UNIVERSITEIT GENT

FACULTEIT ECONOMIE
EN BEDRIJFSKUNDE

TWEEKERKENSTRAAT 2
B-9000 GENT
Tel. : 32 - (0)9 – 264.34.61
Fax. : 32 - (0)9 – 264.35.92

WORKING PAPER

Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing

Dirk Thorleuchter[1]

Dirk Van den Poel[2]

Anita Prinzie[3]

August 2011

2011/733

[1] Fraunhofer INT, D-53879 Euskirchen, Appelsgarten 2, Germany
[2] Corresponding author: Prof. Dr. Dirk Van den Poel, Professor of Marketing Modeling/analytical Customer Relationship Management, Faculty of Economics and Business Administration, dirk.vandenpoel@ugent.be; more papers about customer relationship management can be obtained from the website: www.crm.UGent.be
[3] Visiting Professor at Ghent University

D/2011/7012/38

Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing

Dirk Thorleuchter[a,*,1], Dirk Van den Poel[b], and Anita Prinzie[b]

[a] Fraunhofer INT, D-53879 Euskirchen, Appelsgarten 2, Germany, dirk.thorleuchter@int.fraunhofer.de

[2]Ghent University, Faculty of Economics and Business Administration, B-9000 Gent, Tweekerkenstraat 2, Belgium, dirk.vandenpoel@ugent.be, anita.prinzie@ugent.be

---

[*] *Corresponding author at: Fraunhofer INT, Appelsgarten 2, 53879 Euskirchen, Germany. Tel.: +49 2251 18305; fax: +49 2251 18 38 305*

E-mail address: Dirk.Thorleuchter@int.fraunhofer.de (D. Thorleuchter).

[1] *PhD Candidate, Ghent University, Belgium.*

Abstract

We investigate the issue of predicting new customers as profitable based on information about existing customers in a business-to-business environment. In particular, we show how latent semantic concepts from textual information of existing customers' websites can be used to uncover characteristics of websites of companies that will turn into profitable customers. Hence, the use of predictive analytics will help to identify new potential acquisition targets. Additionally, we show that a regression model based on these concepts is successful in the profitability prediction of new customers. In a case study, the acquisition process of a mail-order company is supported by creating a prioritized list of new customers generated by this approach. It is shown that the density of profitable customers in this list outperforms the density of profitable customers in traditional generated address lists (e. g. from list brokers).

From a managerial point of view, this approach supports the identification of new business customers and helps to estimate the future profitability of these customers in a company. Consequently, the customer acquisition process can be targeted more effectively and efficiently. This leads to a competitive advantage for B2B companies and improves the acquisition process that is time- and cost-consuming with traditionally low conversion rates.

Keywords: B-to-B marketing, Text Mining, Web Mining, Acquisition, SVD

1 Introduction

While products and services are sold by companies with little knowledge or strategy concerning the customers who bought the products in the past a change from this product-centered to a customer-centered environment can be seen today (Coussement & Van den Poel, 2008). This is because for companies it is important to capture and enhance market share while reducing costs. Therefore, they must reconsider the business relationships with their existing customers (Pan & Lee, 2003).

One important aspect is to improve the acquisition of new customers. Normally, this is time- and cost-consuming because it is easier to keep and satisfy existing customers than to attract new ones with a high attrition rate (Reinartz & Kumar, 2003). Therefore, new customers have to be identified who are interested in companies' products and services. This probably can be done in several different ways (e.g. by presenting products and services on fairs, buying addresses from list brokers etc.). However, only a small percentage of potential customers become profitable customers in the future.

In this paper, we propose a new approach to identify new business customers and to predict them as profitable using information of existing customers. For this, existing customers' information is collected from customer relationship management (CRM) systems where customers probably can be divided into different classes e.g. concerning their sales volume. Then existing customers can be classified as profitable customers, if their sales volume over a specific period of time is greater than a specific threshold (Menon, Homburg, & Beutin, 2005).

If we specifically consider on existing customers in a business-to-business environment then we also find information in the CRM system about customers' companies. Nowadays, companies normally present information on internet websites because of the rapid development of IT and the Internet. Many firms rely on Internet websites to provide product information for their customers. Information on existing customers' websites can be crawled and analyzed by use of web structure and content mining approaches. However, the received information consists of masses of unstructured textual information (Coussement & Van den Poel, 2009) and decision makers normally do not use it for acquisition purposes.

This is because the information is not directly usable in a traditional acquisition context and there is often a lack of in-house knowledge on how to analyze this unstructured information for acquisition purposes. Additionally, ready-to-use frameworks are also not available to integrate this information in the acquisition process.

In this approach, textual information of existing customers' websites is analyzed by latent semantic indexing (LSI) to identify specific textual features (concepts). An expectation-maximization (EM) algorithm is used to cluster customers' websites based on the concepts to select prevalent terms from those concepts that mainly occur on the websites of profitable business customers and that seldomly occur on the websites of non-profitable customers. Then, these terms are used as query for web content mining to create a list of further companies with similar concepts on their website. A logistic regression model is built based on the concepts of existing customers' websites to predict the profitability of the new customers from the created list. Comparing this list of potential customers to the traditional acquisition process – e.g. where list brokers' lists of potential customers are used that are expensive - shows that this new approach improves the identification and prediction of new profitable business customers while reducing costs.

This paper contributes to previous research in multiple ways. Firstly, the main contribution of the proposed approach is to show the ability of latent semantic concepts from textual information of existing customers' websites to predict the profitability of new customers (see Sect. 3.5). Secondly, a new web structure/content mining approach is presented to extract relevant information from the websites of existing customers (see Sect. 3.1). Thirdly, a new combined (clustering / web mining) approach is contributed that shows how clustering of websites based on latent semantic concepts can be used to identify prevalent terms and how these terms can be used in a web content mining approach to identify addresses of new potential customers (see Sect. 3.4). Finally, it is shown that using these new addresses in an acquisition process pre-dominates the standard acquisition process e.g. by using addresses of list brokers. Overall, the crawling of new customers using the internet leads to a competitive advantage for B2B companies. Thus, the results contribute to the customer B2B acquisition literature and they testify to the ability of this website-based profitable-customer prediction approach to improve the acquisition process of companies while reducing costs.

2 Related Work

In marketing, we distinguish between transactional and relational approaches. Transactional marketing (Coviello, Brodie, & Munro, 1997) can be defined as an impersonal approach with focus on single point of sale transactions. It describes a company-centric model with an active company and its passive customers, a homogeneous marketplace, and mainly unidirectional information flow from the company to the marketplace / to its customers. In the other direction, little feedback from company's customers to the company can be seen.

In contrast to this, relational marketing focuses on customer retention and satisfaction, rather than single point-of-sale transactions (Kim, 2006; Neslin et al., 2006). Based on relational marketing, information exchange is the main principle in the acquisition of business customers. Its fundamental effect on market growth and structure as well as on new customer acquisition is shown (Naude & Holland, 1996; Verhoef et al., 2010). Further work examines the impact of e-commerce as a new information exchange technology on the acquisition of new business customers (Archer & Yuan, 2000; Baecke & Van den Poel, 2010a; Baecke & Van den Poel, 2010b; De Bock & Van den Poel, 2009; Van den Poel & Buckinx, 2005). Moreover, the impact of word-of-mouth referrals as a traditional information exchange approach is shown on the acquisition of new business customers (Wangenheim & Bayon, 2007).

Related work in the field of web mining focuses on the identification of customer's behaviors in the internet (Bose & Mahapatra, 2001; Bucklin & Gupta, 1992; Lee & Chung, 2003; Park & Chang, 2009) and in the identification of collaborative partners (Engler & Kusiak, 2010).

In contrast to previous work on customer acquisition in a relational B2B context and on web mining, this approach examines latent semantic indexing and web mining as text mining / information retrieval technology for improving the information flow from a company's customers to the company. As a result, the impact of web mining on the acquisition of new business customers can be shown as contribution to the customer B2B acquisition literature.

3 Methodology

Textual information from customer's websites is collected and is transformed in a pre-processing phase to a term-website matrix. After dimension reduction, latent semantic concepts are identified and clustered. Class labels mainly representing profitable customers' websites are used to identify websites of new potential customers. Textual information from these websites is projected into the dimension-reduced latent semantic concept space. A prediction model is built on this concept-space matrix to show that latent semantic concepts from existing customers' websites can be used to predict the profitability of new customers. Fig. 1 shows the methodology of this approach.

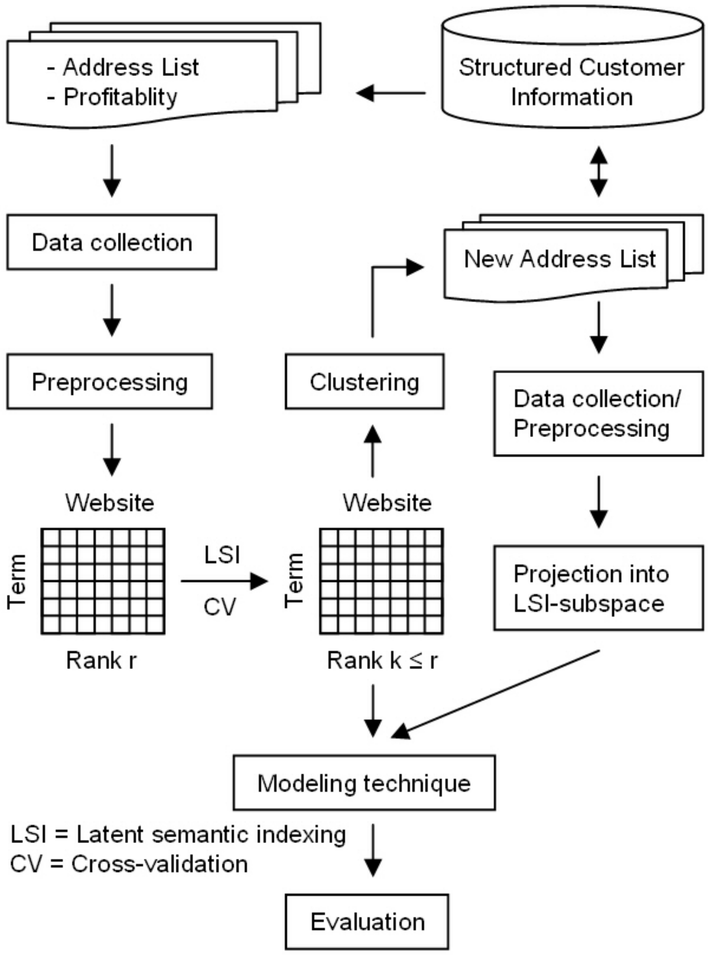

Figure 1: Methodology of the approach

3.1 Data collection

For the data collection phase, structured customer information is needed to collect unstructured content information from customers' websites. Structured customer information can be extracted from CRM systems of a company in which sales volume as well as e-mail addresses or website addresses for each customer is stored. Fig. 2 shows different steps in the data collection phase.

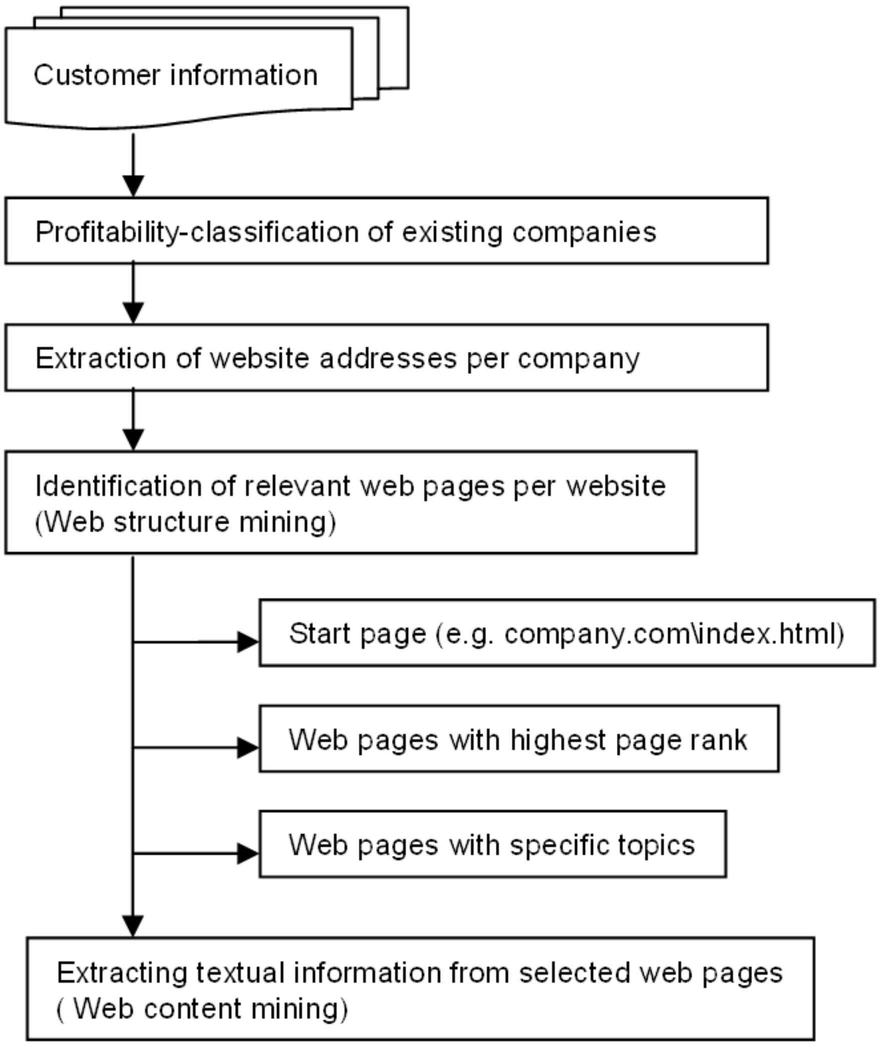

Fig. 2: Different steps in data collection phase.

Information about existing customers' sales volume is used to classify companies as profitable. An aggregation of the sales volume that belongs to the same company is done because several customers probably belong to the same company. Then, companies are assigned to a sales volume and they are defined as profitable if their volume exceeds a specific threshold.

To identify customers' company websites, the website addresses are used or if unavailable, e-mail addresses are converted to website addresses. In general, an e-mail address from a business customer based on the corresponding company website e.g. miller@company-name.com. Therefore, it is often possible to identify the corresponding company website for each customer. If a customer's e-mail address is based on an ISP (internet-service provider) or e-mail provider (e.g. hotmail.com) then his company's website is identified manually, otherwise information about this customer is discarded for further processing.

A website consists of several web pages. To extract information from customers' websites, relevant web pages have to be identified first. This avoids crawling trivial web pages e.g. 'disclaimer', 'privacy / data protection policy', 'sitemap', 'about', 'contact formulary' etc. In general, the starting page of an internet website is relevant. To identify further relevant web pages, the corpus of an internet search engine is used by access to web services. A web service is a software system that is designed to support interoperable machine-to-machine interaction over a network. Frequently, web services are just web-based advanced programming interfaces (APIs). Access to these interfaces is possible over the internet. Then, the requested service is executed, resulting data is ordered by page rank, and it is transferred back to an application that requested the service (Thorleuchter, Van den Poel, & Prinzie, 2010c). A lot of internet search engines offer web services. In this approach, Google is used as well-known internet search engine because its page rank is of high quality and we suppose that Google indexes most commercial websites. For each website, we build a query that is restricted to web pages of the corresponding website and that is additionally restricted on a specific language in a first step. This language restriction is necessary for comparing terms from different web pages and different websites. The query results consist of all indexed web pages ordered by the page rank. For further processing, the starting web page and three further web pages with the highest page rank are selected. Additionally, we suppose that information about a company's history might be relevant for identifying the company as profitable customer. To identify web pages containing this information, we build a query that is restricted to web pages of the corresponding website and in addition contain of specific search terms in a second step. Examples for these search terms are 'founded, history' etc. that have to be translated to the selected language. The resulting web page with the highest page rank is selected for further processing if not already selected in the first step.

With web content mining, information from the selected customers' web pages is extracted. In contrast to the structured information about sales volume for each customer, extracting textual information from customers' websites is highly unstructured. Thus, text pre-processing is necessary to capture the relevant details from this information for integration in the acquisition process.

3.2 Pre-processing

The extracted content information has to be converted into a structured representation as term vectors in a vector-space model. Thus, each web page is represented as a vector of weighted frequencies of designated words (Thorleuchter, Van den Poel, & Prinzie, 2010b). The size of the vector is defined by the number of distinct terms in the dictionary. The importance of a term - with respect to the semantics - is reflected by each corresponding vector component. A vector component is set to its weight if the corresponding term is used in the web page and to zero if the term is not. A collection of these term vectors is used to build a term-by-web page matrix firstly and a term-by-website matrix secondly. The process of converting web pages to a term-by-website matrix is depicted in Fig. 3.

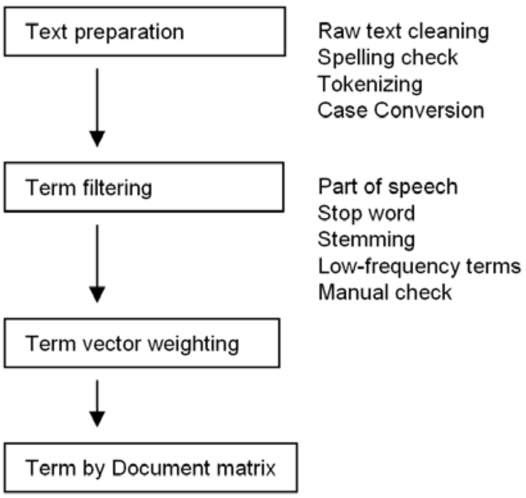

Fig. 3: Different steps in the pre-processing phase.

3.2.1 Text preparation

In the text preparation phase, raw text cleaning is done. For this, images, html-, xml-tags as well as scripting code (e.g. javascript) from the web pages are removed. Additionally, specific characters and punctuation are deleted and typographical errors are corrected by use of a dictionary. With tokenization, all words that are used in the web page can be identified which means texts are separated in terms whereby the term unit is a word. All terms are converted to lower case whereby the first sign is capitalized (case conversion).

3.2.2 Term filtering

The set of different terms in a web page can be reduced by using filtering methods (Thorleuchter, Van den Poel, & Prinzie, 2010a). For further processing, informative terms are selected that belong to a specific syntactic category (nouns, verbs, adjectives and adverbs) by use of part-of-speech tagging. Other (non-informative) terms are discarded. Stop word filtering is the standard filtering method in text mining applications. It is used to remove words that bear little or no content information, like articles, conjunctions, prepositions, etc (Thorleuchter, Van den Poel, & Prinzie, 2010d). Further filtering methods are lemmatization and stemming. A stemmer transforms words to their basic forms named stem by stripping the plural 's' from nouns, the 'ing' from verbs etc. Related words map to the same stem. Stemming is closely related to lemmatization. The difference is that lemmatization uses knowledge of the context to discriminate between words that have different meanings depending on part of speech. Unfortunately, lemmatization is time consuming and still error prone (Thorleuchter, 2008). Therefore, a dictionary-based stemmer is used combined with a set of production rules to give each term a correct stem. The production rules are used when a term is unrecognizable in the dictionary. The term frequencies in textual information follow a Zipf distribution (Zipf, 1949). Half of them appear only once or twice. Thus, those rare terms under these thresholds are deleted that often yield great savings. The last step in term filtering is to check the selected terms manually (Gericke et al., 2009).

3.2.3 Term vector weighting

The selected terms are used to construct a vector of weighted frequencies for each web page. In contrast to term vectors where the component values are raw frequencies of appearance for a term in a web page, the use of term weighting schemes leads to significant improvements in retrieval performance (Sparck Jones, 1972). The weights reflect the importance of a term in a specific web page of the considered web page collection. Large weights are assigned to terms that are used frequently in selected web pages but rarely in the whole web page collection (Salton & Buckley, 1988). Thus a weight $w_{i,j}$ for a term i in web page j is computed by term frequency $tf_{i,j}$ times inverse web page frequency idf_i, which describes the term specificity within the web page collection. In (Salton, Allan, & Buckley, 1994) a weighting scheme was proposed that has meanwhile proven its usability in practice. Besides term frequency – defined as the absolute frequency of term i in web page j - and inverse document frequency - defined as $idf_i := \log(n/ df_i)$ -, a length normalization factor is used to ensure that all documents have equal chances of being retrieved independent of their lengths:

$$w_{i,j} = \frac{tf_{i,j} \cdot \log(n / df_i)}{\sqrt{\sum_{p=1}^{m} tf_{i,j_p}{}^2 \cdot (\log(n / df_{i_p}))^2}} \tag{1}$$

where n is the size of the web page collection D where web pages are represented by term vectors in m-dimensional space, and df_i is the number of web pages in D that contain term i (Chen, Chiu, & Chang, 2005).

3.2.4 Term vector aggregation

As a result, a high-dimensional, weighted term-by-web page matrix is created. However, from the managerial point of view, a prediction is done per customer's company website. Thus, an aggregation of the web pages that belong to the same customer's company website is needed. The aggregated weight of term i for all web pages belonging to a customer's company website j (Coussement & Van den Poel, 2009) is

$$Aw_{i,j} = \sum_{k=1}^{r} w_{i,k} \tag{2}$$

with $w_{i,k}$ equal to the weight of term i in web page k and r equal to the number of web pages belonging to the same customer's company website.

3.3 Concept identification with LSI and singular value decomposition

Using each distinct term as a feature would lead to an unmanageable high dimensionality of the feature space. Additionally, most weights are zero for a customer's company. To reduce the dimension of the feature space LSI is used. LSI groups together related terms (Deerwester et al., 1990) and together with singular value decomposition (SVD) it forms semantic generalizations due to the fact that relationships between terms are recognized by the appearance of terms in similar documents (e.g. web pages). SVD transforms web pages from the high-dimensional feature space to an orthonormal, semantic, latent subspace. Similar terms (keywords) are grouped into concepts. Each concept has a high discriminatory power to other concepts in the reduced feature space.

3.3.1 Feature space dimension reduction

The SVD of a term-by-website (m x n) matrix A with rank r ($r \leq \min(m,n)$) is a transformation into a product of three matrices in form of

$$A = U \Sigma V^t \quad (3)$$

with Σ equal to a diagonal (r x r) matrix containing positive singular values of matrix A where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r$, U equal to the term-concept similarity (m x r) matrix, and V equal to the concept-company similarity (n x r) matrix. The columns of U and V are orthonormal in the Euclidean sense (Yen & Lina, 2010). The weights of the matrix A depended on the latent concepts by

$$w_{i,j} = \sum_{x=1}^{r} U_{i,x} \cdot \Sigma_x \cdot d_{j,x} \quad (4)$$

If $k \leq r$ and the singular values $\lambda_{k+1}, \ldots, \lambda_r$ are small compared to $\lambda_1, \ldots, \lambda_k$ then LSI based on SVD allows a good approximation of A_r with rank r by A_k with rank k. Therefore, LSI dropped the smaller lambda values in Σ by retaining only the first predetermined singular values equal to or greater than k while only the first k columns of U and V were retained.

$$A_k = U_k \Sigma_k V_k^t \quad (5)$$

with U_k, Σ_k and V_k were equal to the k-rank approximation of U, Σ and V, respectively.

The approximated k-rank concept-website similarity matrix V_k contained information on how well a certain website loads on the different k concepts, which reflect the hidden (latent semantic) patterns in the textual information.

3.3.2 Concept dimension selection

The choice of k – the number of concepts - is critical for optimal predictive performance by using SVD. If k is too large then too many irrelevant or unimportant concepts are used for prediction. Otherwise, if k is too small then relevant concepts probably are not considered. The calculation of an optimal number of concepts k can be done using an operational criterion, i.e. a value of k that yields good performance (Chen et al., 2010). In this paper, we use a parameter-selection procedure by constructing several rank-k models, by using a fivefold cross-validation on the training set for each rank-k model, and by selecting the most favorable rank-k model (based on the cross-validated performance) for further analysis. The cross-validation performance is determined by the results of the prediction model (see Sect. 3.5).

3.3.3 Projection of test examples into the LSI-subspace

The meaning of the concepts during testing should stay the same as during training. Consequently, test examples are transformed to term vectors by using the different steps in the pre-processing phase (see Sect. 3.2). Additionally, the projection of the test examples is done into the same semantic latent subspace as created during training (Zhong & Li, 2010). As a result, the term vector A_d is created for each test example and the new concept-website vector can be calculated by

$$V_d = A_d' \cdot U_k \cdot \Sigma_k^{-1} \tag{6}$$

with U_k the k-rank concept-term similarity matrix and Σ_k the diagonal singular value matrix in rank k, both of the original SVD. The new concept-website vector V_d is comparable to the concept-website vectors of the matrix V_k.

3.4 Website clustering

Concepts of the dimension reduced new concept-website matrix reflect the hidden, latent semantic patterns in the textual information from companies' websites that have a high discriminatory power to other concepts. For applying these concepts for acquisition (e.g. to create a list of new potential customers), we have to identify prevalent terms mainly representing concepts from profitable companies' websites and least of all from non-profitable companies' websites. Websites are clustered to identify these terms. An expectation-maximization (EM) algorithm is used for finding maximum likelihood estimates of parameters in a probabilistic model, where the model depends on the dimension-reduced SVD concepts.

As a result, classes contain profitable companies' websites as well as non-profitable company's websites. Class labels represent a number of prevalent terms from the websites assigned to a class. Labels are selected from classes that are mainly assigned to profitable customers' websites. Terms from these labels are used as search query in a web mining approach. Thus, companies can be identified where the selected terms occur on their websites and where the company itself does not occur in the training examples. It can be supposed that their websites contain similar concepts as concepts from the corresponding class and therefore, they probably are profitable customers, too. To evaluate their profitability, textual information from their websites is collected (see Sect. 3.1), pre-processed (see Sect. 3.2), projected into the LSI-subspace (see Sect. 3.3.3), and used as test examples in a prediction modeling approach (see Sect. 3.5).

3.5 Prediction Modeling

As modeling technique, logistic regression is used by producing a maximum likelihood function and by maximizing it in order to become an appropriate fit to the data (Allison, 1999). Logistic regression is conceptually simple (DeLong, DeLong, & Clarke-Pearson, 1988), a closed-form solution for the posterior probabilities is available and it provides quick and robust results in a prediction context (Greiff, 1998). Therefore with a training set of $T = \{(x_i, y_i)\}$ and $i = \{1,2,\ldots,N\}$ and input data $x_i \in R^n$ and corresponding binary target labels $y_i \in \{0,1\}$ (non-profitable, profitable), logistic regression is used to estimate the probability $P(y = 1 \mid x)$ given by

$$P(y = 1 \mid x) = \frac{1}{1 + exp(-(w_0 + wx))} \tag{7}$$

with $x \in R^n$ an n-dimensional input vector (a concept-website vector) as representative for companies' websites load on the concepts, w the parameter vector and w_0 the intercept.

3.6 Evaluation criteria

This evaluation focuses on examining the performance of the prediction model to show that latent semantic concepts from existing customers' websites can be used to predict the profitability of new customers and to show that newly created address lists of potential customers contain more profitable customer addresses than lists from list brokers. This is done with the commonly used criteria: lift, precision, recall, area under the receiver operating characteristics curve (AUC), sensitivity, and specificity.

To evaluate the performance of classification models, lift is the most commonly used performance measure for business applications. It measures the increase in density of the number of profitable new customers relative to the density of new customers in total. For an acquisition process, it is interesting to increase the density of profitable customers, especially in the top 30 percentile of a potential customer list because a new customer acquisition is time- and cost-consuming and budgets / personnel resources for acquisition are often limited. Thus, acquisition managers often focus on a subset of new customers. Practically, all new customers are sorted from most profitable to least profitable by the model. Afterwards, the density of profitable customers from the top 30 percentile can be computed.
Based on the number of positives that are correctly identified (TP), the number of negatives that are classified as positives (FP), the number of positive cases that are identified as negatives (FN), and the number of negative cases that are classified as negatives (TN), we use the sensitivity (TP/(TP+FN)) as the proportion of positive cases that are predicted to be positive, the specificity (TN/(TN + FP)) as the proportion of negative cases that are predicted to be negative, the precision (TP/(TP+FP)) as a measure of exactness or fidelity, and the recall (TP/(TP+TN)) as a measure of completeness. These vary when the threshold value is varied. The receiver operating characteristic curve (ROC) is a two dimensional plot of sensitivity versus (1-specificity). In order to compare the performance of two or more classification models, the AUC is calculated. This measure is used to evaluate the performance of a binary classification system (Hanley & McNeil, 1982). The optimal reduced number of concepts is obtained by optimizing the performance of the predictive model as reflected by a cross-validated AUC.

4 Empirical verification

4.1 Research data

In this study, we forecast the profitability of new customers and we support the identification of new profitable customers for a large German business-to-business mail-order company. The company has a structured, marketing database where information is stored about existing customers and their sales volume, as well as their e-mail addresses.

Based on the information from the structured, marketing database, the company identifies 150,000 customers. An aggregation of customers' affiliation is done because several customers probably belong to the same company. As a result, about 60,000 companies can be identified. This number is reduced to about 35,568 companies for which a corresponding website in German language can be identified. Additionally, the sales volume is calculated summing the incoming orders in the recent year to ensure that companies are currently profitable for the mail-order company. Then, profitable companies are defined with a sales volume exceeding a specific threshold determined by the mail-order company.

The data characteristics are shown in Table 1 for the randomly split training, validation and test set. The training and validation set was used to obtain the optimal SVD dimension and the model estimates, while the test set is used to validate and compare the different models.

	Number of customer groups	Relative percentage
Training set (including validation set):		
Non-profitable customer group website addresses	11,344	45.56
Profitable customer group website addresses	13.,553	54.44
Total	24,897	
Test set:		
Non-profitable customer group website addresses	4,793	44.92
Profitable customer group website addresses	5,878	55.08
Total	10,671	

Table 1: Overview of the website characteristics

4.2 Optimal dimension selection

After the pre-processing phase, a high-dimensional term-by-website matrix was created. To obtain its optimal reduced rank, a cross-validation procedure was applied on the training data (see Fig. 4). The x-axis represents the number of concepts and the y-axis represents the cross-validated AUC under the ROC curve. In the range of 1–50 concepts, the cross-validated AUC was increasing rapidly. From 50 concepts on, it was increasing less rapidly, while in the region around 150 concepts, the cross-validated performance was stabilizing. Including more than 150 concepts resulted in a more complex prediction model, while the AUC hardly increased. Thus, 150 concepts were chosen as the optimal number for representing the textual information in our study. At this point, a good balance was achieved between the number of concepts and the predictive performance.

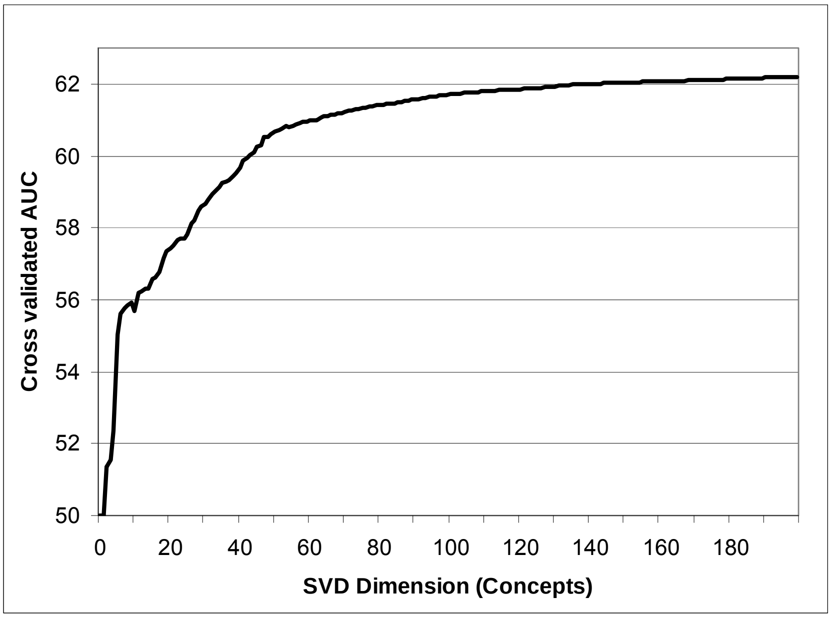

Figure 4: SVD Dimension

Each calculated latent semantic concept shows that the above-chance frequent occurrence of a group of several terms together with the non-occurrence of a further group of several terms on a customer's website can be used to classify this customer as profitable. The terms represent words in stemmed form and they are in German language because only German language websites are considered. Two examples for the interpretation of single SVD dimensions are presented below where the terms are translated to the English language.

Develop (including development, developer etc.) and System (including systems etc.) are two terms that frequently occur together on profitable customers' websites together with the frequent occurrence of following terms (also in stemmed form): Planning, Material, Technique, Build, Product, Machine, Protection, Industry, and Workshop. Further, the following terms should not occur frequently in this context to increase probability of a profitable customer website: Section, History, Experience, Business, Insurance, Energy, Quality and Mobile.

Service (including services, serviced, servicing etc.) and Project (including projects etc.) are two terms that frequently occur together on profitable customers' websites together with the frequent occurrence of following terms (also in stemmed form): Conference, Consulting, Law, Information, Data, Management, Meeting, Union, Contract, Partner, and Staff. Further, the following terms should not occur frequently in this context to increase probability of a profitable customer website: Price, Customer, Offer, Payment, Market, and Tax

The first example could be interpreted as a customer who is interested in workshop equipment and furniture for his production process and the second examples probably shows a customer who is interested in office equipment and furniture. However, it is hard to interpret intuitively why some specific terms should not occur frequently in this context.

4.3 Creating and comparing address lists

In the clustering phase, the EM algorithm identifies seven clusters as well as terms representing cluster labels. Precision (the number of profitable customers' websites over the number of all websites in a cluster) and recall (the number of profitable customers' websites in a cluster over the number of all profitable customers' websites) are computed and clusters are selected with the highest precision values at a recall value over a specific threshold. Terms from the selected cluster labels are used for further processing. As a result, one cluster can be identified with the highest precision and recall value (e.g. 58% precision at 37% recall). Ten terms are extracted from the cluster label (Arbeit, Unternehmen, System, Mitarbeiter, Bereich, Bauen, Technisch, Inhalt, Produkt, Kunde). Heuristically, we are searching for websites that contain at least four of these ten terms by a web search engine API. Companies behind the resulting addresses are manually identified and are added to a list of new potential customers for the mail-order company if they do not occur in the training or validation set. As a result, 160 companies are identified. Comparing these company addresses to addresses from profitable companies from the test set shows that 29 of them (about 18 %) can be classified as profitable (see Table 2). Additionally, 5 of them can be classified as non-profitable. The remaining 127 addresses are used in the acquisition process of the mail-order company. Regardless of the acquisition results - whether further addresses can be classified as profitable or not – a success rate of about 18 % is a good value for this automatically generated list.

	Number of addresses	Relative percentage
Test set A: Addresses generated by this approach		
Non-profitable and non-classified website addresses	131	81.87
Profitable customer group website addresses	29	18.13
Total	160	

Table 2: Overview of the address list characteristics

For evaluation purposes (to determine the frequent baseline), it is critical that the success rate of the traditional acquisition process - by using lists of customers from list brokers - can be estimated. In the year 2008, 3200 company addresses are received from list brokers. The acquisition process leads to 160 profitable customers. The probability that a new customer address leads to a profitable customer, therefore, can be estimated based on an acquisition process as follows: P(A/B) = 160/3200 = about 5 %. According to the acquisition manager, 5 % seems to be a representative value for the success rate of those lists. This low success rate shows the problem for acquisition manager because they get a large amount of addresses but only a few of them lead to profitable acquisitions.

As seen from this example, the density of the number of profitable customers from the address list generated by the approach presented in this paper is about three times larger than the density from list brokers' lists, which companies even have to pay for. Thus, the use of the new potential customer list created by our approach outperforms lists from list brokers and it improves the identification of new profitable business customers while reducing costs.

4.4 Comparing predictive performance

The test set is built on a sample of customers who ordered at least once in the last three years. In contrast to this, the new address list (test set A) additionally includes those who never ordered anything. Thus, it is important to use both, the test set and the test set A to measure the predictive performance.

Overall, Fig. 5 and Fig. 6 show that the predictive performance of the regression model significantly outperforms the baseline because curves from the test sets are situated above the baseline. Fig. 7 also reveals that test sets outperform the baseline at a recall greater than a specific threshold.

Firstly, the cumulative lift curve of the test set and the test set A are above the baseline. Thus, the test sets are able to identify more profitable customers than the baseline within a specific percentile, e.g. the lift value in the top 30 percentile increases from one to 1.21 (test set) and from one to 1.11 (test set A). Secondly, the ROC curve of the test sets lay above the random baseline. Thus, the AUC of the test set (0.6116) and test set A (0.6352) is larger than the baseline (0.5000). This improvement is significant (χ^2=0.02 , d.f.=1, p<0.001). This shows that the model is able to better distinguish profitable from non-profitable customers than the baseline. Thirdly, the precision and recall diagram shows the test sets outperform the baseline at a recall greater than 32 % (test set) and 76 % (test set A). Especially this precision and recall diagram additionally shows that this approach should not be used alone as predictive model but it should be used in addition to conventional acquisition information to transform the acquisition process into a more targeted approach.

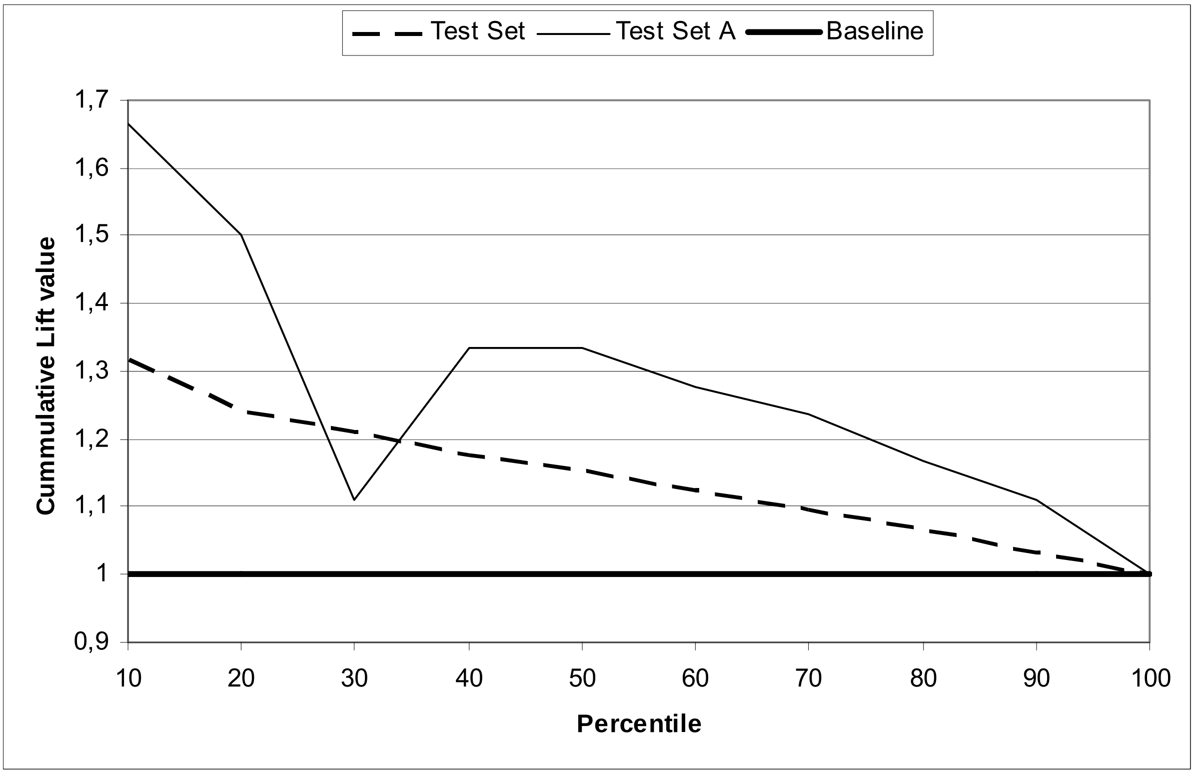

Figure 5: Test sets and baseline lift for the logistic regression model

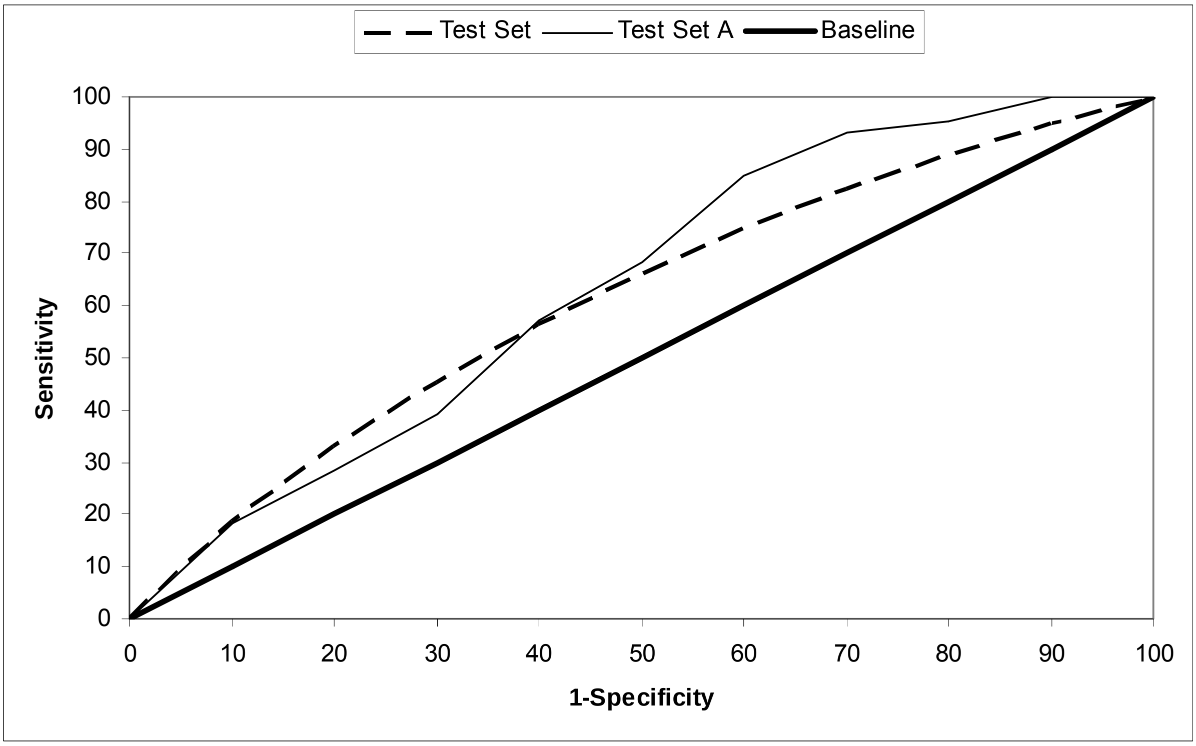

Figure 6: Sensitivity / Specificity Diagram

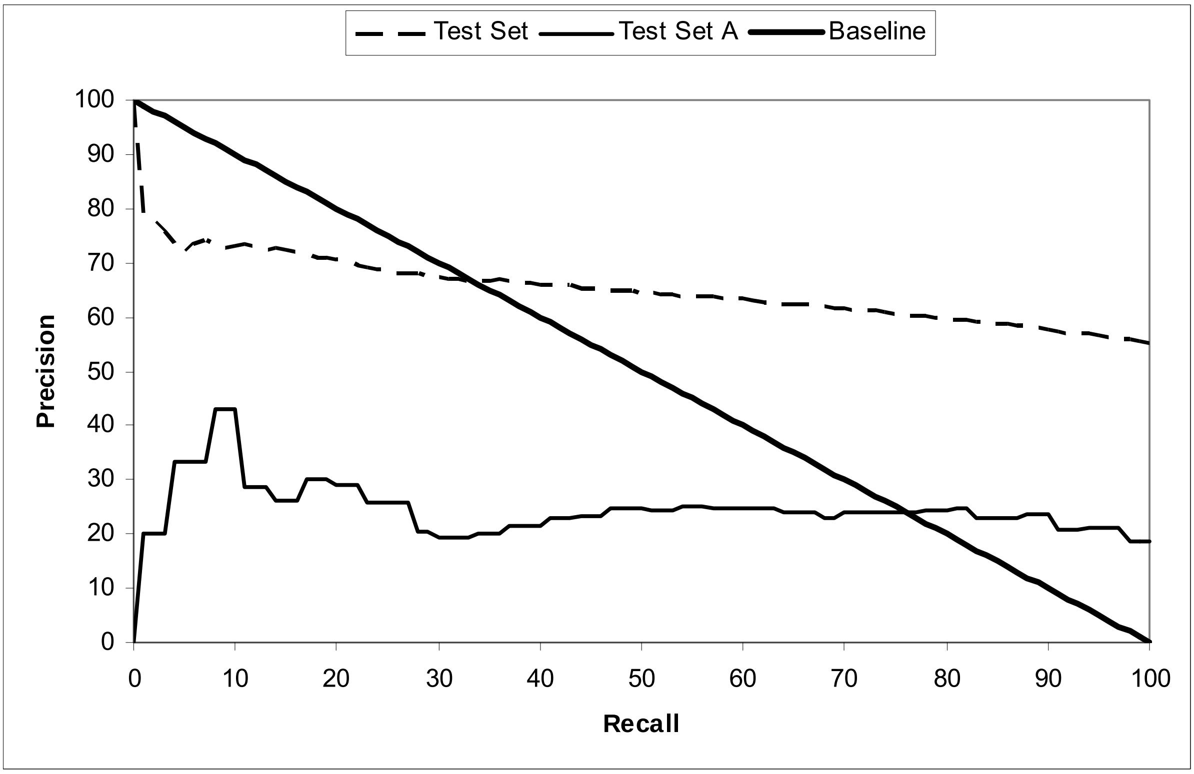

Figure 7: Precision / Recall Diagram

5 Conclusion

In this paper, we demonstrate that using information of existing customers' websites for acquisition purposes helps a B-to-B acquisition manager to identify profitable customers with a higher precision. Consequently, the acquisition process can become more targeted by additionally integrating this textual information. Specific data collection, pre-processing, and dimension reduction steps are required to convert the unstructured textual information into a structured form suitable for profitability prediction. A clustering of websites based on latent semantic concepts leads to the identification of further potential customers that outperforms customers acquired from list brokers by a wide margin. Future work should focus on improving the prediction by adding further unstructured information from existing customers (e.g. e-mails) to the prediction model.

Acknowledgments
We would like to thank the anonymous German mail-order company for their efforts in providing us their data. This project was realized using SAS v9.1.3, SAS Text Miner v5.2, Fraunhofer Idea Web Miner v1.0, and Matlab v7.0.4. For data collection, a self-developed program for web content mining was used. The second author acknowledges the IAP research network grant nr. P6/03 of the Belgian government (Belgian Science Policy).

Bibliography

Allison, P. D. (1999). *Logistic Regression using the SAS System: Theory and Application.* Cary: SAS Institute Inc.

Archer, N., & Yuan, Y. (2000). Managing business-to-business relationships throughout the e-commerce procurement life cycle. *Internet Research: Electronic Networking Applications and Policy,* 10(5), 385-395.

Baecke, P. H., & Van den Poel, D. (2010a). Improving purchasing behavior predictions by data augmentation with situational variables. *International Journal of Information Technology and Decision Making,* Forthcoming.

Baecke, P. H., & Van den Poel, D. (2010b). Data Augmentation by Predicting Spending Pleasure Using Commercially Available External Data. *Journal of Intelligent Information Systems,* Forthcoming.

Bose, I., & Mahapatra, R. K. (2001). Business data mining - a machine learning approach. *Information and Management,* 39(3), 211–225.

Bucklin, R. E., & Gupta, S. (1992). Brand choice, purchase incidence and segmentation: an integrated modeling approach. *Journal of Marketing Research*, 29(2), 201–215.

Chen, M.-C., Chiu, A..L., & Chang, H.H. (2005). Mining changes in customer behavior in retail marketing. *Expert Systems with Applications,* 28(4), 773-781.

Chen, M.-Y., Chu, H.-C. & Chen, Y.-M. (2010). Developing a semantic-enable information retrieval mechanism. Expert Systems with Applications, 37(1), 322-340

Coussement, K., & Van den Poel, D. (2009). Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. *Expert Systems with Applications,* 36, 6127-6134.

Coussement, K., & Van den Poel, D. (2008). Integrating the voice of customers through call center emails into a decision support system for churn prediction. *Information & Management,* 45, 164-174.

Coviello, N., Brodie, R.J., & Munro, H. (1997). Understanding contemporary marketing: Development of a classification scheme. *Journal of Marketing Management,* 13(6), 501-522.

De Bock, K. W., & Van den Poel, D. (2009). Predicting website audience demographics for web advertising targeting using multi-website clickstream data. *Fundamenta Informaticae,* 97, 1-19.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science,* 41(6), 391-407.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics,* 44(3), 837–845.

Engler, J., & Kusiak, A. (2010). Mining Authoritativeness of Collaborative Innovation Partners. *International Journal of Computers, Communications & Control,* V(1), 42-51.

Gericke, W., Thorleuchter, D., Weck, G., Reiländer F., & Loß, D. (2009). Vertrauliche Verarbeitung staatlich eingestufter Information - die Informationstechnologie im Geheimschutz. *Informatik Spektrum,* 32(2), 102-109.

Greiff, W. R. (1998). A theory of term weighting based on exploratory data analysis. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), Proceedings of the 21st *SIGIR Conference*. New York: ACM, pp. 11-19.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology,* 143(1), 29-36.

Kim, Y. S. (2006). Toward a successful CRM: variable selection, sampling and ensemble. *Decision Support Systems,* 41(2), 542–553.

Lee, K.-C., & Chung, N. (2003). Identification of Customer Segmentation Strategies by Using Machine Learning-Oriented Web-mining Technique. *IE Interfaces,* 16(1), 54-62.

Menon, A., Homburg, C., & Beutin, N. (2005). Understanding customer value in business-to-business relationships. *Journal of business-to-business mark*eting*,* 12(2), 1-38.

Naude, P., & Holland, C. (1996). *Relationship Marketing.* London: Paul Chapman Publishing. pp. 40-54.

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. (2006). Defection detection: measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research,* 43(2), 204–211.

Pan, S. L., & Lee, J. N. (2003). Using e-CRM for a unified view of the customer. *Communications of ACM,* 46(4), 95–99.

Park, Y.-J., & Chang, K.-N. (2009). Individual and group behavior-based customer profile model for personalized product recommendation. *Expert Systems with Applications,* 36(2), 1932-1939.

Reinartz, W., & Kumar, V. (2003). The impact of customer relationship characteristics on profitable lifetime duration. *Journal of Marketing,* 67(1), 77–99.

Salton, G., Allan, J., & Buckley, C. (1994). Automatic structuring and retrieval of large text files. *Communications of the ACM,* 37(2), 97–108.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management,* 24(5), 513–523.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation,* 28(1), 11-21.
Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010a). Mining Ideas from Textual Information. *Expert Systems with Applications,* 37(10), 7182-7188.
Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010b). A compared R&D-based and patent-based cross impact analysis for identifying relationships between technologies. *Technological Forecasting and Social Change,* 77(7), 1037-1050.
Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010c). Extracting consumers needs for new products - A web mining approach. In *Proceedings WKDD 2010* (p. 441), Los Alamitos: IEEE Computer Society.
Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010d). Mining innovative ideas to support new product research and development. In H. Locarek-Junge, & C. Weihs (Eds.), *Classification as a Tool for Research* (pp. 587-594). Berlin: Springer-Verlag.
Thorleuchter, D. (2008). Finding Technological Ideas and Inventions with Text Mining and Technique Philosophy. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data Analysis, Machine Learning, and Applications* (pp. 413-420). Berlin: Springer-Verlag.
Van den Poel, D., & Buckinx, W. (2005). Predicting Online-Purchasing Behavior. *European Journal of Operational Research,* 166(2), 557-575.
Verhoef, P. C., Venkatesan, R., McAlister, L., Malthouse, E. C., Krafft, M., & Ganesan, S. (2010). CRM in Data-Rich Multichannel Retailing Environments: A Review and Future Research Directions. *Journal of Interactive Marketing,* 24(2), 121-137.
Wangenheim, F., & Bayon, T. (2007). The chain from customer satisfaction via word-of-mouth referrals to new customer acquisition. Journal *of the Academy of Marketing Science,* 35, 233-249.
Yen, E., & Lina L.-H. (2010). Rubik's cube watermark technology for grayscale images. Expert Systems with Applications, 37(6), 4033-4039.
Zhong, J., & Li, X. (2010). Unified collaborative filtering model based on combination of latent features. Expert Systems with Applications, 37(8), 5666-5672.
Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort.* Cambridge: Addison-Wesley.