LSI based Profitability Prediction of new Customers

Dirk Thorleuchter
Fraunhofer INT - Institute for Technological Trend Analysis
Appelsgarten 2,
53879 Euskirchen, Germany
Thorleuchter@int.fraunhofer.de

Dirk Van den Poel
Ghent University, Faculty of Economics and Business Administration
Tweekerkenstraat 2,
9000 Gent, Belgium
Dirk.VandenPoel@UGent.be

Anita Prinzie
Ghent University, Faculty of Economics and Business Administration
Tweekerkenstraat 2,
9000 Gent, Belgium
Anita.Prinzie@UGent.be

ABSTRACT
This study investigates the impact of website information from existing business customers of a company on the prediction of the profitability of new business customers. Estimating the profitability of new customers is a well-known problem in acquisition management. Thus, the results of this study can be used to advance the acquisition process of a company.

A methodology is provided and the acquisition process of a mail order company is supported by use of this methodology. This case study shows that information of existing customers' websites can be used as successful classifier by modeling the profitability prediction. Thus, new profitable business customers can be acquired by the company.

Keywords
Text mining, Web mining, Information Extraction, Classification.

1. INTRODUCTION
Today, we see a change from a product-centered to a customer-centered environment [6]. Thus, it is important to consider of information about the customers who bought the products [23].

The acquisition of new customers normally is time- and cost-consuming [24]. One reason for this is that many identified customers are not profitable customers in future. Thus, it is important to identify the profitability of new customers.

In this study, we connect both approaches as described above. Additionally, we focus on a Business-to-Business (B2B) environment where profitable business customers should be identified based on information about the customers who bought the products.

A new approach is proposed that predicts new business customers as profitable. This is done by use of information about existing customers. To define customer's profitability (value) we focus on literature where this term in a B2B context is discussed. As a result, customer's profitability is defined as exceeding a sales volume threshold. We are aware that different profitability definitions in literature exist that consider core benefits, add-on benefits, purchasing price, acquisition costs, operations costs, etc. [19]. However, the used definition is just a subset of the further definitions from literature. Thus, if information about existing customers can be used as a successful classifier for the profitability prediction of new customers - in terms of the used profitability definition - then the information can be used as a successful classifier concerning further definitions, too. Thus, we used this simple customer's profitability definition.

In a fist step, we extract information of existing customers from a customer relationship management (CRM) system. This information consists of different aspects e.g. the volume of sales of each existing customer. Based on this information, existing customers are classified as profitable or non-profitable customers by comparing their volume of sales over a specific period to a specific threshold. Further, information about website addresses of existing customers are collected.

In a second step, textual information is crawled from the websites. The information is preprocessed and is analysed by latent semantic indexing (LSI). As a result, specific textual features (concepts) are identified.

In a third step, we build a logistic regression model to show the success of using this information for predicting the profitability of new customers.

2. Related Work
Marketing can be distinguished between two different approaches. The transactional marketing focuses on single point-of-sale transactions [7] while the relational marketing focuses on customer retention and satisfaction [17, 22]. Relational marketing considers the relationships between companies and customers [20]. Thus, in relational marketing the main principle in the acquisition of business customers is the information exchange [14, 17, 21].

Further aspects on the acquisition of new business customers based on e-commerce (as a new information exchange technology) [2] and on word-of-mouth referrals [29].

In the field of web mining, related work can be seen that identifies customer's behaviors in the internet [3, 4, 18, 27] and that identifies collaborative partners [10].

In contrast to previous work, this approach investigates the impact of website information on the acquisition of new business customers as a contribution to the literature concerning acquisition of business customers.

3. METHODOLOGY

We collect textual information from websites of customers and split them in a test and in a training set. Information from the training set is transformed to a term-website matrix in a pre-processing phase. The dimension of the term-website matrix is reduced and hidden textual patterns (concepts) are identified. Then, information of the test set is projected into the concept-space of the dimension reduced term-website matrix. Based on this concept-space matrix, we build a prediction model and we show that the extracted concepts are successful in the profitability prediction of new customers. The methodology is shown in Fig. 1.

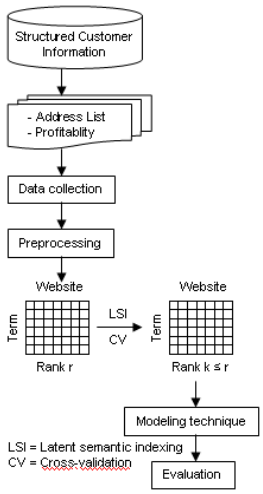

Figure 1. Methodology of the approach.

3.1 Data Collection

We used a web mining approach to extract textual information from customers' websites. Fig. 2 shows different steps in the web mining approach. In a first step, we calculate the profitability of each customer concerning the profitability definition in Sect. 1. In a second step, we identify customers' company websites. However, just relevant web pages from each website are selected to extract information. These are the starting page, web pages with a high page rank as calculated by the search engine Google, and web pages where specific key words occur. Web pages that contain organizational information e.g. 'disclaimer' or 'privacy / data protection policy' are not selected. We use web services [5] and web based advanced programming interfaces (APIs) [32] to collect the information.

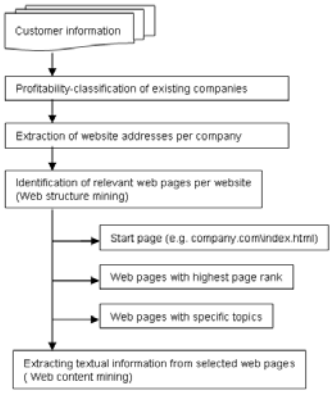

Figure 2. Web mining approach to extract information from customers' websites.

3.2 Preprocessing

After extracting textual information from the websites of the business customers, this information is prepared and it is filtered to represent the information as term vectors in vector space model [30]. In a term vector weighting step, each company is represented as a vector of weighted frequencies of designated words. The number of distinct terms in the dictionary determines the size of the vector. Each vector component represents the semantic importance of a corresponding term as set to its weighted term frequency. Then, a term-by-company matrix is built.

3.2.1 Text preparation

The text preparation phase includes cleaning the raw text (e.g. removing scripting code, images, html-, and xml-tags from the textual information), deleting of specific characters, and correcting of typographical errors by use of a dictionary [33]. Textual information is separated in terms with tokenization [31] where the term unit is word and case conversion is done (terms are converted in lower case whereby the first sign is capitalized) [11].

3.2.2 Term filtering

To reduce the number of different terms in the text, term filtering methods [16] are used. The selection of terms that can be assigned to a specific syntactic category (nouns, verbs, adjectives and adverbs) is done by part-of-speech tagging. Further terms as well as stop words that are non-informative are discarded [34]. A dictionary-based stemmer is used to convert each term to its stem. If a term is not in the dictionary then a set of production rules is applied to give each term a correct stem. In literature, Zipf distribution shows that half of terms appear only once or

twice [30]. Thus, these terms are also non-informative and they are deleted under these thresholds.

3.2.3 Term vector weighting

For each document, a term vector in vector space model is build using the selected terms [29]. Term vectors of weighted frequencies are used instead of raw frequencies because as shown in literature, this leads to significant improvements [28]. A large weight is assigned to a term that occurs frequently in a specific document but rarely in the document collection [26]. Let $w_{i,j}$ be the weight assigned to term i in document j, let n be the number of documents, let m be the number of terms in the vectors (m-dimensional term vectors), and let df_i be the number of documents that contain term i [25]. Then, the weight is calculated by term frequency $tf_{i,j}$ times inverse web page frequency idf_i divided by a length normalization factor [16].

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

3.2.4 Term Vector Aggregation

Each customer's company website consists of several web pages. With term vector aggregation all these web pages that belong to the same customer's company website are aggregated to build a term-by-website matrix.

Let $w_{i,k}$ be the weight of term i in web page k and let r be the number of web pages belonging to the same customer's company website j [6]. Then the aggregated $w_{i,k}$ can be calculated by

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

3.3 Concept identification with LSI and singular value decomposition

The created term-by-website matrix is nearly unmanageable because of its high dimensionality. However, the dimensionality can be reduced by considering the fact that most of the weights are zero and by considering the fact that a semantic generalization helps to group terms together into concepts and thus, to reduce the number of distinct terms. This can be done by using latent semantic indexing (LSI) combined with singular value decomposition (SVD) as method [8].

Let A be the term-by-website (m x n) matrix and r be its rank ($r \leq \min(m,n)$) [12]. The SVD of A can be transformed into a product of three matrices, the term-concept similarity (m x r) matrix U, the concept-website similarity (n x r) matrix V, and a diagonal (r x r) matrix Σ containing positive singular values of matrix A.

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

The rank r of A can be reduced by retaining the first k ($k \leq r$) singular values in Σ and by discarding further positive singular values. The choice of k is critical because it influences the predictive performance. Thus, several rank k-models are constructed on the training examples to select the most favourable rank-k model. A prediction model as described in Sect. 3.4 calculates the predictive performance by integrating the test examples into the same semantic subspace as created by the training examples [8].

3.4 Prediction Modelling

For modelling, we use logistic regression where a maximum likelihood function is produced and maximized [1]. Advantages of logistic regression are the simplicity of the concept [9], the availability of a closed-form solution for the probabilities, and the robustness of the predictive results [13].

Let $T = \{(x_i, y_i)\}$ be a training set, let $i = \{1,2,...,N\}$ be an index, let $x \in R^n$ be an n-dimensional input vector (a concept-company vector) as representative for companies load on the concepts, let w be the parameter vector, let w_0 be the intercept, let $x_i \in R^n$ be input data, and let $y_i \in \{0,1\}$ be the corresponding binary target labels (company information is assigned to a specific security label or not). Then, the probability $P(y = 1 | x)$ is estimated by

$$P(y = 1 \mid x) = \frac{1}{1 + \exp(-(w_0 + wx))} \qquad (4)$$

3.5 Evaluation criteria

An evaluation of the prediction model is done to show that latent semantic concepts from the companies can be used to predict the profitability of the companies in the test set. The performance is examined by use of well-known criteria: the lift, the sensitivity and specificity, and the misclassification rate.

The most commonly used performance measure is the lift that measures the increase in density of companies that are successfully assigned to the group of profitable customers relative to the density of companies that belong to the group of profitable customers in total. Based on the companies that belong to the group of profitable

companies, TP (true positive) is defined as the number of correctly assigned companies and FN (false negative) is defined as the number of incorrectly assigned companies. Based on the companies that do not belong to the group of profitable customers, TN (true negative) is defined as the number of correctly assigned companies and FP (false positive) is defined as the number of incorrectly assigned companies. The sensitivity (TP/(TP+FN)) is defined as the proportion of positive cases that are predicted to be positive and the specificity (TN/(TN + FP)) is defined as the proportion of negative cases that are predicted to be negative.

A two dimensional plot of the sensitivity versus (1-specificity) is named the receiver operation characteristics curve (ROC). The area under the receiver operating characteristics curve (AUC) is used to compare the performance of binary classification models [15]. A cross-validated misclassification rate is used to calculate the optimal number of concepts.

4. EMPIRICAL VERIFICATION

4.1 Research data

For empirical verification, we use the CRM system of a large German mail-order company. The company acts in a B2B environment and existing CRM information for each business customer include website address and volume of sales.

Information concerning 50.000 business customers is collected in a first step. By comparing the extracted website addresses, several customers can be identified that belong to the same company. The volume of sales of these customers are aggregated to calculate a volume of sales per company. This reduces the number of examples to 20.000 companies. A further reduction is necessary, because websites of companies are written in different languages. This causes problems by the identification of textual patterns in the collection of all websites. Thus, companies - where a German language website exists - are selected. As a result, 11.856 companies are split in a training set of 8.299 and in a test set of 3.557 examples.

Further, the aggregated volume of sales per company is used to assign a company to the positive (profitable) or negative (non-profitable) examples.

Table 1 shows the characteristics of the data.

Table 1. Overview of the website characteristics

	Number of customer groups	Relative percentage
Training set (including validation set):		
Non-profitable customer group website addresses	3.781	45,56
Profitable customer group website addresses	4.517	54,44
Total	8.299	
Test set:		
Non-profitable customer group website addresses	1.597	44,92
Profitable customer group website addresses	1.959	55,08
Total	3.557	

4.2 Optimal dimension selection and interpretation

The rank of the high dimensional term-by-company matrix is reduced to obtain the optimal number of SVD dimension (concepts). Thus, a cross-validation procedure on the training data was applied. The x-axis in Fig. 3 represents the number of concepts and the y-axis represents the cross-validated misclassification rate. It can be seen that in the range of 1–50 concepts, the cross-validated misclassification rate was decreasing rapidly. From 50 concepts on, it was decreasing 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 misclassification rate hardly decreased. Thus, 150 concepts were chosen as the optimal number of SVD dimension in our study.

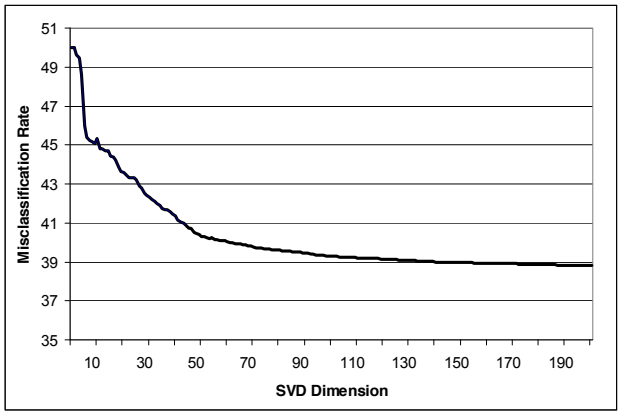

Figure 3. The impact of SVD Dimension on the misclassification rate.

The SVD dimensions represent collocations that means terms that occur together more frequently than it would be expected by chance. Additionally, the SVD dimensions also represent terms that do not occur together with the

collocations even if it would be expected by chance. The terms represent words in German language and in stemmed form because a German stemmer is used as described in Sect. 3.2.2. Below, an example is presented where terms are translated to the English language.

In the positive examples, the stemmed term ‘develop’ (that includes development, developer etc.) and the stemmed term ‘system’ occur together with specific terms e.g. 'planning’, ‘material’, ‘technique’, ‘build’, ‘product’, ‘machine’, and ‘workshop’. However, they do not occur together with further specific terms e.g. ‘section’, ‘history’, ‘insurance’, and ‘energy’. This probably could be seen as a profitable business customer who is interested in workshop equipment and furniture for his production process.

Comparing predictive performance

The predictive performance of the regression model is compared to the frequent baseline. Fig. 4 (the cumulative lift curve) and Fig. 5 (ROC curve) show that the test set outperforms the baseline. The AUC of the test set (63,52) is significantly larger than that of the baseline (50.00). Thus, this approach is able to better distinguish profitable from non-profitable customers than the baseline.

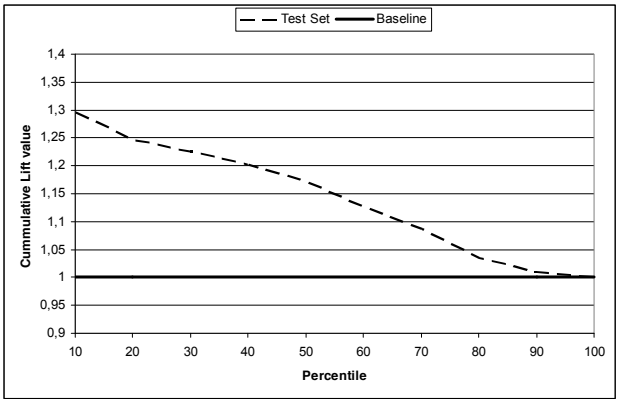

Figure 4. Test set and baseline lift for the logistic regression model

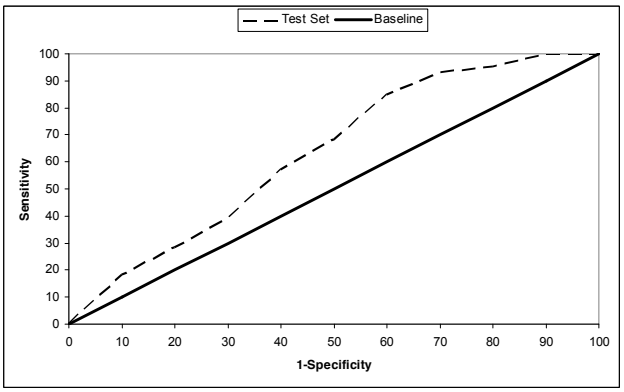

Figure 5. ROC curve

5. CONCLUSION

Using information of existing customers’ websites for acquisition purposes supports acquisition manager by the identification of profitable customers. This prediction should be used in addition to further acquisition means to improve the acquisition process in a company. Avenues of future research could be the improving of the prediction performance by integrating further unstructured information in this approach e.g. information from web logs (blogs) or e-mails.

6. ACKNOWLEDGMENTS

This work is supported by an anonymous German Mail Order Company that provides us data about their business customers. For the processing, SAS v9.1.3, SAS Text Miner v5.2, Fraunhofer Idea Web Miner v1.0, and Matlab v7.0.4 are used.

7. REFERENCES

[1] P.D. Allison. Logistic Regression using the SAS System: Theory and Application. SAS Institute Inc., Cary, NC, 1999.

[2] N. Archer and Y. Yuan. Managing business-to-business relationships throughout the e-commerce procurement life cycle. Internet Res 10(5):385-395, 2000.

[3] I. Bose and R.K. Mahapatra. Business data mining—a machine learning approach. Inform Manage 39(3):211-225, 2001.

[4] R.E. Bucklin and S. Gupta. Brand choice, purchase incidence and segmentation: an integrated modeling approach. J Marketing Res 29(2):201-215, 1992.

[5] D. Carl, J. Clausen, M. Hassler, and A. Zund. Mashups programmieren. pages 51-53. O'Reilly, Germany, 2008.

[6] K. Coussement and D. Van den Poel. Integrating the voice of customers through call center emails into a decision support system for churn prediction. Inform Manage 45(3): 164-174, 2008.

[7] N. Coviello, R.J. Brodie, and H. Munro. Understanding contemporary marketing: Development of a classification scheme. J Marketing Man, 13(6):501-522, 1997.

[8] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis, J Am Soc Inform Sci 41(6):391-407, 1990.

[9] E.R. DeLong, D.M. DeLong, and D.L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a

nonparametric approach. Biometrics 44(3):837-845, 1988.

[10] J. Engler and A. Kusiak. Mining Authoritativeness of Collaborative Innovation Partners. Int J Comput Commun V(1):42-51, 2010.

[11] R. Feldman and J. Sanger. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, page 318. Cambridge University Press, 2007.

[12] S.I. Gass and T. Rapcs. Singular value decomposition in AHP. Eur J Oper Res 154:573-584, 2004.

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

[14] H. Hakansson. Corporate Technological Behaviour: Co-operation and Networks. Routledge, London, 1989.

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

[16] A. Hotho, A. Nürnberger, and G. Paaß. A Brief Survey of Text Mining. LDV Forum 20(1):19-26, 2005.

[17] Y.S. Kim. Toward a successful CRM: variable selection, sampling and ensemble. Decis Support Syst 41(2): 542-553, 2006.

[18] K. Lee and N. Chang. Identification of Customer Segmentation Strategies by Using Machine Learning-Oriented Web-mining Technique. IE Interfaces 16(1):54-62, 2003.

[19] A. Menon, C. Homburg, and N. Beutin. Understanding customer value in business-to-business relationships. J Bus-Bus Mark 12(2):1-38, 2005.

[20] R.M. Morgan and S.D. Hunt. The commitment-trust theory of relationship marketing. J Marketing 58:20-38, 1994.

[21] P. Naude and C. Holland. Relationship Marketing, pages 40-54. Paul Chapman Publishing, London, 1996.

[22] S.A. Neslin, S. Gupta, W. Kamakura, J. Lu, and C. Mason. Defection detection: measuring and understanding the predictive accuracy of customer churn models. J Marketing Res 43(2):204-211, 2006.

[23] S.L. Pan and J.N. Lee. Using e-CRM for a unified view of the customer. Commun ACM 46(4):95-99, 2003.

[24] W. Reinartz, and V. Kumar. The impact of customer relationship characteristics on profitable lifetime duration. J Marketing 67(1):77-99, 2003.

[25] G. Salton, J. Allan, and C. Buckley. Automatic structuring and retrieval of large text files. Commun ACM 37(2):97-108, 1994.

[26] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. Inform Process Manag 24(5):513-523, 1988.

[27] J. Shen, L. Xing, and J. Peng. Study and Application of Web-based Data Mining in E-Business. IEEE Xplore, 2007.

[28] K. Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. J Doc 28(1):11-21, 1972.

[29] D. Thorleuchter, D. Van den Poel, and A. Prinzie. Mining Ideas from Textual Information. Expert Syst Appl 37(10):7182-7188, 2010.

[30] D. Thorleuchter, D. Van den Poel, and A. Prinzie. A compared R&D-based and patent-based cross impact analysis for identifying relationships between technologies. Technol Forecast Soc Change 77(7): 1037-1050, 2010.

[31] D. Thorleuchter, D. Van den Poel, and A. Prinzie. Mining Innovative Ideas to Support new Product Research and Development. In: H. Locarek-Junge and C. Weihs (editors): Classification as a Tool for Research, Proceedings of the 11th IFCS Biennial Conference and 33rd Annual Conference of the Gesellschaft für Klassifikation e.V. Springer, Berlin-Heidelberg-New York, 2010.

[32] D. Thorleuchter, D. Van den Poel, and A. Prinzie. Extracting Consumers Needs for New Products. In: Proceedings of WKDD 2010, pages 440-443. IEEE Computer Society, Los Alamitos, CA, 2010.

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

[34] D. Thorleuchter. Finding technological ideas and inventions with text mining and technique philosophy. In: L. Schmidt-Thieme (editor): Data Analysis, Machine Learning, and Applications, pages 413-420. Springer, Berlin-Heidelberg-New York, 2008.

[35] F. Wangenheim, and T. Bayon. The chain from customer satisfaction via word-of-mouth referrals to new customer acquisition. J of the Acasd. Mark. Sci. 35:233-249, 2007.

[36] G.K. Zipf. Human Behaviour and the Principle of Least Effort. Addison-Wesley, Cambridge, MA, 1949.