and demand management that helps a firm maximize revenue by balancing pricing and inventory controls [94].

1.1 Contextualization and Objectives

Nowadays, hospitality industry and its partners, hotels, airline companies and travel agents are promoting their services on the web. Consequently, the World Wide Web (WWW) as become a global vitrine where specialized sites (e.g. Global Distribution Systems and Online Travel Agents) operate. This modus operandi, providing publicly available information that can be collected, generating large sets of data, that can be used for business intelligence purposes, returning a comparison of offers for similar products.

The objective of Revenue Management is to establish strategies based on the understanding of market dynamics, to anticipate and influence consumer's behavior and to maximize revenue and profits from a fixed resource. The amount of data required to produce optimal decisions is huge, justifying the adoption of the concept currently known as Big Data [36].

This thesis presents some of the work developed in the project "SRM: Smart Revenue Management" financed by QREN I&DT, no. 38962, with promotor VISUALFORMA - Tecnologias de Informação, SA and co-promoter University of the Algarve.

SRM project aimed to develop a set of tools with the ability to extract and store data from secialized websites in real time and with that the information is easily searchable and integratable with a RM system.

The amount of data needed to feed a RM system is huge. As such, four web robots were developed, also known as web crawlers, with the ability to automatically extract data from four sales channels, namely: *Booking.com*, *Expedia*, *TripAdvisor* and *Bestday*. These were the four choosen channels due to their popularity and influence in the hospitality market today.

The data extraction works as following: (a) each crawler begins by filling the website search form with a set of configurable parameters. (b) The search return a set oh hotels which meet the predefined criteria and their information from sales channels

relatively to prices, promotions, reviews and characterization of the hotel is extracted.

The extracted information is initially stored in a NoSQL database (MongoDB [67]). This step is important because it allows to store information as it is extracted, thus guaranteeing the possibility of making reverse engineering in case of consolidation process goes wrong. In the next step, the stored data suffers a mapping process so that information from the four channels (websites) is consolidated and unified in accurate information, without inconsistencies. Finally, the information is stored on a relation SQL database (SQL Server), which allows to have more standardized information and ready for use by any application that is able to read the database.

By now, the extracted data will be used by mathematical prediction models that analyze the prices charged by hotels in recent years and generate a price forecast that hotels will practice in the future according to their sales history. Besides, the extracted data is used to analyze the guests reviews wich are classified between positive and negative reviews. The hotel, with all this information can decide whether to continue to invest in a particular service or even improve a less good service.

This thesis presents four papers resulting from SRM project explaining in more detail the above concepts. An overall contribution to each paper is described bellow, with each one being discussed individually in a complete dedicated chapter, having their own introduction, state of art and conclusions.

- **"A Web Crawler Framework for Revenue Management"** (Martins, Lam, Rodrigues, Cardoso, Serra, 2015) [61]

  This paper presents the implementation of web crawlers that simulates human behavior for data extraction from channels like *Booking.com* with the goal of collect extracted data into a NoSQL database (MongoDB).

- **"Big Data Warehouse Framework for Smart Revenue Management"** (Ramos, Correia, Rodrigues, Martins, Serra, 2015) [77]

  This paper starts by explaining how the data is extracted from the web and

stored into a primary database. Then it describes the use of the mapping rules to convert data from the primary database (NoSQL) to a secondary renational database.

- **"Challenges in Building a Big Data Warehouse Applied to the Hotel Business Intelligence"** (Martins, Ramos, Rodrigues, Cardoso, Lam, Serra, 2015) [62]

  This paper continues the work done on first paper: after information being stored into a MongoDB dabatase then it will be consolidated and segmented into small parts of information. The redundant data is grouped and unified. Finally, the consolidated data is stored into a relation database (SQL server).

- **"Guest Reputation Indexes to Analyze the Hotel's Online Reputation Using Data Extracted from OTAs"** (Choupina, Correia, Ramos, Martins, Serra, 2015) [28]

  This paper describes the work done to classify and segment the reviews of guests into positive and negative, and collect all scores about segments (e.g. location, comfort, cleanliness and facilities). The scores are collected separately from each guest review and will be used to calculate Guest Reputation Indexes.

1.2 General Scope of the Thesis

This dissertation will present some of the papers of the SRM project, resulting in the following structure. The current chapter introduced the theme, the objectives and the contributions of this work.

Chapter 2 explains how crawler accesses the target websites and extracts information about a set of features that characterize the hotels listed there. Additionally it will present the document-oriented database used to store the retrieved information and discuss the usefulness of this framework in the context of the SRM.

Chapter 3 proposes a three stage framework to develop the Big Data Warehouse for the SRM. Namely, (a) the compilation of all available information, in the present case, it

was focus only the extraction of information from the web by a web crawler – raw data. (b) The storing of that raw data in a primary (NoSQL) database, and from that data (c) the conception of a set of functionalities, rules, principles and semantics to select, combine and store in a secondary relational database the meaningful information for the Revenue Management (Big Data Warehouse).

Chapter 4 presents the challenges and some of the necessary steps to overcome the problems associated with the information management and consolidation in a hotel Big Data Warehouse. It explains what rules are used for data consolidation and how we can transform "raw" data into normalized data ready to be readed directly form the database.

Chapter 5 explains how we can use the extracted information from the web to help hoteliers to monitorize their presence in OTAs. It proposes two guest reputation indexes: the Aggregated Guest Reputation Index (AGRI), which shows the positioning of a hotel in different OTAs and it is calculated from the scores obtained by the hotels in those OTAs; And the Semantic Guest Reputation Index (SGRI), which incorporates the social reputation of a hotel and that can be visualized through the development of word clouds or tag clouds.

Finally, Chapter 6 draws some conclusions and presents some future work.

2

A Web Crawler Framework for Revenue Management

Chapter Outline

Smart Revenue Management (SRM) is a project which aims the development of smart automatic techniques for an efficient optimization of occupancy and rates of hotel accommodations, commonly referred to, as Revenue Management. To get the best revenues, the hotel managers must have access to actual and reliable information about the competitive set of the hotels they manage, in order to anticipate and influence consumer's behavior and maximize revenue. One way to get some of the necessary infor-

mation is to inspect the most popular booking and travel websites where hotels promote themselves and consumers make reservations and provide reviews about their experiences. This Chapter presents a web crawler framework to perform automatic extraction of information from those sites, to facilitate the (RM) process of a particular hotel. The crawler periodically accesses the targeted websites and extracts information about a set of features that characterize the hotels listed there. Additionally, we present the document-oriented database used to store the retrieved information and discuss the usefulness of this framework in the context of the SRM system.

2.1 Introduction

The objective of Revenue Management is to establish strategies based on the understanding of market dynamics, to anticipate and influence consumer's behavior and to maximize revenue and profits from a fixed resource. The amount of data required to produce optimal decisions is huge, justifying the adoption of the concept currently known as big data [36].

For a number of years, the hospitality industry and its partners, e.g., Global Distribution Systems (GDS) and Online Travel Agents (OTA), have been promoting their transactional services on the web [42]. These sites provided hotels with certain types of business information on a free partnership base, because they had a great need to promote their services, thus making them willing to promote alliances and to facilitate information, as a means to achieve high growth rates. This was typically a win-win situation but, more recently, this scenario is changing in what concerns hotel-OTAs business relations.

Due to the intense competition, hotel managers are trying to promote their services at OTAs' sites that hold the highest market shares. These predominant OTAs, taking advantage of their privileged position, started to demand extra fees for promotion and for the facilitation of business intelligence data, besides higher booking commissions.

As stated before, hotel managers strive to achieve the best possible revenues but in order to do that, they need to be in possession of actual and reliable information about their competitive set (e.g. hotels with similar location, facilities, class of service, number of rooms, Guest's Reputation Index) and about the corresponding total demand.

One way to get some of the data is to inspect the sites were the hotels of the same competitive set are doing their promotion and bookings. The simplest form, although not the cheapest one, is to contract the access to business data through an API commercialized by the OTA of interest. More complex is to get the information from HTTP, simulating the behavior of a user (not the same as "hacking").

The latter, extraction using a web robot (bot) or crawler, makes it possible to get partial data (e.g., prices, room types, capacity, facilities, amenities, as well as comments from former guests), available on the hotels. The amount of data to extract is huge, since the crawler must run periodically in order to extract updated data.

The OTAs' sites, like *Booking.com* [42], protect their data from those extraction processes for two reasons: (a) They do not want to have bots scrapping their servers, since this will cause an overload of the systems and consequent delays in the normal processes of promoting and booking and; (b) They want to sell the access to the data through the API of their engine site. Nevertheless, research has been conducted on data extraction from web information systems [16, 17, 39, 98], but only a small number of studies have been published on the subject of business to consumer (B2C) [41].

Smart Revenue Management (SRM) is a project in development by the University of the Algarve and VISUALFORMA - Tecnologias de Informação, SA, which aims to develop a set of tools to integrate in a RM system. In this Chapter, we present some part of that tool, namely a web crawler framework for smart RM. The main contribution to that system is the smart web crawler and its associated database to store the extracted information.

2.2 Contextualization and State of the Art

The World Wide Web (WWW) is currently the largest source of information available to the public. The increasing use of the internet, worldwide, has made e-commerce to evolve, thus facilitating the emergence of an evergrowing number of sales channels and business opportunities [15].

Data extraction from the web, including from e-commerce sites, is useful for many types of business analysis, allowing the use of predictive models to enhance the performance of the revenue management system [75]. The extraction of huge amounts of data from the web requires almost impracticable time and effort for humans, thus justifying the creation of mechanisms for its automatic extraction [72].

When the HTML structure of the pages is constant, creating a mechanism that is able to automatically parse and extract data from a particular site is not a difficult task. However, if the site administrator/programmers decide to change the structure of the DOM tree1 or the attributes contained in the tags (e.g., class or id values), it is mandatory to (re)implement/adjust the extraction mechanisms (this is what happens in the websites used to extract the information for the RM).

Over the last decade, several studies have been conducted about the automatic extraction of information from the web. Lerman et al. [56] presented a fully automatic system that can extract lists' and tables' data from web sources through a series of assumptions about the structure of those lists and structures. In [76] it was used an algorithm that requires the user to identify the relevant data to extract. The algorithm uses the Minimum String Edit Distance procedure and is able to identify the tags that are normally used for that type of data. A system that compares two HTML pages to find patterns was proposed in [30]. It uses two web pages with similar structure but with different contents in order to identify patterns between the pages and to create extraction rules.

The Tree-Edit Distance algorithm was used in [79] to find patterns between differ-

ent structures. Initially the HTML pages are converted into a DOM tree. The Tree-Edit Distance applied to two DOM trees computes the distance between two tree, T_a and T_b, which is the cost associated with the minimum number of operations needed to convert the T_a tree into the T_b tree. These operations can be the insertion, the removal or the edition of nodes. Then the system does the page clustering which is the grouping of pages with a certain resemblance. Once the pages are grouped, the tree edit distance is used to generate the extraction pattern applied afterwards. A similar implementation, based on pattern analysis and through DOM trees using the edit distance algorithm, was used in [78].

The VINTs (Visual Information and Tag structure based wrapper generator) was proposed in [98] to extract data from search engines results. The VINTs refers to the visual content of the web page to find regularities / contents order, without being interested on the page's HTML code. Those existing regularities are combined with HTML code regularities to generate the wrappers.

Papadakis et al. [72] presented a way to figure out the format of the information contained in web pages and discover the associated structure. This system consists of three distinct phases. In the first phase, the system transforms the HTML document into a XHTML document through some syntactic corrections, making it "well" structured, and generates the correspondent DOM tree. In the second phase, the regions containing information of interest are segmented. Finally, the third phase consists of mapping the nodes of interest in the original HTML page.

Zhai and Liu [97] presented a system that only requires a sample page labeling. They use a method called Sufficient Match to identify the similitude between the objective page and the main sample page. The ViDE (Vision-based Data Extractor) [58] is a recent method of data mining that relies on the visual aspect of how the data is presented. The ViDE is composed by ViDRE (Vision-Based Data Record Extractor) and ViDIE (Vision-Based Item Data Extractor).This system operates in four distinct steps: First, the visual representation of a sample page is computed, which is then trans-

formed into a visual block tree. Second, the system extracts the data records contained in the visual block tree and the third consists in the separation, supported in semantics, of the extracted data in "data items" and groups. Finally, a set of visual extraction rules to produce the visual wrappers is generated.

However, the previous methods do not take into consideration a major problem, posed by the dynamic pages that use JavaScript. Many Web pages use JavaScript to trigger dynamic changes in the HTML code without any request or response from the web server. In particular, JavaScript is employed in many e-commerce sites to hide information and to difficult the automatic data extraction task. Since these scripts cause changes in the structure of the HTML code only in the client side, it is necessary to interact with the web page so that information becomes visible.

Baumgartner and Ledermiiller [18] presented a method, which overcomes this problem. For that, they have proposed the Lixto Visual Wrapper, which integrates the Mozill browser driver to interact with the web page in order to display the information that is hidden in the backend database. The Lixto Visual Wrapper allows the user to view the page, to extract the data from that page and to interact with it, by sending commands from the keyboard or mouse. Those keyboard and mouse commands are recorded to the element, as well as the XPath, (the path to a given node in a DOM tree) that is to be extracted.

The previous methods are, in general, specific to the data extraction from particular web pages, which are easily transposable to a database schema. For these reasons, most systems store data in XML or relational databases. However, our system aims to extract data from multiple e-commerce sites. The issue is that each vendor has different ways of describing the same product [75]. Therefore, it is almost impracticable to create a well-structured database schema for a set of e-commerce sites.

Thus, our framework uses MongoDB [67], a non-relational database. MongoDB is a NoSQL [78] document-oriented database, with high performance and high reliability. Other characteristics include easy scalability (vertically and horizontally through

replication and auto-sharding, respectively) and map-reduce.

A MongoDB database is structured as a set of collections, which store documents. These documents are BSON objects (binary JSON [47] document format), allowed to have a dynamic schema, i.e., documents in the same collection are not forced to have the same structure. This schema-less property is particularly important in the problem in study since the data retrieved from the sites does not follow, in general, a common design. Furthermore, in our case, the data sources in question are themselves quite dynamic, including and removing new fields very often, which makes it very hard to design and maintain a relational database schema.

2.3 Web crawler framework

As mentioned, the amount of data needed to feed the RM system is huge and the crawlers must run periodically in order to update it. Nevertheless not all the data that is extracted can be used, as is, by the RM models, and from different sites it is possible to extracted different and coincident information from the same hotel.

Due to the above reasons, we choose to apply a crawler to each site (*Booking.com*, *Expedia*, *TripAdvisor*, etc.), extracting periodically all the information existing on the site over different periods of time, simulating different users (2 adults, 2 adults and 1 kid, etc.). In this Chapter, we will only address the crawler and a primary database, where the "raw" extracted data will be stored. Nevertheless, it is important to comprehend the context of the framework, namely that from this raw data, a secondary database is being created using rules and semantics to produce the necessary filtered data for the RM models. Figure 2.1 shows the resumed block diagram of the entire framework, where the two darker blocks (in blue) are the ones that we will address on this Chapter.

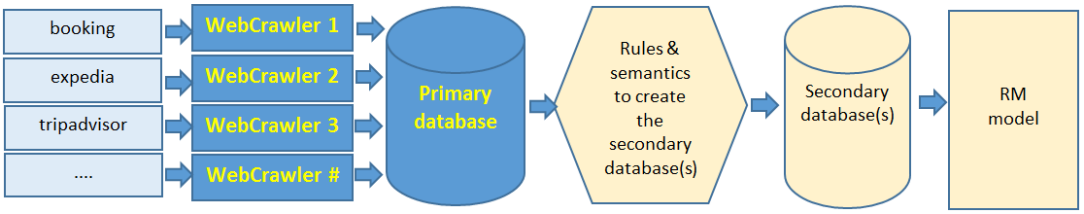


Figure 2.1: Resume block diagram of the SRM.

2.3.1 Implementation of the crawler

As stated in Sec. 2.2, research has been done about methods to automatically extract data from web sites. However, only a few publications mentioned the specific dynamics of ecommerce pages and the ways to deal with them. Most of the e-commerce websites development use JavaScript and AJAX to implement the characteristics associated with ecommerce (e.g., motion, pre-filled data and suggestions). The e-commerce sites are designed for human interaction, disclosing information to the costumer according to the previous inputs. To overcome the complexity of the interaction with this kind of websites we decided to use a webdriver.

The W3C [91] defines the specification for the WebDriver API as a platform and languageneutral interface and associated wire protocol that allows programs or scripts to introspect into, and control the behavior of a web browser [90]. The WebDriver API is primarily intended to allow developers to write tests that automate a browser from a separate controlling process.

In the previous context, ChromeDriver [86] and C# allows simulating the human behavior in the interaction with the browsers/sites. Together they allow to playact the process of inserting text in entry fields and doing clicks on other input elements (e.g., checkboxes, radiobuttons, submit buttons).

We must now recall that the goal is to extract information from websites that show lists of hotels and allow the booking thereof. In general, these sites work in a similar manner: (1) their homepage have a form to allow searching for a type of hotel, city and a period of stay. After fulfilling the form, a (2) list of hotels that match the criteria will

be returned by the site. Clicking on one of these listed hotels, (3) a new page will be showed, exposing information about the selected hotel to the user, namely: available rooms, prices, feature, amenities, policies, guest comments, etc.

It is in the interest of the SRM project to extract information about the list of available rooms and corresponding prices, as well as reviews made by the former customer of the hotels. Almost all websites of this kind provide this data. The main obstacle in building an automatic mechanism for extracting data is the fact that each site provides this information in the manner that it considers most appropriate for a human costumer. Beside this fact, due the market competition, there is no interest of these websites in granting free access to automatic extraction of the information. Some of those sites change, from time to time, their pages layout (design), as well as the attributes values of tags, which contain important information for our purposes.

Our algorithm of extraction is described in the following basic steps:

- **Step 1.** Set the URL to the website to be crawled (e.g., *Booking*, *Expedia*); set the data to fill the website forms (e.g., location, check-in and check-out dates, number of persons, number and type of rooms available) and other parameterizations (e.g., language, currency);

- **Step 2.** Automatically fill an instance of the webdriver which models the behavior of the website to be crawled (using data from Step 1), and do a request to the corresponding server;

- **Step 3.** Store the response of the server request, into a list of links referencing hotels and boarding houses that satisfy the research domain of Step 1;

- **Step 4.** For each link of Step 3, do a second level of request to the server and extract all relevant data (e.g., type of hotel, available rooms, prices, comments and rank position);

The process of finding the important HTML elements (those containing pertinent data for RM), will be described in following sections.

Relevant HTML elements

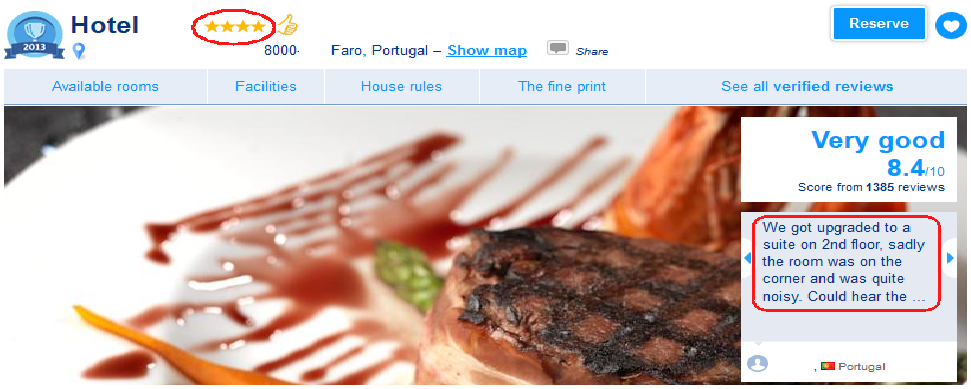
Despite the use of Javascript and AJAX at the front end of the system, it will be a browser rendering the final contents in HTML [91]. Due its intrinsic markup nature, HTML is a language that allows the construction of blocks of nested tags. Thus, in order to extract the value of a specific element with a specific tag, we can use the XPATH definition to do this [91]. The alternative to identify the relevant elements is to locate the attribute associated with a tag.

For instance, the tag <*div class="hotels" id="Hotel1"*> </*div*> contains two attributes; class and id. Their values are "hotels" and "Hotel1". The attributes are defined by pairs of *name* = *value*. The use of the absolute XPATH for accessing a particular element has a main drawback: if there is any change in the page structure the tag might no longer be accessible. To cope with this inconvenience we first use the attributes. If the tag element does not have an attribute, then the relative XPATH will be used. To surpass the problem of absolute XPATH we use a tag from higher level and, from there, the relative XPATH to the target tag is used.

In Fig. 2.2 middle, if we would try to extract the text "We got upgraded to a ..." it is necessary to use the relative XPATH, because the text is inside a <*span*> tag without attributes. The relative XPATH to the text would be *//div[@class='reviews-carousel-scroll'] /div/p/span*, which means: find the tag with the attribute class equal to reviews-carousel-scroll and from it follow the */div/p/span* relative XPATH to the target tag.

The information about the target tags for each website are manually provided and placed in a database. The webmasters of e-commerce websites tend to change the attributes values periodically, although not so frequently, to avoid user/client annoyance. Besides, complete renewals of the website take place from time to time. In this case, there is no other solution than to redo / redefine the target tags.

Nevertheless, for the first case, where the attributes are changed, we use a vector for each target tag, where all the previous used tags are kept, and register how many times these tags has been used. When one of the tags is not found, the crawler goes to the




```
<div class="reviews-carousel-scroll">
  <div class="althotelsDiv2 use_sprites_no_back featured_reviewer">
    <p class="trackit althotelsReview2 fixed_review_height
    fixed_review_top_align review_content" lang="en" data-url="/hotel
    /pt/   .en-gb.html?sid=18c705b8cbd20b48ed2ee738c6831abd;dcid=1;
    dist=0;no_rooms=1;req_adults=2;req_children=0;type=total&;
    rid=401904995#hash-blockdisplay4">
      <span>We got upgraded to a suite on 2nd floor, sadly the room was
      on the corner and was quite noisy. Could hear the ...</span>
    </p>
    <p class="trackit altHotelsReviewer2" style="font-size: 94%;">
    <div class="icon_user_back_container">
  </div>
```

```
<h1 class="item">
  <span id="hp_hotel_name" class="fn">Hotel    </span>
  <span class="nowrap hp__hotel_ratings">
    <span class="hp__hotel_ratings__stars">
      <i class="b-sprite stars ratings_stars_4 star_track" data-track-
      on-mouseover="YdAZJLVINKNMC" title="4-star hotel"></i>
    </span>
    <i class="b-sprite icon_deal_smart jq_tooltip" rel="300"></i>
  </span>
</h1>
```

Figure 2.2: At the top, the rendered view, in the middle, the HTML code that shows the tag without attributes () and at the bottom, the same, but now the tag without explicit information (<i>).

correspondent storage vector and try to find a correspondent tag (from those already used and stored in the vector). Usually, the web programmers employ the change of attributes values in order to block the automatic data extraction mechanisms but as humans, they start reusing values. If the attribute value has never been used a counter is used in order to keep the occurred number of failures, this will allow determining the ratio of unsupervised versus supervised data extraction data.

Data extraction from HTML tags

After the recognition process, it is necessary to extract the information from the HTML tag. The information to extract is located between the begin and end tags of the element (e.g. * Hotel x *). However, some of the information is not showed as explicit text. For instance, some websites show the number of stars of the hotel as an image (Fig. 2.2). In this case we extract a relevant attribute value. The images used in ecommerce website have attributes to help the posing the image in the page and defining the semantic importance of the image, see Fig. 2.2.

2.4 Database implementation

Taking into consideration that distinct websites have distinct designs and structures, we have chosen MongoDB (see Sec. 2.2). MongoBD allows to storage data in collections. In our database there are four collections: AboutHotel, Rooms, Comments and Scores.

All the data concerning the hotel characteristics (e.g., facilities, name, address, and number of star), are kept in the AboutHotel collection. Figure 2.3 shows an example of the data extracted from *TripAdvisor* for a 4 star hotel (in this figure and similar ones, some fields were truncated for questions of size). Data for the same hotel, extracted from *Booking.com* is presented in Fig. 2.4. As we can observe, there are significant differences between sites both on type of information and respective details.

The Rooms collection keeps the information about the rooms of the hotels (e.g., room name, type, capacity, prices/dates, and used taxes). Figure 2.5 presents an example of the data retrieved from *Booking.com*. The Comments collection keeps comments made by formers guests. In this case, the comments are grouped by the type of guest (e.g., couple, family and friends) – see an example on Fig. 2.6. The Score collection keeps the scores given by former guests for evaluations of cleanness, staff kindness/efficiency and comfort of the hotel – see an example on Fig. 2.7.

```
{
"_id":ObjectId("5432c0c9703c3b04260f1ff8"),
"ExtractionDate":ISODate("2014-10-06T16:18:16.170Z"),
"IDFieldsHotel":{
  "_idHotel":ObjectId("5432c0c7a563ee258cf7b75b"),
  "Source":"tripadvisor.ie",
  "HotelName":"xxxxxxxxxxx",
  "Address":"Calle xxxxxxx",
  "Coordinates":"45.xxxx,12.xxx",
  "Stars":4000000
},
"Topics":[
  {"Title":"Amenity", "Content":"Bar / Lounge"},
  {"Title":"Amenity", "Content":"Free Breakfast"},
  {"Title":"Amenity", "Content":"Free High Speed
      Internet ( Wi-Fi )"},
  {"Title":"Amenity", "Content":"Pets Allowed ( Dog /
      Pet Friendly )"},
  {"Title":"Amenity", "Content":"Restaurant"},
  {"Title":"Amenity", "Content":"Room Service"},
  {"Title":"Amenity", "Content":"Suites"},
  {"Title":"Amenity", "Content":"Wheelchair access"},
  {"Title":"Address", "Content":" Calle xxxxxxx"},
  {"Title":"Phone Number", "Content":" 00 xxxxx"},
  {"Title":"Region", "Content":"\r\nItaly xxxxxxx"},
  {"Title":"Price Range (Based on Average Rates)","
      Content":"xxxx"},
  {"Title":"Hotel Class", "Content":"\r\n4 star xxxx 4
      *"},
  {"Title":"Number of rooms", "Content":" 77"},
  {"Title":"Reservation Options", "Content":"\r\
      nTripAdvisor is proud to partner with Booking.
      com, TripOnline SA, Hoteis.com, Agoda and
      Odigeo so you can book your xxxxxxxxxx
      reservations with confidence. We help millions
      of travellers each month to find the perfect
      hotel for both holiday and business trips,
      always with the best discounts and special
      offers."},
  {"Title":"Also Known As",  "Content":"xxxxxxxxxx"}
],
"Photos":[],
"RecommendedHotels":[ "xxxxxxxxxx", "xxxxxxxxxx"]
}
```


Figure 2.3: Example of document stored on collection AboutHotel with information from *Tripadvisor*

```
{
"_id":ObjectId("5432beda703c3b04260f1ca5"),
"ExtractionDate":ISODate("2014-10-06T16:09:59.517Z"),
"IDFieldsHotel":{
  "_idHotel":ObjectId("5432bed7a563ee17546ab81f"),
  "Source":"booking.com",
  "HotelName":"xxxxxxxxxxx",
  "Address":"Calle xxxxxxx",
  "Coordinates":"45.xxxx,12.xxx",
  "Stars":4000000
},
"Topics":[
  {"Title":"Description", "Content":"xxxxxxx is a
      former xxxx with well-kept gardens, (...)."},
  {"Title":"Low rates", "Content":"Best price for 1
      guest:  126.65 (for 1 night)  (...)"},
  {"Title":"578 verified reviews", "Content":"We
      verify all reviews from guests who (...)"},
  {"Title":"Bedroom", "Content":"Wardrobe/Closet"},
  {"Title":"Outdoors","Content":"Garden, Terrace"},
  {"Title":"Media & Technology", "Content":"Pay-per-
      view Channels, Flat-screen TV, Telephone"},
  {"Title":"Food & Drink", "Content":"Restaurant, Bar,
       Breakfast in the room, Restaurant (...)"},
  {"Title":"Internet", "Content":"Free! WiFi (...)"},
  {"Title":"Parking",  "Content":"No parking (...)."},
  {"Title":"Services", "Content":"Room service, 24-
      hour front desk, Express check-in/(...)"},
  {"Title":"General", "Content":"Newspapers, Safety
      deposit box, Non-smoking rooms, (...)"},
  {"Title":"Languages spoken", "Content":"Italian,
      English"},
  {"Title":"Check-in", "Content":"From 14:00 hours"},
  {"Title":"Check-out", "Content":"Until 12:00 hour"},
  {"Title":"Cancellation/prepayment", "Content":"
      Cancellation and prepayment policies(...)"},
  {"Title":"Children and extra beds", "Content":"All
      children are welcome."},
  {"Title":"Children and extra beds", "Content":"Free!
       One child under 4 years stays (...)"},
  {"Title":"Children and extra beds","Content":"One
      older child or adult is charged 40 %(...)"},
  {"Title":"Children and extra beds", "Content":"The
      maximum number of extra beds/children's(...)"},
  {"Title":"Children and extra beds", "Content":"Any
      type of extra bed or child's cot/crib (...)"},
  {"Title":"Children and extra beds","Content":"
      Supplements are not calculated (...)"},
  {"Title":"Pets","Content":"Pets are allowed (...)"},
  {"Title":"Groups","Content":"When booking more than
      5 rooms, different policies (...)"},
  {"Title":"Cards accepted at this property","Content"
      :"Hover over the credit cards for more info."},
  {"Title":"The fine print","Content":"In case of
      early departure, the hotel may charge (...)"}
],
"Photos":[ "http://r-ec.bstatic.com/images/hotel/max30
    0/320/xxxxxx.jpg", (...)],
"RecommendedHotels":[ xxxxxxx, xxxxxxx, xxxxxxx]
}
```

Figure 2.4: Example of document stored on collection AboutHotel with information retrieved from *Booking.com*.

```
{
"_id":ObjectId("5432bee5703c3b04260f1ca6"),
"IDFieldsHotel":{
  "_idHotel":ObjectId("5432bedea563ee17546ab820"),
  "Source":"booking.com",
  "HotelName":"xxxxxx",
  "Address":"Callexxxxxxxx",
  "Coordinates":"12.xxxx,45.xxxx",
  "Stars":4000000
},
"Search":{
  "Location":"Faro, Portugal",
  "NumberOfAdults":2,
  "NumberOfChildren":0,
  "NumberOfRooms":1,
  "CheckinDelayNights":10,
  "DifferenceBetweenCheckinCheckout":1,
  "CheckinDate":ISODate("2014-10-15T23:00:00.000Z"),
  "CheckoutDate":ISODate("2014-10-16T23:00:00.000Z")
},
"ExtractionDate":ISODate("2014-10-06T16:10:06.777Z"),
"RoomName":"Double or Twin Room",
"Description":[
  {"Title":"Facility", "Content":"Air Conditioning"},
  {"Title":"Facility", "Content":"Soundproofing"},
  {"Title":"Facility", "Content":"Flat-screen TV"},
  {"Title":"Facility", "Content":"Free WiFi"},
  {"Title":"Info","Content":"Air-conditioned en suite
       room with modern design. Offers a flat-screen
       TV with national and international channels and
        a minibar.\r"},
  {"Title":"Info", "Content":"\r"},
  {"Title":"Info", "Content":"Please specify bed
       preference when booking.\r"},
  {"Title":"Room facilities", "Content":" Safety
       Deposit Box, Air Conditioning(...)"},
  (...)
  ],
"TariffList":[
  {"Conditions":["Non-refundable", "Breakfast included
       "],
   "Tax":[ {"Title":"Included","Content":" 10% VAT"},{
        "Title":"Not included","Content":"  4.50 city
        tax per person per night."}],
   "MaxOccupancy":[{"Title":"Adults", "Content":"2"}],
   "OldPrice":{"_t":"TitleValue", "Title":" ", "Value"
       :17400},
   "NewPrice":{"_t":"TitleValue","Title":" ","Value":1
       5810}
  },
 (...)
]
}
```

Figure 2.5: Example of document stored on collection Room (from *Booking.com*).

```
{
"_id" : ObjectId("5432bfc3703c3b04260f1cab"),
"IDFieldsHotel" : {
  "_idHotel" : ObjectId("5432bee9a563ee17546ab821"),
  "Source" : "booking.com",
  "HotelName" : "xxxx",
  "Address" : "xxx",
  "Coordinates" : "12.xxx,45.xxxx",
  "Stars" : 4000000
},
"ExtractionDate" : ISODate("2014-10-06T16:10:20.214Z")
    ,
"ReviewerType" : "Couples",
"CommentDate" : "5 October 2014",
"ReviewerLocation" : "Canada",
"CommentTitle" : "Good",
"CommentList" : [
    {
        "Title" : "Negative",
        "Content" : "I was disappointed (...)."
    }
],
"CommentScore" : 790
}
```

Figure 2.6: Example of document stored on collection Comments (from *Booking.com*).

Each of these collections has a general structure capable of storing data from this type of websites described above. By general structure, we mean the subject lists and tables of data for each website, kept with no dependencies in name or content manner, see Fig. 2.8.

In order to keep the independence from the websites structures, the documents have keyvalue structure for "title" and "content". The "title" will possess the name of the subject in the layout of website and the "content" will be the content of that subject.

The proposed database is the primary database, with the propose of retrieving and storing has much data as possible from the websites. Later, in dependence of the aim/purpose of research, a large number of studies can be conducted. Data can be grabbed from the database and clustered in a single unit of information or, it can be used to create time series, in order to analyze past performance or to predict future trends (see Fig. 2.1).

```
{
"_id" : ObjectId("5432c3f8703c3b04260f2380"),
"IDFieldsHotel" : {
  "_idHotel" : ObjectId("5432c3f6a563ee258cf7b75e"),
  "Source" : "tripadvisor.ie",
  "HotelName" : "xxxxxxxxxx",
  "Address" : "xxxxxxxx",
  "Coordinates" : "45.xxxx,12.xxxx",
  "Stars" : 4000000
},
"ExtractionDate" : ISODate("2014-10-06T16:31:52.706Z")
    ,
"ScoresList" : [
  {"Title" : "Sleep Quality", "Value" : 450},
  {"Title" : "Location","Value" : 400},
  {"Title" : "Rooms","Value" : 450},
  {"Title" : "Service","Value" : 450},
  {"Title" : "Value","Value" : 400},
  {"Title" : "Cleanliness","Value" : 450}
],
"ReviewerType" : "All",
"ReviewersCount" : 923
}
```

Figure 2.7: Example of document stored on collection Scores (from *TripAdvisor*).

2.5 Discussion

The usefulness of the RM system is based on the availability of data. In order to satisfy this need, the crawlers must run periodically in order to collect it in a suitable and updated manner. If this procedure is repeated with short intervals, the hoteliers can get valuable information to organize data series that can be used with predictive algorithms to decide on the best prices and service-mix strategy, in order to obtain higher revenues. In this Chapter, a web crawler framework for RM was presented, aiming to demonstrate that it is possible to automatically "browse" e-commerce websites, identify the relevant elements and extract them to a NoSQL database.

We were able to overcome the interaction of JavaScript and AJAX by using a webdriver and, although the extraction of data can not be a fully unsupervised process, human supervision is only required if page layouts are modified. In the future we pretend to improve the web crawler, so it can detect and "understand" layout modifications, adapt to them without the need of human supervision, provide metrics on

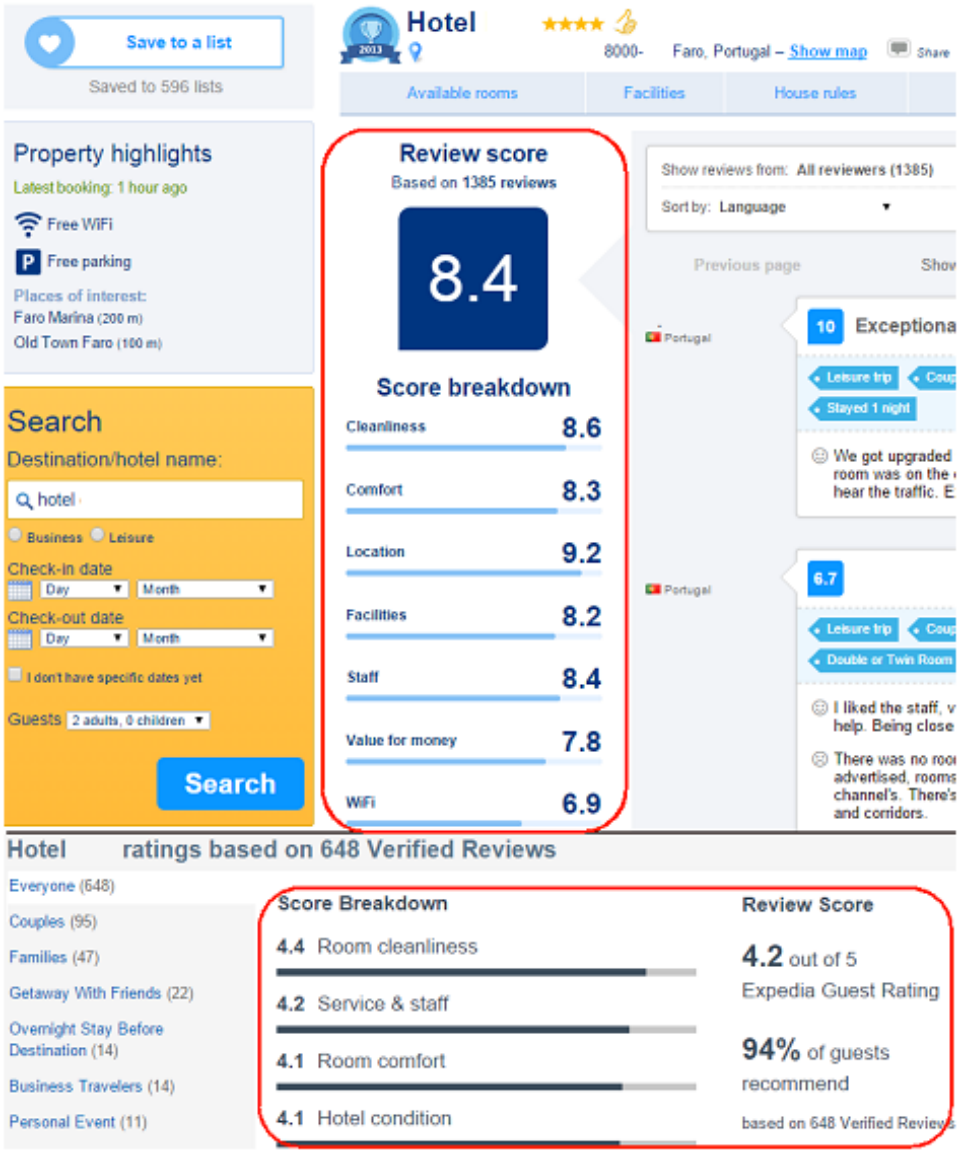


Figure 2.8: Two examples of different review scores for the same hotel on *Booking.com* (top) and *Expedia* (bottom).

the history scheme for the replacement of the attribute values (tags), and register the correspondent metrics in the database. All these metrics will be computed when the full integration of the system occurs, i.e., when the primary database (presented in this Chapter) is fully integrated with the secondary databases (see Chap. 4), which integrate the information extracted from the web crawlers of the different sites and provide formatted data for business intelligence purposes. These secondary database will be the one where the mathematical models of the RM will be based.

3

Big Data Warehouse Framework for Smart Revenue Management

Chapter Outline

Revenue Management's most cited definitions is probably "to sell the right accommodation to the right customer, at the right time and the right price, with optimal satisfaction for customers and hoteliers". Smart Revenue Management (SRM) is a project, which aims the development of smart automatic techniques for an efficient optimization of occupancy and rates of hotel accommodations, commonly referred to, as revenue management. One of the objectives of this project is to demonstrate that

the collection of Big Data, followed by an appropriate assembly of functionalities, will make possible to generate a Data Warehouse necessary to produce high quality business intelligence and analytics. This will be achieved through the collection of data extracted from a variety of sources, including from the web. This Chapter proposes a three stage framework to develop the Big Data Warehouse for the SRM. Namely, the compilation of all available information, in the present case, it was focus only the extraction of information from the web by a web crawler – raw data. The storing of that raw data in a primary NoSQL database, and from that data the conception of a set of functionalities, rules, principles and semantics to select, combine and store in a secondary relational database the meaningful information for the Revenue Management (Big Data Warehouse). The last stage will be the principal focus of this Chapter. In this context, clues will also be giving how to compile information for Business Intelligence. All these functionalities contribute to a holistic framework that, in the future, will make it possible to anticipate customers and competitor's behavior, fundamental elements to fulfill the Revenue Management.

3.1 Introduction

In the area of hospitality, the information to be managed has very specific features since it comes from different activities related to the tourism sector, such as facilities and transportation, among others. It is constantly undergoing changes as, for example, the tariffs offered to potential customers who want to book a room.

To the hotel, it is relevant that marketers and managers have access to intelligence, and make the best use of it [74]. These professionals have invested heavily in recent years, organizing strong scientific teams, including statisticians and database (DB) experts, well equipped to build and analyze the contents of their Data Warehouses.

However, the development and use of internal data sources is no longer sufficient to ensure competitive advantage [23]. This type of Data Warehouse consists of infor-

mation from the transactions that occur within the organization, while, nowadays, it is necessary to consider the current trend that favors the development and use of Big Data Warehouse architectures consisting of internal and external data sets [35, 65].

The concepts associated with Big Data [74] are describe as technologies that promise to fulfill a fundamental tenet of research in information systems, which are to provide the right information to the right receiver in the right volume and quality at the right time. Following the same path, the concept of Big Data Warehouse refers commonly to the activity of collecting, integrating, and storing large volumes of data coming from data sources, which may contain both structured and unstructured data. Volume alone does not imply Big Data. Other specific issues are related to the velocity in generating data, their variety and complexity [35].

Nowadays, hospitality industry and its partners, hotels, airline companies and travel agents are promoting their services on the web. Consequently, the World Wide Web (WWW) has become a global vitrine where specialized sites, e.g., Global Distribution Systems (GDS) and Online Travel Agents (OTA) operate, thus, providing publicly available information that can be collected, generating large sets of data, that can be used for business intelligence purposes, providing a comparison of offers for similar products.

In the early days of web-based business, data could be freely acquired from specialized websites, because it was in the business company's interest to promote their products [44]. However, nowadays, this panorama is rapidly changing, and information is not free and easy to collect. Nevertheless, hotel marketers need to have access to this kind of information, to define their revenue management policies and to redefine their business tactics and strategies, by using Business Intelligence and analytical techniques to promote and sell their rooms, at the best possible price to the right costumers.

Smart Revenue Management (SRM) is a project in development by the University of the Algarve and VISUALFORMA - Tecnologias de Informação, SA, which aims to

develop a set of tools to integrate in a Revenue Management (RM) system. This Chapter, presents the conceptual and some practical stages in development to construct a Big Data Warehouse (BDW), that will allows the detection of knowledge and the development business intelligence analytics applications.

The article is structured as follows: besides the introduction, the Sec. 3.2 presents a thorough contextualization of the subject of study. Section 3.3 highlights the relevance of Big Data Warehouse, mainly to the hospitality and tourism organizations. Section 3.4, presents the process to develop business analytic tools, based on the BDW, including the analyses of the challenges in hand and the proposed solution to solve it. Finally, we will present some discussion, conclusions and suggestions for future work.

3.2 Contextualization and State of Art

In the current society, information, creativity and knowledge play an important role in any organizational process and strategy. To cope with globalization, it is essential to use mechanisms that allow the collection and treatment of essential information for the organizations. The optimization of that information in a differentiated way for management tactical and strategic purposes is essential in all organization levels; which aims the reduction of uncertainty in the decision-making processes and track the most sensitive parameters of the organizational performance [52].

Such mechanisms/stages, in the case of the Smart Revenue Management project, include: (a) the automatic collection of information from several sources, including the internal Data Warehouse (DW) of the hotel, but also from the web (using a web crawlers). (b) The storage of the extracted information, and the (c) selection of the most relevant information to the business, taking into consideration the data model suitable for storage, and for the (future) analyses and information treatments.

The analyses to be considered are associated with business analytics, where advanced analytic techniques operate on big data sets. The Big Data analytics is really

about two things - big data and analytics - plus how the two have teamed up to generate today one of the most profound emerging trends in Business Intelligence (BI) [81].

In this Chapter we will not focus in the extraction of data from the internal sources of the hotel (Data Warehouse, Property Management System (PMS), etc.), we will focus only the extraction of information from an external source – the web. For the automatic collection of information from the web (a), a set of crawlers [18, 72, 75] must run periodically in order to produce suitable data [61], nevertheless not all the data that is extracted can be used in all hospitality business models, and from different sites (*Booking.com*, *Expedia*, *TripAdvisor*, etc.) it is possible to extracted different and coincident information from the same hotel.

In the SRM project, a different crawler was used for each site: *Booking.com*, *Expedia*, *TripAdvisor*, etc.; for more details see [61]. The crawler extracts periodically all the information existing in each site about each hotel, over different periods of time, and considering different types of users (2 adults, 1 adult and a kid, etc.), which generates an huge amount of "raw" data, that needs to be stored [61].

Related to the storage (b), means getting and store a high volume and data variety at high speed. To store this information it is usually necessary dynamic storage databases, the one chosen was the MongoDB database [61, 78], which is a NoSQL document-oriented database, structured as a set of collections that store documents, it also presents high performance, high reliability, easy scalability and map-reduce, etc.

The last stage (c) consists on the combination and selection of the relevant information from (a) and (b) for the business, in general, is the constructing of the Data Warehouse [70]. Due to the different collections, the integrating, velocity, and the storing large volumes of data coming from the GDS, OTA, internal (DW, PMS), etc. in reality it is a Big Data Warehouse [35, 65]. It is also necessary to consider data models tailored to the needs of the organization, both in terms of features to consider and in terms of information storage structure; as well as semantic concepts [21] to perform a

suitable data storage, according to the structure defined. Another important aspect is the information stored in (b), and semantically analyzed to store in the BDW (c) must include the social networks, that define the online reputation (OR) of a product or organization, to develop personalized recommendations and address various customer purchasing behavior [70].

As already mentioned, to access and use the information considered as Big Data, it was contemplated a set of technologies, as for example a NoSQL database. However, the relational database (RDB) continues to be the more prevalent [70] data storage, which allows viewing of data from multiple formats and for different stakeholders, even the ones that their activity is not related to technology. In this sense (not only, as we will see along the text), it is necessary to integrate the information form the MongoDB database in a RDB database, that allows the storage of a collection of data and the access, management and information processing, where the different professionals are able to use and access the data, in a variety of formats.

To do the above transformation, it is necessary to use data models to transform unstructured information in structured, i.e., in relational database models (RDBM). In situations where it is not possible to structure the data present in the NoSQL DB in a RDB is necessary to consider the concepts of semantics for a suitable data processing and conversion, and only later the storage in an appropriate structure.

3.3 Big Data Warehouse for SRM

The first phase of the generic architecture of the framework is presented in Fig. 3.1 (for the second phase see Sec. 3.4, as well as for the explanation of the "..." appearing in Fig. 3.1), from (a) the extraction of "all" the information available in the web, by the web crawlers [61], to (b) the storing of all that raw data in a primary database – MongoDB [61]. (c) The creation of semantic models, lexical databases, data models, rules and principles to select and combine the relevant information (in this case for

the Revenue Management), and store this information in a secondary database (RDB). The final integration of these three steps (with the ones presented in Sec. 3.4, Fig. 3.9) is the Big Data Warehouse for SRM. Again, we call the attention that in this article, we do not integrate the information from internal sources of each hotel, but in the SRM project, they are being considered (see Fig. 3.9).

In this Chapter we will focus on (c) – the last stage in the implementation of the Big Data Warehouse, being already presented stages (a) and (b) in [61], nevertheless, for the better comprehension of the following Sections it is necessary a brief explanation and examples about stage (b).

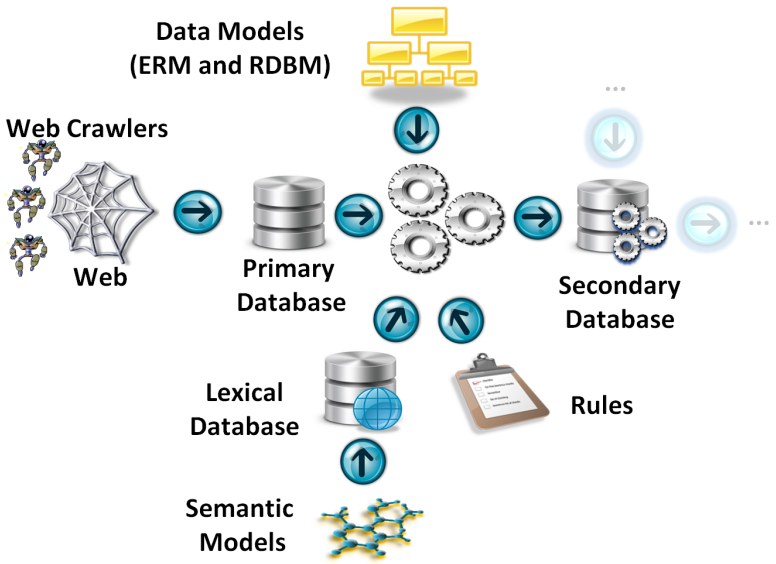


Figure 3.1: Web extraction, selection and conversion of information for the SRM Big Data Warehouse; see text, and Sec. 3.4 for the "..." explanation.

3.3.1 Data Models

As already mentioned, the extracted raw data from the different web crawlers was stored in a MongoDB. Figure 3.2 presents one example of information retrieved form one site (*Expedia*), belonging to the collection Room (see the remaining collections, and details in [61], extracted from a specific hotel at a particular date. Different sites (*Book-*

ing.com, *Expedia*, etc.) presents similar and different information extracted about the same topic [61]. In addition, with the structure presented in Fig. 3.2 it is not possible to make analyses of relationship with other data, for example the price, nor reading the information is intuitive.

To overcome this problem it was considered a model entity-relationship [27], which allows describing reality in terms of a collection of objects and the interaction between them. Taking into account the information presented in Fig. 3.2 and the concepts of entity-relationship model (ERM) was conducted the analysis of the information system, and has been defined the respective data model, (whose result is presented in a small part in Fig. 3.3).

Figure 3.3 shows the association between Rooms and Hotel, where the "..." represents generically other related entities with the hotel and for which is also being collected information. The entity Rooms have some attributes represented in the figure. Namely, RoomName, NewPrice (price with discount), OldPrice (price without discount), NumberOfAdults (number of adults that can be considered to book the room), NumberOfChildren (number of children that can be considered to book the room), and "..." which represent the others characteristics that are also relevant, but aren't represented in the figure.

The next step is to transform the ERM in a structure that it is possible to implement in a RDB. After the analysis, we considered the design of the system, and transform the ERM in a RDBM [29],considering the concepts associated with this data model, where an elementary object will be a table and the associationbetween them will be transformed by specific rules.

The result that ending the conception of an information system, is designated by the specification of the systems and is concretized bythe data model to implementin the database system considered, as presented in the Fig. 3.4.

In the end of the information system conceptualization, the data model includes the tables to create and the relationship to consider between them. In Fig. 3.4, the table

```
{
   "_id" : ObjectId("5423f668703c3b04260f0585"),
   "_idHotel" : ObjectId("5423f659a563ee1338ba3484"),
   "Source" : "expedia.ie",
   "ExtractionDate" : ISODate("2014-09-25T11:02:59.005Z"),
   "Search" : {
      "Location" : "       , Portugal",
      "NumberOfAdults" : 2,
      "NumberOfChildren" : 0,
      "NumberOfRooms" : 1,
      "CheckinDelayNights" : 0,
      "DifferenceBetweenCheckinCheckout" : 1,
      "CheckinDate" : ISODate("2014-09-25T11:02:59.005Z"),
      "CheckoutDate" : ISODate("2014-09-26T11:02:59.005Z")
   },
   "RoomName" : "Apartment, 2 Bedrooms",
   "Description" : [
      {
         "Title" : "paragraph-hack",
         "Content" : "1 queen and 2 single\r\nThis room opens to a
furnished balcony. The Select Comfort bed and pillow … This room
is Non-Smoking."
      }
   ],
   "TariffList" : [
      {
         "Conditions" : [
            " FREE Valet Parking",
            "FREE Cancellation before Mon, 13 Oct"
         ],
         "Tax" : [],
         "MaxOccupancy" : [
            {
               "Title" : "max-occupancy",
               "Content" : "Max Occupancy: 4 guests (up to 3
children, 2 infants)"
            }
         ],
         "OldPrice" : {
            "_t" : "TitleValue",
            "Title" : "€",
            "Value" : 16111
         },
         "NewPrice" : {
            "_t" : "TitleValue",
            "Title" : "€",
            "Value" : 14500
         }
      }
   ]
}
```


Figure 3.2: Example of the information about the collection Rooms extracted from *Expedia*, stored in the MongoDB.

Rooms represents the entity Rooms in Fig. 3.3 and the fields that belong to the table, in Fig. 3.4, are corresponding to the attributes of the Rooms entity in the Fig. 3.3.

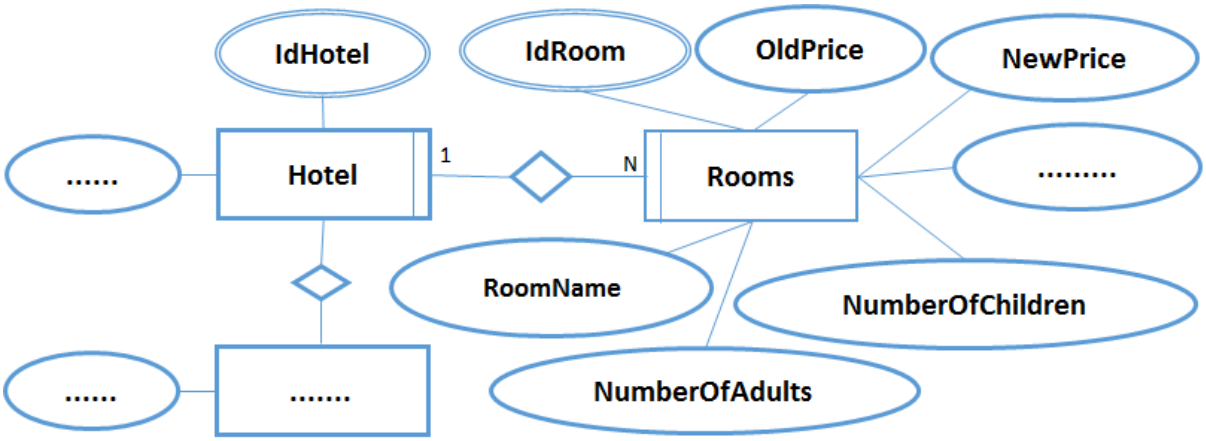


Figure 3.3: Excerpt of the ERM.

The next step is the development of the RDB, or also called the secondary database. In this database is where we will deposit the data collected from the MongoDB, the primary database, according to the logic structure defined by the data model, presented in the Fig. 3.4. In the data transformation from a NoSQL database to a RDB there are some challenges that the application developer has to face, such as the insertion of the adequate data in the appropriate fields.

For example, and returning again to Fig. 3.2, considering the information that is formatted in bold, it is possible to see that the content to be inserted in the field NumberOfAdults is 2, the number of children that is permitted in the room is zero, the price with discounts is 145.00 euros (in Fig. 3.2 shown as an integer), which is the content of the field NewPrice.

But there are other cases, that aren't so trivial, for example, consider that we have a table At your hotel which store the information about the hotel features (property features), for example the number of restaurants, the number of swimming pools, among others considered relevant to the business; see an example in Fig. 3.5, formatted in bold. To insert these attributes as indicated before, the application developer have to consider the concepts of semantic to find the right information, in the content of MongoDB, so they can be insert in the appropriate field of the RBDM.

Again, we call the attention that even the same information, e.g., the number of

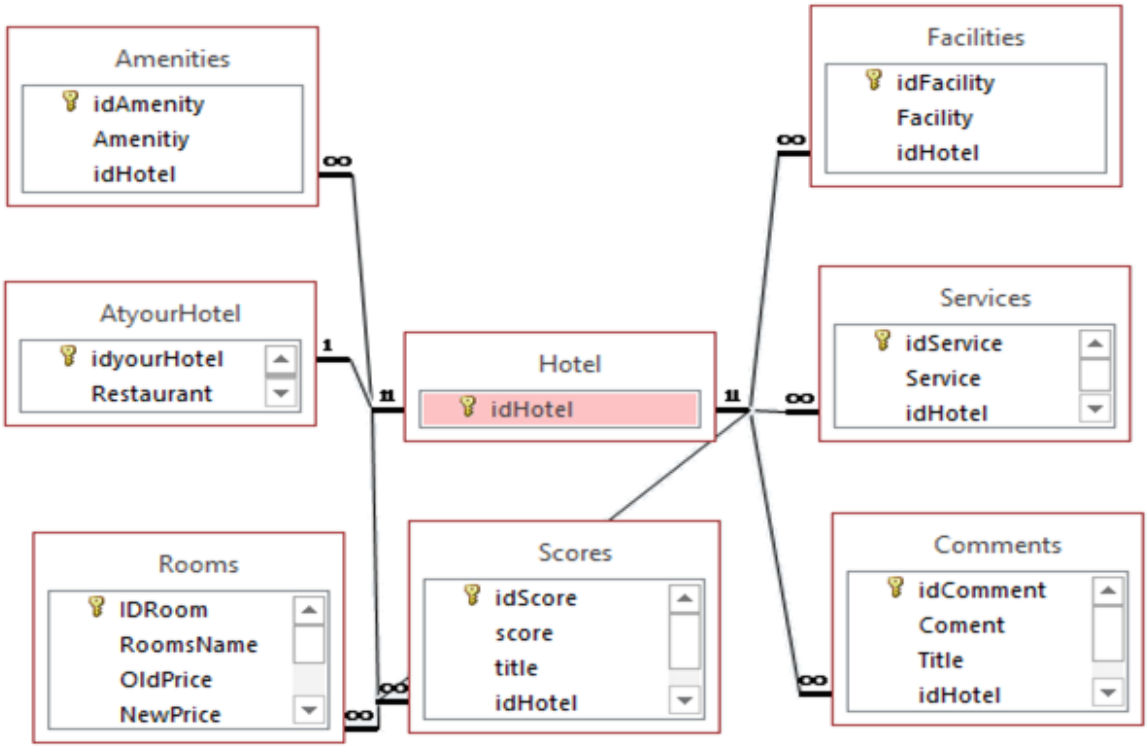


Figure 3.4: Excerpt of the RBDM.

stars of a hotel was retrieved from different forms (text, image, image captions) by the web crawler in each site, and even in the same site, it changes along the time, see details in [61]. Nevertheless, in the field Stars in the collection AboutHotel a number will be available.

For some type of data, during the transformation of the data stored in the MongoDB to the RDB, it will be necessary to map the fields of the first DB into the fields of the second DB. These mappings will not be direct and straight because there is no normalization in the notation used by the different producers of information for the web. For example, at the date this article was written, *Booking.com* uses Review Score from 0-10, and Score Breakdown in 7 fields, the *Expedia* shows ReviewScore from 0-5, and Score Breakdown in 4 fields.

On the other hand, as it is known, live languages consist of phrases and words with multiple meanings of difficult understanding for computational systems. Also, the utilization of plurals instead of singulars can worsen this problem. For the inter-

{
"_id" : ObjectId("5423f65b703c3b04260f0584"),
"_idHotel" : ObjectId("5423f659a563ee1338ba3484"),
"Source" : "expedia.ie",
……..
{
"Title" : "At your hotel",
"Content" : " features a full-service spa, **3 outdoor swimming pools**, an outdoor tennis court, ... room(s)\r\nMeeting rooms\r\nMultilingual staff\r\nFree valet parking\r\nPoolside bar\r\nBeach bar\r\nPorter/bellhop\r\nLuggage storage\r\nArea shuttle (surcharge)\r\n**Number of restaurants - 2**\r\nConference center\r\nOutdoor tennis court\r\nSauna\r\nSpa services on site\r\nSteam room\r\nWedding services\r\nHair salon\r\nBeach/pool umbrellas\r\nTours/ticket assistance\r\nWireless Internet access - surcharge\r\nWired (high-speed) Internet access - surcharge\r\nNumber of outdoor pools - 3\r\nChildren's club\r\nRoom service … tub\r\nSupervised childcare/activities\r\nChildren's pool\r\nIndoor pool\r\nHide"
}
…..
}

Figure 3.5: Example of the information about the hotel property stored in the MongoDB, extracted from *Expedia*.

pretation of the meaning of a sentence by a computational application, we need more than a dictionary because language is polysemic, i.e., the same word or phrase can acquire various meanings according to the context in which it operates.

3.3.2 Lexical database, semantics and ontology

A lexical database as the WordNet, developed by Princeton University (WN.Pr) [38, 64] as a Natural Language Processing (NLP) application, can help to interpret the meaning of the sentences (see also Sec. 3.4.1). Lexical information is not organized in word forms, but in word meanings, which is consistent with the human representations of meaning, and their processing in the brain.

As mentioned, there is no normalization of the information used and displayed in

the websites, it can happen that two websites as e.g., *Booking.com* and *Tripadvisor*, use different words to designate the same facilities or amenities. For example, they can use different words to designate the same type of room. This can be solved using the mentioned lexical database, which will be responsible for the mappings between the MongoDB and the RDB, and taking in consideration the semantic web concepts; see the structure in Fig. 3.1.

A different support can come from the semantic web [21]. In the semantic web, the organization ofthe pages structure is different from the current web, as shown in Fig. 3.6. In the semantic web, the structure consists of software, documents, libraries, images, concepts and people, in the case of current web, each document provides hyperlinks to other documents, which may, or may not be linked between them. The semantic web seeks to understand the meaning more than the content present on the page [96], in order to identify the existing knowledge on the web through in a way that is understandable to all (canonical form, that is, in its simplest form).

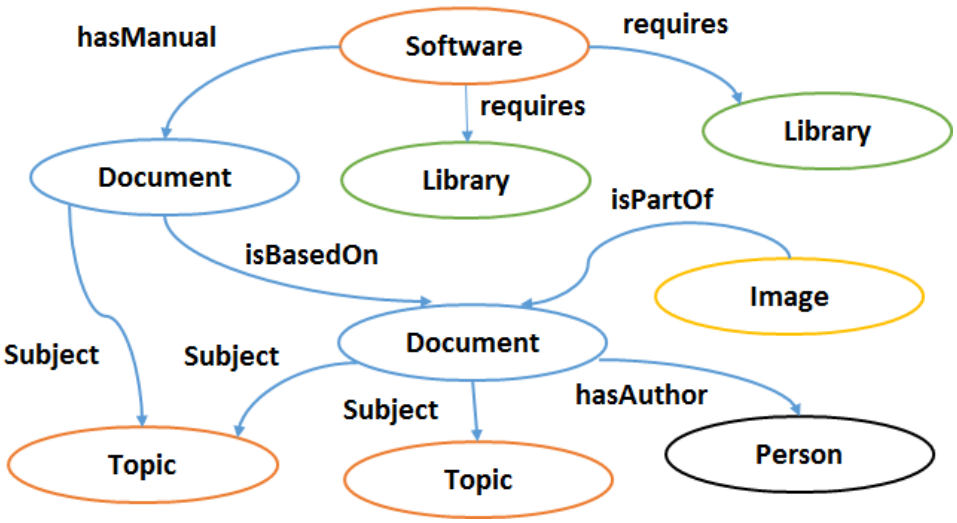


Figure 3.6: Semantic web, example of the organization of the page structure.

According to W3C [89], the goal of the semantic web is to create a universal medium for data exchange. In the semantic web environment requires the ability to represent and manage the content on the web in the form semantics, i.e., allow an agent to learn the meaning of a term by appointment of a formalization of terms based on metadata,

ontologies, or other concepts considered to generate knowledge. Moreover, the extraction of information from a collection of documents has to be done in order to satisfy the needs of the user. This extraction is made in documents written in natural language, stored, represented and organized in different types of systems [33], as presented in the Fig. 3.7 (see also [20]).

An ontology consists of a set of classes, relations, instances and axioms, where the classes represent concepts that belong to a domain which describes the ontology relationships and represent the association between the elements of the ontology, the instances are used to represent a particular element of the class and finally, the axioms are assumed to be true statements [33]. The layer of ontologies is one of the most important because it is responsible for providing the necessary expressiveness to the representation of ontologies. An example of a multilingual ontology for the hospitality sector can be seen in [59]. Figure 3.8 presents a small extract of one ontology related with tourism. In the figure, the term "is a" means that the name "is part of", for example, "Cultural" is part of the "Tourism", and so on.

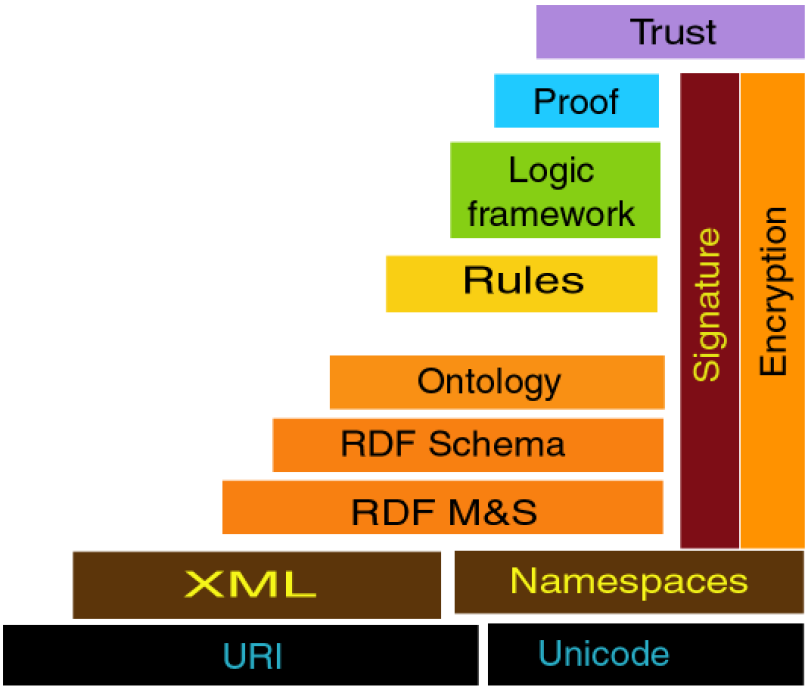


Figure 3.7: Semantic web layers [20].

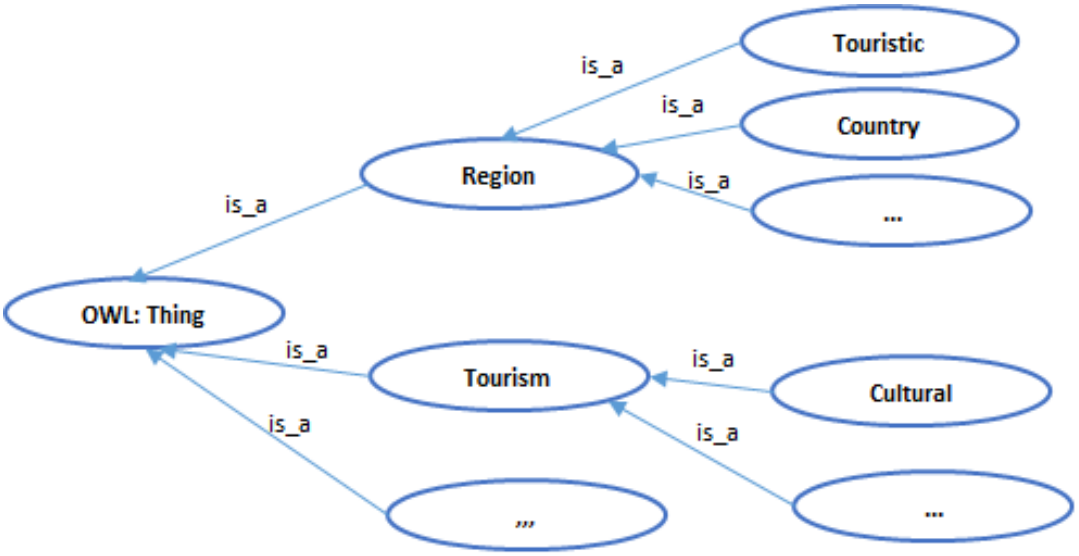


Figure 3.8: Tourism ontology example.

The linguistic ontologies are referenced by their application in natural language processing systems. To work with ontologies are used designated languages for Web Ontology Language (OWL). There are different types of languages, recommended by the World Wide Web Consortium (W3C) [89] to work with different levels of semantic expressivity, as for example the OWL Lite, OWL DL, and OWL Full. In addition, the semantic web will make it possible to find information in the extracted data from the web.

However, when the goal is to extract information from the collected data, stored in the Big Data Warehouse, it is necessary to consider other features.

3.4 Extracting information for BI

In the above Sections (see also Fig. 3.1), we show the path of the information from the source, in this case the web, to the secondary database - Big Data Warehouse. In this Section, we concentrate on complementing the information in the secondary database and the extracting of information from the BDW for the BI. To extract information for the Big Data Warehouse, the use of traditional methods of analysis are no longer

adequate [70], making it is necessary to consider new tools, some also described in Sec. 3.3.2.

Figure 3.9 shows the second phase of the generic architecture of the framework, i.e., represents the process to extract knowledge from the data, that is relevant to control and manage the organization and to support the decision maker in the context of Business Intelligence. In the secondary DB (or the SRM Big Data Warehouse), as already mention, and we reinforce the subject, it is necessary to include the Data Warehouse of each hotel (were the SRM will be applied), and use these information system, with the forecasting and RM models most adequate to the business model of each hotel.

With all these data, and with analytic models, and sentiment analysis, an analytic background is created, one, which will make possible for the decision maker to have access to the analytical tools that will facilitate the creation of the business intelligence environment, such as reporting, forecasting and cubing for data analysis.

The analytical tools can be considered for the development of thematic or segmented subsets of the Big Data Warehouse, called Data Marts (DM) to analyze and manage specific areas, such as RM, or Online Reputation or the Customer Relationship Management (CRM) as represented in the Fig. 3.9.

Business Intelligence is a way to identify new opportunities and implementing an effective strategy based on insights, or intelligence, it can offer to the business a competitive intelligence that give a market advantage and a long term stability [80]. The competitive intelligence is developed by considering a set of techniques and tools for the transformation of data into meaningful and useful information for business analysis purposes [80].

In the Business Intelligence process, it is necessary to take in consideration two steps: (a) the extraction of knowledge and the (b) assessment of the intelligence extracted from that knowledge.

- **Knowledge extraction** - a huge information collected is relevant, but it is necessary to use adequate tools to produce knowledge about the organization. The

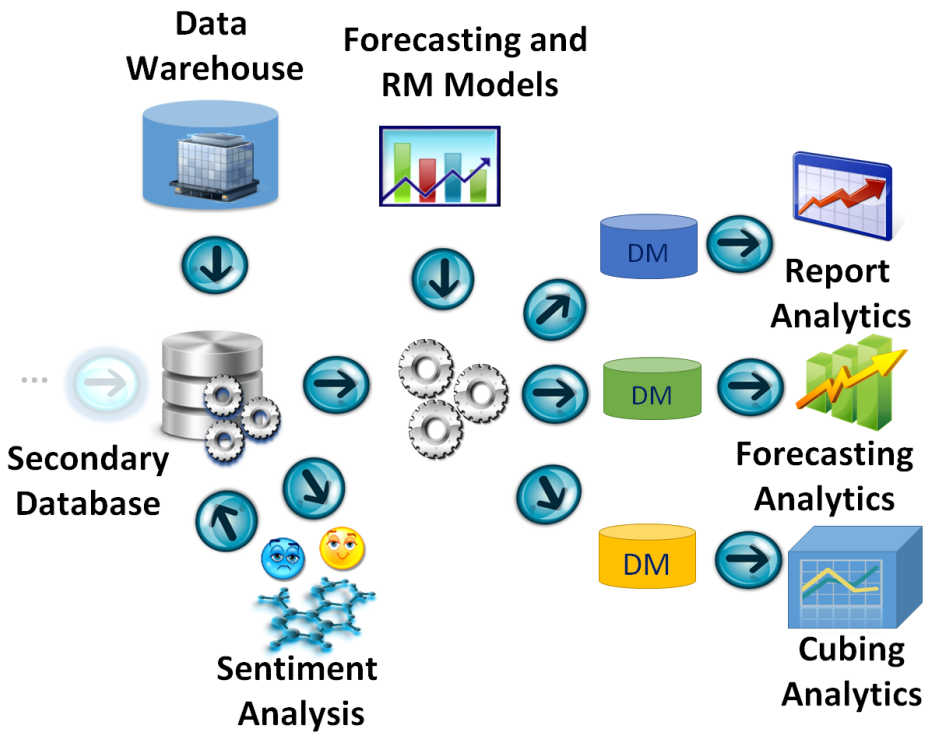


Figure 3.9: Extracting information from SRM Big Data Warehouse for BI (second phase); see text, and Sec. 3.3, Fig. 3.1 for the "..." explanation.

reports and fixed dashboards [70] produced by the Data Warehouse are limited solutions compared with the results from the big data analytics, which can include ad hoc queries and discoveries of meaningful relationships between the data. Furthermore, to extract knowledge from the information stored it is necessary to include the data from the Data Warehouse, that have information about the transactional operations of organization, as presented in the Fig. 3.9, and to include predictive models to help to define the future behavior of the consumers.

- **Intelligence Assessment** - the possibility to extract business value for the organization have captivated several researchers and stakeholders [60, 63, 70, 74, 92]. There are several techniques that are actually considered in the big data environment [70]: (a) recommendation systems, for example in social networks when refers "People you may know likes de hotel X". (b) Analysis of social networks, to identify the influence over others. (c) Analysis of new products, to test new

products or ideas and obtain instant feedback. (d) Analyses competitors' pricings, to compare with their prices. (e) Sentiment analysis, which permit to define the costumer sentiment towards products, services, destinations, hotels; among others.

3.4.1 Sentiment analysis

The sentiment analysis or opinion mining techniques is probably the biggest challenge in the second phase of the SRM Big Data Warehouse, it comprises an area of NLP, computational linguistics and text mining [60], and refers to a set of techniques that deals with data about opinions and tries to obtain valuable information from them [57]. It is constituted by a group of computational techniques used to extract, sort, understand and evaluate the opinions expressed by users about products, services, destinations, cruise companies, hotels, among others; from textual sources. It can be used, for example, to understand the opinions of the hotel clients or product consumers. The emergence of the semantics has created many opportunities to understand the views of the consumers on marketing campaigns and preference for products. Some of the concepts already presented in the Sec. 3.3.2 can be used to extract the characteristics and to identify the opinion associated with those characteristics, which may be positive or negative [57, 60].

The semantic is the key, not only one, to find information in the content of textual field in primary DB (MongoDB), but also to consider the customers opinion data (feedbacks) and apply a sentiment analysis to produce intelligence associated to the organization, which is a challenging subject ofinvestigation and with difficult implementation. One of reasons is because of the nature of the associated tourism information, whether referring to data of thehotels, transportation, or entertainment. Another reason, and most relevant, is identify the adequate methodology to apply the ontologies to the tourism and hospitality information. This difficulty is present in the storage of the information from the MongoDB to the secondary database, as in the analyses

referred before, with more impact in the sentiment analyses.

Nevertheless, there is at least one solution to solve this challenging problem; it is to consider a lexical database that can help to interpret the different meanings and to find the synonyms of words. WordNet, [9, 11, 38, 64] (see also Sec. 3.3.2) can be seen as a "dictionary of meaning," integrating the functions of a dictionary and a thesaurus. As the data extracted by the web crawler can be in different languages: English, Portuguese or Spanish, the adoption of the various adjustments to the WordNet to other languages can be one of the approaches. The other one could be to translate all the data extracted by the web crawler, independently of the language in which it is, to the same language, for instance, to English. By default, the web crawler searches the information in English, nevertheless users comments can appear in several languages.

If the decision is not to translate, in Global Wordnet Association Website [11], there is information about the languages, the name of the resources and the type of license on the various adaptations available worldwide. For instance, adaptations to the Portuguese are three, two for the Portuguese of Portugal (Onto.PT and WordNet.PT) and one for Portuguese of Brazil (OpenWN-EN) [4, 68].

In WordNet, nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms, called synsets, each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. Nouns can be connected through hypernyms/hyponyms and meronyms/holonyms relationships; verbs are organized by troponym/hyperrnym and entailment relations; while adjectives are linked to their antonyms, and relational adjectives point to their related nouns. Finally, adverbs mostly derived from adjectives and are linked to them via a pertainym relation.

There are several API for WordNet, as for Java, C# and Python [3, 5, 10]. One of them that is particularly well known is the Java API for WordNet Searching (JAWS) [2], which is an API that provides Java applications with the ability to retrieve data from the WordNet database. Futhermore, for each synset, WordNet shows the several re-

lationships. Hypernyms are the synsets that are more general and the hyponyms are the synsets that are more specific. The synsets and the relationship hypernyms/hyponyms are the principal relationships that will be expected to be used for the SRM project.

However, there are others types of relationships. The holonyms/meronyms relationship is one of them, where the holonyms are used to denote a whole and the meronyms are used to denote things that are a part of something. For example, given: "floor," "wall" and "room light" were some of the meronyms found and "building" and "edifice" were the holonyms reached.

The WordNet 3.0 contains 155.287 distinct words, distributed in 117.659 synsets, resulting in 206.941 pairs of word/meaning [9, 11].

In resume, WordNet and the other adaptations of this lexical database to other languages are popular NLP applications, which allow disambiguating senses of words, measuring their relatedness to others and defining and describing their meaning. In this research, these lexical databases are used in the normalization of the data during the transformation of the primary DB into the secondary DB. However, WordNet cannot be used for a complete semantic analysis of a text or corpus, which may require detecting and processing sentiments. To do this, sentiment analysis or opinion mining can solve this problem.

3.5 Discussion

The Big Data collected about the environment that surrounds an organization, mainly in the hospitality industry, and apply to them big data analytic tools is a powerful way to support the decision maker and t control the organization.

Big Data are mainly velocity, volume and variety [70], the huge amounts of data, collected from different sources and high velocity will, in conjunction with the data of the organization itself, constitute the Big Data Warehouse. It is a necessary informa-

tion system once the travel business is in constant change, and the stakeholders need to visualize the information business in real time, to detect urgent situations and automate with immediate answers [50]. For example, with new policies applied to the rooms, dynamic pricing, taking in consideration the competitive set of the hotel.

With the concepts associated with the development of an information system it is possible to develop and implement a Big Data Warehouse. However, the traditional Data Warehouse as to complemented with external sources, for that it is necessary to take in considerations some technologies such as: web crawlers and NoSQL databases. By other side and in the context of the present work, the concepts and techniques of semantics have to be included in the system to overcome the problems that are founded in the creation of a system with this dimension, this variety, and this quick analytics tools. The WordNet was considered as a semantic tool to solve some limitations; however, this solution is not fully adequate to a sentiment analysis or opinion mining.

This work is an asset to the hotel managers and marketers and tourism stakeholders, as it features a set of stages and intermediate phases and steps necessary for creating a Big Data Warehouse and the development of big data analytic tools whose capabilities for managers and for business intelligence are obvious and go with the current trend in the society.

In terms of future work, consist in the completing the development of the application/software, once part is already under development, and presenting promising results (see [61]), and solve the limitations regarding the sentiment analysis, that will be addressed and developed through the use of concepts and semantic techniques.

4

Challenges in Building a Big Data Warehouse Applied to the Hotel Business Intelligence

Chapter Outline

All the tourism stakeholders, mainly the hoteliers, need a state of the art online information system to reply to the customers searches with the necessary and updated information. On the other hand, the hoteliers also need information about their organization competitive set, which implies having access to information about the clients,

competitors, and all the stakeholders associated to the hospitality activity. To satisfy the hotels and consumer needs, it is essential to have access to a Business Intelligence (BI) system, which consolidates all the relevant data to be used by the analytical tools. In this sense, BI systems have some challenges related to the definition of an adequate methodology to integrate and store the retrieved data into a hospitality Big Data Warehouse. This Chapter presents the challenges and some of the necessary steps to overcome the problems associated with the information management and consolidation in a hotel Big Data Warehouse.

4.1 Introduction

Common travelers plan their vacations or travels using the Internet to search for information about tourism products that they intend to consume, such as accommodation, transportation and entertainment. In addition, they search for information about other traveler's opinions, to know if they had a good experience in the destinations, they intend to visit. Therefore, it is common to have travelers making their decisions, about what they want to experiment in their holydays, using the information they have access to in the internet. In resume, they will buy according to their preferences and the opinions of others travelers [55].

In this environment, for all the tourism professionals, in particular for hoteliers, marketers and organization managers, it is relevant to have access to analytical information of what are the traveler's commentaries, for instance, which commentaries emerge in the internet, and which were more considered by the consumer in the moment of their decision. These commentaries become part of the online reputation of the tourism destination. For the hotel, the online reputation and the travelers who produce the commentaries are considered as hotel representatives or hotel brand agents [84, 85].

For the hotel decision makers, the access to a technological analytic tool that collects

the online information about their business and their competitor (which integrates analyses to support the decision maker) is mandatory. The need to define strategic actions to create new values or increase their competitiveness, can make the difference between success and failure, once the analytic tool can achieve new ideas to implement new actions and create new values to the consumers [13].

However, the creation of tools with these characteristics, which will coexist in a technological architecture, have some challenges to overcome. One of those challenges is the definition of the procedure to store and manage the data collected from different web sources [74]. Maybe even more important, it is the information consolidation in a Big Data Warehouse (BDW), in order to be used by the analytical components in a context of Business Intelligence (BI) and support the different kinds of management in a hotel business [77]. In hotel revenue management activities, which is supported on the manager's experience and available information, must be constantly updated. These challenges assume a great relevance due to their nature.

In addition, in a BDW some tasks must be performed: (i) integration of the information collected from different internet sources. (ii) Update and maintenance of the new information gathered from the internet (once the information about accommodation is constantly changing), e.g., the rooms prices or the promotions offered in according with the customer profile. (iii) Maintenance of the historical data, which will be relevant to the hotel big data analytics, in order to achieve the business intelligence systems requirements [77].

Based on a primary database [61, 77] founded by data collected by WebCrawlers from different web sources (such as *Expedia* or *Booking.com*), this Chapter presents the challenges and the necessary steps to overcome the problems associated with the definition of an adequate structure and suitable process of information consolidation in a BDW, here designated as a secondary database (see also [77]). The collected data is then to be used by the analytical tools to fulfil the needs of a Hotel Business Intelligence Manager.

The article is structured as follows: after the introduction, the second section presents the contextualization of the subject of the study. The third section highlights the procedure to the consolidation of the extracted data in a BDW. Finally, we will present some conclusions and suggestion for future work.

4.2 Contextualization

In the area of tourism and hospitality, the volume of information associated with tourism activity, to represent all the different activities related to the tourism sector, is huge. When the tourists plan their travels they search for information about aspects that are related with their preferences, as well as the opinions about the same touristic destinations made by others travelers. From the moment when the destination is chosen and the purchase occur, they are buying a product based on the information mainly available in the web [71]. The relevant touristic information that exists on the internet is updated constantly and has to be managed several times a day to reflect the business needs and the consumer needs.

For the organizations, mainly those related with hospitality, it is important to have access to the information associated with their business, both in the supplier and in terms of demand perspective. This information allows to better understand and analyze the entire environment that surround the organization, including the customer's preferences, online reputation, business trends, among others.

These organization environments (on the internet) are associated with the hotel activity and are characterized by a big volume of data [35, 66]. The processed data comes from different sources, with great diversity, in an unstructured format. In this context, nowadays, the concepts associated with Big Data have to be considered by the organizations in general [77], and have higher relevance to the ones whose business is situated in the hospitality activity.

For the hospitality, the stakeholders need an information systems, which provide

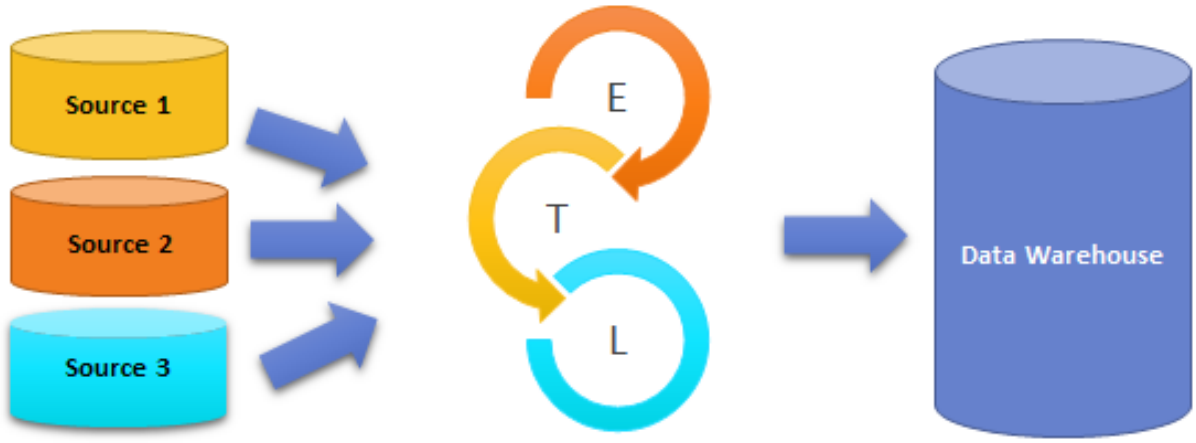


Figure 4.1: Extraction, Transformation and Load Data in a Business Intelligence System.

the right information to the right receiver in the right volume and quality, at the right time [83]. Furthermore, to the managers and marketers, to have access to information with these characteristics, can make the difference in their ability to increase the competitiveness and ensure their survival, in a society that all days emerge new competitors and new trends that influence the consumer and offer new values that satisfy their preferences and needs.

It is therefore important and necessary to consider analytical tools that integrate historical analysis from several years, with functions to support decision-making in terms of management and in terms of strategy. Those tools permit the construction and development of Key Performance Indicators (KPI) [31, 48, 73] for the hoteliers and marketers.

For all the touristic organizations, and in particular to the ones related to the hospitality activities, the knowledge management and the BI are the areas that contributes to improve the quality and quantity of information. With that information on hand, it is possible to increase the business and organization performance, supported on decisions that are more accurate and to define expert strategically plans to the organization [82].

The BI systems are supported on a set of phases that includes several kinds of

technology and concepts, namely: (i) Data Integration, (ii) Data Warehousing, (iii) Online Analytical Processing Cubes, (iv) Data Mining methods, and (v) Analytic Tools.

- **(i) Data Integration**, is constituted by the Extraction, Transformation and Load (ETL) process that integrate the data that exist in different sources in a Data Warehouse [40], with all the information that is pertinent to the organization, from internal and external sources (see Fig. 4.1, and also [77]). After the identification of the several sources (e.g., search engines), the data is selected and extracted in the extraction process. The following step is the data transformation, which include tasks such as the cleaning and standardization of data, among others that contributes to the integration of all data that is relevant to consider in the data warehouse to support the analytical tools. The last task, the load, is related with the storing and refreshment of the data in the Data Warehouse.

- **(ii) Data Warehousing**, includes the technology to manage and store the data in a "Data Warehouse". In some cases, business decision makers can also consider the use of Data Marts which are databases constituted by an organization subset of the data, generally related with a department or an activity, which can be independent or dependent of the Data Warehouse [82]. This division allows to fulfil the needs of the organization, taking in consideration the adequate information structure multidimensional model.

- **(iii) Online Analytical Processing (OLAP)** [40, 46], permits the creation of cubes to explore the information in the Data Warehouse, or in the Data Marts (see Fig. 4.2). These technologies allow to analyze the information on different business perspectives (dimensions).

- **(iv) Data Mining methods**, is optional on the Business Intelligence Systems and can be used together with the OLAP cubes. The Data Mining methods consist in the application of artificial intelligence algorithms to discover knowledge in the historical data and, at the same time, to make forecasting to different areas

or activities in the organization. The data mining tasks can be divide in two: descriptive and predictive. The descriptive task is considered to identify rules, which characterizes the data, and includes several techniques, such as Clustering and Summarization. The predictive task is pertinent to identify new models that define a variable behavior, which can be used to estimate the future variable values [82].

- **(v) Analytic Tools**, it allows the analytical investigation about the organization data producing enterprise reporting (also called management reporting) or dashboards, which may take the form of graphics, text and tables. The outcomes of the analytical tools can include results from data mining tasks, interactive queries, key performance indicators, cubes, balanced scorecards, forecasting methods, among others [45, 82].

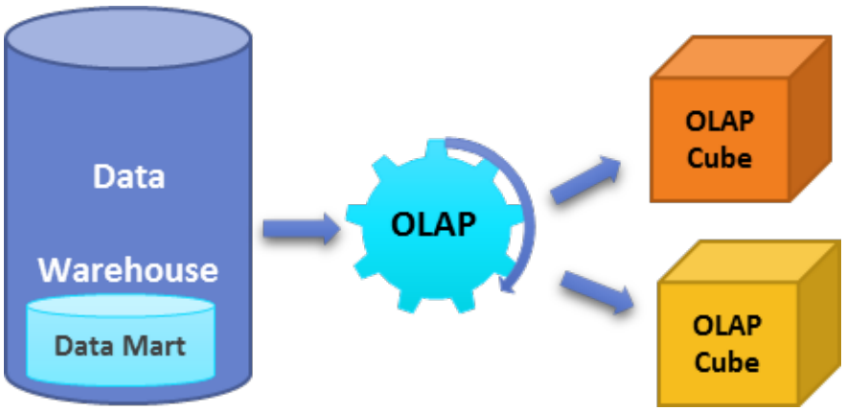


Figure 4.2: Online Analytical Processing in a Business Intelligence System.

However, all of these kinds of technology that coexist in this kind of architecture have a main challenge related to the definition of an adequate methodology to integrate, consolidate and store the data in a BDW [88], which can be relevant to be used by the analytical tools to achieve the requirements of a hospitality business intelligence system.

In the hospitality activity, this challenge assumes a great relevance due to the nature of this activity that is completely supported on information, as presented before. In

addition, to achieve the data consolidation, is necessary to find a suitable process to store and manage the data in a secondary database, the BDW. In this case, we are considering that a primary database is constituted, e.g., by the data collected by a WebCrawler from several sites in the web, such as *Booking.com*. A major problem arises from the fact that different sources have different structure and different meaning to the same hotel features [61]. For example, the "cleanliness" and "room cleanliness" appear in different sources.

In our investigation, the collection of data in the primary database (see Fig. 4.3) was collected by a web robot or crawler [61, 95] and stored in a NoSQL database, namely MongoDB [67]. The data is then consolidated in a posterior step by integrating it in a secondary database, a BDW [40, 77, 95], as presented above. It is the secondary database that the hotel big data analytics is done.

However, the ETL process in a hospitality BDW takes a higher scale of complexity once the data was collected from different unstructured web channels where, as already mentioned, data with the same meaning can be classified with different designations, which raises the need of consolidation of the extracted data to ensure that the information stored in our BDW is consistent and reliable to the hotel business.

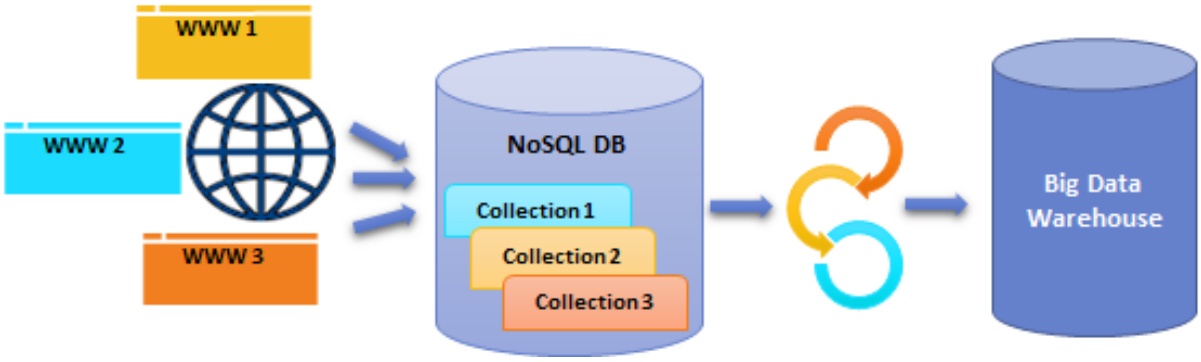


Figure 4.3: Data integration in a Big Data Warehouse.

4.3 Consolidation of extracted data

In a first phase, the data was extracted from different web channels and stored in a MongoDB database (DB1), the primary database. Four MongoDB collections were used: AboutHotel, Rooms, Comments, and Scores [61]. The collection designated by AboutHotel contains the hotels characterization, which includes information about hotel name, location, hotel features and rooms amenities. The collection Rooms has the information about the rooms and prices, namely including data about the room's name, and number of adults and children that are permitted in the room. In Comments are the data concerning the reviews of the hotels, which includes information about the customer segment and their country of origin. Finally, in the Score collection are the reviews that tourists have attributed to hotels which contributes to define the hotel online reputation.

The DB1 serves as an intermediate database between the web sites / web robots and the BDW, also called secondary database (DB2), which is a relational database. To create the DB2, it is necessary to apply a set of rules, namely: define the data forwarding rules, identify possible conversions, in order to make them readable data for a particular application. In the following sections will be addressed the techniques used for this purpose (see Fig. 4.4).

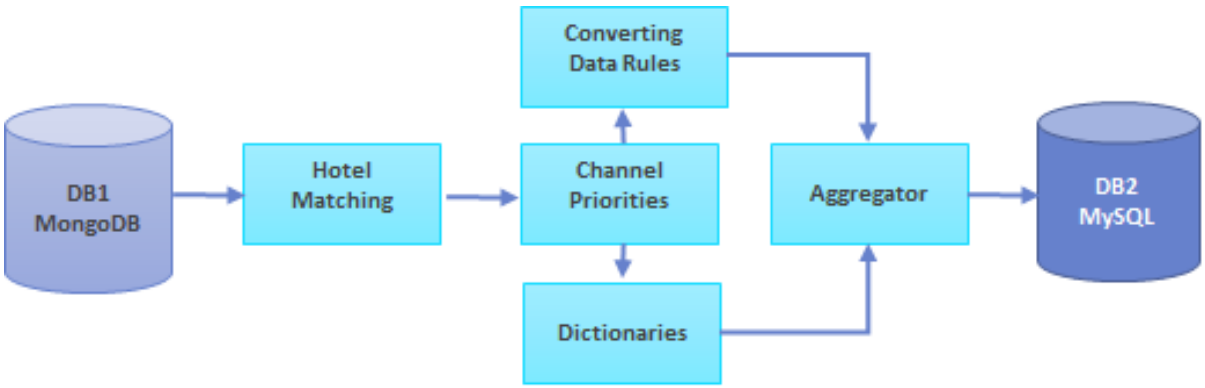


Figure 4.4: Diagram explaining the process for processing data to be consolidated.

4.3.1 Reading information from DB1

The first step of data consolidation is the reading of data from collections stored in DB1. To do that, it is necessary to know where the data is stored in DB1 and load it into the memory of the consolidation program. This data will be processed and converted, in an adequate way to be stored in the DB2, without the preoccupation of having an inappropriate elimination or changes in the original data extracted by web robots.

4.3.2 Data conversion rules

The vast majority of DB1 data is in string format. Although some fields are straight copies from DB1 to DB2, sometimes it is necessary to make some conversions, for instance to numerical values.

An example are the dates of the reviews for a particular hotel. For instance, on *Booking.com* dates appear as "22 March 2015", and therefore it is necessary convert the date to a valid format to store in DB2, i.e., in the format "22/05/2015". Another conversion that is necessary to do are the GPS coordinates of the hotels. These GPS coordinates were extracted from the web, in format "latitude, longitude" where latitude and longitude are decimal numbers. For this reason, the string need to be converted into two separate decimal fields.

4.3.3 Data dictionaries

As stated in Sec. 4.3.2, the most of the data stored in DB1 is free text, i.e., text written by a human, see [61]. To give meaning to those texts, the consolidation program should use data dictionaries.

These dictionaries are added manually by the user and updated each time the user finds a new synonym for a word. The dictionaries are stored in a MongoDB collection and are structured in four fields: Source, Type, Word and Alias (see Fig. 4.5). The field

Source indicates the channel (e.g., *Booking.com*) where that particular dictionary should be consulted. The Type indicates the context in which particular dictionary should be consulted. For instance "Type": "Amenities" means that this dictionary should be consulted when looking for the hotel amenities in a sentence. The Word field concerns the word that is being searched. Finally, the Alias field is a list of synonyms of the word that is necessary to find.

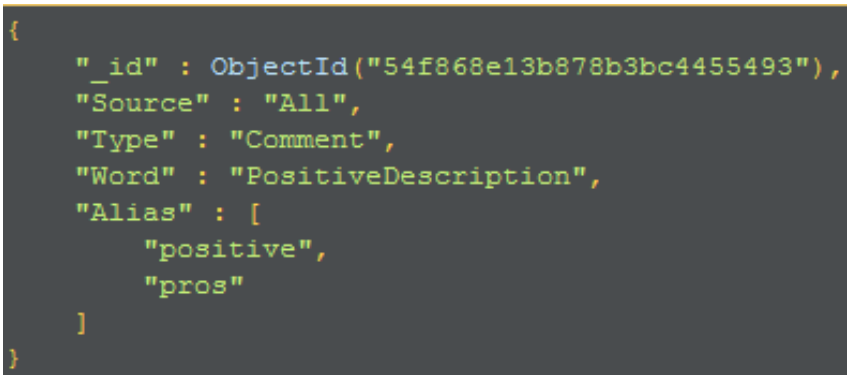


Figure 4.5: Example of a MongoDB (JSON) document related to "comments".

In the hotel reviews, each channel has different forms to display the customer's opinions, and do it differently from each other. There are channels in which the comments are divided into "positive" and "negative", and there are others where it is just free text. The comments where there is no distinction between positive and negative are stored directly in DB2 without any treatment. Case the channels contains comments separate in positive and negative, it is necessary to refer to data dictionaries in order to distinguish between a positive and a negative review. This is necessary due to the different designations that each channel gives to the positive and negative comments. For example, in *Booking.com* the terms are "positive" and "negative", while in *Expedia* are "pros" and "cons" (see Fig. 4.6).

4.3.4 Correspondence between extracted hotels indifferent channels

As already mentioned, the web robots are concerned to extract data from web, but not to process/analyse it. The same hotel extracted from different sources is not identified

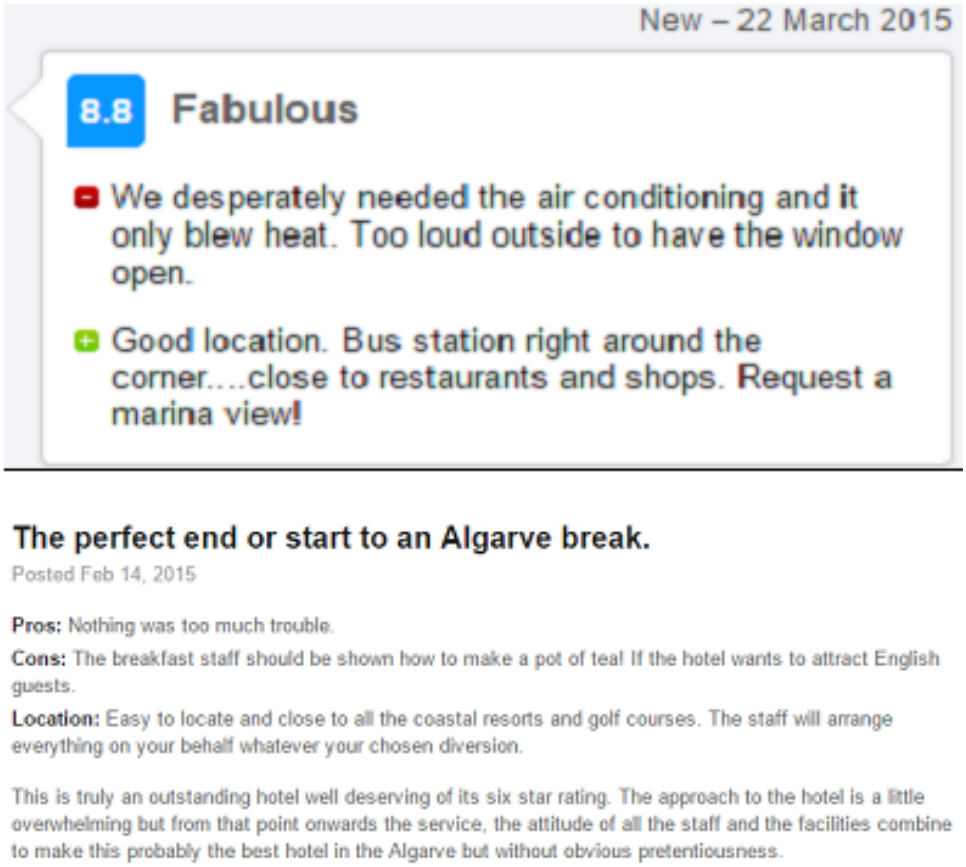


Figure 4.6: Example of review comments (*Booking.com* appears above and *Expedia* appears below).

as being the same hotel, i.e., the IDs of the documents of the databases are different.

In other words, when a web robot extracts the hotels, there is no warranty that the Hotel X, e.g., in *Booking.com* is the same as the Hotel X on *Expedia*. This happens because the names of the hotels can change slightly according to the channel. What happens is that the same hotel will get two different ID, depending on the channel from where its data was extracted. This is a serious problem, because it is essential to make a correct consolidation of the hotel data, for the other existing tables that depend on the table of the hotels.

To solve this problem was developed an algorithm that uses the name, address and hotel's GPS coordinates to match between hotels extracted from different channels. A function (MatchStrings) was developed to help the algorithm to verify the degree of similarity between two strings. This function detects the number of words in the same

strings and, also the ratio (range 0 to 1) versus the number of equal words in each string. The ratio equation is given by $R = 0.5 \times EQW \times (L1 + L2) / (L1 \times L2)$, where EQW is the number of equal words between the strings, and L1 and L2 are the number of words in each string, string1 and string2 respectively.

The hotel matching algorithm has the purpose of finding the correspondence between hotels in different channels and works as follows (see Fig. 4.7):

i. Search for hotels that have the same name and address of Hotel X. If true ($R > 0.99$), the hotel ID is found, go to vi). If the search returns empty then go to step (ii).

ii. Search hotels that have the same GPS coordinates of Hotel X (equal to the second decimal place). Only checks to the second decimal place between channels because the coordinates range slightly changes from the third decimal place, both latitude and longitude.

iii. Use the MatchStrings function to check the degree of similarity between the Hotel X, name and the names of each of the hotels resulting from the previous step, returning only hotels with a degree of similarity from a given threshold (e.g., $R > 0.5$).

iv. The same as step (iii) but now with the addresses of the hotels.

v. If the above steps do not result a match, then it is considered that the hotel was not yet acquired from any other channel. If there are hotels that match the search, then all of those hotels receive the same ID. Exceptions can still occurs, for instance two hotels in the same street one side with the other and with a very similar name. In those cases, both hotels ID will be the same. In those exceptions, it is impossible to know with absolute sure that the hotel is the same, or if they are two different hotels.

vi. Hotel ID defined.

4.3.5 Channels priority

Once the data is extracted from different channels, it happens that there are sometimes redundant data. For example, after the application of the MatchStrings algorithm (explained in Sec. 4.3.4), it is possible to obtain several distinct names for the same hotel.

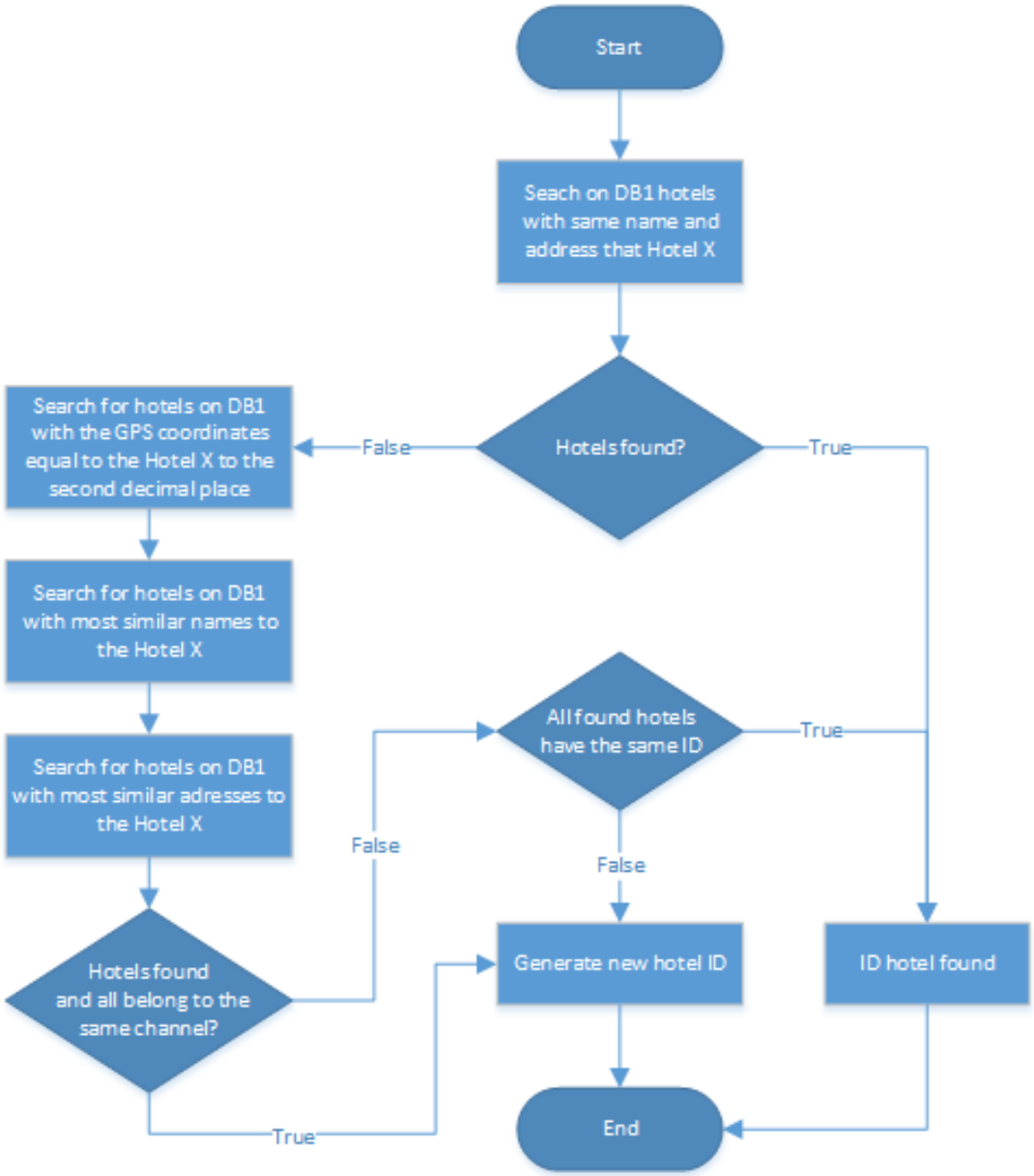


Figure 4.7: Flowchart of the algorithm to find hotel match in different channels. The "Hotel X" is the unknown hotel that we wish to find a correspondence.

In this case, the administrator will decide the name to select and stored in DB2. For this reason, it is necessary to define priorities to the channels, to withdraw and to

store the information in DB2. In other words, if a hotel exists in four different channels and the priority to the channel is set, in descending order, as channel 1 $\rightarrow$ channel 2 $\rightarrow$ channel 3 $\rightarrow$ channel 4, then the priority channel should be channel 1 and so the data stored will be from that channel. Another example in which these priorities must be defined is the number of stars of the hotels. Usually there is no consensus on the number of hotel's stars from channel to channel, i.e., an hotel that has four stars on a channel can have three stars in another channel. Then, once again, the rule of priority to decide which channel is more reliable is applied.

This mechanism will only be applied in cases where there are evidence of ambiguities. In the case of the comments, for example, this is not necessary since every review is a different opinion and all comments must be stored, even if there are two identical comments, what matters is that they are two different people.

4.3.6 Routing rules / data flow

After the extracted data from DB1 were transformed by all conversion rules, dictionaries, correspondence and channel priorities it is necessary route them to the right place. Once there are some tables in DB2, which have relations of many-to-many and for these cases, it is necessary to comply with a data storage order.

For instance, consider the three tables: Hotel, Amenities and HotelAmenities, where Hotel and Amenities are tables, which are related to each other by a many-to-many relation, supported on the HotelAmenities table. To store the data in these tables, first, the data are stored in Hotel and Amenities to create the corresponding IDs, and after that, will be stored the data in HotelAmenities table (see Fig. 4.8).

4.3.7 DB2 Information storage

Finally, the data will be stored in DB2. When storing the data in DB2, the program searches in DB1 the origin of such data and updates the consolidation date. Thus, it is

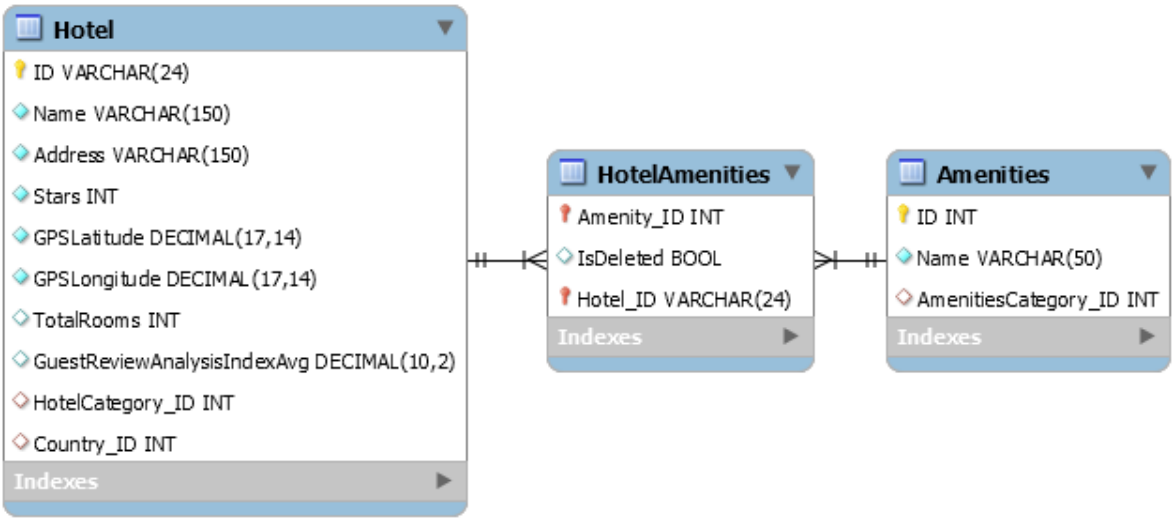


Figure 4.8: Schemes of the relationships between Hotel, HotelAmenities and Amenities tables.

possible to know later the date and time when it was performed the consolidation for the data. Once stored in DB2, the data can be used directly for the final application in the BI system.

4.4 Discussion

The requirements of the hospitality BI system are very specific, with well-defined needs, and have to be developed to integrate analytical tools applied to historical data. The data includes the internal and external information that is pertinent to the hotel. The objective is to provide the decision makers with timely relevant data and a shared vision of the future and knowledge that encompasses the decision makers' resolution and create intelligence, providing a BI system to the organization [82].

In a hospitality BI system, which includes data from several web sources, with different formats and structure, it is essential to consider the development of a BDW. In this kind of Data Warehouse, the data integration phase starts when the web crawler collects information that is presented in relevant websites, related with the hotel business, and store it in collections of different kind of information (in a NoSQL database).

The NoSQL database, also called in our case as primary database, is constituted by collections of data that are stored in an unstructured format and aren't consolidate, which represent a problem to implement a BDW. Therefore, it is necessary to clean and transform the data, and after that to upload the consolidated extracted data into a database, also called a secondary database. This secondary database permits the development and implementation of analytical tools, which includes OLAP and Data Mining, to elaborate enterprise reporting which supports the hotel decision maker activities.

The process of data transform and consolidate the data in an adequate structure and format in the secondary database is constituted by several tasks: i) reading information from DB1, ii) data conversion rules; iii) data dictionaries, iv) correspondence between extracted hotels in different channels, v) channels priority, vi) routing rules/-data flow, and vii) storage information on DB2.

The consolidation of information extracted from the web is a task that needs supervision from time to time. Each time a channel, add a new word that does not appear in a data dictionary it is necessary to add it to the data dictionary. It is expected that over time the data dictionary start becoming increasingly complete and thus the number of new words that may arise will decrease. Consequently, the consolidation system will become more stable over time.

5

Guest Reputation Indexes to Analyze the Hotel's Online Reputation Using Data Extracted from OTAs

Chapter Outline

Nowadays many travelers use online travel agency (OTAs) to book flights, hotel rooms, rent-a-cars, cruises or entire vacation packages. Usually OTAs allow their users to give scores and to write reviews about what was used. Each OTA defines the terms and conditions for guest rating or review score and hoteliers are giving increasing impor-

tance to the scores and reviews their guests do in OTAs. This Chapter proposes two guest reputation index to help hoteliers to monitorize their presence in OTAs. The Aggregated Guest Reputation Index (AGRI), which shows the positioning of a hotel in different OTAs and it is calculated from the scores obtained by the hotels in those OTAs. Another one, the Semantic Guest Reputation Index (SGRI), which incorporates the social reputation of a hotel and that can be visualized through the development of word clouds or tag clouds. Examples of usage of these indexes are given with data extracted from 5-stars hotels in the Algarve, south region of Portugal, that are available on Booking.com and Expedia.

5.1 Introduction

The management of rates on hotel management is becoming increasingly complex and it is very difficult to understand the value that hotels present, in a geographical area or in a class of services with similar features.

With the quantity of information that daily circulates through the web and a number of users estimated at 3 billion in 2015 [49], there is a lot of information about competitors, the hospitality industry and about consumers trends. This information is increasingly more accessible to organizations at lower costs, presenting a new challenge on creating platforms that are able to deal with this huge amount of information that organizations have at their disposal. This is the "big data challenge" [26], that allows that organizations have turned their focus to collect information not only from internal sources, but from external ones [25].

The web 2.0 with its strong interactive component allows its users to consult the static contents and to share and exchange information within the virtual community, which is extremely dynamic and influential in the consumers decision making. Virtual relationships currently established between the hospitality industry and its guests provide valuable information, allowing a continuous assessment by the hotel managers

in its management decisions, guests' feedback and the behavior of their competitors [32].

The sale of hotel rooms by online channel, particularly through the various OTAs (Online Travel Agencies) that exist in the market, assumes an increasingly importance [43]. Many travelers consult different websites before booking online or to contact a hotel booking service, which reinforces the idea of the increasingly important role that the OTAs have in choosing a particular hotel.

OTAs are the fastest growing segment of the travel industry. *Booking.com*, *Expedia* [54], Travelocity, Priceline. Orbitz and Kayak are some examples of OTAs. Travelers can use these OTAs to search for flights, hotel rooms, rent-a-cars, and so on. For example, *Expedia* collect and aggregate data from thousands of travel service providers, allowing to book flights, hotel rooms, rental cars, cruises or entire vacation packages.

More and more booking traffic is to be carried over the traditional channels (travel agencies) to the individual customers and to corporate travel planners, which use the online intermediaries (OTAs) for information queries and to obtain pricing information and online reputation of the hotel [24].

This Chapter proposes two different guest reputation indexes. The first one, the Aggregated Guest Reputation Index (AGRI), which shows the positioning of a hotel in different OTAs and it is calculated from the scores obtained by the hotels in those OTAs. The second one, the Semantic Guest Reputation Index (SGRI), which incorporates the social reputation of a hotel and that can be visualized through the development of word clouds (also known as tag clouds), which enables a facilitated and a graphically attractive visualization of the characteristics most mentioned by the guests of a hotel in their reviews in the OTAs.

The AGRI and SGRI can be considered as a new Key Performance Indicators (KPI), to be included in techniques for an efficient optimization of occupancy and rates of hotel accommodations, known as Smart Revenue Management (SRM) [61, 77].

Prices and types of rooms, capacity, facilities, amenities, and reviews from the hotel

guests, among others are some of the functionalities extracted by webcrawlers [61, 77], which run periodically through the webpages of the different OTAs, over different periods of time, in order to get suitable data. For this work, two different OTAs were analyzed: *Booking.com* and *Expedia*.

After the extraction of the information available in OTAs, it is necessary to perform its analysis to make available to hoteliers of valid and easily legible information about their hotels and about their competitive set, in order to enable valuable and quick decision-making.

This Chapter is structured as follows: Sec. 5.2 presents two scenarios for the calculation of an aggregated guest reputation index, while Sec. 5.3 explains a semantic guest reputation index, which can be developed using the sentiment analysis or opinion mining approach or in a simpler manner, using word clouds. Finally, Sec. 5.4 presents the conclusion and some guidelines for future work.

5.2 Aggregated Guest Reputation Index (AGRI)

As presented in [61, 77], the webcrawler performs the extraction of several items with information about the hotel in a given OTA, for instance: the available rooms, prices, features, amenities, policies, guest reviews and so on. On *Booking.com* and *Expedia*, only the person who booked and completed a stay at that hotel can write reviews and/or gives scores to that hotel. On *Expedia*, this rate is called the Guest Rating; on *Booking.com* is considered the Review Score. Any of these two designations are used in this Chapter. On TripAdvisor [53] any person can leave a review about a hotel, a restaurant, and so on. They do not need to book and to complete a stay in that hotel. This is one of the reasons why TripAdvisor is not considered in the calculation of the proposed indexes in this Chapter [12].

The webcrawler also extracts the review score that is based on a given number of reviews and the score breakdown that rate several information dimensions. On *Book-*

ing.com these dimensions are: Cleanliness, Comfort, Location, Facilities, Staff, Value for Money and Free Wi-Fi; on *Expedia* they are: Room Cleanliness, Service & Staff, Room Comfort and Hotel Condition.

Another important information related to guest reviews is the customer segment that the guest belongs. For example, *Booking.com* displays the following segments: All reviewers, Families, Couples, Group of friends, Solo travelers, Business travelers, while *Expedia* shows the following ones: Everyone, Couples, Families, Getaway with friends, Business travelers, Overnight stay before destination, Personal event, Spa, Golf and other.

In the process of organizing the information extracted by the webcrawler, it was found that the Review Score on *Booking.com* does not correspond to the average of the ratings of each dimension. Figure 5.1 shows an example for a hotel on *Booking.com*, where the Review Score does not correspond to the average of the Score Breakdown.

On the contrary, the Guest Rating on *Expedia* corresponds to the average of the scores of the different dimensions analyzed. Figure 5.2 displays an example of the calculation of the Guest Rating of a hotel on *Expedia*.

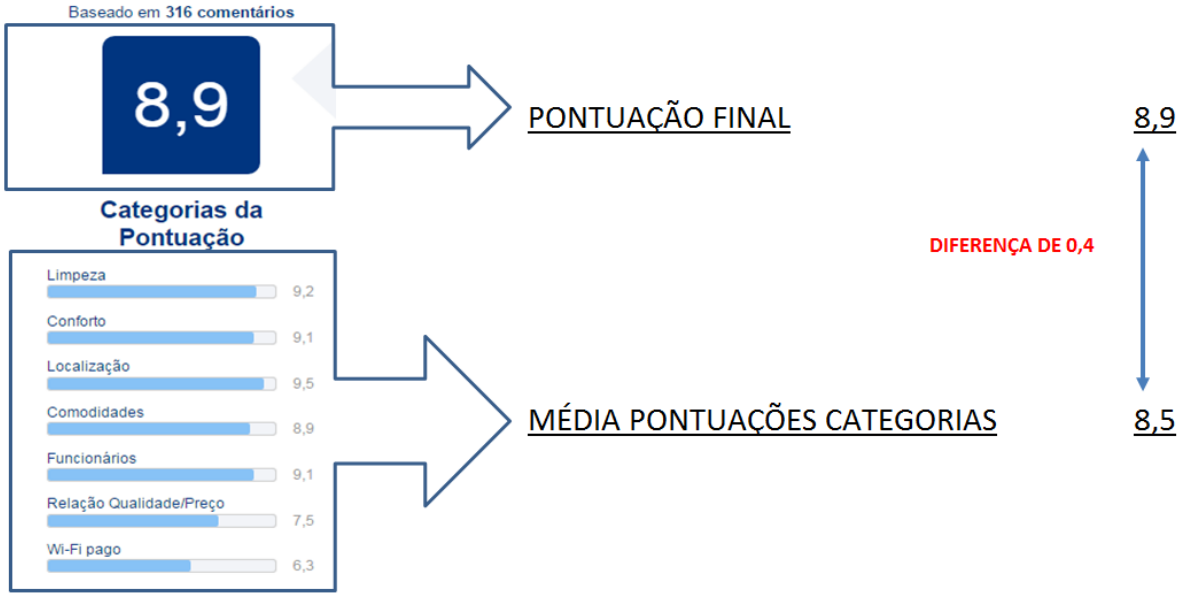


Figure 5.1: Differences between Review Score and Score Breakdown.

To overcome these differences that exists with *Booking.com*, a first approach to the

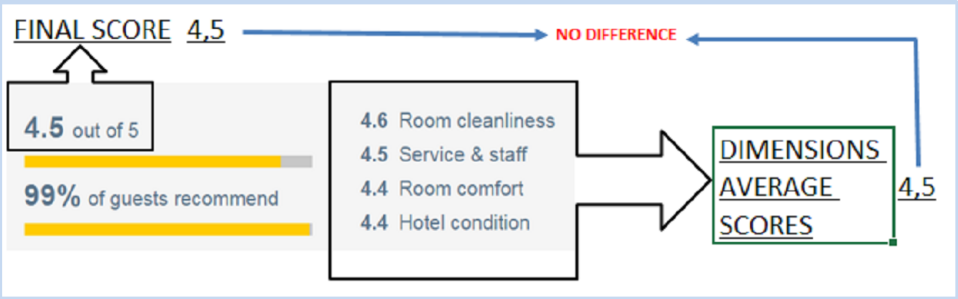


Figure 5.2: Guest Rating corresponds to the average of the dimensions scores.

calculation of an Aggregated Guest Reputation Index (AGRI) was done using the following items:

1. Number of reviews by OTA;
2. Weight of each OTA in the total number of reviews;
3. Review Score by OTA;
4. Review Score by Segment;
5. Number of reviews per Segment;
6. Weight of each Segment in the total number of reviews.

Some of these items are extracted by the webcrawler, others are calculated, as items 2) and 6). As the number of reviews in one OTA can be different from that number of reviews in another OTA, the weight of each OTA in the total number of reviews and the weight of each segment in the total number of reviews are considered and calculated.

Another important aspect is that OTAs can use different rating scales. While on *Booking.com* scores are assigned on a scale of 1-10, on *Expedia* the scale chosen is 1-5. So, the normalization of the scale has to be done. In this Chapter, the normalization or standardisation to the 1-10 scale was chosen for two reasons: the first one because it is considered easierto have 1-10 scale than 1-5; the other one is because *Booking.com* continues to be considered the number one OTA in the world.

Table 5.1 displays an example for a hotel for the two OTAs considered, for the dimensions of each OTA and showing only one segment, the Families one. The other segments are not displayed only for lack of space. Furthermore, it was considered that the Solo travelers of *Booking.com* match the Personal Event of *Expedia*. Finally, the numbers from 1 to 6 showed in the columns of Tab. 5.1 correspond to the items presented for the calculation of the AGRI.

HOTEL						
	Reviews			Families		
	3	1	2	5	4	6
	Score	Nº	Weight Number Total Reviews	Score	Nº Reviews	%
GRI	**4,5**	**865**	**100%**	**4,2**	**130**	**15,0%**
Booking	4,3	540	62,4%	4,2	79	14,6%
Cleanliness	4,5	540	62,39%	4,3	79	14,6%
Comfort	4,3			4,2		
Location	4,9			4,8		
Facilities	4,2			4,1		
Staff	4,5			4,2		
Value for Money	3,7			3,6		
Expedia	4,6	325	37,6%	4,2	51	15,7%
Room Cleanliness	4,7	325	37,55%	4,4	51	15,7%
Service & Staff	4,6			3,9		
Room Comfort	4,5			4,1		
Hotel Condition	4,6			4,4		

Table 5.1: Review Score and one dimension of Score Breakdown.

Having in mind this information, several scenarios can be drawn. Table 5.2 shows an example of one of them: the calculation of an AGRI as the weighted average of the scores obtained in two OTAs using the weight that each OTA has in the total number of analysed reviews.

Another scenario can be using a weighting factor that can be defined by the user. In this case, the weighted factor "number of bookings received YearTo-Date by each channel" (Tables 5.3 and 5.4) was used.

Taking into account the need to present reliable results and to allow a scalability

Scenario 1

OTA	Nº Reviews	% Total Nº Reviews	Score	Normalized Score	Weigthed Average 1
Booking	200	65,10%	9,2	9,2	9,3
Expedia	118	34,90%	4,8	9,6	
Total	**338**	**100,00%**			

Table 5.2: Calculation of the weighted average using a weighted factor in function of total number of reviews.

OTA	Nº bookings received by each channel YTD	Weighted Factor
Booking	100	91%
Expedia	10	9%
Total	**110**	**100%**

Table 5.3: Calculation of a weighted factor in function of the number of bookings received YearTo-Date (YTD) by each channel.

Scenario 2

OTA	Weighted Factor	Score	Normalized Score	Weighted Average 2
Booking	91%	9,2	9,2	9,2
Expedia	9%	4,8	9,6	
Total	***100%***			

Table 5.4: Calculation of weighted average using weighted factor defined by the user in function of number of bookings received Year-To-Date by each channel.

and a rapid information integration in any hotel revenue management system, several scenarios for the calculation of the AGRI can be proposed. The hoteliers have to choose the scenario that for them provided the most consistent information with the reality of their hotel units.

It is important to refer that the information displayed by the OTAs changes very rapidly. OTAs are constantly improving the interface and changing the type of information displayed. For this reason, the dimensions and segments presented in this Chapter for each OTA can change from one day to another.

5.2.1 Application of the AGRI to Algarve 5-stars hotels

Next, the calculation of the AGRI, using the first scenario is demonstrated with the information extracted by the webcrawler for forty 5-stars hotels of the Algarve region, in the south of Portugal, that are available on *Booking.com* and on *Expedia*.

The hotel designation, score and number of reviews were extracted by the webcrawler for each OTA and are displayed in Tab. 5.5 The calculation of the AGRI was performed using the total number of reviews of the analyzed hotels in the two OTAs considered.

The AGRI can be used to develop KPIs (Key Performance Indicators), which can provide valuable information about the hotel positioning compared with the other 5-stars hotels segment or compared with the competitive set.

For each hotel, the hotelier can analyse the weight that each OTA has on the number of reviews posted about the hotel. For example, for all the forty 5-stars hotels of the Algarve that are available on *Booking.com* and *Expedia*, Fig. 5.3 shows that *Expedia* generated 31,5% of the total number of reviews, while *Booking.com* generated 68,5%. As expected, this reinforces the perception that *Booking.com* takes an increasingly important role in the number of bookings generated by OTAs. It is important to stress the importance of choosing a common time horizon to the analysis to be done.

Figure 5.4 displays the first ten hotels displayed in Tab. 5.5. Using the information provided by the hotel revenue management system or by the Property Management System (PMS), it is possible and recommended an analysis of the number of bookings raised by each OTA and the number of reviews that were generated by this same channel, allowing to analyse what is the type of guests that is more interactive and participative.

Figure 5.5 shows the positioning of one hotel, denoted as "hotel1" in relation to the average of the forty 5-stars hotels segment (score 8,8) and also in relation to its most direct competitors, its competitive set, which are highlighted with yellow colour in Tab. 5.5. This is an example of a KPI that can be performed and that is easy to read

HOTEL	Booking Nº reviews	Booking Weighted Factor	Booking Score	Expedia Nº reviews	Expedia Weighted Factor	Expedia Score	Norm. Score	Reviews Total	AGRI
Hotel1	116	65,2%	9,5	62	34,8%	4,8	9,6	178	9,5
Hotel2	226	65,5%	9,2	119	34,5%	4,8	9,6	345	9,3
Hotel3	484	56,9%	9	367	43,1%	4,8	9,6	851	9,3
Hotel4	210	53,8%	8,9	180	46,2%	4,8	9,6	390	9,2
Hotel5	371	91,6%	9,2	34	8,4%	4,6	9,2	405	9,2
Hotel6	501	81,7%	9,2	112	18,3%	4,6	9,2	613	9,2
Hotel7	174	70,2%	9	74	29,8%	4,8	9,6	248	9,2
Hotel8	35	66,0%	9	18	34,0%	4,7	9,4	53	9,1
Hotel9	78	91,8%	9,1	7	8,2%	4,6	9,2	85	9,1
Hotel10	690	66,5%	8,9	347	33,5%	4,7	9,4	1037	9,1
Hotel11	30	66,7%	8,9	15	33,3%	4,7	9,4	45	9,1
Hotel12	188	67,4%	9	91	32,6%	4,6	9,2	279	9,1
Hotel13	140	67,3%	8,9	68	32,7%	4,7	9,3	208	9,0
Hotel14	237	76,5%	8,9	73	23,5%	4,7	9,4	310	9,0
Hotel15	477	61,8%	8,9	295	38,2%	4,6	9,2	772	9,0
Hotel16	148	55,8%	8,7	117	44,2%	4,7	9,4	265	9,0
Hotel17	200	39,5%	8,6	306	60,5%	4,6	9,2	506	9,0
Hotel18	150	48,7%	8,6	158	51,3%	4,6	9,2	308	8,9
Hotel19	147	82,1%	8,8	32	17,9%	4,5	9,0	179	8,8
Hotel20	91	84,3%	8,8	17	15,7%	4,5	9,0	108	8,8
Hotel21	399	76,4%	8,7	123	23,6%	4,6	9,2	522	8,8
Hotel22	644	71,6%	8,8	256	28,4%	4,4	8,8	900	8,8
Hotel23	404	80,5%	8,7	98	19,5%	4,6	9,2	502	8,8
Hotel24	88	57,5%	8,5	65	42,5%	4,6	9,2	153	8,8
Hotel25	667	92,0%	8,7	58	8,0%	4,5	9,0	725	8,7
Hotel26	508	95,5%	8,7	24	4,5%	4,4	8,8	532	8,7
Hotel27	365	60,9%	8,4	234	39,1%	4,5	9,0	599	8,6
Hotel28	593	76,5%	8,5	182	23,5%	4,5	9,0	775	8,6
Hotel29	165	62,0%	8,5	101	38,0%	4,4	8,8	266	8,6
Hotel30	430	48,7%	8,5	453	51,3%	4,3	8,6	883	8,6
Hotel31	71	28,2%	8,4	181	71,8%	4,3	8,6	252	8,5
Hotel32	145	97,3%	8,5	4	2,7%	4,8	9,6	149	8,5
Hotel33	169	38,5%	8	270	61,5%	4,4	8,8	439	8,5
Hotel34	411	89,5%	8,5	48	10,5%	4,1	8,2	459	8,5
Hotel35	116	61,1%	8,4	74	38,9%	4,2	8,4	190	8,4
Hotel36	54	56,8%	8,4	41	43,2%	4,2	8,4	95	8,4
Hotel37	318	86,9%	8,3	48	13,1%	4	8,0	366	8,3
Hotel38	171	77,4%	8,2	50	22,6%	4,2	8,4	221	8,2
Hotel39	215	68,7%	8,1	98	31,3%	4,2	8,4	313	8,2
Hotel40	467	69,7%	8	203	30,3%	4,2	8,4	670	8,1

Table 5.5: AGRI calculation for 5-stars hotels of the Algarve that are available on *Booking.com* and on *Expedia*.

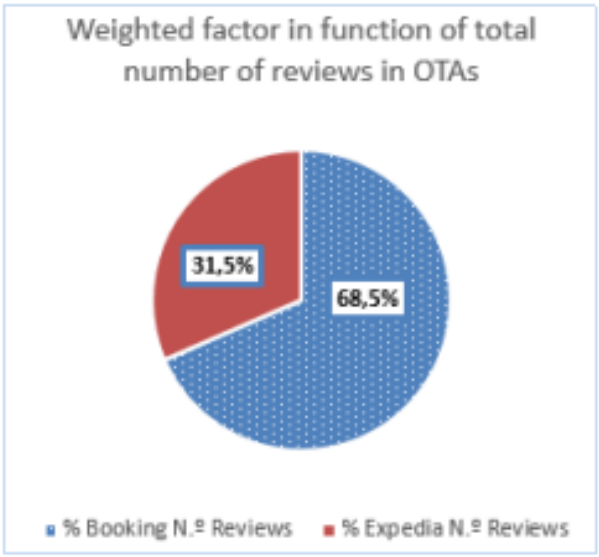


Figure 5.3: Weighted factor calculated using the total number of reviews of the Algarve 5-stars hotels on *Booking.com* and on *Expedia*.

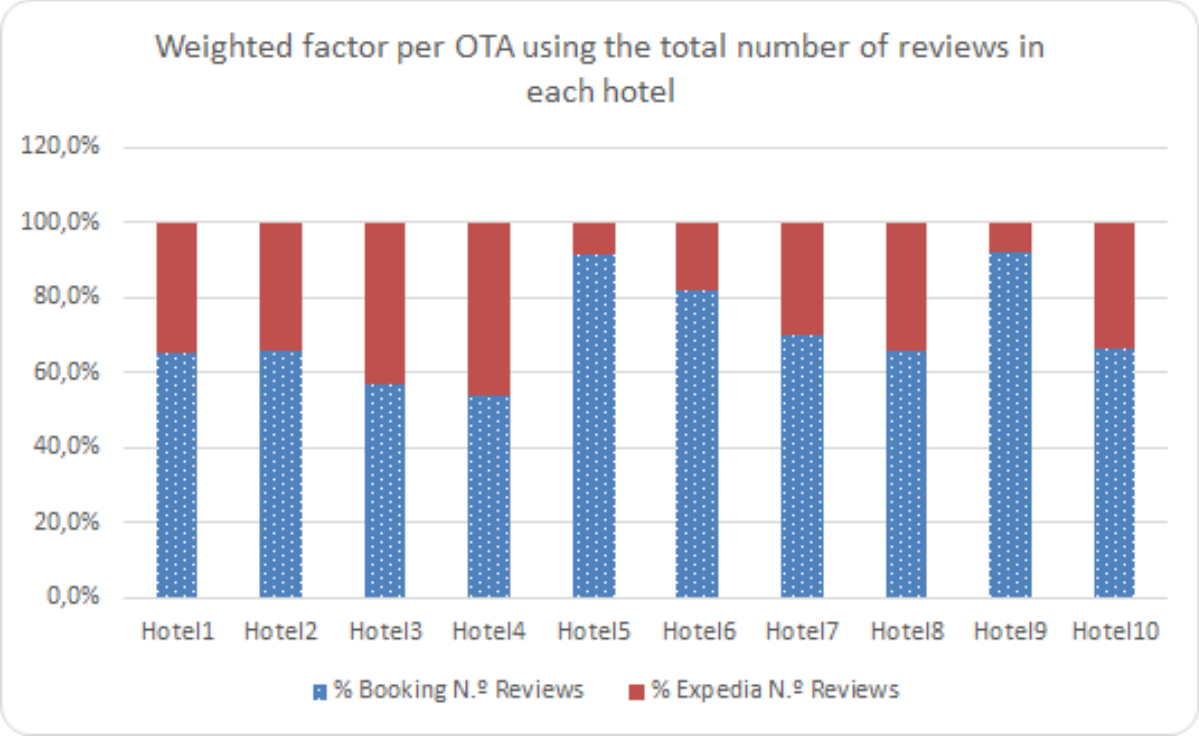


Figure 5.4: Weighted factor per OTA using the total number of reviews of each hotel.

and to understand.

Finally, saving these KPIs along time, allows that the hotelier can check the evolution of the AGRIs over time and compared it to the competitive set.

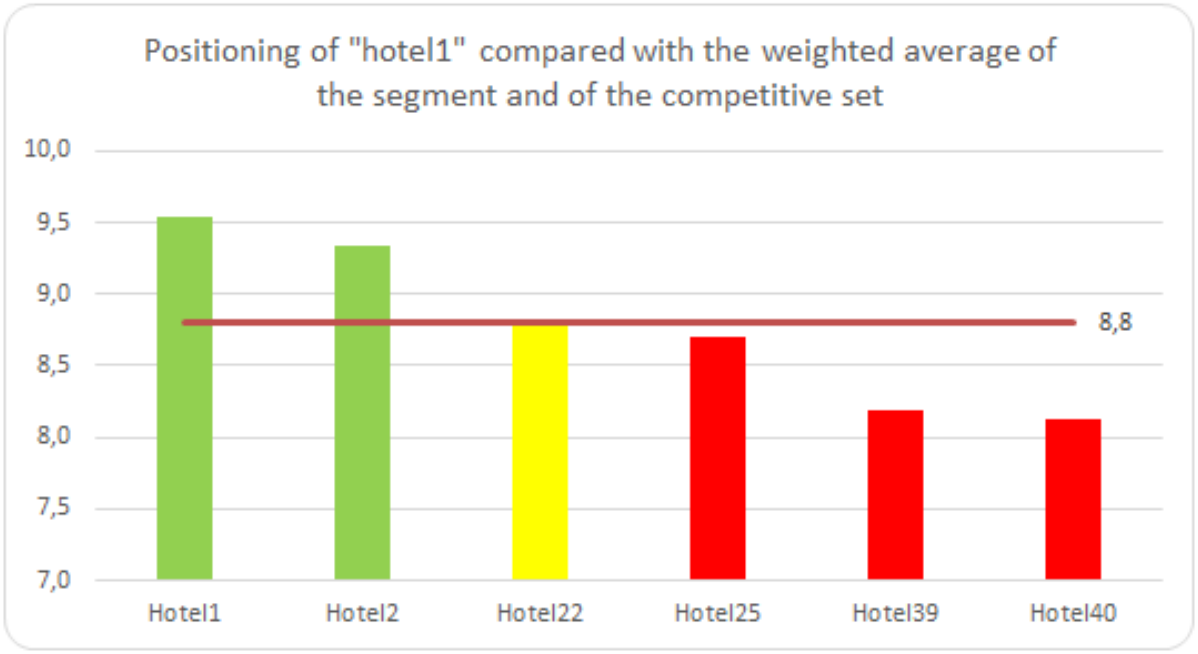


Figure 5.5: Positioning of "Hotel1" compared with the weighted average of the segment of 5-stars hotels of the Algarve and of the competitive set.

5.3 Semantic Guest Reputation Index (SGRI)

In addition to the analysis that was done using the AGRI proposed above, the text of the reviews extracted by the webcrawler in each OTA can be used in a different manner and need a different analysis.

The reviews reflect the opinions of the guests and can highlight aspects and items that are more or less valued for them.

Each OTA has its specifications on how it is possible to write reviews on the website. For example, on *Booking.com* the guest can give a positive review that is displayed with a green plus sign + or a negative one as - grey minus sign. For *Expedia*, the designation is different, the positive reviews aredesignated as "Pros", the negative as "Cons" and there is a different designation "Location", where reviews about the localization can be given.

Another important aspect of the reviews is the language used to write them. On *Booking.com* there is the possibility to have reviews in 17 languages as can be seen in Fig. 5.6.

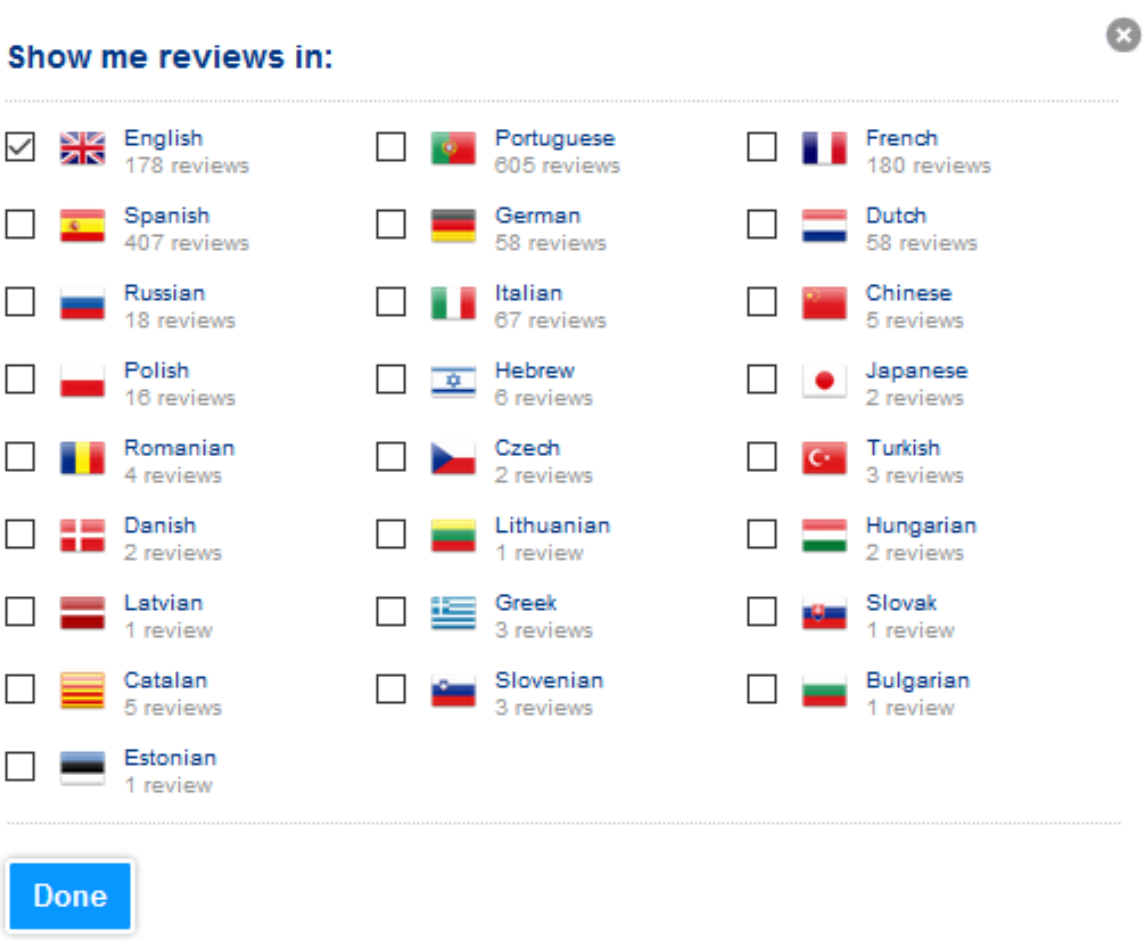


Figure 5.6: Languages of the reviews on *Booking.com*.

On *Expedia* the languages used are 13. Czech, Russian, Polish and Slovenian are not considered on *Expedia* but used on *Booking.com*.

Another important difference between *Booking.com* and *Expedia* is concerning to the reviews showed, *Booking.com* shows the reviews posted by guests during the 14 past months whether *Expedia* never delete reviews apart if the hotel ask it (for example following a refurbishment or a property change of ownership).

A relevant aspect is that reviews give valuable information about the hotel, what is going well or bad with the hotel, related to the various dimensions presented before. Recently, *Booking.com* changed the way as reviews are showed. They began to display a summary of the reviews, given information about the number of positive and negative reviews about Location, Staff, Price, Bathroom, and so on. Figure 5.7 shows an

example for a hotel.

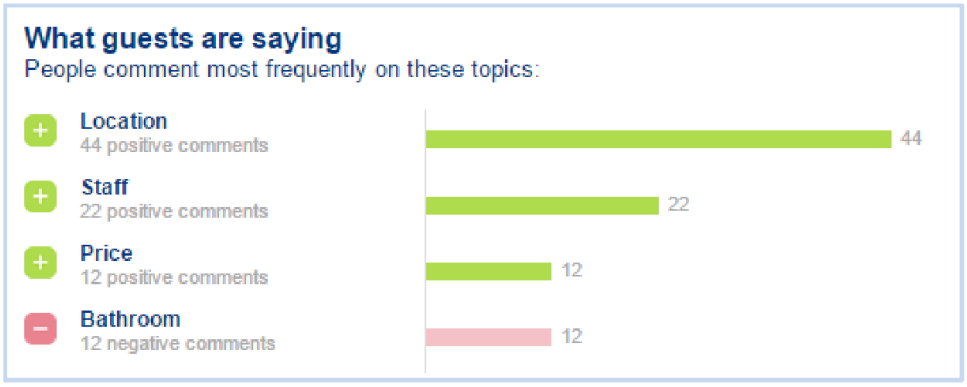


Figure 5.7: Total number of positive and negative reviews for a hotel on *Booking.com*.

It is also important to note that the review can be considered positive or negative by the guest, but the text of the review itself can give a slightly different information, which can be important for the hotelier. For instance, the positive review displayed in Fig. 5.8 (extracted from *Booking.com*) says that staff is friendly but minimal and it was considered by the guest as a positive review.

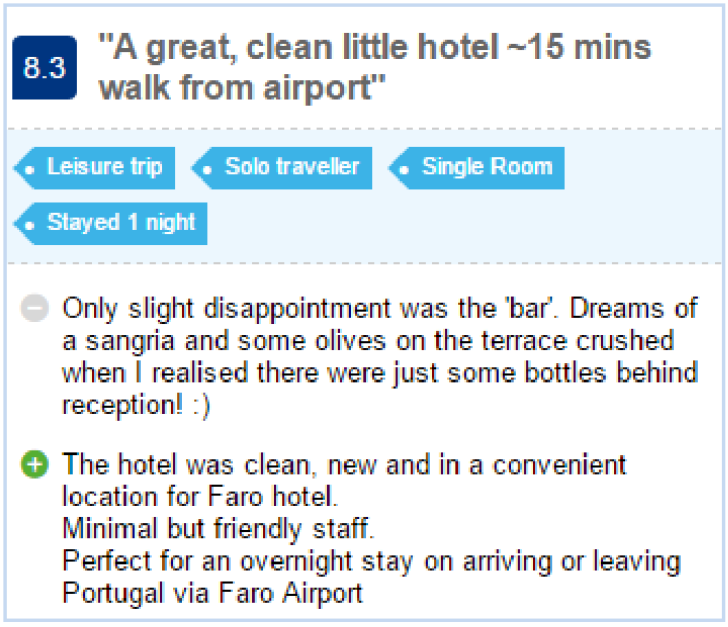


Figure 5.8: Example of a positive and a negative review for a hotel on *Booking.com*.

Finally, there are techniques that allow to extract and to evaluate the sentiment expressed in textual data. Sentiment Analysis (also known as Opinion Mining) allows this evaluation.

5.3.1 Sentiment Analysis or Opinion Mining

Sentiment analysis or Opinion Mining refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in different source materials, generally from text. These fields of knowledge are at the crossroads of information retrieval and computational linguistics and have a rich set of applications [37], since ranging from tracking users' opinions about products or services, to customer relationship management until analysis of hotel guest reviews on OTAs.

To get good results with Sentiment Analysis (or Opinion Mining), it is necessary to implement previously several techniques to sentiment analysis itself, which can help to get the polarity of the text.

As the text in reviews is normally written in an informal way, it is necessary the preprocessing of the text to correct grammatical and orthographical errors, which can difficult the search of relevant information [87].

Generally, Sentiment Analysis involves several phases: extraction and preprocessing of text; natural processing language and sentiment analysis itself.

In the extraction and text preprocessing the abbreviations and linguistics contractions are corrected in order to obtain words that exists in a given language. In informal text, it can also occur the repetition of letters in words to give emphasis (for example, "baaaaaad" instead of "bad").

The natural language processing also involve several steps [87], since the division of the text in simpler terms (tokenization) until complex ones as parsing (phrase chunking). Another step is POS Tagging (part-of-speech tagging), which determines the grammatical class of each component of the analyzed sentences. The Apache OpenNLP library [1] is the most used software for the processing ofnatural language. It is a machine learning based toolkit, which supports the most common NLP tasks and also includes maximum entropy and perceptron based machine learning.

Finally, in the sentiment analysis phase the subjects of the text are identified. The

names that expose the subjects of the text and the adjectives that characterize those subjects as positive or negatives are analyzed in this phase. The terms are analyzed according to their grammatical class. These operations use databases and lexical resources. SentiWordNet [14, 51] is an example of a tool that can be used to perform the Sentiment Analysis (or Opinion Mining). SentiWordNet extends WordNet's usability by another dimension. WordNet as explained in [11, 77] is a "dictionary of meanings", which integrates the functions of a dictionary and a thesaurus. In WordNet, nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms, called synsets, each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.

Figure 5.9 displays a flowchart for the Sentiment Analysis process applied on reviews using SentiWordNet. According to [51], after preprocessing the text, it is reduced to its contents words in a normalized form. For each of the words, SentiWordNet retrieves the synsets that contain each word. If SentiWordNet does not find any synset for that word, the sentiment score is defined to zero. On the contrary, if more than one synset are returned, the word sense disambiguation is necessary. According to those authors, there are several ways to perform word sense disambiguation using WordNet, one of them is using the Lesk algorithms, which disambiguate calculating overlaps of the context words and the synsets' glosses. Finally scores are then given. Generally the scale [-1.0; 1.0] is used, the -1.0 corresponding to the most negative sentiment, 0 to a neutral sentiment and 1.0 the most positive.

Other solutions are acquiring existing software in the market. These solutions allow an easiest implementation of the Sentiment Analysis process, however implies an extra cost with the acquisition of this software. Table 5.6 listed three of several software available in the market. These solutions are currently used by several leading companies of different areas such as hospitality, consulting and social networking among others.

A different and easier approach to the semantic guest reputation index is to use

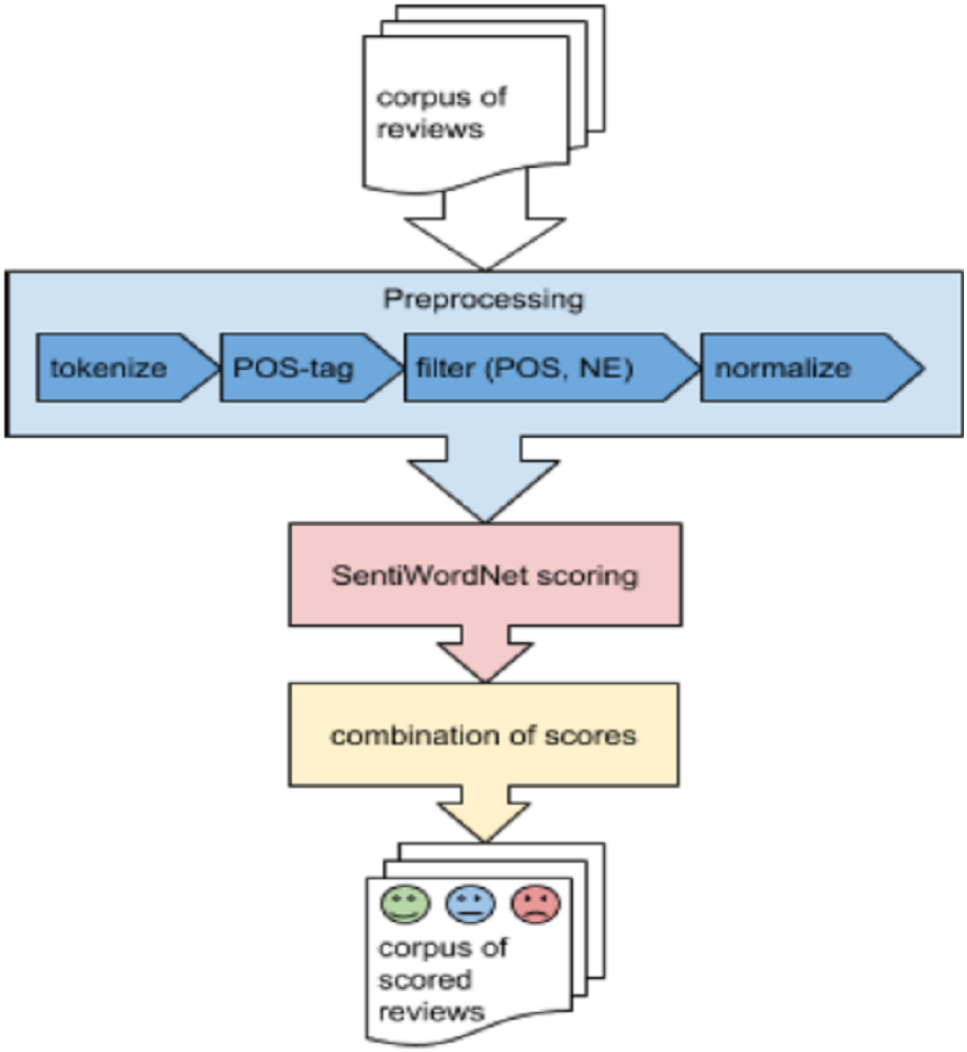


Figure 5.9: Flowchart for Sentiment Analysis (or Opinion Mining) process applied on reviews using SentiWordNet. Source: [51].

word clouds to graphically see the most mentioned words present in the reviews.

5.3.2 Word Clouds

Word clouds (also known as tag clouds or text clouds) is a visual representation for text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text. The frequency of each word/tag can be shown with a different font size or color. Wordle, Tagxedo, Tagcrowd and Wordaizer are examples of word clouds software available in the market.

Software	Languages	Features
Repustate	English, French, Spanish, German, Italian, Russian, Chinese and Arabic	Sentimental Analysis; Semantic Analysis; Repustate Server; Repustate's Excel plugin
ReviewPro	Source language of comments	Global Review Index™ (GRI); Revenue Optimizer; Guest Survey; Sentiment Analysis
TrustYou	Semantic analysis in more than 20 languages with an accuracy of 90-95% for most languages	Reputation Marketing; Reputation Surveys; Reputation Monitoring; Mata Review API

Table 5.6: Software for sentiment analysis or opinion mining of guest reviews. Source: [6, 7, 8].

In this Chapter, the Wordaizer [11] software was used to analyse the extracted guest reviews. This software allows a word counting, revealing the items most cited by guests in their reviews. Obviously, the software does not identify sentiment, feeling or opinion associated to reviews. As was explained in Sec. 5.3.1, *Booking.com* and *Expedia* have the reviews classified in positive (pros) and negatives (cons), so two different reviews groups(Positives and Negatives) were created and analysed with Wordaizer. Twenty five rev iews on *Booking.com* and on *Expedia* of one of the forty 5-stars hotels of the Algarve are presented in Figures 5.10 and 5.11.

As can be seen in Figure 5.10, room, staff, hotel, breakfast, excellent, great, restaurant, facilities, services are some of the nouns and adjectives most used in the positive reviews.

On the other hand, swimming pool, little, hotel, euros, warm, breakfast, nothing, reception, short are of the nouns and adjectives most used in the negative reviews.

In conclusion, word clouds can give to hoteliers an easy and fast way to visualize the positive and negative reviews.

Figure 5.10: Word cloud of positive reviews.

Figure 5.11: Word cloud of negative reviews.

5.4 Conclusion

This Chapter presents recent developments to hotel's online reputation management, which aims the development of smart automatic techniques for an efficient optimization of occupancy and rates of hotel accommodations [61, 77].

In this Chapter two different guest reputation indexes were proposed. The Aggregated Guest Reputation Index (AGRI), which shows the positioning of a hotel in different OTAs and that is calculated from the scores obtained by the hotels in those OTAs. The Semantic Guest Reputation Index (SGRI), which incorporates the reviews given by the hotel guests who booked the hotel room using an OTA.

The AGRI proposed can use two different scenarios: the first one that calculated the AGRI as the weighted average of the scores obtained in various OTAs using the weight that each OTA has in the total number of reviews analysed. The second one, using a weighting factor that can be defined by the hotelier.

The SGRI can also be developed using two approaches. One, using Sentiment Analysis (or opinion mining) that identifies the sentiment, feeling or opinion expressed in reviews; other, the analysis and visualization of word clouds (or tag clouds) that graphically shows the words most cited by guests in the reviews. Each one of the approaches can give valuable information to hoteliers to monitorize the social reputation and positioning of hotels in OTA. Furthermore, hoteliers can anticipate and influence consumer behavior in order to maximize revenue.

The results achieved in this Chapter open multiple paths for future work. One of them is to study and to compare the software presented in Tab 5.6, the other one is to implement the Sentiment Analysis using SentiWordNet or other similar software, and createa new indicator that integrates the concepts of AGRI, SGRI and social networks into a single indicator.

6

Conclusions

This thesis presented a set of works that are part of a RM framework, aiming to demonstrate that it is possible to automatically browse e-commerce websites, indentify the relevant DOM elements and extract them to a database. The usefulness of the RM system is suported on the availability of data. In order to satisfy this need for data, the web crawlers must run periodically to collect information in a suitable and updated manner.

This document shows a solution to overcome some of the major dificulties in data extraction, namely the interaction of Javascript and AJAX, by using a webdriver. Furthermore, although the extraction of data can not be a fully unsupervised process in the proposed solution, human actions are only required if page layouts are modified

since last extraction. We can extract information about hotel characterization, prices, amenities and reputation from four different web channels, *Booking.com*, *Expedia*, *Tripadvisor* and *Bestday*. Moreover each time the web crawlers inspect a web page, the HTML tags, which are filled with the target information, are stored into the database, creating an history of tags (as mentioned on Sec. 2.3.1). This history allows to give some intelligence to the web robots, in the sense that they try a set of HTML tags until the DOM element is found in the webpage. The more times the robots analize the websites the more intelligent they will be.

The extracted information allows to implement a Big Data Warehouse (BDW) which consists mainly in huge amounts of data, collected from different sources. To properly create that BDW, concepts and techniques of semantics were also included in the software, overcoming the problems that are found in the creation of an analytics tools system with this dimension, this variety, and velocity requirements.

In this work, the data integration phase, within the Data Warehouse, starts when web crawlers collect information from relevant websites, related with the hotel business, and store it in collections of a NoSQL database.

The NoSQL database in our study, also called primary database, is constituted by collections of data that are stored in an unstructured format (JSON) and are not consolidated, presenting a problem in the implementation of the BDW. Therefore, it is necessary to clean and transform the data, and after that to upload the consolidated information into a relational database, also called secondary database. This secondary database allows an easiest the development and implementation of analytical tools, which includes OLAP and Data Mining, to elaborate enterprise reportings which support the hotel's decision maker activities.

The information retrieved by the web robots is stored exactly how it was extract from the websites. This feature allows to do reverse engineering in case of something went wrong after data consolidation. For instance, if some badly consolidated data is attained we can readapt the consolidation rules to treat the data from primary database

again without the need to re-extract the data from the websites.

The consolidation of information extracted from the web is a task that needs supervision from time to time. For example, new words have to be added to the data dictionary whenever not present (see Sec. 4.3.3). Over time is expected that the data dictionary will become increasingly complete and thus, the number of new words that may arise will decrease. Consequently, the consolidation system will become more stable, as happens with the HTML tags history introduced before.

The correspondence between extracted hotels, refered on Sec. 4.3.4, is a process that will fail some times generating duplicated hotels on the database. This problem happens in rare cases, as for example when the information of an hotel is present in distinct sites with similar names and the GPS coordinates are very close. This problem was already improved, as will be explained in Future Work Section.

Finally, two different guest reputation indexes were proposed. The (i) Aggregated Guest Reputation Index (AGRI) which shows the positioning of an hotel in different OTAs being calculated from the scores obtained by the hotels in those OTAs, and the (ii) Semantic Guest Reputation Index (SGRI) which incorporates the reviews given by the hotel's guests who booked a room using an OTA. WordNet was considered as a semantic tool to help to calculte this last index, however this solution is not fully adequate to a sentiment analysis or opinion mining. The semantic analisys proved to be the most dificult part to implement mainly because there was many ways to say the same thing.

6.1 Future Work

The development of SRM project continued after the publication of the papers that make up this dissertation. For this reason some of the problems presented on Sec. 6 were already improved.

The following items show some of the progress that was done after this publica-

tions that form this work:

- **Scheduling extraction** - Using background jobs and workers (Hangfire [69]) it is possible to launch the web robots extraction automaticaly in specific hours of the day and schedule how many times the robots will run. Other advantage of using this is that the crawlers can be launched in parallel and extract huge amounts of data in short time.

- **Web Crawler searching criteria** - The search results are constant for the same search criteria for a period. For instance, if we search on *Booking.com* for Check-in: 22/11/2016, Check-out: 23/11/2016, 1 room and 1 adult the results are the same for a period of time, until the hotel changes the prices. Taking advantage of this, it is possible to program robots to use the results of already extracted data instead of access the website again to get the same data.

- **Data dictionaries multi language** - The data dictionaries presented in Sec. 4.3.3 were thought to store data just in one language, English. This limitation was solved and the language can now be automatically set by the *web robots* when extracting information from channels. This upgrade allows to improve the reputation indexes tools by using words in many languages.

- **Identify hotel address by GPS coordinates and vice-versa** - Using the Geolocation API [34] of Google it is possible to convert GPS coordinates into Addresses and vice-versa. This API can resolve the problem of duplicated hotels refered in Sec. 6 by giving extra information about his location. With the complete location information we can decide more easily when two hotels are the same.

- **Auto detect language** - To create de data dictionaries in many languages it is necessary to know in what language the web robot is extracting the information. To auxiliate in this task, was created an algorithm that detects the language

with a sentence as input. This algorithm uses the most popular expressions of languages to detect the language.

In the future, it is intended to study and to compare the software presented in Tab. 5.6 and to implement the Sentiment Analysis using SentiWordNet or other similar software, and create a new indicator that integrates the concepts of AGRI, SGRI and social networks into a single indicator. The objective of this software it is to create a graphic interface where the user can configure extraction criteria and scheduling the web robots and see the extracted information treated and presented in tables and graphics to compare them.

6.2 Publications

List of the published articles:

- **Martins, D.**, Lam, R., Rodrigues, J.M.F., Cardoso, P.J.S., Serra, F. (2015) **A Web Crawler Framework for Revenue Management**, In Proc. 14th Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases (AIKED '15), in Advances in Electrical and Computer Engineering, Tenerife, Canary Islands, Spain, 10-12 Jan, pp. 88-97. ISBN: 978-1-61804-279-8.

- Ramos, C.M.Q., Correia, M.B., Rodrigues, J.M.F., **Martins, D.**, Serra, F. (2015) **Big Data Warehouse Framework for Smart Revenue Management**. In Proc. 3rd NAUN Int. Conf. on Management, Marketing, Tourism, Retail, Finance and Computer Applications (MATREFC '15), in Advances in Environmental Science and Energy Planning, Tenerife, Canary Islands, Spain, 10-12 Jan., pp. 13-22. ISBN: 978-1-61804-280-4.

- **Martins, D.**, Ramos, C.M.Q, Rodrigues, J.M.F., Cardoso, P.J.S., Lam, R., Serra, F. (2015) **Challenges in Building a Big Data Warehouse Applied to the Hotel**

Business Intelligence, In Proc. 6th Int. Conf. on Applied Informatics and Computing Theory (AICT'15), in Recent Research in Applied Informatics, Salerno, Italy, 27-29 June, pp. 110-117. ISBN: 978-1-61804-313-9.

- Choupina, R., Correia, M.B., Ramos, C.M.Q, **Martins, D.**, Serra, F. (2015) **Guest Reputation Indexes to Analyze the Hotel's Online Reputation Using Data Extracted from OTAs**, in Proc. 6th Int. Conf. on Applied Informatics and Computing Theory (AICT'15), in Recent Research in Applied Informatics, Salerno, Italy, 27-29 June, pp. 50-59 ISBN: 978-1-61804-313-9.

- Ramos, C.M.Q., **Martins, D.**, Serra, F., Lam, R., Cardoso, P.J.S., Correia, M.B., Rodrigues, J.M.F. (2017) **Framework for Hospitality Big Data Warehouse: the implementation of an efficient Hospitality Business Intelligence System**, International Journal of Information Systems in the Service Sector (IJISSS) 9(2), (accpeted for publication 2017).

Bibliography

[1] Apache OpenNLP. https://opennlp.apache.org/. [Accessed 28/10/2015].

[2] Java API for WordNet Searching (JAWS). http://lyle.smu.edu/~tspell/jaws/. [Accessed 21/11/2014].

[3] JWNL (Java WordNet Library). http://sourceforge.net/projects/jwordnet/. [Accessed 21/11/2014].

[4] OntoLingua. http://www.ksl.stanford.edu/software/ontolingua/. [Accessed 25/06/2015].

[5] PyWordNet. http://osteele.com/projects/pywordnet/. [Accessed 21/11/2014].

[6] Repustate - Enterprise scale text analytics. https://www.repustate.com/. [Accessed 01/05/2015].

[7] Reviewpro - Guest Intelligence. http://www.reviewpro.com/. [Accessed 03/05/2015].

[8] Trustyou. http://www.trustyou.com/. [Accessed 01/05/2015].

[9] WebONTO. http://projects.kmi.open.ac.uk/webonto/. [Accessed 25/10/2014].

[10] WordNet CSharp. https://wordnet.codeplex.com/. [Accessed 21/11/2014].

[11] Wordnet Princeton. http://wordnet.princeton.edu/. [Accessed 25/10/2014].

[12] R. Ali and Skift. The Top Online Travel Booking Sites for January 2014. http://skift.com/2014/02/24/the-toponline-travel-booking-sites-for-january-2014/. [Accessed 31 05 2015].

[13] Chris K Anderson, Sherri Kimes, and Bill Carroll. Teaching Revenue Management at the Cornell University School of Hotel Administration. *INFORMS Transactions on Education*, 9(3):109–116, 2009.

[14] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *LREC*, volume 10, pages 2200–2204, 2010.

[15] Robert Baumgartner, Oliver Frölich, Georg Gottlob, Patrick Harz, Marcus Herzog, and Peter Lehmann. Web data extraction for business intelligence: the lixto approach. In *Proc. of BTW 2005*, pages 30–47, 2005.

[16] Robert Baumgartner, Wolfgang Gatterbauer, and Georg Gottlob. Web data extraction system. In *Encyclopedia of Database Systems*, pages 3465–3471. Springer, 2009.

[17] Robert Baumgartner, Georg Gottlob, and Marcus Herzog. Scalable web data extraction for online market intelligence. *Proceedings of the VLDB Endowment*, 2(2):1512–1523, 2009.

[18] Robert Baumgartner and G Ledermiiller. Deepweb navigation in web data extraction. In *Computational Intelligence for Modelling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on*, volume 2, pages 698–703. IEEE, 2005.

[19] T. Berners-Lee. The World Wide Web: Past, Present and Future. `https://www.w3.org/People/Berners-Lee/1996/ppf.html`, 1996.

[20] Tim Berners-Lee. Enabling Standards & Technologies - Layer Cake. `http://www.w3.org/2002/Talks/04sweb/slide12-0.html`. [Accessed 20/11/2014].

[21] Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, 284(5):1–5, 2001.

[22] Ricardo Buettner. Predicting user behavior in electronic markets based on personality-mining in large online social networks. *Electronic Markets*, pages 1–19, 2016.

[23] Carlos Caldeira. *Data Warehousing*, 2012. Edições Sílabo.

[24] Bill Carroll and Judy Siguaw. The evolution of electronic distribution: Effects on hotels and intermediaries. *Cornell Hospitality Quarterly*, 44(4):38, 2003.

[25] Malu Castellanos, Florian Daniel, Irene Garrigós, and Jose-Norberto Mazón. Business Intelligence and the Web. *Information Systems Frontiers*, 15(3):307, 2013.

[26] Surajit Chaudhuri, Umeshwar Dayal, and Vivek Narasayya. An overview of business intelligence technology. *Communications of the ACM*, 54(8):88–98, 2011.

[27] Peter Pin-Shan Chen. The entity-relationship model—toward a unified view of data. *ACM Transactions on Database Systems (TODS)*, 1(1):9–36, 1976.

[28] R. Choupina, M.B. Correia, C.M.Q Ramos, D. Martins, and F. Serra. Guest Reputation Indexes to Analyze the Hotel's Online Reputation Using Data Extracted from OTAs. *Recent Researches in Applied Informatics*, pages 50–59, 2015.

[29] Edgar F Codd. A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377–387, 1970.

[30] Valter Crescenzi, Giansalvatore Mecca, Paolo Merialdo, et al. Roadrunner: Towards automatic data extraction from large web sites. In *Proc. VLDB*, volume 1, pages 109–118, 2001.

[31] Ljubica Knežević Cvelbar and Larry Dwyer. An importance–performance analysis of sustainability factors for long-term strategy planning in Slovenian hotels. *Journal of sustainable tourism*, 21(3):487–504, 2013.

[32] B. Andrlic D. Ruzic and I. Ruzic. Web 2.0 promotion techniques in hospitality industry. *International Journal of Management Cases*, 13(4):310–319, 2011.

[33] Claudia Deco, Cristina Bender, and Adrián Ponce. Proposal of an ontology based web search engine. In *XIV Congreso Argentino de Ciencias de la Computación*, 2008.

[34] Google Developers. The Google Maps Geolocation API. `https://www.developers.google.com/maps/documentation/geolocation/intro`. [Accessed 03/08/2016].

[35] Francesco Di Tria, Ezio Lefons, and Filippo Tangorra. Big Data Warehouse Automatic Design Methodology. *Big Data Management, Technologies, and Applications*, pages 115–149, 2014.

[36] Alex Dietz. "Big Data" Revenue Management. `http://blogs.sas.com/content/hospitality/2013/05/17/big-data-revenue-management/`, 2014. [Accessed 28/10/2014].

[37] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422. Citeseer, 2006.

[38] Christiane Fellbaum et al. WordNet: An electronic lexical database MIT Press. *Cambridge MA*, 1998.

[39] Emilio Ferrara, Pasquale De Meo, Giacomo Fiumara, and Robert Baumgartner. Web data extraction, applications and techniques: A survey. *Knowledge-Based Systems*, 70(0):301 – 323, 2014.

[40] Matthias Fuchs, Wolfram Höpken, and Maria Lexhagen. Big data analytics for knowledge generation in tourism destinations–A case from Sweden. *Journal of Destination Marketing & Management*, 3(4):198–209, 2014.

[41] Ali Ghobadi and Maseud Rahgozar. An Ontologybased Semantic Extraction Approach for B2C eCommerce. *The International Arab Journal of Information Technology*, 8(2):163–170, 2011.

[42] GuestCentric.com. Booking.com: Your worst best friend? `http://www.guestcentric.com/bookingcom-your-worst-best-friend/`, 2014. [Accessed 28/10/2014].

[43] Jin-Xing Hao, Yan Yu, Rob Law, and Davis Ka Chio Fong. A genetic algorithm-based learning approach to understand customer satisfaction with OTA websites. *Tourism Management*, 48:231–241, 2015.

[44] Caryl Helsel and Kathleen Cullen. Dynamic Packaging–2005 White Paper series. *Hotel Electronic Distribution Network Association HEDNA, the SolutionZ Group, VA*, 2005.

[45] G. Hill. A guide to enterprise reporting. `http://ghill.customer.netspace.net.au/reporting/definition.html`. [Accessed 31/03/2015].

[46] Wolfram Höpken, Matthias Fuchs, Gerhard Höll, Dimitri Keil, and Maria Lexhagen. *Multi-dimensional data modelling for a tourism destination data warehouse*. Springer, 2013.

[47] JSON. Javascript Object Notation. `http://www.json.org/`, 2014. [Accessed 28/10/2014].

[48] Devkant Kala and Satish Chandra Bagri. Balanced Scorecard Usage and Performance of Hotels: A Study from the Tourist State of Uttarakhand, India. *Asia-Pacific Journal of Innovation in Hospitality and Tourism (APJIHT)*, 3(2):1–21, 2014.

[49] M. Kende. Internet Society Global Internet Report 2014. `https://www.internetsociety.org/sites/default/files/Global_Internet_Report_2014_0.pdf`. [Accessed 25/04/2015].

[50] TaKeaways Key. The Forrester Wave: Big Data Streaming Analytics Platforms. 2014.

[51] J. Kreutzer and N. Witte. Opinion Mining using SentiWordNet. `http://stp.lingfil.uu.se/~santinim/sais/Ass1_Essays/Neele_Julia_SentiWordNet_V01`. [Accessed 01/05/2015].

[52] Kenneth C Laudon and Jane P Laudon. Management information systems: managing the digital firm. *New Jersey*, 8, 2004.

[53] Rob Law. Internet and Tourism—Part XXI: TripAdvisor. *Journal of Travel & Tourism Marketing*, 20(1):75–77, 2006.

[54] Rob Law and Freddy Chen. Internet in travel and tourism-part II: Expedia. *Journal of Travel & Tourism Marketing*, 9(4):83–87, 2000.

[55] Kwang-Ho Lee and Sunghyup Sean Hyun. A model of behavioral intentions to follow online travel advice based on social and emotional loneliness scales in the context of online travel communities: The moderating role of emotional expressivity. *Tourism Management*, 48:426–438, 2015.

[56] Kristina Lerman, Craig Knoblock, and Steven Minton. Automatic data extraction from lists and tables in web sources. In *IJCAI-2001 Workshop on Adaptive Text Extraction and Mining*, volume 98, 2001.

[57] Bing Liu. *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Science & Business Media, 2007.

[58] Wei Liu, Xiaofeng Meng, and Weiyi Meng. Vide: A vision-based approach for deep web data extraction. *IEEE Tr. on Knowledge and Data Engineering*, 22(3):447–460, 2010.

[59] Chaves Marcirio and Trojahn Cassia. Towards a multilingual ontology for ontology-driven content mining in social web sites. Proceedings of the ISWC 2010 Workshops, Volume I, 1st International Workshop on Cross-Cultural and Cross-Lingual Aspects of the Semantic Web, 2010.

[60] Edison Marrese-Taylor, Juan D Velásquez, and Felipe Bravo-Marquez. A novel deterministic approach for aspect-based opinion mining in tourism products reviews. *Expert Systems with Applications*, 41(17):7764–7775, 2014.

[61] D. Martins, R. Lam, J.M.F. Rodrigues, P.J.S. Cardoso, and F. Serra. A Web Crawler Framework for Revenue Management. In *14th International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases*, pages 88–97, 2015.

[62] D. Martins, C.M.Q Ramos, J.M.F. Rodrigues, P.J.S. Cardoso, R. Lam, and F. Serra. Challenges in Building a Big Data Warehouse Applied to the Hotel Business Intelligence. *Recent Research in Applied Informatics*, pages 110–117, 2015.

[63] Kevin May. Crawling is the new API – a legal and technical rough guide for the travel industry. `http://www.tnooz.com/article/APInew-crawling-legal-technical-guide/`. [Accessed 21/11/2014].

[64] George A Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[65] Soumendra Mohanty, Madhu Jagadeesh, and Harsha Srivatsa. *Big Data imperatives: enterprise Big Data warehouse, BI implementations and analytics*. Apress, 2013.

[66] Soumendra Mohanty, Madhu Jagadeesh, and Harsha Srivatsa. *Big Data imperatives: enterprise Big Data warehouse, BI implementations and analytics*. Apress, 2013.

[67] Inc. MongoDB. MongoDB for GIANT Ideas. `http://www.mongodb.com/`, 2014. [Accessed 28/10/2014].

[68] Jorge Morato, Miguel Angel Marzal, Juan Lloréns, and José Moreiro. Wordnet applications. In *Global Wordnet Conference*, volume 2, pages 270–278, 2004.

[69] Sergey Odinokov. Hangfire – Background jobs and workers for ASP.NET. `http://www.hangfire.io/`, 2013. [Accessed 10/07/2016].

[70] Bob Offutt. Big Data: Redefining Travel Business Decision Making. A White Paper Sponsored by UNIT4 Business, Phocuswright, 2014.

[71] Harmen Oppewal, Twan Huybers, and Geoffrey I Crouch. Tourist destination and experience choice: A choice experimental analysis of decision sequence effects. *Tourism Management*, 48:467–476, 2015.

[72] Nikolaos K Papadakis, Dimitrios Skoutas, Konstantinos Raftopoulos, and Theodora A Varvarigou. Stavies: A system for information extraction from unknown web data sources through automatic web wrapper generation using clustering techniques. *Knowledge and Data Engineering, IEEE Transactions on*, 17(12):1638–1652, 2005.

[73] David Parmenter. *Key performance indicators: developing, implementing, and using winning KPIs*. John Wiley & Sons, 2015.

[74] Tim Peter. Use hotel data to drive growth. `http://www.hotelnewsnow.com/Article/14553/Use-hotel-data-to-drive-growth`, 2014. [Accessed 21/05/2015].

[75] Taofen Qiu and Tianqi Yang. Automatic information extraction from E-Commerce web sites. In *Proc. Int. Conf. on E-Business and E-Government (ICEE)*, pages 1399–1402. IEEE, 2010.

[76] Budi Rahardjo and Roland HC Yap. Automatic information extraction from web pages. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 430–431. ACM, 2001.

[77] C.M.Q. Ramos, M.B. Correia, J.M.F. Rodrigues, D. Martins, and F. Serra. Big data warehouse framework for smart revenue management. *Advances in Environmental Science and Energy Planning*, pages 13–22, 2015.

[78] Eric Redmond and Jim R Wilson. *Seven databases in seven weeks: a guide to modern databases and the NoSQL movement*. Pragmatic Bookshelf, 2012.

[79] Davi de Castro Reis, Paulo Braz Golgher, Altigran Soares Silva, and AlbertoF Laender. Automatic web news extraction using tree edit distance. In *Proceedings of the 13th international conference on World Wide Web*, pages 502–511. ACM, 2004.

[80] Olivia Parr Rud. *Business intelligence success factors: tools for aligning your business in the global economy*, volume 18. John Wiley & Sons, 2009.

[81] Philip Russom et al. Big data analytics. *TDWI Best Practices Report, Fourth Quarter*, pages 1–35, 2011.

[82] Maribel Yasmina Santos and Isabel Ramos. *Business Intelligence: Tecnologias da informação na gestão de conhecimento*. FCA-Editora de Informática, 2006.

[83] Michael Schermann, Holmer Hemsen, Christoph Buchmüller, Till Bitter, Helmut Krcmar, Volker Markl, and Thomas Hoeren. Big Data. *Business & Information Systems Engineering*, 6(5):261–266, 2014.

[84] Maja Šerić, Irene Gil-Saura, and Alejandro Mollá-Descals. Loyalty in high-quality hotels of Croatia: From marketing initiatives to customer brand loyalty creation. *Journal of Relationship Marketing*, 12(2):114–140, 2013.

[85] Marianna Sigala. Exploiting Web 2.0 for new service development: findings and implications from the Greek tourism industry. *International Journal of Tourism Research*, 14(6):551–566, 2012.

[86] Chromium team. Chromedriver. `https://code.google.com/p/selenium/wiki/ChromeDriver`, 2014. [Accessed 28/10/2014].

[87] Diogo Teixeira and Isabel Azevedo. Análise de opiniões expressas nas redes sociais. *RISTI-Revista Ibérica de Sistemas e Tecnologias de Informação*, (8):53–65, 2011.

[88] Mladen Varga and Katarina Curko. Some aspects of information systems integration. In *MIPRO, 2012 Proceedings of the 35th International Convention*, pages 1583–1588. IEEE, 2012.

[89] W3C. World Wide Web Consortium. `http://www.w3.org/`. [Accessed 28/10/2014].

[90] W3C. w3c.org/webdriver. `http://www.w3.org/TR/webdriver/`, 2014. [Accessed 25/11/2014].

[91] W3C. w3c.org/xpath. `http://www.w3.org/TR/xpath/`, 2014. [Accessed 28/10/2014].

[92] Albert Weichselbraun, Stefan Gindl, and Arno Scharl. Enriching semantic knowledge bases for opinion mining in big data applications. *Knowledge-based systems*, 69:78–85, 2014.

[93] Michael J Welch, Junghoo Cho, and Christopher Olston. Search result diversity for informational queries. In *Proceedings of the 20th international conference on World wide web*, pages 237–246. ACM, 2011.

[94] Jochen Wirtz, Sheryl E Kimes, Jeannette Ho Pheng Theng, and Paul Patterson. Revenue management: Resolving potential customer conflicts. *Journal of Revenue and Pricing Management*, 2(3):216–226, 2003.

[95] Zheng Xiang, Zvi Schwartz, John H Gerdes, and Muzaffer Uysal. What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management*, 44:120–130, 2015.

[96] YY Yao, Ning Zhong, Jiming Liu, and Setsuo Ohsuga. Web Intelligence (WI) research challenges and trends in the new information age. In *Web intelligence: Research and development*, pages 1–17. Springer, 2001.

[97] Yanhong Zhai and Bing Liu. Extracting web data using instance-based learning. In *Web Information Systems Engineering–WISE 2005*, pages 318–331. Springer, 2005.

[98] Hongkun Zhao, Weiyi Meng, Zonghuan Wu, Vijay Raghavan, and Clement Yu. Fully automatic wrapper generation for search engines. In *Proceedings of the 14th international conference on World Wide Web*, pages 66–75. ACM, 2005.