Integrating expert knowledge and multilingual web crawling data in a lead qualification system

J. D'Haen [a,*], D. Van den Poel [a], D. Thorleuchter [b], D.F. Benoit [a]

[a] Department of Marketing, Faculty of Economics and Business Administration, Tweekerkenstraat 2, 9000 Ghent, Belgium
[b] Fraunhofer Institute for Technological Trend Analysis, Appelsgarten 2, 53879 Euskirchen, Germany

ARTICLE INFO

ABSTRACT

Qualifying prospects as leads to contact is a complex exercise. Sales representatives often do not have the time or resources to rationally select the best leads to call. As a result, they rely on gut feeling and arbitrary rules to qualify leads. Model-based decision support systems make this process less subjective. Standard input for such an automated lead qualification system is commercial data. Commercial data, however, tends to be expensive and of ambiguous quality due to missing information. This study proposes web crawling data in combination with expert knowledge as an alternative. Web crawling data is freely available and of higher quality as it is generated by companies themselves. Potential customers use websites as a main information source, so companies benefit from correct and complete websites. Expert knowledge, on the other hand, augments web crawling data by inserting specific information. Web data consists of text that is converted to numbers using text mining techniques that make an abstraction of the text. A field experiment was conducted to test how a decision support system based on web crawling data and expert knowledge compares to a basic decision support system within an international energy retailer. Results verify the added value of the proposed approach.

1. Introduction

Customer relationship management (CRM) is centered on the full customer life cycle using acquisition, development, retention and win-back strategies. The focus in this study is on customer acquisition, which is inherent to any company. Customers are lost for various reasons forcing companies to rely on winning new customers to counterbalance this loss [4,33]. As a result, prospecting initiatives such as cold calling are a continuous requirement to create opportunities in the sales process [49]. Qualifying prospects as leads to contact is, however, a complex exercise [17,49]. Sales representatives rarely have sufficient time or resources to rationally select the best leads to call [67]. As a result, customer acquisition is dictated by arbitrary decision rules based on gut feeling [17]. This hampers the acquisition process, with precious time and money lost on irrelevant leads. Moreover, sales representatives often complain about the quality of the leads they receive from marketing [45]. Thus, an automated decision support system is necessary to provide sales with quality leads. Such a system takes a prospect list as input and uses an array of statistical and data mining techniques to qualify those prospects that are most likely to become a customer as

leads to contact. As a result, sales representatives have higher faith in the quality of leads they receive, making them more motivated to follow up on those leads [50].

To develop a useful lead qualification system, two criteria have to be met. Quality data is needed and a model is required to discover relations hidden in this data. The main challenge lies, however, in the former. This refers to the well known "garbage-in, garbage-out" principle that a model can only be as good as the data that is used to train it [8]. A shortage of data is inherent to customer acquisition [7,67]. As there is no internal information on prospects, companies depend on external data sources for acquisition modeling. The case study at hand focuses on the B2B side of an energy retailer. In B2B lead qualification the external data sources entail mostly firmographic data [37]. Firmographics contain key business demographics such as industry or number of employees and are mainly purchased through specialized vendors [64]. Yet, commercial data tends to be expensive, while providing poor quality due to missing information.

Nowadays, the internet provides a wealth of data. It has had a significant impact on CRM due to its high-speed and cost effectiveness [32, 35]. For example, acquisition costs can be lowered by using online channels for prospecting instead of the more expensive offline channels [12]. As a result, internet is increasingly being used as a medium for customer management [65]. Internet data, and more specifically a company's website, is assumed to be more complete than external, commercial data as its content is generated by the company itself [43]. Websites

* Corresponding author.
E-mail addresses: Jeroen.DHaen@UGent.be (J. D'Haen), Dirk.VandenPoel@UGent.be (D. Van den Poel), Dirk.Thorleuchter@int.fraunhofer.de (D. Thorleuchter), Dries.Benoit@UGent.be (D.F. Benoit).

are used by companies to communicate information about themselves to (potential) customers. Thus, it is in their own interest to make this information as complete and detailed as possible. The web crawling data is further augmented by including expert knowledge based variables. Managerial expertise is always implicitly present in data mining. From the problem definition to the selection of the best model, experts intervene in the data mining process. However, the explicit *integration* of expert domain knowledge into data mining models is far less apparent and under-investigated in literature [15]. Expert knowledge is especially relevant in a text mining context as text mining techniques rely on a conversion of pure text to more abstract concepts. This abstraction is a double-edged sword. On the one hand, it reduces noise by grouping words into a concept. On the other hand, concrete information is removed as the individual words disappear. Integrating expert knowledge preserves the noise reduction advantages, while introducing specific expert knowledge information.

Previous research suggests that web crawling data is a quality input source for customer acquisition decision support systems. Its performance is tested on historical acquisition data. This research applies web crawling data in a real-life experiment in an energy retailing context. As a result, only leads (companies in this case) are selected that have a website. Sales representatives receive a random selection of leads that are scored by the decision support system using the web crawling data and expert knowledge. Sales representatives did not receive this score not to bias results. In the post hoc analyses, a distinction is made between the top scoring leads and the remaining leads. The results of the experiment are compared to the results of the company internal decision support system which is based on a basic segmentation. Fig. 1 provides a general overview of the different prospect to qualified lead strategies in lead qualification decision support systems. Note that gut feeling, although presented separately, can always penetrate decision support systems, for example when a calling agent decides to not follow up on some of the selected leads.

The aim of this paper is multifold. Firstly, a decision support system is presented to improve the qualification of leads process that is dominated by gut feeling and basic segmentation. To facilitate the implementation of the decision support system in business, an algorithm is provided to search for website addresses. Second, it presents, to the best of the authors' knowledge, the first real-life field experiment using a decision support system for lead qualification using web crawling data. To date, web crawling data for lead qualification, and in extension, customer acquisition models are rarely used in academic literature. As a result, the available studies are limited to tests on the data. Finally, this study integrates expert domain knowledge with data mining modeling. Here, expert knowledge is used as an additional data augmentation strategy. The Conclusion and discussion section elaborates on how expert knowledge complements web crawling data.

The remainder of the paper is structured as follows. First, a short overview of related studies is presented. Second, the methodology of the web crawling algorithm and subsequent text mining is elaborated. Next, the data of the test case is described. Third, results are discussed. Then, a conclusion and discussion section is provided. The paper ends with limitations and suggestions for future research.

2. Related work

Customer acquisition remains to date an under researched topic [17]. Few theoretical and application studies are available in literature. The most common applied lead qualification strategy is profiling models using commercially available data [30,53,64]. Profiles of current customers are created based on a fixed set of features and prospects are matched to these profiles to select leads to contact [14]. If historic data is available on which contacted leads became customer, supervised techniques can be applied that weight the different features [26]. In a previous study, however, it was shown that an alternative data source, originating from web crawling, could provide better results in these profiling models [18]. A company's website provides abundant information on the company itself such as size and industry, which is similar to the information present in commercial data. Yet, the information on a company's website is more complete, which is the main issue of commercial data.

Variety is one of the cornerstones of the current big data hype [9]. It hints at the fact that data usually exists in multiple forms or originates from different sources. A single source data input is becoming more and more rare and even problematic, especially in customer acquisition settings [6]. Thus, different data sources are often combined, which is also known as data augmentation. Expert knowledge is a data source that can be integrated with data mining throughout the whole process from problem definition to implementation [36]. It has, for example, been applied to create intuitively interpretable models [40] or improve rule induction using heuristics [1]. Bayesian methods are especially suited to integrate expert opinions in data mining models [15]. With respect to text mining research, expert knowledge is mainly used in sentiment analysis. Either dictionaries defining positive and negative words or expert-labeled documents are used as input [42]. This paper, in contrast, introduces expert knowledge during the data gathering step of the web mining application. Specific search queries are used on a website that represent a website's "activity" level (see Section 3.1 for more information). These queries are based on expert domain
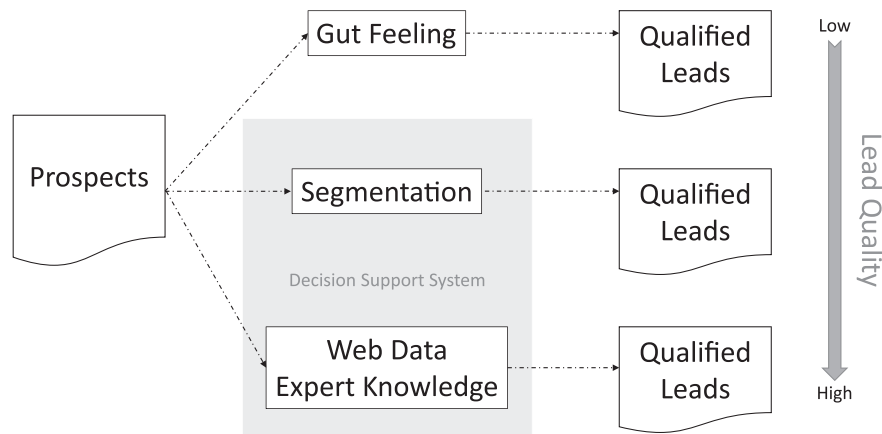

Fig. 1. Three strategies for lead qualification.