Analyzing Website Content for Improved R&T Collaboration Planning

Dirk Thorleuchter[1] and Dirk Van den Poel[2]

[1] Fraunhofer INT, Appelsgarten 2, D-53879 Euskirchen, Germany
Dirk.Thorleuchter@int.fraunhofer.de
[2] Ghent University, Faculty of Economics and Business Administration, Tweekerkenstraat 2, B-9000 Gent, Belgium
Dirk.Vandenpoel@ugent.be

Abstract. A well-known problem in research and technology (R&T) planning is the selection of suited R&T collaboration partners. We investigate the use of textual information from the website content of possible collaboration candidates to identify their suitability. This improves the selection of collaboration partners and it enables a successful processing of R&T-projects. In a case study 'defense R&T', organizations and companies that have proven their suitability as collaboration partner in former R&T projects are selected (positive examples) as well as organizations and companies that have not. Latent semantic indexing with singular value decomposition and logistic regression modeling is used to identify semantic textual patterns from their websites' content. As a result of prediction modeling, some of these textual patterns are successful in predicting new organizations or companies as (un-) suited R&T collaboration partners. These results support the acquisition of new collaboration partners and thus, they are valuable for the planning of R&T.

Keywords: Collaboration, Research, Technology, Semantic Classification, Text Mining, Defense.

1 Introduction

At present, projects in research and technology (R&T) are often complex [1] and their successful processing requires the collaboration with external partners (organizations or companies) [2]. Based on partners' capabilities, organizations or companies can be classified as suited or unsuited collaboration partners [3]. Thus, an important aspect for the R&T planning of an organization is the selection of suited collaboration partners to enable a successful processing of R&T-projects [4].

Normally, the selection of suited collaboration partners is done by considering gained experience of the organization where the potential candidates have already proven their suitability as collaboration partners in former R&T-projects [5]. This often leads to the selection of already known partners and this excludes the selection of new and unknown partners for future projects [6].

Literature introduces the use of textual information from companies' websites to predict the success of companies in business-to-consumer (B2C) environment [7,8]. Recent research has shown that this information also can be used to predict the success of business-to-business (B2B) commerce transactions [9,10]. Whereas R&T collaborations represent specific B2B commerce transactions, textual website content possibly can be used to predict suited R&T collaboration partners.

Thus, we propose a new methodology that creates such a prediction model to show its success for this specific B2B area. In detail, we investigate textual website content of suited R&T collaboration partners that have proven their suitability in current or former projects (positive examples). We also investigate textual website content of organizations and companies that have not involved in R&T projects in present or past (negative examples). Semantic textual patterns are extracted from the content and prediction modeling is applied.

A case study shows that specific textual patterns can be used for a successful assigning of defense based R&T organizations and companies to the positive or negative examples. As a result, some textual patterns represent existing success factors as known from e-commerce literature while others represent new success factors that specifically can be used to predict R&T collaboration partners. Thus, the created prediction model supports research planners by acquiring new R&T collaboration partners.

2 Methodology

This new methodology consists of several steps as depicted in Fig. 1. The first step is to create two lists of organizations and companies as suited or as unsuited collaboration partners based on experiences of the past. Elements of these lists are divided in training set and test set.

In a second step, the websites standing behind the organizations and companies are identified and textual information from websites' content is crawled. Crawling is done based on a web content mining approach because each website consists of several web pages and only some of them are relevant [11]. Trivial web pages e.g. 'disclaimer' or 'sitemap' are discarded. The identification of relevant web pages is done by selecting the starting page of a website as well as by selecting the four most frequently visited web pages. Whereas information about page visitors is not available, we use results from Google ranking algorithm as indicator for highly visited web pages where high ranked web pages are selected [12]. Additionally, it is supposed that website information about R&T activities might be relevant for identifying the organization or company as suited collaboration partner. Thus, web pages where specific terms ('research', 'technology', 'high-tech', 'science' etc.) occur are selected.

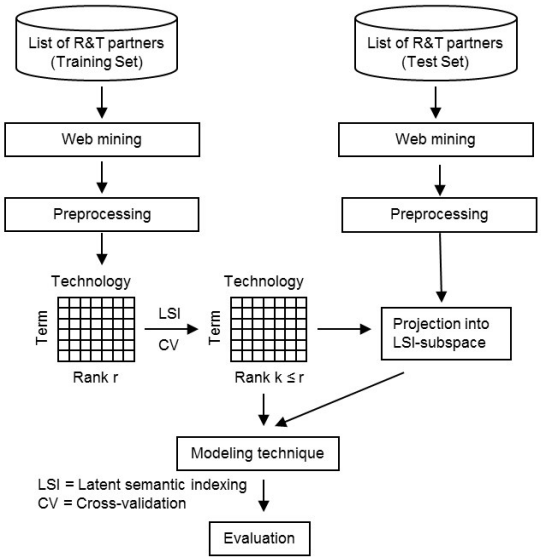

Fig. 1. Processing of the approach.

The third step prepares the information crawled from the selected web pages to obtain a specific granularity [13-17]. Typographical errors are corrected, tokenization is used, and all terms are converted in lower case [18]. Further, part-of-speech tagging, stop word filtering, and stemming is applied as well as Zipf distribution [19]. A term vector for each web page is created based on vector space model [20]. Vectors are aggregated to build a vector for each website. Weighted term frequencies are calculated for the vector components because they lead to a significant improvement [21].

The weight of a term is at its maximum if the term occurs very frequently in a small number of websites and if the term does not occur in the other websites. The weights are calculated by multiplying term frequency and inverse object frequency and by dividing it with a length normalization factor [22]. As a result, a term-by-website matrix is created based on the vectors. The matrix consists of high dimensionality and many components of the matrix are zero.

In the fourth step, latent semantic indexing (LSI) is used to reduce the high dimensionality [23]. This summarizes textual patterns concerning their aspect of meaning (semantic textual patterns). Dimensionality reduction can be done by matrix factorization techniques. Thus, we use singular value decomposition (SVD) as the commonly used matrix factorization technique. The term-by-website matrix A is a (m x n) matrix where m equals the number of terms and n the number of websites. The rank r of matrix A is smaller than min(m,n). To reduce the rank r of the term-by-website matrix A to k, SVD splits the matrix in the product of three matrices: a) a term-dimension (m x r) matrix U that shows the impact of each term on the semantic dimensions, b) a website-dimension (n x r) matrix V that shows the impact of each website on the semantic dimensions, and c) a positive singular values (r x r) matrix Σ where the singular values are positioned on the diagonal.

$$A = U \Sigma V^t \quad (1)$$

The rank r of the term-by-website matrix to k can be reduced by selecting the first k singular values and by discarding further singular values.

The value of k is selected based on an optimal predictive performance of the semantic textual patterns. Several rank-k models are constructed using a parameter-selection procedure based on backward selection as main approach in stepwise regression [24,25]. For each rank-k model, a fivefold cross-validation is applied on the training set. We use logistic regression [26] as predictive modeling technique where a maximum likelihood function is produced and maximized. As performance measure, the area under the receiver operating characteristics (ROC) curve (AUC) is used [27-29]. To obtain the optimal number of k, the predictive performance of the model as calculated by the cross-validated AUC is optimized [30].

The fifth step applies prediction modeling on the test set to measure the performance of the approach. Logistic regression is used to identify textual patterns that are characteristic for both the positive and the negative examples. These textual patterns represent success factors and they are compared to existing website success factors known from literature [12]. As a result, existing success factors from literature can be identified and it can be shown that they are also successful in predicting R&T collaboration partners. Further, success factors that are not mentioned in literature can be identified to introduce them as new success factors to the scientific community. For the evaluation, commonly used evaluation criteria are used in a last step: the cumulated lift [31], the precision and recall [32], the cross-validated AUC, as well as the sensitivity and specificity [33].

3 Case Study

In a case study 'Defense R&T', organizations and companies are identified that have been involved as collaboration partners in R&T projects funded by German Ministry of Defense [34,35]. Whereas a participation in a project is always agreed by manual evaluation of research planners, they are assigned to the positive examples. A large number of further organizations and companies with no relation to defense R&T projects are taken over from an existing study in literature [9] as negative examples. The aim of the case study is to identify semantic textual patterns that can be used to predict new (unseen) organizations and companies as member of positive examples. To prevent language translation problems, the case study is restricted to websites in German language. 5.315 positive examples are identified and they are split in 3.720 training examples and 1.595 test examples. 24.897 negative examples are selected and split in 17.428 training examples and in 7.469 test examples. The relative percentage is 17,5% (positive examples) to 82,5% (negative example).

An optimal predictive performance was reached by setting k to 24 dimensions. Four dimensions can be identified as classifier for the positive examples while two dimensions have large impact on the negative examples. Each dimension consists of several textual patterns and thus, it might represent one or several success factors.

Based on a comparing of the identified success factors from the seven dimensions to the success factors from e-commerce literature, we have found three factors that are successful to predict the positive examples (the trustworthiness, reliability and security') and two factors that are successful to predict the negative examples (the website interactivity and the technical support of a website).

We also have identified the occurrence of further success factors (High quality of website content, Usefulness of website content, Website Usability, Website responsiveness, Wide product choice, Website customization, Website promotion initiatives, Addressing emotions, and Website design). However they cannot be used to predict the positive or the negative examples.

Last, further factors have been identified from the dimensions that are not mentioned in literature before. These factors are also successful in predicting the positive examples: These factors can be labeled by 'Data protection policy', 'Conflict monitoring', 'International collaboration', 'Emergency response', and 'Demonstrators und Prototypes'. This is because these factors are of particular interest in defense R&T context.

4 Evaluation

The methodology is applied. Based on the results of the test set, an evaluation is done by use of well-known evaluation criteria: the lift, the precision and recall, the ROC curve that is based on sensitivity and specificity, and the AUC. The results from the test set are compared to the frequent baseline.

The cumulated lift lays above the frequent baseline and it shows that the test set is able to identify more suited collaboration partners than the random baseline within a specific percentile, e.g. the cumulated lift value in the top 10 (30) percentile increases from one to 1.24 (1.13). In the precision and recall diagram, the test set outperforms the baseline at a recall greater than 40 %. The ROC curve of the test set lies above the random baseline and the cross validated AUC of the test set (0.6245) is larger than the baseline (0.5000). This improvement is significant (χ^2=0.02, d.f.=1, $p<0.001$). Overall, the evaluation shows that the proposed methodology outperforms the baseline.

5 Conclusion

The results show that using information from websites' content helps research planners to identify suited R&T collaboration partners with a higher precision. Thus, the proposed methodology can support decision makers to improved research planning. It is shown that some e-commerce success factors are also successful in R&T collaboration planning and that some e-commerce success factors are not successful. Further, new success factors are identified and proposed to the scientific literature.

The AUC of the test set (62%) shows that this approach should not be used alone as predictive model for a collaboration decision. However, the identification process can

become more targeted by additionally integrating this information in the decision process as one variable among others.

References

1. Thorleuchter, D., Van den Poel, D.: Technology classification with latent semantic indexing. Expert Syst. Appl. 40 (5), 1786-1795 (2013)
2. Bammer, G.: Enhancing research collaborations: Three key management challenges. Res Pol 37 (5), 875-887 (2008)
3. Heinze, T., Kuhlmann, S.: Across institutional boundaries?: Research collaboration in German public sector nanoscience. Res Policy 37 (5), 888-899 (2008)
4. Thorleuchter, D., Van den Poel, D.: Semantic Technology Classification. In: International Conference on Uncertainty Reasoning and Knowledge Engineering, pp. 36–39. IEEE Press, New York (2011)
5. Van Rijnsoever, F.J., Hessels, L.K.: Factors associated with disciplinary and interdisciplinary research collaboration. Res Policy 40 (3), 463-472 (2011)
6. He, Z.L., Geng, X.S., Campbell-Hunt, C.: Research collaboration and research output: A longitudinal study of 65 biomedical scientists in a New Zealand university. Res Policy 38 (2), 306-317 (2009)
7. Ballantine, J., Levy, M., Powell, P.: Evaluating information systems in small and medium-sized enterprises: issues and evidence. Eur. J. Inform. Syst. 7, 241– 251 (1998)
8. Serafeimidis, V., Smithson, S.: Information systems evaluation as an organizational institution - experience from a case study. Inform. Syst. J. 13, 251–274 (2003)
9. Thorleuchter, D., Van den Poel, D., Prinzie, A.: Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing. Expert Syst. Appl. 39 (3), 2597–2605 (2012)
10. Thorleuchter, D., Van den Poel, D.: Using Webcrawling of Publicly-Available Websites to Assess E-Commerce Relationships. In: Service Research and Innovation Institute 2011 (SRII 2011), pp. 402--410. IEEE Press, New York (2012)
11. Lihui, C., Lian, C. W.: Using Web structure and summarisation techniques for Web content mining. Inform Process Manag 41 (5), 1225-1242 (2005)
12. Thorleuchter, D., Van den Poel, D.: Predicting e-commerce company success by mining the text of its publicly-accessible website. Expert Syst. Appl. 39 (17), 13026-13034 (2012)
13. Thorleuchter, D., Weck, G., Van den Poel, D.: Granular Deleting in Multi Level Security Models - an Electronic Engineering approach. In: International Conference on Mechanical and Electronic Engineering. LNEE, vol. 177, pp. 609--614. Springer, Berlin (2012)
14. Thorleuchter, D., Weck, G., Van den Poel, D.: Usability based Modeling for Advanced IT-Security - an Electronic Engineering approach. In: International Conference on Mechanical and Electronic Engineering. LNEE, vol. 177, pp. 615--619. Springer, Berlin (2012)
15. Thorleuchter, D., Van den Poel, D.: High Granular Multi-Level-Security Model for Improved Usability. In: 2nd International Conference on System science, Engineering design and Manufacturing informatization, pp. 191--194. IEEE Press, New York (2011)
16. Gericke, W., Thorleuchter, D., Weck, G., Reilaender, F., Loss, D.: Vertrauliche Verarbeitung staatlich eingestufter Information - die Informationstechnologie im Geheimschutz. Informatik Spektrum 32 (2), 102--109 (2009)
17. Thorleuchter, D., Van den Poel, D.: Rapid Scenario Generation with Generic Systems. In: Management Sciences and Information Technology. Lecture Notes in Information Technology. IERI, Delaware (2012), in press

18. Zheng, H.T., Kang, B.Y., Kim, H.G.: Exploiting noun phrases and semantic relationships for text document clustering. Inform Sciences 179 (13), 2249-2262 (2009)
19. Thorleuchter, D., Schulze, J., Van den Poel, D.: Improved emergency management by loosely coupled logistic system. In: Future Security 2012. CCIS, vol. 318, pp. 5--8. Springer, Berlin (2012)
20. Thorleuchter, D., Van den Poel, D.: Companies Website Optimising concerning Consumer's searching for new Products. In: International Conference on Uncertainty Reasoning and Knowledge Engineering, pp. 40--43. IEEE Press, New York (2011)
21. Sparck Jones, K.: Collection properties influencing automatic term classification performance. Inform Storage Retrieval 9 (9), 499-513 (1973)
22. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inform Process Manag 24 (5), 513-523 (1988)
23. Zhang, W., Yoshida, T., Tang, X.: A comparative study of TF*IDF, LSI and multi-words for text classification. Expert Syst. Appl. 38 (3), 2758-2765 (2011)
24. Thorleuchter, D., Van den Poel, D.: Extraction of Ideas from Microsystems Technology. In: Computer Science and Information Engineering. AISC, vol. 168, pp. 563--568. Springer, Berlin (2012)
25. Thorleuchter, D., Herberz, S. and Van den Poel, D.: Mining Social Behavior Ideas of Przewalski Horses. In: 3rd International Symposium on Computer, Communication, Control and Automation. LNEE, vol. 121, pp. 649--656. Springer, Berlin (2012)
26. D'Haen, J., Van den Poel, D., Thorleuchter, D.: Predicting Customer Profitability During Acquisition: Finding the Optimal Combination of Data Source and Data Mining Technique. Expert Syst. Appl. (2013), in Press, doi:10.1016/j.eswa.2012.10.023
27. Chen, M.Y., Chu, H.C., Chen, Y.M.: Developing a semantic-enable information retrieval mechanism. Expert Syst. Appl. 37 (1), 322-340 (2010)
28. DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44 (3), 837–845 (1988)
29. Hanley, J.A. McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 131 (1), 29–36 (1982)
30. Thorleuchter, D., Van den Poel, D.: Using NMF for analyzing war logs. In: Future Security 2012. CCIS, vol. 318, pp. 73--76. Springer, Berlin (2012)
31. Thorleuchter, D., Van den Poel, D.: Protecting Research and Technology from Espionage. Expert Syst. Appl. (2013), in Press, doi:10.1016/j.eswa.2012.12.051
32. Thorleuchter, D., Van den Poel, D., Prinzie, A.: Mining Ideas from Textual Information. Expert Syst. Appl. 37 (10), 7182-7188 (2010)
33. Thorleuchter, D., Van den Poel, D.: Improved Multilevel Security with Latent Semantic Indexing. Expert Syst. Appl. 39 (18), 13462-13471 (2012)
34. Thorleuchter, D., Van den Poel, D., Prinzie, A.: 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)
35. Thorleuchter, D., Van den Poel, D.: Web Mining based Extraction of Problem Solution Ideas. Expert Syst. Appl. In press (2013)