UNIVERSITÉ DU QUÉBEC À MONTRÉAL

BENEFITS OF THE APPLICATION OF WEB-MINING METHODS AND TECHNIQUES FOR THE FIELD OF ANALYTICAL CUSTOMER RELATIONSHIP MANAGEMENT OF THE MARKETING FUNCTION IN A KNOWLEDGE MANAGEMENT PERSPECTIVE

THESIS

PRESENTED

AS PARTIAL REQUIREMENT

OF THE MASTERS IN MANAGEMENT SCIENCES

BY MYRIAM

ERTZ

DECEMBER 2012

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

LES BENEFICES DE L'APPLICATION DES METHODES ET TECHNIQUES DE WEB-MINING POUR LA BRANCHE DE LA GESTION DE LA RELATION CLIENT ANALYTIQUE DE LA FONCTION MARKETING DANS UNE PERSPECTIVE DE GESTION DE LA CONNAISSANCE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE ÈS SCIENCES DE LA GESTION

PAR MYRIAM

ERTZ

DÉCEMBRE 2012

REMERCIEMENTS

J'ai rédigé le présent mémoire dans le cadre de ma Maîtrise ès Sciences de la Gestion (MSG), au sein de l'École des Sciences de la Gestion (ÉSG) de l'Université du Québec à Montréal (UQÀM).

Bien que je fusse en charge de réaliser la part du lion de ce projet, beaucoup de personnes y ont été impliquées, à des degrés variables, contribuant, de fait, directement ou indirectement à la réalisation du présent travail de recherche. Je souhaite leur exprimer ma sincère gratitude à tous.

Tout d'abord, je remercie mon superviseur, Prof. Dr. Graf, professeur titulaire au sein du Département Marketing de l'ÉSG UQÀM. Il me supporta généreusement dans la conduite de mon projet, me fournit un appui académique et théorique précieux et fit preuve de flexibilité et d'une grande mansuétude dans mes requêtes. Je remercie également les Prof. Dr. Zuccaro et Prof. Dr. Daghfous, qui m'ont fourni les outils et la méthodologie nécessaires pour approfondir le domaine si vaste mais non moins fascinant du data-mining. Mes pensées vont également au Prof. Dr. Arcand pour son initiation à la dimension virtuelle et multicanal du marketing, laquelle m'a permis d'approfondir la perspective managériale du projet. Je suis reconnaissante à toutes ces personnes, et plus largement, aux Départements Marketing de l'ÉSG UQÀM, de l'Université Laval, des Hautes Études Commerciales (HEC) Montréal et de l'Université de Sherbrooke (UDS) pour m'avoir aidé à relever le défi de trouver des répondants qualifiés pour participer à l'étude exploratoire que comporte ce projet.

Je suis également reconnaissante à l'Association Canadienne du Marketing (ACM), et à tous les membres du Customer Insight & Analytics Council de l'ACM, en particulier à M. Jordan Sandler et à MME. Sandra Singer, pour les contacts professionnels pertinents qu'ils ont pu me fournir dans le domaine du web mining, au Canada. De même, je remercie les membres du Chapitre de la Recherche Marketing de l'Association Américaine du Marketing (AMA), en particulier Sima Dahl, consultante chez Parlay Communications ainsi que John Luna, gestionnaire des affaires internet de l'AMA, pour m'avoir introduite à leurs réseaux professionnels d'experts en web-mining et de fait, pour m'avoir soutenue dans ma recherche de répondants. Je remercie également Regina

Malina, directrice du marketing chez Intact Insurance, Richard Boire, co-fondateur de Boire Filler Group, Peter Edry, gestionnaire et développeur des ventes chez Brilig.com, ainsi que Baskhar Patel, analyste d'affaires chez Cybersoft, pour m'avoir accordé un peu de leur temps afin de me donner une vision éclairée des défis et potentialités du web-mining.

Il va sans dire que la réussite de ce projet reposait fortement sur les avis et opinions de professionnels experts en web-mining, ou tout du moins, en data-mining. Ainsi, je remercie chaleureusement l'ensemble des répondants ayant participé avec enthousiasme et spontanéité aux entretiens en profondeur. Certains ont déjà été cités précédemment et je tenais bien évidemment à remercier tous les autres.

Au niveau personnel, je remercie ma famille pour leur soutien, leur compréhension et la patience dont ils ont fait preuve tout au long de mon projet de recherche.

J'omets certainement de mentionner des personnes et je m'excuse de n'avoir pu les citer ici. Qu'ils soient assurés que ma sincère gratitude leur revient à tous.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS, ACRONYMES AND INITIALS

aCRM	Analytical Customer Relationship Management
ADSL	Asymmetric Digital Subscriber Line
AI	Artificial Intelligence
AR	Augmented Reality
BI	Business Intelligence
BT	Behavioral Targeting
CAD	Canadian Dollars
CAR	Classification Association Rule
CEFRIO	CEntre Facilitant la Recherche et l'Innovation dans les Organisations
CF	Collaborative Filtering
CFI	Closed Frequent Itemset
CLF	Common Log Format
CLV	Customer Lifetime Value
CPM	Cost Per Mille
CPC	Cost Per Click
CPA	Cost Per Action
CPI	Cost Per Impression
CPT	Cost Per Thousand
CRM	Customer Relationship Management
CSF	Critical Success Factor
CTI	Click Through Interest
CTR	Click-Through Rate
DA	Discriminant Analysis
DB	Database
DM	Data Mining

DW	Data Warehousing
eCRM	Electronic Customer Relationship Management
ERP	Enterprise Resource Planning
ERPS	Enterprise Resource Planning System
ESML	Earth Science Markup Language
ETL	Extract Transform Load
FAQ	Frequently-Asked Question
GLM	General Linear Model
GMT	Greenwich Mean Time
HTTP	Hypertext Transfer Protocol
IE	Information Extraction
IP	Internet Protocol
IR	Information Retrieval
IS	Information System
ISP	Internet Service Provider
IT	Information Technology
KDD	Knowledge Discovery in Databases
KM	Knowledge Management
KSF	Key Success Factor
LR	Logistic Regression
mCRM	Mobile Customer Relationship Management
MFI	Maximum Frequent Itemset
MIS	Marketing Information System
ML	Machine Learning
NLP	Natural Language Processing
NPV	Net Present Value

oCRM	Operational Customer Relationship Management
OECD	Organisation for Economic Co-operation and Development
OLAP	On-Line Analytical Processing
P	Proposition
PPC	Pay Per Click
RFM	Recency Frequency Monetary
ROMI	Return On Marketing Investment
RQ	Research Question
RS	Recommendation System
sCRM	Social Customer Relationship Management
SEO	Search Engine Optimization
SP	Sequential Patterns
SPSS	Statistical Package for Social Sciences
SVM	Support Vector Machine
SVR	Support Vector Regression
tCRM	Technical Customer Relationship Management
URL	Uniform Resource Locator
VTR	View Through Rate
WCM	Web Content Mining
WM	Web Mining
WSM	Web Structure Mining
WUM	Web Usage Mining
WWW	World Wide Web
XLF	Extended Log Format
XML	Extended Markup Language

LIST OF EQUATIONS

LIST OF SYMBOLS

AC	Acquisition Cost
	Margin produced by the customer in each time period n
	Cost of marketing and serving the customer,
p	Probability the customer will not defect in one year
N	Total number of years or time periods
Σ	Sum

RÉSUMÉ

Le Web Mining (WM) reste une technologie relativement méconnue. Toutefois, si elle est utilisée adéquatement, elle s'avère être d'une grande utilité pour l'identification des profils et des comportements des clients prospects et existants, dans un contexte internet. Les avancées techniques du WM améliorent grandement le volet analytique de la Gestion de la Relation Client (GRC). Cette étude suit une approche exploratoire afin de déterminer si le WM atteint, à lui seul, tous les objectifs fondamentaux de la GRC, ou le cas échéant, devrait être utilisé de manière conjointe avec la recherche marketing traditionnelle et les méthodes classiques de la GRC analytique (GRCa) pour optimiser la GRC, et de fait le marketing, dans un contexte internet. La connaissance obtenue par le WM peut ensuite être administrée au sein de l'organisation dans un cadre de Gestion de la Connaissance (GC), afin d'optimiser les relations avec les clients nouveaux et/ou existants, améliorer leur expérience client et ultimement, leur fournir de la meilleure valeur. Dans un cadre de recherche exploratoire, des entrevues semi-structurés et en profondeur furent menées afin d'obtenir le point de vue de plusieurs experts en (web) data mining. L'étude révéla que le WM est bien approprié pour segmenter les clients prospects et existants, pour comprendre les comportements transactionnels en ligne des clients existants et prospects, ainsi que pour déterminer le statut de loyauté (ou de défection) des clients existants. Il est à constitue, à ce titre, un outil d'une redoutable efficacité prédictive par le biais de la classification et de l'estimation, mais aussi descriptive par le biais de la segmentation et de l'association. En revanche, le WM est moins performant dans la compréhension des dimensions sous-jacentes, moins évidentes du comportement client. L'utilisation du WM est moins appropriée pour remplir des objectifs liés à la description de la manière dont les clients existants ou prospects développent loyauté, satisfaction, défection ou attachement envers une enseigne sur internet. Cet exercice est d'autant plus difficile que la communication multicanale dans laquelle évoluent les consommateurs a une forte influence sur les relations qu'ils développent avec une marque. Ainsi le comportement en ligne ne serait qu'une transposition ou tout du moins une extension du comportement du consommateur lorsqu'il n'est pas en ligne. Le WM est également un outil relativement incomplet pour identifier le développement de la défection vers et depuis les concurrents ainsi que le développement de la loyauté envers ces derniers. Le WM nécessite toujours d'être complété par la recherche marketing traditionnelle afin d'atteindre ces objectives plus difficiles mais essentiels de la GRCa. Finalement, les conclusions de cette recherché sont principalement dirigées à l'encontre des firmes et des gestionnaires plus que du côté des clients-internautes, car ces premiers plus que ces derniers possèdent les ressources et les processus pour mettre en œuvre les projets de recherche en WM décrits.

Mots-clés

WEB MINING, GESTION DE LA CONNAISSANCE, GESTION DE LA RELATION CLIENT, DONNÉES INTERNET, COMPORTEMENT DU CONSOMMATEUR, FORAGE DE DONNÉES, CONNAISSANCE DU CONSOMMATEUR

ABSTRACT

Web Mining (WM), remains a relatively unknown technology. However, if used appropriately, it can be of great use to the identification of existing or prospective customers' profiles and behaviours online. The recent technical advances in the field of Web-Mining enhance tremendously the analytical side of Customer Relationship Management (CRM), which is still usually related to a simple transactional function. This study, follows an exploratory approach to assess whether Web-Mining fulfills, alone, all the core objectives of CRM and thus should be used in conjunction with traditional marketing research and other classic aCRM methods to optimize CRM, and hence marketing, in a web context. WM-derivated knowledge can then be managed within the company in a Knowledge Management (KM) framework, in order to optimize relationships with new and/or existing customers, improve their customer experience and ultimately deliver greater value to customers. Following an exploratory research design type, several in-depth semi-structured interviews were conducted in order to get the insight of several (web) data mining experts. The study revealed that web-mining is very well suited to profile existing and prospective web customers, to understand transactional web behaviour(s) (navigation patterns, amount of purchases by week, by month, by region, cross-selling and up-selling opportunities, etc.) of existing and prospective customers and to determine loyalty (or defection) statuses of existing web customers. Therefore, web-mining is a tremendous efficient predictive tool via the wide array of classification and estimation potentialities that it offers. It is also a formidable efficient descriptive tool since it offers a plethora of clustering/segmenting as well as association potentialities. Nevertheless, Web-Mining does not well in understanding the less obvious and underlying dimensions of customer behaviour, the is for instance describing how existing or prospective customers develop satisfaction, loyalty, defection and attachment on the web. Also, Web-Mining is relatively weak a tool to identify defection patterns to and from competitors as well as loyalty patterns towards them. Web-Mining still needs to be complemented with traditional marketing research in order to reach those more difficult but essential objectives of aCRM. Eventually, the conclusions of this piece of research are mainly aimed at companies and managers' intention, because they more than the client have the resources and processes to implement the discussed Web-Mining research projects.

Keywords

WEB MINING, KNOWLEDGE MANAGEMENT, CUSTOMER RELATIONSHIP MANAGEMENT, WEB DATA, CUSTOMER BEHAVIOR, DATA MINING, CUSTOMER KNOWLEDGE.

INTRODUCTION

INTRODUCTION

Web-Mining (WM) techniques have been developed in order to analyze massive quantities of data and information spread through the internet. However, WM remains a mysterious, almost occult field for any neophyte. In the popular imagination it is often perceived as an entanglement of algorithms, artificial intelligence, machine learning and sophisticated statistics from which, one does not always grasp the depth and scope but finds it useful to discover "a diamond in a coal pile"[1]. Paradoxically, in an increasingly rationalized world, technology appears as a blue chip. WM is, therefore, appealing because since it is so complex and cabbalistic it must necessarily be a reliable method to obtain relevant results. Data-Mining (DM) and hence, WM would be a kind of recipe which extracts totally new, ready-to-use knowledge from databases. WM is, however, simultaneously a gold mine and a minefield, depending on how the organization implements it.

According to Shaw *et al.* (2001), marketing depends more and more on the web for customer data. Meanwhile, businesses tend to rely heavily on integrated processes which manage every aspect of customer relationships, such as Customer Relationship Management (CRM) systems. As such, WM needs to be considered as a fundamental part of the CRM issue. WM techniques offer undoubtedly technical advantages, cheap shortcuts and effective heuristics where costly marketing research devices would have been required or impossible to implement. Besides, traditional research methods are becoming increasingly more difficult to administer (Mihai, 2009). Surveys are perceived as being too long to complete, incurring high amounts of errors and biases that arise before, during and after the sampling and fieldwork processes. Surveys are also relatively loosely given the many adjustments and weightings required to obtain sound data (Kotler & Keller, 2006; Malhotra, 2010). This is a serious drawback, especially in a hypercompetitive socio-economic context, where time is money and any waste of time is a waste of money (Déry, 2010). As Winston Churchill said, "polls are the worst way of measuring public opinion except for all the others".

Instead, WM focuses on data generated in real-life situations, which is very close to the observation techniques used in primary marketing research. When applying WM, individuals

[1] SAS website: <http://www.sas.com/>

are generally unaware of the fact that the data they generate on the web will be used for analytical purposes, which decreases thus the risk of biases (social desirability, etc.) that may arise in experimental contexts. WM techniques offer a unique alternative. They avoid the hassle of primary research processes (Mihai, 2009 and "handle large, complete, integer, reliable, cheap data in a time-effective manner, by generating refined knowledge from gross data". They may even leverage information that might not have been obtained with traditional primary research, the "holy grail" of hidden information. As such, WM is a valuable component of the marketing research process and of the broader Knowledge Management (KM) framework.

The organization is a living organism in constant interaction with its environment and it needs to adapt to contextual trends and changes that occur in the environment (Rothschild, 1990; Kauffmann & Shapiro, 2001). Business will increasingly be done over the internet. Now is the time for organizations whose business models are not 100% web-based, to shift from their traditional view of the web being a minor complement to their sales to that of considering the web as a crucial distribution channel just like the others, if not more importantly, depending on the business' industry, products and/or services sold. In that respect, acquiring knowledge from, about and for consumers as well as savvy knowledge dissemination throughout the multiple organizational layers, in a highly efficient Knowledge Management perspective, is a prerequisite for becoming web-based sales champions. Along with, or in place of traditional marketing research, WM would be an ideal candidate to be used in a knowledge-enabled customer relationship management framework. Development, extraction, interpretation, dissemination and usage of such knowledge should, in fact, be done in a KM framework for optimizing returns.

However, despite all the advantages of WM for increasing web-based sales, WM still suffers from many drawbacks especially at the operational level. In addition, it needs to be further integrated into the CRM applet of the marketing function as well as into the KM framework of the entire organization. Companies that adopt market-oriented strategies to better compete on the global marketplace, need to assess objectively and critically the benefits of the WM methods and techniques and their utility as inputs of the organizational decision-making process. The whole challenge for organizations is to be able to diffuse and disseminate

efficiently the information derived from WM by being holistically integrated to the aCRM applet of CRM systems and to KM systems. Ultimately, as long as organizational culture (software) is consistent with organizational structure (hardware) (Allaire & Firsirotou, 1984), a cutting edge competitive advantage may be developed with such use of WM.

The purpose of this research is to analyze the benefits of *WM methods and techniques* on the *analytical CRM (aCRM)* applet of the marketing function, in the framework of a *Knowledge Management (KM) perspective*. More specifically, it seeks to understand how WM might be optimally integrated to the blended aCRM and KM framework to leverage insightful knowledge about existing and prospective web customers who interact with online businesses.

The first part consists of a thorough review of the relevant literature relating primarily to the concept of WM and aCRM integrated into the KM framework. The derived specific conceptual framework will be comprised of the research propositions to which this piece of research aims at providing insightful answers. The second part describes the research methodology comprised of the research design and data gathering processes. The fourth part discusses the results of the research process and the fifth part presents the academic and managerial implications of the findings. The conclusion closes the research project and provides limitations as well as directions for future research on the application of WM to the marketing domain.

ENVIRONMENTAL CONTEXT

Past information and projections

WM is both about internet and Data-Mining (DM). Internet was developed to enable simple access to static data. However, this initial conception evolved rapidly toward an on-demand, dynamic point-of-access to multimedia resources (catalogues, products), a concept which is heavily used for commercial purposes (Bazsalicza & Naim, 2001). Internet penetration reaches 77% in Canada (CEFRIO, 2011). From 2010 to 2015, in the Canadian economy only, retail commerce spending (from domestic and foreign sites) will increase by 87% reaching

CAD 30.9 billion[2]. Online shoppers aged 16 or more, who browsed, researched or compared products on the internet but have not necessarily bought them online, will represent 86.3% of internet users aged 16 or more, in 2015 (80.7% in 2010)[3]. Online buyers aged 16 or more, who have made at least one purchase online will represent 68.9% of internet users aged 16 or more, in 2015 (52.7% in 2010)[4]. In 2015, online buyers in Canada will spend, on average, CAD 1,928 per annum, a 7.1% increase from 2010[5]. It seems that from the customer's perspective, this is a structural shift and not merely a conjectural hype. The first interaction with a company will increasingly be done over the internet. The customer relationship moves irresistibly toward a digital level.

This may be a drawback because of the lack of human interaction, limited product selection, reservations about buying products sight unseen or high shipping costs. However, such an evolution may also be an opportunity since it is possible to offer entirely personalized contents (morphing), unique products and services, free shipping, flexible exchange and returns policies, multichannel convenience and superior customer service according to web users' profiles[6]. Not to mention the infinite possibilities offered by Augmented Reality (AR)[7] settings online. Also, the profile and behavior of web customers and shoppers can be analyzed into details with visitors' browsing traces. Here is where data-mining (DM) comes into play. DM roots are traced back along three major family lines: classical statistics (univariate and multivariate data analyses), Artificial Intelligence (AI) (application of human-thought-like processing to statistical problems by using Relational Database Management Systems, as opposed to statistics), and Machine Learning (ML) techniques applied to business applications, by letting computers learn about the data they study (blending of the

---

[2] Retail Ecommerce Spending in Canada, 2009-2015, eMarketer February 2011 (retrieved on 1-5-2011).
[3] Ibid.
[4] Ibid.
[5] Ibid.
[6] Canada Retail eCommerce Forecast: Measured Growth Ahead, 2009-2015, eMarketer February 2011 (retrieved on 16-5-2011).
[7] Augmented Reality is a live, (in)direct view of a physical, real-world environment whose elements are augmented by computer-generated sensory input such as sound, video, graphics or GPS data (a view of reality is modified by a computer).

AI heuristics with advanced statistical analysis)[8]. DM unifies historical and recent developments in machine learning, AI and statistics, to discover previously unknown patterns from large data sets. DM complements and even surpasses conventional research. In 2008, 35% of marketers found DM to be the most important marketing research technique to pay attention to. This proportion jumped to 43% in 2009 and is expected to grow for the next decades[9]. DM is also the technique that marketers use most often. 59% of them used it regularly in 2009 and 16% would like to use it in the near future[10]. WM is therefore the unique junction between the internet/e-commerce boom and the highly dynamic DM field.

Resources and constrains

The World Wide Web is a large and ever-growing database, which is a fertile area for WM research (Van Wel & Royakkers, 2004). In the framework of CRM, mining (frequent) patterns from large, unstructured and heterogeneous data streams constitutes challenging problems for a wide range of online applications *e.g.* retail chain data, network traffic, click-stream, telecommunication data, e-business or stock data analysis (Tanbeer *et al.,* 2009).

WM also requires whole website reconfigurations due to ill-structured designs. There are budgetary costs associated with such practices as well as maintenance times shutting down the website for unknown periods of time (Mihai, 2009). Consequently, website traffic, e-commerce activities, business image and hence profitability may suffer from WM activities. Besides, integrating WM procedures and techniques does not ultimately guarantee the discovery of relevant patterns in the data. There are many constrains that arise during the WM process. In practice, whenever the quality of the data is bad, the whole WM process is void (Rafea & El-Beltagy, 2006). Also, only a small portion of the information on the web is truly relevant and useful (Sivaramakrishnan & Balakrishnan, 2009). There is a huge job of sifting through the data involved with WM. It is time-consuming and may not necessarily yield the expected returns. Another critical problem of WM relates to the difficulty of

[8] A brief history of Data-Mining: < http://www.data-mining-software.com/data_mining_history.htm> (retrieved on 1-5-2011).
[9] Marketing Executives Networking Group (MENG), "MENG Marketing Trends Report 2009", January 5, 2009 (retrieved on 1-5-2011).
[10] Next Gen Market Research Group (NGMR), "NGMR MR Trends Survey I", January 7, 2010 (retrieved on 1-5-2011).

identifying users since two users can use the same computer, the same browser, the same IP (Internet Protocol) address and look at similar sets of pages, making it hard to analyze the behavior of one single user (Mihai, 2009).

WM entails also a whole set of normative issues relating to the protection of privacy and individuality (Wahlstrom *et al.,* 2010; Van Wel & Royakkers, 2004). The data may also be used for other purposes than those for which they were originally intended to (Adams, 2009). This raises serious ethical problems. Also, WM techniques are not self-sufficient. They will always require experts who can develop, implement WM projects and translate outputs into managerial terms.

Consumer behavior

WM is concerned with the discovery of patterns and trends in web visitors' online behaviors. The term "internet users" refers to those individuals who have access to the internet and who navigate on it for all sorts of purposes. Web users utilize the internet for sending messages, looking for information, participating in social networks, watching videos or listening to music (Baeza-Yates, 2009). In Canada, consumers were more willing to buy online in most product and service categories in 2009 than two years ago. The leading products purchased online by online buyers include travel services and arrangements, books/magazines/online newspapers, Clothing/jewelry/accessories or music[11]. For all product categories, web users use primarily the internet for researching items. Purchasing items comes in the second place[12]. Internet is, therefore, an important information source for web users. Also, people become more comfortable and accept it culturally to buy products they were usually reluctant to purchase online such as home electronics, apparel, luxury items and even bulky packaged goods such as diapers or cosmetics[13]. Besides, group buying sites such as Groupon, mobile shopping, free online shipping offers (Ali Azaria's Well.ca) have also attracted more low- and middle-income consumers who discover e-commerce's convenience, lower prices and

---

[11] Statistics Canada, « Canadian Internet Use Survey 2009", as cited in press release Sept.27, 2010 (retrieved on 16-05-2011)

[12] Aegis Mdeia, Vizeum and Carat Canada, "Consumer Connection Study (CCS", by Toluna, NOV.24 2010 (retrieved on 16-05-2011).

[13] Canada Retail Ecommerce Forecast: Measured Growth Ahead, 2009-2015, eMarketer February 2011 (retrieved on 16-5-2011).

broader product selection[14]. This precludes good prospects for a thorough increase of e-commerce, which will inflate the volume of web data available to web-based or web-extended businesses for WM purposes.

Legal environment

Intrinsic applications of WM practices, *i.e.*, behavioral targeting, geographic targeting, etc., are subject to a certain amount of criticism and objections because they violate ethical standards to various extents (Beales, 2010). WM may threaten such values as people's privacy or "the quality or condition of being secluded from the presence or view of others" (Vedder, 1999). WM may also threaten a strong Western value, individuality, or "the quality of being an individual; a human being regarded as a unique personality" (Van Wel & Royakkers, 2004). By the action of pressure groups such as consumer associations or lobbies, the authorities are led to develop legislations that limit abuses in web analytics practices. In the US, Congress is considering restrictive regulation to limit the use of information gleaned about people through their online information and behavior, by marketers[15]. Back in 1980, the OECD (Organization for Economic Cooperation Development) had already formulated some internationally accepted principles regarding the collection, use and unveiling of personal data[16]. In Europe, the OECD guidelines correspond to the European Directive 95/46/EC of the European Parliament (Council of 24 October 1995), and every European Union member state has to implement these basic guidelines in their national laws (Van Wel & Royakkers, 2004). Online data collection is increasingly scrutinized by governmental entities which discuss with panelists about the fact that advances in marketing targeting require more personal information (Learmonth, 2010). With regards to individuality, WM can lead to "de-individualization, which is defined as the tendency to judge and treat people on the basis of group characteristics instead of their own individual characteristics and

---

[14] Ibid.

[15] Learmonth, M. (2010). Holy Grail of Targeting is Fuel for Privacy Battle, Advertising Age, vol. 81, issue 12: <http://www.allbusiness.com/marketing-advertising/marketing-advertising/14183750-1.html> (retrieved on 23-3-2011).

[16] An on-line version can be found on: <http://www.oecd.org/dsti/sti/it/secur/prod/PRIV-EN.HTM> (retrieved on 23-3-2011).

merits" (Vedder, 1999). Privacy concerns should be handled now so that they do not grow bigger and lead to government regulation (Learmonth, 2010).

Economic environment

During the 1990s the focus has shifted from competition as developed by Porter, Juran, Deming, Covey or Quinn, to management techniques and organizations redesign. We are still witnessing the aftermath of strategic refocusing (Hamel & Pralahad, 1994), reengineering (Hammer & Champy, 1993) or restructuring (Goshal & Bartlett, 1997). The competitive heritage from the 70s-80s is also still very vivid and leads companies to become "adaptive", "agile", "lean", "co-opetitive" "cooperative" or "strategically aligned", to stay in tune with and to answer to environmental trends and shifts. In addition, the digital revolution calls irresistibly for a digital adaptation of increasingly more digitalized business models. This evolution leads to a hypermodern management where strategic and logistic problems arise in order to serve volatile, fragmented, globalized and socially-sensitive hyper-customers (Déry, 2010). The global economic context has turned into a hypercompetitive arena in which most western markets are becoming fragmented into multiple micro-segments where the level of targeting can become as low as that of the individual and converges toward hyper-specialization (Bousquet *et al.*, 2007). E-commerce offers unique opportunities in that respect. It enables mass customization for targeted sales and marketing through particle or one-to-one marketing (Kotler & Keller, 2006). To develop efficient e-tailorized marketing activities, businesses need to be fully aware of all their customers' aspects. Traditional marketing research has usually been and still is critical for efficient and effective studies on consumer behavior. But technological progresses have made it possible to analyze these behaviors as they take place on the spot, in digital contexts (the internet).

Marketing and technological competences

The integrated, accurate and systematic study of consumers has been enhanced by the emergence of automated Management and Marketing Information Systems (MIS)[17] which

[17] For an introduction to Marketing Information Systems see Chapter 4: Marketing Information Systems and Marketing Research", pp.94-117 of the book Marketing Management by Kotler, P. (1991).

dashboard the information of the organization into a holistic perspective and enable Database Marketing practices. In addition to operational databases, marketing databases deal specifically with model scores (*e.g.* balanced scorecards), appended data, Recency-Frequency-Monetary method (RFM-method), Customer Lifetime Value (CLV), promotions, responses, surveys and preferences[18].

Increased capacities in computer technology, computation means, database warehousing, and the constant progresses in the analysis of very large datasets have made Knowledge Discovery in the Database (KDD) very popular amongst marketing research departments (Adams, 2009). KDD enables the usage of Data-Mining (DM), the "non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data" (al-Fayyad, 1996). DM has become a must at the top of organizational Business Intelligence (BI) architectures[19]. WM is an extension of the initial DM field, embedded into a KDD framework, which is itself a procedure for obtaining BI. In a BI model, data is transformed into information, which gives rise to organizational knowledge, which is acquired and disseminated through the organization by means of Knowledge Management (KM), acting as a powerful platform for tactical and strategic decision-making purposes (Ranjan & Bhatnagar, 2011).

Since the mid-1990s, better prospects for DM in market research relate to new technology, specifically the internet, which includes analyzing WWW traffic and click-stream data, e-commerce and search marketing (Chiu & Tavella, 2008; Ortega & Aguillo, 2009). Marketers are then able to measure return on marketing investments (from web campaigns), develop web personalization, etc. To enhance customer service and relationships with key stakeholders, e-commerce is thus increasingly bundled with other organizational legacy systems such as CRM and the like, technologically-underpinned by KM and which tighten

---

[18] The practice of Database Marketing is fully developed in Hugues (2005). Strategic Database Marketing: <http://books.google.ca/books?hl=fr&lr=&id=Ws_cBzKDk8MC&oi=fnd&pg=PR7&dq=strategic+database+marketing+hugues&ots=TgCYjSUN4I&sig=1G7h6Z7uLHvHCsH1jrsrcaoYJfc#v=onepage&q&f=false> (retrieved on 23-3-2011).

[19] A thorough review of Business Intelligence systems, knowledge management and Enterprise Systems (ES) can be found in Turban et al. (2010). Decision Support and Business Intelligence Systems. Prentice Hall Press.

interdepartmental linkages and operations into BI meta-structures[20]. This enables the implementation of a (holistic) marketing concept as developed by Kotler and Keller (2006), integrating the marketing function organization-wide by means of IT deployment. It will be of utmost interest to analyze how integrating WM in such structures leverages critical business knowledge and creates, ultimately, an inimitable competitive advantage.

To perform WM, access to online information and analytics software is a prerequisite. According to a review by the KDnuggets, popular software include SPSS WM Clementine, which enables the user "to extract web events, including online campaign results, and to use this online behavior in Clementine's predictive modeling environment". SAS' Webhound on the other hand "analyzes web site traffic to answer questions like: who is visiting; how long do they stay? What are they looking at?". Megaputer PolyAnalyst integrates WM together with data and text mining because it does not have a separate module for WM demonstrated, while ClickTracks is a web metric tool that makes online behavior visible by showing information in context to the user (Zhang & Segall, 2008). Free and/or open-source software include Analog or ht://Miner[21].

Besides technical and technological gear, WM requires also mastering standard and sophisticated statistical techniques and procedures, especially those relating to DM (Machine Learning, Artificial Intelligence, etc.), to design appropriate WM procedures and interpret results adequately. The statistics-savvy expert must also possess knowledge of Data Warehousing (DW) and Information Systems (IS) in order to be able to Extract, Transform and Load (ETL model of BI) the data from their respective databases for exploratory and analytical purposes. Finally, a certain level of expertise with WM software is also required. Although decision-making based on WM outputs is generally made by top executives, the knowledge derived from WM should ideally be accessible to all relevant employees throughout the organization. In practice, it is very rare that one unique person possesses all those requirements. Rather, several people from different departments are generally gathered into cross-functional teams for a WM project.

---

[20] Sandeep Walia (2008), CRM and e-commerce the integration payoff : < http://www.crmbuyer.com/story/64103.html> (retrieved on 23-3-2011).

[21] For an exhaustive review of all the WM softwares available see also KDnuggets : http://www.kdnuggets.com/software/web-mining.html (retrieved on 23-3-2011).

PART ONE

LITERATURE REVIEW

CHAPTER I

CONCEPTUALIZATION OF WEB-MINING

1.1. From Data-Mining (DM) to Web-Mining (WM)

DM is « the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner » (Witten & Frank, 2005). DM is a mature technology suitable for a variety of primary and secondary analysis tasks on large data sets, especially in Customer Relationship Management (CRM) (Adams, 2009). WM is the application of DM methods and techniques to data that originate from the web.

In that respect, performing DM on the web entails slight differences compared to offline DM. The web data are inherently different from standardized data issued from data warehouses and data marts. They are log files (standard, cookies, specific), content about clients, about transactions, pre-defined rules (ontology base) and hyperlinks (web architecture). Their dissimilarity in nature and structure – some of them are qualitative and nominal, while others are quantitative and metric, discrete or continuous – make them a vast heterogeneous whole with little communalities to start with, and which requires tremendous pre-processing and cleansing to perform patterns discovery and analysis on a normalized scale. Secondly, the actions to be taken are individual in nature, because it is sought to personalize and customize the web relationship to each individual web visitor.

Therefore, when applied to the marketing science, WM seeks primarily to optimize the "internet relationship" (Bazsalicza & Naim, 2001) which is divided into several steps: entering the web site, surfing on the web site, choosing (pages, products, links, etc.), buying,

informing, returning to the web sites. Improving each of these steps requires using the appropriate type of data and analysis to obtain the desired output and use that output to improve the internet relationship step which is sought to be optimized. According to Bazsalicza and Naim (2001), the web experience becomes thus personalized and adaptive to end users to these steps of the internet relationship by:

- Identifying market segments and identifying the attributes of high value prospects
- Identifying key attributes of web clients for each product
- Selecting promotion strategies best suited to one particular web customer segment
- Improving web customers targeting
- Testing and identifying marketing actions having the most important impact
- Identifying customers most likely to be interested by new products
- Identifying best prospects online for a service
- Improving products cross-selling and up-selling on a web site
- Reducing costs and improving quality of contacts with customers
- Identifying the reasons of a churn or attrition and improve customer loyalty and retention
- Maximizing the impact of online advertising

Input formats of WM consist typically of intra-site and inter-site traffic data such as log files for global analyses (many users' browsing information on one or more websites), cookies for individual analysis (one user's browsing behavior on one or more website) and personal identification information for personal analysis (one user's browsing behavior on one specific site); transactional data (purchases, etc.); Information Systems data which may come from offline sources (ERP, etc.); or various other content data such as emails, registration data, images, videos, etc. These formats support data from different nature such as Recency Monetary Frequency or RFM-values (cross-tabulating recency of the last purchase in the period studied with the frequency of purchases in that period and examining the distribution of purchases in each intersection), profitability or economic value (difference between profits from a customer, segment, market, etc. and all the costs incurred to generate those profits), lifetime values (updated net value - including all the "predictives" such as propensity,

attrition, risk, etc. - of the expected future financial transactions with a customer, segment, market, etc.), data on products and contracts (numbers, mean product life, etc.), channel (preferred channel for ordering, etc.), relational data (interactions with the online business) attitudinal data (customer loyalty patterns), psychographic data (lifestyle, personality, values, risk aversion, knowledge, focus of interest, opinions and behavior, social styles), technical data (type of customer, payer status, etc.), sociodemographic or geographic data (Tufféry, 2011).

Outputs of WM consist of reports, segments, ready-to-use models, intelligent rules or direct recommendations for actions (Chiu & Tavella, 2008). They can be used for re-designing web site(s) (reorganization), enriching the database(s) with models to generate personalized contents according to a segment or score and integrating models into the database directly. Enriching the database requires to perform regularly batch analyses. However, integrating the model directly to the website, *i.e.,* to the application server, enables to apply the model in a real-time fashion to any new web user which saves time and costs (Bazsalicza & Naim, 2001). WM has therefore become a critical element of the marketing research process because it provides useful market(ing) intelligence (Büchner *et al.,* 1999).

1.2. The Web-Mining branches

The WM research area is a converging area from several research communities, such as Database (DB), Information Retrieval (IR)[22], Artificial Intelligence (AI)[23], Machine Learning (ML) and Natural Language Processing (NLP)[24] (Kosala & Blockeel, 2000). WM can be defined as the *"whole of DM and related techniques that are used to automatically discover and extract information from web documents and services"* (Cooley *et al.,* 1997; Kosala &

[22] In their book "Introduction to Information Retrieval", Manning, Raghavan and Schütze (2008) define Information Retrieval (IR) as finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

[23] According to Poole, Mackworth and Goebel (1998), Artificial Intelligence (AI) is a branch of computer science that "studies and designs intelligent agents."

[24] Eugene Charniak (1984) defines NLP in his book "Introduction to Artificial Intelligence" as the field of computer science and linguistics concerned with the interactions between computers and human (natural) languages.

Blockeel, 2000; Srivastava *et al.,* 2000). WM can be divided into three main categories (Madria *et al.,* 1999): Web Content Mining (WCM), Web Usage Mining (WUM) and Web Structure Mining (WSM). The three WM categories, summarized in Figure 1, can be used in isolation or in combination in a specific WM project, depending on the desired outcome. In contrast to content and structure data, web usage data is, however, not necessarily publicly available (Custers, 2001). The provider of access to web sites and the owner of the web sites visited by the user are in theory the only parties who are able to produce transaction logs (Van Wel & Royakkers, 2004). Web usage data is privately-owned and can neither be analyzed by agents (robots, spiders, etc.) or stored into databases, nor be purchased from third parties (tracking reports, psychographic surveys filled out, etc.). Each of these categories will be detailed below in more details.

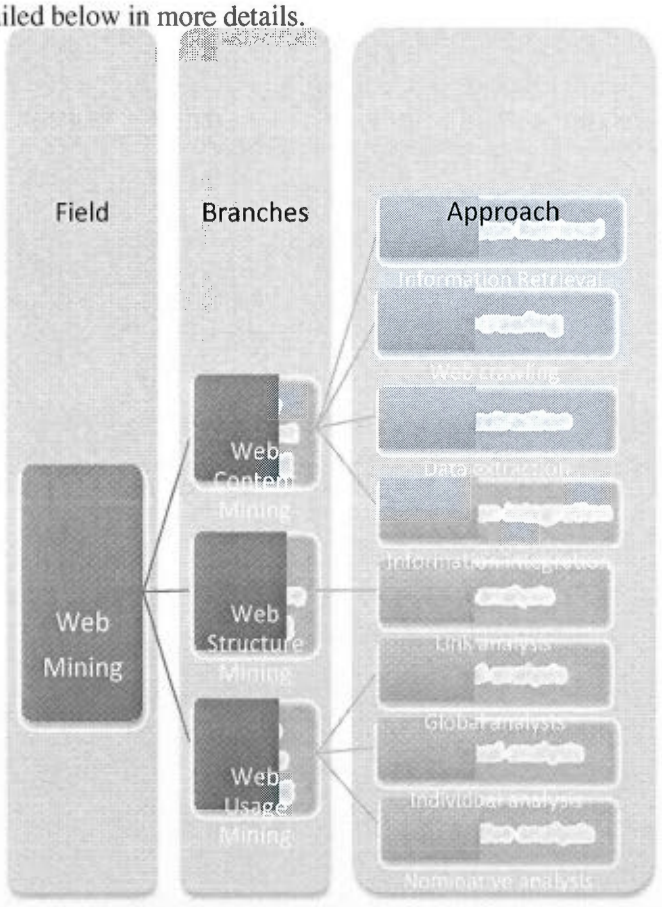

Figure 1.1 The branches of Web-Mining and their respective types of approaches. (Adapted from Srivastava *et al.,* 2000; Liu, 2011; Tufféry, 2011.)

1.2.1. Web Usage Mining (WUM)

Beyond, mere review of web site statistics, reports, OLAP (On-Line Analytical Processing) results and visualization of counting software data, the huge amount of data stored in backup systems needs to be mined to improve a firm's profit (Clendaniel, 2002). WUM is one of the most important approaches of WM in that respect. It aims at fulfilling the following tasks (Pabarskaite & Raudys, 2007; Tufféry, 2007):

- Reorganizing and optimizing the web site for fast and easy customer access
- Improving links and navigation patterns
- Attracting more advertisements capital by intelligent adverts
- Turning viewers into customers by better site architecture
- Monitoring the efficiency of the web site

WUM is a branch of WM that is concerned with the analysis of usage data, *i.e.* usage of how a web application is used (Pierrakos *et al.,* 2003). DM techniques are applied to the discovery of patterns based on *web log data* (Cho & Kim, 2004). WUM discovers the users' various access models by using the *log files* (web log data), which register all the clicks by each user during a web application interaction (Liu, 2007). More specifically, the log file is a written text recorded on the server of a web site and in which a line is written for each new request (change of page, downloading of a file, etc.) of the web user (Tufféry, 2007). However, these log files can also be collected at the level of the client (client log), or a proxy cache (proxy log), but the nature and structure of these log files remains basically the same which means that they can be combined together after standardization and pre-processing procedures for common analyses (Srivastava *et al.,* 2000).

<u>The major WUM approaches</u>

WM can be divided into three major approaches: global analyses, individual analyses and nominative analyses (Tufféry, 2007). The log files are very well-suited to overall analyses because they encompass general types of information. Some navigation information may be

useful although it is not possible to know to which web user to attach them *i.e.,* knowing that 40% of web users who visit page A also visit page B or that 20% of web users who visit page A immediately visit page C afterwards, does not enable organizations to draw one-to-one marketing strategies. But, they allow them to optimize the navigation in their web site or allow them to put ad banners or links at the right spot in the web site (Tufféry, 2007). With global analysis, organizations can group the web pages by their content, but they can construct taxonomies of users, according to number of pages visited, the files downloaded, etc., while matching those customer typologies with the customer databases of the business (Tufféry, 2011).

Global analyses

In typical global analyses, log files are used to record the user's behavior and how they interact with an application from the instant they access a site to the moment they leave the site (Vakali & Pallis, 2006). A site owner can actually see what a web user is looking at (Khabaza, 2000). The Common Log Format (CLF) contains the IP address of the client computer, the date and the time of the request, the type of request, the requested URL, the HTTP protocol, the return code of the server as well as the size (in bits) of the envoy (Tufféry, 2011). An example of a CLF is depicted below:

```
130.5.48.74 [22/May/2006:12:16:57 - 0100] "GET
    /content/index.htm  HTTP/1.1" 200 1243
```

This fragment of a log file, indicates a successful request (return code = 200) of downloading (GET) an item of 1243 bits, the 22nd of May 2006, at 12:16 with a time jetlag of -1 hour (-0100) from the Greenwich Mean Time (GMT) measure. More elaborated log formats include the Extended Log Format (XLF) and other formats for specific web sites such as securitized web sites (Mihai, 2009). The XLF indicates also the "referrer" (the web page from which the web user originates), the "user agent" (the browser), the operating system, the user login

between the IP address and the date (if it is known, but generally it is not) and other parameters[25], as depicted below:

```
130.5.48.74 - [22/May/2006:12:16:57 - 0100]
"GET/content/news.htm HTTP/1.1" 200 4504 "/content/index.htm"
"Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)".
```

In that case, the web user originated from the web page /content/index.htm with Internet Explorer 6.0 installed on Windows XP SP2 (Tufféry, 2007).

The whole process of collecting log files is done with Information Retrieval (IR) (Kosala & Blockeel, 2000). Basic web analytics tools such as Google Analytics, Web Trends, and the like, monitor and record constantly data generated by web users on a web site. These data are presented in basic descriptive statistics, advanced counting techniques such as OLAP[26], other knowledge query mechanisms and intelligent agents (Cio *et al.*, 2007). Given that log files are huge (hundreds of mega octets a day), several items are deleted such as pages visited by less than 5 IP addresses, image files (.jpeg, .gif, .jpg...), scripts, robots/agents accesses, links testers, outliers, (Tufféry, 2007). Tools such as Webalizer, Awstats or Analog monitor, extract and report in a structured way the most important information derived from those log files[27]. They record users' visits on the web such as user IP address, the date of the visit, the time of the beginning and of the end of the visit (working hours/night/week-end/holidays), the browser type (Internet Explorer, Firefox, Netscape, Opera...), the operating system (Windows, Linux, Mac...), the geographic origin, the visited pages, the number of visited pages, the average time spent on each web page, and the average number of clicks for further click-stream analysis (Mihai, 2009). Once the web log files have been selected and preprocessed, there are basically two main approaches for WUM (Mihai, 2009): (1) Mapping

---

[25] For a more detailed description of the various *return codes* and *types of request* that are generally visible in log files, see also Tufféry (2011), p.638-639.

[26] OLAP is an SQL-based methodology specifically used by datawarehouses (DW) to formulate and execute complex user queries, by executing complex read-based queries, providing multidimensional views (incl. GROUP BY) and aggregative functions (Roll up, Drill down, slice, dice, pivot-charts, etc). OLAP can be designed as ROLAP (Relational OLAP), MOLAP (Multidimensional OLAP) and HOLAP (Hybrid OLAP). See also, Cios et al., 2007.

[27] Software such as SPSS' Clementine ensures a highly automated cleaning process.

log data into relational tables: adopting a data mining technique to be performed on the data set partitioned in the relational tables, (2) Using log files directly: utilizing special preprocessing techniques for applying DM techniques, which refers to WUM.

Individual analyses

Individual analyses on the other hand, reveal for instance that web users, who look at the web page about cheese on a web site, also visit the web page about wine in the two following weeks. For such *one-to-one analyses* which integrate temporal parameters, log files are no longer sufficient. Other files are used such as *cookies* (files recorded on the user's hard disk while visiting a web site) which assign a login to the user's computer and other information about the navigation pattern of the user (the number of visited pages, entry pages, exit pages, originating web sites, downloaded files, some nominative information asked by the web site) and whenever the computer connects to a given web site the login in the cookie is transferred to the server and the content of the cookie can be automatically updated and propose personalized content, advertisements or web pages that may be of interest to the web user (Tufféry, 2007). These cookies however, can be easily deleted, blocked or refused by fire-walls, cleaning software or browsers, and may represent a threat to privacy. As for an IP address, the major drawback of a cookie lies in the fact that it is attached to a computer and that a computer may as well be used by one user or by one hundred different users which makes it difficult to effectively target a user on a one-to-one basis (Bazsalicza & Naim, 2001). For the log file, the difficulty lies in the fact that the IP address, the primary identifier of the user, is often non-permanent because it is dynamically attributed by the Internet Service Provider (ISP) at the moment of the connection except for ADSL connections (Tufféry, 2007). The user may also switch of computer or use the professional network of a company which may create double profiles. Individual analyses incur several disadvantages that need to be overcome for more insightful results.

Nominative analyses

Nominative analyses overcome the drawbacks of individual analysis. Users personally log in on a web site (online banking, etc.) with their unique login and password. This is generally possible when the web user is a known client of the organization that runs the web site.

Except in rare fraud cases, organizations can be sure of the identity of the web user and are thus able to add information about web behavior and patterns of the web users, in order to enrich the offline data that the organization already possesses about those clients. Also, the login step demonstrates of the genuine interest of the client for specific services or products and this information may be used for scoring purposes in order to predict the behavior of an online client (Tufféry, 2007).

WUM benefits

Web personalization

Adapting the content and structure of a web site according to a specific user in a mass-customization perspective leads to a more customer-centered approach. It is perhaps one of the most important achievements of WUM and it is often regarded as being an indispensable part of e-commerce (Van Wel & Royakkers, 2004). Web personalization is therefore a crucial element for effective (e-)marketing activities since customers find relevant information more quickly, feel that their needs are better answered and are therefore more satisfied and likely to be loyal to the organization on a long term basis. This leads to increased revenues and improved profitability. Automatic web personalization recommendation can be proposed by discovering usage profiles and frequent item sets selected by web users (Mobasher *et al.*, 2000a; Mobasher *et al.*, 2000b). Using the Customer Lifetime Value (CLF) and rating-based Collaborative Filtering (CF) methods, result in better Recommendation Systems (RS) and hence web personalization for the web user (Liu & Shih, 2005a; Liu & Shih, 2005b; Albadvi & Shahbazi, 2010). In addition to web server logs, Srivastava *et al.* (2000), also used client-side data (client logs) and web proxy caching (proxy logs) in order to characterize the web usage of visitors, modify the site, improve the system and personalize the web pages accordingly. This study also intends to include all the information derived from WUM tasks into Business Intelligence (BI) processes. Fenstermacher & Ginsburg (2003) argue that client-side data (client logs) can also be collected with standard office productivity tools, instead of only web browsers. In Joshi (2001a) and Joshi (2001b), it is argued that personalization should be done in three phases: (1) finding natural groups of clusters with clustering techniques, (2) studying which URLs tend to be requested together with

association rules and (3) analyzing the order in which URLs tend to be accessed with sequential analysis. WUM for web personalization purposes should even be done as soon as possible in order to track and accurately classify users' access patterns in real-time (Shahabi & Banaei-Kashani, 2002).

Search engines personalization

WUM is also used to create personalized search engines, which can understand a person's search queries in a personal way by analyzing and profiling the user's search behavior (Van Wel & Royakkers, 2004). Srivastava *et al.* (2000) used WUM for system improvement and site modification. Mining sequences of events and the routes connecting those events also betters web evaluation and web design (Spiliopoulou, 2000).

Trends prediction

The visitors' web usage patterns and trends for e-commerce might also be predicted (Abraham & Ramos, 2003). Extracting usage patterns enables organizations to determine the life-time value of customers, deploy marketing strategies across products and monitor the effectiveness of marketing campaigns (Cooley *et al.*, 2003). Analyzing user behavior and patterns can also be used to better respond to web users' needs and preferences for higher profits, *i.e.*, insert sweepstakes or web promotions.

Customer profiling

Also, many advantages of WUM are based on customer profiling (Van Wel & Royakkers, 2004). It is often more cost efficient to look at a group of web users instead of looking at each individual, because groups are cheaper and easier to approach (for instance by placing an ad in the right magazine instead of mailing every individual member of the group) (Custers, 2001). These so-called group profiles can also be of added value to individual profiles, because of the fact that some individual characteristics only become clear after looking at the individual from a group perspective (see also Custers, 2001). Clustering web users, web pages and identifying frequent access paths can thus also be used for marketing strategies, personalization and web site adaptation (Song & Shepperd, 2006). The characteristics of web usage can be used as a basis for clustering and segmentation (Srivastava *et al.*, 2000). Web

browsing patterns can also be detected from a group point of view with web user clustering (Song & Shepperd, 2006). Such patterns can complement web users' profiles that had already been developed beforehand.

Important software specifically dedicated to WUM includes, WUM, Netmind, SiteHelper, Oracle9iAS, SETA or Tellim (Pierrakos *et al.*, 2003).

1.2.2. Web Content Mining (WCM)

WCM is performed by extracting useful information from the content of a web page or web site (Zhang & Segall, 2008). The web content may consist of text, image, audio or video data, but the text format is the most widely used for WCM (Yeh *et al.*, 2009). WCM includes, in fact, extraction of structured data/information from web pages, identification, match, and integration of semantically similar data, opinion extraction from online sources, concept hierarchy, ontology and knowledge integration (Zhang & Segall, 2008). As for WUM, WCM is useful for marketing research since it allows marketers to learn about interests and preferences of existing or potential customers who are increasingly more averse to provide private information about them or answer to long surveys. WCM enables the collection and processing of customers' data that may be used in a customer-centered approach for direct marketing or other commercial activities. There are typically two types of web mining strategies: those that improve on the content search of other tools like search engines and those that directly mine the content of documents for better direct marketing strategies (Galeas, 2008).

Search Engine Optimization

The discipline of Search Engine Optimization (SEO) makes tremendous use of WCM in order to improve browsers, identify web sites that are of high general interest (authorities), learn which topics are related to each other by occurrence of links between topics, find related words by them often occurring in the same page, find keywords that are most typed by web users when looking for a particular topic (particularly relevant for optimizing web sites' natural and/or paying referencing and indexation) or improve web filters (Van Wel &

Royakkers, 2004). Zhang *et al.* (2003) developed a system of content-based image retrieval based on relevance feedback to refine the query or similarity measures in image search process to improve the retrieval performance compared to traditional approaches. The desired knowledge is often difficult to find in the large amount of information available. Chen *et al.* (2005) developed a prototype of intelligent search called iSEARCH to accurately represent knowledge in web documents. A good data preparation improves the performance of data-mining algorithms (Zhang & Segall, 2008). Mahboubi *et al.* (2007) transformed thus multiform and heterogeneous data into a unified format (the XML, Extended Markup Language) for warehousing purposes. Graves *et al.* (2007) handled heterogeneous data formats by using the Earth Science Markup Language (ESML) to create a DM web service designed specifically for science data.

Direct marketing strategies optimization

Content mining tools track down online misuse of brands and perform overall product reputation mining (Pang & Lee, 2008). In fact, Opinions can be extracted from various web sources with structured data extraction, information integration and Information Extraction (IE) from unstructured text, such as customer written comments, in order to perform opinion mining (Liu *et al.*, 2006). Product opinion mining is, therefore, particularly interesting in that respect, since it tracks and monitors the evolution of web users' opinions about specific goods, services, brands, websites, etc. WCM tools also analyze the content of competitive web sites in detail to gain valuable market competitive intelligence and eventually strategic advantages (Popescu & Etzioni, 2005). Mining the content of answers to online or offline marketing campaigns also enables organizations to evaluate those marketing campaigns (Zhang & Segall, 2008). With mining tools, online curriculum vitae or personal homepages can be collected and after interpreting the unstructured text information on these formats into business intelligence stored in a database, they can be used for marketing purposes (Lau *et al.*, 2005). Profiles on potential customers can be produced and more detailed information is added to profiles of (prospect) customers allowing organizations to prevent churn by better retaining their existing customers (Van Wel & Royakkers, 2004).

1.2.3. Web Structure Mining (WSM)

Instead of using log files such as in WUM or web page/site contents such as in WCM, WM can be performed by using the hyperlink structure of the web as an information source (Zhang & Segall, 2008). The process called WSM may be used for mapping user navigation patterns, web browsing patterns for e-commerce or discovering knowledge from hypertext data. Basically, WSM investigates how the web documents are structured and discovers the model underlying the link structure of the World Wide Web (Chaudhary, 2011). According to Furnkranz (2005), the web may be viewed as a graph with the documents as nodes and the hyperlinks between them as edges. The graph view can be used for effective retrieval and classification (Zhang & Segall, 2008). Consequently, WSM is particularly well-suited for mining and mapping navigation/browsing patterns and performing Search Engine Optimization (SEO) as with WCM and web site redesign.

Web site redesign

A given web site and the pages it encompasses are generally optimized to be as efficient and effective as possible for the web user. Fang and Sheng (2004) addressed that issue at the portal page level. They elaborated a heuristic approach to hyperlink selection based on relationships among hyperlinks (structural relationships that can be extracted from a web site and access relationship that can be discovered from a web log). Guan and McMullan (2005) designed a bookmark structure that allows individuals or groups of users to access the bookmark from anywhere on the internet and which includes more features (URL, document type, document title, keywords, date added, date last visited, etc.).

Search Engine Optimization (SEO)

The graph structure of the WWW on which draws the graph theory can be exploited to improve retrieval performance and classification since many search engines use graph properties in ranking their query results. In fact, the information of predecessor pages (pages that have a hyperlink pointing to the target page) can be used for enhancing text classification performance (Furnkranz, 2005). On the contrary, hypertext links, which are links that direct to another page or set of pages from the same website (in opposition to hyperlinks, which

direct to other pages or other sets of pages from other websites), can also be mined in order to discover knowledge with network analysis and machine learning for better SEO (Chakrabarti, 2003).

Discovery of navigation patterns

The sequence in which a web user selects and reviews web items (hyperlinks or hypertext links) can be of great use in order to derive standard and original navigation patterns. User navigation patterns can be mapped into visual tools that improve the understanding of the structure of the web site and the navigation behaviors of web users (Smith & Ng, 2003). Web users can be clustered according to the order in which web pages are requested and the different lengths of clustering sequences (Hay *et al.,* 2004). Song and Shepperd (2006) view the typology of a web site as a directed graph in which they can cluster users and URLs. They mine web browsing patterns for e-commerce purposes by identifying frequent access paths and sets of pages visited.

As for WCM, WSM complements WUM by shifting the focus from automatic log files that are empty of qualitative insight, toward the content (WCM) and the structure (WSM) that lies behind those log files. In short, WUM discovers user access patterns from usage logs; WCM mines knowledge from page contents and WSM discovers knowledge from hyperlinks (Liu, 2007, 2011). The different WM methods can be summarized as in Figure .

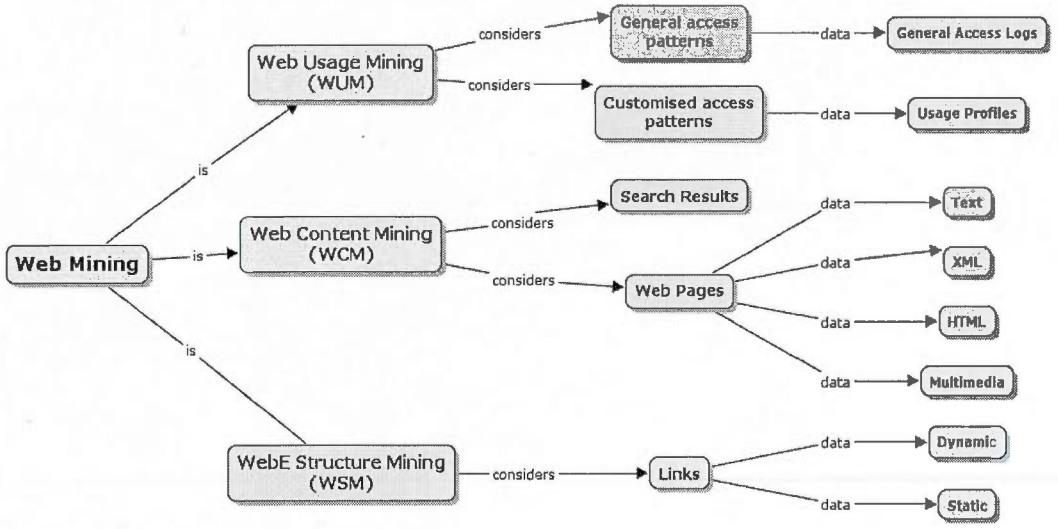

Figure 1.2 Conceptual map of Web-Mining. (Source: Inf@Vis.)

1.3. The Web-Mining process

1.3.1. The theoretical WM procedure

Based on Kosala & Blockeel's (2000) work, Zhang and Segall (2008) draw a five-stage WM process, which is particularly specific to WUM, but can be used in both WCM and WSM frameworks, as well:

[1] Resource finding and retrieving – the task of retrieving intended web documents.

[2] Information selection and preprocessing – automatically selecting and pre-processing specific information from retrieved web sources, which includes (Albadvi & Shahbazi, 2010):

- Data cleansing, user identification, session identification, path completion

[3] Patterns recognition and analysis – automatically discovering general patterns at individual web sites as well as across multiple sites by the use of descriptive and/or predictive WM techniques.

[4] Validation and interpretation – optimizing of the output(s) obtained in step 3 after applying it to test samples or by using other fine-tuning techniques.

[5] Visualization and implementation – implementing the output of the WM process in real-world web applications.

The fifth step, originally called "visualization" only, has been adapted here to business contexts by adding the implementation dimension. In fact, the effective integration of the resulting WM outputs to the current organizational web systems is of utmost importance since it will determine the returns and the profitability of the whole WM project, regardless of the ultimate desired results, *e.g.,* database enrichment, website redesign, search engine optimization and model integration to the website (development), etc. A new algorithm may be highly effective in segmenting new web users but it may become completely useless if it is poorly implemented.

1.3.2. Pattern discovery and analysis methods

Pattern discovery draws upon methods and algorithms developed from several fields such as statistics, DM, machine learning and pattern recognition (Srivastava *et al.,* 2000). Pattern discovery is related to step 3 in the WM process. In that piece of research the stress will be put on the added-value of pattern discovery and analysis methods and techniques for the aCRM process of organizations.

WM methods are similar to DM methods. Actually, in WM, DM methods and techniques are merely applied to the web data, which is also why WM is called Web Data Mining. For the sake of this research, these methods will be referred to WM methods. These are traditionally divided into 4 main categories summarized in Table 1

Association analysis is mainly used to find all co-occurrence relationships called associations, among data items (Liu, 2007). It thus analyzes the pages visited by a web user as well as the products bought by a customer on a single visit online (Tufféry, 2011). The classic application of association rule mining is the market basket analysis (Agrawal *et al.,*

1993). In association analysis, the challenge is to process enormous volumes of data and to pick out previously unknown and relevant associations from the overwhelming majority of irrelevant or previously known associations (Liu, 2007).

Clustering refers to segmentation and this definition is also used in the field of WM. Clustering is therefore, the *"division of a population of customers into a certain number of subgroups, as homogeneous as possible, in order to enable organizations to better adapt their marketing strategies to each of these subgroups"* (Lendrevie & Lindon, 1993). Both clustering and association analysis are descriptive methods that follow an unsupervised learning approach.

Classification is one of the two major predictive and supervised learning methods of DM. Classifying behavior or attitudes is particularly valuable in marketing to know for instance if a prospect might become a customer, if a customer is going to remain loyal to the company or if a customer may not pay the bills. Scoring applications, *e.g.,* attrition scores, epitomize that method. It intends to identify if a customer might renew a given contract subscription, continue buying from the company, etc. The class variable (customer or non-customer) is determined by other various variables, usually qualitative, which may predict the class variable. Then cases are gathered where class variable and the other pre-defined variables are available. A model is then built in order to link the class variable and the explaining variables. This model is eventually tested before being widespread and used for other applications (Bazsalicza & Naim, 2001).

Prediction deals specifically with continuous variables (the probability of a purchase and the amount of money spent on that purchase, etc.), while classification deals specifically with discrete variables that are of a qualitative type (client or non-client, loyal or non-loyal, etc.). This second major supervised learning method is complementary to the classification method. It implies to identify the class variable (amount of purchase), and other variables, usually quantitative ones, which may explain the class variable. As for the classification method, cases are gathered where both the class variable and the predicting variables are available. A model is built to link these two types of variables and the model is verified before being generalized.

Each task entails a number of specific techniques that are applicable depending on the type of project undertaken and the outputs that are sought. Bazsalicza and Naim (2001) underline the importance of taking into account the whole scope and context in which the WM project takes place in order to determine the most adequate technique to be used depending on data types, outputs sought, etc. According to them, "in the framework of a specific mission, analyzing detailed, relevant data that are available to an organization in order to infer the most rational actions to be implemented, in order to reach profitability, will probably be the best approach". DM and hence WM is always a process linked to a specific organizational objective.

Table 1.1

Categorization of Web-Mining methods

METHODS	GOAL	EXAMPLE
DESCRIPTIVE MODELING		
Clustering	Creating homogeneous groups	Segmenting consumers on two variables: "the amount of time spent on the website" and "the number of items downloaded"
Association	Finding sets of data items that occur together frequently (association rules) or that occur together frequently in some sequences (sequential patterns)	Finding regularities in the web data: in WUM, association rule mining can be used to find users' visit and purchase patterns. Sequential patterns can be used to find users' navigation patterns by analyzing clickstreams in server logs
PREDICTIVE MODELING		
Classification	Explain or predict the qualitative characteristic of an individual based on other qualitative	In the banking context, connecting socio-demographic characteristics of a customer to his/her possession of a specific financial product

	characteristics of that individual	
Prediction	Explain or predict the quantitative characteristic of an individual based on other quantitative characteristics of that individual	In the telecommunication context, connecting consumption of customers to their quantitative socio-demographics

Source: adapted from Bazsalicza and Naim (2001). Data-Mining pour le Web. Eyrolles; Liu, B. (2007). Web Data Mining. Springer; Tufféry, S. (2011). Data-mining and statistics for decision-making. Wiley.

1.3.3. Web-Mining techniques

It is not intended to analyze into details the whole range of techniques available to organizations in order to perform WM activities. However, a brief description of the most important techniques is provided in this section.

1.3.3.1. Association analysis techniques

Association analysis aims at finding the most frequent combined values of a set of variables of a data set (Tufféry, 2011). It is perhaps the most important model invented and extensively studied by the database and DM community (Liu, 2007). It comprises association rules mining and sequential patterns mining.

Association rules mining is a fundamental DM task. It was first introduced in 1993 by Agrawal, Imielinski and Swami and is commonly referred to "basket analysis", which strives to analyze customer buying patterns by finding associations between items that customers put into their baskets (Cios *et al.*, 2007). Association rules are discovered by using "transactional data" (Bazsalicza & Naim, 2001). In the context of WUM these transactional data do not only refer to products but also to web pages (Tufféry, 2011). Association rules can therefore be used to find how items purchased by customers on a web site are associated, but they are also

used to find word co-occurrence relationships and web usage patterns (Liu, 2011). A rule is therefore an expression of the form: if *Condition*, then *Result* (Tufféry, 2011). Association rules should have a relevant *"support index"* (frequency of the entire rule with respect to the set of items) (Agrawal *et al.,* 1993). That means, the percentage of transactions that contain both the Condition and the Result such as "if *newsletter subscription* and *use of recommender systems* (Condition), then *online purchase* (Result)" (Liu, 2007; Tufféry, 2011). Association rules should also have a *"confidence index"* (strength of implication in the rule) (Agrawal *et al.,* 1993). This is the probability of having the Condition and the Result divided by the probability of the Condition (Liu, 2007; Tufféry, 2001). The two types of association rules that can be mined are single-dimensional association rules (takes into account one predicate such as "buy") and multidimensional association rules (takes into account multiple predicates such as "major", "takes_course", "level") (Cios *et al.*, 2007). Each transaction is a set of web pages visited on a web site by a web user. The set (*I*) is the set of all web pages that the web sites encompasses and the sub-whole or "basket" is the log file which gathers all the visit data of a website by an individual (Bazsalicza & Naim, 2001). Below is a list that shows a set of seven transactions in the form of web pages viewed by the same or several different individual(s):

Transaction 1: Homepage, Cosmetics, Map
Transaction 2: Homepage, Contact
Transaction 3: Contact, Press
Transaction 4: Homepage, Cosmetics, Contact
Transaction 5: Homepage, Cosmetics, Forum, Contact, Map
Transaction 6: Cosmetics, Forum, Map
Transaction 7: Cosmetics, Map, Forum

If the user-specified support index "minsup" is 30% and the confidence index "minconf" is 80%, the following association rule: Cosmetics, Forum (Condition) → Map (Result) [support = 3/7, confidence = 3/3] is valid as it supports 42.84% of the cases (> 30%). In 3 out of the 7 transactions, Cosmetics, Forum and Map pages appear and in 2 out of those 3 transactions do both Cosmetics and Forum appear before Map. The confidence index is thus 100% (>80%).

The following rule is also valid: Forum → Map, Cosmetics [sup = 3/7, conf = 3/3]. This is a very basic form of association but more association (sophisticated) rules can be discovered.

The best known mining algorithm is the Apriori algorithm which first generates all frequent itemsets (an itemset has transactions above the minimum support index) and then all confident association rules from the frequent itemsets (a rule with a confidence above the minimum confidence support) (Agrawal & Srikant, 1994). However, avoiding the extraction of an overwhelming knowledge is of primary importance as it guarantees extra value knowledge, usefulness and reliability (Bouchahda, Ben Yahia & Slimani, 2006). To generate parsimonious sets of rules, managers use multiple minimum support indexes to find rules involving both frequent and rare itemsets (Liu *et al.*, 1999; Liu, 2007). The MS-Apriori algorithm integrates Multiple Minimum Item Supports (MIS) (Wang *et al.*, 2000; Seno & Karypis, 2005; Xiong *et al.*, 2003; Liu *et al.*, 1999). It is also possible to mine Maximal Frequent Itemsets (MFIs) to avoid that association rule mining generates a too big number of frequent itemsets and rules (Bayardo, 1998; Lin & Kedem, 1998; Agarwal *et al.*, 1999; Bürdick *et al.*, 2001). Another approach, which consists of mining Closed Frequent Itemsets (CFIs) has been deemed more effective since it provides a lossless concise representation of all frequent itemsets (Zaki & Hsiao, 2002; Pasquier *et al.*, 1999; Wang *et al.*, 2003). Both are concise representations of Frequent Itemsets. New algorithms have been developed since then such as the mining of Generalized Association Rules (Srikant & Agrawal, 1995), Multilevel Association Rules to find strong rules at the high level(s) and finding weak rules at the lower-level(s) (Han & Fu, 1995), Class Association Rules (CARs) for finding rules with fixed target items (*e.g.* how users activities such as search queries or pages clicked by web users, are associated with advertisements that they may like to view), where the Consequent (Result) has a single item (Liu *et al.*, 2006; Liu, 2007). Such associations enable managers to develop, for instance, advanced online advertisement techniques such as Behavioral Targeting (BT) in online Display Advertising, to increase the Click Through Rate (CTR) (Yan *et al.*, 2009).

Beyond the issue of generating Frequent Itemsets, other types of association rules have also been developed such as cyclic association rules (Ozden *et al.*, 1998), periodic patterns (Yang

et al., 2004), negative association rules, which aims at finding products that are not bought or pages that are not visited (Wang *et al.*, 2004), association rules with numerical/quantitative variables, which considers items that assume a range of values instead of just two values (Webb, 2001), high-performance rule mining (Buehrer *et al.*, 2006), inter-dimension association rules (one or more of the predicates is repeated), hybrid-dimension association rule (incorporates some of the predicates multiple times) (Cios *et al.*, 2007) or association rule mining from bioinformatics data with a very large number of attributes and a very small number of transactions (Cong *et al.*, 2005). New techniques also help managers to identify those rules that are most relevant among the huge number of generated rules and to classify them (Padmanabhan & Tuzhilin, 2000; Tan *et al.*, 2002; Tuzhilin & Adomavicius, 2002; Wang *et al.*, 2003; Yan *et al.*, 2005).

Sequential Patterns Mining. Association rule mining reveals which products tend to be bought together most often but not in which order. Likewise, it does not reveal the order of the web pages visited together most often. Sequential Association Rules (Sequential Patterns Mining) overcome that drawback by identifying which pages have been viewed or which products have been bought simultaneously and in which order (Bazsalicza & Naim, 2001). In WUM specifically, it is useful to find navigational patterns in a website from sequences of page visits of users and for these applications, classic association rules are not appropriate (Liu, 2007). The fundamental algorithm of Sequential Patterns (SP) called GSP aims at finding all sequences (ordered list of itemsets) that have a user-specified minimum support (fraction of total data sequences that contains the sequence) (Liu, 2007). The period of the analysis is relative since the time frame can be predetermined (by day, week, month, etc.) (Bazsalicza & Naim, 2001). Each sequence is a sequential pattern. The MS-GSP allows the use of multiple minimum supports (Srikant & Agrawal, 1996). The PrefixSpan algorithm for single minimum support and its extended version MS-PS for multiple minimum supports, causes less memory and computational problems (Pei *et al.*, 2001). As for association rules, once SPs are generated, it is then possible to generate many types of rules such as sequential rules, label sequential rules and class sequential rules, analogous to CARs (Liu, 2007).

Collaborative Filtering (CF) algorithms constitute an application of association analyses outputs by providing recommendations to a web user. The recommendation can follow a *random system* by just recommending whatever item comes up-less effective-or it can follow the *principal system* which recommends the item(s) that entail most hits or that are best evaluated by all other users (Adomavicius & Tuzhilin, 2005). *Social* or *CF-based algorithms* identify the proximity of the tastes of the individual with the tastes of other similar individuals by typically taking into account ratings, remarks or comments from those individuals (Shih & Liu, 2008). C*ontent-based algorithms* make recommendations on the basis of the content generated by web users themselves, such as previous purchases, history of queries, etc. (Albadvi & Shahbazi, 2010). Another type of filter for CF is the *socio-demographic profile* which would recommend to a user the choices made by another user with similar socio-demographic characteristics. A combination of different types of filters is also possible (Bazsalicza & Naim, 2001). Recently, the users' Customer Lifetime Value (CLV) and Recency-Frequency-Monetary (RFM) profiles have been integrated to rating-based CF in order to generate more relevant recommendations to customers (Albadvi & Shahbazi, 2010). The primary goal is to recommend to web users how to do things by showing them how others did. Such techniques are particularly used by websites selling cultural products such as Amazon, epitomized by the famous catchphrase "those who bought that item also bought..." (Bazsalicza & Naim, 2001).

Bayesian networks form another technique which is a graphical representation of the knowledge structure of an individual. The Bayes decision theory is based on the assumption that the classification of patterns (the decision problem) is expressed in probabilistic terms (Cios *et al.*, 2007). Knowledge is quantified by probabilities in the form of a chart which gathers all the probabilities regarding all binary combination of variables (Saporta, 2006). A Bayesian network is therefore a graph (structure) and tables of probabilities of each node conditioned by its parameters (Beck & Naim, 1999). This technique offers the advantage of reasoning with incomplete information as is often the case with web data (Bazsalicza & Naim, 2001).

1.3.3.2. Supervised learning techniques for classification and prediction

Supervised learning is inductive by nature and can be used for classification purposes. Inductive learning means going from the data to the model, as opposed to deductive which goes from the model to the data (Tufféry, 2011). Inductive learning hence infers generalized information or knowledge from the data by searching for regularities among the data and it is correct for the given data but merely plausible outside the data (Cios *et al.*, 2007). The two main predictive approaches are classification (discrimination where the dependent variable is nominal or categorical) and prediction (regression where the dependent variable is continuous) which is different from forecasting. In supervised learning the class labels and the class of an individual are provided and thus known a priori (Liu, 2011). The objective is to estimate the value of a dependent variable (class) relating to an individual/object as a function of the value of other variables (attributes) relating to the same individual/object (Tufféry, 2011). The function can thus be used to predict the class values/labels of the future data (Liu, 2011). First, a training data set is used to learn the classification or prediction model and the model accuracy is then evaluated on a test set, as depicted in Figure 1.2 below.

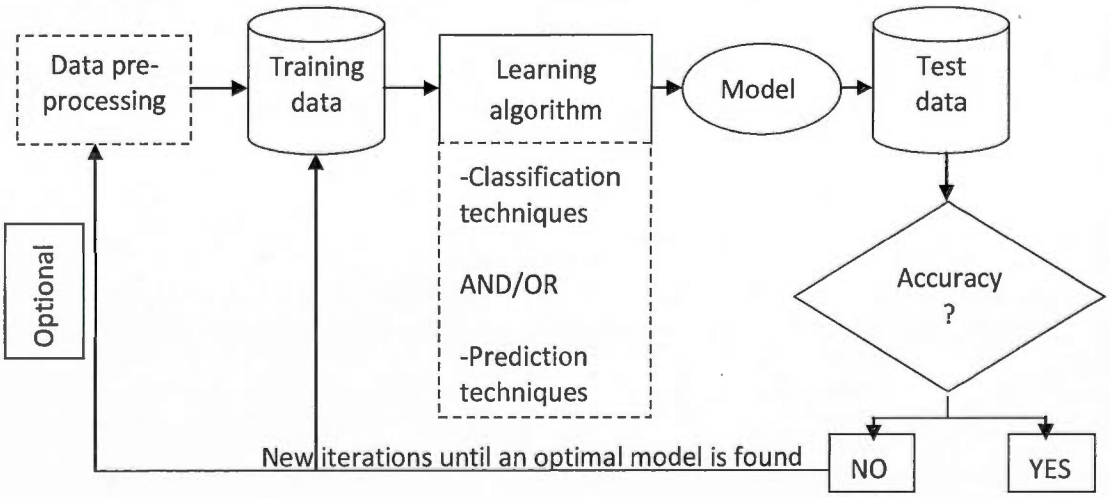

Figure 1.2 The supervised learning process. (Adapted from Lefébure and Venturini, 2001.)

Classification and prediction techniques:

Decision tree induction: Decision trees originate from Hunt's Concept Learning Systems, which emulates human learning by finding key distinguishing features between two categories (Cios *et al.,* 2007). Its underlying "Divide-and-Conquer Strategy", stipulates that at each step, all attributes are evaluated and one is selected as the most discriminating to divide the data into *m* disjoint subsets (*m* being the number of values of the attribute) (Hunt, 1962). Decision trees are useful to detect criteria for dividing the individual of a population into a predetermined number of classes, usually 2 (Tufféry, 2011). In order to classify items into several categories it is necessary to search, in an iterative fashion, the most informative criteria with regards to that classification (Bazsalicza & Naim, 2001).

At the beginning, all the observations are at the root and as the tree increases, the observations are sub-divided recursively (Quinlan, 1992). A decision tree is then built in three steps:

(1) The researcher chooses among an array of heuristic methods (Chi-square, Gini index, (ordered) Twoing, entropy, etc.) to determine the precise criterion or separation criterion which chooses variables and the separation condition on variables to divide individuals into *n* classes; as well as the stop criterion which determines when the growth of the tree can be halted (Tufféry, 2011),

(2) Individuals are distributed among the leaves and are assigned to a given class, creating very deep trees with many tree leaves covering few observations which is very accurate to predict the training set but not the unseen test set (overfitting) (Mitchell, 1997),

(3) Rules are pruned in order to simplify them and decrease overfitting. Pruning can take several forms: testing each sub-tree of the maximum tree on the test sample and keep the sub-tree giving the lowest error rate (CART) (Mitchell, 1997), deleting branches or sub-trees and replace them with leaves of majority classes

(Quinlan, 1992) or cross-validating by combining error rates found for all possible sub-trees to choose the best possible tree (Tufféry, 2011).

The resulting model is then evaluated for accuracy on the test set. It is represented as a tree with a root at the top and many leaves at the bottom. Terminal nodes (leaves) consist of individuals of a single class and an observation is assigned to a leaf (class) when it conforms to all the rules for reaching this leaf (Tufféry, 2011).

The first decision tree algorithm, AID, was developed by Morgan and Sonquist in 1963 and was based on ANOVA. New developments include CART (Classification And Regression Tree) which handles all variables, preferably with many categories (Breiman *et al.*, 1984). The CHAID (Chi-Square Automation Interaction Detection) handles only discrete or qualitative variables and uses the chi-square to define the most significant variables of each node which produces wider rather than deeper trees (Hartigan, 1975). The C5.0 is an improvement of the earlier ID3 and C4.5 trees that can be applied to all kinds of variables (Quinlan, 1998).

A decision tree can be converted to an IF…THEN set of rules since each path from the root to the leaf forms a rule: decision nodes along the path form the conditions of the rule and the leaf node (class) forms the consequent (Liu, 2007). The tree paths rules are thus mutually exclusive and exhaustive and the decision tree generalizes the data, while the challenge is to find the best tree that is small, general and accurate (Hyafil & Rivest, 1976). One advantage of decision trees to that respect is that unlike the association rules technique, they only find a subset of rules that exist in data, whereas association rules technique finds all the rules subject to minimum support and confidence constraints (Liu, 2011).

The major drawback of decision trees lie in the slight variations that may arise in the data and which generate other nodes choices and may lead to irreversible schemes since each node determines the others (Bazsalicza & Naim, 2001). Also, decision trees generate very complex (long) rules, which are hard to prune, they generate a large number of corresponding rules that can become excessively large unless pruning techniques are used to make the rules more comprehensible require large amounts, and they require large amounts of memory to store the entire tree for deriving the rules (Cios *et al.*, 2007).

Rule induction: it is possible to learn classification rules directly without decision tree by using a sequential covering algorithm that builds an ordered list of rules (not association rules) or decision list (Rivest, 1987). Unlike the divide-and-conquer strategy underpinning decision trees, rule induction is based on a separate-and-conquer strategy which evaluates all attribute-value pairs (conditions), much larger in number than the number of attributes, and selects only one, which is slower than the divide-and-conquer strategy (Liu, 2007). Either a rule is found in each iteration (Clark & Niblett, 1989) or all rules for each class are learnt together (Quinlan, 1990). The basic idea of rule induction algorithms is to learn a list of rules (decision list) sequentially, one at a time, to cover the training data and after each rule is learned, the training observations covered by the rule are removed, while only the remaining data are used to find subsequent rules (Furnkranz & Widmer, 1994; Brunk & Pazzani, 1991). A rule covers an observation if the observation satisfies the condition of the rule (Cohen, 1995).

Such rules are easy to comprehend, their output can be easily written in first-order logic format, or directly used as a knowledge base in expert systems, the background knowledge can be easily added into a set of rules, and they are modular and independent, *i.e.,* a single rule can be understood without reference to other rules (Cios *et al.*, 2007). This last advantage is also a disadvantage for rules since they do not show relationships between each other. *Hybrid algorithms* involving both decision trees and rule algorithms include CLIP4, CN2, C5.0, genetic algorithms (Cios *et al.*, 2007).

Classification Based on Association (CBA system) rules: Normal Association Rules (NARs) applied on transaction data only (search queries, web pages clicked by each visitor, etc.) have no fixed class and any item can appear on the right or left hand-side of the rule (for recommendations, only one item appears on the right-hand side of a rule). NARs predict what a customer is likely to buy and when a customer purchases some products, the system will recommend him/her other related products based on what he/she has already purchased (Liu, 2007). At the prediction (recommendation) time, given a past transaction, all rules that cover the transaction are selected and the strongest rule is chosen and the item on the right-hand side of the rule (Consequent) is predicted and recommended to the user (Liu, 2011). Multiple minimum supports increase recommendation power (Mobasher *et al.*, 2001). The advantage

is that association rules can predict any item since any item can be the class item (Consequent).

A Class Association Rules (CARs), on the other hand, has only one fixed class on the right-hand side of the rule (Consequent) and can be used for classification directly (Liu *et al.,* 2006) even with continuous variables (Al-Fayyad & Irani, 1993). They may be redundant and need to be pruned (Liu *et al.,* 1998), while multiple minimum class supports can also be assigned to different classes (Mobasher *et al.,* 2001). CAR rules can be used to build a classifier by using the strongest rule (Li *et al.*, 1998), selecting a subset of the rules (Cong *et al.,* 2005) or combining multiple rules together (Zaki & Aggarwal, 2003; Wang *et al.,* 2000). Further, CARs can be used as features to augment the original data or form a new data set which is then fed into a traditional classification algorithm (decision tree, etc.), because CARs capture multi-attribute correlations with class labels not found with other classification algorithm (Antonie & Zaiane, 2002; Deshpande & Karypis, 2002).

Bayesian methods.

-The Naïve Bayesian Classifier: takes a probabilistic point of view since it estimates class posterior probabilities given a test example so that the class with the highest probability is assigned to the example (Domingos & Pazzani, 1997). If a training data set has two attributes A (with categories h, g and m), B (with categories b, s and q) and the class C (with categories t or f), the goal is to compute all the probability values of the type: $Pr(A = m| C = t) = 2/5$, $Pr(B = b | C = t) = 1/5$, etc., to learn a naïve Bayesian classifier in order to predict the class of the test case (Kohavi *et al.,* 1997; Langley *et al.,* 1992). The Naïve Bayesian Classifier (NBC) makes the naïve assumption that variables are independent conditionally on the dependent variable. That is, the presence (absence) of a class feature is unrelated to the presence (absence) of any other feature, given the class variable. For example, a web user may be considered "goal-oriented" if, (s)he views product information pages, uses electronic decision aid agents or spends more time on each page than an average user (Guéguen *et al.,* 1996; Naim *et al.,* 2007). Even though these features depend on each other or upon the existence of each other, a NBC considers all of these properties to contribute independently to the probability that this web user can be classified as "goal-oriented" and not as "hedonist" or "shallow", as defined by Moe (2003) taxonomy of web users. The NBC is actually very

efficient despite the naivety of its assumptions (Zhang, 2004). It is, however, outperformed by ensemble methods (boosted trees, random forests, bagged trees), Support Vector Machines and neural networks (Caruana & Niculescu-Mizil, 2006).

Due to the rapid growth of online documents in organizations and on the web, automated document classification is an important issue tackled by the NBC (Liu, 2007). A specific naïve Bayesian learning method uses text characteristics to classify future documents (McCallum & Nigam, 1998). It assumes that each document is generated by a parametric distribution governed by a set of hidden parameters. The training data is used to estimate these parameters which are then applied to classify test documents by calculating the posterior probability that the distribution associated with a class would have generated the given document (Lewis & Gale, 1994; Nigam *et al.,* 2000). Classification becomes a matter of selecting the most probable case (McCallum & Nigam, 1998).

-Bayesian networks: suppose that it is the class which is considered as being the cause of the observed variables and not the opposite as in cross-selling, for instance (Saporta, 2006). Bayesian networks constitute a probabilistic graphical model (directed acyclic graph) relating the dependent variable to the (discrete) independent variable with arrows indicating their conditional dependencies (Naim *et al.,* 2007; Guéguen *et al.,* 1996). The network is constructed by creating links between the dependent variable and some independent variables. The network then enables to calculate the conditional probabilities of the variables with respect to each other (Thomas *et al.,* 2002; Baesens *et al.,* 2004). If no category is specified for any independent variable, the network supplies the a priori probability of each category of the dependent variable. For example, the probability of the "reimbursing" and the "defaulting" categories of a credit risk score. In an online credit scoring application, it can thus be decided that credit should be approved (or refused) if the conditional probability of the category "good payer" of the dependent variable is above (or below) a given score threshold (Chang *et al.*, 2000). Such a dynamic model is useful to build adaptive questionnaires for online credit or insurance applications, because customers may easily abandon the process if it is too long (Tufféry, 2011).

Artificial Neural Networks (ANNs): ANNs are based on reproducing the capacities of Human logic reasoning in a computer (Davalo & Naim, 1991). They can be used for both

classification and prediction. Their non-linearity allows building input-output associations by means of learning through multiple layers of analysis, the neurons layers (Cios *et al.*, 2007). The most common application is to associate a prediction (output) to a set of observed variables (inputs) and the association is not direct but goes through several intermediary neurons which represents non-linear relationships (Bazsalicza & Naim, 2001). The MultiLayer Perceptron (MLP) learns for instance by associating each individual of the learning set with unit output values of 1 per unit corresponding to the (known) class of the individual and 0 for the other units (Rumelhart *et al.*, 1986). The Radial Basis Function (RBF), on the other hand, is better-suited if two classes are to be predicted. There is only one unit in the output layer. The value 0 for this unit corresponds to one class, and the value of 1 to the other class (Lukaszyk, 2004). Consequently, the network is entirely described by the loadings of the connections between neurons and adjusting the network means thus to find the configuration of loadings (parameters) which best approaches the requested input-output association (Davalo & Naim, 1991). This is mostly used in the banking context where financial institutions seek to predict clients who may default refunding loans according to clients' characteristics (SPSS 17.0). Neural networks can be further etymologized into two broad categories: "feedforward" with no loops and connections within the same layers (Kohonen SOM network, Radial Basis Function network, probabilistic networks...) and "recurrent" with possible feedback loops (Hopefield network…) (Lovelace & Cios, 2007; Swiercz *et al.*, 2006; Cios *et al.*, 2007).

Support Vector Machines (SVMs) : in 1992 and 1995, Vapnik *et al.* introduced SVMs, which are especially suited for high dimensional data. In Vapnik *et al.*'s (1992) "separable case", the linear learning system builds a two-class classifier. It assumes that linearly separated observations can be cut off from each other by a linear boundary (Tufféry, 2011). The SVM finds a so-called hyperplane or decision boundary which separates the positive and negative training observations, and observations are assigned to a given surface (one side of the hyperplane) depending on their linear score (Vapnik *et al.*, 1992). There are an infinite number of possible boundaries but SVMs allow to select the optimal hyperplane that maximizes the width of the margin between the two groups to be discriminated (goodness-of-fit of the model) (Vapnik & Cortes, 1995) and it is as distant as possible from all the

observations. In practice, training data has some degree of randomness and noise (errors, etc.) which make the two populations to be discriminated not perfectly separated but overlapping. To allow noise in the data, the constraints of the linear separable case SVM must be relaxed (Liu, 2007). This can be done by either introducing an error (slack variable) defined for each observation on the wrong side of the boundary by measuring the distance separating it from the margin (Vapnik & Cortes, 1995), or, by moving to a space having a dimension high enough for there to be a linear separation (the scaler product called a *kernel* function can be polynomial, Gaussian, sigmoid or linear) (Scholkopf & Smola, 2002). SVM is thus a linear learning system that finds the maximum margin decision boundary to separate positive and negative observations (Cristianini & Showe-Taylor, 2000). It is however a binary classification with only two classes and additional techniques are used for multidimensional classes (Dietterich & Bakiri, 1995). Continuous independent variables require a variant of SVM called Support Vector Regression (SVR) (Drucker *et al.*, 1996). Hyperplanes are hard to understand and kernel functions make it even more difficult, this is why SVMs are used in applications where human understanding is not required (Tufféry, 2011).

Genetic algorithms: Genetic algorithms are biomimetic since they aim at reproducing the mechanisms of natural selection by selecting the rules (genes) best adapted to classification (or prediction) and by crossing and mutating them until a sufficiently predictive model is obtained (Holland, 1975). According to Lumer and Faieta (1994), genetic algorithms are issued in three steps. Firstly, a random generation of initial rules takes place. Secondly, the best rules that maximize a given fitness function are selected and finally new rules are generated by mutating and crossing the existing rules. Both genetic algorithms and SVMs are not widely available in software and are very lengthy (Azzag *et al.*, 2003). As such, these techniques remain rarely used in WM.

Discriminant analysis (D.A.): D.A. aims at finding a representation of the individuals which provides the best separation between groups (descriptive discriminant analysis) and finding the rules for assigning individuals to their groups (predictive discriminant analysis) (Hai *et al.*, 2006). For instance, a set of individuals or pages are characterized by a qualitative dependent variable Y and quantitative independent variables X. The representation of the

relations between Y and X correspond to the descriptive method while finding the rules for predicting the category of Y starting from the values of X corresponds to the predictive method (Tufféry, 2011).

Logistic Regression (L.R.): unlike D.A., L.R. can handle dependent variables with two values, 3 or more ordered values and nominal values while independent variables can be qualitative or quantitative. Its S-shaped distribution function is particularly relevant in marketing regarding sales of a new product (Tufféry, 2011). Variant of L.R. include logit, probit (normit), ordinal L.R. and log-log (Gombit), which differ by their link function and transfer function (Nakache & Confais, 2003). It is also possible to perform L.R. on individuals with different weights, with correlated data (*e.g.* longitudinal data, time series), ordinal L.R., multinomial L.R., (Tennenhaus, 2000), the General Linear Model (GLM) (Nelder & Wedderburn, 1972) or to model rare events with poisson regression, the generalized additive model even more general than the GLM (Hastie & Tibshirani, 1990). The regression can be performed by using adjustment by regression technique which aims at selecting a function that best approximates the observations X (explaining variables), from a wide range of given functions family in order to explain a given variable called Y (Bazsalicza & Naim, 2001). If a linear function is selected then the regression will be linear although this is not systematic (Saporta, 2006). This most prominent approach is the minimization method of least squares of errors which uses as a criterion a sum of squares of modeling errors (Cios *et al.*, 2007). Regression models can further be divided into a linear and a nonlinear category but this will not be developed here.

K nearest neighbor.

This transductive learning type, also called "lazy learning" as opposed to the other "eager learning" techniques exposed previously, does not produce a model from the data (Liu, 2007). Rather, when a test case is introduced, the algorithm compares it to all training observations to compute similarity or distance between them. The k most similar or closest observations (k nearest neighbors) are then selected and an observation is then assigned to the class that is the most frequent class (majority class) (Liu, 2011). For instance, a web page (web user) is assigned to the class "sport", because it is mainly surrounded by "sport pages"

(web users who are keen on sport). This technique is slow it also requires a lot of storage and computer capacity (Tufféry, 2011) but performs equally well as SVMs in text classification (Yang & Liu, 1999).

Bootstrapping and ensemble methods

Instead of using one single classification or prediction technique to develop one unique model, another approach consists of using machine learning meta-algorithms to build many classifiers and combining them to produce a better classifier or choosing the best classifier according to the context (Liu, 2011). One approach to ensemble methods, called bootstrapping, aims at combining models which disagree in their predictions (Efron & Tibshirani, 1997). The meta-algorithm then disrupts their learning, either by changing the learning sample or by keeping the same sample and varying the learning parameters (Hansen & Salamon, 1990). For example, combining neural networks built on the same sample by varying only the parameters and typology of each base network (Opitz & Maclin, 1999). However, combining randomization of the learning sample with randomization of the parameters, known as the "random forests" technique produces even better performances than mere bootstrapping (Breiman, 2001). Bagging (Bootstrap AGGregatING) remedies the lack of robustness of unstable classifier *e.g.* decision trees, neural networks, etc. and is thus less useful for stable classifiers. It increases the generalization of the model especially when there are few individuals to be modeled and it also decreases the nuisance caused by outliers (Breiman, 1996). Bagging constructs a family of models on m bootstrap samples and aggregates the predictions of each model by voting (in case of a classification situation) or by averaging (in case of a regression situation) (Breiman, 1996). Boosting seeks to know whether a set of weak learners create a strong learner (Kearns, 1998). The classification algorithm, AdaBoost, is applied successively to versions of the initial training sample which are modified at each step to allow for classification errors of the preceding step. The classifiers (which may be weak) constructed in this way are then combined to produce a stronger classifier (Freund & Schapire, 1996). It concentrates its action on individuals that are hard to model and behave in ways harder to predict (Tufféry, 2011). It is also very resistant to overfitting (Friedman *et al.*, 2000).

1.3.3.3. Unsupervised learning techniques for clustering

Partition-based clustering, also called *objective function-based clustering* relies on a certain objective function whose minimization is supposed to lead to the "discovery" of structure existing in the data (Cios *et al.*, 2007). In partition-based clustering the number of clusters is pre-defined and the analysis proceeds with the optimization of the objective function, although one is never sure what type of structure to expect and hence what should be the most suitable form of the objective function (Cios *et al.*, 2007).One approach is the *k*-means algorithm which seeks to group the individuals in *k* partitions, *k* being predetermined by the researcher (Bazsalicza & Naim, 2001). The criteria optimized here is intra-class and inter-class inertia (Saporta, 2006). Each cluster is constituted of points which are nearer of a given center than all the other centers (forming other clusters). Other major clustering approaches include the Kernel-based clustering, the K-mediods algorithm, the Fuzzy C-means algorithm, the Density-based algorithm and the Clustering Using REpresentatives (CURE). Partition-based clustering does not provide an optimal partitioning of the whole dataset in *k* classes, but it is powerful for analyzing large datasets of more than 1000 individuals (Nasraoui *et al.*, 2008).

Hierarchical clustering lies in the successive development of clusters through successive splits (starting with one single cluster that is an entire data set) or with individual points treated as initial clusters, which are kept on merging (this process leads to the concept of agglomerative clustering) (Cios *et al.*, 2007). The hierarchical ascending technique seeks to sequentially gather "near" items (Benzécri, 1983). The distance between two items of a same group is smaller than that of two items of two different clusters (Bazcalicza & Naim, 2001). Without going into the details, the distance between two such clusters is computed with the single link method, the complete link method or the average link method (Cios *et al.*, 2007). Unlike partition-based clustering, the hierarchical technique does not perform well in large datasets but provides very easy determination of the number of clusters thanks to the dendogram and the histogram of level indices (Nasraoui *et al.*, 2008). The resulting modeling is typically represented by a dendogram which is a binary tree with a distinguished root that

has all the data items at its leaves in a bottom-up (agglomerative) mode or a top-down mode (divisive) (Cios *et al.*, 2007).

Hybrid techniques combining both the partition-based and hierarchical techniques can be performed to overcome that difficulty (see also Saporta, 2006).

The *model-based clustering* represents another approach in which a certain probabilistic model of the data is assumed and it is sought to estimate the parameters of that model (Cios *et al.*, 2007). The mixture density model for instance assumes that the data are a result of a mixture of *c* sources of data and each of these sources is treated as a potential cluster in probabilistic terms (Saporta, 2006). Eventually, the maximum likelihood estimation is used in order to discover the clusters (Cios *et al.*, 2007).

The *Principal Component Analysis* (PCA), based on the factor analysis approach, generates the best "summary" of a whole dataset (Bazsalicza & Naim, 2001). While not being specifically designed for clustering, PCA can be used as a complementary technique before or after clustering techniques especially if the dataset is very large. PCA seeks to evaluate whether the scores on a number of X variables, *e.g.*, the value attributed to an individual on a given characteristic, may be explained by a smaller number of latent variables, called components (Warner, 2008). All these techniques refer to unsupervised modeling, which means that it is sought to organize the data without orientating the results according to a given economic criteria (Bazsalicza & Naim, 2001).

1.3.4. The outputs of data-mining methods and techniques

Combined to Online Analytical Processing (OLAP), visualization tools, knowledge query mechanisms and intelligent agents, these pattern discovery and analysis techniques generate path analysis, association rules, sequential patterns, clusters and classification rules which will be directly integrated to websites for cross-selling (as on Amazon, etc.), creation of dynamic content per visitor or per segments of visitors; for web site redesign and optimization; for enriching current databases on customers with new models and information.

Possible actions that could be undertaken for integrating the results of a WM project are summarized in Figure .

Table 1.2

Integration of the modeling results

Method used	**Type of result**	**Type of action**	**Example of action**
Counting	Report	Redesign	Reorganizing the website
Clustering	Clusters (segments)	Redesign	Reorganizing the website according to the discovered clusters
	Model	Enriching the database	Application of the model: integration of the code of the cluster in the customer database
		Development	Integration of the model directly on the website server
Association analysis	Association rules	Redesign	Reorganizing the website, *e.g.*, according to the results of the basket analysis (link to the diapers page on the baby clothes page)
		Development	Recommendations module
	Sequential association and Collaborative Filtering	Development	Recommendations module
	Bayesian models	Development	Recommendation modules
Classification	Scores	Enriching the	Application of the model:

		database	integration of the score in the customer database
	Model	Development	Recommendations module
Prediction	Model	Enriching the database	Integration of the score in the customer database for future predictions

Source: adapted from Bazsalicza and Naim (2001).

1.3.5. The applications of WM

It has been seen that WM is a set of methods and techniques derived from DM, which aims at performing descriptive modeling by means of clustering, classification, association and prediction methods and techniques. WM can be done by using different types of gross data: log files, requiring the WUM method; search results and web pages such as text etc., requiring WCM; and links, requiring WSM. Text-Mining, (Micro)Blog-Mining, Natural Language Processing (NLP) as well as Semantic Research can also be used to add further value to marketing activities but these techniques are beyond the scope of this research.

It has been seen that most aCRM goals and objectives may be achieved by using WM methods and techniques. However, despite its many advantages and the various trends that point toward an increased use of the internet over the years for almost all kinds of transactions, WM remains still relatively under-exploited to fully reach aCRM goals and objectives. Nonetheless, WM has various applications which should make it a relevant part for any business analytics project, if not an integrated component of a meta-aCRM structure. Such a highly-integrated meta-structure would form one module of an organizational Information System (IS) tightly interlinked for leveraging optimal Business Intelligence (BI).

Given the tremendous amount of information provided by web log files WUM is particularly well-suited for user modeling such as web content personalization, web site reorganization, prefetching and caching, e-commerce and business intelligence (Facca & Lanzi, 2003). WUM focuses on extracting usage patterns from web logs in order to derive useful marketing

intelligence (Büchner *et al.,* 1999; Cooley *et al.,* 1997; Cooley *et al.,* 1999; Spiliopoulou *et al.,* 1999; Zaiane *et al.,* 1998), to discover aggregate profiles for the customization and optimization of web sites and search engines (Nasraoui *et al.,* 1999; Srivastava *et al.,* 2000; Mobasher *et al.,* 2000b; Mircan & Sitar-Taut, 2009; Pierrakos & Paliouras, 2010), and to enhance the effectiveness of Collaborative Filtering (CF) approaches (Shardanand & Maes, 1995; Herlocker *et al.,* 1999; Albadvi & Shahbazi, 2010). However, WUM should ideally be combined to WCM and WSM for optimal returns as such an approach yields more complete results, a 360°-view of the WM process. A more detailed classification of the main applications of WM is given in Table 3. DM and hence WM are considered to play a crucial role in providing a true Knowledge Management (KM) environment in the organization (Ranjan & Bhatnagar, 2011). WM uses as input raw operational data collected from the web in order to derive useful knowledge which is used as a basis for subsequent decision-making purposes. Table 3 does not cover the wide array of applications for which WM may be used. Many more may be developed and a scientific taxonomy is not the purpose of this piece of research. However, it is a tentative attempt to compile and gather the great variety of literature about WM applications in a non-exhaustive way to epitomize those most important benefits that a business might obtain by using WM in a knowledge acquisition perspective.

Table 1.3

Major applications of Web-Mining

Web Usage Mining	Web Content Mining	Web Structure Mining
-Web personalization • Mass customization • Adaptive websites • Recommender Systems • Collaborative Filtering -Search engine personalization • Adjusting web filters to individual	-Seach Engine Optimization (SEO) • Improving browers • Learning which topics are related to each other by occurrence of links between topics • Finding related words by them often occurring in the same page • Finding keywords	-Web site redesign • Discovery of structural relationships • Discovery of access relationships • Identifying web sites of high general interest (authorities) -Search Engine Optimization (SEO)

preferences • (Re)organizing web site for fast and easy customer access • Improvig links and navigation patterns -Online customer profiling -Direct online marketing • Intelligent advertisements • Increasing the value of each visitor • Turning viewers into customers through better site architecture • Web site efficiency monitoring	that are most typed by web users when looking for a particular topic • Improving web filters -Direct marketing strategies • Discovering online misuse of brands • Product reputation mining • Opinion mining • Competitive analysis • Evaluating marketing campaigns • Detailed (potential) customer profiling • Preventing churn	• Improving retrieval performances • Optimizing classification -Discovery of Navigation and browsing patterns • Understanding web users' behaviors • Clustering web users • Identifying frequent access paths and visited pages

Source: adapted from Srivastava *et al.* (2000); Van Wel & Royakkers (2001); Zhang & Segall (2008).

This chapter put the stress on WM. The different branches of WM have been exposed, as well as the major WM methods and techniques and their most frequent types of applications in real-world settings. The last chart highlighted the various fields of applications in web marketing and more broadly, e-commerce.

If well-implemented, WM leverages tremendous potentialities to reach the micro-segmentation level and develop a true 1-to-1 relationship with a website's customers.

However, using WM as a standalone tool, disconnected from overall business processes is risky business. In fact, today's company processes lean more towards convergence and integration rather than separation and isolation. WM remains as useful a tool as the robustness of the framework in which it is meant to work. WM is a mean and not an end. This is what is so often misunderstood about WM tools, and leads to corporate disasters. WM

needs to be integrated, used and managed in congruence with other business processes and be aligned on business and corporate designs, capacities and strategies.

The relational marketing paradigm calls for customer-centric marketing strategies that are market-oriented and capitalizes on knowledge and relationships to build stronger and more profitable ties with customers and other business partners. The Customer Relationship Management is a powerful business strategy that has been fine-tuned for the last 30 years to do just that: improving identification, conversion, acquisition and retention of customers (preferably the most profitable ones) and to better differentiate and customize products or services, as well as better interact with customers (Tiwana, 2001). CRM is everything but a series of software or a technical solution. Rather, it is a true business strategy combining business processes and technology to understand customers of a company from several perspectives in order to differentiate products and services of that company, competitively (see also Tiwana, 2001).

DM is already a cornerstone of CRM and more specifically the analytical side of CRM, called aCRM. DM finds relations in the data that are hidden or non-obvious a priori (Schiff, 2009). As such DM derives useful predictive analytics to enable granular marketing strategies. While DM is mainly applied to data obtained from so-called offline sources (ERPS, SCM, POS terminals, marketing research, syndicated services, and the like), it may be reasonably though that data emanating from online sources, generated on the internet, and analyzed with WM will drastically improve brick-and-click companies' understanding of their customers and prospects. For pure-players WM will lie at the core of their aCRM. In that respect, the next chapter focuses on what it is meant by aCRM (analytical Customer Relationship Management), and emphasizes more on how WM is related to aCRM.

Developing CRM means focusing on customers. However, Operations Management call for stronger and more integrated ties with suppliers and more generally all key partners, or stakeholders of a company for effective and efficient supply chain management, production management, lean and agile networks of companies. WM could be used in a Knowledge-enabled Customer Relationship Management (KCRM) perspective as well. While Customer Knowledge Management (CKM) developed by WM typically aims at generating value-creating lock-ins, channel knowledge should be developed as well by WM to strengthen

relationships and collaborative efficiency and effectiveness, with business partners (Tiwana, 2001). This would help going beyond the mere Knowledge-enabled Customer Relationship Management (KCRM) with customers, to reaching a more valuable Knowledge-enabled *Collaborative* Relationship Management (KCRM): with customers and partners.

CHAPTER 2

CUSTOMER RELATIONSHIP MANAGEMENT AND KNOWLEDGE MANAGEMENT

2.1. Defining CRM and aCRM

2.1.1. What is CRM?

The customer is a strategic element in a company's downstream supply chain (Xu & Walton, 2005). The changes of attitudes, behavior, preferences of customers need to be identified by businesses in order to remain profitable. Knowledge is the only meaningful resource (Drucker, 1996). Gaining this knowledge is becoming an important differentiator for competitive advantage (Paiva *et al.,* 2002). This is why managing customer relationships, is of utmost importance for market-oriented companies. CRM is a process designed to collect data related to customers, to grasp features of customers and to apply those qualities in specific marketing activities (Swift, 2001). Theoretically, CRM leverages and exploits interactions with customers to maximize customer satisfaction, ensure business returns and enhance customer profitability (Xu & Walton, 2005). More specifically, it is a combination of business processes and technologies in order to understand a company's customer from different perspectives in order to differentiate products and services of that company competitively (Tiwana, 2001). However, in practice, it appears that CRM has become more of a buzzword (Luck & Lancaster, 2003). There appears to be a fundamental problem in CRM research at present because no common image regarding what CRM is actually exists (Choy *et al.,* 2003). It appears to mean different things to different people (Paulissen *et al.,* 2007). In fact, CRM is difficult to conceptualize because it is a cross-disciplinary field of

research which includes marketing, business management, IT and Information Systems (Ngai, 2005). The CRM concept is also relatively new since it appeared at the end of the 1990s (Romano & Fjermstad, 2002a,b) and has kept on increasing in importance in academia since then, until 2004 (Wahlberg *et al.,* 2009).

Wahlberg *et al.* (2009) argue that there are four different perspectives applied to the phenomenon of CRM in the literature: (1) CRM being a matter of integrating business processes in an organization; (2) CRM being a matter of customer-focused business strategy; (3) CRM being a matter of customer knowledge management; and (4) CRM being a matter of technology-enabled customer information management activities, including Strategic CRM (sCRM), Analytical CRM (aCRM), Operational CRM (oCRM), Collaborative CRM (cCRM), Technical CRM (tCRM) and according to some authors Electronic CRM (e-CRM) (Xu & Walton, 2005; Romano & Fjermestad, 2002a,b; Bazsalicza & Naim, 2001). The latter perspective (number 4) has been mostly used in academia and hence it will be used throughout this piece of research, which focuses namely on the Analytical CRM applet of CRM (aCRM).

2.1.2. The CRM typology

In accordance with the perspective depicting CRM as a matter of technology-enabled customer information management activity, there are six different CRM branches which encompass salient topics each:

(1) **Strategic CRM (sCRM):** focuses primarily on the integration, implementation - of a customer-centric view and customer focus - and the coordination of all the activities related to architecture, design, leadership and change management required for the planning, implementation, integration, monitoring and control of a CRM system (Wahlberg *et al.,* 2009).

(2) **Operational CRM (oCRM):** collecting customer data through a whole range of touch points such as contact centers, contact management, mail, fax, sales force, web, etc. to

store and organize them in a customer-centric database which is made available to all users who interact with the customer (Xu & Walton, 2005; Tiwana, 2001). CRM systems are mainly used for such types of operational activities. Gaining customer knowledge to provide strategically important customer information to other departments is not perceived as important as improving operational efficiency (Campbell, 2003; Rowley, 2004).

(3) **Collaborative CRM (cCRM):** integrating CRM systems with enterprise-wide systems (ERP, IS, etc.) to allow greater responsiveness to customers throughout the supply chain (Kracklauer & Mills, 2004) by driving sales through every channel from call center to the web (e-commerce), mobiles (m-commerce) channels or e-learning (Wahlberg *et al.*, 2009).

(4) **Electronic CRM (eCRM):** a web-centric approach to synchronizing customer relationships across communication channels, business functions, and audiences (Forrester Research, 2001). Also includes the increasing Mobile CRM (mCRM) function, which is highly relevant for mobile marketing.

(5) **Technical CRM (tCRM):** focuses on all the aspects relating to hardware platforms, software components or tools; databases and protocols; physical storage and structure, as well as communications network (Chan, 2005).

(6) **Analytical CRM (aCRM):** according to Wahlberg *et al.* (2009) it was the largest branch of CRM until very recently when interest shifted to sCRM given the high failure rates of CRM systems implementations worldwide for reasons related primarily to strategic alignment (Berg, 2001; Cleary 2003). The aCRM branch develops a customer knowledge database from the systematic collecting and storing of customer data which can be perceived as an asset to the enterprises (Wahlberg *et al.,* 2009). It provides a 360° view of the customer (Kotorov, 2002). In fact, aCRM analyses customer data through a range of analytical tools such as standard statistical techniques and data-mining to discover those non-obvious or non-linear patterns in the data; and machine learning (ML) and

Artificial Intelligence (AI) learning techniques which may be combined to DM[28] in order to generate customer profiles, identify behavior patterns, determine satisfaction level and support customer segmentation to develop appropriate marketing and promotion strategies (Xu & Walton, 2005). The field is therefore dominated by the DM concept which collects primarily offline "islands of customers' data" throughout the organization, with data warehousing techniques (Wahlberg *et al.,* 2009). Online web data (Hyperlinks, web log files, texts, images, and so forth) complement those "islands of offline data", by completing them with additional information on customers and prospects, obtained from the internet. The offline-based aCRM process encompasses powerful predictive techniques such as DM, forecasting and scoring (Kimball & Ross, 2002). They typically segment customers more effectively or optimize offerings to better fit customers' buying profiles (Xu & Walton, 2005). Adding web data to the aCRM is thus expected to optimize that process since it will leverage more powerful marketing capacities.

2.1.3. Integrating WM to aCRM

According to Ranjan and Bhatnagar (2011), the aCRM objectives that are relevant to marketing are:

- Helping in cross-selling and up-selling
- Allowing targeted marketing
- Helping in segmenting customers on the basis of fixed criteria

Xu and Walton (2005) developed a more exhaustive framework including the tasks of aCRM according to the "who" or profiling tasks and the "how" or the patterns identification tasks, of simultaneously existing customers (internal) and prospects customers (external). The framework for customer knowledge acquisition is presented in Figure 2..

[28] aCRM explained plainly on 12manage website: < http://www.12manage.com/methods_analytical_crm.html> (retrieved on 13-05-2011).

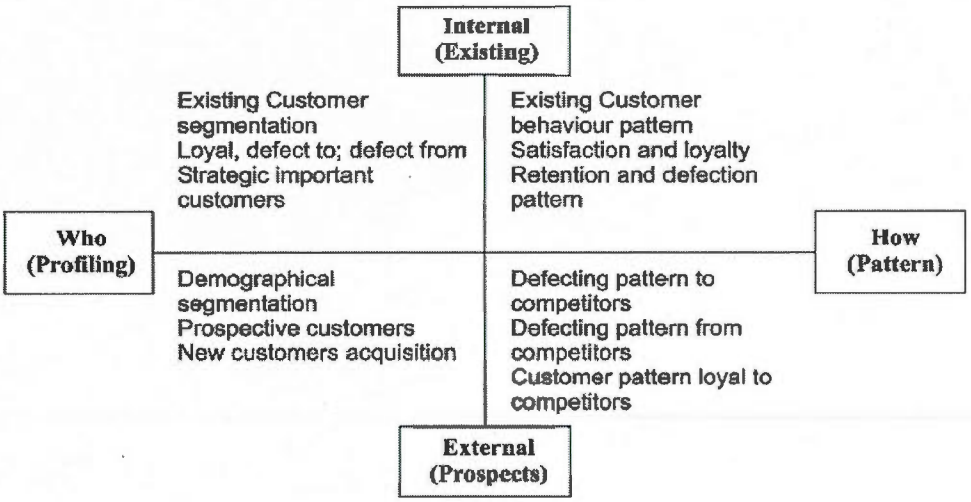

Figure 2.1 aCRM for a customer knowledge acquisition framework. (Source: Xu & Walton, 2005.)

By adapting Xu and Walton's (2005) model, it is possible to develop a framework of aCRM for web users' knowledge acquisition. Figure 5 below illustrates this shift. WM methods used on web data enable to achieve the main aCRM objectives which can be placed in a two-dimension conceptual map: the "prospects" axis (ordinate) ranges from prospects to existing customers, while the "research type" axis (abscissa) ranges from who profiling study (who) to behaviour study (how). The use of WM methods and techniques is assumed to turn operational web data into meaningful and relevant knowledge of current and prospective web customers. Integrating the WM process into the aCRM process provides a formal and structured guideline which improves efficiency and effectiveness of the WM process while providing insightful knowledge to the aCRM applet of CRM.

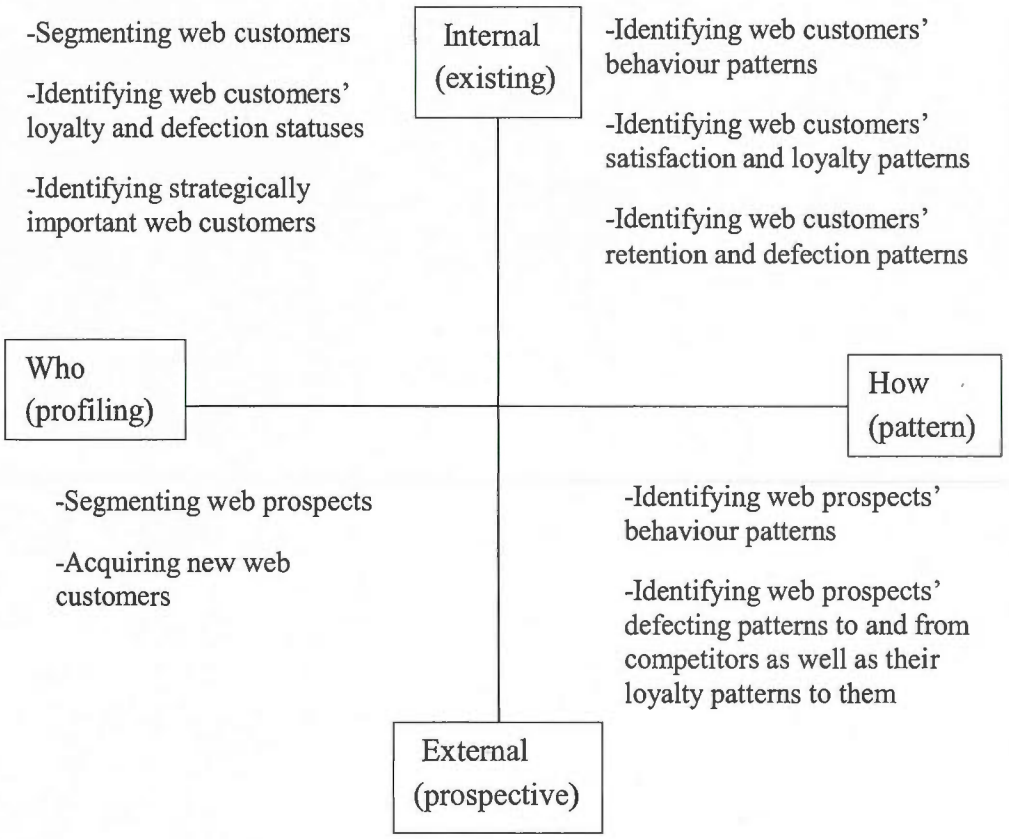

Figure 2.2 aCRM for a web users' knowledge acquisition framework. (Adapted from Xu and Walton, 2005.)

2.2. Defining Knowledge Management (KM)

2.2.1. Integrating WM-enabled aCRM into KM

Knowledge Management (KM) is the process of managing organizational knowledge to create value and sustain the competitive advantage through creating, communicating and applying the knowledge acquired from interactions with customers to maximize business growth and value (Tiwana, 2001). Previous research has emphasized the necessity to blend a Knowledge Management (KM) framework, DM and aCRM, to add value to the business, which contributes to the firm's competitive advantage (Ranjan & Bhatnagar, 2011). In fact KM focused on aCRM (and hence DM) develops intangible assets such as knowledge capital

and relationship capital to create and deliver innovative products and services; manage and strengthen relationships with new and existing customers, partners and suppliers and improves work practices and processes related to the customer (Tiwana, 2001). It is a must to develop high path dependency capabilities.

aCRM capitalizes on oCRM and eCRM/mCRM which gather respectively *knowledge about customers*, and *knowledge from customers* (Lei & Tang, 2005). That knowledge can be obtained from web server access logs, proxy server logs, browser logs, user profiles, registration data, cookies or user queries (Mihai, 2009), but also search results, web pages content (WCM) or hyperlinks or hypertext links (WSM). The information retrieved by aCRM through the eCRM/mCRM branches can be used for providing customers the benefits they are in need (improvement of a website, products or services recommendations, etc.). This is the *knowledge for customers* (Lei & Tang, 2005), as depicted in the lower box in Figure 2., which may impact positively the web experience.

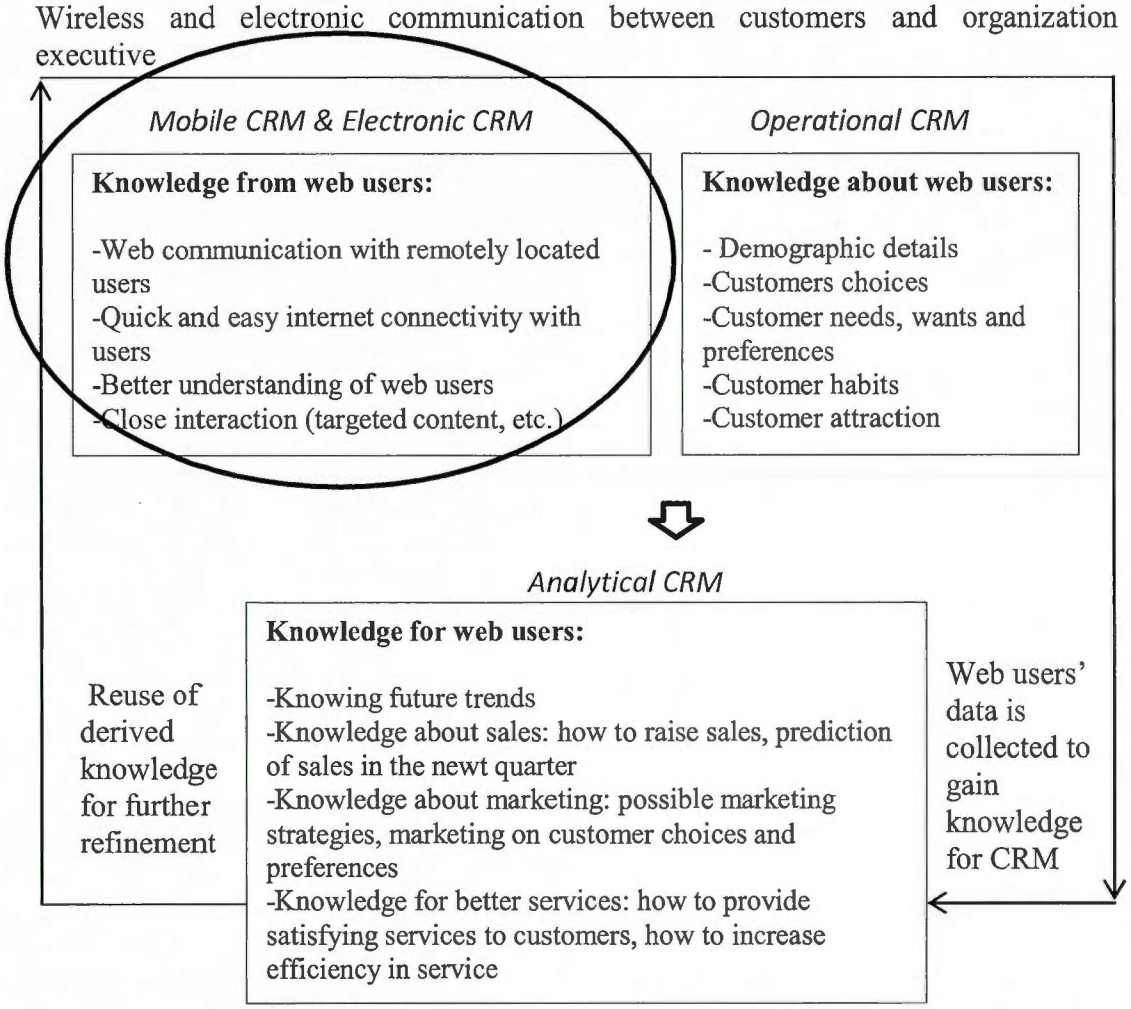

Figure 2.3 Technology-enabled KM for aCRM. (Adapted from Lei & Tang (2005) and Ranjan & Bhatnagar (2011).)

The increase in web data volumes and the fast-growing capacities of warehouses offer unique opportunities to be seized through a better understanding of customers by integrating both WM processes and aCRM altogether. By leveraging the strengths of WM, it is assumed that the major aCRM problems, identified through Xu and Walton's (2005) framework, can be better fulfilled in a web perspective. The four major methods of WM, namely the descriptive-oriented methods of clustering and association analysis; and predictive-oriented methods of classification and prediction (Bazsalicza & Naim, 2001), are congruent with aCRM

objectives displayed in Figure 2.. Unlike oCRM and eCRM/mCRM, the outcome of aCRM is to provide *information for customers*. This is the information gathered by analytical means such as WM, which should ideally be combined to DM-analyzed offline data about customers and other "islands of customer data" obtained from other legacy or enterprise-wide systems, *i.e.*, Enterprise Resource Planning (ERP), Sales Force Automation (SFA), Supply Chain Management (SCM), Vendor Relationship Management (VRM), Material Requirements Planning (MRP), Master Production Schedule (MPS), Decision Support Systems (DSSs), Supplier Relationship Management (SRM), etc. (Chan, 2005). Software engineering approaches may even enable interoperable services, hence allowing information moving between each enterprise software package through Service Oriented Architecture (SOA), Enterprise Integration Applications (IAE) or Software as a Service (SaaS) design principles (Bedell *et al.*, 2007).

But such an analysis goes beyond the scope of this piece of research. The focus, here, is on utilizing WM tools on web data issued by existing and prospective customers online in order to provide an additional digital information dimension to the aCRM framework so that this framework may provide systematic and powerful knowledge for improving the web operations of businesses be they *brick-and-clicks* or *pure players*. As depicted in figure 7, ultimately, such a WM-integrated framework will extract "knowledge for customers" such as knowing future trends, predictions for sales, predictions for marketing and solutions for better services (Xu & Walton, 2005). This knowledge will lead to multi-level improvements and may in return be analyzed again with WM methods and techniques closing the loop and developing a virtuous circle which may lead to a knowledge-driven competitive advantage

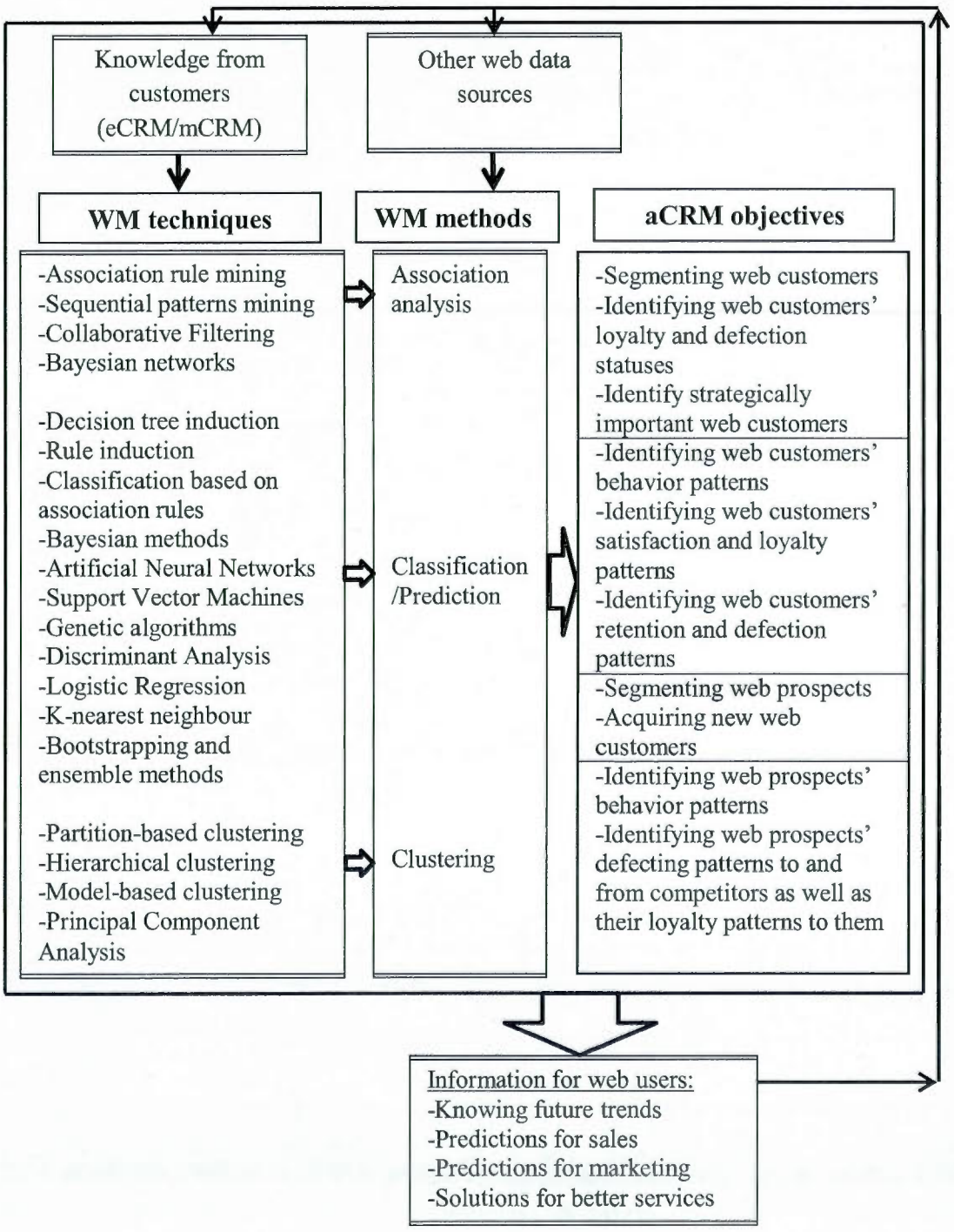
Knowledge from customers (eCRM/mCRM)
Other web data sources
WM techniques
WM methods
aCRM objectives
-Association rule mining
-Sequential patterns mining
-Collaborative Filtering
-Bayesian networks
Association analysis
-Segmenting web customers
-Identifying web customers' loyalty and defection statuses
-Identify strategically important web customers
-Decision tree induction
-Rule induction
-Classification based on association rules
-Bayesian methods
-Artificial Neural Networks
-Support Vector Machines
-Genetic algorithms
-Discriminant Analysis
-Logistic Regression
-K-nearest neighbour
-Bootstrapping and ensemble methods
Classification /Prediction
-Identifying web customers' behavior patterns
-Identifying web customers' satisfaction and loyalty patterns
-Identifying web customers' retention and defection patterns
-Segmenting web prospects
-Acquiring new web customers
-Partition-based clustering
-Hierarchical clustering
-Model-based clustering
-Principal Component Analysis
Clustering
-Identifying web prospects' behavior patterns
-Identifying web prospects' defecting patterns to and from competitors as well as their loyalty patterns to them
Information for web users:
-Knowing future trends
-Predictions for sales
-Predictions for marketing
-Solutions for better services

Figure 2.4 WM-enabled framework for achieving aCRM objectives. (adapted from Lei & Tang (2005) and Ranjan & Bhatnagar (2011).

2.2.2. Modeling of the specific conceptual framework

DND Intranet (2004) defined KM as an integrated systematic approach which, when applied to an organization, enables the optimal use of timely, accurate and relevant information. It also facilitates knowledge discovery and innovation, fosters the development of a learning organization and enhances understanding by integrating all sources of information, as well as individual and collective knowledge and experience. Acquiring knowledge is an essential element in the aCRM process but the systematic and cross-functional dissemination of knowledge among the organization refers to the KM part of the Business Intelligence process. It is therefore critical to develop a framework for aCRM considering WM tools and techniques from a KM perspective. The acquired knowledge will be used as an input for decision-making purposes. A standard framework for integrating WM in the broader KM perspective is a useful tool for synthesizing the process but it must be underlined that each WM project is inherently specific and may require adjustments. Hence, prior to any analysis, it is necessary to identify the objectives of the project and determine the resources allocated to achieve the objectives. A plan needs also to be developed to orientate the process toward the right type of data collection and analysis. The following procedure, illustrated in figure 8, describes the tentative model for knowledge-enabled aCRM using WM methods and techniques, based on, and adapted from research of Hameed (2004), Lei & Tang (2005), Xu & Walton (2005), Ranjan & Bhatnagar (2011), Zhang & Segall (2008), Wahlberg *et al.* (2009) and Albadvi & Shahbazi (2010).

(1) ***Collection of relevant web data from multiple web sources:*** the web data is generated by customers or prospects who access to the web site of a business from various touch points and sources as well as by different means. The data encompasses general access logs and usage profiles, generally used for WUM; search results and web pages (text, XML, HTML and multimedia formats), generally used for WCM; dynamic and static internet links, generally used for WSM. This process refers to the two first stages of the WM process developed by Zhang and Segall (2008), namely *resource finding and*

retrieving as well as *information selection and preprocessing* which consist of selecting and cleaning data from gross web log files, registration data, documents, usage attributes, etc. (depending on which type of WM methods are used afterwards), but also identifying user, identifying session and path completion, which is particularly relevant for WUM (Albadvi & Shahbazi, 2010).

(2) ***Integrated Data:*** the clean data is gathered into an Integrated Database. This formatted data consists thus primarily of operational web data and provides information pertaining to web users, which can be used as such without analysis. This type of real-time information can be useful for marketing, sales or by any department which might need plain web data (Xu & Walton, 2005).

(3) ***WM descriptive and predictive techniques usage for exploring and analyzing the web data:*** the formatted data needs to be transformed according to the type of WM technique(s) that will be applied to it. Discovery of unknown and useful information from the formatted data, hence patterns discovery, is done primarily with mere descriptive statistical (path) analysis. On the more interesting inferential level, there are four major WM methods known as clustering, association analysis, classification and prediction which encompass various techniques. The process of selecting the appropriate type(s) of analysis is primarily defined by the WM outcome(s) sought (web personalization, building profiles, etc.). Descriptive techniques such as clustering and association analysis will be used to perform segmentation and cross-selling/up-selling tasks, which lead to profiling, portfolio building, and increasing web sales, etc. Predictive techniques are better suited for classification and regression tasks, hence concentrating on the behavior analysis of existing and/or prospect customers as depicted in Xu & Walton's (2005) framework. This stage refers to the *patterns recognition and analysis* step of Zhang and Segall's (2008) WM process. Once a model, segment or score is derived from the data, the model is tested on the remaining part of the data, the test set. This refers to the *validation and interpretation* steps but it is included in the present exploratory and analytical step since it only (in)validates the resulting models.

(4) ***Knowledge data warehouse of web user information:*** the knowledge data warehouse differs from the data warehouse since it contains analyzed and validated data and is built after the discovery of patterns in the formatted data typically stored in the data warehouse. The freshly-acquired knowledge is then disseminated across the organization, especially to marketing and sales through Management Information Systems (MISs), Business Intelligence (BI) processes. Higher in the hierarchy, managers can access these insightful WM-generated metrics, models and overall knowledge through high-end managerial Dashboards, Balanced Scorecards, Business Performance Management (BPM), Business Activity Monitoring (BAM), Business Rules Engines (BRE) systems which enable changing business rules in real-time (Bedell *et al.*, 2007). These systems can be connected to the knowledge data warehouse especially if it is an in-company item as opposed to a warehouse available at the server level, *i.e.*, Cloud Computing services. Information Systems monitor overall business aspects and are thus fed in real-time with streams or batches of knowledge acquired through WM projects. It is important to store safely the web user knowledge so that it may be used anytime by executives and employees at multiple hierarchical levels, as a support for operational, tactical and strategic decision-making processes.

(5) ***KM of web user data:*** the knowledge obtained is *knowledge for web users*, be they customers or prospective customers of the business (Lei & Tang, 2005). Other applications also include *information about web users* used in oCRM or cCRM but also *information from web users* typically addressed with mCRM/eCRM.

The acquired knowledge can then be used in eCRM and mCRM because these CRM fields are highly concerned with any types of web-based interactions. The acquired knowledge can also be used, to a lesser extent, in oCRM. However, the bulk of information is typically answering web-oriented aCRM questions and should be made available throughout the organization by means of a KM framework. As recommended by Ranjan and Bhatnagar (2011), the whole process of aCRM is managed through the KM framework as follows (Shaw *et al.*, 2001): (a) creation of knowledge about web users, (b) storage of knowledge about web users, (c) dissemination of knowledge about web users, (d) application of knowledge about web users. The KM framework needs to be integrated to the three-level model for

knowledge-enabled aCRM using WM. It is assumed that such an approach will considerably reduce the issue of losing not only gross data but also knowledge at all steps of the knowledge creation process. Also, the achievement of aCRM objectives can be done in a more automatic and systematic manner which reduces cost and time, while minimizing all kinds of waste. It is important to stress that such a framework for knowledge-enabled aCRM using WM methods and techniques intends primarily to derive knowledge from web data, which originated from web users be they customers or not and which seeks therefore to provide knowledge for those web users in order to optimize the CRM process occurring on the internet channel. Consequently, the model deals specifically with evolving online web data and not offline data which are stored in offline databases. While most authors in the literature focused on the broad area of aCRM in a KM framework applied to offline data (Ranjan & Bhatnagar, 2011; Xu & Walton, 2005; Kimball & Ross, 2002; Tiwana, 2001), the following tentative framework is a proposition for a framework applied to web-sourced data.

The knowledge-enabled aCRM using the WM model is summarized into three levels, in figure 8 below. We can see that these three levels are:

(01) *Data store of web data warehouse:* collection of relevant data from multiple web sources and the Integrated Data stages

(02) *Data discovery and analysis:* WM descriptive and predictive techniques usage for exploring and analyzing web data stage

(03) *KM applications to aCRM:* Knowledge data warehouse of web user information and KM of web user data stages

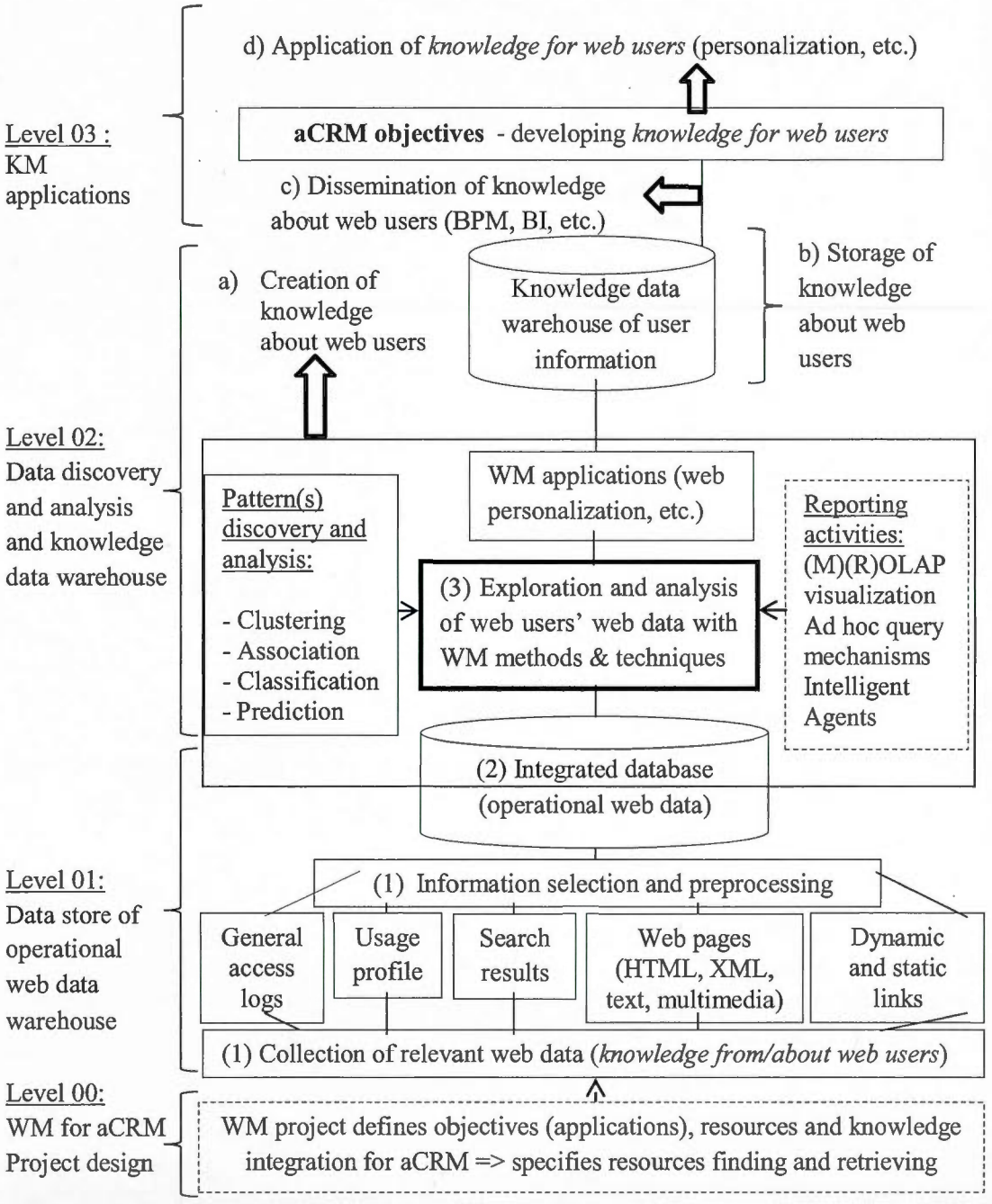

Figure 2.5 Framework for knowledge-enabled aCRM using WM methods and techniques

2.3. Qualitative research components

Web user knowledge storage, creation, dissemination and application are all managed through the model exposed previously. WM methods and techniques are used to reach aCRM objectives (KM applications) in a KM procedure.

The research components are developed below. Given a company A, the *existing web customers* group refers to those customers who purchased at least once from the web channel of A at any of its digital touch points. These customers purchase from the business through digital but not necessarily through non-digital channels. The *prospective web customers* group refers to those customers who have never purchased from the web channel of A at any of its digital touch points. These customers are not necessarily customers of A but if they are, they purchase from A through any channel except through the web channel.

Component 1: Usage of WM applied to the profiling of existing web customers (WHO – INTERNAL dyad)

Identifying to what extent it may be possible to use the main WM methods (and techniques) in order to analyze current web customers' profiles, strategic importance, loyalty and defection statuses for further identification of web customer churn rates and implementation of segmentation and targeting e-marketing activities on the web.

Component 2: Usage of WM applied to the identification of existing customers patterns on the web (HOW – INTERNAL dyad)

Identifying to what extent it may be possible to use the main WM methods (and techniques) in order to pinpoint existing web customers' behavior patterns, as well as their satisfaction and loyalty patterns and their retention and defection patterns, to optimize, among others, cross-selling/up-selling offerings.

Component 3: Usage of WM applied to the profiling of prospective customers on the web (WHO – EXTERNAL dyad)

Identifying to what extent it may be possible to use the main WM methods (and techniques) in order to analyze prospective web customers' profiles, preferences, needs, habits, etc. in order to acquire them.

Component 4: Usage of WM applied to the identification of prospective web customers patterns on the web (HOW – EXTERNAL dyad)

Identifying to what extent it may be possible to use WM methods (and techniques) in order to pinpoint prospective web customers' behavior patterns, as well as their defecting patterns to and from competitors, and their loyalty patterns to competitors.

2.3.1. Research questions and propositions

In the following, the abbreviation "RQ" stands for "Research Question" and "P" stands for "Proposition".

Component 1: Usage of WM applied to the profiling of existing web customers (WHO – INTERNAL DYAD 1)

RQ 1: [WHO – INTERNAL DYAD 1a] to what extent do Clustering and Classification methods, applied to web data (web log data, etc.), provide accurate profiles of existing web customers?

- P1: Web data generated by existing web customers are sufficiently detailed and accurate to provide a strong basis for the creation of precise profiles about those existing web customers.
- P2: Clustering and Classification techniques applied to web data (web log data, search results and web pages, etc.) create homogeneous groups of existing web customers.

RQ 2: [WHO – INTERNAL DYAD 1b] To what extent do classification and prediction methods, applied on web data, identify the profit-cost ratio and the Recency-Monetary-

Frequency (RFM) of purchases made by individuals, on which the Customer Lifetime Value (CLV) is based, to identify strategically important, existing web customers?

- P1: Web data generated by existing web customers encompass enough information about the profit-cost, the Recency-Monetary-Frequency (RFM)-based CLV of purchases made by existing web customers, which contributes to identify those strategically important customers
- P2: Classification and prediction methods applied to web data predict the value of a given web customer to identify strategically important existing web customers.

RQ3: [WHO – INTERNAL DYAD 1c] To what extent do classification and clustering methods identify existing web customers' loyalty or defection statuses?

- P1: Web data generated by existing web customers indicate whether a web customer is loyal to a given business or defects from that business.
- P2: Classification and clustering methods applied to web data predict membership of an individual to the loyal or defecting customer group.

Component 2: Usage of WM applied to the identification of existing web customers patterns on the web (HOW – INTERNAL DYAD 2)

RQ4: [HOW – INTERNAL DYAD 2a] To what extent do clustering, association analysis, classification and prediction methods applied to web data identify existing web customers' behavior on the web?

- P1: Web data generated by existing web customers highlight the particular behavioral patterns of existing web customers when they are navigating on the internet.
- P2: Clustering, association analysis, classification and prediction methods applied to web data provide descriptive and predictive modeling of web customers' behavior on the internet.

RQ5: [HOW – INTERNAL DYAD 2b] To what extent do clustering, association analysis, classification and prediction methods applied to web data capture how existing web customers develop satisfaction and loyalty on the internet?

- P1: Web data generated by existing web customers describe the existing web customers' satisfaction and loyalty patterns on the internet.
- P2: Clustering, association analysis, classification and prediction methods applied to web data grasp the dynamics of existing web customers' satisfaction and loyalty patterns on the internet.

RQ6: [HOW – INTERNAL DYAD 2c] To what extent do clustering, association analysis, classification and prediction methods identify how existing web customers' remain attached to or defect from a given business on the internet?

- P1: Web data generated by existing web customers describe existing web customers' retention and defection patterns on the internet.
- P2: Clustering, association analysis, classification and prediction methods applied to web data capture the dynamics of existing web customers' retention and defection patterns on the internet.

Component 3: Usage of WM applied to the profiling of prospective customers on the web

RQ7: [WHO – EXTERNAL DYAD 3a] To what extent do classification and prediction methods segment prospective web customers?

- P1: Web data generated by prospective web customers are sufficiently detailed and accurate to provide a strong basis for the creation of precise profiles about those prospective web customers.
- P2: Classification and prediction methods applied to web data create homogeneous groups of prospective web customers.

RQ8: [WHO – EXTERNAL DYAD 3b] To what extent do clustering, association analysis and classification methods provide insightful information about prospective web customers' preferences, needs, habits etc. to develop targeted e-marketing and e-commerce strategies to acquire them?

- **P1:** Web data generated by prospective web customers are sufficiently detailed and accurate to provide various and complementary characteristics about those prospective web customers such as their preferences, needs, habits, to be used for acquiring those customers.
- **P2:** Clustering, association analysis and classification methods generate relevant information about prospective web customers' characteristics which can be used for further targeted marketing and sales efforts to acquire them.

Component 4: Usage of WM applied to the identification of prospective web customers patterns on the web

RQ9: [HOW – EXTERNAL 4a] To what extent do clustering, association analysis, classification and prediction methods, applied to web data, identify prospective web customers' behavior on the internet?

- **P1:** Web data generated by prospective web customers highlight the particular behavioral patterns of prospective web customers when they are navigating on the internet.
- **P2:** Clustering, association analysis, classification and prediction applied to web data provide descriptive and predictive modeling of prospective web customers' behavior patterns on the internet.

RQ10: [HOW – EXTERNAL 4b] To what extent do clustering, association analysis, classification and prediction methods applied to web data identify how prospective web customers defect to and from competitors as well as how they are loyal to competitors on the internet?

- **P1:** Web data generated by prospective web customers highlight prospective customers' defection patterns to and from competitors as well as their loyalty patterns to competitors on the internet.
- **P2:** Clustering, association analysis, classification and prediction methods applied to web data identify prospective web customers' defection patterns to and from competitors as well as their loyalty patterns to competitors, on the internet.

2.4. Specific Conceptual framework

This study focuses on exploring the benefits yielded by WM methods and techniques embedded into a KM procedure for achieving the major aCRM tasks and objectives relating to existing and prospective customers who interact with a business through any touch points of the internet channel. The scope of research is limited to the web context and thus web customers. It is beyond the scope of this research to analyze the many technical and technological aspects of WM in details and the technical adequacy of specific WM methods/techniques to achieve given aCRM objectives. However, the research strives to emphasize the valuable knowledge that can be extracted from web data through a systematic and formal approach which integrates standard DM techniques applied to web data in broader enterprise-wide KM and aCRM processes. aCRM is the missing link in CRM strategy to implement BI in the organization (Hall, 2004). It has been acknowledged that DM represents the lion's share of the aCRM process (Wahlberg *et al.,* 2009) and that aCRM needs to be built on DM techniques and KM approach to meet business challenges (Xu & Walton, 2005). DM manipulates offline and often rather static data. Businesses generally focus on that type of research without integrating the gigantic flow of web data. The latter is thus under-exploited, although it has been demonstrated through extensive research that, if well-processed, web data provide extremely critical information for businesses (Liu, 2007; Liu, 2011). It is argued that WM is another crucial component of a DM-built aCRM-KM framework since it provides valuable information about web patterns of customers and/or prospects online. It provides the missing piece of the puzzle for a true 360° view of organizations' consumers. WM turns the fast-growing gross web data into precious *knowledge for customers* (Lei & Tang, 2005), to improve dramatically cross-disciplinary holistic marketing and sales functions, in a relational marketing paradigm.

The scope of this research is limited to analyzing the benefits of WM methods and techniques on the analytical CRM (aCRM) applet of the marketing function, in the framework of a Knowledge Management (KM) perspective. That is, to understand how WM might be optimally integrated to the blended aCRM and KM framework to leverage insightful knowledge about existing and prospective web customers who interact with online businesses.

The *WM block* is comprised of web data types, methods and techniques which result in specific WM applications. Given the fact that methods encompass techniques, the research propositions investigate the benefits of WM methods to reach aCRM objectives. Different sorts of web data may be used. General access logs and usage profiles, are both considered for WUM. Search results and web pages with content such as text, HTML, XML (RSS feed, etc.), multimedia (audio, video, etc.) are considered for WCM, while dynamic and static web links are considered for WSM. OLAP/visualization tools, knowledge query mechanisms and intelligent agents are used to describe statistically the web patterns (Cooley *et al.*, 1997). On a more analytical level, there are four types of WM methods applied to these web data for discovering and analyzing meaningful patterns, namely clustering, association analysis, classification and prediction (Bazsalicza & Naim, 2001; Srivastava *et al.*, 2000). Combining those methods and techniques enables to anticipating future trends, smoothly predicting sales (how to raise online sales, prediction for online sales in the next quarter according to seasonal and other contingent effects), optimizing marketing solutions (possible e-marketing strategies, marketing on customer choices and preferences, marketing according to prediction by WM tools) and developing solutions for better customer service (how to provide satisfied services online to customers, how to increase efficiency in online service(s)) (Lei & Tang, 2005: Ranjan & Bhatnagar 2011).

The *aCRM block* is comprised of the major tasks and objectives that are typically fulfilled by aCRM in a knowledge acquisition perspective (Wahlberg *et al.*, 2009), as identified in Xu and Walton's (2005) framework, and adapted to a web-oriented approach:

(1) The first dyad focuses on profiling existing (internal) web customers, which requires segmenting existing web customers, identifying their loyalty and defection statuses, and identifying strategically important web customers.

(2) The second dyad focuses on identifying existing (internal) customers' patterns on the web which requires tracking and monitoring existing web customers' behavior patterns, identifying web customers' satisfaction and loyalty patterns and identifying retention and defection patterns.

(3) The third dyad focuses on profiling prospective (external) web customers, which requires segmenting prospective web customers and identifying their intrinsic characteristics to make them become existing (internal) customers.

(4) The fourth dyad focuses on identifying prospective (external) customers' patterns on the web which requires tracking and monitoring prospective web customers' behavior patterns, identifying defecting patterns to competitors, identifying defecting patterns from competitors, and identifying patterns of customers loyal to competitors.

The *KM framework* as defined by the DND Intranet (2004), integrates the two previous blocks together, enabling congruency between them, in an online context. Hence, the KM framework is the blueprint for the whole WM process from scratch to high-end application. The knowledge-enabled aCRM framework provides *knowledge for customers* that can lead to one or several of the nine WM applications listed in table 3, depending on the pre-defined WM project. It is the knowledge used to enhance customers' experience, customers' flow and overall customers' satisfaction. The whole process of WM is managed through the KM framework: *customer knowledge creation*, *customer knowledge storage*, *customer knowledge dissemination* and *customer knowledge application* to aCRM, other CRM applets and more broadly to every corporate process or function needing such highly competitive knowledge. Figure presents the conceptual framework as the summary of all three blocks put together. The study focuses primarily on the Knowledge creation and storage stages in the KM framework. It seeks to understand to what extent the four methods of WM, when applied to web data, are successful in achieving the 12 major aCRM objectives for e-commerce (transactional) websites. The section "results" investigates in-depth to what extent WM methods and techniques are suitable for achieving each of these 12 major aCRM objectives. A final refined schema will eventually represent which aCRM objectives may indeed be achieved by means of WM methods and techniques.

Figure 9 represents the specific conceptual framework of this study as developed thus far.

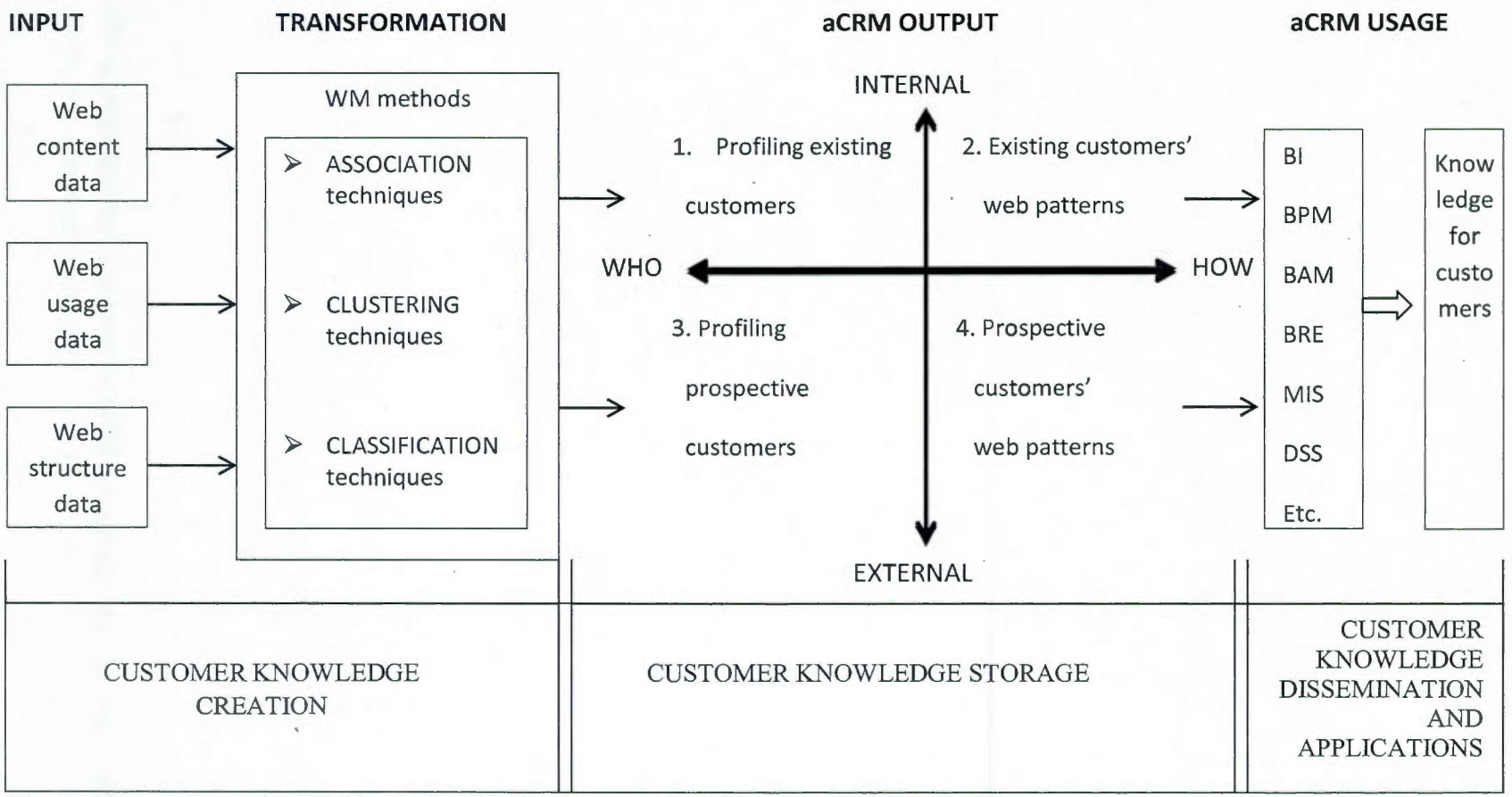

Figure 2.6 A tentative WM-built aCRM-KM framework. (specific conceptual framework)

PART TWO

DESCRIPTION OF THE RESEARCH PROJECT

CHAPTER 3

RESEARCH METHODOLOGY

3.1. Research design

3.1.1. Research paradigm

A paradigm is a constellation of techniques, values, beliefs, etc. shared by a given community[29]. While natural sciences and artificial sciences have well-established and strong research paradigm bases, administration sciences tend to have a weaker determination of their research paradigm. It can be either positivist or constructivist. The current piece of research derives data from sensory experience and logically treats these data while considering both of these processes as being the exclusive source of authentic knowledge. In fact, because of the lack of research on the subject of WM in a marketing perspective, there are very theoretical frameworks which could have given rise to empirical testing. Also, introspective and intuitional attempts to gain knowledge are rejected. For all those reasons, this study is embedded into a positivist paradigm with its subsequent ontology, epistemology and methodology.

3.1.2. Ontological orientation

Ontology refers to the study of the general properties of what exists (Behling, 1980). According to Laverty (2003), from an ontological perspective, positivist frameworks view reality as something 'out there' to be apprehended (Denzin & Lincoln, 2000). There

[29] Wikipedia: http://fr.wikipedia.org/wiki/Paradigme (retrieved on 04-02-2012).

is an assumption that the world is structured by law like generalities that can be identified, predicted, manipulated or controlled to yield universal statements of scientific theory (Munhall, 1989). Polkinghorne (1983) described this as a 'received' view of science, as something apart from ourselves that we receive and can study, rather than as something we create.

On the other hand, the interpretivist framework of inquiry supports the ontological perspective of the belief in the existence of not just one reality, but of multiple realities that are constructed and can be altered by the knower (Laverty, 2003). Reality is not something 'out there', but rather something that is local and specifically constructed. Realities are not more or less true, rather they are simply more or less informed (Denzin & Lincoln, 2000). Polkinghorne (1983) described this paradigm as an attitude about knowledge, not a school of thought. Knowledge is seen as the best understandings we have been able to produce thus far, not a statement of what is ultimately real.

The current piece of research is built upon a literature review that leads toward the construction of a general and specific model. The underlying assumption is that of a natural determinism. The real is ruled by immutable laws, observable cause-and-effect mechanisms that could eventually be scientifically measured. As a matter of fact, it does not correspond to the relativist ontology, which states that there are multiple realities that are socially constructed and non-governed by natural and causal laws. Therefore, the study rather belongs to the realist ontology type. In fact, this study acknowledges the existence of an objective reality which can be understood by means of observations for internal validity.

3.1.3. Epistemological orientation

According to Laverty (2003), from an epistemological stance, the positivist tradition saw a duality between the object of inquiry and the inquirer. Researchers are described as attempting to assume a stance of a disinterested scientist (Denzin & Lincoln, 2000). The researcher is seen as being able to obtain a viewpoint, devoid of values or biases (Polkinghorne, 1983).

Epistemologically, the constructivist framework sees a relationship between the knower and the known. The notion of value-free research has been challenged as questionable and it is believed that attempts to attain such a stance have resulted in the loss of certain kinds of knowledge about human experience, such as meaning making (Cotterill & Letherby, 1993; Jagger, 1989). Polkinghorne (1983) viewed research as a human activity in which the researcher as knower is central and Denzin and Lincoln (2000) viewed the investigator and the investigated as interactively linked in the creation of findings, with the investigator as a passionate participant (Laverty, 2003).

Since this study is embedded in the positivist paradigm, both the observer and the observed phenomenon are separable. The researcher is independent from its object. Consequently, this study does not belong to the monist subjectivist or interpretivist type which assumes a strict non-separability between the researcher and the researched object. Rather, this study follows an objective dualist epistemology where the researcher has no influence on the researched object.

3.1.4. Methodological orientation

Methodology is a guideline system for solving a problem, with specific components such as tasks, phases, methods, techniques and tools (Irny & Rose, 2005). Methodologically, specific methods are utilized to try to ensure the absence of the investigator's influence or bias, as this is perceived as a threat to the validity of the results (Laverty, 2003). Consequently, benchmarks of internal/external validity, reliability and objectivity have been developed to facilitate this process (Denzin & Lincoln, 2000).

A set of research questions and their related research propositions are formulated and meant to be answered by a pool of experts to reach natural determination. Consequently the method that will be used in order to develop knowledge is not hermeneutic with hermeneutic method assumptions. Knowledge is not constructed conjointly (co-constructed) between the researcher and the respondents in a continuous (re)iterative, (re)analytical and critical process leading to the development of co-constructed representations emic-based (interpretations by the members of the phenomenon that is analyzed) or etic-based (interpretations by the researcher) interpretations. Rather, the present research approach follows the logical-empirical approach based on the cartesian

principle of analytic division and breakdown (Osborne, 1994). It follows a hypothetico-deductive model as determined by Whewell in 1837. Some forms of syllogism are also used as part of the logical reasoning type of the realist ontology (Frede, 1975).

3.1.5. Design type

Conducting formal research on WM in marketing is still a relatively new field. Integrating WM to the aCRM/KM frameworks is even less researched. There are relatively few information and frameworks of reference available. Frameworks need to be adapted from adjacent research. The present work is, therefore, a radical innovation in marketing research since it:

> "creates a dominant new design that incorporates links between the previously-identified new core concepts" (Henderson & Clark, 1990).

The research design is exploratory, in the form of in-depth semi-structured interviews, because there is an increasing interest in the application of DM to aCRM and KM/BI frameworks, but little information exists in the literature about integrating WM to aCRM-KM frameworks. Much advances have been made in the field of computer sciences, systems intelligence and artificial intelligence but few studies sought to explore the combination of DM/WM tools to Intelligent Systems such as CRM for instance and even less explored the effect of the synergy between the DM-CRM combinations and their tight intertwining into KM frameworks. Exploratory research provides insightful information, discovers ideas while improving the understanding of the phenomenon in order to generate tentative results which will form the basis for further exploratory or confirmatory research (Malhotra, 2010). Consequently, the exploratory purpose of this research is to:

- Identify possible ways of action for tackling the identified problem
- Isolate key variables and relationships for future research on the subject
- Provide information for the development of an approach to the identified problem
- Establish priorities for future research on the subject

It must be stressed that, although following a scientific approach, the methodology is not intended to provide generalizable results, but answers to research questions supported by inferred research propositions.

More specifically, about 45 respondents were directly approached to participate to the in-depth, semi-structured interviews. 11 of them took part in the project, but one of the interviews did not yield the expected level of insight and its results were thus not included in the analysis. The response rate is thus 33%. The interview guide that was used for conducting the discussion is included in appendix 1.

The interviews were recorded and then entirely transcribed. The verbatims were then placed in an answer matrix by themes. Hence for a given research theme (*e.g.* usage of WM for segmenting existing web customers) the answers of the 10 respondents were analyzed, aggregated if similar, polarized (*e.g.* most answers tend towards negative or positive, etc.), summarized and finally the most important elements were extracted with relevant quotations associated to them as support.

3.1.6. Information needed

The information sought is broadly defined, the research process is flexible, versatile and unstructured, and it will be analyzed qualitatively. Table 3.1 shows the information needed for each research proposition related to each research question.

Table 3.1

Description of the information needed per research proposition

RESEARCH QUESTIONS	RESEARCH PROPOSITIONS	INFORMATION NEEDED
Component 1: Usage of WM applied to the profiling of existing web customers (WHO – INTERNAL DYAD 1)		
RQ 1: [WHO –	**RP1:** Web data generated	*Characteristics of the web*

INTERNAL DYAD 1a] To what extent do clustering and classification methods applied to web data (web log data, etc.) provide accurate profiles of existing web customers?	by existing web customers are sufficiently detailed and accurate to provide a strong basis for the creation of precise profiles about those existing web customers.	*data generated by existing web customers make this data useful for developing accurate profiles of existing web customers.*
	RP2: Clustering and classification methods applied to web data (web log data, search results and web pages, etc.) create homogeneous groups of existing web customers	*Features and attributes of the clustering and classification methods make them useful for creating groups of similar existing web customers.*
RQ2: [WHO – INTERNAL DYAD 1b] To what extent do classification and regression methods applied to web data, identify the profit-cost ratio and the Recency-Monetary-Frequency (RFM) of purchases made by individuals, on which the Customer Lifetime Value (CLV) is based, to identify strategically important, existing web customers?	**RP1:** Web data generated by existing web customers encompass enough information about the profit-cost, the Recency-Monetary-Frequency (RFM)-based CLV of purchases made by existing web customers, which contributes to identify those strategically important customers	*Extent to which web data display the recency-monetary-frequency (RFM) values of an existing web customer's purchase(s) online as well as the ratio of costs needed for stimulating the purchase on the profits received from the purchase.*
	RP2: Classification and prediction methods applied to web data predict the value of a given web customer to identify strategically important existing web customers.	*Features and attributes of classification and prediction methods make them useful for identifying strategically important existing web customers.*
RQ3: [WHO – INTERNAL DYAD 1c]	**RP1:** Web data generated by existing web customers	*Extent to which the web data display indications*

To what extent do clustering and classification methods identify existing web customers' loyalty or defection statuses?	indicate whether an existing web customer is loyal to a given business or defects from that business.	*about the loyalty or the defection of an existing web customer from a business.*
	RP2: Clustering and classification methods applied to web data predict membership of an individual to the loyal or defecting customers group.	*Features and attributes of the clustering and classification methods make them useful for predicting the membership of an existing web customer to the loyal or the defecting customers group.*
Component 2: Usage of WM applied to the identification of existing customers' patterns on the web (HOW – INTERNAL DYAD 2)		
RQ4: [HOW – INTERNAL DYAD 2a] To what extent do clustering, association analysis, classification and prediction methods applied to web data identify existing web customers' behavior on the internet?	**RP1:** Web data generated by existing web customers highlight the particular browsing behavior of existing web customers when they are navigating on the internet.	*Characteristics of the web data generated by existing web customers which display how those customers behave on the internet.*
	RP2: Clustering, association analysis, classification and prediction methods applied to web data provide descriptive and predictive modeling of web customers' consumer behavior on the internet.	*Features and attributes of clustering, association analysis, classification and prediction methods make them useful for describing and predicting the behavior of existing web customers online.*
RQ5: [HOW – INTERNAL DYAD 2b] To what extent do	**RP1:** Web data generated by existing web customers describe the existing web	*Characteristics of the web data generated by existing web customers which*

clustering, association analysis, classification and prediction methods applied to web data capture how existing web customers develop satisfaction and loyalty on the internet?	customers' satisfaction and loyalty patterns on the internet.	*describe the patterns that make those customers satisfied and loyal on the internet.*
	RP2: Clustering, association analysis, classification and prediction methods applied to web data grasp the dynamics of existing web customers' satisfaction and loyalty patterns on the internet.	*Features and attributes of the clustering, association analysis, classification and prediction methods make them useful for identifying the patterns that make existing web customers satisfied and loyal on the internet.*
RQ6: [HOW – INTERNAL DYAD 2c] To what extent do clustering, association analysis, classification and prediction methods identify how existing web customers' remain attached to or defect from a given business on the internet?	**RP1:** Web data generated by existing web customers describe existing web customers' retention and defection patterns on the internet.	*Characteristics of the web data generated by existing web customers which describe the patterns that make those customers retained to the business or defect from it on the internet.*
	RP2: Clustering, association analysis, classification and prediction methods applied to web data capture the dynamics of existing web customers' retention and defection patterns on the internet.	*Features and attributes of the clustering, association analysis, classification and prediction methods make them useful for identifying how existing web customers retain to the business or defect from it on the web.*
Component 3: Usage of WM applied to the profiling of prospective customers on the web		
RQ7: [WHO –	**RP1:** Web data generated	*Characteristics of the web*

EXTERNAL DYAD 3a] To what extent do classification and prediction methods segment prospective web customers?	by prospective web customers are sufficiently detailed and accurate to provide a strong basis for the creation of precise profiles about those prospective web customers.	*data generated by prospective web customers which make this data useful for developing accurate profiles of prospective web customers.*
	RP2: Classification and prediction methods applied to web data create homogeneous groups of prospective web customers	*Features and attributes of the classification and prediction methods make them useful for creating accurate profiles of prospective web customers.*
RQ8: [WHO – EXTERNAL DYAD 3b] To what extent do clustering, association analysis and classification methods provide insightful information about prospective web customers' preferences, needs, habits etc. to develop targeted e-marketing and e-commerce strategies to acquire them?	**RP1:** Web data generated by prospective web customers are sufficiently detailed and accurate to provide various and complementary characteristics about those prospective web customers such as their preferences, needs, habits, to be used for acquiring those customers	*Extent to which the web data generated by prospective web customers display prospective web customers' characteristics i.e. preferences, needs, habits etc.*
	RP2: Clustering, association analysis and classification methods generate relevant information about prospective web customers' characteristics which can be used for further targeted marketing and sales efforts to acquire them.	*Features and attributes of the clustering, association analysis and classification methods make them useful for identifying preferences, needs, habits and other characteristics of prospective web customers.*
Component 4 : Usage of WM applied to the identification of prospective web		

customers patterns on the web		
RQ9: [HOW – EXTERNAL 4a] To what extent do clustering, cross-selling, classification and regression methods applied to web data identify prospective web customers' behavior on the internet?	**RP1:** Web data generated by prospective web customers highlight the particular browsing behavior of prospective web customers when they are navigating on the internet.	*Characteristics of the web data generated by prospective web customers which display how those customers behave on the internet.*
	RP2: Clustering, association analysis, classification and prediction methods applied to web data provide descriptive and predictive modeling of prospective web customers' consumer behavior on the internet.	*Features and attributes of the clustering, association analysis, classification and prediction methods make them useful for describing and predicting the behavior of prospective web customers online.*
RQ10: [HOW – EXTERNAL 4b] To what extent do clustering, association analysis classification and prediction methods applied to web data identify how prospective web customers defect to and from competitors as well as how they are loyal to competitors on the internet?	**RP1:** Web data generated by prospective web customers highlight prospective customers' defection patterns to and from competitors	*Characteristics of the web data generated by prospective web customers which display how those customers defect to and from competitors*
	RP2: Clustering, association analysis, classification and prediction methods applied to web data identify prospective web customers' loyalty patterns to competitors, on the internet.	*Features and attributes of the clustering, association analysis classification and prediction methods are useful for identifying how prospective web customers' remain loyal to competitors on the internet.*

3.1.7. Survey methodology

A semi-structured survey has been developed based on the secondary data obtained from the literature review, published material, digitalized material, and netnographical results (LinkedIn expert groups, Viadeo experts network, WM forums and chats, etc.) on the subject; as well as on the primary data obtained from a first round of open and informal interviews with marketing professionals from the Customer Insight & Analytics Council of the Canadian Marketing Association (CMA). Some respondents also agreed to answer the formal subsequent qualitative survey.

Qualitative research was undertaken in the form of individual in-depth interviews with marketing practitioners and academicians, in order to answer to the main research questions and modify or complete their related research propositions. Respondents answer to open-ended questions freely and without any limitation as regards to the length and content of the answers. It is the purpose to collect the true insight of participants, without any bias which may be caused by the researcher or the respondent's field of force. Therefore no comments or criticism are made with regards to their answers and distracting objects are put aside (*i.e.,* cellphone switched off).

The qualitative procedure is direct with in-depth interviews because the problematic is very specific, structured and requires all the attention of the respondent. Focus groups are thus eliminated since this technique tends to provide spontaneous and quick answers from group-energized respondents who are subject to snowball effects from other respondents. Also, focus groups offer limited structure in the discussion which is at the opposite of what is expected. Indirect projective techniques are also left aside since it is not the purpose to discover hidden motivations or needs in the unconsciousness of respondents.

3.1.8. Formulation of the qualitative gathering process and the required tools

A total of 11 qualitative direct in-depth interviews were conducted. Among those, one interview did not yield the required level of insight and was thus not considered for the analysis. Therefore, the results of 10 interviews were used in the framework of this study. Although a low figure, this is appropriate for a non-representative and non-generalizable

exploratory study. Also, the content given by respondents proved to be rich enough in order to build a strong insight into the subject. The respondents were recruited by means of a non-probabilistic sampling method based on judgment. Their selection was based, primarily, on their specific knowledge of web data mining methods and techniques and, to a lesser extent, on their overall web analytics skills.

The type of survey method with which respondents were investigated was not fundamentally important, and multiple tools could be used for conducting those interviews, although e-mail surveys generally tend to provide more honest, well-thought and structured results (Coderre *et al.*, 2004). Consequently, given the length and concentration-intensive aspect of the survey, the survey was administered either personally in face-to-face interviews or via electronic means, attached to e-mails. Given the geographical dispersion of respondents among Canada, the U.S.A., and even Europe, the e-mail solution took precedence over face-to-face interviews in some instances. It also facilitated further semantic analyses of answers since fully-completed surveys were automatically digitalized. The mail method was not considered since it required additional resources such as postal fares, which were not available for this study. It also incurred the risk of lost mail. Besides, mail surveys implied large delays in sending and receiving the survey. The phone method was not really appropriate either, because it was too long a survey to be completed on the phone, moreover recording issues might have risen.

The data collected from the research process was analyzed qualitatively in order to extract key variables and coincident statements to derive common meaning and answer the research propositions in a positivist approach.

3.2. Data gathering process

3.2.1. Gathering method used

The data gathered from the surveys was either hand-written text if the interview was conducted face-to-face or digitally-written text if the survey has been sent by e-mail. Some respondents were chosen by the Canadian Marketing Association (CMA) according

to their level of expertise with the subject. The Marketing Research and Intelligence Association (MRIA) also provided some relevant respondents. Additional respondents were selected by the researcher in a convenience sampling precodure.

3.2.2. Gatherers' identification

The researcher was alone in the data gathering process. Besides, there were no intentions to hire people to conduct interviews or analyze the answers. In fact, the scale and scope of the output data proved to be small enough to be processed by one single person thoroughly.

3.2.3. Control mechanisms to ensure good data quality

Although the data is not quantitative in nature and does not intend to be generalizable, it is nevertheless important to have quality in the open answers. The major constrain in submitting the survey by (e-)mail is monitoring answer patterns, for face-to-face interviews the problem arises rather on the logistics aspect of organizing meetings. It is recognized that pushing respondents too hard to get them answer the survey might lead to poor insights and hence decrease quality in the answers (Malhotra, 2010). Therefore, no more than two recalls were made to non-respondents who complete the survey per e-mail, to avoid the non-response bias. During face-to-face interviews, respondents were free no to answer a specific question. This was never the case though.

Also, no incentives were offered to respondents. This may explain to some extent the very low rate of acceptance to participate to the study, albeit level of expertise and level of commitment might account for the strongest of the reasons for the low answer rate. It was expected that a non-incentivized driven approach would sift through opportunistic behaviors, which may arise when rewards are offered. People without any knowledge on the subject might thus not be tempted to answer abruptly to the survey just to get the reward. Professional respondents might also not have the legal right to accept such rewards since it would be perceived as bribery, or other forms of unethical behavior. Conversely, the lack of incentives might discourage knowledgeable experts to answer to the survey because of the perceived non-usefulness to do so, especially if they go through

hectic quarters. Weighting the pros and cons of each approach it was determined that it would be wiser not to offer any incentive.

Eventually, the survey started with a brief description of the research project to which participants took part. To ensure that each respondent has a homogeneous and similar understanding of the 4 WM methods under study, a glossary included a short definition for each of the 4 WM methods. For an obvious reason of conciseness the definitions of all the WM techniques that are covered by the different WM methods were not listed. Clear understanding of the questions, however, relied on the respondents only.

PART THREE

RESEARCH RESULTS

CHAPTER 4

RESULTS

4.1. Profile of the respondents

Table , below, provides information about the profile of respondents whose answers were deemed satisfactory with regards to the research project on hand.

Table 4.1

Profiles of respondents

Job title	Company sector	Area of expertise	Age range	Income range
Marketing and brand development director	Insurance	Marketing Communication	36 – 55 years old	80 000 $ +
Marketing research agency co-owner	Marketing research and analytics	Marketing Communication	55+ years old	80 000 $ +
Web business manager	Academia	Marketing Communication IT Other	36 - 55 years old	80 000 $ +
Database Administrator (DBA)	IT	IT	36 – 55 years old	80 000 $ +
Scholar	Academia	Marketing Communication IT	36 – 55 years old	80 000 $ +
Marketing research agency director	Marketing research & IT	IT Engineering R&D	36 – 55 years old	60 000 - 79 999 $
Scholar	Academia	Marketing Communication	36 – 55 years old	80 000 $ +
Scholar	Academia	Other: statistics and data-mining	36 – 55 years old	N.S.
Scholar	Academia	Marketing Communication	N.S.	N.S.
Scholar	Academia	Marketing Communication Other: statistics, scoring and Data-mining	N.S.	N.S.
Scholar	Academia	Marketing Communication	N.S.	N.S.

4.2. Themes of the research

This section exposes the output generated by the qualitative interviews with professionals and academicians. It aims at identifying to what extent the WM methods and techniques (association, clustering, classification and prediction) are beneficial to the achievement of the following taxonomy of aCRM objectives for e-commerce (transactional) websites, according to Xu and Walton's (2005) adjusted framework of reference:

THEME 1: Profiling of existing web customers on the internet

1. Segmentation of existing web customers of a website
2. Identification of the strategically important existing web customers of a website
3. Identification of existing web customers' loyalty and defection statuses on a website
4. Conclusion on WM-enabled profiling of existing web customers on a website

THEME 2: Identifying existing web customers' behavior on the internet

5. Identification of existing web customers' behavior on a website
6. Identification of existing web customers' satisfaction and loyalty patterns on a website
7. Identification of existing web customers' retention and defection patterns on a website
8. Conclusion on WM-enabled identification of existing web customers' behavior on a website

THEME 3: Profiling of prospective web customers on the internet

7. Segmentation of prospective web customers of a website

8. Collection of information about prospective web customers' preferences, needs, habits, etc. to develop e-marketing strategies and acquire them
9. Conclusion on WM-enabled profiling of prospective web customers of a website

THEME 4: Identifying prospective web customers' behavior on the internet

9. Identification of prospective web customers' behavior patterns on a website
10. Identification of prospective web customers' defection patterns to and from competitors' websites as well as how they remain loyal to competitors

11. Conclusion on WM-enabled identification of prospective web customers' behavior on one or more website(s)

4.3. Profiling of existing web customers on the internet

Early on, marketers acknowledged the fact that they were incapable of satisfying all the needs and wants of all the individuals on the market. Instead, they identified commonly shared characteristics among individuals to divide them in smaller homogeneous segments which were more manageable. Web businesses face the same dilemma but the potentialities of the web have much more avenues to offer in that respect. This section examines how WM tools contribute to optimally grasp these web opportunities.

4.3.1. Segmenting existing web customers of a website

Segmentation can be based on a wide array of attributes (Tufféry, 2011). The level of segmentation starts from masses referring to global marketing strategies and narrows down to segments, niches, personas and eventually to one unique individual, referring to personalization or 1-to-1 marketing strategies (Bousquet *et al.*, 2007). Personalization is relatively difficult to implement in traditional off-line channels, *e.g.* stores, etc., but the internet offers the unique opportunity to treat each customer differently, and not as part of approximate segments or niches (Tiwana, 2001). Customers' unique needs and wants can be understood thanks to advanced forensics such as WM tools which allow to draw highly detailed profiles and use them for crafting targeted marketing responses on the spot. It is in that sense that the term segmentation should be understood here. While there will always be initial groups of customers that can be called segments, the high amount of web data available about customers allows marketers to go even more micro until the granular level of one customer, enabling the holy grail of CRM namely: one-to-one marketing. It all depends on the capacities and objectives of the web business.

The current section examines to what extent WM tools actually enable marketers to go that surgical in their web marketing efforts.

It appears that identifying existing web customers' profiles is the easiest aCRM task that can be performed by using WM tools.

All respondents agreed upon the fact that WM tools were appropriate to segment existing customers of a website. WM yields tremendous results when applied on offline data. So do these tools when applied on online data, *i.e.,* clickstream data, demographics, psychographics and the like, to focus on user-centered design. They allow building those famous "personas" or: "fictional characters created to represent the different user types within a targeted set of attributes, to be representative of a specific segment" (Jenkinson, 1997). In a web context, personas will be typically created on the basis of similar website usage, to absorb customer data in a palatable format (Pruitt & Adlin, 2006). However, one of the recurring pitfalls in correctly implementing WM methods to optimize personalization is the inability to recognize every unique user behind a PC. Users should systematically be requested to log in. Under such circumstances, log files and cookies can be manipulated with confidence, enabling cross tabulation and aggregation of the web data as well as enrichment of the customer database to draw accurate profiles of existing customers.

Various types of customers' segmentation may be done or may already exist. According to a scholar, the web visitors of a web site can be divided into 4 main generic categories:

1. Those who enter once and leave
2. Those who enter, visit and leave
3. Those who enter, visit, leave, enter, visit, buy and leave (1 or a few transactions)
4. Those who enter, navigate, buy, enter, navigate, buy (many transactions): loyal

Within these 4 categories there are sub-categories based on the type of products bought or the amount of spending and so forth. All of these (sub-)categories need to be managed differently. Two respondents indicated that a web manager should know to which category a web user belongs to as soon as (s)he enters the website to adapt the offering to the user's profile. To go even further, although some users belong to the same (sub-)category they do not behave uniformly. They should be identified and managed differently as well. This is where personalization steps in. Global classification narrows down until the micro-classification level with personalization as the ultimate step of the analysis. By focusing on categories 3 and 4, the existing web customer constitutes the grain to be managed as one segment. In the words of a scholar, this drives for instance promotion strategies:

> "A user belonging to the 4th category and to the sub-group of "frequent CD buyers" will be more responsive to a promotion on CDs than another of the 4th category, who belongs to the "infrequent apparel buyers" subgroup."

Such forensics yield patterns of probability or risk toward the fact that a specific user will behave in a specific way, *e.g.,* purchase a product, be offended by an offering. One respondent underlined that what is important is the number of specifiers on the data (how well the data is explained, categorized and labeled). The more specifiers, the more precise or surgical web managers can be about applying assumptions for a particular type of profile the user falls into. WM tools are largely reliable but one respondent indicated that:

> "Reliability of the results can be tested with other sampling methods to improve confirm or aid reliability."

The differences in results obtained with multiple sampling techniques should not be statistically significant (Malhotra, 2010). DM is less interested in the mechanics of the techniques, the statistical correctness of models and does not require assumptions to be made about the data (Lefébure & Venturini, 2001). But the outputs need to make sense and therefore big statistically significant differences from one sample to the other cause confusion and weaken the outputs.

However, despite its great usefulness, WM is subject to various methodological caveats. First, one academician advanced that there are many, overwhelming, quantities of data available which need to be exploited. But some limitations apply to avoid the "garbage-in garbage-out" pitfall. Accuracy of the WM output always depends on the data quality.

Second, two respondents indicated that the data needs also to be drawn from large and homogeneous datasets regarding the context of the study.

Third, one respondent emphasized that they need to include many specifiers. A marketing director added that the data should be used directionally when needed. The researcher might step in to interpret the data in some sense that is consistent with the research project objectives.

Fourth, a major issue refers also to the identification of the actual user behind the data. Geographic information may not be reflective of the user's location but of the ISP's location. Also, both cookies and log files track the activities of a single computer not that

of a single computer user. Even with inferential techniques such as regression or classification to identify users, the error rates incurred by these techniques make absolute user recognition unfeasible. To make up for that drawback, a database administrator suggested triangulating multiple data types from various sources to get a sense or a feeling of the web user:

> "I look at EVERYTHING. In addition to clustered and classified, I review the condition of the internet for that time period (checking router and switch health) I monitor the condition of our web servers, the service of our ISP, I look-up any news event, I check dates for opening and closing of schools, corporations and national holiday, I look at their IP addresses to get a better feel for origins, looking for trends and I sometimes tie in census department data to get an in-depth picture."

Another respondent advanced that prior to WM analysis, a good start is to review descriptive statistics provided by web analytics tools such as Google Analytics who owns information on almost anyone accessing the web. Just like descriptive statistics, *e.g.,* frequency tables, web analytics are used to explore the data prior to an inferential univariate or multivariate analysis, web analytics are used to explore web data and get a primary feeling of what they tell us. WM can then be applied to such website traffic statistics in addition to the log files. Such an approach surely yields great insight but incurs considerable amounts of time and resources. It appears that the more granular the marketer wishes to go the more work and resources are needed. These considerations need to be balanced depending on the details one seeks. In that respect, classification and clustering can be the start or the end of the WM project.

Table 6 below provides validation of the research questions and research propositions with summarized answers to the research question. RP1 and RP2 are validated. Consequently RQ1 stipulating that WM methods and techniques provide accurate profiles of existing web customers is validated.

Table 4.2

Validation of Research Question 1

RESEARCH QUESTIONS	RESEARCH PROPOSITIONS	VALIDITY OF THE RP	ANSWERS TO THE RQ
Component 1: Usage of WM applied to the profiling of existing web customers (WHO – INTERNAL DYAD 1)			
RQ 1: [WHO – INTERNAL DYAD 1a] To what extent do clustering and classification methods applied to web data (web log data, etc.) provide accurate profiles of existing web customers?	**RP1:** Web data generated by existing web customers are sufficiently detailed and accurate to provide a strong basis for the creation of precise profiles about those existing web customers.	*Valid*	*Internal and/or external web data should be large, granular, issued by logged in web users, of good quality, possibly triangulated with offline data, and analyzed with WM to do accurate segmentation of existing web customers*
	RP2: Clustering and classification methods applied to web data (web log data, search results and web pages, etc.) create homogeneous groups of existing web customers	*Valid*	

CONCLUSION –Figure shows visually the WM-enabled process of existing customers profiling as reported by respondents. Segmentation is depicted as a continuum ranging from overall segmentation with limited differentiation, to personalized segmentation with optimal differentiation adapted to one unique individual. Depending on the level of segmentation one seeks to achieve, input data range from the least to the most detailed and insightful data types, namely: IP address, ISP information, log files, cookies and other navigation files, clickstream data, web analytics providing advanced website traffic statistics, to transactional data. The latter is the most detailed and accurate data for performing customer profiling because it typically requires the customer to log in. Additional external data can be retrieved or purchased such as census data, third parties' advanced web analytics data, demographics, psychographics and other market- and consumer-related knowledge.

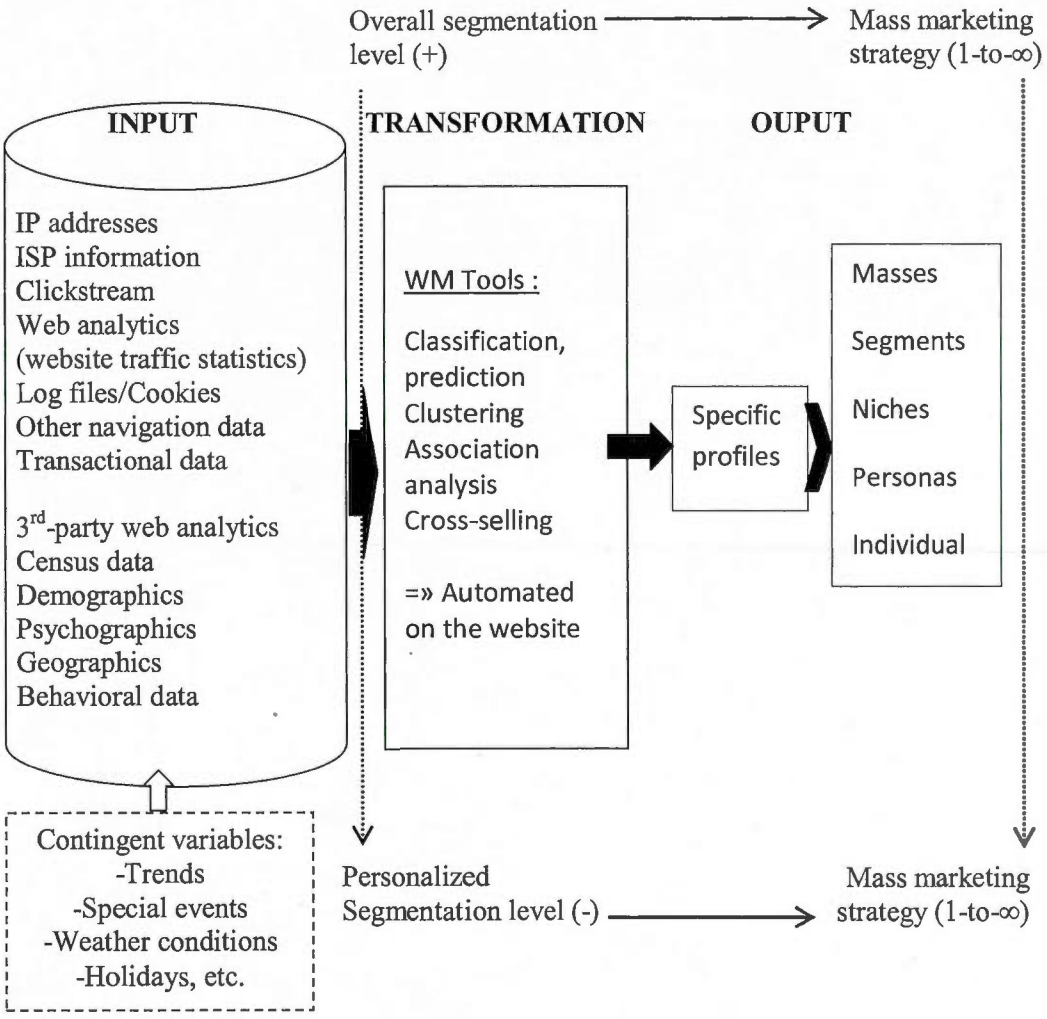

Figure 4.1 WM-enabled segmentation of existing web customers

Contingent variables such as trends, holiday periods, special events, weather conditions, etc. should be used in conjunction to internal and external data types during the WM analytical step in order to triangulate the knowledge, pinpoint congruence or contradictions, and get an in-depth picture of existing web customers. Contingencies should also help in interpreting internal and/or external data as well as WM results. The data should also be large, homogeneous regarding the context of the research and include many specifiers. WM tools are useful but the quality and quantity of the data really determine the appropriateness of the output.

WM tools have traditionally been used in an iterative fashion. They were built based on the input data and then applied to the new data. A lot of time was incurred during both of

these steps. Recent advances in the WM community have made WM tools more actionable on the spot, *i.e.,* directly on the website. They should be propelled directly on the website in order to dynamically feeding and enriching databases that encompass customers' profiles. Automation also allows classifying the customer in a predefined business-specific taxonomy which could range from masses, segments, niches, personas to the individual, depending on the segmentation strategy chosen. The most powerful of these strategies, namely the personalization strategy based on the individual allows immediate dynamic customization be it of the web page, products displayed, options proposed, etc.

4.3.2. Identification of the strategically-important existing web customers

"All customers are not equal": this is the basic principle of value-based marketing and those rules also apply on the internet. Identifying the Recency Frequency and Monetary (RFM) value of existing customer's purchases enables to determine individual Customer Lifetime Value (CLV) in Net Present Value (NPV). The acquisition and retention costs associated with one web customer should not exceed the CLV of that customer (Davis, 2005). Therefore, it is important to assess the value of individual customers and relationships in order to ensure customer profitability (Farris *et al.*, 2006). It permits to determine, for instance, the relevance of a direct e-mail marketing strategy, that is, to which customers to send to, with what content, for which expected response rate, etc. (Jeffery, 2010). *Recency* refers to the length of time since the last purchase and is used to track changes in the number of active customers (Farris *et al.*, 2006). *Frequency* refers to the pace at which a customer buys from a website and *Monetary* refers to the value of the purchase. RFM is part of the CLV. The CLV formula is as follows:

Equation 4.1 CLV formula

Minus AC refers to the Acquisition Cost that needs to be deducted from the income generated by the customer; is the margin produced by the customer in each time period n, is the cost of marketing and serving the customer, p is the probability the

customer will not defect in one year; and N is the total number of years or time periods (The Greek sigma means sum) (Jeffery, 2010). The RFM represents the gross revenue generated by one customer. When the costs incurred to serve that customer are deducted from it, it helps determining the margin ( ) produced by the customer in n timeframe. Activity-Based Costing (ABC) technique is particularly well-suited to track accurately the costs incurred for one unique customer (Kimball & Ross, 2001)

Out of the 10 respondents, only 1 indicated that WM tools may not be appropriate to identify RFM accurately because the web datasets may be too small for effective DM and statistical analysis. However, another respondent who usually performs DM on offline data emphasized that these techniques work well on online data, it all depends on how the data is collected. As for offline data warehousing, the preference should go towards non-aggregated data, granular or atomic data since these cannot be broken down any further and best describe a given action (Kimball & Ross, 2001). An academician indicated that these techniques enable to make highly targeted offers to a visitor with a minimum of information about that visitor, *e.g.,* historical data, transactional data, product information (recreational, informational), options chosen, and to see how these elements correlate to other customers' data. He explains:

> "With the first click on a website, it is possible to predict the development of a task, that is, if the customer is to become loyal (in terms of business rules) or not and consequently adapt the website step-by-step accordingly."

Definitions of loyalty and thresholds for customer profitability as well as business rules are specific to a particular business but one respondent indicated that as a rule of thumb it takes a frequency of 15 visits on a website for a prospect to make a first purchase or any desirable action, and become a customer. Three respondents indicated that both classification and clustering provide indications as to the value and frequency of purchase and can determine if a customer buys often and in which quantity. These models categorize visitors as soon as they enter the website. An academician emphasized that:

> "RFM data allow building predictive models to foresee individuals' actions."

Prediction and classification techniques may therefore also reasonably well estimate the p value of the CLV equation, namely the "probability that the customer will not defect in a year". Based on that knowledge, a customer's CLV can be computed instantaneously on

the website in terms of NPV, to classify the customer into a given category, predict his/her actions, determine which marketing actions should be applicable to him/her and provide him/her with a personalized content. A web business manager suggested that in that respect applying DM to web data provides richer information bases than pure offline data which is too static.

Another respondent added that WM methods are also richer than web analytics tools alone. The latter only provide descriptive indications on pages viewed, frequency of pages viewed and other information. It is not possible to identify precisely transactions and determine taxonomies of web users with simple statistics. Customization is even less feasible. The predictive scope of WM is a powerful asset. Offline data and web analytics act as complements or substitutive sources for data useful in the computation of the CLV but WM remains central an element of the analytical process.

One respondent supported a research axis which will be developed further. Namely that since classification is used to develop the best model to predict the future behavior of a visitor based on his/her past and current profile, "the interest does not lie in existing customers but in those we would like to acquire, guide and better serve". The capability of predicting behavior enables to better serve the customer. He states that:

> "It is prediction that is interesting with those techniques. We can see whether the visitor is small, medium or heavy; casual or regular; which products he buys together and in what sequence (association rules). We can do better customer management."

There are, however, a number of caveats and pitfalls when using WM. It has been evoked to choose the right data with the right usage. Nevertheless, a marketing director pointed out that the analysis is often longitudinal and cookie-based resulting in a lot of inaccuracies because cookies do not tell if it is always the same PC user. The data should only be analyzed if it comes from logged in customers. Fortunately though, this is generally the case for transactional data, *i.e.,* log-in sections require users to identify prior to any purchase on e-commerce websites.

Another respondent referred to the conceptual issue of the CLV concept. The RFM component of CLV informs whether a customer is spendthrift or feels bad about making more casual purchases, but people shop under infinite circumstances and for many

reasons, *e.g.,* special occasions, on behalf of someone else, for re-selling, etc. Nevertheless, it is difficult to predict the likelihood that someone buys something for another reason than casual shopping. An 8-year old who buys regularly a 6-pack of beer, at the corner shop, on behalf of his father, would be labeled a loyal beer consumer with an accurate CLV attached to him. This may sound awkward, but it is exactly what many marketers do every day in web contexts by associating transactional web data to individual users without grasping the depth and scope of customer's buying processes. CLV is a relative metric that cannot predict the strategic importance of customers in absolute terms because it does not integrate the underlying buying decision process of the customer. WM tools are appropriate to analyze RFM and CLV, but as in brick-and-mortar shopping contexts, the underlying dimensions of consumer behavior remain largely unknown. WM cannot go that far yet.

Table 7 below provides validation of RQ 2 and its research propositions with summarized answers to the research question. RP1 and RP2 are validated. Consequently RQ2 stipulating that WM methods and techniques enable to identify the RFM-based CLV of existing web customers to filter out the strategically important ones of them, is supported.

Table 4.3

Validation of Research Question 2

RESEARCH QUESTIONS	**RESEARCH PROPOSITIONS**	**VALIDITY OF THE RP**	**ANSWERS TO THE RQ**
Component 1: Usage of WM applied to the profiling of existing web customers (WHO – INTERNAL DYAD 1)			
RQ2: [WHO – INTERNAL DYAD 1b] To what extent do classification and regression methods applied to web data, identify the	**RP1:** Web data generated by existing web customers encompass enough information about the profit-cost, the Recency-Monetary-Frequency (RFM)-based CLV of purchases made by existing web customers, which contributes to identify	*Valid*	*Internal and/or external web data should be large, granular, issued by logged in web users, of good quality, possibly triangulated*

profit-cost ratio and the Recency-Monetary-Frequency (RFM) of purchases made by individuals, on which the Customer Lifetime Value (CLV) is based, to identify strategically important, existing web customers?	those strategically important customers		*with offline data, and analyzed with WM to determine existing web customers' RFM, CLV and strategic importance*
	RP2: Classification and prediction methods applied to web data predict the value of a given web customer to identify strategically important existing web customers.	*Valid*	

CONCLUSION Figure shows how WM contributes to identifying strategically important customers. Determining existing customers' RFM to compute their CLV requires little information that can be obtained from a variety of sources such as transactional data, cookies, log files, and other historical data and to a lesser extent web analytics data. The dataset needs however to be large, customers should log in and the data used in the analytical process should preferably be as granular as possible. This is useful for conducting further dimensional modeling such as OLAP (mining)[30] (Kimball & Ross, 2001), which is another major component of BI along with DM and relational reporting (Pareek, 2007). If supplemented with additional useful offline data, CLV can be computed automatically with WM on the spot, in order to determine a customers' specific category, his/her likely behavior in the future and craft relevant e-marketing strategies to maximize returns. Business rules can be derived from aggregating huge datasets of such profiles to be used for further model fine-tuning, personalization and customization purposes.

[30] OnLine Analytical Processing (OLAP) aims at analyzing instantaneously information along several axes (dimensions) in order to produce analytical reports such as marketing, financial budgeting, forecasting, etc. OLAP mining integrates on-line analytical processing (OLAP) with data mining so that mining can be performed in different portions of databases or data warehouses and at different levels of abstraction at user's finger tips (Codd, 1993).

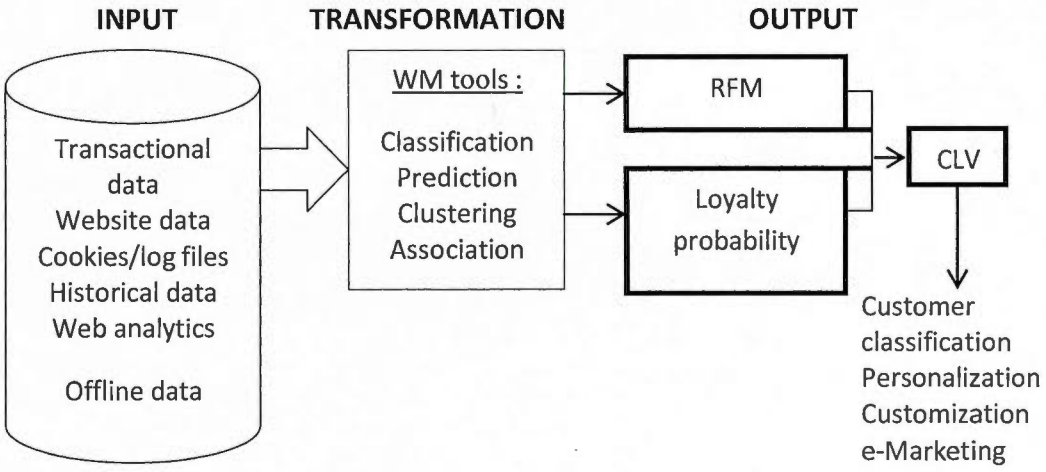

Figure 4.2 WM-enabled determination of strategically important customers

4.3.3. Identifying existing web customers' loyalty and defection statuses from a website

Increasing customer retention through increased customer loyalty is important because keeping existing customers is less expensive than acquiring new ones (Farris *et al.*, 2010). Increased loyalty is also strongly linked to an increased profitability since a 5% improvement in customer retention can cause an increase in profitability between 25% and 85% in NPV, depending upon the industry (Sasser & Reichheld, 1990). In fact, only during later transactions, when the cost of serving loyal customers falls, do relationships generate profits (Buchanan & Gilles, 1990). The concept of e-loyalty is even more crucial to web companies since the Internet is a nearly perfect market where information is instantaneous and buyers can compare the offerings of sellers worldwide (Kuttner, 1998). The ultimate result of reduction in information asymmetries between sellers and buyers is fierce price competition and vanishing brand loyalty (Srinivasan *et al.*, 2002). This is why it is of utmost importance that web businesses identify as quickly as possible the current status of each customer in order to determine whether e-marketing strategies are effective and whether the business remains profitable. Identifying antecedents, *i.e.*, company-specific e-business factors that appear to impact loyalty, and consequences of loyalty, is an additional step that businesses may undertake to boost loyalty and lower attrition (Srinivasan *et al.*, 2002).

It is widely acknowledged that both concepts of loyalty and defection can be expressed in numerous ways. Organizations use company-specific variables to express both concepts, rendering benchmarking attempts void even across organizations of the same industry. Even worse, both concepts are measured and determined by variables that may be specific to an organization's business model, industry, market or a combination of all of these.

For our purpose "loyalty" integrates both an attitudinal and a behavioral dimension, as suggested by the literature (Engel & Blackwell, 1982; Jacoby, 1971; Assael, 1992; Keller, 1993; Gremler, 1995). It is defined as a customer's favorable attitude toward the e-retailer that results in repeat buying behavior (Srinivasan *et al.,* 2002). The frequency of purchase may be very variable and even volatile from the perspective of one website to another but as long as a customer is retained (s)he is considered to be loyal whatever his/her degree of loyalty. Even though a customer shifts partially his/her account from the website to a competitor's website, reduces his/her share of the wallet with the website, as long as (s)he repeats even partial purchases (s)he remains loyal. As a matter of fact, for the purpose of this research, "defection" does not cover the concept of partial customer churn which is still a form of loyalty since the customer remains attached to the business (Buckinx & Van den Poel, 2005). Rather, defection is defined as the customer's total switch to another website in a voluntary or involuntary fashion (Van den Poel & Burez, 2006). Loyalty and defection are thus two poles of one continuum on which customers can be placed to determine their loyalty status.

These semantic and semiotic variations do not affect the power of WM tools to identify the loyalty vs. defection status of an existing web customer but there are limitations and caveats. This section examines to what extent WM tools allow determining the situation or status of an individual at a specific point in time. Either the customer is loyal *i.e.,* purchasing in a more or less frequent pattern, or has defected and ceased therefore all transactional activities.

One respondent indicated that instead of using WM, merely reviewing bounce rates may be a wiser approach. They show the percentage of visitors who entered the website and bounced, *i.e.,* they left the website, instead of viewing other pages within the same site. It appears that, as long as we are concerned with the status of an individual, WM may be

too sophisticated a tool because it may be better suited to Root Cause Analyses (RCA)[31] of loyalty/defection. Also, interpretation of the bounce rate should be conducted in light of the website objectives and of the definition of conversion as stipulated by management. A high bounce rate is not always a sign of poor loyalty/high defection. In an e-commerce context, bounce rates should be interpreted in correlation with purchase conversion rates. As a rule of thumb, it is hard to get a bounce rate under 20%, anything higher than 35 % is cause for concern and above 50% it is worrying (Kaushik, 2009).

Another respondent indicated that identifying loyalty/defection statuses is not as easily done even with WM tools because each customer has a specific purchase cycle.

> "A regular customer of Amazon may not visit Amazon's website for a year but this does not mean he left, he may buy again in one year or more."

The risk would be to consider this customer as having defected and when he comes back, the business might view him as a new customer and tailor its e-marketing strategies accordingly to him as a newcomer. Each customer has a different loyalty rating. To avoid such pitfalls a better strategy would be to use the visitor's IP address. By using the IP address, it may be possible to identify the client that connects to the host server and if that IP address visits the website once, stays a few seconds, leaves and never comes back, it may be highly probable that this visitor had a bad experience or did not have the money to buy and left. Such a user would be labeled as defective. However, if that IP address returns this may not be indicative of the visitor's loyalty either since it does not tell us whether it is the same unique user returning. The computer might be used by other people and here lies the biggest limitation of IP address usage. Another approach is to use cookies which can now be adjusted to the fact that people delete them.

Most other respondents reported that WM do in fact enable to identify existing web customers' loyalty or defection statuses. Two respondents indicated that classification techniques, when detailed enough, are very useful for that purpose. Besides, they indicate the extent to which a customer might buy, return or defect. The intrinsic predictive goal of classification determines not only the status of one customer as being either loyal or defective but also whether this status may evolve in the future. Is the customer continuing to buy on a more or less frequent basis (loyalty) or stopping all transactions (defection)?

---

[31] RCA identifies the root causes of problems and further discovers the points of leverage where patterns of problematic behaviours originate and can be changed.

Again, this always depends on how loyalty is defined. An additional element would be to consider loyalty in a specific timeframe. That is, arbitrarily determined by the business based on past research or business rules, as suggested by one respondent:

> "We observe users' behaviors: as soon as a customer who used to be loyal in a specific timeframe stops buying or visiting (*i.e.,* stops displaying the company-specific loyalty behavior), we might deduce he/she left."

Hence, by dividing the total sample of existing customers randomly into two subsamples, one of these subsamples (analysis sample) is used to determine which variables have caused defection or loyalty (discriminant function); the second (holdout sample) is used to test the discriminant function (Hair *et al.*, 2006), it is then possible to take another portion of customers for which we seek to know the current status and compare their profiles with those of the training set and the variables of interest and infer on their current status on the website. Such a cross-validation approach is not absolutely accurate as a respondent indicated it. However, classification provides a good basis for further study and follow-up. Three respondents indicated that when classification is combined to clustering, the power of classification is enhanced. A database administrator puts it like this:

> "Clustering provides two or more axes from which to deduce likelihood of status. Combining both techniques increases the assurance that a user viewing a particular object at a particular time will behave according to a certain pattern. Such knowledge drives surgical or customized marketing as well as website customization, because it drives design efforts to represent or accommodate a greater percentage of predicted behavior most of the time."

All respondents indicated that browsing history data are particularly suited for such an approach and may even provide additional insight into what caused defection. A business can thus determine two groups of users: those who are loyal and those who defected according to the business' definition of loyalty. An academician highlighted that cross-selling such data of one user with those of other customers who are similar to him/her may leverage insight into whether that user might engage with the website or not. From that point, it is possible to review details of users in order to understand why some left and why others stayed: do both groups share similar characteristics within groups?

And/or dissimilar attributes between them? Etc. An academician indicated that the reasons may arise from such variables as profiles or interests:

> "We try to explain defection, before that you must trace it. That's why it is necessary to see profiles to see what may have caused defection or loyalty."

Two respondents indicated that even the earliest activities of a customer on a website are highly indicative of the overall engagement and loyalty as in the words of a marketing director:

> "If we identify a group of customers who come to a website and leave from the home page every time and/or spend only a few seconds on the website, we can say their web experience is not great, they are not finding what they are looking for easily and so their web portion of the overall customer experience will suffer, translating into a high chance of defection."

This refers to the fundamental notion of customer satisfaction as a primary predictor of customer loyalty. The level of satisfaction depends on prior expectations of overall quality compared to the actual performance received (Parasuraman *et al.,* 1985). If the recent experience exceeds prior expectations, customer satisfaction is likely to be high (Gronröos, 1994). One academician pinpointed that statement saying that the level of satisfaction can also be a clue as to the level of loyalty, although this might be more difficult to determine in a purely web context.

He added that recent advances in the DM field extended the DM tools so that even partially known data such as censored data, *e.g.,* missing data, can be mined by using survival trees or other types of survival/failure time analysis. Web data are often of the censored type and the satisfaction as well as the loyalty status can also be labeled as such. A censored data is one for which the value of the measurement is not known or lies on an interval between two values or below/above one of the value. Since loyalty and defection represent two extremes on a continuum, the challenge is to find where each individual lies on that scale. Likewise, another respondent indicated that the techniques cited are not relevant and that it may be better to use hazard regression models. The objective it to predict a risk or hazard which is defection and the input variables to predict that status are company specific. Survival and regression models both were previously identified as of

the prediction method type[32], as they inform on the probability of a given risk for a business, such as attrition. No back office work and no delays between actual activities and output of the analysis are incurred.

At any rate, the activities of a visitor must be identified and analyzed as quickly as possible (in real-time would be a must-have) and "those preferences must be identified on time and the appropriate modalities that accommodate the preferences should be applied on an equally timely manner, before losing the customization opportunity."

One respondent indicated that in order to conduct proper WM analyses the web data needs also to be well cleaned up and properly used for analysis:

> "It is crucial to have the right web data in the format easy to analyze and above all which corresponds to the definition of loyalty and defection of the company. When resources are available in conjunction with other data sources, then WM is definitely appropriate."

This confirms that the "resource finding and retrieving" as well as the "information selection and preprocessing" phases are both of utmost importance as reported by Kosala and Blockeel (2000) and Zhang and Segall (2008) in their 5-stage WM process.

Table 8 below provides validation of RQ 3 and its research propositions with summarized answers to the research question. RP1 and RP2 are validated. Consequently RQ 3 stipulating that WM methods and techniques enable to identify existing web customers' loyalty or defection statuses.

[32] The added value of such a technique is that it allows to determine the status of a customer by calculating a score in real-time even directly on the website when the user is visiting it.

Table 4.4

Validation of Research Question 3

RESEARCH QUESTIONS	RESEARCH PROPOSITIONS	VALIDITY OF THE RP	ANSWERS TO THE RQ
Component 1: Usage of WM applied to the profiling of existing web customers (WHO – INTERNAL DYAD 1)			
RQ3: [WHO – INTERNAL DYAD 1c] To what extent do clustering and classification methods identify existing web customers' loyalty or defection statuses?	**RP1:** Web data generated by existing web customers indicate whether an existing web customer is loyal to a given business or defects from that business.	*Valid*	*Internal and/or external web data should be large, granular, issued by logged in web users, of good quality, possibly triangulated with offline data, and analyzed with WM to identify existing web customers' loyalty/defecti on statuses*
	RP2: Clustering and classification methods applied to web data predict membership of an individual to the loyal or defecting customers group.	*Valid*	

CONCLUSION - Although useful and highly effective, traditional WM tools that imply back office work and long delays between time of computation and time of output launching, seem no longer relevant anymore. As shown in Figure, the speed of pace on the internet calls for automated methods that compute in real-time the loyalty or defection status of a given customer based on his/her history of browsing activities, interests, or any other variable that defines loyalty or defection for a given web business. Such methods also compute in real-time the loyalty or defection risk and possibly even their extent

depending once again on how loyalty is defined. These methods are however of limited use if the data is not well-retrieved, well-cleansed and heterogeneous enough to integrate a wide range of loyalty predictors and facets. Businesses need thus to provide particular care to the data-preprocessing and need to integrate extended WM tools to obtain existing customers' current loyalty or defection status. They may use traditional tools as useful complements to flesh out the analyses by providing additional insight into the characteristics of those who leave/stay, opportunities for customization, cross-selling purposes and surgical marketing strategies.

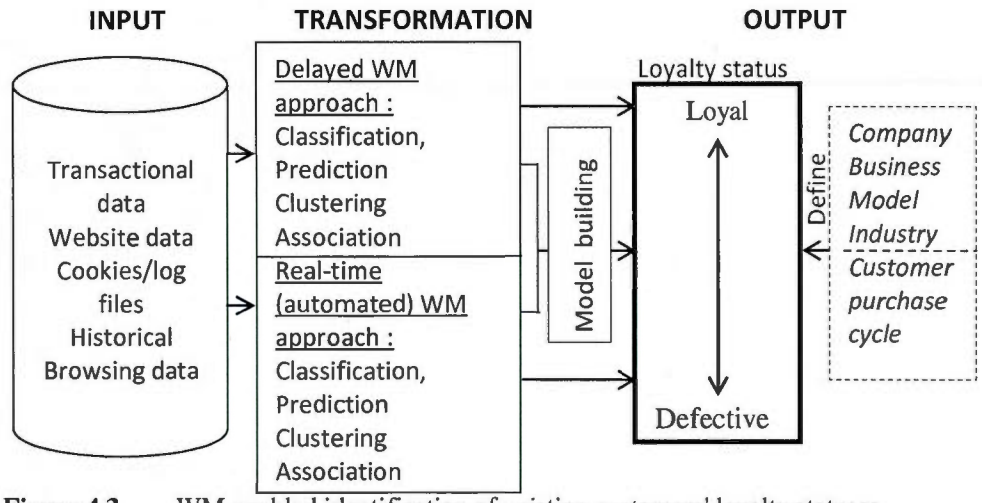

Figure 4.3 WM-enabled identification of existing customers' loyalty statuses

4.3.4. Conclusion on WM-enabled profiling of existing web customers on a website

The business has to choose data with a certain level of detailing in accordance with the desired level of segmentation which may range from overall segmentation to personalized segmentation. In Figure 13, internal data have been classified in ascending order of precision, *i.e.* from least precise (IP address) to most precise (transactional data). It has been suggested to integrate offline data originating from Business Intelligence systems as well. External data act as complements to internal data and useful correlations and cross-tabulations can be made between both data types in order to get a better initial picture of a

given customer, *e.g.,* cross tabulating a customer's longitudinal browsing data for a 1-year time period with his/her demographic characteristics.

The data needs to be objectively good, because the quality of the WM output will always depend on the quality of the input data fed into analytical processes. It has been suggested to collect preferably granular data (low level of detail) with many specifiers. Also, the datasets need to be large and homogeneous regarding the research context. The data should on the other hand originate from as heterogeneous sources as possible in order to increase accuracy checking and provide multiple perspectives of an individual. For further accuracy, customers should be required to log in in order to attach with confidence the data to specific individuals. Besides, the data should be well-retrieved, selected, cleansed and pre-processed prior to the analytical stage.

In order to get an even more precise estimation of users, other contingent data with an effect on internal data may be considered. This is in line with what is suggested in the literature. By doing so, web managers can view market dynamics from various perspectives and arrive at "triangulated" strategies and solutions (Farris *et al.,* 2006). Additionally, with multiple metrics and data, marketers can use each as a check on the others. In this way, they can maximize the accuracy of their knowledge (Katz & Hauser, 1998; Farris *et al.,* 2006). However, the more details a business seeks to obtain, in order to go personal with its customers, the more time and resources it will cost. The trade-off has to be made, ideally on basis of a cost-benefit analysis.

Once data have been well selected, retrieved, pre-processed, cleansed and triangulated, they can be used as such for reporting, budgeting or other business-related activities. In our case, they will be more valuable input for advanced analytics. This is where WM steps in. It has been determined that WM can be performed by means of two approaches: the slow-and-steady approach which implies a relatively long delay between data collection and model generation because it needs human intervention. This approach seems out-dated as current hardware and technology advances propelled the real-time WM approach. Data are entered into the WM system and a model is generated or updated on the spot without extensive delays. It seems however that the slow-and-steady approach is necessary to build the model the first time and the real-time approach may be preferred afterwards once the model is developed in order to automatically process and store new

data streams and produce the desired outputs. It is not clear though to what extent each and every WM technique enables automatic creation of models on the spot to bypass the thorough slow-and-steady approach. Additional research should estimate to what extent WM methods truly enable automatic model creation and implementation directly on the website and without human interaction. Anyway, the 4 types of techniques can be performed in one way or the other, depending first and foremost on the business' level of technology acceptance, objectives, procedures and resources availability.

Either approach allows determining customer loyalty by identifying the loyalty status. It also determines customer profitability by determining some elements of the CLV equation, namely the RFM and loyalty probability of a customer. Eventually, the specific profile of a customer can be drawn for subsequent segmentation purposes. Both the loyalty status and the CLV may also constitute core variables to draw segments. The figure shows that both profiles and loyalty status have an impact on the CLV entity. Although not studied in the present piece of work, respondents indicated that profiles and level of loyalty displayed by a customer have an obvious impact on his/her future profitability and WM is also used to determine to what extent both independent sets of variables influence the dependent variable of a customer's NPV. Loyalty, profitability and segmentation enable then a company to draw very detailed customer profiles and issue dynamic response frameworks directly on the website. Once and every time a customer with a specific profile enters the website, content, layout and structure can be rearranged to reach company-specific objectives regarding that given customer. The example below is a simulated real-life case developed by the researcher based on the findings relating to the first theme of profiling existing web customers of a website. It summarizes the potentialities of WM as identified so far:

> *GameQuest.com is an online video game website which supports new and browser games portals, multiplayer and single player gaming platforms while offering a plethora of goods and services to gamers such as technical assistance, chat rooms, forums, FAQs, gamers' communities, customized clothing, tokens and items, etc. Users can play for free to limited versions of the games or open an account and pay variable fees for variable levels of full versions of the games. One video player has opened a basic membership account and visits GameQuest on average every 2 days which corresponds, in terms of the website's definition*

of loyalty, to a medium-high loyalty status. CLV may be calculated based on past purchase cycles providing RFM and future loyalty probability. A regression analysis is performed to compute the future loyalty probability. The resulting CLV is estimated to be lower than the business' average Cost of Acquisition for customers. A sound business rule suggests that a CLV should always be superior to the AC for that given customer in order for a business to remain profitable (Farris et al., 2006).

The customer's data comprised of browsing history, clickstream data, purchase history, log files and IP address are used as input to perform WM classification analyses such as neural networks and discriminant analysis. This compares the user's profiles to other similar users in order to assign a score to the customer (scoring) as to his current loyalty status and also probability of playing at higher frequencies and moving to a superior gaming account. The customer's score is low and predictive results indicate he is not going to increase gaming frequencies, nor may there be any up-selling opportunity. Although he was classified as a medium-high loyal customer, WM analyses took other variables into account to determine exactly the user's current loyalty status. Additional data elements enabled to get information as to his zip code, some psychographics, favourite games as being heroic-fantasy RPG games, etc.

A Clustering analyses with other users' specific profiles provided 150 niches. Based on the data it had on hand, the company chose the niche-based targeting level and divides thus users into niches which are smaller than segments without reaching ultimate personalization. This strategy is thus halfway between granular and segment marketing. For each niche, specific marketing strategies are crafted. Clustering revealed that the user belongs to the "bargain hunting maverick gamer" niche. These players typically subscribe to the lowest account available while seeking to play arcade and free games but above all top-level RPG games usually in test mode and for long periods of time but without moving to suggested upscale accounts. Besides, they overload servers, request frequent technical assistance from staff, bother other loyal players and even sometimes commit fraud (also detected thanks to WM methods) to earn additional points that need to be bought otherwise.

Based on that knowledge, the business can adjust many marketing variables. Regarding direct marketing for instance, the business adjusts itself to the user by

not sending him premium offers. The business can also implement dynamic response frameworks in the website architecture to automatically adapt to the user by restricting access to online technical assistance and to specific posts on the forum: pricing premium accounts at lower fares for determined periods of time: offering free gifts in exchange of membership acceptance; displaying low-cost arcade games on top and high-scale RPG games at the bottom; gearing towards other similar games thanks to recommendation systems. A lot more variables are adjusted likewise.

The first theme corresponding to the first meta-objective of profiling existing web customers of a website, as identified in Xu and Walton's (2005) adjusted framework, can be fulfilled by using WM. Traditional descriptive statistics are useful complementary tools in that respect. In the specific context of the web, WM provides a richer view of the customer as well as personalization features that mere database marketing or web analytics cannot provide.

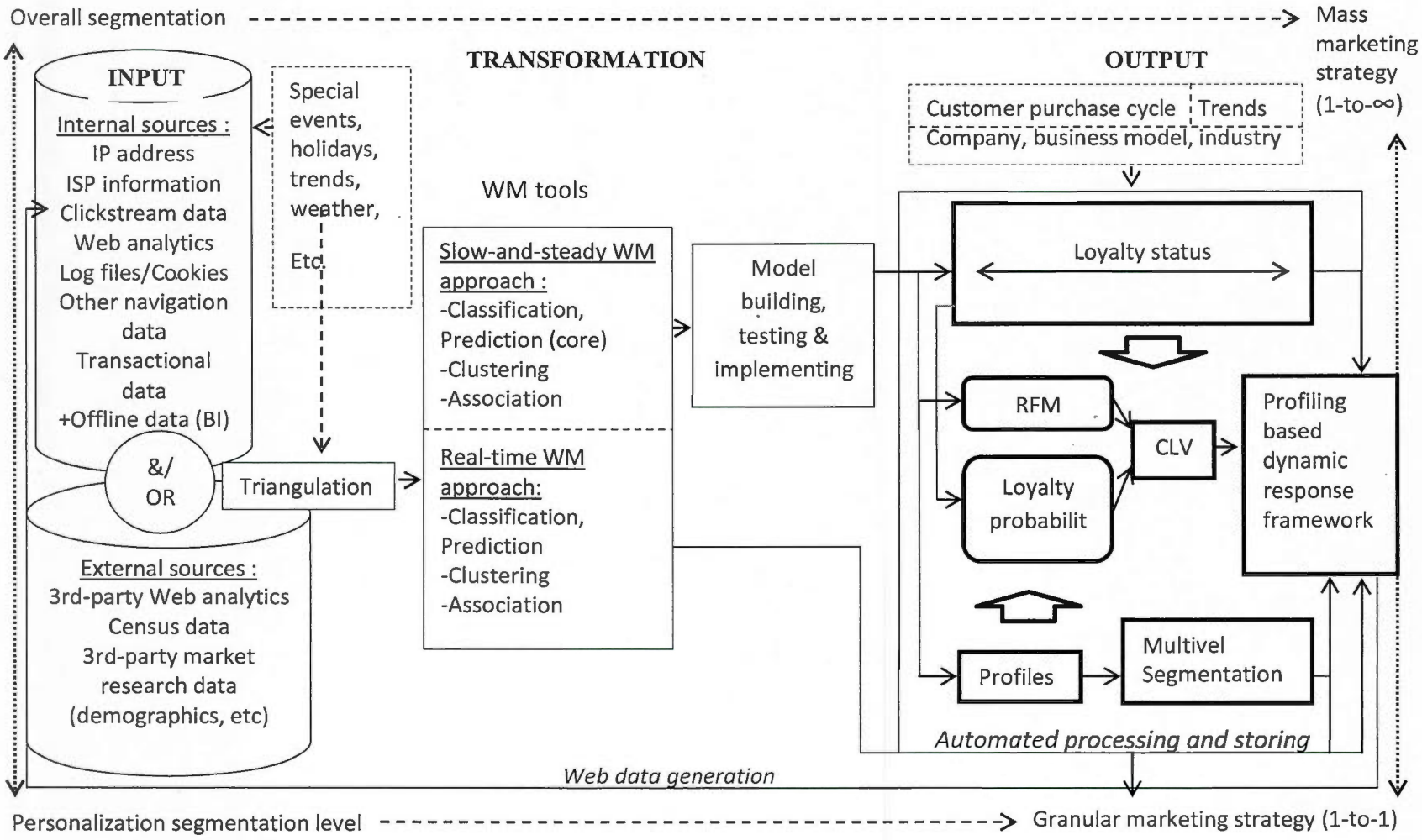

Figure 4.4 WM-enabled profiling of existing web customers on a website.

4.4. Identifying existing web customers' behaviors on the internet

4.4.1. Identifying existing web customers' behavior on a website

The profile, the RFM or the loyalty status of an existing web customer, constitute descriptive and static data that may be very useful in order to trace the dynamic moves, paths of the so-called "web user behavior". The more diversified the input data, the richer the view of user behavior.

Most respondents agreed on saying WM tools enable identification of existing web customers' behaviors on a website but recognized that there were several limitations. Only one respondent indicated that there are no caveats or pitfalls.

WM tools are not inappropriate per se but 2 respondents indicated that the WM methods cited such as classification, regression, etc. may be too elaborated. Instead, descriptive methods such as path analysis, frequency tables or clickstream analysis may be more suitable to draw behavioral patterns. An academician puts it like this:

> "Clickstream analysis is powerful because it tracks the actual behavior of people, it cannot lie. Other listed techniques are less important for that."

The respondents who admitted that WM tools were appropriate to determine web users' behaviors indicated that by using individual axes such as classification and additional axes such as association web administrators can identify behaviors on the training set and predict behaviors of new users hence identifying web customers' behavior automatically. Another respondent emphasized that such previously learned rules and profiles can then be used as a basis for identifying future behaviors. One respondent went further by stating that identifying behaviors offers in turn the start of customer profiling and the level of detailing is up to the company. An academician indicated that:

> "Extrapolation and cross-tabulating the data with that of other users it is possible to get an idea of future behaviors."

One respondent indicated that traditional WM approaches such as discriminant analysis for classification or logistic regression for prediction, etc. may be somewhat useless in the current web context. Now, such new techniques as neural networks are preferred to

automatically identify users' behavior on websites and extract meaning out of it. This is done in a much more automated, fast and easy fashion than other more classical tools.

However, WM, whatever the recency of the technique used, may be more or less useful under a certain number of conditions. A marketing director stated that the web data needs to be properly extracted and right tools need to be attached to the website for that purpose. She further added that in order to avoid the "unique user issue" (tracking the behavior of one unique user and not one unique computer), analysis is best done if it is based on data collected on websites where web customers are required to log in before performing any activity. This remains very rare since customers are usually only required to register when engaging into a commercial transaction and not before. They may also be required to register the first time they visit the website and could then be automatically logged in each and every subsequent time they visit the website (ex: YouTube).

Two respondents also reminded that e-behavior is all about browsing history and purchase history. In a web context the term behavior refers in fact to the "browsing behavior", *i.e.,* time spent on a web page, clicks on a specific web link, etc. and there is no psychology involved. An academician stated that:

> "There is no other form of behavior that can be studied on the internet up to now than browsing strategy. On a website you only have an interface where you can do all sorts of things depending on how you wish to use the internet. Web administrators can only obtain browsing statistics and RFM data."

It is clear that consumer behavior, which could be termed "e-consumer behavior" vis-à-vis websites cannot be studied as such. However, Mobasher and Nasraoui (2011) highlighted that user-provided content, especially written may by much helpful in that respect. In fact, the web 2.0 propelled user interaction and interactivity with the web to make them more active and exclusive producers of web content in the form of blogs, micro-blogs, forums, discussion groups, social networks or other live chat sessions. The availability of such big and unstructured data permits sentiment analysis and opinion mining which reveal some aspects of some specific variables of customer behavior especially regarding emotions or feelings towards a product, organization and brand. Although still in its infancy, such WM approaches, much more focused on the content (WCM) provided by users instead of their web usage (WUM), may reveal additional

layers of their personalities and lifestyles in order to better grasp their behavior not only online but also in other contexts.

Regarding the hardware, new technologies may also be developed such as iris-recognition devices integrated onto computers to conduct eye-monitoring studies. Dilatation of the pupils may reveal emotional, physical or even feelings and attitudes. Behaviors could be inferred from such data. Beyond the obvious technological issues such studies may pose, they also cause ethical dilemma. Also, it may not be that clear whether e-consumer behavior constitutes a strict transposition of a consumer's behavior in a non-web context into a web context. Some differences may arise between both. People may be less subject to social desirability or social control when evolving in a web context, as compared to their behavior in real-life. The research on that field is still in its early infancy.

Consequently, today it appears that the more information on browsing history and purchase history, the better the acquisition of existing customers' behaviors. But two respondents added that the different data must converge, be numerous and diversified in nature (clicks, behaviors, etc.). A certain level of homogeneity in content however (*i.e.* demographic homogeneity among the user for instance) may remarkably facilitate the identification of clear patterns of behavior. A marketing research director shaded that statement by recalling that:

> "Usually data sets are not large and homogeneous enough to generate reliable behavior patterns and as a matter of fact predict behavioral patterns."

Two respondents suggested conducting traditional surveys among customers to get a more comprehensive insight of the customers by also identifying their consumer behavior (attitudes, perceptions, emotions, etc.), *e.g.,* perceptions of an online credit application service, etc. Actual customer behavior is tracked online in the form of browsing and purchase history and may then be combined to survey results in order to perform a triangulation. An academician explains it as follows:

> "If we see that 90% of what people tell in surveys corresponds to what they actually do, it may be enough to ask them a priori, in order to predict their behaviors. However, if 90% of what people report does not correspond to what they really do, surveys turn to be useless and burden the website."

Surveys also incur error rates and various biases, which constitute precisely the reasons why WM tools are used instead. The trade-off between using either survey data, web browsing data or both needs to be carefully balanced in light of the business' industry, objectives, culture, customer type and business model.

Table 9 below provides validation of RQ 4 and its research propositions with summarized answers to the research question. RP1is validated but RP2, isn't. Consequently RQ 3 stipulating that WM methods and techniques enable to identify existing web customers' behaviour on the internet is partially-validated.

Table 4.5

Validation of Research Question 4

RESEARCH QUESTIONS	RESEARCH PROPOSITIONS	VALIDITY OF THE RP	ANSWERS TO THE RQ
Component 2: Usage of WM applied to the identification of existing customers' patterns on the web (HOW – INTERNAL DYAD 2)			
RQ4: [HOW – INTERNAL DYAD 2a] To what extent do clustering, association analysis, classification and prediction methods applied to web data identify existing web customers' behavior on the internet?	**RP1:** Web data generated by existing web customers highlight the particular browsing behavior of existing web customers when they are navigating on the internet.	*Valid*	*Internal and external web data should be large, granular, issued by logged in web users, of good quality, and analyzed with WM to identify existing web customers' browsing behavior. Traditional market research*
	RP2: Clustering, association analysis, classification and prediction methods applied to web data provide descriptive and predictive modeling of web customers' consumer behavior (personality, motivation, lifestyle, perception,	*Not valid*	

	attitudes, emotions, satisfaction etc.), on the internet.		*outputs and analyses should complement WM to detect existing web customers' consumer behavior*

CONCLUSION – Behavior in a web context refers strictly to browsing and purchase behavior, a very limited insight of the broader "e-customer behavior", which cannot be studied by WM tools only. Figure is a summary of respondents' opinions on the subject. It shows that descriptive methods such as frequency tables, path analysis or cross tabulation seem as appropriate as more sophisticated WM tools. They also serve as inputs for WM. Among those WM tools, it appears that the more automated techniques tend to be faster, easier to use and offer real-time insight into existing web customers' behavioral patterns. Consequently, descriptive methods should be used to get an initial insight of behaviors. Then WM provides more sophisticated explanations of behavior such as sequential patterns for instance. Nowadays, automated processing in real-time is preferred to the more traditional model building and deployment approach. In addition, traditional surveys should be used as complements to these tools, in a triangulation perspective, in order to get additional insight into customers' psychological dimensions on cognitive, affective and conative levels. The input web data should ideally be converging, large, homogeneous enough as well as properly extracted and ideally resulting from customers who log in. This will enable web managers to anticipate specific offer for each different customer, enhance the web environment according to each unique web user through customization or even measure the effects of a web marketing campaign.

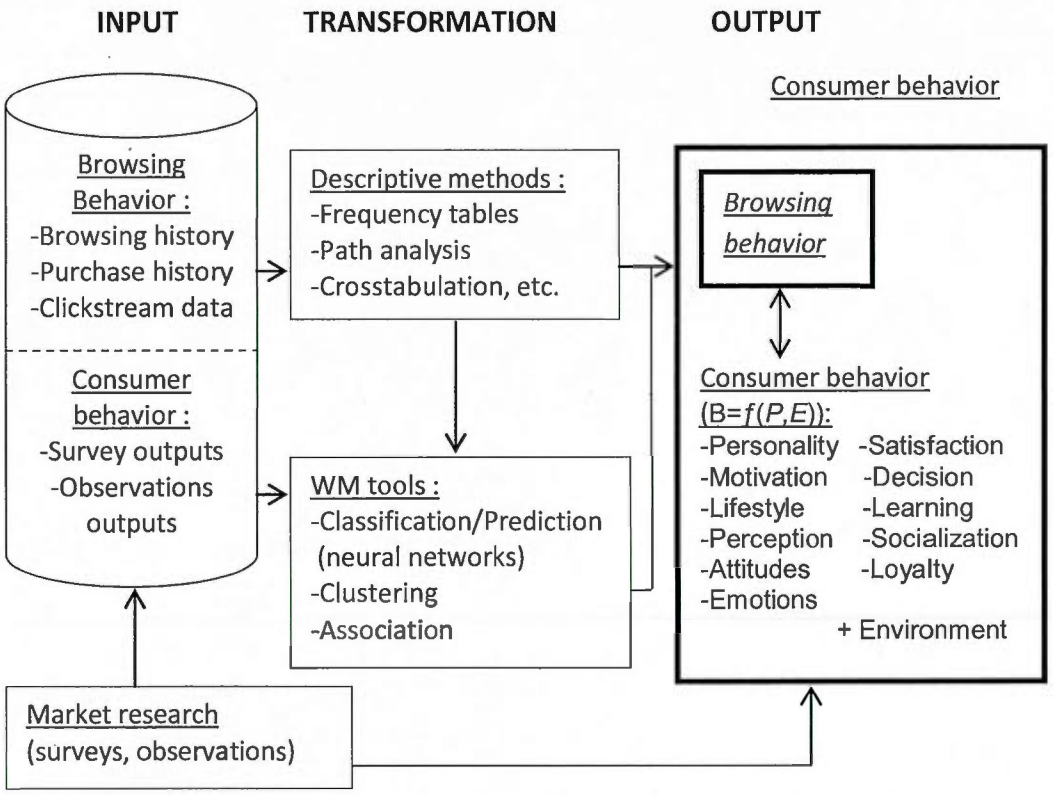

Figure 4.5 WM-enabled identification of existing web customers' behaviours.

4.4.2. Identifying how existing web customers develop satisfaction and loyalty on a website

It has been seen previously that the level of satisfaction is a strong indicator of a customer's level of loyalty. In order to increase satisfaction and loyalty, a specific branch of marketing called loyalty marketing relies on word-of-mouth and advertising to develop both satisfaction and loyalty to increase profitability (Carrol & Reichheld, 1992). Incentive and inducement programs such as frequent flyers, Card Linked Offers, prizes, premiums or other point-based reward programs (Molony, 2006) are implemented to make customers develop a certain level of satisfaction which should in turn trigger loyalty development. Loyal and satisfied customers are then expected to do referral and advocacy through (e-)word of mouth i.e., viral marketing (Reichheld, 1996). In a web context, viral marketing is particularly powerful because consumer-based eWOM is considered to be

more trustworthy than company-controlled communications and mainstream media[33]. The former is perceived as being a more reliable and credible promotional source than the latters, at least in web contexts (Henning-Thurau & Walsh, 2004).

Provided the business has an efficient product/service marketing strategy, a well-implemented loyalty business model not only contributes extensively to branding but also retains current customers especially core customers, while also acquiring new ones (Stuart, 2007). Reichheld (1996) dubbed it the "power of extension" of loyalty marketing. It is thus of utmost importance to understand how current customers develop loyalty and how the business might take advantage of that knowledge to increase loyalty among existing customers and possibly lock in prospects as well.

Incentive and inducement programs are specifically aimed at tracking the evolution, hence the development, of customers' satisfaction and loyalty over time. They may be used in online contexts as well. In fact, most consumer reward programs have been extended online into a multichannel setting to accommodate customers. But what about online businesses (be they pure players or multichannel activists) wishing to grasp how their customers develop satisfaction and loyalty? Incentive and inducement programs generate useful web data that can be mined extensively to discover hidden loyalty patterns and understand how existing customers come to patronize a specific web business.

Three respondents recalled that there are many definitions of loyalty and satisfaction. Out of the 10 respondents, 9 indicated that it is easy to determine how loyalty is built by using WM tools. An academician reported that:

> "By using explanatory variables such as behavior, profile, etc. as input we can explain the dependent variable which is loyalty and that can be seen as a continuum (loyal to defective). All behaviors explain how we become loyal or disloyal."

[33] A survey from InSites Consulting group revealed that 38% of social networking users say posts from other consumers are most credible, followed by posts from brands themselves (32%), the media (7%) and marketers (3%).

There are enough points of reference to define users and web managers need to pinpoint those identifiers that best explain satisfaction and loyalty altogether. Three other respondents indicated usage of historical usage data. A web business manager outlined the following:

> "Customers develop loyalty and satisfaction if the business presents new entertaining or necessary products through engaging user experience and value proposition. WM tools allow determining how an enhanced value proposition, for example, increases loyalty/satisfaction by means of identifying correlations between both of these elements. They also allow reaching these important tasks, but this is another application."

WM provides detailed analyses over time of how users develop loyalty by giving them a loyalty rating for instance. Loyalty may also be estimated thanks to the intensity of usage. A visitor who uses the website frequently may indicate he is more loyal than another who does not visit the website anymore. Only one respondent indicated that this may not be known without surveys and surveys are actually distasteful to users especially in a web context. So, according to that respondent one can only know by looking whether the customer comes back at all as WM methods are good for presentation/description and prediction but not for determining such complicated patterns as loyalty and satisfaction development. A database administrator adds:

> "To gauge satisfaction it is better to see whether visitors return and whether they invite others (advocacy, referral) spontaneously, be it in the framework of a consumer referral reward program or not. Identifying those users is the best strategy to help a web concept go viral."

WM may thus not be a prerequisite to develop viral marketing and may even be less efficient than traditional use of surveys. But these are repulsive to web users. More descriptive techniques may thus be preferred for that purpose. One respondent indicated however that WM is truly insightful, hence less directional and more accurate, only if customers are required to log in on the website to identify unique visitors. Web behavior trends over time can thus be compared for loyal customers as compared to those who left the website.

Surprisingly, almost all respondents reported that it is almost impossible or at least very difficult to determine how respondents develop satisfaction, nor is it possible to determine their satisfaction level in a web context. Despite the fact that theoretically-speaking, satisfaction is an antecedent to loyalty, it is harder to determine it in a web context than it is for loyalty. One respondent stated that satisfaction is more of a psychological and emotional concept and such type of data cannot be grasped over the web except if users are asked to fill in a survey online or offline. Usage of surveys to determine satisfaction development was reported by two other respondents. An academician explains it this way:

> "Satisfaction development is harder to identify since it is a construct with hidden dimensions which may be hard to reveal and to understand with pure cross tabulation of data in a web context (convergence is not enough), although it may significantly impact users' behavior online. Emotions, for instance, influence tremendously satisfaction, loyalty and engagement toward a website.

A marketing director further adds that the satisfaction development and level is even harder to identify for each unique customer. For another respondent, this debate is futile since the fact that users come back to the website indicates they are satisfied with it and it is a measure of the website quality. There might be no need to make the concept of satisfaction more complex than it is. Satisfaction is always relative since we do not really know on what it is based (On design? Who can rate objectively design?). Amazon's customers are satisfied with its website because they come back to it and they developed satisfaction because Amazon was the first mover, the pioneer in book e-tailing. Most respondents agreed on the fact that satisfaction just like loyalty might be defined in several different ways and that it is far too subjective to be reasonably quantified. It is thus not necessarily well-explained through WM tools and remains relative, as an academician puts it:

> "A user might buy on a website by obligation, lack of choice or on behalf of someone else without appreciating the website."

This refers to the caveat of CLV identified in the previous section. Both the CLV and loyalty development patterns are great but superficial hints as to the customer's buying process. They do by no means reveal satisfaction and other affective traits of the customer.

Table 10 below provides validation of RQ 5 and its research propositions with summarized answers to the research question. Both RP1 and RP2 are partially validated. Consequently RQ 5 stipulating that WM methods and techniques enable to identify how existing web customers develop satisfaction and loyalty on the internet, is partially validated.

Table 4.6

Validation of Research Question 5

RESEARCH QUESTIONS	RESEARCH PROPOSITIONS	VALIDITY OF THE RP	ANSWERS TO THE RQ
Component 2: Usage of WM applied to the identification of existing customers' patterns on the web (HOW – INTERNAL DYAD 2)			
RQ5: [HOW – INTERNAL DYAD 2b] To what extent do clustering, association analysis, classification and prediction methods applied to web data capture how existing web customers develop satisfaction and loyalty on the internet?	**RP1:** Web data generated by existing web customers describe the existing web customers' satisfaction and loyalty patterns on the internet.	*Partially valid*	*Internal and external web data should be large, granular, issued by logged in web users, of good quality, and analyzed with WM to identify existing web customers' loyalty development. Traditional market research outputs and analyses should complement WM to identify existing web customers' loyalty development*
	RP2: Clustering, association analysis, classification and prediction methods applied to web data grasp the dynamics of existing web customers' satisfaction and loyalty patterns on the internet.	*Partially valid*	

CONCLUSION – As in Figure, company-specific attributes such as order of entry in the industry/sector, value proposition, product attributes, user experience offer or level of entertainment influence directly the customer-specific attributes relating to emotions, attitude, etc. toward the website. Customer attributes may not be grasped per se unless market research in the form of surveys are conducted among web users, but the web data do reflect partially those elements through web browsing behavior (intensity of usage, historical usage data, etc.), profile data (filling in of online forms, etc.), which are more accurate and less directional if customers are required to log in. These data are very useful to identify the paths of loyalty development by using both WM but also more descriptive tools such as frequency tables. However, these tools tell little about satisfaction development since satisfaction is a more complex construct with hidden and underlying facets that may be hard to get on the internet, requiring market research in the form of surveys instead. Although not very useful to tackle the satisfaction development process, WM and descriptive statistics help gauging the efficiency of a planned customer loyalty marketing-if a deliberate loyalty strategy was deployed at all- which in turn enables to drive powerful viral marketing. If no specific loyalty marketing effort was engaged, WM is useful to determine loyalty development based on the criteria of loyalty development defined by the business.

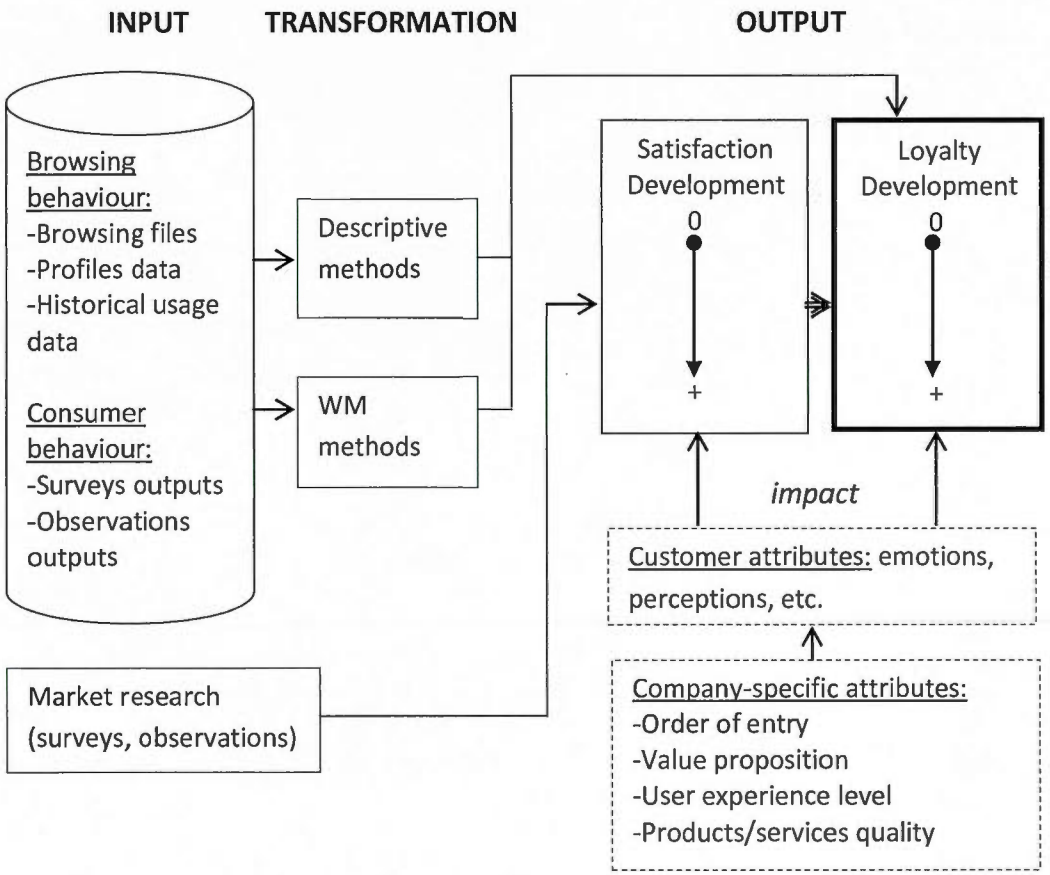

Figure 4.6 WM-enabled identification of existing customers' loyalty development patterns.

4.4.3. Identifying how existing web customers remain attached to- or defect from a website

Once a customer has become loyal to a business, in company-specific loyalty terms, it is also essential to track the evolution of that loyalty pattern. A customer who remains loyal over time is termed attached to the business. On the contrary, a customer who stops displaying the desired loyalty behavior is termed defecting from the business. The process by which existing web customers remain engaged with a given business (patronization) and how they are led to defect from it are closely linked to both concepts of satisfaction and loyalty. By understanding how web users become satisfied and/or how they become loyal, be they satisfied or not, it is possible to identify how they develop attachment or defection for a website. It has been seen previously that WM tools allow identification of loyalty development, but not satisfaction development. The question here is how WM can understand attachment development and defection development, with or without the satisfaction variable.

In case of missing satisfaction information, the WM project may lead to variable inaccuracies. Some customers might be identified as remaining attached to a website while actually not being satisfied (lock-in contracts, lack of alternatives, etc.). Conversely, some might be identified as defectors while they actually remain satisfied (money shortage to keep up with the price increases, etc.). Despite the missing link of satisfaction development-which could be identified with traditional market surveys though-identification of the formation of attachment or defection is not only interesting but also important. It empowers managers to predict, prevent or promote such behaviors (depending on the situation) among newer or older users and develop the appropriate web strategies to maximize the retention rate, in the first place and the customer acquisition rate, in a second place. Out of 10 respondents, 9 indicated WM are useful in that respect. Only one respondent reported that WM tools provide limited insight as to the way in which a customer patronizes or leaves the websites. Three respondents considered defection as an inverted measure of attachment that may be explained by the very same factors (explanatory variables). Design may appeal and cause attachment for one customer whiles it may irritate another to make him leave. One academician put it this way:

> "Starting to look why some people defect also tells why others remain attached to the business."

Defection may be easy to measure and WM helps also to understand how it develops. Out of the three, one however indicated that WM, although handful, may not be used often for these kind of analyses because not all companies can afford the proper data preprocessing tools and analytical resources. It may be more prevalent among companies that have all or very large parts of their business online or very large organizations such as MNCs for example.

Two respondents justified the usage of WM tools to identify attachment and defection formation by indicating that web data are available for that purpose. In fact, data on iteration of visits and purchases during a specific timeframe and users' interaction with a company's communications combined to the comparison with other consumers' previous behaviors determine how the user comes to be attached or defect from the website. A web business manager further adds:

> "Web data such as search keywords, user experience (navigation and usability) are translations of web users' interests and wants. Businesses can use WM to follow-up to determine at which point a user started using the website more frequently, what may have caused the purchase of an item, and other similar analyses which, combined together, may not provide the absolute but at least partial explanation of attachment or defection behaviours."

According to that user, such knowledge should drive layout design and architecture of the website to increase usability and user experience for increased patronizing. Another respondent acknowledged that classification was very appropriate for that.

As expected, 3 of the respondents who admitted that WM tools provide valuable insight into analysis of attachment/defection development, outlined that attachment/(defection), in the affective dimension as resulting from (dis)satisfaction, is harder to measure. A user can defect while remaining satisfied or remain attached while being unsatisfied. WM determines only statistically and quantitatively both satisfaction and loyalty, but tells little about the qualitative aspects. In fact, WM is useful for transactional studies based on transactional profiles (when did the customer buy? What? Etc.) but they are insufficient to understand why people leave or stay. Everything related to motivations or other

affective dimensions are almost impossible to obtain. There will always be a need for supporting research in the form of contacting or observing customers, to understand customers' reasons for leaving or remaining attached to a website in affective and emotional terms. Continuous and ongoing satisfaction research would be needed. This has already been underlined previously as being difficult to set up in a web context as it can irritate or distract users who wanted to make a purchase for instance.

Another respondent advocated thinking and analyzing outside the "website box" to get an overall view of attachment and defection patterns. Internet is a very free and open space that allows quick moves, increases availability of multiple information sources and as such reduces information asymmetries. Thinking of attachment and defection without considering other (dis)similar websites would be a failure to recognize the tight interlinks and intertwining of the world wide web. A database administrator states that:

> "You would have to stand outside and watch the return rate vs. defect rate of a demographic across similar sites/opportunities: where did everyone go? Is the key question to ask. The perspective of an individual business may not be seen until you step away from it and look at it alongside similar opportunities, competitors and even diverse contexts elsewhere."

He adds that this can be done by using WM analysis method, namely classification/prediction, clustering and association both in supervised or in unsupervised fashions. The question is how to get sufficient and appropriate data in order to analyze such insightful but complex patterns? Web analytics alone do not reveal confidential data of competitors such as defection rates or customer acquisition rates which are essential to draw a comprehensive image of customers' moves on the web. Also, this approach still does neither tell the motivations of customers, nor the underlying satisfaction which drives the development of attachment or defection. Such an approach would only determine that a given number of our customers do also visit competitors' websites at a certain frequency, for a given period of time and knowing whether this influences their patronizing of our website can only be inferred but not determined in absolute terms.

Table below provides validation of RQ 4 and its research propositions with summarized answers to the research question. RP1 is validated and RP2 is not validated.
Consequently, RQ4 stipulating that WM methods and techniques enable to identify existing web customers' behaviours on the internet.

Table 11 below provides validation of RQ 6 and its research propositions with summarized answers to the research question. Both RP1 and RP2 are partially validated. Consequently RQ 6 stipulating that WM methods and techniques enable to identify how existing web customers remain attached to or defect from a website, is partially validated.

Table 4.7

Validation of Research Question 6

RESEARCH QUESTIONS	RESEARCH PROPOSITIONS	VALIDITY OF THE RP	ANSWERS TO THE RQ
Component 2: Usage of WM applied to the identification of existing customers' patterns on the web (HOW – INTERNAL DYAD 2)			
RQ6: [HOW – INTERNAL DYAD 2c] To what extent do clustering, association analysis, classification and prediction methods identify how existing web customers' remain attached to or defect from a given business on the internet (patronizing)?	**RP1:** Web data generated by existing web customers describe existing web customers' retention and defection patterns on the internet.	*Partially valid*	*Internal and external web data should be large, granular, issued by logged in web users, of good quality, and analyzed with WM to identify existing web customers' transactional loyalty development. Traditional market research outputs and analyses should complement WM to identify existing web customers' emotional and actual loyalty development (patronizing)*
	RP2: Clustering, association analysis, classification and prediction methods applied to web data capture the dynamics of existing web customers' retention and defection patterns on the internet.	*Partially valid*	

CONCLUSION – Systems are not conceived to understand how and even less why customers leave. Such knowledge can only be obtained from intensive data aggregation and analysis to get a converging image through triangulation and inferences based on analytical outputs. As displayed in

Figure, the input data have to be rich and diversified enough (search keywords, user experience data, transactional data, communications interactions, browsing behavior, visits information). WM tools are then appropriate only to determine transactional loyalty patterns, that is, how existing web customers develop attachment or defection in terms of continuous purchases, visits or desired transactional actions. If actions are ongoing, the customer is deemed attached and if they stop he is deemed defecting. The emotional and affective aspects of loyalty cannot be grasped with WM tools alone. Market research needs to complement WM to provide additional insight into transactional and to determine emotional patronizing and hence provide an actual patronizing view encompassing the nuances of observed transactional patronizing. Contingent variables that influence attachment or defection such as cash availability of the respondent, seasonality, special occasions, etc. and their effect on loyalty can also be better grasped through market research. WM can also be used to a lesser extent.

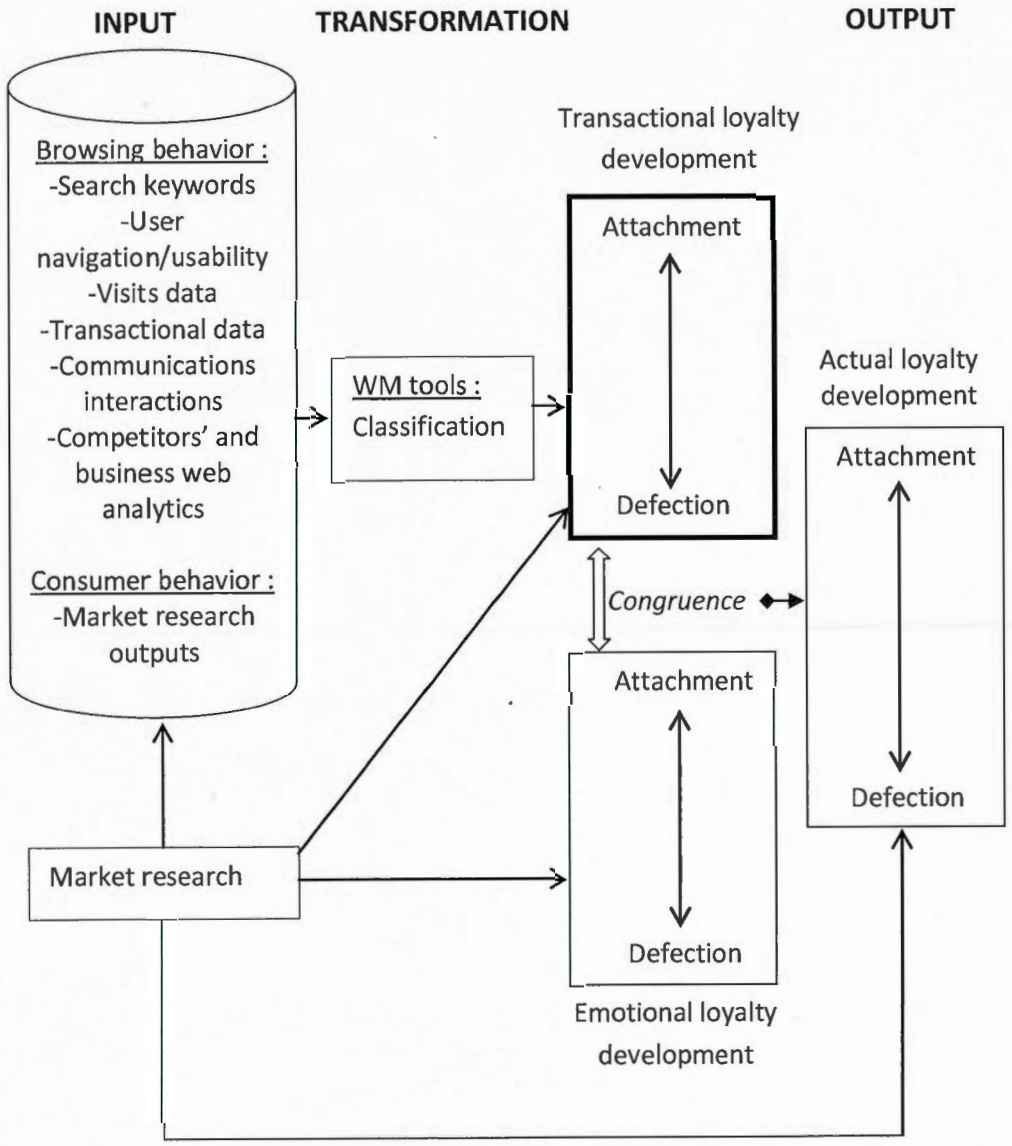

Figure 4.7 WM-enabled identification of existing customers' attachment or defection patterns.

4.4.4. Conclusion on WM-enabled identification of existing web customers' behaviors on a website

In a knowledge-enabled CRM framework, WM offers limited but insightful information on the behaviors of existing web customers. It has been established that, the input data used in the WM project should be large, homogeneous, and extracted from logged in customers. As shown in Figure, the data used as input for analyses may be: (1) internal, *e.g.,* search keywords, communications interactions, web analytics, transactional data,

browsing history, profiles data and clickstream data; or (2) external, *e.g.,* competitors' web analytics or 3rd-parties data. Some previous or current market research data resulting from surveys or observations can also be integrated to the WM process provided they are of acceptable quality and conform to the specifications of the pre-processing step. Both descriptive statistics and WM methods can be used. Descriptives are particularly useful to identify customers' behavior in a less complex fashion than do WM tools. They may also be used as exploratory tools to have a first impression of the data that will be further processed with WM tools.

The concept of behavior has actually two facets in a web context: (a) the browsing behavior facet and (b) the consumer behavior facet. The first refers to the actual, objective, observable and factual actions of an individual on a website that is everything (s)he does and that can be recorded in multiple ways for straight analysis. On the other hand, consumer behavior refers to the complex intermingle of all those underlying layers that constitute the human being and that are widely studied in a discipline that bears the same name. It encompasses thus such various elements as personality, motivations, lifestyle, perceptions, attitudes, emotions, satisfaction and loyalty, decision process, learning and socialization, if not more. Customer behavior can be influenced by company-specific attributes such as order of entry, level of user experience offered, etc. For instance, first-movers (order of entry variable) tend to be more appreciated by customers because they are perceived as being pioneers and tend to be more liked by the public than followers and late entrants (Srinivasan, 1988; Kerin *et al.*, 1992; Van Hippel, 1984).

Both types of behaviors surely influence each other but it remains unclear to what extent the typical customer behavior displayed in offline contexts, *i.e.,* everywhere but on internet channels, differs from the browsing behavior displayed on internet channels and even on different types of internet channels. Is it merely transposable? Or is there such thing as an e-consumer behaviour that an individual switches on when navigating on the internet? Additional research would be needed on that subject.

Further, it has been found out that while browsing behavior can be identified very accurately by WM tools and to a lesser extent by descriptive methods, consumer behavior aspects cannot be that easily determined with WM. At least, not yet. Market research is still required for that purpose. The level of congruence between browsing behavior and

consumer behavior cannot be specifically determined, nor can the resulting actual satisfaction development be pinpointed either with WM alone. Again market research is necessary to determine actual satisfaction development. Satisfaction is the major antecedent of loyalty. Without knowledge about the satisfaction level of customers though, WM tools enable marketers to identify how web customers develop transactional loyalty and how they nurture that transactional loyalty on a continuous basis by remaining attached or by defecting. Transactional loyalty, however, tells little about the underlying buying process of the customer. Emotional loyalty refers to the liking or not liking of a website on a continuous basis by remaining thus emotionally attached or not. Emotional loyalty development cannot be grasped by means of WM tools alone and market research is also needed here. Congruence between transactional and emotional loyalty development exhibits the actual or real loyalty development. Again, without market research this is impossible to get with WM tools alone. The example below is a simulated real-life case developed by the researcher based on the findings relating to the second theme of identifying existing web customers' behavior on the internet. It summarizes the potentialities of WM as identified so far:

> *An example will illustrate this application of WM. Drinkco.com is a website selling various kinds of functional drinks namely sports and energy drinks in multiple formats. The business decides to focus on logged in customers to understand their loyalty formation patterns. Although transaction-oriented, the website is very entertaining and offers a high level of user experience to increase the flow and satisfaction level for increased sales. This has another advantage, the company collects many different kinds of data: search keywords entered in the index section, browsing history, clickstream data, web analytics about all users, and more individual data such as transaction information and user-generated content provided into various forms and communications interactions such as emails, online polls and so forth.*
>
> *To get an overview of users' loyalty, the business reviews frequency tables of visits and purchases to make a ratio. For every 10 visits a purchase is made, which is a very good figure in the industry. The purchase data are then used in an association analysis by means of sequential patterns with the PrefixSpan algorithm to discover how items purchased by customers are associated and the sequence in which the items are purchased (Liu, 2011). About 56 relevant*

sequences fulfilling minimum multiple confidence and support levels have been identified. It essentially says that 35% of customers who buy Red Bull first, then buy Monster, then Amp and eventually less known brands such as Guru and continue buying on frequent bases. In comparison, 8% of customers buy firstly Guru, then Hype and NOS before leaving. This tells that the most attached customers are initially attracted by the market leader brand and develop loyalty over time, assuming they develop satisfaction too, by trying other brands narrowing down until niche-oriented beverages. Big brands leverage the demand for smaller brands creating a "long tail effect". On the other end of the continuum, defectors are rather usually attracted by small brands first and seek additional smaller brands. The website is not a niche-oriented spot though and thus these customers defect from the website because they do not find what they want. In addition, market basket analysis revealed transactional loyalty patterns which allowed building decision aids tools based on the past purchases and on purchases made by groups belonging to similar loyalty development groups.

WM tools revealed insightful knowledge about transactional loyalty patterns but are not suited to identify dimensions of consumer behaviour. Drinkco conducts several market research offline and online to identify motivations, perceptions, emotions, preferences, interests and satisfaction development of loyal and disloyal clients. It then compares clients' transactional loyalty development patterns to the emotional loyalty development patterns, which was asked to customers. The level of congruence between both elements indicates the actual loyalty development of clients. Results are taken cautiously by management though because response rates for online surveys were very low, whole sections were often intentionally left blank and outliers emerged, all of which contributed to decrease the reliability and statistical validity of outputs.

Management knows the limitations of WM tools usage but finds great utility in the transactional loyalty development model. The knowledge-enabled business knows that marketing efforts should be geared toward consolidating a strong base of big energy drinks brands and integrating more smaller brands of energy drinks which can be proposed to customers in the form of recommended beverages or which can be advertised on the website, or on other websites through behavioural advertising as well as on emails or profiles of social networks visited by customers. Creative cross-selling and up-selling strategies should be

implemented based on that finding in order to retain big brands customers tempted to discover new tastes and flavors by seeking less conventional brands as well as smaller brands customers who wish to stick to "underground", "out-of-the-mainstream" beverages.

The second theme corresponding to the second meta-objective of identifying existing web customers' behaviors on the internet, as identified in Xu and Walton's (2005) adjusted framework, can be fulfilled by using WM. Traditional descriptive statistics are also useful complementary tools in that respect. Market research remains a cornerstone for identification of the more psychological and psychographical aspects of web customers behaviour.

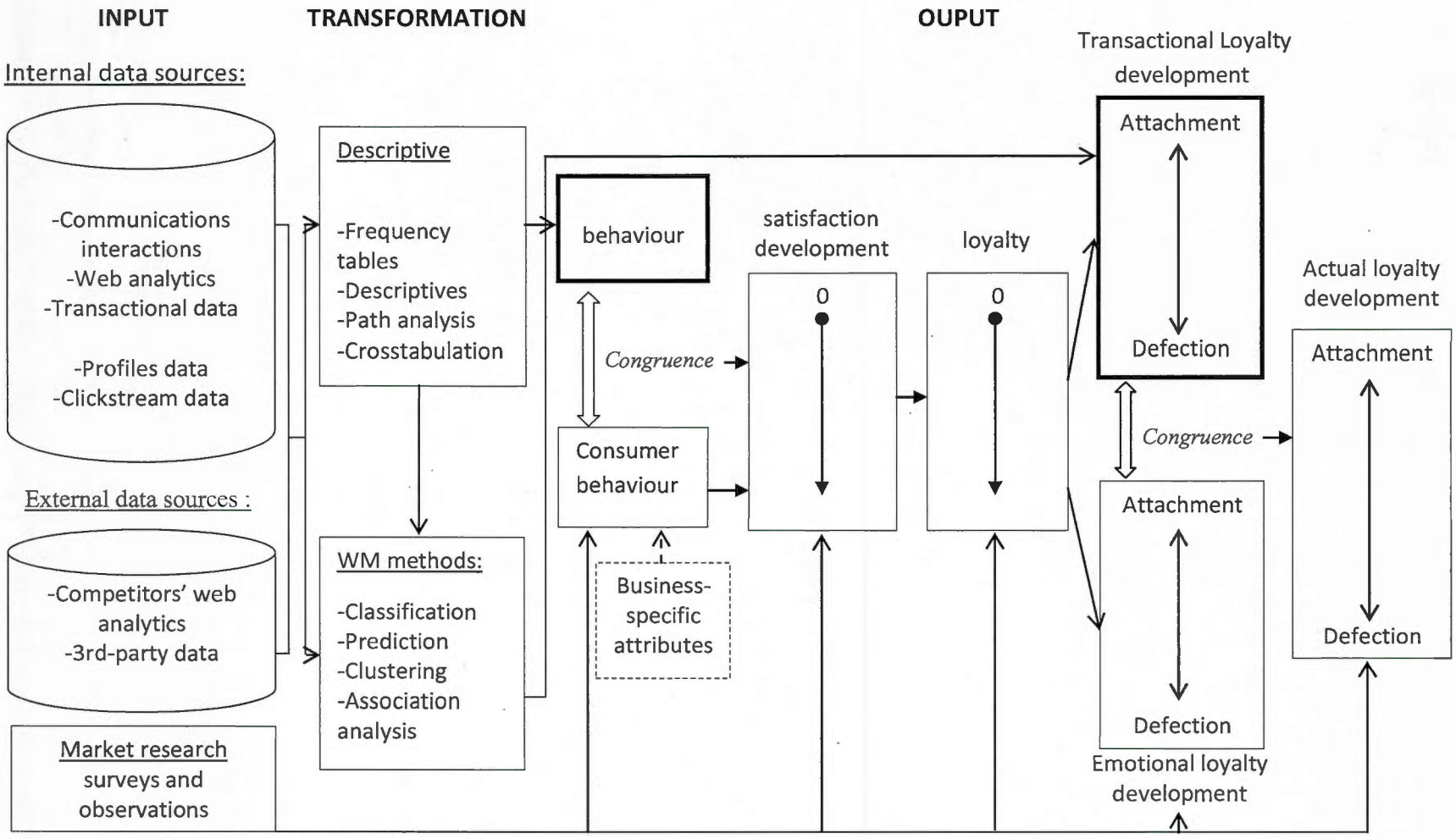

Figure 4.8 WM-enabled identification of existing web customers' behaviors on a website.

4.5. Profiling of prospective web customers on the internet

Prospective customers of a website are defined as those who visit the website without having ever made a purchase on it. Any other definition of a prospect would be difficult to make in a web context. Consequently, all visitors are considered prospects as long as they have not made a purchase, or performed the desired transactional action. If they do, they are converted into existing customers who were the focus of the 2 previous sections.

This section aims at identifying how WM tools profile prospective web customers and how they reveal their behavioral patterns.

4.5.1. Segmenting prospective web customers of a website

Segmentation of prospects is very similar to that of existing customers. The main difference lies in the lack of prospects' transactional data which provide additional insight into their needs, wants, interests or preferences (Tufféry, 2011). Such knowledge is therefore usually not inferred from transactional data, but rather from other sources such as browsing behavior, clickstreams, etc.

Out of the 10 respondents, 9 indicated that WM methods were effective in identifying the profile of prospects. Eight respondents agreed on the value of classification for segmentation purposes. Three of them added that when combined to regression such technique offers tremendous insight into the interests of prospects which enables to segment them, whereas the fourth respondent indicated that regression was not that useful in that respect. An academician further explains:

> "This can be summarized as follows: show me how you click on the website and I will tell you who you are and what you may be likely to do on the website as well as your future behavior."

The focus on prospects, that is, the segmentation of prospects alone can also be the result of the generalization of the model developed from a statistically representative sample of the customer database. If the website has enough assurance that both customers and prospects do share similar characteristics, those profiles extracted from the customers'

database might be partially or totally congruent with that of prospects as one respondent puts it:

> "If we can segment existing customers we can segment prospects too."

The trick lies in precisely determining on which basis prospects are to be segmented. Two respondents recognized that there are different types of segmentation across different companies and even industries, *i.e.,* young vs. elderly prospects. There are also different definitions of what a prospect may be. For an online casino website, prospects under 18 years of age would never constitute a target of interest, doing so would even constitute a crime. Also, if a user comes for the first time on a website, can he/she be considered as a prospect? It may thus be less easy to segment prospects albeit not impossible. A database administrator indicates:

> "Segmenting prospects is possible if segment means interests of a customer and channel him into a concept where he/she has a higher propensity to spend money. In that case, WM tools are appropriate."

For another respondent it is also possible to segment prospects if web managers possess some sociodemographic and socioeconomic data about them in addition to browsing behaviors. In that case, if web managers see prospects returning to the site several times, they may issue sharper profiles to optimize both online and offline customer acquisition strategies. In a multichannel context this could mean contacting the sales rep of the prospects' local division to let him get in touch with the prospect. An academician puts it this way:

> "The potential market (prospects) is not only online or offline, it can and is usually both. It is important to have an overall, a global view of the market."

WM enables thus to draw powerful multichannel strategies. However, WM might be volatile because individuals are users not demographics, they evolve continuously. According to a respondent, classification and regression techniques should thus be combined to clustering to draw a more stable but momentary picture of prospects. In addition, one respondent acknowledged the fact that WM tools become actually really useful once prospects have been converted into (loyal) customers since in such instances segmentation becomes sharper and more accurate. Another respondent suggested the same idea and added that WM methods work better for converted prospects who log in.

Only one respondent stated that WM tools are unreliable to profile prospects because the data sets used to analyze prospective web customers are of limited scope and scale. Not only are they too small in quantitative terms with too few observations especially on small websites (scale) but they present also limited web users' characteristics which could be used for segmentation purposes (scope). This contradicts the statement of other respondents who declared that a great variety of attributes about prospects could be obtained even in a web context. One respondent for instance warned about the fact that prospects may not really be interested in buying products or services offered by a website to call them truly prospects. However, he added that if one wants to dig further there are other means to obtain insightful data on prospects. Facebook Inc. and other third party companies can sell highly detailed profiles of individuals to companies.

Prospects can be better identified provided they have an active Facebook account or do not opt-out from ad networks and other data exchange marketplaces. Today, social networks, which are true products of the interactive web 2.0 era as well as weak legislation regarding data confidentiality and privacy online, created plenty of opportunities to buy and aggregate highly detailed knowledge about almost every aspect of a human's life (professional dimensions with LinkedIn, private life with FB, leisure/hobbies with FB again but also Flickr, YouTube, Dailymotion, etc.). Using such data web businesses can develop highly customized advertising tailored to the specific needs and interests of prospective customers at a specific point in time and diffuse them on all the web pages the prospects might visit. This display advertising technique known as behavioral targeting has proven to be highly effective at increasing publishers' revenues through increased click-through rates by 670% (Yan *et al.,* 2009). As a result, there are higher conversion rates and increased revenues for web businesses as well (Beales, 2010).

Such data files may however be too costly for small businesses with limited marketing resources. But the fact is they do exist, so do the techniques to analyze them, and increasingly also if they are generated by a small number of observations. Segmentation of prospects might slowly become more feasible for more companies on the web.

Surprisingly, no respondent mentioned surveys as complementary tools for segmenting prospects. Market research may not appear to be absolutely needed in that case. Profiles can be well-developed from online databases only.

Table 12 below provides validation of RQ 7 and its research propositions with summarized answers to the research question. Both RP1 and RP2 are validated. Consequently RQ 7 stipulating that WM methods and techniques enable to segment prospective web customers, is validated.

Table 4.8

Validation of Research Question 7.

RESEARCH QUESTIONS	RESEARCH PROPOSITIONS	VALIDITY OF THE RP	ANSWERS TO THE RQ
Component 3: Usage of WM applied to the profiling of prospective customers on the web			
RQ7: [WHO – EXTERNAL DYAD 3a] To what extent do classification and prediction methods segment prospective web customers?	**RP1:** Web data generated by prospective web customers are sufficiently detailed and accurate to provide a strong basis for the creation of precise profiles about those prospective web customers.	*Valid*	*Internal and external web data should be large, granular, issued by logged in web users, of good quality, and analyzed with WM to segment prospective web customers' and do accurate profiling*
	RP2: Classification and prediction methods applied to web data create homogeneous groups of prospective web customers	*Valid*	

CONCLUSION – Website managers own useful browsing information, clickstreams and socio-demographic as well as socioeconomic information collected from forms, registrations, etc. in order to draw useful profiles with classic WM tools. As shown in Figure, the level of segmentation determines the resulting marketing strategy from mass to individual. Social networks which are gold mines of precious unique profile information provide additional attributes on which to segment prospects or garner a given

individual's profile. Gaining useful insight of prospects, is a stepping-stone for segmenting them and crafting strategies accordingly to acquire them in multichannel settings, through tailored display advertising (BT, demographic targeting, geographic targeting, etc.) or direct marketing campaigns, the very core of CRM.

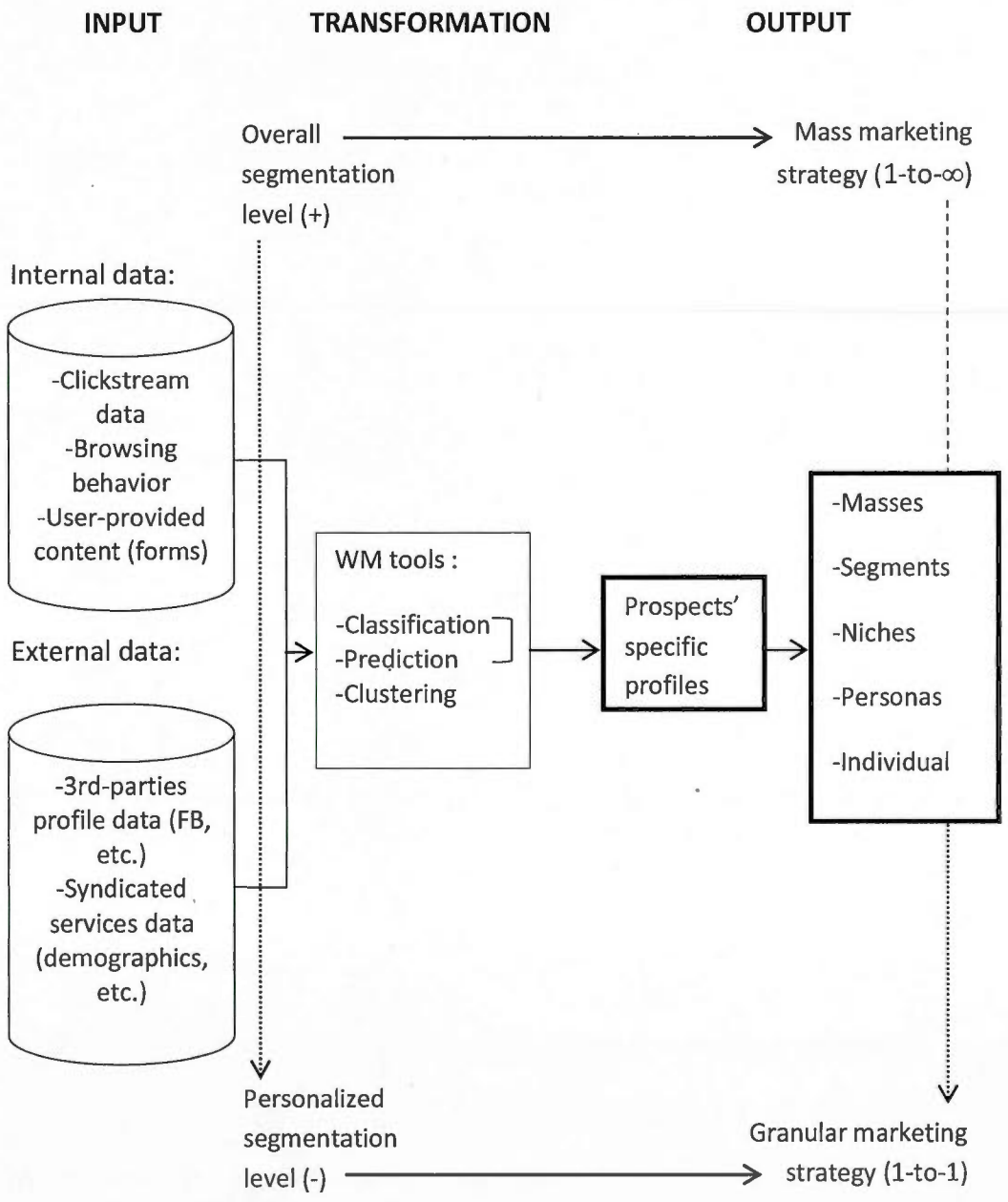

Figure 4.9 WM-enabled profiling of prospects.

4.5.2. Collecting information about prospective web customers' preferences, needs, habits, etc. to develop targeted e-marketing strategies and acquire them

This aCRM objective is closely linked to the concept of segmentation. Segmentation may be based on a multitude of attributes such as sociodemographic data, geographic data, psychographic data, attitudinal or behavioral patterns, etc. In their original framework, Xu and Walton (2005) based segmentation on demographic data for instance. In order to develop e-marketing strategies and especially direct marketing strategies, as seen before, web managers need to have previous knowledge of their prospects' profiles which may be comprised of all sorts of characteristics. However, the most useful knowledge refers to prospects' preferences, current needs and wants. If profiling does not take into account the wide array of prospects' interests and if segmentation is not based on such attributes, drawing relevant e-marketing strategies to acquire prospects might not be flawless. Because, segmentation does not necessarily take into account these attributes it is thus essential to determine if they can be identified at all by using WM tools.

All respondents agreed on that point with slight nuances though. The overall sentiment is that clustering is particularly appropriate because it is unsupervised, contrarily to classification which is supervised. Unsupervised seems more appropriate in an exploratory context such as that of discovering preferences, needs, etc. One respondent however suggested that simpler descriptive methods may also be appropriate if not more than WM. Another respondent emphasizes that WM actually permits to develop a more dynamic view by displaying behavioral patterns, which descriptives do not allow to do. In fact, interests, preferences, etc. are always captured in browsing behavior. It is always regarding browsing that web managers might infer such strategic data as those related to overall interests and habits. He puts it like this:

> "Traffic analysis, browsing paths of prospect and longitudinal statistics allow to crosstabulate prospects and existing customers to identify correlations. This approach permits to link actual transactions to the website traffic, in other words to develop a relation between browsing volume (generated by both prospects and customers) and final transaction data (generated by old or new customers only) to develop such statistics as: for each additional unique new 2000 prospects, 1 purchase is made."

Web descriptive statistics such as web analytics do in fact provide many data on individual users' interests and habits but it remains on a static schema. WM permits to develop a more dynamic view by displaying behavior and as such WM is as appropriate if not more than descriptive statistics. From that point, predictions can be made as to prospects' future likely behaviors to develop offers and focus on e-marketing strategies (newsletters, articles, ads, etc.), based on identified prospects' web habits and trends to attract them.

WM tools should also be used in conjunction to develop real-time "response frameworks" to dynamically modify the approach taken with the web user on the fly and not post-hoc, *i.e.,* morphing. A database administrator suggests:

> "The usefulness of traditional slow-and-steady methods that keep up with the flight of individuals in their demographics flocks will become as obsolete as the Pony Express is to communications. Now users become more savvy and internet grows consequently in its ability to respond to user preferences with dynamic approach to cross-selling based on regression/classification and the use of clustering that responds on the day or the hour according to current trends, which is far more valuable than traditional approach."

This is also why, as a respondent puts that, WM provides great insights in an online context but only limited insights into off line for the prospects' preferences might be related to fonts, website layout color, wallpaper, add-ins, and so forth. However, most businesses are mostly interested in prospects' hobbies, leisure, activities or opinions in order to match them with their own products and services offerings. The rise of social networks has much helped in that respect because they are spots where people reveal their true preferences, habits and needs without even noticing it. An academician highlighted that:

> "FB does not sell demographic data they mainly sell behavioral data : links on which users click, ads they watch, what they "like", what their friends "like" and all activities put at their disposal for their own pleasure, while they are actually sharp measurement and evaluation tools of behaviors. These are then sold."

Attitudes and behaviors on the web are translated into web data that provide hints as to the interests and habits of people and these data are increasingly more available through social networks. Web businesses can identify IP addresses of their users and ask third-parties to sell them behavioral data on the user behind that IP address. Using social networks allows to identifying the unique user(s) hidden behind mere numeric IP addresses. More comprehensive and accurate pictures of users can then be drawn. Highly tailored advertisements should be produced and submitted to prospects.

However, usage of such data and application of WM tools to derive useful meaning and tailor sharp advertising may not necessarily increase conversion rates of websites as well as profits, at least not in the short run. One respondent indicated that it is more useful to do branding and conditioning on the longer run:

> "If you are sent many ads corresponding to your tastes, even though you might not have the budget now, you will have it one day when you will start working and then you might consider those offers again (recall) because you will already have a favorable idea of the brand due to these targeted actions. You can convert prospects into customers in the longer run too."

WM tools are however also subject to limitations in identifying habits and interests of prospects. One limit refers to the architecture of a website. The more structured, the better the analysis. Lots of data are also available but not all may be relevant to identify needs, habits and preferences. They may be too overwhelming, calling for extra amounts of time spent sifting useful from unusable data or even worse, leveraging false habits and interests. Another respondent indicated that WM is appropriate if datasets are large but this is rare.

Table 13 below provides validation of RQ 8 and its research propositions with summarized answers to the research question. Both RP1 and RP2 are validated. Consequently RQ 8 stipulating that WM methods and techniques enable to identify prospective web customers' attitudinal characteristics to acquire them, is validated.

Table 4.9

Validation of Research Question 8

RESEARCH QUESTIONS	RESEARCH PROPOSITIONS	VALIDITY OF THE RP	ANSWERS TO THE RQ
Component 3: Usage of WM applied to the profiling of prospective customers on the web			
RQ8: [WHO – EXTERNAL DYAD 3b] To what extent do clustering, association analysis and classification methods provide insightful information about prospective web customers' preferences, needs, habits etc. to develop targeted e-marketing and e-commerce strategies to acquire them?	**RP1:** Web data generated by prospective web customers are sufficiently detailed and accurate to provide various and complementary characteristics about those prospective web customers such as their preferences, needs, habits, to be used for acquiring those customers	*Valid*	*Internal and external web data should be large, granular, issued by logged in web users, of good quality, and analyzed with WM to identify prospective web customers' needs, wants, habits, preferences, etc. to enable crafting and deployment of targeted e-marketing strategies to acquire prospective web customers*
	RP2: Clustering, association analysis and classification methods generate relevant information about prospective web customers' characteristics which can be used for further targeted marketing and sales efforts to acquire them.	*Valid*	

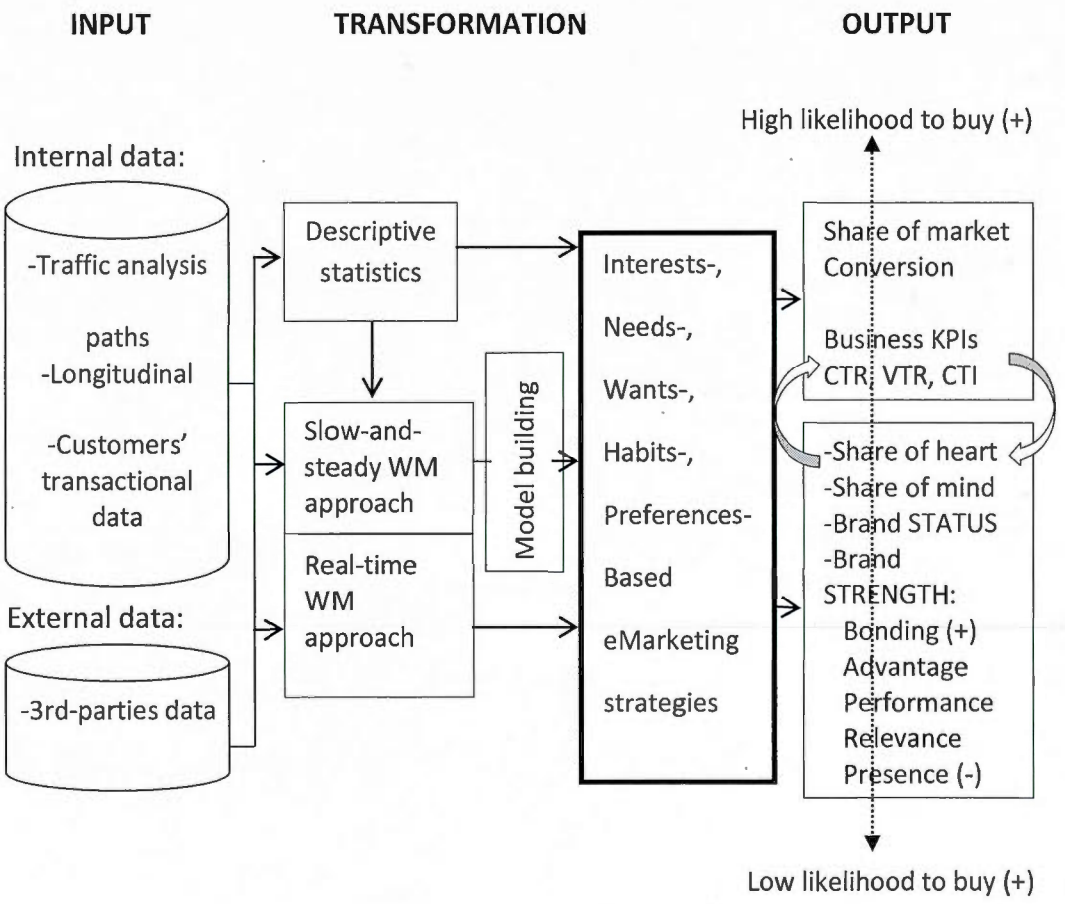

Figure 4.10 WM-enabled identification of prospects' attitudinal characteristics to acquire them

CONCLUSION – Prospects may be segmented against a multitude of attributes (VALS, Rokeach models on values and lifestyles, AIO model, etc.). Such data can be gleaned successfully in a web context with slight differences from real-world data in some instances. Browsing paths, traffic data inform about a prospect's web behavior and hence interests, habits, needs and the like. Transactional data provide additional and sharper insight into that respect and can be correlated to browsing volume to triangulate both data blocks and pinpoint true interests and habits. WM tools and to a lesser extent descriptive statistics enable to discover these behavioral patterns which are hints to habits and interests. It appears that WM tools can be subdivided into two approaches: one "slow-and-steady" approach that analyzes the data and produces business models, rules, algorithm post-hoc and another "on the fly" approach that produces in real-time dynamic

response frameworks (models, rules, etc.) directly applicable on the website. The latter approach seems more efficient especially in the fast-moving web context. Figure illustrates the fact that outputs generated by both approaches have effects in the short as well as in the long run. In the short term, better knowledge of prospects' needs etc. enables businesses to build highly tailored e-marketing strategies (morphing, BT, etc.) increasing CTR, Click Trough Interests (CTI) or the views divided by the clicks and View-Through Rates (VTR) or the views divided by the shows, the conversion rates and acquisition rates. Impact of those strategies need not be instantaneous but may also impact prospects in the longer run through branding and conditioning to increase brand awareness, recall, recognition, equity to increase the brands', products', organisation's share of prospects' minds and hopefully share of heart (Kotler & Keller, 2006), so that they consider them during future decision-making processes. Their likelihood to buy is thus low at the moment but may increase substantially in the longer term when more favorable conditions or settings enable the web visitor to be more spendthrift.

4.5.3. Conclusion on WM-enabled Profiling of prospective web customers of a website

Unlike existing customers, prospects are harder to profile because the company lacks fundamental data such as transaction information that bring most insight into individuals' profiles. It is also more difficult to track individual unique prospects on a website since they do not necessarily log in. The company needs to find a way to attach input data to each unique user with confidence. Various datasets such as traffic analysis, browsing behaviour, clickstream data, longitudinal statistics (descriptive statistics done over time), profiles data, user-provided content and existing customers' transactional data can then be used as input as long as they are large, well retrieved, selected, cleansed and pre-processed.

The business usually transposes the segmentation level applied to existing customers to prospects. As previously identified, the segmentation continuum ranges from masses to the individual. The targeting level chosen determines the resulting marketing strategy used to attract and retain prospects after WM outputs have revealed knowledge. Both descriptive methods and WM can be used. Frequency counts and so forth provide an

initial overview of the data as well as an aggregated input for WM processes. The transformation process can be done through the slow-and-steady approach of the real-time approach as this was determined for existing customers.

Either way, it has been established that WM contributes to draw very accurate specific prospects' profiles. The classification technique is very useful for estimating the missing transactional aspects of prospects. By using existing customers' transactional information, predictive models can be developed in order to anticipate the likely action or attribute of a prospect. Prospects are assigned a score which classifies them in groups identified post hoc. WM also allows determining prospects interests, needs, wants, habits and preferences which are more attitudinal and behavioural types of data. Based on the derived knowledge the business can build segmentation schemes, *e.g.*, psychographics-based niche segmenting, etc. The choice of the variables to be used as group descriptors, as well as segmentation techniques and other arbitrary considerations remain under the control of the manager.

In figure 20, those outputs that WM enables to obtain are framed in bold. Prospects' profiles together with their identified needs, interests, wants, habits and preferences drive the business' eMarketing acquisition efforts. It is based on that knowledge that the company can craft relevant strategies that have high chances to be considered by prospects because they correspond to their tastes and needs.

The effect of these acquisition strategies may be twofold: in the short run and in the long run. Most businesses will seek to get the short run impact which affects mostly the bottom line. In fact, deriving accurate prospects' profile and offering them a value proposition tailored to their unique being, increases the likelihood that these prospects will engage in the activity desired by the company be it a video game account subscription, a book purchase, or whatever else. In web advertising, behavioural advertising capitalizes on WM techniques to enhance core strategic indexes such as Clickthrough Rates, Viewthrough Rates or Clickthrough Interest and boosts conversion rates (Yan *et al.,* 2009; Fulgoni, 2008). More clicks on ads means also more revenues for publishers (owners of the website displaying ads) or any other entity that hosts ads, enabling them to also enhance the quality of their website(s) (Beales, 2010). For e-business announcers increased click rates also increase the Return on Marketing

Investment (ROMI) comprised of Cost Per Mille, effective Cost Per Mille, Cost per Thousand, Cost Per Impression, Cost Per Click/Pay Per Click or Cost per Action, depending on the pricing model chosen. Such a virtuous circle makes e-commerce more efficiently-run. This is in line with what Pfeiffer and Zinnhauer (2010) asserted, namely that online advertising is especially effective in later-funnel stages (order, repeat visits, repeat order[34]), but should still be combined to classic media to support upper-funnel marketing-related activities (impressions, clicks, offers, calls-to-action). WM-driven eMarketing tactics and strategies may therefore have a tremendous impact on e-business' KPIs and market shares in the short term.

Meanwhile, additional long-lasting and incrementally-built impacts develop in the longer run. In the literature it has in fact been recognized that highly targeted online campaigns contribute to brand equity by enhancing brand status, namely brand recall, awareness, sympathy, intent to use and first choice (Pfeiffer & Zinnhauer, 2010; Hollis, 2005). Attitudinal aspect of brand equity (*i.e.,* brand strength) can be estimated via the BrandDynamics framework (Dyson *et al.,* 1996). Other measurement tools may apply. Online advertising contributes to increase the five levels of attitudinal loyalty to a brand to heighten the probability of a person buying a brand (Hollis, 2005). The lowest level is *presence* (knowledge of what the brand stands for). It corresponds to the lowest prospects' likelihood to buy; the *relevance* step is defined by negative drivers of loyalty (*e.g.* cheap brand thought to be of bad quality, etc.). In *performance*, the prospect agrees the brand provides acceptable levels of performance on basic criteria (Daniel's (1961) Critical Success Factors (CSFs)). In *advantage*, the prospect admits that the brand displays long-lasting and ongoing advantages over competitors (Daniel's (1961) Key Success Factors (KSFs)). Eventually, *bonding* takes account of prospects' relative classification of the brand's advantages (or KSFs) and the degree to which they believe the brand shares the same endorsement than others. By moving up the pyramid as shown in figure 20, the prospects' likelihood to buy increases accordingly. Both brand status and brand strength contributes to capture more of prospects' share of the mind and ultimately share of the heart (Kotler & Keller, 2006).

---

[34] This is based on Viewmark Marketing Agency's specific sales funnel model: customers start with impressions and continue through clicks, stickiness efforts, repeat visits and clicks, engagement offers, calls-to-action, contact or call-back, order, repeat visits and repeat orders: www.viewmark.com

The online marketing campaign remains framed in the segmentation level and subsequent acquisition strategy that have been chosen by the business. Mass segmentation and its subsequent global acquisition strategy will typically lead to less relevant and interesting marketing campaigns for prospects. This approach may nurture their knowledge of the brand but also decrease their purchase intentions. On the contrary, personalized segmentation and its subsequent granular acquisition strategy drives prospects to the bonding stage so that they agree by themselves that the brand is superior, different and worth the dollars to be spent. A simulated real-life case, developed by the researcher, will shed some light on the practical application of WM to profile prospective web customers of a website.

> *Reinhardt&Kuntz.com is a furniture wholesale broker website. It buys and sells furniture entirely on the internet. The company reaches out to upper-middle class individuals with a pronounced taste for robust and long-lasting German-made furniture. It does not have a very detailed segmentation strategy though and uses primarily its website to attract and retain masses of potential customers. Existing customers are automatically identified and logged in on the website. From its web analytics reports, it discovered that 75% of its website traffic is done by non-existing customers. R&K would like to draw a profile of the diverse users who come and leave without purchasing. Since the company focuses more on logistics, procurement and relationship management with suppliers, it does not collect much data about web users. It can use its existing customers' transactional data, and prospects' browsing paths and clickstream data. Some prospects subscribed to the company's newsletter and provided their email. The company buys some additional data about them from Facebook and Twitter corporations.*
>
> *Descriptive analyses have been done prior to the data gathering with web analytics. Multiple WM techniques are used such as classification and prediction which use existing customers' browsing and clickstream data and correlates them with prospects' data. Cluster analyses are also conducted to form prospects groupings. A special interest is granted to prospects who display similar browsing activities than existing customers, especially frequent buyers. The company does not have a very advanced IT environment and processes the data in a slow-and-steady fashion. After model building and testing on test sets of prospects they are implemented to draw specific profiles of prospects.*

The vast majority of prospects share (browsing) similarities with low frequency buyers and may belong likewise to lower social categories. Additional data from social networks indicated they are usually undergraduates, postgraduate students or young professionals, liking and clicking preferably bargain-related advertisements (e.g. Groupon) and using the web for almost any purchase type. More refined data provided insight into interests, habits and preferences of these prospects which were aggregated to the specific profile that the company had built so far. It learns for instance, that many of them are "slacktivist" or activists with concerns about the environment, wood logging, earth warming and other sustainable development issues; they signed online petitions, belong to online discussion groups and leave posts on these subjects. Meanwhile, they live busy lifestyles, display hedonist behaviours and are keen on state-of-the-art technology.

Based on that knowledge R&K drafted a knowledge-driven acquisition strategy that should capitalize on social networks and mobile technology to leverage the demand from these young prospects. It develops an online advertising campaign based on fun, auto-derision and green marketing. Ads will be displayed on profiles of prospects as well as on the website they tend to visit most. R&K expects that only a limited portion of prospects will reach the final stage of the sales funnel and become repeat customers of the website. However, it primarily seeks to build brand equity among that public. The prospective target market may not have the money know but as they start working and earn more money, prospects might consider again those ads or simply recall them.

The third theme corresponding to the third meta-objective of profiling prospective web customers of a website, as identified in Xu and Walton's (2005) adjusted framework, can be fulfilled by using WM. Traditional descriptive statistics as well as data about prospects issued by third parties or syndicated services constitute useful complementary tools and material in that respect.

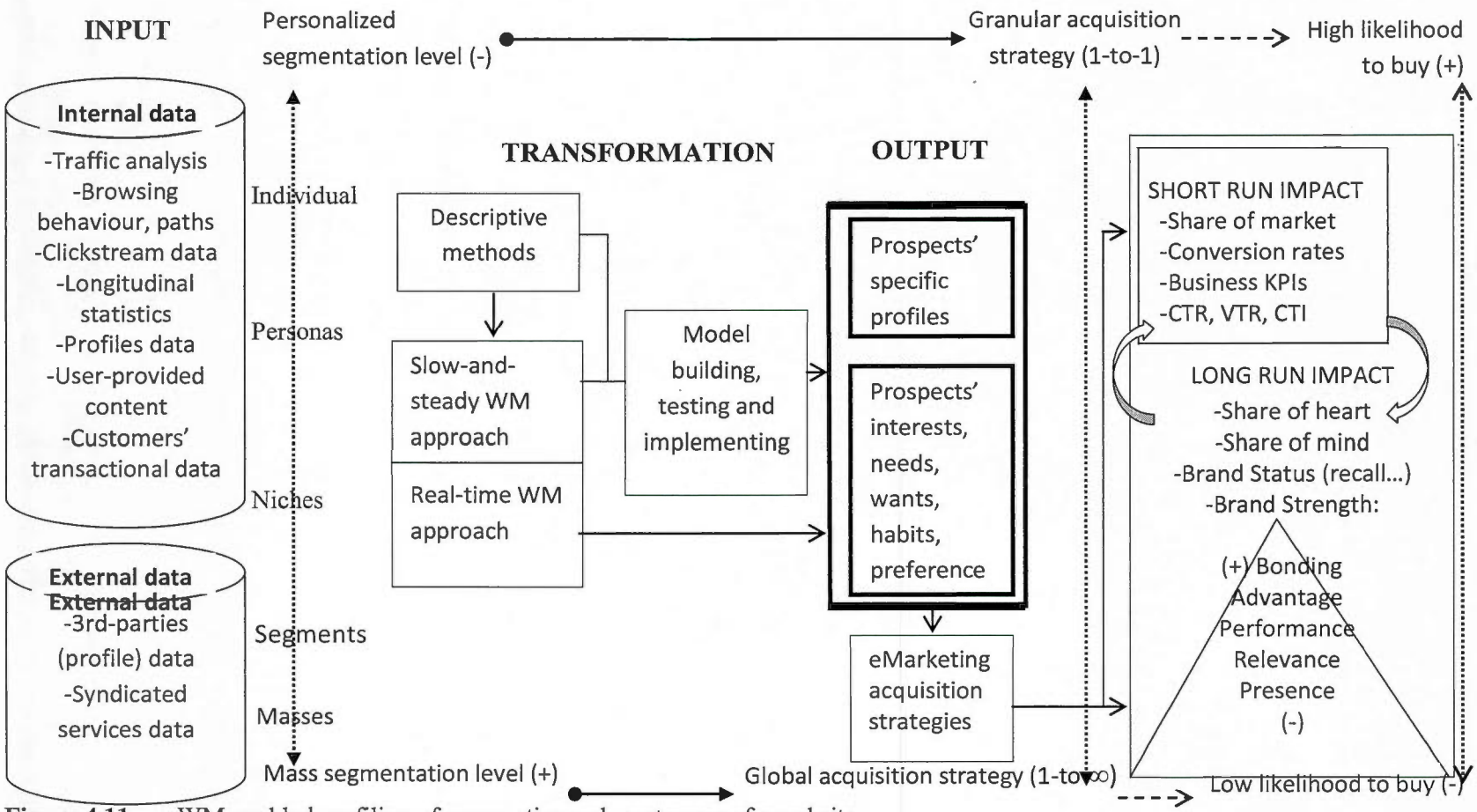

Figure 4.11 WM-enabled profiling of prospective web customers of a website.

4.6. Identifying prospective web customers' behaviors on the internet

Just like existing web customers' behavioral patterns may be observed on a website, those referring to prospects may also be subject to in-depth analyses. This part investigates to what extent WM tools enable to identify the "HOW" aspect of prospects, that is how they behave on a given website as well as on competitors' website.

4.6.1. Identifying prospective web customers' behavior patterns on a website

All respondents agreed upon the fact that prospects' behavior may be identified on a website thanks to WM tools. Four respondents recalled that real and observable behavioral data are highly useful because they show what prospects did on the website, which informs about their interests and habits as specified previously. All other psychological aspects of their behaviors cannot be grasped by means of WM tools because these are underlying dimensions. One respondent indicated that even for prospects it is possible to collect highly specific web data by contacting third parties. Such knowledge collected by means of the data marts can be even more reinforced by integrating external data from web 3rd-parties (FB, etc.) or from syndicated services such as Mosaic or PRIZM data. These types of data do in fact allow websites to segment their prospects according to shared demographic, lifestyle and behavioral traits. Once that information has been integrated into the system they may be cross-tabulated with the socioeconomic profile and behavioral data obtained on the website and from third-parties. The credit card is also very useful in that respect since it gives plenty of information on the individual's credit score, postal code, and so on.

However, credit card information is rarely known for prospects since they have not made any purchase yet. Even if they do so, credit card information remains highly sensitive and confidential. Also, before purchasing costly environics analytics and/or third-parties' data, the company has to be sure as to the unique identity of the prospect. It may be that the prospect identified on a website is connected on a computer that is freely available to students and hence the activities tracked on the IP address of that computer do not refer to one unique user but to several hundred students'. One respondent indicated that recourse to buying external data may only be relevant in the case where web users, both existing customers and prospects, need to log in. Two respondents also indicated that WM tools are very appropriate but the usefulness of their output always depends on the quality and

quantity of the data. One respondent added that the more homogeneous (demographically or in purpose) the visitors' data, the more accurate the browsing profile of prospects. Another underlined that this depends essentially on how well the website is structured, the more categories, the better.

Regarding the methods, two respondents highlighted that classification was appropriate, but one thought regression was not. One academician explains the use of classification methods in the context of identifying prospects' potential loyalty and disloyalty:

> "WM tools help identify the profile of current loyal and disloyal customers. Based on that, it is possible, for every new prospect, to compare his profile to those of loyal and disloyal customers. This tells whether the prospect may behave loyally or disloyally and helps manage that prospect accordingly."

One respondent indicated that additional axes such as cross-selling provide richer insight into the prospect and help enhance the web environment accordingly through customization for instance. However, one respondent advocated a more proactive and efficient approach to the use of WM tools for customization purposes by stating that the methods mentioned come from the 70s and that currently neural networks are more widely used because they determine in real-time and for variable timeframes the moves and behaviors of prospects:

> "There is no researcher entering one-by-one the variables into the system to generate a model/algorithm. Rather, systems are entirely automated and do not require human intervention anymore."

In that case, systems transmit the information about prospects and their behaviors in a mechanic fashion. The model predicts, classifies, etc. automatically and that knowledge can be accessed immediately. It is not very clear however to what extent newly discovered techniques such as neural networks do completely replace the older ones such as discriminant analysis for classification for instance and if they replace all other techniques at all. Each technique and its variants has a different purposes or accommodates for different types of data inputs. Is it that such techniques can provide the output usually generated by all other techniques by integrating all different types and kinds of data inputs or are they just other types of WM tools that are better suited to real-time analysis but by producing exactly the same output as classic WM techniques? In

fact, "on-the-fly" techniques do not leverage more and different kinds of outputs than traditional techniques although they may be better suited to particular situations. They just differ in the way they produce those outputs: in real-time and not post hoc as do traditional tools. At any rate, it appears that such dynamic approaches are now more common in web contexts than the older back-office approaches. They become central in the WM analytical process as well as in the aCRM framework. Other techniques will always be useful as well depending on the specific analytical situation (small samples, small homogeneity, etc.). This should be carefully balanced by the web administration team.

Contrarily to the question on identifying existing web customers' behavioral patterns, no respondent indicated the recourse to traditional online and/or offline surveys to discover more about prospects' psychological dimensions of behavior. This may be useful as well. Nor did any respondent indicate to use descriptive statistics to discover behavioral paths of prospects. Are both approaches less desirable in the case of discovering the behaviors of prospects or is this the result of respondents' oblivion? While it has been seen before that descriptive statistics are less suitable to discover dynamic paths such as behavioral patterns, the use of surveys may also be less feasible in a prospective perspective since very few is known about prospects. Surveys are already distasteful to web users that may be current and loyal consumers, submitting surveys to users who have never made any purchase on the website may discourage them to do so, give them a bad image or a bad feeling about the company, hence be even more repulsive to them than to existing customers. Both traditional surveys and descriptive statistics may therefore be less considered in the case of prospects. Less shall be known about prospects' consumer e-behavior, unless they provide deliberately such information through other channels such as social networks for instance where people provide usually very personal information about themselves regarding their motivations, perceptions, lifestyles and any other consumer behavior variables. Such knowledge may be directly usable to understand prospects' psychologically-related behavior or may be mined to build consistent behaviors. Such applications as sentiment analysis and opinion mining are one of many practices that typically allow to understanding time-related aspects of consumer behavior as well as their evolution through time (Nasraoui & Mobasher, 2011).

Table 14 below provides validation of RQ 9 and its research propositions with summarized answers to the research question. RP1 is validated but RP2, isn't validated. Consequently RQ 9 stipulating that WM methods and techniques enable to identify prospective web customers' behaviour on the internet is partially validated.

Table 4.10

Validation of Research Question 9.

RESEARCH QUESTIONS	RESEARCH PROPOSITIONS	VALIDITY OF THE RP	ANSWERS TO THE RQ
Component 4 : Usage of WM applied to the identification of prospective web customers patterns on the web			
RQ9: [HOW – EXTERNAL 4a] To what extent do clustering, association analysis, classification and prediction methods applied to web data identify prospective web customers' behavior on the internet?	**RP1:** Web data generated by prospective web customers highlight the particular browsing behavior of prospective web customers when they are navigating on the internet.	*Valid*	*Internal and external web data should be large, granular, issued by logged in web users, of good quality, and analyzed with WM to identify prospective web customers' segment-based browsing behavior. Traditional market research outputs and analyses should complement WM and related techniques of sentiment analysis and opinion mining to detect prospective web customers' consumer behavior*
	RP2: Clustering, association analysis, classification and prediction methods applied to web data provide descriptive and predictive modeling of prospective web customers' consumer behavior on the internet.	*Not valid*	

CONCLUSION – Behavioral data may be internal but need to be rather homogeneous, large, of good quality, generated on a well-structured website and generated by users who are requested to log in whatever their purchase status. When possible and if applicable, internal data need to be reinforced with external data such as environics (PRIZM, Mosaic, etc.), credit card information and social networks data (Facebook, etc.). The latter are of special value since they provide insightful content into consumer psychological behavior and not only browsing behavior as do other data types. Through advanced WM applications such as opinion mining or sentiment analysis they do in fact allow partially or totally to grasp prospects' elements of consumer behavior, *i.e.,* attitudes, perceptions, and the like. Websites can leverage such data as well if they are designed to capture content from web users (forums, blogs, micro blogs, etc.). In fact, every website requiring customer to produce content be it in written, video or audio formats, such websites may be able to draw highly specific prospects' psychologically-related behavior in addition to pure browsing behavior. Behavioral browsing patterns can be produced with traditional WM tools but appear to be increasingly detected automatically by means of more sophisticated WM tools that allow for real-time behavior detection and subsequent customization. Figure represents the summary of WM-enabled identification of prospects' browsing behavior.

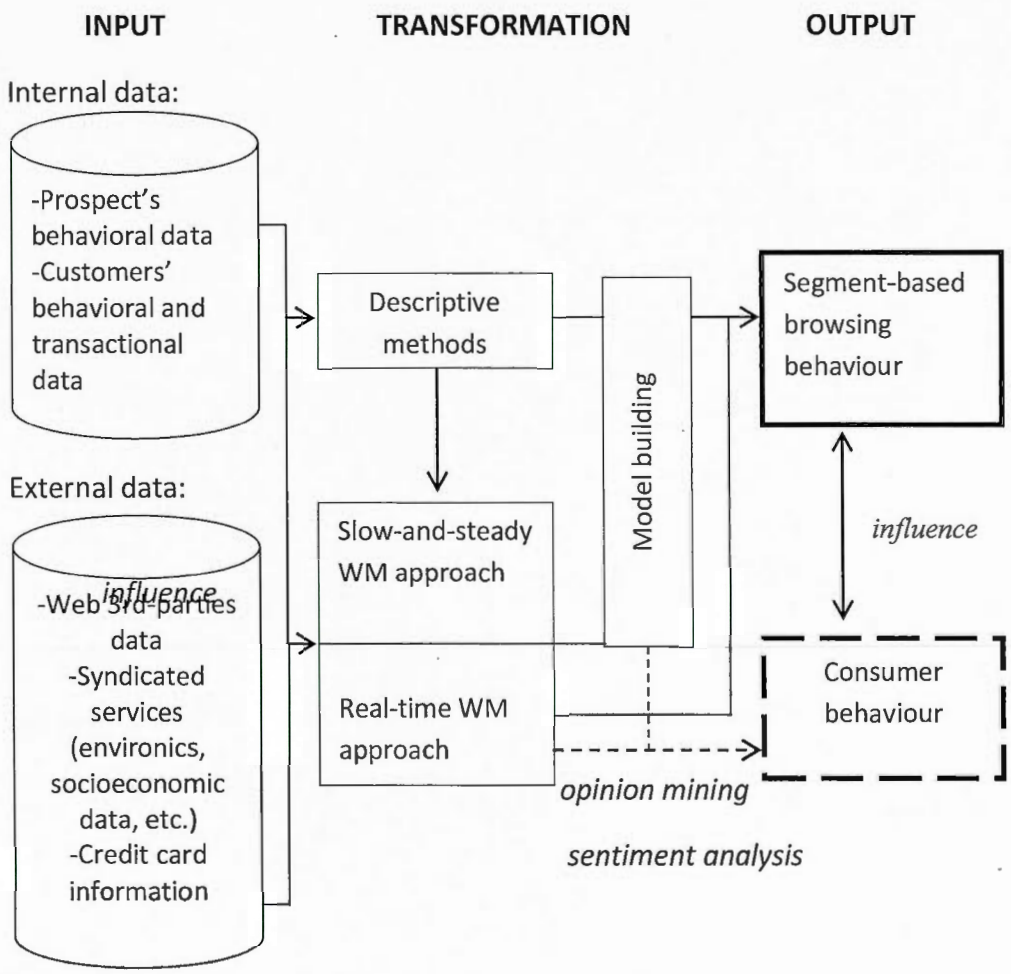

Figure 4.12 WM-enabled identification of prospects' browsing behavior.

4.6.2. Identifying how prospective customers defect to and from competitors' websites as well as how they remain loyal to competitors

Prospects are not (yet) regular customers of a website. Instead, they may be loyal customers of a few other websites while also gleaning information from multiple similar, competitive or disparate websites looking for cheaper prices, better products, etc. Some may also defect from competitors' website possibly because they saw better value proposition elsewhere. From a unique website's point of view it is of high interest to capture and understand the status of prospects with regards to competitive website and more interestingly to see how prospects actually move to and from competitors. If we take a company's website as a point of reference, by "defect to and from" it is meant that the prospect stops buying from one competitor's website to buy from another competitor's website but not on the company's website. By "remain loyal to" it is meant that the prospect buys on a specific basis on the competitor's website and not on that of the referent company's website. The point is to understand how prospects defect or remain loyal. By understanding how prospects move from one website to another and under which circumstances, a company might discover attributes that trigger loyalty or attrition behaviors and use that knowledge to craft knowledge-enabled strategies to make prospects defect from competitors and become totally or partially loyal to the company's website.

Thinking outside the website box, as suggested by a respondent in the context of identifying loyalty patterns of existing customers, should also be carried out on prospects. As expected, this appears to be the most difficult aCRM task to fulfill. Three out of ten respondents indicated that this is impossible to do and those who stated that it is possible reported many limitations.

The common sentiment of respondents is that it is already difficult to determine the loyalty status of an existing web customer on the whole spectrum of the websites that internet encompasses. It is even harder to track the behavior of that existing customer to understand how he/she became loyal or defected from a website. Chances are close to null when attempting to do the same for prospective customers. Not to mention that loyalty is defined in multiple ways across different websites and while one company considers a frequency of 1 purchase a year as a high score on the loyalty scale, another might consider it to be churn. Apart from metrics definition, the challenge lies in

following behaviors of prospects outside a controllable environment, which is the company's website. This cannot possibly be done with the sole data of one website because it is too limited a dataset and does not include complete customer behavior across the infinite web. A marketing director indicated that:

> "WM are not in cause, it is the data that is not available."

An academician shared that belief indicating however that the data is not unavailable but rare. Two respondents indicated that getting such knowledge may be achieved by traditional market research or third-party data, *i.e.,* ComScore. While it seems impossible to track prospects' behaviors and hence loyalty forming patterns towards competitors, it is nevertheless interesting to check whether prospects consider competitors in their decision process. This may be done by verifying whether they visit the company's website and then that of competitors or the opposite, at which frequency, etc.

Despite the difficulties, seven respondents indicated WM could help in identifying prospects' loyalty patterns in competitors' contexts. They do not provide clear-cut answers as to the why and the underlying motives and reasons of individuals on the web, but rather hints and clues as primary material to make inferences and deductions. For instance, one respondent indicated that the factor missing in the WM approach of this research is the concept of trends. Contents of websites must follow and adjust to the trends. If a website does not, then it might explain why prospects develop loyalty for competitors. They offered what they were looking for. One respondent advized supplementing internal company data with other data, methods and sources. Another reported log files are sufficient but it is difficult to access competitors' log file databases. Albeit not having access to browsing information of competitors' websites, three respondents indicated that a company can use web analytics (Google Analytics, etc.) and online statistics from third parties (AC Nielsen, etc.) in order to gather descriptive statistics about frequency rates of visits, number of additional or less customers on a competitors' website etc. Companies track and sell those data on which a company can build to increase visibility, positioning or increase visits of the website. One academician explains:

> "Many companies do not know their navigation information is freely and openly available to anyone. When opening a store on Sainte-Catherine Street in Montréal, you cannot hide it. The same holds for websites. The internet is an

> open space, if one opens a website on it, he/she cannot hide it and the data about the web traffic (number of people who enter the site, etc.) is collected by Google, mainly, and openly displayed online."

These are useful statistics but they only display superficial facts such as the additional number of customers on a competitor's website and they do not inform about why they defected or they remained loyal, especially if they come from the company's website or not. The behavioral path of prospects across the web is not tracked unless people are required to log in. In that case, the analysis is eased and more accurate but this is hard to implement across several (competitive) websites. Currently it may only be possible to put the stress on one website. Another respondent indicated that such information may be biased. At the end of the day a company only has a good insight of its own problematic but not into that of its competitors which remains highly strategic and sensitive information. Industrial spying remains the only viable option to get in-depth information.

In that vein, a respondent indicated that the problem of fulfilling that aCRM task is not transactional or methodological, but it is conceptual, ethical and moral. In order to identify defection and loyalty patterns accurately, one will have to access very confidential data. So, such data exist although not many people are aware of it. An academician explains:

> "Amazon or Google have huge platforms to track web patterns, so do banks and credit card companies because they own information about their customers financial moves. These data are thus highly confidential and only exploitable with WM tools by these companies. There may be restrictions and regulations about these kinds of information which would require customers' consent."

The recent rise of consumer privacy protection movements indicates that consumers are not willing to make such information public. It seems again that the data exists and the tools as well, but the system does not allow both elements to meet except in some restrictive instances.

Unless having access to highly confidential and private prospects' information, companies may not be able to identify totally how prospects build loyalty or defect to and from competitors' websites. To do so, the internet should evolve into a comprehensive and cohesive conglomerate where all e-commerce websites information is free-flowing

and openly available to anyone. Web traffic and behavioral patterns would be exchanged on virtual marketplaces freely or at a cost, web users would be required to register on it with their own personal log in and password to avoid many-users pitfalls. But then again to what extent may sensitive information be freely available, to whom, at which cost, for how long? Such questions would remain and underline the conceptual issue which is at the core of the problem. Some giant companies such as Procter & Gamble (P&G) encompassing different and often competitive brands, divisions, products/services, may be able to develop such types of analyses. However, looking into defection and loyalty development in a web context may only tell about cannibalization and not necessarily competition from external companies.

Table 15 below provides validation of RQ 10 and its research propositions with summarized answers to the research question. RP1 is partially validated but RP2, isn't validated. Consequently RQ 10 stipulating that WM methods and techniques enable to identify prospective web customers' behaviour on the internet is partially validated.

Table 4.11

Validation of Research Question 10

RESEARCH QUESTIONS	RESEARCH PROPOSITIONS	VALIDITY OF THE RP	ANSWERS TO THE RQ
Component 4 : Usage of WM applied to the identification of prospective web customers patterns on the web			
RQ10: [HOW – EXTERNAL 4b] To what extent do clustering, association analysis classification and prediction methods applied to web data identify how	**RP1:** Web data generated by prospective web customers highlight prospective customers' defection patterns to and from competitors	*Partially valid*	*Internal and external web data should be large, granular, issued by logged in web users, of good quality, and analyzed with WM to identify prospective existing web customers' factual moves on a given website but also to and from competitors but do*

prospective web customers defect to and from competitors as well as how they are loyal to competitors on the internet?	**RP2:** Clustering, association analysis, classification and prediction methods applied to web data identify prospective web customers' loyalty patterns to competitors, on the internet.	*Not valid*	*not identify underlying reasons and motives for doing so, that is their consumer behavior as well as their loyalty patterns to competitors. Traditional market research outputs and analyses should complement WM to detect prospective web customers' consumer behavior and their loyalty patterns to competitors. Additional data types are also needed to track prospects behaviors across multiple websites*

CONCLUSION – Identifying prospects' defection and loyalty patterns with regards to competitors is one of the hardest if not the hardest task to fulfill. Currently, web managers do not have access to all competitors' data in order to develop an accurate analysis. Internal types of data such as browsing information ideally resulting from logged in users and market research conducted by the website owners may be combined with, compared to or cross-tabulated with additional third-party data, and/or web analytics tools that record prospects' activities on diverse website across the internet in order to determine how they develop loyalty or attrition. Based on that companies can triangulate and infer conclusions but without absolute certainty. Access to more confidential data such as financial transactional purchase would close the gap by providing highly strategic information on a customers' behavior on the website in transactional perspectives. However, underlying motives and reasons are still not grasped even with those methods which are also debatable as regards to their intrusiveness and pervasiveness.

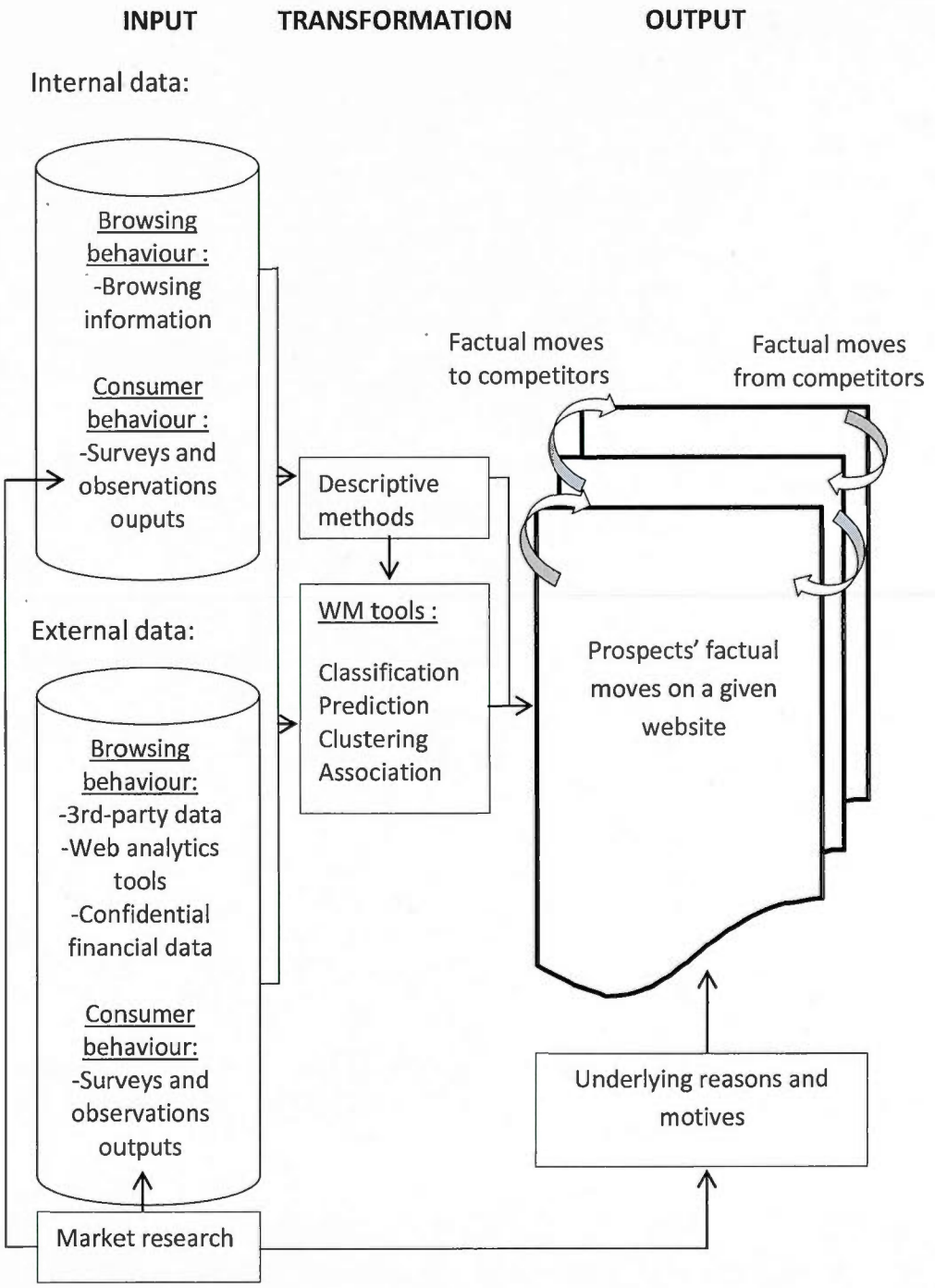

Figure 4.13 WM-enabled identification of prospects' loyalty patterns to and from competitors' websites.

4.6.3. Conclusion on WM-enabled identification of prospective web customers' behavior on one or more website(s)

Some visitors are prospects because they are not customers of the website. It may thus be reasonably well-assumed that they are prospects because they are customers of other competitive websites. Identifying prospects' behaviour is typically more difficult to do than for existing web customers. The latter are more easily identified and tracked because they provide more information about themselves and usually log in automatically or manually on the website. However, prospects might also register and log in on a website. Their resulting browsing behaviours are the most accurate and useful types of data that can be used in a rigorous WM process. Ideally, input data should therefore originate from logged in prospects Customers' transaction data can also be used just like it was done for profiling. External datasets such as web 3rd parties data, syndicated services data or credit card information come complement internal data types.

As depicted in Figure, WM and, to a lesser extent, descriptive methods, are then used on these data types in order to generate useful models and outputs to understand the prospects' behaviours. Unless logged in, prospects' behaviour is better estimated in terms of segments of prospects. To each segment corresponds a behaviour type. As for existing customers though, behaviour only refers to the browsing behaviour. We understand how people navigate and use the web, hence how they behave on the website but not the underlying affective and conative dimensions that trigger and guide the visit. Psychological and emotional aspects which are termed "consumer behaviour" in figure 23, can only be discovered by means of online or offline market research (surveys) and subsequent statistical analysis. However, recent advances in opinion mining and sentiment analysis[35] enable to understand partial aspects of consumer behaviour such as the prospects' attitude which can be his/her judgement or evaluation, affective or emotional state. These techniques help determining the polarity of a prospect-provided content (product reviews, ratings, recommendations, etc.) as positive, negative or neutral. More sophisticated sentiment classification (beyond polarity) seeks to identify more sentimental and emotional states (Snyder & Barzilay, 2007). Businesses can use these techniques to market their products, identify new opportunities and manage their reputations.

---

[35] Both expressions refer to the use of text analytics, computational linguistics and natural language processing (NLP) to identify and extract subjective information in source materials.

Consumer behaviour and browsing behaviour influence each other. As for existing customers, it should be understood whether prospects tend to display a specific e-consumer behaviour or transpose their real-life consumer behaviour to the web context, or if it is just a blending of both aspects, the prospect might well transpose his/her consumer behaviour while adjusting few of its component variables. Identifying behaviours of prospects is already a difficult task however it has been sought to discover if WM was useful to also discover their behaviour on competitors' websites. However, WM is not useful to understand how prospects defect to or from competitors' websites. The difficulty resides in the fact that moves and patterns across several websites need to be captured and available to the business. Web analytics may provide factual statistics relating to prospects paths but the browsing behaviour on competitors' website cannot be tracked since a business does not have access to its competitors' data. The same data input as for the analysis of one unique website can be used. In addition and if applicable, additional data sources such as financial data that are typically held by credit card companies can be useful complements because they track the purchasing activities of prospects across the whole web and not only one website. It should be investigated further to what extent access to these type of information may still be legal and how such data types might provide additional insight into how prospects move to and from competitors.

Both WM approaches provide factual information about prospects' loyalty patterns across similar or different competitors' website. In fact, analyses based on web analytics data provide insight into facts, the number of entries or clicks on a website, the number of pages viewed, etc. Correlations, path analysis, regression or clustering analysis are applicable to such data types. However, this does not tell anything about how prospects' loyalty patterns towards and across several websites. Unless computer screens may be filmed internally and output may be stored and processed analytically, it is not possible to discover the true evolution of prospects among diverse website. Even so underlying reasons and motives remain unknown. Market research remains an essential although burdensome component to grasp the depth and scope of prospects' motivations as well as satisfaction and loyalty levels across different website. Asking people may incur error rates but correlating survey outputs with filmed web sessions, one might ultimately be able to draw an accurate picture of prospects loyalty to, on and across multiple websites.

A final simulated real-life case that illustrates the fourth theme of identifying prospective web customers' behaviour on one or more website(s), is provided below:

> *PluggedIn is a social website that connects job-seekers and employers. It offers variable job-seeker accounts which encompass such services as live chat with employers, tailored job ads, job market newsletters, firms' contact information, interest group discussions, and other activities to make the job market more efficient and promote professional networking.*
>
> *Top management noticed that although many users visit the website and have a profile, they do not buy accounts and thus remain prospects. It is assumed that most of these prospects already have an account on bigger, more well-established social work-related websites and hence use PluggedIn out of curiosity among the many other mushrooming social networks on the internet. The business would like to know more about how customers develop loyalty or defection towards competitive websites. Two other competitors' websites were chosen for the analysis. PluggedIn has a vast amount of browsing and transactional history to track prospects' profiles on the website, because they are required to log in and all their actions are recorded and displayed on the website so that others see what a particular visitor did. This helps a great deal to understand the browsing behaviour of segments of prospects. WM techniques such as prediction and classification reveal which prospects are the most likely to buy an account and for how long as well as those who are least likely to buy and defect quickly. For each aspect of the browsing behaviour segments are created. For instance, there are segments for usage of live chat (heavy, medium, low users) or for group discussion participation (shouter, chatter, quiet, mute), etc. WM is also used to understand aspects of consumer behaviour. Since prospects can express their opinions and meanings on an infinite variety of work-related subjects, sentiment analysis is conducted in order to see the intensity of their expression and further to derive the polarity of their comments, reviews, status updates, etc. Opinion mining reveals that overall prospects feel strongly affected by the slow recovery of the financial crisis that hit in 2008-9. Some lost their jobs, many are graduates and delayed their entry on the workplace, others started their own business. Overall the sentiment is rather negative and this may explain why few prospects convert to customers on PluggedIn.*

Although it appears the website has a strong potential customer base, it would like to know if prospects are also reluctant to buy an account because they already subscribed to other websites such as the market leader, which pioneered the concept and enjoys the first-mover advantage. Hopefully, this knowledge will provide hints as to the right things to do in order to attract and retain these prospects. PluggedIn has limited access to competitors' websites' browsing data and transactional information but it collects a substantial amount of web analytics. Compared to PluggedIn's, these data reveal that prospects spend more time on a competitor' web page than on theirs, path analysis reveals also that prospects come from more diversified websites which have competitors' links on their web page. Usually, prospects go to competitors' websites first and then come directly to PluggedIn. Visitors tend to visit all their different professional social networks in a row to check for news or updates but spend less time on PluggedIn. From these factual loyalty patterns across the 3 different websites, PluggedIn managers infer that the lack of differentiation of their website might be in cause. People will always prefer the pioneer market leader instead of the copycat. Consumer behaviour analysis previously also revealed that prospects generally struggled financially and although highly motivated to find a job, they couldn't afford subscribing an account on a regular basis. These are the underlying motives and reasons combined to consumer behaviour which explain the factual loyalty patterns observed among the 3 different websites. This knowledge drove management to focus on the existing customer base through up-selling, selling more higher-priced services. Market research was conducted among prospects to identify the maximum amount they were ready to pay for an account. A more appropriate pricing strategy was implemented with variations from one country to another. A financial viability analysis also revealed that it was financially sound to price the account 10% lower than that of competitors. An aggressive marketing campaign would attract at least 5% of the competitors' customer base, to convert them into loyal and regular consumers. More partnerships were also developed with more diversified websites to put PluggedIn's logo link on their websites to increase the number of trails that lead users to PluggedIn. Additional creative features and options were added to make the website more differentiated from its competitors. It positioned itself as being a

place of choice to promote, support and enable entrepreneurship and personal initiatives.

The fourth theme corresponding to the fourth meta-objective of identifying prospective web customers' behaviour on one or more website(s), as identified in Xu and Walton's (2005) adjusted framework, can be fulfilled by using WM. Traditional descriptive statistics are useful complementary tools in that respect. As for existing web customers, market research, be it done in-house or out-house, provides the missing link of prospects' "consumer behaviour" information as well as their underlying reasons and motives for navigating on a given website and moving from one website to another.

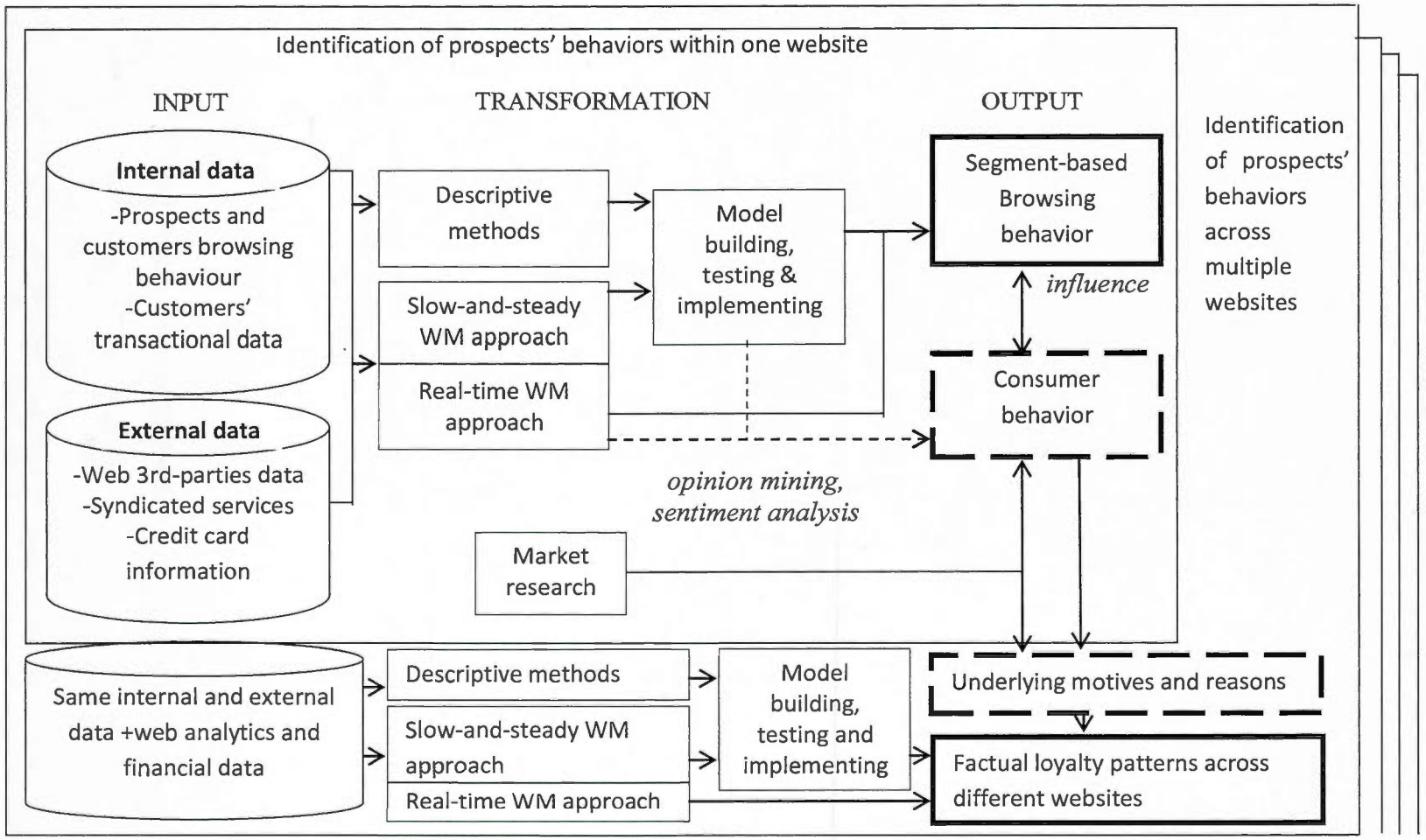

Figure 4.14 WM-enabled identification of prospects' loyalty patterns across the internet.

177

PART FIVE

DISCUSSIONS OF THE RESULTS

5.1. Review of the potentialities of WM-enabled aCRM in a KM perspective

Figure shows Xu and Walton' (2005) aCRM for a web users' knowledge acquisition framework, as depicted in the literature review section. The four quadrants refer to the four themes that were investigated previously. The top left quadrant refers to the profiling of existing web customers theme (theme 1). The top right quadrant refers to the identification of existing web customers' web behavior theme (theme 2). The bottom left quadrant refers to the profiling of prospective web customers theme (theme 3). Finally, the bottom right quadrant refers to the identification of prospective web customers' web behavior theme (theme 4).

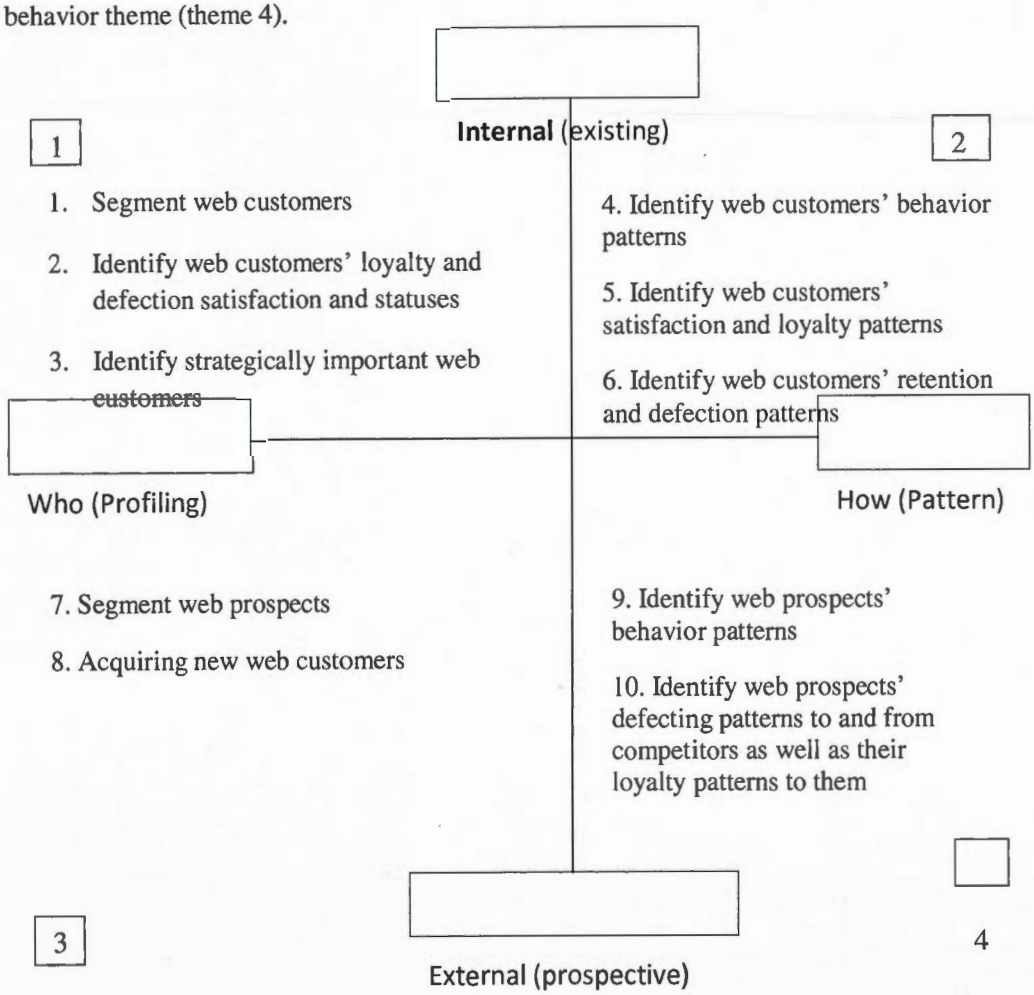

Figure 5.1 aCRM for a web users' knowledge acquisition framework.

The following 4 sub-sections determine for each of these 4 meta-objectives (themes) respectively the outputs that can be generated by means of WM in order to reach the objectives of each meta-objective. They strictly focus on the outputs that can be generated by means of WM alone without surveys and/or observations, based on the exploratory findings. The following provides a good estimation of the current capabilities in the field of DM as applied to web data in order to generate useful information and ultimately knowledge and allow businesses to build stronger and more profitable customer relationships.

5.1.1. WM-enabled profiling of existing web customers

Figure 3 depicts the potentialities of WM to reach the first aCRM meta-objective (theme) of profiling existing web customers. Depending on the level of segmentation sought by the company and, therefore, the strategic marketing orientation resulting from the targeting process, WM allows to determining the loyalty status (defection vs. Loyalty), the profile based on a multitude of attributes as well as the strategic importance of existing web customers (unprofitable vs. profitable). WM enables thus to fulfill the 3 first objectives of aCRM. Knowledge about these 3 elements is useful to develop business-specific profiles ranging on a continuum from low to high, given the segmentation level that is sought. The profile of a given web customer can thus be used to predict that customers' future loyalty, but also future profitability to develop dynamic response frameworks in real-time and appropriate to the profile of the customer.

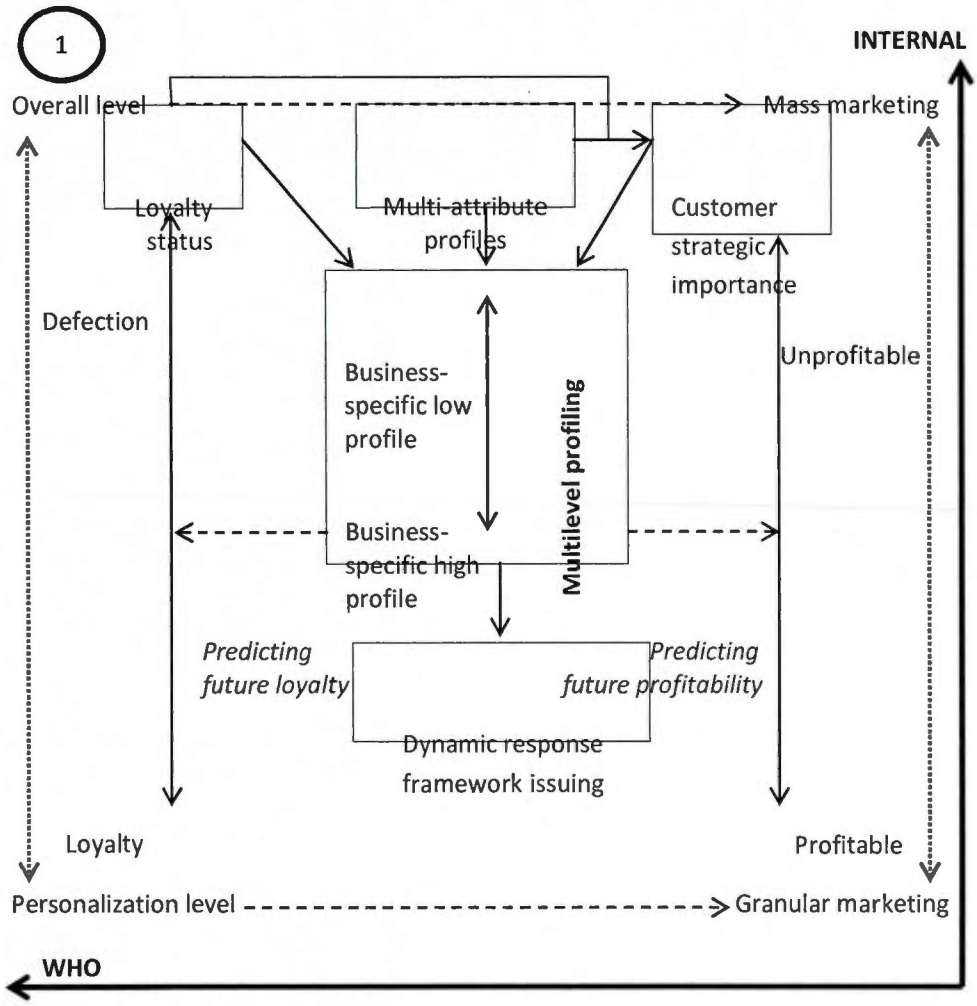

Figure 3.2 WM-enabled profiling of existing web customers [WHO – INTERNAL DYAD]

5.1.2. WM-enabled identification of existing web customers' behaviours

Figure depicts the potentialities of WM to reach the second aCRM meta-objective (theme) of identifying existing web customers' behaviors. WM is well-suited to track operational and transactional aspects of customers' behavior online but not underlying reasons, motives or preferences. Market research is better suited to discover these psychological and emotional facets of online customers' behaviors. WM enables thus to fulfill partially the fourth objective of aCRM. In fact, it cannot determine the consumer behavior variables of customers with much acuracy; it can however identify their

browsing behavior. Also, WM can only partially fulfill the fifth objective, namely identification of satisfaction and loyalty development patterns. In fact, it can only identify the transactional loyalty development but not the satisfaction development and the emotional loyalty development. Eventually, it cannot fulfill the sixth objective of aCRM, namely identifying retention and defection patterns of web customers that is their patronizing of a web business.

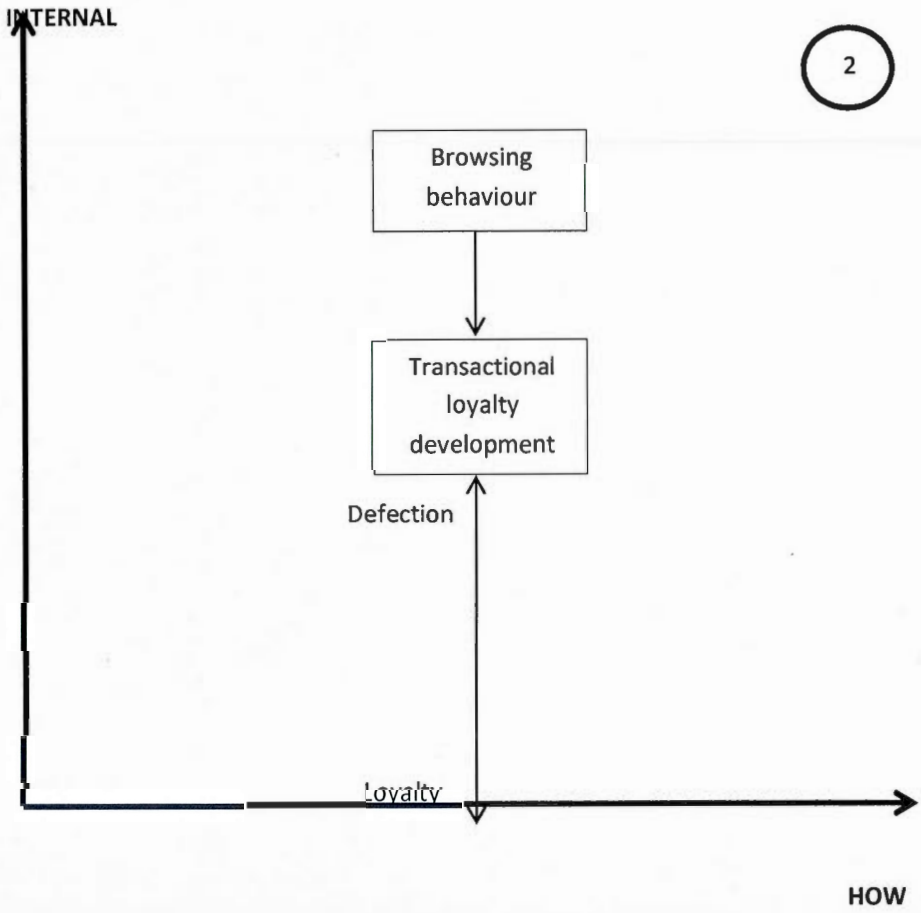

Figure 5.3 WM-enabled identification of customers' behaviors [HOW – INTERNAL DYAD]

5.1.3. WM-enabled profiling of prospective web customers

Figure depicts the potentialities of WM to reach the third aCRM meta-objective (theme) of profiling prospective web customers on the internet. WM is well-suited to fulfill the seventh objective of segmenting prospective web customers as well as the eight objective of discovering prospects' needs, wants, habits, preferences and interests. Depending on the segmentation level sought, web businesses can combine segmentation attributes, needs, etc., to predict potential loyalty statuses of prospects on the business-specific loyalty continuum (loyal vs. defective) as well as their potential profitability on the business-specific profitability continuum (profitable vs. unprofitable). These attributes and knowledge on prospects enable businesses to draw business-specific profile (profiling) and craft profile-specific eMarketing acquisition strategies in accordance with the pre-defined segmentation level. These strategies, especially web advertising may have short term effects on the bottom-line or more long-term effects such as branding (brand status and strength) as well as increase in share of hearts or share of minds.

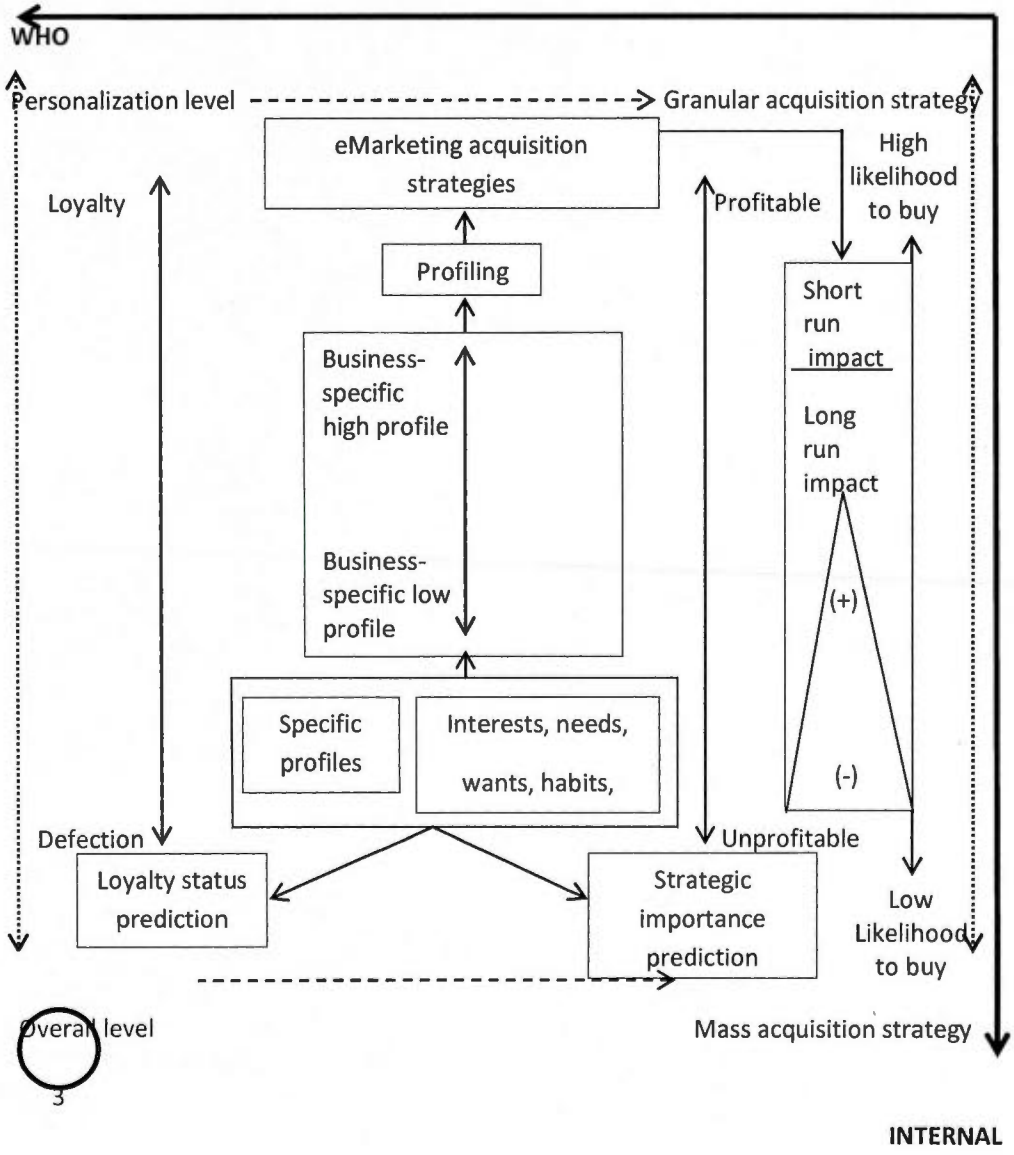

Figure 5.4 WM-enabled profiling of prospective web customers [WHO-EXTERNAL DYAD]

5.1.4. WM-enabled identification of prospective web customers' behaviours

Figure depicts WM potentialities to reach the fourth and final aCRM meta-objective (theme) of identifying prospective web customers' behaviors. WM is well-suited to track operational and transactional aspects of prospects' behaviors online and therefore their browsing behavior, to derive segments. However, underlying reasons, motives or

preferences, *i.e.*, consumer behavior, are better investigated by means of market research, which unveils psychological and emotional facets of prospects' behaviors. WM enables thus to fulfill partially the fourth objective of aCRM. In fact, while, it cannot determine the "consumer behavior" variables of prospects with much accuracy, it can however identify their browsing behavior. Moving beyond the framework of one unique website, WM can also only partially fulfill the objectives of identifying prospective web customers' defection patterns to and from competitors as well as their loyalty patterns to these (dis)similar competitors, across the internet. Actually, WM is well-suited to identify factual loyalty patterns across websites but is relatively ineffective in determining the underlying reasons and motives that drive and justify these moves on the web. Market research should be used in support of WM in that respect. Therefore, WM can only partially fulfill the tenth, eleventh, and twelfth aCRM objective.

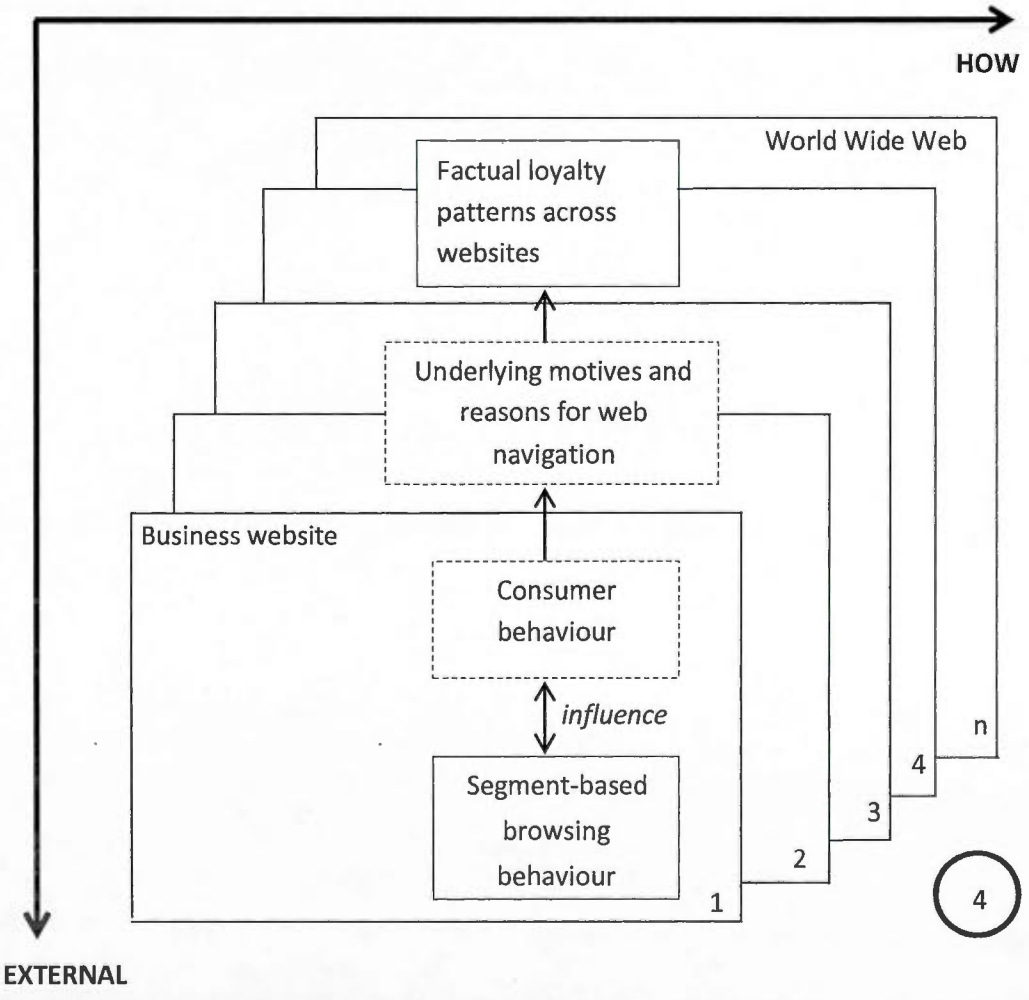

Figure 5.5 WM-enabled identification of prospects' loyalty patterns across the web [HOW – EXTERNAL DYAD]

By putting all 4 quadrants together we can integrate them to Xu and Walton's (2005) adjusted framework. Figure represents the current possibilities regarding WM-enabled achievement of aCRM generic objectives in a KM perspective.

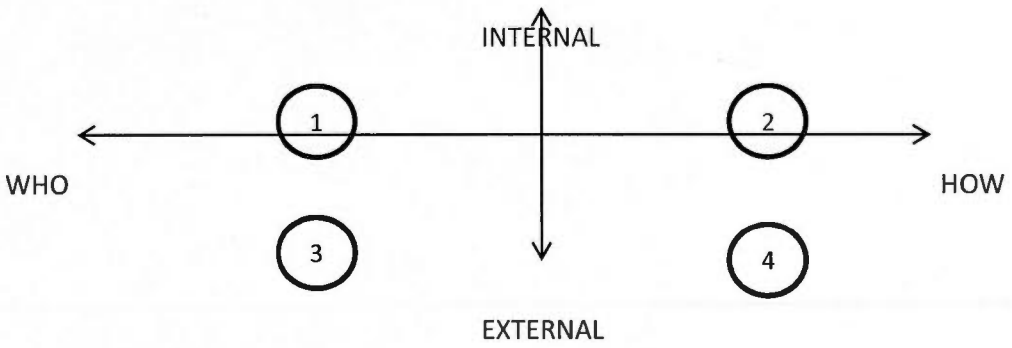

Figure 5.6 WM-enabled achievement of aCRM objectives in a KM perspective

5.2. Managerial implications

This study sought to dig in the tremendous potentialities of WM as integrated to the blended aCRM-KM framework to leverage insightful knowledge about existing and prospective web customers who interact with online businesses, especially e-commerce portals. The research project was thus geared toward investigating the benefits of WM methods and techniques on the analytical CRM (aCRM) applet of the marketing function, in the framework of a Knowledge Management (KM) perspective.

In the light of this study, it is very important for managers and practitioners to understand that web-mining is not an end in itself. It is a combination of varied tools and techniques that enable experienced teams to discover useful meaning out of big structured or unstructured data. Managers should be aware of the garbage in-garbage out trap, as all information is not necessarily good to process. Data produced by logged in customers appear to be the best kind of data since navigation patterns can be precisely attached to a specific customer. Companies' databases should also allow for high volumes of data entries, since higher quantities of data leverage better results.

Research findings reasserted that WM is a tool among many others to be used for an infinite number of projects or practical applications. While WM is very well-suited to discover the profiles of existing as well as prospective web customers of an e-commerce website, these methods and techniques are of relatively limited use to dig further in-depth the specific behavior of web users. WM can momentarily only process those tangible, operational, transactional data which refer to browsing history, clickstreams, logging history, transactional activities, etc. Consequently, it is an extremely powerful approach to analyze navigation patterns, web paths, uncover hidden or a priori non-obvious relationships in the web data. It informs thus well about the specific browsing behavior and factual loyalty patterns development of both existing and prospective web customers.

Nevertheless, WM suffers serious limitations as regards to the discovery of underlying reasons, motivations, preferences, and more broadly all the variables referring to the so-called "consumer behavior" in marketing, *i.e.*, internal and external influences that impact the web browsing behavior of both existing and prospective web customers of an e-commerce website. Additional market research conducted either internally by the web business or externally by third-party marketing agencies, consulting firms or other types of syndicated services remains a cornerstone for understanding emotional, psychological and even psychographical aspects of web users navigating on one website but also across two or more websites. It should be added though that market research is extremely difficult to implement in a web context since participation to web-based surveys are extremely low (Malhotra, 2010). Also, bothering current customers with compelling web surveys might create annoyance, irritation, lower levels of satisfaction and ultimately disloyalty and increased churn, not to mention the resulting bad publicity through eWOM or traditional WOM.

This does not mean that WM is to be deemed unable to evolve to grasp these consumer behaviors. Actually there are very promising avenues for future development in that respect. Opinion mining and overall sentiment analysis laid the foundations for determining the polarity or direction of web users as well as the magnitude of their opinions, thoughts, etc. informing about their perceptions, their likes, dislikes, etc. Semantic analysis, Natural Language Processing and other Artificial Intelligence techniques which focus on the qualitative aspects of web content rather than on core quantitative and structured data are to become increasingly more integrated to the traditional WM techniques in order to close the gap on consumer behavior. Liu's (2007)

Bible on Web-Mining, written in 2007, did almost have nothing about opinion mining and sentiment analysis. In 2011, that is, only 4 years later, Bamshad Mobasher and Olfa Nasraoui, two eminent researchers in the Web-Mining sciences, dedicated a full chapter to the potentialities of these techniques especially with the rise of social media and networks embedded in the web 2.0. It may reasonably be assumed that in a couple of years, these more qualitatively-oriented techniques, applied to big, unstructured, qualitative data, will grow in importance and in practical usage calling for further investigation fields in the discipline of WM. Research avenues are definitely plentiful in that area.

In the specific context of marketing, this is great news. Academic research evolution coupled to tremendous technological advances in hardware, software as well as server and web data hosting capacities mean that additional data types will be processed, more often, more efficiently and will yield more insightful results. Brick-and-mortar businesse managers can use web knowledge to devise smarter web strategies to better acquire, develop, and keep their web customers be they already offline customers or not. Enhanced websites that offer more intuitive and high-level web experiences will also increase the flow of web users, and eventually their satisfaction and ultimately their loyalty, increasing both bottom lines through increased share of the wallet, optimized portfolio building and branding capacities. Not only, will the web become a powerful customer puller by converting prospects into buyers, but it will also nurture current customers' loyalty and comfort them with their patronizing of a specific web entity. This is exactly the ultimate goal of CRM and, if well-applied and deployed, WM is in congruence with CRM. For pure-players, whose business model relies entirely on web activities, it is even more crucial that they use WM to develop powerful relationship capital and knowledge capital to develop their customer base. This holds for both B2B and B2C contexts.

The devil is in the details for it is mainly in the WM project planning and the quality and availability of the data as well as the robustness of the WM analytical process that most benefits will be derived from WM. This is a statement that came back on a regular basis. in the exploratory interviews. Regarding the availability of the data, it remains unclear to what extent highly specific data such as credit card information, and other financial activities data streams which show web users' factual loyalty, defection and overall patronizing patterns on the internet are available. While not encompassing the emotional

and psychological aspects of prospective or existing web customers, they do tell a lot about their preferences, likes, dislikes, etc., at least in transactional terms. This is already quite much for CRM which is primarily transaction-focused. Access to these types of information may be very restrictive, very costly, if not legally forbidden. Also, ethical and moral issues arise as not everything can or should be mined irrespectively of web users' privacy and safety. WM and DM are often under the flashlights for inquiring subversively into people's private life without them knowing about these activities. Business ethics research streams brought the matter on the academia table but not much is known or wanted to be known as to the actual limitations of WM/DM usage. This is another area of interest to be investigated.

5.3. Limitations and research avenues

This research is very broadly-defined it took the 4 major WM methods to determine their effective use for reaching a generic taxonomy of 12 aCRM objectives as determined by Xu and Walton (2005). A number of limitations are worth being noticed and should give rise to future research avenues.

First, it has been seen that there is a slow-and-steady approach to WM which implies a traditional analytical approach with data collection, pre-processing, analysis and model development and deployment. Another approach, the so-called "real-time WM approach", tends toward the automation objective as devised in most BI frameworks. Data issued by web users is processed on the spot and the user's data is used instantaneously to develop his/her profile, categorize him/her and offer him/her personalized content, recommendation services, intelligent adaptation, morphing or augmented reality capacities. It seems, however, that the slow-and-steady approach is necessary to build the model in the first place and the real-time approach may be preferred afterwards once the model is developed in order to automatically process and store new data streams and produce the desired outputs. As a matter of fact, it is not clear to what extent each and every WM method, or technique, enables automatic creation of models on the spot to bypass the thorough slow-and-steady approach. Additional research should estimate to what extent WM methods truly enable automatic model creation and implementation-deployment directly on the website, without human interaction.

Second, not each method has been investigated in-depth in order to see its specific contribution to reach each objective. In fact, some discrepancies may arise as some

techniques might be better suited to reach a wider range of objectives as some others. Also, the aCRM framework used is not unique, other types of aCRM objectives can be thought of and these have not been taken into account in that research project. Therefore, additional research could focus on investigating the relative benefits yielded by one specific method and its number of associated techniques for reaching Xu and Walton' (2005) generic taxonomy of aCRM objectives adjusted to the web context or other more web-specific categorizations of aCRM objectives. Also, only supervised and non-supervised WM methods and techniques, have been considered. However, the literature also indicates that there are a bunch of semi-supervised learning methods and techniques, mainly a combination of supervised and non-supervised approaches. Their usefulness could also be investigated as regards to fulfilling aCRM objectives in a web context.

Third, this study focused entirely on WM, *i.e.*, DM applied to an online data stream to fulfill aCRM objectives. However, in practice, most brick-and-clicks will make a blended usage of both web data and other more offline data types to answer aCRM problematic. A company will typically use its call center outputs, command entry systems, POS terminals, database leads, etc (Tiwana, 2001). Clickstream data, customer interaction logs, and other web data only represent an infinitesimal portion of the overall data flow that is used in aCRM, in top of the core off-line issued data. A hybridization of both data types to reach aCRM objectives is beyond the scope of this research but could be investigated, once both online and offline frameworks are validated theoretically and empirically. This may represent an ambitious project because of the high number of independents and contingents as well as the many relationships that are to be investigated.

Fourth, Data Warehouses (DWs) host the data to be mined with Data-Mining and other modeling tools such as scoring, forecasting, and more descriptive reporting systems, query tools and analytical applications (operational, tactic and strategic) (Kimball and Ross, 2004; 2008). The rise of the internet gave way to another important data repository which has kept on gaining in popularity for the last decade, the Web House (WH). WM is to become to the WH what DM is to the DW. It remains unclear to what extent businesses will use both either separately or together in an integrative orientation as BI tends to tell businesses to do so. There is a data convergence for increased data processing effectiveness and efficiency. However, if the WH and WM are to become separate entities from the DW, then it is of utmost importance to develop a new stream of research dedicated to the WH-WM bloc and which sheds light on the beneficial aspects of such a

structure for marketing. This is especially true for brick-and-clicks which operate both offline and, to variable extents, online.

Fifth, the sample surveyed in this research project is very limited. In fact, only 11 respondents were surveyed through in-depth interviews. Strength of the research tough is that these respondents are not all located in Canada but also in the USA and Europe providing a richer and broader view on WM capacities as divergences may exist in WM usage between these different geographic regions. Nevertheless, another research with more respondents could be undertaken and results can be compared to identify discrepancies. The ratio business practitioners and scholars should be kept as close to 1 as possible since both types of experts have different views on WM usage, one typically practically-oriented while the other is far more theoretically-oriented, obviously both should be mutually complementary and enforcing.

Sixth, both the browsing behavior and the consumer behavior (motivations, perceptions, etc.) of either an existing or a prospective web customer influence each other. A colorful web page might diffuse warmth and joy, and hopefully increase the web user's fun and experience. Over time, positive emotions resulting from the nice website navigation will act as an instrumental conditioner so that whenever the web user sees the website logo, name or the like, he/she will remember the nice experience and develop intentions to go back there to feel the positive emotions again. However, should the web user be in an extremely bad mood, this might affect his web browsing behavior on this same website, despite the engaging user experience. Consequently, it remains unclear to what extent the typical customer behavior displayed in offline contexts, *i.e.,* everywhere but on the internet channels, differs from the browsing behavior displayed on the internet channels and even on different types of internet channels. Is it merely transposable? Or is there such thing as a separate e-consumer behavior that an individual switches on when navigating on the internet? A blend of the two previous approaches seems more likely but additional research is needed on that subject to identify those elements that tend to change most once on the internet, to what extent, magnitude or direction and most interestingly, could WM be appropriate to track such changes?

Finally, although both WM approaches (slow-and-steady and real-time) apply to identify segment-based browsing behaviour, it remains unclear to what extent a real-time approach applies to sentiment analysis in order to detect consumer behaviour. Sentiment

analysis builds on techniques that go beyond traditional WM applications. It uses machine learning such as latent semantic analysis, support vector machines or bag of words (Turney, 2002). They do semantic analyses which require human intervention starting with model building, refinement, trial-and-error until deployment. It appears that some open source software automate sentiment analysis on large collections of texts, *e.g.,* online news, discussion groups, online reviews, blogs and social media (Dey & Mirajul Haque, 2008). Nevertheless, in line with Wright's (2009) managerial suggestion to mine the web for feelings and not for facts, additional research should determine to what extent sentiment analysis is useful for marketing to automate processes of sifting through the noise, understanding conversations, predicting future developments, identifying relevant content as well as the most influential opinion holders. Business sciences and especially marketing will always be needed in order to leverage that knowledge appropriately, and for optimal returns.

CONCLUSION

CONCLUSION

This study sought to investigate the integration of Knowledge Discovery in the Database (KDD), epitomized by WM methods and techniques, to the marketing function, and more specifically the analytical Customer Relationship Management (aCRM) applet of marketing, embedded in a Knowledge Management (KM) perspective. The innovativeness of that study called for an exploratory research design focused on developing research propositions that could be investigated further.

Given the current potentialities of WM, the availability and the quality of the data, only 12 out of the 20 initial research propositions were validated for further investigation, 5 were partially validated and 3 were not validated. The relationships identified in 5 models out of the 10 (referring to 5 research questions out of 10) presented can be tested empirically for external validity and generalizability. These models refer primarily to the aCRM meta-objectives (themes) of profiling, respectively, existing web customers and prospective web customers both in real-time or in a slow-and-steady fashion. The remaining 5 models referring to the 5 other research questions appeared to be partially valid. These refer to the 2 other meta-objectives (themes) of identifying web behavior patterns of existing web customers and prospective web customers, respectively. Web users' behaviors on the web have two facets, namely transactional and browsing behavior inferred from search words, hyperlinks clicking, clickstream analyses, and other browsing or purchasing history. On the other hand, there is also a more psychological and emotional facet of web users' behaviors referred to "consumer behaviour". It typically encompasses their motivations, preferences, reasons, emotions, satisfaction levels, etc. While WM is well-suited to identify browsing behaviors, it is not yet capable of identifying the latent, hidden dimensions of actual browsing behaviors. Recent advances in WM such as opinion mining, sentiment analyses and other semantics-related analyses have attempted to increase proficiency in exploring and discovering these more latent elements of web users' behaviors. However, these techniques are expert systems, very costly, calling for appropriate infrastructures and human resources, making their use still very limited. Future opportunities lie in democratizing access to such analytical tools and processes. But, for the moment WM activities still need to be complemented with traditional market research, at least with regards to identifying web users' psychological and emotional behavior on the internet.

For optimized operational efficiency and tactical as well as strategic decision-making purposes, WM research projects should be devised in a KM framework. Web businesses need first to determine the objectives, resources and limitations of the rolling out a specific WM project. Relevant web data should then be collected from multiple sources and customer touch points according to business capacities, objectives and data availabilities. Once retrieved, selected, pre-processed (cleaned, transformed, standardized, etc.), the data should be gathered into an integrated operational database. Both of these activities constitute the data store of the web data warehouse stage of the KM process. The second stage of the process consists in using WM techniques and descriptives to explore and analyze the knowledge from web users typically obtained from mCRM and eCRM, as well as the knowledge about web users typically obtained from oCRM. This is the data discovery and analysis process, hence aCRM. At this point, useful knowledge for web users is generated to be used for the benefit of web users. This is where WM is beneficial primarily to segment existing and prospective web users and to a lesser extent discover their browsing behavior, transactional loyalty patterns on a website and factual loyalty patterns across websites. Traditional market research remains necessary to reach the other objectives of the framework. Finally, the third stage of the KM process aims at storing the knowledge for web users derived from stage 2 into a knowledge data warehouse which can be used again in eCRM, mCRM and oCRM, and for any customer interaction and touch points of the business be it online or offline. The process goes iteratively in a circular fashion for incremental improvements afterwards.

Marketing goes increasingly technologically-, analytically-driven, it goes also social and mobile and might even go glocal in case of multinational corporations. This means that the traditional marketing perspectives should change and integrate increasingly more technological and technical advances to stay tuned and contribute effectively to business development. It is also a matter of accountability. It has often been and still is being argued that marketing cannot be measured accurately and consequently is not worth the high amount of money that is invested in it. This motto often results in poorly-designed marketing activities rendering organizations vulnerable to fierce market competition. These are excellent candidates for corporate disasters. Technology should be blueprinted in the marketing function. Simultaneously, research in marketing should increasingly capitalize on technological shifts and advances especially in IT, in statistics-analytics, but also in the finance and accounting fields to better track and monitor usage of marketing

expenses as well as the resulting returns that are attributable to marketing activities. Marketing evolves increasingly in an internally (and even externally) borderless company. The sole concept of integrated and holistic marketing as devised by Kotler and Keller (2006) calls for such a fresh look from the marketing research community on primarily a technology-enabled and ultimately knowledge-enabled marketing function. WM and technology as a whole is not an end in itself but a tool to gather highly competitive knowledge and disseminate it pervasively to business parts. Surely each and every function of marketing from data collection to tracking, controlling and monitoring but also strategic activities of segmenting, targeting, positioning, differentiating and the more tactical 4Ps, will be somehow impacted by the availability of relevant knowledge obtained from state-of-the-art technology-enabled aCRM-KM and ultimately BI.

APPENDIX A

INTERVIEW GUIDE

INTERVIEW GUIDE

INTRODUCTION AND INSTRUCTIONS FOR SURVEY RESPONDENTS

This research seeks to better understand the usefulness of the web data mining methods (and techniques) as well as the benefits that they generate for the Analytical Customer Relationship Management (aCRM) applet of the Marketing/Sales function. It has been established that companies neglect or do not necessarily have the means to analyze in-depth the vast amount of gross data generated by their web visitors when they visit the organization's website. Some organizations outsource completely the management of their website as well as the adjacent database management. However, it appears that those "web data" may be very useful in order to reach the aCRM objectives. Data mining applied to rapidly evolving data streams, such as web data, would therefore be a critical complement to information derived from data-mining applied to offline data. It is sought to discover your opinion on how the major web data mining methods (clustering, cross-selling, classification, regression) contribute to reach major aCRM objectives. There is no right or wrong answer to the questions. The main purpose is to give you the opportunity to express yourself freely and openly. You are also very welcome to propose other methods for the achievement of the aCRM objectives. All the answers that you will provide are kept confidential and will only be used for the research project on hand.

LEXICON – Web data mining methods

"Clustering" - "Grouping together users or data items (pages) with similar characteristics". For example, segmenting customers on two variables: "the amount of time spent on the website" and "the number of items downloaded".

"Cross-selling" – "Connecting the individual(s')'s characteristics, behaviors or preferences". For example, selling an additional product to a customer based on previous purchases identified thanks to a basket analysis.

"Classification" - "Explaining or predicting the qualitative characteristic of an individual based on other qualitative characteristics of that individual". For example, in the banking context, connecting socio-demographic characteristics of a customer to his/her possession of a specific financial product.

"Regression" – "Explaining or predicting the quantitative characteristic of an individual based on other quantitative characteristics of that individual". For example, in the telecommunication context, connecting customers' consumption of airtime to their quantitative socio-demographics.

DISCUSSION GUIDE

1. **Usage of WM applied to the profiling of existing web customers**

[RP1a] In your opinion, to what extent do clustering and classification methods applied to web data (web log data, etc.), provide accurate profiles of the existing web customers of a web site?

[RP1b] According to you, to what extent do classification and regression methods applied on web data, identify the recency, the value and the frequency of an individual's purchases online, in order to identify strategically important customers?

[RP1c] For you, to what extent do classification and clustering methods identify existing web customers' loyalty or defection statuses from a web site?

2. Usage of WM applied to the identification of existing customers behaviors on the web

[RP2a] According to you, to what extent do clustering, cross-selling, classification and regression methods, applied to web data, identify how the existing web customers behave on a web site?

[RP2b] In your opinion, to what extent do clustering, cross-selling, classification and regression methods, applied to web data, identify how existing web customers develop satisfaction and loyalty on the internet?

[RP2c] For you, to what extent do clustering, cross-selling, classification and regression methods identify how existing web customers remain attached to- or defect from a given business on the internet?

3. Usage of WM applied to the profiling of prospective customers on the web

[RP3a] For you, to what extent do classification and regression methods enable to segment prospective web customers?

[RP3b] According to you, to what extent do clustering, cross-selling and classification methods provide insightful information about prospective web customers' preferences, needs, habits etc. to develop targeted e-marketing strategies to acquire them?

4. Usage of WM applied to the identification of prospective web customers patterns on the web

[RP4a] In your opinion, to what extent do clustering, cross-selling, classification and regression methods, applied to web data, identify how prospective web customers behave on a given website?

[RP4b] In your opinion, to what extent do clustering, cross-selling, classification and regression methods, applied to web data, identify how prospective web customers defect to and from competitors as well as how they are loyal to competitors on the internet?

(by "defect to" it is meant the web user stops buying from the business and starts buying on the competitor's website; by "defect from" it is meant the web user stops buying on the competitor's website and starts buying on the business' website; and by "loyal to" it is meant the web customer buys regularly on the competitor's website and not on the business website)

5. Demographics (non-compulsory):

5.1. What is your age?

- 18 – 25 years old
- 26-35 years old
- 36-55 years old
- 55 years +

5.2. What is your individual annual income?

- 0 – 20,000 $
- 20,000 – 39,999 $
- 40 000 – 59 999 $
- 60 000 – 79 999 $
- 80 000$ +

5.3. What is (are) your area(s) of expertise?

- Finance
- Procurement
- Human Resources
- Operations Management and Logistics
- Marketing/Communication
- Sales
- Accounting
- Information Technology (IT)
- Engineering
- R&D
- Other

BIBLIOGRAPHY

Abraham, A. and Ramos, V. (2003), Web Usage Mining Using Artificial Ant Colony Clustering and Linear Genetic Programming. In *Evolutionary Computation*, pp. 1384-1391.

Adams, N.M. (2009), "Perspectives on Data-Mining," *International Journal of Marketing Research*, 52(1), pp.11-19.

Adomavicius, G. and Tuzhilin, A. (2005), "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 12(6), pp.734-749.

Agarwal, R., Aggarwal, C. and Prasad, V. (1999), "A Tree Projection Algorithm for Generation of Frequent Itemsets," In *Proc. of the High Performance Data Mining Workshop*.

Agrawal, R., Imielinski, T. and Swami, A.N. (1993), "Mining Association Rules Between Sets of Items in Large Databases," In *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data (SIGMOD'93)*, pp.207-216.

Agrawal, R. and Srikant, R. (1994), "Fast Algorithms for Mining Association Rules," In *Proc. of the 20th Intl. Conf. On Very Large Data Bases (VLDB '94)*, pp.487-499.

Albadvi, A. and Shahbazi, M. (2010), "Integrating Rating-based Collaborative Filtering with Customer Lifetime Value: New Product Recommendation Technique," *Intelligent Data Analysis*, vol.14, pp.143-155.

Al-Fayyad, U.M. and Irani, K.B. (1993), "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning," In *Proc. of the Intl. Joint Conf. on Artificial Intelligence*, pp. 102–1027.

Al-Fayyad, U.M. (1996), "Data-Mining and Knowledge Discovery: Making Sense Out of Data", *IEEE Expert*, 11(5), pp.20-25.

Allaire, Y. and Firsirotu, M. (1984), "La stratégie en deux temps, trois mouvements," *Gestion-Revue internationale de gestion*, 9(2).

Antonie, M. and Zaiane, O. (2002), "Text Document Categorization by Term Association," In *Proc. of IEEE Intl. Conf. on Data Mining.*

Assael, H. (1992), *Consumer Behavior and Marketing Action*, Boston: PWS-Kent Publishing Company.

Azzag, H., Picarougne, F., Guinot, C. and Venturini, G. (2003), "Un Survol des Algorithms Biomimétiques pour la Classification," : <http://www.antsearch.univ-tours.fr/publi/azzag04survol.pdf>

Baesens, B., van Gestel, T., Suykens, J.A.K., Viaene, S., Vanthienen, J., Dedene, G., de Mooret, B. and Vandewalle, J. (2004),"Benchmarking Least Squares Support Vector Machine Classifiers," *Machine Learning*, 54(1), pp.5-32.

Baeza-Yates, R. (2009), "Tendencias en Minería de Datos de la Web," *El Profesional de la Información*, 18(1).

Bayardo, R.J. (1998), "Efficiently Mining Long Patterns from Databases," In *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data (SIGMOD '98)*, pp. 85-93.

Bazsalicza, M. and Naim, P. (2001), *Data Mining pour le Web: Profiling, Filtrage Collaboratif, Personnalisation Client*. Eyrolles.

Beales, H. (2010), "The Value of Behavioral Targeting," Network Advertising Initiative (NAI):
<http://www.networkadvertising.org/index.asp> (last visited Mar. 3, 2012)

Beck, A. and Naim, P. (1999), *Les Réseaux Bayésiens*. Editions Eyrolles.

Bedell, J., Brobst, S. and Markarian, J. (2007), *Critical Success Factors Deploying Pervasive BI*, Informatica-Teradata-MicroStrategy.

Behling, O. (1980), "The Case for the Natural Science Model for Research in Organizational Behavior and Organizational Theory," *Academy of Management Review*, 5(4), pp.483-490.

Benzécri, J.P. (1983), *Histoire et Préhistoire de l'Analyse des Données*. Dunod, Paris.

Berg, H. (2001), "The Stresses of CRM Installations," *ComputerWorld,* 15 January 2001.

Bouchahda, A., Ben Yahia, S. and Slimani, Y. (2006), "Une approche pour l'extraction des itemsets (fermés) fréquents," URL : <http://www.cari-info.org/actes2006/136.pdf>

Bousquet, J., Lachance, Y., Laferté, S., Marticotte, F. (2007), Marketing stratégique. Chenelière Éducation.

Breiman, L., Friedman, J.H., Olshen, R. and Stone, C.J. (1984), *Classification and Regression Trees*. Chapman and Hall, New York.

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 24(2), pp. 123–140.

Breiman, L. (2001), "Random Forests," *Machine Learning*, 45(1), pp. 5–32.

Brunk, C.A. and Pazzani, M.J. (1991), "An Investigation of Noise-Tolerant Relational Concept Learning Algorithms," In *Proc. of the 8th Intl. Workshop on Machine Learning*, pp. 389–393.

Büchner, A.G., Anand, S.S. Mulvenna, M.D. and Hughes, J.G. (1999), "Discovering Internet Marketing Intelligence through Web Log Mining," In *Proc.* Unicom99 Data Mining & Datawarehousing: Realising the full Value of Business Data, pp. 127-138.

Buckinx, Wouter & Van den Poel, Dirk, (2005), "Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting," European Journal of Operational Research, Elsevier, 164(1), pp.252-268, July.

Buehrer, G., Parthasaraty, A. and Ghoting, A. (2006), "Out-of-core frequent pattern mining on a commodity PC," In *Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (SIGMOD '06)*, pp. 86-95.

Bürdick, D., Calimlim, M. and Gehrke, J. (2001), "MAFIA: A Maximal Frequent Itemset Algorithm for Transactional Databases," In *Proc. of the Intl. Conf. on Data Engineering (ICDE '01)*.

Campbell, A. (2003), "Creating Customer Knowledge: Managing Customer Relationship Management Programs Strategically," *Industrial Marketing Management,* 32(5), pp.375-383.

Carrol, P. and Reichheld, F. (1992), "The Fallacy of Customer Retention," *Journal of Retail Banking*, 13(4).

Caruana, R. and Niculescu-Mizil, A. (2006), "An Empirical Comparison of Supervised Learning Algorithms," in *Proceedings of the 23 rd International Conference on Machine Learning '06.*

CEFRIO (2011), "Rapport Nettendances," avril 2011.

Chakrabarti, S. (2003), *Mining the Web: discovering knowledge from hypertext data.* Morgan Kaufmann Publishers.

Chan, J. (2005), "Toward a Unified View of Customer Relationship Management," *Journal of American Academy of Business*, 6(1), pp.32-38.

Chang, K.C., Fund, R., Lucas, A., Oliver, R. and Shikaloff, N. (2000), "Bayesian networks applied to credit scoring," *IMA Journal of Management Mathematics*, 11(1), pp. 1-18.

Chaudhary, S. (2011), "Usage of Web-Mining in Management Research," *Indian Journal of Commerce & Management Studies*, 2(3), pp.101-108.

Chen, J., Qin, Z., Liu, Y., Lu, J. (2005), "Particle Swarm Optimization with Local Search," *Intl Conf. on Neural Networks and Brain, ICNN&B '05*, pp.481-484.

Chiu; S. and Tavella, D. (2008), *Data Mining and Market Intelligence for Optimal Marketing Returns*. Butterworth-Heinemann.

Cho, Y.H. and Kim, J.K. (2004), "Applications of Web Usage Mining and Product Taxonomy to Collaborative Recommendations in E-commerce," *Expert System Applications*, 26(2), pp.233-246.

Choy, K.L., Fan, K.K. and Lo, V. (2003), "Development of an Intelligent Customer-Supplier Relationshp Management System: the Application of Case-based Reasoning," *Industrial Management & Data Systems*, 103(4), pp.263-274.

Cios, K., Pedrycz, W., Swiniarski and Kurgan, L. (2007), *Data Mining: a Knowledge Discovery Approach*. Springer.

Clark, P. and Niblett, T. (1989), "The CN2 Induction Algorithm," *Machine Learning*, Vol. 3, pp. 261–283.

Cleary, C. (2003), "Strategic Issues in Customer Relationship Management (CRM) Implementation," *Business Process Management Journal*, 9(5), pp.592-602.

Clendaniel, T.S. (2002), Profitability and Web Data Mining: Avoiding the Path to Red Ink, published on TDAN.com, 1 january 2002.

Coderre, F., Mathieu, A. and St-Laurent, N. (2004), "Comparison of the quality of qualitative data obtained through telephone, postal and email surveys," *International Journal of Marketing Research*, 46(3), pp.347-357.

Cohen, W.W. (1995), "Fast Effective Rule Induction," In *Proc. of 12th Intl. Conf. on Machine Learning* (*ICML'95*), pp. 115–123.

Cong, G., Tan, K.-L., Tung, A.K.H. and Xu, X. (2005), "Mining Top-k Covering Rule Groups for Gene Expression Data," In *Proc. of ACM SIGMOD Intl. Conf. on Management of Data (SIGMOD '05)*, pp.670-681.

Cooley, R., Mobasher, B and Srivastava, J. (1997), "Web-Mining: Information and Pattern Discovery on the World Wide Web," *ICTAI.*

Cooley, R. Mobasher, B. and Srivastava, J. (1999), "Data Preparation for Mining World Wide Web Browsing Patterns," *Knowledge and Information Systems*, 1(1), pp. 5–32.

Cooley R., Mobasher B., Srivastava J., (2003), "Data Preparation for Mining World Wide Web Browsing Patterns," *Knowledge and Information Systems*, 1(1), pp. 5-32.

Cortes, C. and Vapnik, V. (1995), "Support vector networks," *Machine Learning*, 20(3), pp. 273–297.

Cotterill, P., & Letherby, G. (1993), "Weaving stories: Personal auto/biographies in feminist Research," *Sociology,* 27(1), pp.67-79.

Cristianini, N. and Shawe-Taylor, J. (2000), *An Introduction to Support Vector Machines*. Cambridge University Press.

Custers, B. (2001), Data mining and group profiling on the Internet. In *Ethics and the Internet,* A. Vedder (ed.), Antwerpen, Groningen, Oxford: Intersentia, 87-104.

Daniel, R. (1961), "Management Information Crisis," *Harvard Business Review*, Sept.-Oct.

Davalo, E. and Naim, P. (1991). *Des Réseaux de Neurones*. Editions Eyrolles.

Davis, J. (2005), *Measuring Marketing: 103 Key Metrics Every Marketer Needs*. Wiley.

Denzin, N.K. and Lincoln, Y.S. (2000), *The Qualitative Handbook, 2nd Edition*. Sage Publications.

Déry, R. (2010), *Le Management*. JFD.

Deshpande, M. and Karypis, G. (2002), "Using conjunction of attribute values for classification," In *Proc. of the ACM Intl. Conf. on Information and Knowledge Management (CIKM'02)*, pp. 356-364.

Dey, L. and Mirajul Haque, S.K. (2008), "Opinion Mining from Noisy Text Data," In *AND '08 Proc. Of the Second Workshop on Analytics for Noisy Unstructured Text Data.*

Dietterich, T.G. and Bakiri, G. (1995), "Solving Multiclass Learning Problems via Error-Correcting Output Codes," *J. of Artificial Intelligence Research*, Vol.2. pp. 263–286.

DND Intranet (2004), <http://onlineanswers.com> (retrieved on 16-05-2011)

Domingos, P. and Pazzani (1997), "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss," *Machine Learning,* 29(2–3), pp. 103–130, 1997.

Drucker, P. (1996), "The Information Executives Truly Need," *Harvard Business Review*, January-February, pp.54-62.

Dyson P., Farr A. and Hollis N. (1996), "Understanding, Measuring and Using Brand Equity," *Journal of Adverstising Research*, 36(6), pp.9-21.

Efron, B. and Tibshirani, R. (1997), "Improvements on cross-validation: The .632+ bootstrap method," *J. Amer. Statist. Assoc., Mathematical Reviews* (MathSciNet), Vol. 92, pp. 548—560.

Engel, J.F. and Blackwell, R.D. (1982), *Consumer Behavior, Fourth edition*, Chicago: The Dryden Press.

Facca, F.M. and Lanzi, P.L. (2003), "Recent Developments in Web Usage Mining Research", *Data Warehousing and Knowledge Discovery, 5th International Conference DaWak '03.*

Fang, X. and Sheng, O.R.L. (2004), "Link Selector: A Web Mining Approach to Hyperlink Selection for Web Portals," *ACM Transactions on Internet Technology (TOIT)*, 4(2), pp.209-237.

Farris, P.W., Bendle, N.T., Pfeifer, P.E., Reibstein, D.J. (2006), *Marketing Metrics: 50+ Metrics Every Executive Should Master*. Wharton School Publishing.

Fenstermacher, K. and Ginsburg, M. (2003), "Client-side monitoring for web-mining", *Journal of the American Society for Information Science and Technology,* 57(7), pp.625-637.

Forrester Research Group (2001), "Glossary", available at: <www.forrester.com>

Frede, M. (1975), Stoic vs. Peripatetic Syllogistic. *Archive for the History of Philosophy,* Vol.56, pp.99-124.

Freund, Y. and Schapire, R.E. (1996), "Experiments with a New Boosting Algorithm," In *Proc. of the 13th Intl. Conf. on Machine Learning* (*ICML'96*), pp. 148–156.

Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000), "Using Bayesian Networks to Analyze Expression Data," *Journal of Computational Biology*, 7(3-4), pp.601-620.

Fulgoni, G.M. (2008), *"How Online Advertizing Works Whither the Click,"* ComScore, prepared for Empirical Generalizations in Advertising Conference for Industry and Academia, The Wharton School, Philadelphia, PA.

Furnkranz, J. and Widmer, G. (1994), "Incremental Reduced Error Pruning," In *Proc. of the Eleventh Intl. Conf. Machine Learning*, pp. 70–77.

Furnkranz, J. (2005), *Data Mining and Knowledge Discovery Handbook*, SpringerLink.

Galeas, P. (2008), Patricio Galeas.Org: < http://www.galeas.de/webmining.html> (retrieved on 16-04-2011).

Ghoshal, S. and Bartlett, C.A. (1997), *The Individualized Corporation: A Fundamentally New Approach to Management*. HarperBusiness.

Gremler, D. D. (1995), "*The effect of satisfaction, switching costs, and interpersonal bonds on service loyalty*," Unpublished doctoral dissertation, Arizona State University, Tucson, Arizona.

Grönroos, C. (1994), "From marketing mix to relationship marketing: towards a paradigm shift in marketing," *Management Decision*, 32(2), pp. 4-20.

Guan, S.U. and McMullen, P. (2005), "Organizing Information on the Next Generation Web – Design and Implementation of a Structure," *International Journal of Information Technology & Decision-Making*, 4(1), pp.97-115.

Guba, E.G. and Lincoln, Y.S. (1989), *Fourth Generation Evaluation*. Sage, London.

Guéguen, L. and Datcu, M. (1996), "Spatio-Temporal Pattern Clustering Method Based on Information Bottleneck Principle," Competence Center in the field of Information Extraction and Image Understanding for Earth Observation: <http://biblio.telecom-paristech.fr/cgi-bin/download.cgi?id=6403>

Hair, J., Black, W., Babin, B., Anderson, R., and Tatham, R. (2006), *Multivariate Data Analysis*, 6th ed. Pearson Prentice Hall, Upper Saddle River, New Jersey.

Hall, J. (2004), "BI: The missing link in your CRM strategy", *DM Review*, 14(36).

Hameed, I. (2004), "*Knowledge Management and Business Intelligence; what is the difference?*".

Hamel, G. and Prahalad, C.K. (1994), *Competing for the Future.*

Hammer, M. and Champy, J. (1993), *Reengineering the Corporation: A Manifesto for Business Revolution.*

Han, J., Fu, Y. (1995), "Discovery of Multi-Level Association Rules fron Large Databases," In *Proc. of the 21st Intl. Conf. on Very Large Data Bases (VLDB '05)*, pp.420-431.

Hansen, L.K. and Salamon, P. (1990), "Neural Network Ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 12*, pp. 993-1001.

Hartigan, J.A. (1975), *Clustering Algorithms*. John Wiley & Sons Inc.

Hastie, T. & Tibshirani, R. (1990), *Generalized Additive Models*. Chapman and Hall.

Hauser, J.R. and Katz, G. (1998), "Metrics: You Are What You Measure," *European Management Journal*, 16(5), (October), pp.516-528.

Hay, B., Wetset, G. and Vanhoof, K. (2004), "Mining Navigation Patterns Using a Sequence Alignment Method," *Knowledge and Information Systems*, 6(2), pp.150-163.

Henderson, R.M. and Clark, K.B. (1990), "Architectural Innovation: The Reconfiguration of Existing Product Technologies and the Failure of the Established Firms," *Administrative Science Quarterly*, 35(1), pp.9-30.

Hennig-Thurau, T., Gwinner, K.P., Walsh, G. and Gremler, D.D. (2004), "Electronic Word-Of-Mouth Via Consumer-Opinion Platforms: What Motivates Consumers to Articulate Themselves On The Internet," *Journal Of Interactive Marketing*, 18(1), ABI/INFORM Global, pp.38-52.

Herlocker, J.L., Konstan, J.A., Borchers, A. and Riedl, J. (1999), "An Algorithmic Framework Collaborative Filtering", In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp.230-237.

Holland, J.H., (1975), *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor.

Hollis, N. (2005), "Ten Years of Learning How Online Advertising Builds Brands," *Journal of Marketing Research*, June, pp.255-268.

Hunt, E.B. (1962), *Concept learning: An information processing problem.* New York: Wiley.

Hyafil, L. and Rivest, R.L. (1976), "Constructing Optimal Binary Decision Trees is NPComplete," *Information Processing Letters* 5, pp. 15–17.

Irny, S.I. and Rose, A.A. (2005), "Designing a Strategic Information Systems Planning Methodology for Malaysian Institutes of Higher Learning (isp-ipta)," *Issues in Information System*, 6(1).

Jacoby,J. (1971), "A Model of Multi-Brand Loyalty," *Journal of Advertising Research*, 11(June), pp.25-31.

Jagger, A. (1989), "Love and knowledge: emotion in feminist epistemology," In A. Garry & M. Pearsall (Eds.), *Women, knowledge and reality: explorations in feminist philosophy* (pp.129-155). Boston: Unwin Hyman.

Jeffery, M. (2010), *Data-Driven Marketing: The 15 Metrics Everyone in Marketing Should Know*. Kellogg School of Management.

Jenkinson, A. (1997), *CustomerPrints: Defining the Essentials of the Consumer: The Essential Guide to what they are, why and how to do them and even how to use them.* Truffles, OgilvyOne.

Joshi, A. (2001a), "Web-mining": <www.cs.umbc.edu/~ajoshi/web_mine>

Joshi, A. (2001b), "Web/data mining and personalization," University of Maryland Baltimore County (UMBC) eBiquity Research Area: <http://ebiquity.umbc.edu/edu/project/html/id/17/Web-Data-Mining-and-Personalization>

Kauffman, S. & R. Shapiro (2001), "The biology of invention: A conversation with Stuart Kauffman and Robert Shapiro," Growing the adaptive enterprise, Issue 4: <http://www.businessinnovation.ey.com/journal/issue4/features/biology/body.html>

Kaushik, A. (2009), *Web Analytics 2.0: The Art of Online Accountability and Science of Customer Centricity*, Sybex.

Kearns, M. (1998), "Efficient Noise-Tolerant Learning from Statistical Queries," *Journal of the ACM*, Vol. 45, pp. 983–1006.

Keller K. L. (1993), "Conceptualizing, Measuring, and Managing Customer-Based Brand Equity," *Journal of Marketing Research*, Vol.29, pp.1-22.

Kerin, R.A., Varadarajan, P.R. and Peterson, R.A. (1992), "First-Mover Advantage: A Synthesis, Conceptual Framework, and Research Propositions," *Journal of Marketing*, 56(4), pp.33.-52.

Khabaza, T. (2000), "As E-asy as falling off a web-log: data mining hits the web," in *Proceedings of the fourth international conference on the practical applications of knowledge discovery and data mining*, Manchester UK, April 2000, The Practical Application Company. <http://www.mining.dk/SPSS/Nyheder/as_easy.htm>

Kimball, R. and Ross, M. (2002), *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 2nd Edition*. Wiley.

Kohavi, R. Becker, B. and Sommerfield, D. (1997), "Improving Simple Bayes," In *Proc. Of European Conference on Machine Learning (ECML'97)*.

Kosala, R. and Blockeel, H. (2000), "Web Mining Research: A Survey", *ACM SIGKDD Explorations*, Vol.2, pp.1-15.

Kotler, P. and Keller, K.L. (2006), *Marketing Management, 12th Edition.*

Kotorov, R. (2002), "Ubiquitous Organisation: Organisational Design for e-CRM », *Business Process Management Journal*, 8(3), pp.218-232.

Kracklauer, A.H. and Mills, D.Q. (2004), *Collaborative Customer Relationship Management: Taking CRM to the Next Level*. Springer, Berlin.

Kuttner, R (1998), "The net: a market too perfect for profits," *BusinessWeek*, vol.3577, May 11, pp.20.

Langley, P. Iba, W. and Thompson. K. (1992), "An Analysis of Bayesian Classifiers," In *Proc. of the 10th National Conf. on Artificial Intelligence* (*AAAI'92*), pp. 223–22.

Lau K-N, Lee K.-H., Ho Y. (2005), "Text Mining for the Hotel Industry," *Cornell Hotel and Restaurant Administration Quarterly*, 46(3), pp. 344-362.

Laverty, S. M. (2003), "Hermeneutic phenomenology and phenomenology: A comparison of historical and methodological considerations," *International Journal of Qualitative Methods,* 2(3).

Learmonth, M. (2010), "Holy Grail of Targeting is Fuel for Privacy Battle," *Advertising Age*, Vol. 81, Edition 1.

Lefébure, R. and Venturini, G. (2001), *Le Data Mining, 2ème Édition.* Les Éditions Eyrolles, Paris.

Lei, Y. and Tang, B. (2005), "Study on KM based CRM system," *Proceedings of the Sixth Wuhan International Conference on E-Business-Innovation Management Track.*

Lendrevie, J. and Lindon, D. (1993), *Mercator – Théorie et Pratique du Marketing*. Editions Dalloz.

Lewis, D. and Gale. W. (1994), "A Sequential Algorithm for Training Text Classifiers". In *Proc. of the ACM SIGIR Intl. Conf. on Research and Development in Information Retrieval* (*SIGIR'94*), pp. 3–12.

Lin, D.-I. and Kedem, Z.M. (1998), "Pincer-Search: A New Algorithm for Discovering the Max Mum Frequent Set," In *Proc. of the 6th Intl. Conf. Extending Database Technology (EDBT '98).*

Liu, B., Hsu, W. and Ma, Y. (1998), "Integrating Classification and Association Rule Mining," In *Proc. of Knowledge Discovery and Data Mining* (*KDD'98*), pp. 80–86.

Liu, B., Hsu, W. and Ma, Y. (1999), "Mining Association Rules with Multiple Minimum Supports," In *Proc. of Intl. Conf. on Knowledge Discovery and Data-Mining (KDD '99)*, pp.337-341.

Liu, B., Zhao, K., Benkler, J. and Xiao, W. (2006), "Rule Interestingness Analysis Using OLAP Operations," In *Proc. of the 12th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD '06)*, pp.297-306.

Liu, B. (2007), *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data.* Springer. Verlag Press.

Liu, B. (2011), *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data, 2nd Edition.* Springer. Verlag Press.

Liu, D.R. and Shih, Y.Y. (2005a), "Hybrid approaches to product recommendation based on customer lifetime value," *Journal of Systems and Software*, 77(2), pp.181-191.

Liu, D.R. and Shih, Y.Y. (2005b), "Integrating AHP and data-mining for product recommendation based on customer lifetime value," *Information & Management*, Vol.42, pp.387-400.

Lovelace, J. and Cios, K.J. (2007), "A very simple spiking neuron model that allows for efficient modeling of complex systems," *Neural Computation*, Vol.19, pp.1-26.

Luck, D. and Lancaster, G. (2003), "E-CRM: Customer Relationship Marketing in the Hotel Industry," *Managerial Auditing Journal*, 18(3), pp.213-231.

Łukaszyk, S. (2004), "A new concept of probability metric and its applications in approximation of scattered data sets," *Computational Mechanics*, Vol.33, pp.299-304.

Lumer, E.D. and Faieta, B. (1994), "Diversity and adaptation in populations of clustering ants," In *Proceedings of the Third International Conference on Simulation of Adaptive Behaviour*, pp. 501–508.

Madria, S.K., Bhowmick, S.S., Ng, W.-K. and Lim, E.P. (1999), "Research Issues in Web mining," *Lecture Notes in Computer Science*, 1676:303-312.

Mahboubi, H., Darmont, J. and Aouiche, K. (2007), "Un Index de Jointure pour les Entrepôts de Données XML," *ARXIV*, Cornell University Library. Malhotra,

N.K. (2010). *Marketing Research: an Applied Orientation*. Pearson.

McCallum, A. and Nigam. K. (1998), "A Comparison of Event Models for Naïve Bayes Text Classification," In *Proc. of the AAAI–98 Workshop on Learning for Text Categorization.*

Mihai, I. (2009), "Web-Mining in E-commerce," *Annals of the University of Oradea, Economic Science Series,* pp.959-962.

Mircan, D. and Sitar-Taut, D.A. (2009), "Preprocessing and Content/Navigational Pages Identification as Premises for an Extended Web Usage Mining Model Development," *Informatica Economica*, 13(4).

Mitchell, T. (1997), *Machine Learning*. McGraw Hill.

Mobasher, B., Cooley, R. and Srivastava, J. (2000a), "Automatic Personalization based on web usage mining," *Communication ACM*, 43(8), pp.142-151.

Mobasher, B., Dai, H., Luo, T., Sun, Y. and Zhu, J. (2000b), "Integrating Web Usage and Content Mining for More Effective Personalization", in Bauknecht, K., Madria,

S.K. and Pernul, G. (2000). *EC-Web*, LNCS 1875, Springer-Verlag Berlin Heidelberg, pp.165-176.

Mobasher, B., Dai, H., Luo, T. and Nakagawa, N. (2001), "Effective Personalization Based on Association Rule Discovery from Web Usage Data," In *Proc. of the 3rd ACM Workshop on Web Information and Data Management (WIDM01)*, pp. 9–15.

Mobasher, B. and Nasraoui, O. (2011), "CHAPTER 12: Web Usage Mining," invited book chapter in "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)," *Second Edition*, July 2011, by Bing Liu.

Moe, W. (2003), "Buying, Searching, or Browsing: Differentiating Between Online Shoppers Using In-Store Navigational Clickstream," *Journal of Consumer Psychology*, 13(1-2), pp.29-39.

Moloney, C.X. (2006), *"Winning Your Customers' Loyalty: The Best Tools, Techniques and Practices,"* AMA Workshop Event(s). San Diego.

Morgan, J.N. and Sonquist, J.A. (1963), "Problems in the Analysis of Survey Data, and a Proposal," Journal of the American Statistical Association, Vol.58, pp.415-435.

Munhall, P. (1989), "Philosophical ponderings on qualitative research methods in nursing," *Nursing Science Quarterly, 2(1)*, pp.20-28.

Naïm, P., Wuillemin, P.H., Leray, P., Pourret, O. and Becker, A. (2007), *Réseaux Bayésiens*. Collections Algorithmes.

Nakache, J.P. and Confais, J. (2003), *Statistique Explicative Appliquée*. Éditions Technip.

Nasraoui, O., Krishnapuram, R. and Joshi, A. (1999), "Mining Web Access Logs Using a Relational Clustering Algorithm Based on a Robust Estimator," *Proceedings 8th International World Wide Web Conference (WWW '99)*, pp.40-41.

Nasraoui, O., Soliman, M., Saka, E., Badia, A. and Germain, R. (2008), A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites," IEEE Transactions on Knowledge and Data Engineering, 20(2), pp.202-215.

Nelder, J.A. and Wedderburn, R.W.M. (1972), "Generalized Linear Models," in *J.R. Statist. Soc. A.*, Vol.135, Part 3.

Ngai, E.W.T. (2005), "Customer Relationship Management Research(1992-2002). An Academic Literature Review and Classification," *Marketing Intelligence & Planning*, 23(6), pp.582-605.

Nigam, K. McCallum, A. Thrun, S. and Mitchell, T. (2000), "Text Classification from Labeled and Unlabeled Documents Using EM," *Machine Learning*, 39(2-3), pp. 103–134.

Opitz, D.and Maclin, R. (1999), "Popular Ensemble Methods : An Empirical Study," *Journal of Artificial Intelligence Research,* Vol.11, pp.169-198.

Ortega, J.L. and Aguillo, I.F. (2009), "Minería del Uso de Webs," *El Profesional de la Información*, 18(1), pp.20-26.

Osborne, J. (1994), "Some similarities and differences among phenomenological and other methods of psychological qualitative research," *Canadian Psychology,* 35(2), pp.167-189.

Ozden, B., Ramaswamy, S. and Silberschatz, A. (1998), "Cyclic Association Rules," In *Proc. Intl. Conf. Data Engineering (ICDE '98)*, pp. 412-421.

Pabarskaite, Z. and Raudys, A. (2007), "A process of knowledge discovery from web log data: Systematization and critical review," *Journal of Intelligent Information System*, 28(1), pp.79-104.

Padmanabhan, B. and Tuzhilin, A. (2000), "Small is Beautiful: Discovering the Minimal Set of Unexpected Patterns," In *Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD '00)*, pp.54-63.

Paiva, E.L. Roth, A.V. and Fensterseifer, J.E. (2002), "Focusing Information in Manufacturing a Knowledge Management Perspective," *Industrial Management & Data Systems*, 102(7), pp.381-389.

Pang, B. and Lee, L. (2008), "Opinion Mining and Sentiment Analysis," Foundations and Trends in Information Retrieval, 2(1-2), pp.1-135.

Parasuraman, A., Zeithaml, V.A. and Berry, L.L. (1985), "A Conceptual Model of Service Quality and Its Implications for Future Research," *Journal of Marketing*, Vol.49, pp.41-50.

Pareek, D. (2007), "Business Intelligence for Telecommunications," *CRC Press*, pp.294.

Pasquier, N., Bastide, Y., Taouil, R. and Lakhal, L. (1999), "Discovering Frequent Closed Itemsets for Association Rules," In *Proc. of the 7th Intl. Conf. on Database Theory*, pp. 398-416.

Paulissen, K., Mills, K., Brengman, M., Fjermestad, J. and Romano, N.C. Jr. (2007), "Voids in the Current CRM Literature: Academic Literature Review and Classification (2000-2005)," In *Proceedings of the 40th Annual Hawaii International Convergence on System Sciences.*

Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U. and Hsu, M.-C. (2001), "Prefix-Span: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth," In *Proc. of the 2001 Int. Conf. Data Engineering (ICDE'01)*, pp. 215–224.

Pfeiffer, M. and Zinnhauer, M. (2010), "Can Old Media Enhance New Media? How Traditional Advertising Pays off for an Online Social Network," *Journal of Advertising Research*, pp.42-49.

Pierrakos, D., Paliouras, G., Papatheodorou, C. and Spyropoulos, C.D. (2003), "Web Usage Mining as a Tool for Personalization: A Survey," *User Modeling and User-Adapted Interaction*, 13(4), pp. 311-372.

Pierrakos, D. and Paliouras, G. (2010), "Personalizing Web Directories With The Aid of Web Usage Data", *IEEE Transactions on Knowledge and Data Engineering,* 22(9), pp.1331-1344.

Polkinghorne, D. (1983), *Methodology for the human sciences: Systems of inquiry*. Albany: State University of New York Press.

Popescu, A-M. and Etzioni, O. (2005), "Extracting Product Features and Opinions from Reviews," In *Proc. of Conf. on Human Language Technology and Empirical Methods in Natural Language Processing*, pp.339-346.

Pruitt, J. and Adlin, T. (2006), *The Persona Lifecycle: Keeping People in Mind Throughout Product Design*. Morgan Kaufman.

Quinlan, J.R. (1990), "Learning Logical Definitions from Relations," *Machine Learning*, Vol.5, pp. 239–266.

Quinlan, J.R. (1992), *C4.5: Program for Machine Learning*. Morgan Kaufmann.

Rafea, A. and El-Beltagy, S. (2006), "Data Mining Center of Excellence – Organizational Web Mining," First annual technical report TR/COE_WM/9.0/12/2006, sept. 2005 to aug.2006: <http://www.claes.sci.eg/coe_wm/First_annual_Tech_report.pdf>

Ranjan, J. and Bhatnagar, V. (2011), "Role of Knowledge Management and Analytical CRM in business Data Mining based framework," *The Learning Organization*, 18(2), pp.131-148.

Reichheld, F. (1996), *The Loyalty Effect*, Harvard Business School Press, Boston.

Rivest, R.L. (1987), "Learning Decision Lists," *Machine Learning*, 2(3), pp. 229–246.

Romano, N.C. and Fjermestad, J. (2002a), "Electronic Commerce Customer Relationship Management: An Assessment of Research," *International Journal of Electronic Commerce*, Vol.2, pp.61-113.

Romano, N.C. and Fjermestad, J. (2000b), "Electronic Commerce Customer Relationship Management: A Research Agenda," *Information Technology and Management*, Vol.4, pp.233-258.

Rothschild, M. (1990), Preface to Bionomics, <http://bionomics.org/text/resource/preface.html>

Rowley, J. (2004), "Partnering Paradigms? Knowledge Management and Relationship Marketing," *Industrial Management & Data Systems*, 104(2), pp.149-157.

Rumelhart, D., Hinton, G. and Williams. R. (1996), "Learning Internal Representations by Error Propagation," In D. Rumelhart and J. McClelland (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Volume 1, Chapter 8, pp. 318–362.

Saporta, G. (2006), *Probabilités, analyse des données et statistique*. Technip.

Sasser, W. and Reichheld, F. (1990), "Zero Defects: Quality Comes to Services," *Harvard Business Review*, Sept.-Oct., pp.105-111.

Scholkopf, B. and Smola, A. (2002), *Learning with Kernels*. MIT Press.

Seno, M. and Karypis, G. (2005), "Finding Frequent Patterns Using Length-Decreasing Support Constraints. Data Mining and Knowledge Dicsovery, 10(3), pp.197-228.

Schiff, M.A. (2009), "Business Intelligence: A Guide For Midsize Companies," SAP Business Objects, SAP White Paper BI.

Shahabi, C. and Banaei-Kashani, F. (2002), "A Framework for Efficient and Anonymous Web Usage Mining Based on Client-Side Tracking," In Proc. WEBKDD'01 revised papers from the 3rd Intl Workshop on Mining Web Log Data Across All Customers Touch Points, Springer-Verlag, London, pp.113-144.

Shardanand, U. and Maes (1995), "Social information filtering: Algorithms for automating word of mouth," In Proceedings of CHI'95 Human Factors in
Computing Systems, pp.210-217.

Shaw, M.J., Subramaniam, C., Tan, G.W. and Welge, M.E. (2001), "Knowledge Management and Data Mining for Marketing," *Decision Support Systems*, Vol.31, pp.127-137.

Shih, Y.Y. and Liu, D.R. (2008), "Product recommendation approaches: Collaborative filtering via customer lifetime value and customer demands," *Expert Systems with Applications,* Vol. 35, pp. 350–360.

Sivaramakrishnan, J. & Balakrishnan, V. (2009), "Web mining Functions in an Academic Search Application," *Informatica Economica*, 13(3), pp.132-139.

Smith, K.A. and Ng, A. (2003), "Web Page Clustering Using a Self-Organizing Map of User Navigation Patterns," *Decision Support Systems*, 35(2), pp.245-256.

Snyder, B. and Barzilay, R. (2007), "Multiple aspect ranking using the Good Grief algorithm," In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics.*

Song, Q., Shepperd, M., Cartwright, M. and Mair, C. (2006), "Software Defect Association Mining and Defect Correction Effort Prediction," *Software Engineering, IEEE Transactions*, 32(2), pp.69-82.

Spiliopoulou, M., Pohle, C. and Faulstich, L.C. (1999), "Improving the Effectiveness of a Web Site with Web Usage Mining," *Proceedings of the International Workshop on Web Usage Analysis and User Proling, WEBKDD '99.*

Spiliopoulou, M. (2000), "Web usage mining for web site evaluation". *Communication ACM*, 43(8), pp.127-134.

Srikant, R. and Agrawal, R. (1995), "Mining Generalized Association Rules," In *Proc. of the 21st Intl. Conf. on Very Large Data Bases (VLDB '95)*, pp.407-419.

Srikant, R. and Agrawal, R. (1996), "Mining Quantitative Association Rules in Large Relational Tables," In *Proc. of the ACM SIGMOD Conf. on Management of Data (SIGMOD'96)*.

Srinivasan, S., Anderson, R. and Kishore, P. (1998), "Customer loyalty in e-commerce: An exploration of its antecedents and consequences," *Journal of Retailing,* 78(1), pp.41–50.

Srinivasan, S.S., Anderson, R. and Ponnavolu, K. (2002), "Customer loyalty in ecommerce. An exploration of its antecedents and consequences," *Journal of Retailing*, Vol. 78, No.1, pp.41-50.

Srivastava, J., Cooley, R., Deshpande, M. and Tan, P.N. (2000), "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," *SIGKDD Explorations*, 1(2), pp.12-23.

Stuart, E. (2007), "No Such Thing as Loyalty," iclploylty.com.

Swiercz, W., Cios, K.J., Staley, K., Kurgan, Accurso, F. and Sagel, S. (2006), "New synaptic plasticity rule for networks of spiking neurons," *Nature neuroscience*, Vol.3.

Swift, R.S. (2001). *Accelerating Customer Relationship Using CRM and Relationship Technologies.* Prentice-Hall, Englewood Cliffs, NJ.

Tan, P.N., Kumar, V. and Srivastava, J. (2002), "Selecting the Right Interestingness Measure for Association Patterns," In *Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD'02)*, pp. 32-41.

Tanbeer, S.K., Ahmed, C.F., Jeong, B.S. and Lee, Y.K. (2009), "Sliding Window-based Frequent Pattern Mining over Data Streams," *Information Sciences*, Vol.179, pp.3843-3865.

Tennenhaus M. (2000), *La Régression Logistique PLS*, Journées d'Études en Statistique, Modèles Statistiques pour données Qualitatives.

York, T.L., Durrett, R.T., Tanksley, S. and Nielsen, R. (2002), *"Bayesian and Maximum Likelihood Estimation of Genetic Maps,"*: <http://www.creem.st-and.ac.uk/len/papers/ThomasSCOS32003.pdf>

Tiwana, A. (2001), *The Essential Guide to Knowledge Management: The CRM and E-Business Applications*. Prentice Hall.

Tuzhilin, A. and Adomavicius. G. (2002), "Handling very Large Numbers of Association Rules in the Analysis of Microarray Data," In *Proc. of ACM SIGKDD Intl. Conf. On Knowledge Discovery and Data Mining (KDD'02)*, pp. 396–404.

Tufféry, F. (2007), *Data-Mining et Statistiques Décisionnelles : l'Intelligence des Données*. Editions Technip.

Tufféry, S. (2011), *Data mining and Statistics for Decision Making*. Wiley.

Turney, P.D. (2002), "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," In *Proceedings of the Association for Computational Linguistics 40th Anniversary Meeting*. Association for Computational Linguistics, New Brunswick, NJ.

Vakali, A. and Pallis, G. (2006), *Web Data Management Practices: Emerging Techniques and Technologies*. Idea Group Publishing.

Van Den Poel, D. and Burez, J. (2006), "CRM at a Pay-TV Company: Using Analytical Models to Reduce Customer Attrition by Targeted Marketing for Subscription Services," *Expert Systems with Applications*, 32(2), pp.277-288.

Van Hippel, E. (1984), "Generation and Evaluation of Novel Product Concepts via Analysis of Experienced Users," Marketing Science Institute, Cambridge, Mass.

Van Wel, L. & Royakkers, L. (2004), "Ethical Issues in Web Data-Mining," *Ethics and Information Technology*, Vol.6, pp.129-140.

Vapnik, V., Boser, B.E. and Guyon, I.M. (1992), "A Training Algorithm for Optimal Margin Classifiers," in *5th Annual Workshop on Computational Learning Theory*, Pittsburgh, ACM, pp.144-152.

Vapnik, V. (1995), *The Nature of Statistical Learning Theory*. Springer.

Vedder, A. (1999), "KDD: The challenge to individualism," *Ethics and Information Technology*, Vol.1, pp.275-281.

Wahlberg, O., Strandberg, C., Sundberg, H. and Sandberg, K.W. (2009), "Trends, Topics and Under-researched Areas in CRM Research," *International Journal of Public Information Systems*, Vol.3, pp.191-208.

Wahlstrom, M., Twardowska, I., Walder, I., Kaartinen, T. and Drielsma, J.A. (2010), "European Waste Characterization Standards for the Prevention of Acid/Neutral Rock Drainage," *IMWA 2010.*

Wang, K., He, Y. and Han, J. (2000), "Mining Frequent Itemsets Using Support Constraints," In *Proc. of 26th Intl. Conf. on Very Large Data Bases (VLDB '00)*, pp.43-52.

Wang, J., Han, J. and Pei, J. (2003), "Closet+: Searching for the Best Strategies for Mining Frequent Closed Itemsets," In *Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD '03)*, pp. 236-245.

Wang, W., Yang, J. and Yu, P.S. (2004), "WAR: Weighted Association Rules for Item Intensities," *Knowledge and Information Systems*, 6(2), pp. 203-229.

Webb, I.G. (2001), "Discovering Association with Numeric Variables," In *Proc. of the SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD '01)*, pp. 383-388.

Whewhell, W. (1837), *History of the Inductive Science.*

Wright, A. (2009), "Mining the Web for Feelings, Not Facts," *The New-York Times*, August 23, 2009.

Witten, I.H. and Frank, E. (2005), *Data-Mining: Practical Machine Learning Tools and Techniques*. Elsevier.

Xiong, H., Tan, P.-N. and Kumar, V. (2003), "Mining Strong Affinity Association Patterns in Data Sets with Skewed Support Distribution," *In Proc. of the 3rd IEEE Intl. Conf. on Data Mining (ICDM '03).*

Xu, M. and Walton, J. (2005), "Gaining Customer Knowledge through Analytical CRM", *Industrial Management and Data Systems*, 105(7), pp.955-971.

Yan, J., Liu, N., Wang, G., Zhang, W., Jiang, Y. and Chen, Z. (2009), "How Much Can Behavioral Targeting Help Online Advertising," In *Proc. Intl. World Wide Web Conference (WWW '09)*, pp. 261-270.

Yan, X., Cheng, H., Han, J. and Xin, D. (2005), "Summarizing Itemset Patterns: a Profile-Based Approach," In *Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD'05)*, pp. 314-323.

Yang, Q., Zhang, H.H. and Li., T. (2001), "Mining web logs for prediction models in WWW caching and prefetching," In *Proceedings of the seventh ACM SIGKDD International conference on knowledge discovery and data mining*, ACM New York.

Yang, J., Wang, W. and Yu, P. (2004), "Mining Surprising Periodic Patterns," *Data Mining and Knowledge Discovery*, 9(2), pp. 189-216.

Yang, Y. and Liu, X. (1999), "A Re-Examination of Text Categorization Methods," In *Proc. of the ACM SIGIR Intl. Conf. Research and Development in Information Retrieval (SIGIR'99)*, pp. 42–49.

Yeh, I.C., Lien, C.H., Ting, T.M. and Liu, C.H. (2009), "Applications of Web-Mining for Marketing of Online Bookstores," *Expert Systems with Applications*, No.36, pp.11249-11256.

Zaiane, O., Xin, M. and Han, J. (1998), "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs," *Proceedings Advances in Digital Libraries (ADL '98)*, pp.19-29.

Zaki, M.J. and Hsiao, C. (2002), "Charm: An Efficient Algorithm for Closed Association Rule Mining," In *Proc. of SIAM Conf. on Data Mining*.

Zaki, M.J. and Aggarwal, C.C. (2003), "XRules: an Effective Structural Classifier for XML Data," In *Proc. of the Ninth ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD'03)*, pp. 316–325.

Zhang, S., Zhang, C. and Yang, Q. (2003), "Data Preparation for Data-Mining," *Applied Artificial Intelligence: An International Journal*, 17(5-6), pp.375-381.

Zhang, N.L. (2004), "Hierarchical latent class models for cluster analysis," *Journal of Machine Learning Research*, Vol. 5, pp.697-723.

Zhang, Q. and Segall, R.S. (2008), "Web Mining: A Survey of Current Research, Techniques, and Software," *International Journal of Information Technology and Decision-Making*, 7(4), pp.683-720.