Working Paper

Integrating Expert Knowledge and Multilingual Web Crawling Data in a Lead Qualification System

J. D'Haen, D. Van den Poel, D. Thorleuchter, D. F. Benoit

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 Kamakura et al. (2005); Ang and Buttle (2006). As a result, prospecting initiatives such as cold calling are a continuous requirement to create opportunities in the sales process Rodriguez et al. (2012). Qualifying prospects as leads to contact is, however, a complex exercise Rodriguez et al. (2012); D'Haen and Van den Poel (2013). Sales representatives rarely have sufficient time or resources to rationally select the best leads to call Yu and Cai (2007). As a result, customer acquisition is dictated by arbitrary decision rules based on gut feeling D'Haen and Van den Poel (2013). 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 Oliva (2006). 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 Sabnis et al. (2013).

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 Baesens et al. (2009). A shortage of data is inherent to customer acquisition Yu and Cai (2007); Baecke and Van den Poel (2012). 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 Laiderman (2005). Firmographics contain key business demographics such as industry or number of employees and are mainly purchased through specialized vendors Wilson (2006). 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 Kimiloglu and Zarali (2009); Kalaignanam et al. (2008). For example, acquisition costs can be lowered by using online channels for prospecting instead of the more expensive offline channels Chelariu and Sangtani (2009). As a result, internet is increasingly being used as a medium for customer management Wright et al. (2002). 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 Melville et al. (2008). Websites 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 Coussement et al. (2015). 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. Figure 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

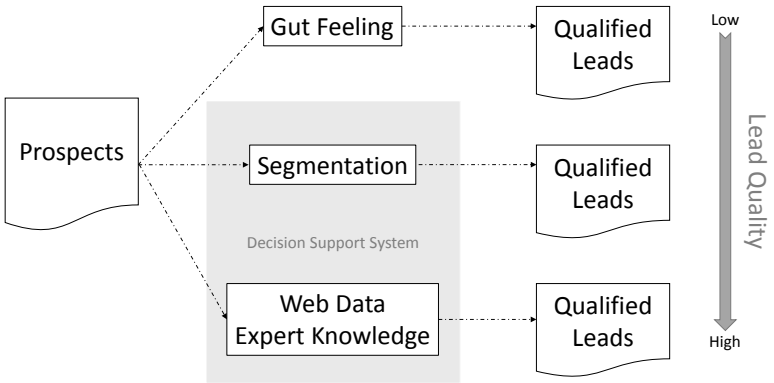

Figure 1: Three Strategies for Lead Qualification

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 D'Haen and Van den Poel (2013). Few theoretical and application studies are available in literature. The most common applied lead qualification strategy are profiling models using commercially available data Jackson (2005); Setnes and Kaymak (2001); Wilson (2006). 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 Chou (2000). If historic data is available on which contacted leads became customer, supervised techniques can be applied that weight the different features Gutierrez et al. (2010). 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 D'Haen et al. (2013). 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 Beyer and Laney (2012). 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 Baecke and Van den Poel (2011). 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 Kopanas et al. (2002). It has, for example, been applied to create intuitively interpretable models Lima et al. (2009) or improve rule induction using heuristics Alonso et al. (2002). Bayesian methods are especially suited to integrate expert opinions in data mining models Coussement et al. (2015). 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 Melville et al. (2009). 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 knowledge by sales, making them domain specific. Domain knowledge is information that is known beforehand, based on previous studies or years of experience Anand et al. (1995). Sales indicated certain company/website characteristics they assumed to indicate a good lead. The expert domain knowledge augments the web crawling data by introducing specific semantic based variables, information that might get lost during the text mining phase.

3 Methodology

While commercial data can readily be used in data mining models, textual data first needs to be preprocessed (Section 3.2). This preprocessing entails for example the stemming of words and counting of words within documents (see Section 3.2). Next, statistical techniques are used to discover which words concurrently appear across leads' websites and group them together in a so-called "concept" (Section 3.3 and 3.4). The result is a dataset with a score on each concept per lead. The higher the score, the more a concept is present on the website of that specific lead. In a final step, a predictive model is trained to discover which concepts are more related to quoted than unquoted leads. As such, web crawling models basically do the same as models using commercial data. Commercial data models predict quoted versus unquoted leads based on, for example, company size, turnover and industry. In a web crawling model this data is embedded in the different concepts. Yet, the important difference is that web crawling data is more complete and reliable as compared to commercial data. To apply the model, new prospect websites are crawled. The text of these websites are extracted and go through the same steps mentioned above. Finally, the different concept scores are inputted into the trained model to get a probability of quotation per prospect.

3.1 Web Crawling

The first step in the web crawling process is to obtain website addresses for the list of potential leads. Two different external commercial databases were

initially consulted to match companies with a web address. Both were able to match between 8 to 9 % of leads. To increase the matching rate and avoid costs related with acquiring commercial data, a website address matching algorithm was developed (see Algorithm 1).

The input of the algorithm is a list of (B2B) leads. More specifically, the company name and city are used as input. Both are employed as a search string in an online search engine. Next, a Levenshtein distance is applied to the root of the website addresses of the returned search results and company name. The root of a website address is defined as the characters between "www" and the suffix such as ".com". The Levenshtein distance measures to what degree sequences of characters are different Nederstigt et al. (2014); Levenshtein (1966). It calculates the number of operations needed to convert a character sequence into another character sequence Lee et al. (2014). If the Levenshtein distance is lower than a predefined threshold, the returned website address is assumed to be the company's address. A manual pre-testing of different thresholds suggested a value of 10 was optimal. Such a threshold is sufficiently conservative to prevent wrong website addresses to be selected as the company's address. If no match was discovered, business directory websites (such as http://www.business-directory-uk.co.uk, http://www.europages.com, http://businessdirectory.bizjournals.com) were selected from the returned results. The public information available on these websites was crawled to retrieve the website address. To prevent overloading the search engine and business directory websites, random waiting times are introduced between each crawl.

Algorithm 1 Website Address Search Algorithm

Define:
leads: A non-empty set containing 2-tuples $\{name, city\}$
SearchEngine($name, city$): Outputs list of URLs, result of using *name* and *city* as search input
D_L: Levenshtein distance
Business_directory: List of business directory websites
Crawl: Extracting text from a website

for $i \leftarrow 1$, $count(leads)$ **do**
    $URL_i \leftarrow$ SearchEngine($name_i, city_i$)

    if Any($D_L(name_i, URL_i) < 10$) **then**
        $website \leftarrow URLs_j$, *where*
        $j = \arg\min_i(D_L(name, URLs_i))$
    else
        for all $Business\_directory \in URL_i$ **do**
            if $website \in$ Crawl($business\_directory$) **then**
                Return($website$)
            else
                Return($website$) $= NULL$
            end if
        end for
    end if
    Random Wait
end for

Next, the identified websites were crawled, extracting the HTML code behind a web page. The homepage and first level links are crawled. The HTML code is parsed to extract the non-HTML text. During this parsing step the activity variables are also created. One group of variables is related to social media. They measure whether sites such as "Facebook", "LinkedIn" and "Twitter" are mentioned on the website. A second group investigates whether the website contains a contact form for visitors to fill in or contact details are present. Sales proposed these characteristics because they reflect to what degree a company is active on the internet. Instead of simply having a website on which potential customer can look up information, they provide visitors opportunities to contact them. The premis of company experts is that companies with these characteristics are also more open to be contacted by a sales team.

The website search and web crawling algorithm were both implemented in SAS Base 9.4M2.

3.2 Text Preprocessing

Texts are typically a collection of unstructured data Weiss et al. (2005). However, several methods exist to convert text into a more structured form Hogenboom et al. (2014). In the following paragraphs the methods utilized to convert the textual data into a more structured, numeric variant are discussed. Each crawled website is called a document and is basically a set of words. A set of documents is called a corpus Srivastava and Sahami (2010). A corpus can be depicted as a term-document matrix of n terms, m documents, and a term frequency tf of term t in document d:

$$A = \begin{bmatrix} tf(t_1, d_1) & \cdots & tf(t_1, d_m) \\ \vdots & \ddots & \vdots \\ tf(t_n, d_1) & \cdots & tf(t_n, d_m) \end{bmatrix} \tag{1}$$

Before a corpus can be analyzed, pre-processing steps are necessary to standardize the text Thorleuchter et al. (2012); Gupta and Lehal (2009). Stop words and numbers are removed, and stemming is applied. Stemming will convert a word to its root form, grouping words with the same conceptual meaning such as *run*, *runner*, *running*.

Often, weights are added to reflect the importance of a term in a document compared to the whole corpus, which has been proven to increase robustness Robertson (2004); Salton and Buckley (1988). More specifically, high frequency terms that appear in most documents need a lower weight as they are less relevant Hao et al. (2014). Low frequency terms, on the other hand, that are present in a limited, specific set of documents are more likely to be relevant, and thus need a higher weight Salton and Buckley (1988). The weight W of a term t in a document d is defined as:

$$W_{t,d} = tf(t,d) log(\frac{N}{df}) \tag{2}$$

with N the total number of documents in the corpus and df the number of documents in which term t appears Salton and Buckley (1988); Thorleuchter et al. (2012); Wu et al. (2008); Salton and Yang (1973); Wei et al. (2014). This weighting scheme is also known as Term Frequency Inverse Document Frequency (Tf-Idf).

At this point, the document-term matrix is characterized by a high sparsity and a high amount of noise. Part of the sparsity and related noise is reduced by removing rare words from the document-term matrix. Including too many irrelevant words is shown to potentially decrease the performance of the model Melville et al. (2008). Words are deleted that have a certain percentage, as defined by the *sparsity reduction parameter*, of empty cells or higher. For example, a parameter value of 0.80 means that all words are deleted that do not appear in at least 80% of all documents. This sparsity reduction parameter is rarely discussed in research as part of the model optimization. However, the Results section will indicate that the setting of this sparsity reduction parameter has an

important impact on the results. By excluding irrelevant words the majority of sparsity is reduced.

The remainder of the sparsity and noise will be resolved by techniques that make an abstraction of the text into concepts.

3.3 Latent Semantic Analysis

The standard technique to convert text to numbers was developed by Deerwester et al. (1990) and is called latent semantic analysis (LSA). It uses the implicit higher-order structure that is present in the relation between terms and documents. By going to a higher level of abstraction, remaining noise, as well as sparsity, are reduced. The statistical backbone of LSA is a singular value decomposition (SVD). The SVD of an m-by-n matrix A has the following form:

$$A = U\Sigma V^T \tag{3}$$

where U (an m-by-m orthogonal matrix) and V (an n-by-n orthogonal matrix) are the left and right singular vectors and Σ is an m-by-n diagonal matrix containing the singular values of A Forsythe et al. (1977). In LSA, the SVD is applied to the term-document matrix Deerwester et al. (1990); Foltz (1996). In that case U contains the document vectors and V holds the term vectors. The first k singular values of are retained to produce a rank-reduced version of A. This reduction is applied to approximate the original document-term matrix, but not match it perfectly:

$$A \approx A_k = U_k \Sigma_k V_k^T \tag{4}$$

This rank-reduced A_k is closest to A in the sense that it minimizes the sum of squares of the difference between A and A_k Wall et al. (2003). Alternatively put, the original A is mapped to a new k-dimensional space. By reducing the dimensionality, vectors of similar documents become more similar as well Foltz (1996). Figure 2 presents a graphical representation of the singular value decomposition with rank-reduction.

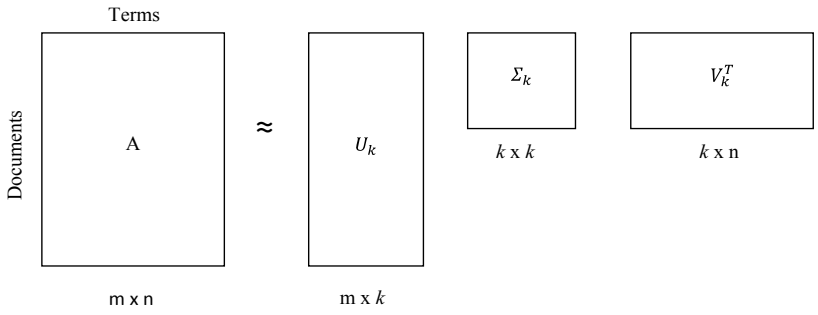

Figure 2: Rank-reduced SVD of a Document-Term matrix

The most popular method to represent new documents in the new k-dimensional space is called folding-in. Folding-in is done using the following equation:

$$\hat{d}_i = \Sigma_k^{-1} U_k^T d_i \tag{5}$$

where d_i represents a new document vector and $\hat{d}_i$ is that document vector mapped onto the existing k-dimensional space Wei et al. (2008). However, folding-in mainly stems from the past where computational power was fairly limited, as it is a less-expensive method Zha and Simon (1999). Moreover, folding-in has the disadvantage of losing information present in newer document vectors. Thus, as this study does not face computational power issues, it simply re-estimates the SVD after adding new document vectors.

3.4 Spherical Clustering

The available textual data on websites has increased rapidly Fersini et al. (2014), making the internet currently the largest data repository Gopal et al. (2011). At the same time, the internet has become a highly multilingual environment Kelly-Holmes (2006). The disadvantage of LSA is that it is known to perform worse in multilingual environments Chew et al. (2007); Bader and Chew (2010). The reason is that it tends to cluster languages and not concepts in multilingual data. An alternative method to LSA in a multilingual environment is clustering. The goal of clustering data is to discover "natural" groups that are inherent in the data Jain (2010). More specifically, it clusters objects into K groups based on a certain similarity measure. Applied on text, clustering discovers latent concepts in unstructured text documents Dhillon and Modha (2001), similar to what LSA does.

The standard similarity measure in a k-means clustering algorithm is Euclidean distance Jain (2010). However, Euclidean distance is not optimal for high dimensional data Strehl et al. (2000). In contrast, the inner product, or cosine similarity is invariant to vector length Strehl et al. (2000); Huang (2008). As a result, documents that have different totals, but the same composition will be similar in terms of their cosine similarity. Thus, cosine similarity, is a more appropriate measure of similarity as compared to Euclidean distance Dhillon and Modha (2001). Dhillon and Modha (2001) altered the Euclidean k-means algorithm and referred to it as the spherical k-means algorithm. It minimizes the following equation:

$$\sum_i (1 - cos(d_i, p_{c_i})) \tag{6}$$

with d representing the document vectors and p the prototype of cluster c(i) Buchta et al. (2012). The membership μ_{ij} of document i to cluster j can be denoted as follows:

$$\mu_{ij} = \begin{cases} 1, \, if \, c_i \, = \, j \\ 0, \, otherwise \end{cases} \tag{7}$$

The standard spherical k-means algorithm can be specified by combining equations 6 and 7:

$$MIN \sum_{i,j} \mu_{ij}(1 - cos(d_i, p_j)) \tag{8}$$

However, documents tend to contain multiple concepts Chau and Yeh (2004). As a result, assigning documents to a single crisp cluster is not appropriate. Instead, a fuzzy clustering algorithm is utilized. Equation 8 can easily be extended to the fuzzy clustering problem as follows:

$$MIN \sum_{i,j} \mu_{ij}^{m}(1 - cos(d_i, p_j)) \tag{9}$$

where m $>$ 1, and softness increases with increasing m values Buchta et al. (2012). The optimal softness setting is discussed in Section 5. New observations can be clustered according to equation 10:

$$\frac{(1 - cos(d_i, p_j))^{\frac{-1}{m-1}}}{\sum_i (1 - cos(d_i, p_j))^{\frac{-1}{m-1}}} \tag{10}$$

All text mining related analyses were performed in R using the "tm" Feinerer et al. (2008), "textir" Taddy (2013) and "lsa" Wild (2015) packages.

To summarize, there are three input data sources: LSA, spherical clustering and expert knowledge. The LSA data is, per company, a vector of length k, which depends on the selected rank reduction (see section 3.3). The values of the cells range between -1 and 1, the former suggesting a concept is absent on a website and the latter suggesting a concept is present. The spherical clustering is, per company, a vector of which the length depends on the amount of clusters retained. The values of the cell range from 0 to 1 and each vector sums to 1. Each cell represents to what degree a company belongs to a specific cluster. The expert knowledge is a set of binary variables. They signify whether, for example, a contact form is present on a website or not.

3.5 Modeling

Logistic regression, a traditional and robust classification technique Andriosopoulos et al. (2012), was used to build different models. Moreover, more complex models did not improve predictive power. A 10-fold cross validation was employed to measure the performance. K-fold cross validation divides the data in k folds of a similar size, where the model is trained on $k - 1$ folds and tested on the remaining fold Rodriguez et al. (2010). This process is repeated so that each fold operated as a test set and the model performance is calculated as the average over all test sets. AUC was selected as performance measure. AUC tests the model over all prediction levels, making it a good performance measure when no specific selection level is known upfront. In the experiment, qualified leads will be added to a queue according to the capacity of calling agents. If a queue becomes empty, more leads are added. A second way in which the quality

will be tested, is the harmonic mean of precision and recall, also known as the F_1-measure Powers (2011):

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \tag{11}$$

The F_1-measure will be used to assess the performance of the best model. AUC is a threshold independent performance measure. This means that it assesses the ranking quality of models. The F_1-measure, in contrast, uses crisp predictions as input. Thus, a threshold is needed to convert the probability to a 0/1 class. The threshold is defined by using a Kolmogorov-Smirnov test on the training dataset to find threshold were the deviation between good and bad leads is maximal.

4 Data

The company in this study is an international energy retailer. Here, we focus on the B2B side of the organization. The energy market is different compared to other markets in the sense that the product cannot be differentiated, it is a commodity. Gas and electricity will be identical, no matter what energy retailer it is bought from. As a result, the price of the commodity becomes crucial Simkin and Dibb (2011). This, however, does not necessarily mean that if a certain retailer becomes the cheapest, all customers switch to this supplier. Other factors such as habit or price insensitivity might also play a role. Due to these specific market characteristics, quotations, rather than contracts, are of interest as contracts are for a great deal determined by price. The goal is to increase the number of quotations, which is assumed to eventually increase the number of contracts as well.

Currently, the energy retailer uses a decision support system based on a basic segmentation (see figure 1) using industry to randomly select purchased leads to contact. The vendor that sells them the list of leads also provides additional data on those leads. However, this data is extremely limited in number of variables. Only five are of interest for a lead qualification model. Moreover, the data quality is low due to a high prevalence of missing values. As a result, building a model using the available, purchased data is not recommended. Testing further showed that models using commercial data performed consistently worse compared to those using web data (AUC of 0.526).

Historical data on quotations of 2013 and part of 2014 were used as input for the crawling model. On average, 1.37 % of these contacted leads received a quotation. 44 % of the leads could be matched with a website address, which is significantly higher than the proportion that could be matched using commercial databases (between 8 to 9 %, see Section 3.1). 22 % of this list could could not be crawled due to pure flash content or limited text, among others. However, the quotation conversion rate remained relatively stable, 1.67 %. If a more severe difference was observed, crawl-ability or simply having a website would potentially play a role in the quotability of a lead.

5 Results

First the parameter optimization is discussed, next the main conclusion of the decision support system are presented. The optimization process consists of finding the optimal combination of three parameters: the sparsity reduction level, cluster softness and total number of clusters. Figures 3-6 on the next page summarize this process. On average, the maximum obtainable AUC per softness setting is around 0.6. However, if the softness parameter is set to 1.4, the maximum AUC drops below 0.6. This indicates that at this point the clusters are becoming too similar to successfully differentiate quoted from unquoted leads. Moreover, an inverse relation between cluster fuzziness (i.e., the softness parameter) and sparsity level in terms of their AUC can be detected. For lower fuzziness levels, the maximal AUC is found in higher sparsity levels and vice versa. Increasing the sparsity means adding data that is potentially noisy. Highly sparse entries are most likely anomalies that are irrelevant. Decreasing fuzziness protects against an increased noisiness.

The optimal parameter combination is a fuzziness factor of 1.2 in combination with a 0.9 sparsity level and 38 clusters. This combination leads to an AUC of 0.62. Intuitively, this can be interpreted as follows. In 62 % of the cases, a randomly selected good lead has a higher probability than a randomly selected bad lead. This is, especially for lead qualification cases, a good result. Table 1 shows that the F_1-measure is low. Yet, it is clear that this is due to a low precision and not recall. Limiting the amount of false negatives (high recall value) is in our study more important than reducing the false positives (high precision value). The number of good leads is limited, thus finding all of them is crucial.

Table 1: Crossvalidated F_1-measure

F_1-Measure	Precision	Recall
0,0388	0,0202	0,6316

A model containing only expert knowledge has a cross validated AUC of 0.55. Using the optimal sparsity and softness setting, combining expert knowledge with clustering is compared to only using clustering. Figure 7 illustrates the added value of these expert knowledge based variables. For the first 40 to 50 clusters the additional variables lead to an increase in AUC. After about 50 clusters, the increase is diminished. Clusters make abstract groups of word combinations, while the expert knowledge is more specific. Increasing the number of clusters decreases the abstraction, making the expert knowledge based variables obsolete. At this point, the expert knowledge information is embedded in the clusters. As a result, there is no more added value of expert knowledge.

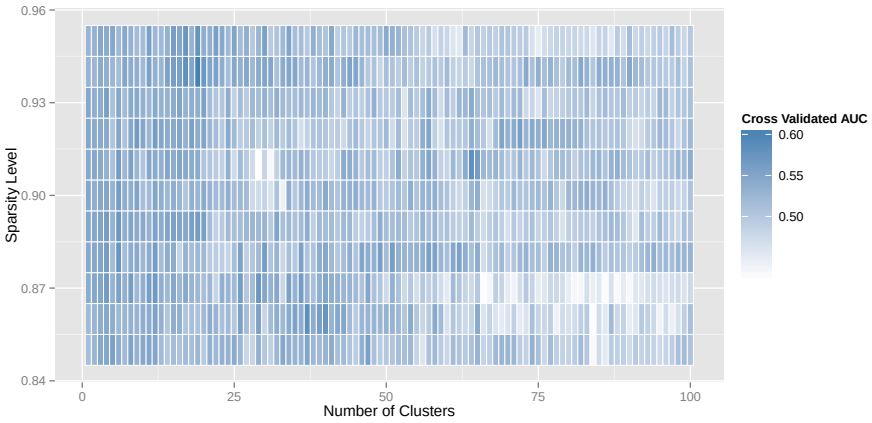

Figure 3: AUC of the Number of Clusters - Sparsity for 1.1 Softness

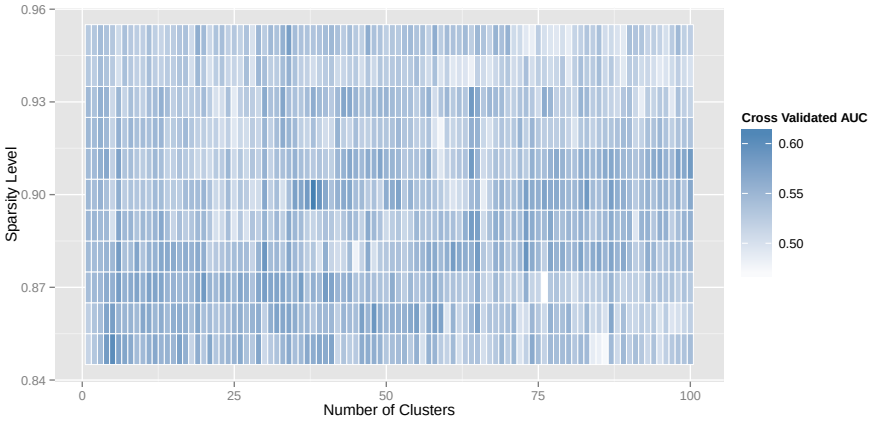

Figure 4: AUC of the Number of Clusters - Sparsity for 1.2 Softness

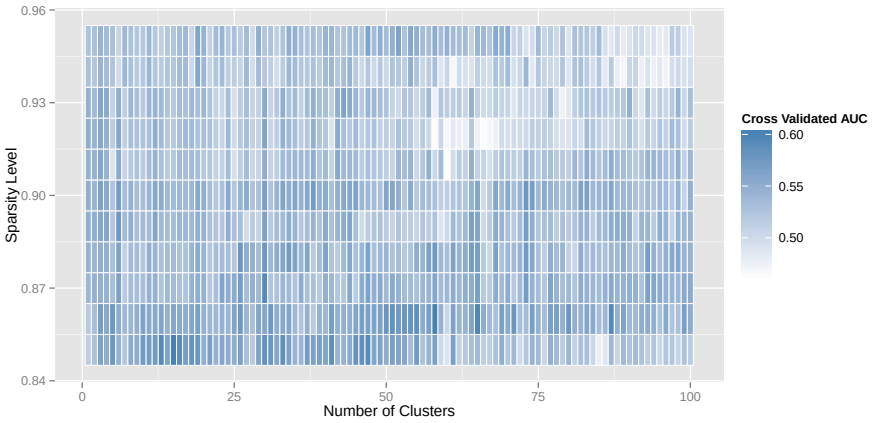

Figure 5: AUC of the Number of Clusters - Sparsity for 1.3 Softness

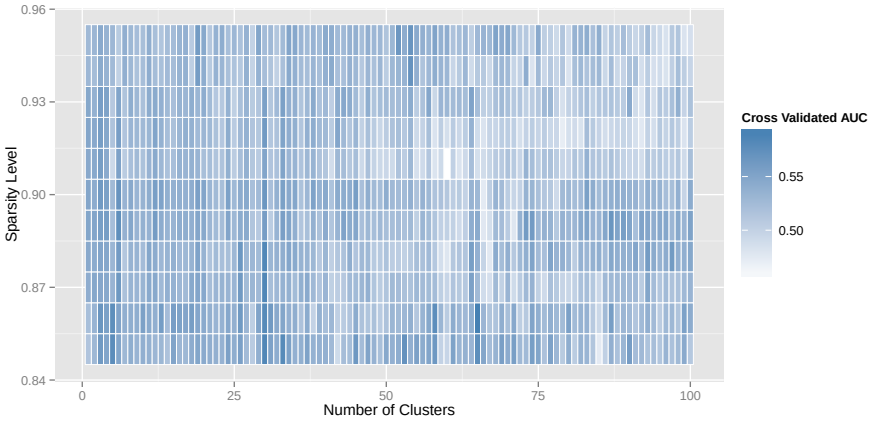

Figure 6: AUC of the Number of Clusters - Sparsity for 1.4 Softness

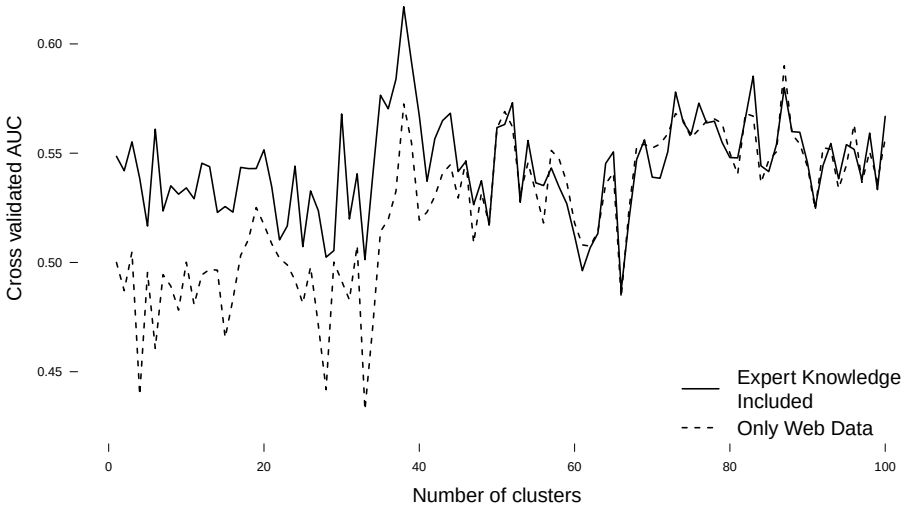

Figure 7: Added Value of Expert Domain Knowledge

The AUC of the optimal model was compared to the standard LSA technique (see Figure 8). Only the first 100 dimensions of the LSA are shown as adding more dimensions did not increase performance. These results suggest that LSA is not able to reach a comparable performance as spherical clustering. A possible explanation is that LSA clusters languages and not concepts in a multilingual environment (see Methodology section).

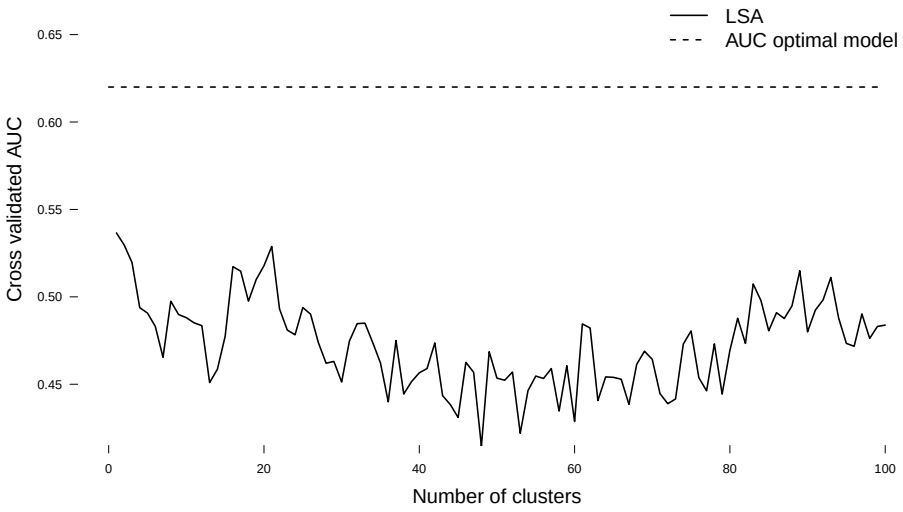

Figure 8: LSA Performance

Table 2 exemplifies how LSA groups languages. Two of the top dimensions (in terms of their explained variance) are selected and the ten words scoring highest on each dimension are shown. The selection of words show that rather than creating multilingual dimensions based on a certain concept, dimensions are created based on language without any concept behind it. Note that the words in the table are stemmed versions and not the original words (see Methodology section).

Table 2: LSA in a multilingual setting

Dutch Dimension	English Dimension
jarenlang	brow
vooral	product
onderstaand	compan
specifiek	project
daarom	collect
bestat	privac
werkt	industr
breng	web
voornam	download
stell	market

In a next step, the optimal parameters were used to score a list of prospects into qualified leads. The prospects are clustered according to equation 10. Next, these clusters are fed into the logistic regression model that delivered the highest AUC. The prospects were divided in four groups based on their score. A random selection of the top group was used as a "best leads" group. An equally sized random selection of the second and third group was selected as a "rest leads" group. The lowest scoring group was ignored. As this was a real test, it was decided not to include low scoring leads as they are assumed to be irrelevant, thus limiting the cost of the test. During four months these leads were contacted by phone and their lead – quoted lead conversion ratio was monitored. The results are presented in Table 3 and are compared to the baseline of 1.37 % (see Data Section).

Table 3: Lead - Quoted lead conversion ratio

Top Leads	Rest Leads	*Baseline*
6.4 %	2.8 %	*1.37 %*

The top scored leads have a conversion ratio from lead to quoted lead of 6.4%. The rest of the leads, on the other hand, have a conversion ratio of 2.8%. This difference is statistically significant at the 0.05 level.

6 Conclusion and Discussion

Sales representatives regularly receive lists that contain a vast amount of leads. To reduce these number of leads to a feasible amount to contact, arbitrary rules are applied. Often these rules are based on mere gut feeling. To assist sales representatives in selecting leads based on a more sound reasoning, decision support systems are created. These systems use historic information on leads contacted in the past to qualify prospects as new leads to contact. A crucial input for any decision support system is quality data, which is a known issue in customer acquisition models. A standard input for these decision support systems is commercial data. However, this data is known for its limited quality, mainly due to missing information. Recently, web crawling data emerged as an alternative data source. This paper provides a first real life test of web crawling data as a quality input for a decision support system for lead qualification. The test case is special due to the industry it is situated in. The energy market is characterized by products that are commodities. This makes the creation of lead qualification systems more difficult as there is no product or brand differentiation possible. It is expected that the proposed model provides better results in different markets as the lead - no lead profiles can include this product and/or brand differentiation. To facilitate future research, a website address search algorithm is provided. Matching website addresses for a huge list of leads is often a bottle neck. The algorithm should stimulate other researchers to implement and potentially improve the algorithm.

The data was further augmented by incorporating expert knowledge into the model. Up until around 50 clusters, expert domain knowledge leads to an increase in AUC. After that, the additional expert information is imbedded in the clusters and performance decreases due to an increase in noise. Tests on historical data indicate that commercial data was not able to match the performance of the web crawling model. This illustrates the quality problem of commercial data and further stresses the importance of finding alternative data sources. The test case validated the quality of the web crawling approach. The top scoring leads had a better conversion rate than the lower scoring leads (6.4% versus 2.8% respectively). Due to economic reasons the worst scoring leads were excluded, explaining the slight difference between the rest group and the expected conversion rate.

The results of this study further illustrate the importance of parameter optimization. More specifically, it is investigated how failing to optimize the combination of parameters can lead to suboptimal results. An important parameter in this study, that is often ignored in literature, is sparsity. Sparsity is an initial noise filter. Including too many sparse entries in the document-term matrix prevents data mining techniques to discover useful information, even if those techniques include a dimension reduction element. On the other hand, excluding too many words potentially removes important data. Moreover, the results suggest an inverse relation between cluster fuzziness and sparsity. Lower fuzziness levels allow a higher sparsity and vice versa. The optimal parameter combination was set at a 1.2 fuzziness factor, 0.9 sparsity level and 38 clusters.

This combination lead to an AUC of 0.62. Moreover, it was illustrated how LSA tends to group languages rather than concepts, making it less suitable for multilingual web data.

7 Limitations and Future Research

The web crawling and expert knowledge data sources are the core of the approach in this study. However, at the same time, they identify both the strength and weakness of the approach. Only leads with a website that can be crawled are potential candidates as input for the model. Moreover, text embedded in HTML is a necessary requirement for the proposed approach to work. As a result, leads with flash websites or image based websites are excluded as well.

The current website search algorithm utilizes the Levenshtein distance to compare potential website addresses with the company name (see Section 3.1). However, it fails to take website address length into account. For example, two substitutions in a string of length 3 do not have the same impact as two substitutions in a string of length 14 Marzal and Vidal (1993). The selected threshold in this study is designed for company names of average length (average length defined by the average length in our subset) and might, as a result, not be appropriate for company names strongly deviating from this average. An improvement might be to use a normalized Levenshtein distance instead. Such a distance measure takes string lengths into account.

The concern of reliability on websites is a potential input for future research. Decision support systems for lead qualification of web crawling and non-web crawling based leads could be integrated into a single model. For example, web and non-web leads can be matched using commercial data, while their qualification score is based on the web data model.

A further improvement on the data augmentation part can be achieved by crawling social media data of companies. Social media are a highly interactive and high-speed medium, providing more up-to-date information.

8 References

References

Alonso, F., Caraa-Valente, J. P., Gonzlez, A. L., Montes, C., 2002. Combining expert knowledge and data mining in a medical diagnosis domain. Expert Systems with Applications 23 (4), 367 – 375.

Anand, S. S., Bell, D. A., Hughes, J. G., 1995. The role of domain knowledge in data mining. In: Proceedings of the Fourth International Conference on Information and Knowledge Management. CIKM '95. ACM, New York, NY, USA, pp. 37–43.

Andriosopoulos, D., Gaganis, C., Pasiouras, F., Zopounidis, C., 2012. An application of multicriteria decision aid models in the prediction of open market share repurchases. Omega 40, 882–890.

Ang, L., Buttle, F., 2006. Managing for successful customer acquisition: an exploration. Journal of marketing management 22, 295–317.

Bader, B. W., Chew, P. A., 2010. Algebraic techniques for multilingual document clustering. In: Kogan, M. . J. (Ed.), Text Mining: Applications and Theory. John Wiley & Sons, pp. 21–35.

Baecke, P., Van den Poel, D., 2011. Data augmentation by predicting spending pleasure using commercially available external data. Journal of Intelligent Information Systems 36 (3), 367–383.

Baecke, P., Van den Poel, D., 2012. Including spatial interdependence in customer acquisition models: A cross-category comparison. Expert Systems with Applications 39 (15), 12105 – 12113.

Baesens, B., Mues, C., Martens, D., Vanthienen, J., 2009. 50 years of data mining and or: upcoming trends and challenges. Journal of the Operational Research Society, S16–S23.

Beyer, M. A., Laney, D., 2012. The importance of 'big data': a definition. Stamford, CT: Gartner.

Buchta, C., Kober, M., Feinerer, I., Hornik, K., 2012. Spherical k-means clustering. Journal of Statistical Software 50, 1–22.

Chau, R., Yeh, C. H., 2004. A multilingual text mining approach to web cross-lingual text retrieval. Knowledge-Based Systems 17, 219–227.

Chelariu, C., Sangtani, V., 2009. Relational governance in b2b electronic marketplaces: an updated typology. Journal of Business & Industrial Marketing 24 (2), 108–118.

Chew, P. A., Bader, B. W., Kolda, T. G., Abdelali, A., 2007. Cross-language information retrieval using parafac2. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACMACM, pp. 143–152.

Chou, P. B., 2000. Identifying prospective customers. In: Proceedings of the 6th ACM SIGKDD international conference on Knowledge discovery and data mining. ACMACM, pp. 447–456.

Coussement, K., Benoit, D., Antioco, M., 2015. A bayesian approach for incorporating expert opinions into decision support systems: A case study of online consumer-satisfaction detection. Decision Support Systems, –.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R., 1990. Indexing by latent semantic analysis. Journal of the American Society for Information Science 41, 391–407.

D'Haen, J., Van den Poel, D., 2013. Model-supported business-to-business prospect prediction based on an iterative customer acquisition framework. Industrial Marketing Management 42, 544–551.

D'Haen, J., Van den Poel, D., Thorleuchter, D., 2013. Predicting customer profitability during acquisition: Finding the optimal combination of data source and data mining technique. Expert systems with applications 40, 2007–2012.

Dhillon, I. S., Modha, D. S., 2001. Concept decompositions for large sparse text data using clustering. Machine Learning 42, 143–175.

Feinerer, I., Hornik, K., Meyer, D., 2008. Text mining infrastructure in r. Journal of Statistical Software 25 (5), 1–54.

Fersini, E., Messina, E., Pozzi, F., 2014. Sentiment analysis: Bayesian ensemble learning. Decision Support Systems 68 (0), 26 – 38.

Foltz, P., 1996. Latent semantic analysis for text-based research. Behavior Research Methods, Instruments, & Computers 28, 197–202.

Forsythe, G. E., Malcolm, M. A., Moler, C. B., 1977. Least squares and the singular value decomposition. In: Computer Methods for Mathematical Computations. Englewood Cliffs: Prentice-Hall, Inc. Englewood Cliffs: Prentice-Hall, Inc, pp. 192–239.

Gopal, R., Marsden, J. R., Vanthienen, J., 2011. Information miningreflections on recent advancements and the road ahead in data, text, and media mining. Decision Support Systems 51 (4), 727–731.

Gupta, V., Lehal, G. S., 2009. A survey of text mining techniques and applications. Journal of emerging technologies in web intelligence 1, 60–76.

Gutierrez, P. A., Segovia-Vargas, M. J., Salcedo-Sanz, S., Hervas-Martinez, C., Sanchis, A., Portilla-Figueras, J. A. e. a., 2010. Hybridizing logistic regression with product unit and rbf networks for accurate detection and prediction of banking crises. Omega 38, 333–344.

Hao, J., Yan, Y., Gong, L., Wang, G., Lin, J., 2014. Knowledge map-based method for domain knowledge browsing. Decision Support Systems 61 (0), 106 – 114.

Hogenboom, A., Heerschop, B., Frasincar, F., Kaymak, U., de Jong, F., 2014. Multi-lingual support for lexicon-based sentiment analysis guided by semantics. Decision Support Systems 62 (0), 43 – 53.

Huang, A., 2008. Similarity measures for text document clustering. In: Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008). Christchurch, Christchurch, New Zealand, pp. 49–56.

Jackson, T. W., 2005. Crm: From 'art to science'. Journal of Database Marketing & Customer Strategy Management 13, 76–92.

Jain, A. K., 2010. Data clustering: 50 years beyond k-means. Pattern Recognition Letters 31, 651–666.

Kalaignanam, K., Kushwaha, T., Varadarajan, P., 2008. Marketing operations efficiency and the internet: An organizing framework. Journal of Business Research 61 (4), 300 – 308.

Kamakura, W., Mela, C. F., Ansari, A., Bodapati, A., Fader, P., Iyengar, R. e. a., 2005. Choice models and customer relationship management. Marketing letters 16, 279–300.

Kelly-Holmes, H., 2006. Multilingualism and commercial language practices on the internet. Journal of Sociolinguistics 10 (4), 507–519.

Kimiloglu, H., Zarali, H., 2009. What signifies success in e-crm? Marketing Intelligence & Planning 27 (2), 246–267.

Kopanas, I., Avouris, N., Daskalaki, S., 2002. The role of domain knowledge in a large scale data mining project. In: Vlahavas, I., Spyropoulos, C. (Eds.), Methods and Applications of Artificial Intelligence. Vol. 2308 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 288–299.

Laiderman, J., 12 2005. A structured approach to b2b segmentation. Journal of Database Marketing & Customer Strategy Management 13 (1), 64–75.

Lee, A. J., Yang, F.-C., Tsai, H.-C., Lai, Y.-Y., 2014. Discovering content-based behavioral roles in social networks. Decision Support Systems 59 (0), 250 – 261.

Levenshtein, V. I., 1966. Binary codes capable of correcting deletions, insertions and reversals. In: Soviet Physics Doklady. Vol. 10. p. 707.

Lima, E., Mues, C., Baesens, B., 08 2009. Domain knowledge integration in data mining using decision tables: case studies in churn prediction. The Journal of the Operational Research Society 60 (8), 1096–1106.

Marzal, A., Vidal, E., 1993. Computation of normalized edit distance and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence 15 (9), 926–932.

Melville, P., Gryc, W., Lawrence, R. D., 2009. Sentiment analysis of blogs by combining lexical knowledge with text classification. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 1275–1284.

Melville, P., Rosset, S., Lawrence, R. D., 2008. Customer targeting models using actively-selected web content. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 946–953.

Nederstigt, L. J., Aanen, S. S., Vandic, D., Frasincar, F., 2014. Floppies: A framework for large-scale ontology population of product information from tabular data in e-commerce stores. Decision Support Systems 59 (0), 296 – 311.

Oliva, R. A., 2006. The three key linkages: improving the connections between marketing and sales. Journal of Business & Industrial Marketing 21 (6), 395–398.

Powers, D. M., 2011. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

Robertson, S., 2004. Understanding inverse document frequency: On theoretical arguments for idf. Journal of documentation 60, 503–520.

Rodriguez, J. D., Perez, A., Lozano, J. A., 2010. Sensitivity analysis of k-fold cross validation in prediction error estimation. Pattern Analysis and Machine Intelligence, IEEE Transactions on 32, 569–575.

Rodriguez, M., Peterson, R. M., Krishnan, V., 2012. Social media's influence on business-to-business sales performance. Journal of Personal Selling & Sales Management 32, 365–378.

Sabnis, G., Chatterjee, S. C., Grewal, R., Lilien, G. L., 2013. The sales lead black hole: On sales reps' follow-up of marketing leads. Journal of Marketing 77, 52–67.

Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. Information Processing & Management 24, 513–523.

Salton, G., Yang, C. S., 1973. On the specification of term values in automatic indexing. Journal of documentation 29, 351–372.

Setnes, M., Kaymak, U., 2001. Fuzzy modeling of client preference from large data sets: An application to target selection in direct marketing. IEEE Transactions on Fuzzy Systems 9, 153–163.

Simkin, L., Dibb, S., 2011. Segmenting the energy market: problems and successes. Marketing Intelligence & Planning 29 (6), 580–592.

Srivastava, A., Sahami, M., 2010. Text Mining: Classification, Clustering, and Applications. Taylor & Francis.

Strehl, A., Ghosh, J., Mooney, R., 2000. Impact of similarity measures on web-page clustering. In: Workshop on Artificial Intelligence for Web Search (AAAI 2000). pp. 58–64.

Taddy, M., 2013. Multinomial inverse regression for text analysis. Journal of the American Statistical Association 108 (503), 755–770.

Thorleuchter, D., Van den Poel, D., Prinzie, A., 2012. Analyzing existing customers websites to improve the customer acquisition process as well as the profitability prediction in b-to-b marketing. Expert systems with applications 39, 2597–2605.

Wall, M. E., Rechtsteiner, A., Rocha, L. M., 2003. Singular value decomposition and principal component analysis. In: D. P. Berrar, W. D., Granzow, M. (Eds.), A Practical Approach to Microarray Data Analysis. Kluwer, Norwell, pp. 91–109.

Wei, C. P., Yang, C. C., Lin, C. M., 2008. A latent semantic indexing-based approach to multilingual document clustering. Decision Support Systems 45, 606–620.

Wei, C.-P., Yang, C.-S., Lee, C.-H., Shi, H., Yang, C. C., 2014. Exploiting polylingual documents for improving text categorization effectiveness. Decision Support Systems 57 (0), 64 – 76.

Weiss, S. M., Indurkhya, N., Zhang, T., Damerau, F. J., 2005. Text Mining: Predictive Methods for Analyzing Unstructured Information. Springer, New York.

Wild, F., 2015. lsa: Latent Semantic Analysis. R package version 0.73.1.
URL `http://CRAN.R-project.org/package=lsa`

Wilson, R. D., 2006. Developing new business strategies in b2b markets by combining crm concepts and online databases. Competitiveness Review: An International Business Journal incorporating Journal of Global Competitiveness 16, 38–43.

Wright, L. T., Stone, M., Abbott, J., 2002. The crm imperativepractice vs theory in the telecommunications industry. The Journal of Database Marketing 9 (4), 339–349.

Wu, H. C., Luk, P., W., R., Wong, K. F., Kwok, K. L., 2008. Interpreting tf-idf term weights as making relevance decisions. ACM Transactions on Information Systems 26, 1–37.

Yu, Y. P., Cai, S. Q., 2007. A new approach to customer targeting under conditions of information shortage. Marketing intelligence & planning 25, 343–359.

Zha, H., Simon, H. D., 1999. On updating problems in latent semantic indexing. SIAM Journal on Scientific Computing 21, 782–791.