Mining Social Behavior Ideas of Przewalski Horses

Dirk Thorleuchter [1], Sarah Herberz [1], and Dirk Van den Poel [2]

[1] Fraunhofer INT, Appelsgarten 2, D-53879 Euskirchen, Germany
[2] Ghent University, Faculty of Economics and Business Administration, B-9000 Gent, Tweekerkenstraat 2, Belgium
Dirk.Thorleuchter@int.fraunhofer.de, Sarah.Herberz@int.fraunhofer.de
dirk.vandenpoel@ugent.be, http://www.crm.UGent.be

Abstract. Literature introduces idea mining as an approach for extracting interesting ideas from textual information. Related research focuses on extracting technological ideas as starting point for future technological research and development activities. Thus, it is limited to the technological domain. The algorithms standing behind idea mining also are optimized for the technological domain.
In contrast to previous research, this work transfers idea mining to the social behavior domain by selecting and adapting parameters of the idea mining algorithm. Forward selection as main approach in stepwise regression is used to choose the predictive variables based on their statistical significance. Grid search is used to optimize the parameter values. A case study shows that these optimized idea mining parameters are successful in extracting social behavior ideas of animals in this case of Przewalski horses. Based on these findings, differences between technological ideas and social behavior ideas can be shown.

Keywords. Idea Mining, Social Behavior, Przewalski Horses, Textmining, Knowledge Discovery

Introduction

Idea mining is known as an approach for extracting interesting ideas from textual information. Current work identifies technological ideas from scientific publications [1] or patents [2] and from internet blogs [3, 4]. Further, extended approaches for innovative potential identification [5] and for semantic classification [6] of the extracted ideas are introduced. The idea mining approach is optimized for calculating new ideas from the technological domain, where an idea is defined by a combination of means and ends. Identifying new ideas can be done by searching for new means appearing together with known ends or by searching for new ends appearing together with known means.

The idea definition is not restricted to the technological domain. E.g., ideas from the social behavior domain of animals can be found in scientific publications and they consist of textual patterns in which means and ends occur together, too. However, the way means and ends are defined or formulated probably differs by technological

domain. Thus, ideas from different domains probably need different parameters or at least different parameter values for a successful extraction of new ideas.

Since the idea mining approach has been applied successfully in many case studies, this work takes over existing parameters one by one by considering their statistical significance and, if significant, by adapting their values for the use of different domains. This approach is called 'forward selection procedure' in literature [7]. It is supported by applying a grid search and a 5-fold cross-validation [8].

In a case study, the social behavior domain of Przewalski horses, also known as Takhi (Equusferus przewalskii), is used. Przewalski horses are a species successfully saved from extinction by breeding in captivity [9]. By now, they have been reintroduced to their native habitat. Behavioral observations in captivity and in the wild lead to interesting ideas concerning the social behavior of these horses.

It is shown here that - comparing to the baseline - extracting social behavior ideas of Przewalski horses is successful by use of the newly selected and adapted parameter values. The results outperform those obtained with commonly used values that are optimized for the technological domain.

Background – The Idea Mining Classifier

The rationale behind the idea mining approach as described in [10] is based on an idea definition from technique philosophy. There, an idea is defined as a combination of a means and an end. A systematic search for means and ends occurring together leads to the identification of ideas. Means and ends themselves are textual patterns containing domain-specific terms. Further, the identification of new ideas depends on existing context information where textual patterns representing known ideas occur. By considering given context information, known means and ends can be identified and compared to new textual information to identify new ideas where a known means appears together with an unknown end or vice versa. The idea mining classifier compares a textual pattern from a new text to all textual patterns from the given context information. It identifies known and unknown means and ends and based on this identification it decides whether the textual pattern represents a new idea or not.

Idea mining uses several parameters for the classification decision. Let α be a set of terms from a new text and let β be a set of terms from the context where at least one term is in both sets. Let $p = |\alpha|$ be the number of terms in α and let $q = |\alpha \cap \beta|$ be the number of terms existing in both sets. Then, m_1 is defined as measure for a well-balanced known and unknown term distribution.

$$m_1 = \begin{cases} \frac{2 \cdot (p - q)}{p} & (q \geq \frac{p}{2}) \\ \frac{2 \cdot q}{p} & (q < \frac{p}{2}) \end{cases} \quad (1)$$

Let z be a percentage. Let δ be a set of most z % frequent terms without stop words in context. Let $r = |\alpha \cap \beta \cap \delta|$ be the number of frequent terms existing in both sets. Then, m_2 is defined as measure for frequent occurrence of known terms in context.

$$m_2 = \frac{r}{q} \tag{2}$$

Let φ be a set of z % frequent terms without stop word in the new text. Let $s = |\alpha \cap \bar{\beta} \cap \varphi|$ be the number of frequent terms only existing in α. Then, m_3 is defined as measure for frequent occurrence of unknown terms in the new text.

$$m_3 = \frac{s}{p - q} \tag{3}$$

Let λ be a set of characteristic terms (e.g. higher, better etc.). Let $t = |\alpha \cap \lambda|$ be the number of these terms in α. We define m_4 as measure for changing means and purposes.

$$m_4 = \begin{cases} 1 & (t > 0) \\ 0 & (t = 0) \end{cases} \tag{4}$$

Let $g_1, \ldots, g_4 \geq 0$ be weighting factors with $g_1 + g_2 + g_3 + g_4 = 1$. Let m be the sum of all four submeasures multiplied by weighting factors.

$$m = \begin{cases} g_1 m_1 + g_2 m_2 + g_3 m_3 + g_4 m_4 & (p \neq q) \\ 0 & (p = q) \end{cases} \tag{5}$$

Let â be a threshold. A textual pattern represented by α is classified as new idea regarding a context represented by β if $m \geq$ â.

Let u and v be percentages and let l be an integer value. The length of a textual pattern depends on the weights of its terms. A term is assigned the weight u if the term is a stop word. If not, a term is assigned the weight v. Terms are added to a textual pattern one by one while the sum of all term weights in the pattern is smaller than l.

Methodology

Fig. 1 shows the methodology of this approach.

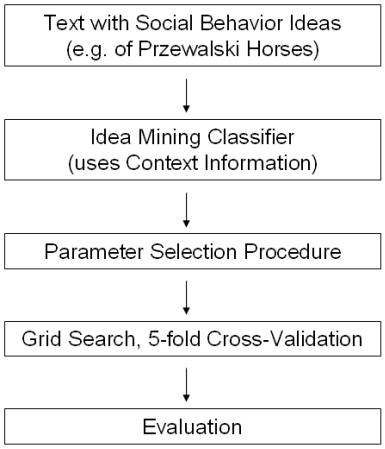

Fig. 1. Mining social behavior ideas is done in different steps.

A new text is selected that contains many ideas concerning the used domain. The idea mining classifier also uses a second text as context information. The latter also consists of ideas about the used domain. However, some ideas can be found in both texts while others appear in the new text or the context only. Thus, known ideas can be distinguished from new ideas and the results (the extracted ideas from the new text) can be assigned to the following categories: true positives, false positives, and false negatives for evaluation purposes.

The idea mining classifier consists of parameters used to determine the length of the text patterns and of parameters used for calculating the idea mining measure. The length of the text patterns depends on parameter l and on a term weighting scheme that is calculated by the percentages u and v. For calculating the idea mining measure, six parameters have to be set: (g_1, g_2, g_3, g_4, â, and z) with g_1 being the parameter for the balance of known and unknown term distribution, g_2 the parameter for the frequent occurrence of known terms in context, g_3 the parameter for the frequent occurrence of unknown terms in the new text, and g_4 the parameter for the occurrence of specific terms. Further, the parameter â is used as threshold for the classification decision and with the percentage z frequent terms are distinguished from non-frequent terms. Thus, all nine parameters (l, u, v, g_1, g_2, g_3, g_4, â, and z) are important for the performance of the idea mining classifier.

Parameter Selection

A non-optimized selection of these parameters leads to significantly low values for the precision and recall as shown by [1]. Thus, a parameter selection procedure is needed for computing an optimized parameter set that results in the highest performance using unseen data for the idea mining classification.

To calculate an optimized parameter set with nine parameters is computationally expensive. In addition, the resulting models are less comprehensible if they are based on too many parameters. To avoid these disadvantages, variable selection [11] is used

to reduce the number of parameters for the resulting model. Parameters with high predictive performance are selected while the others are discarded. The parameters are ordered based on their χ^2-statistic (highest first) and then, they are included in the resulting model one by one (forward-selection procedure) until a stopping rule is satisfied [12].

Grid search

Using a 'grid search' on the selected parameters has been suggested by [13]. This grid search uses the training data and is based on an n-fold cross-validation approach. It leads to an optimized parameter set that results in the highest performance using unseen data. For this reason our study uses the proposed grid search on the parameters. We define discrete sequences of the parameters. All combinations of the parameters have to be considered and the set that results in the best cross-validated F-measure is selected.

In contrast to [13] that proposes the use of the best cross-validated accuracy, we decided to use the F-measure. The F-measure is based on the precision measure, calculated by the number of true positive elements divided by the sum of true positive and false positive elements. Furthermore, it is based on the recall measure, calculated by the number of true positive elements divided by the sum of true positive and false negative elements. Both measures do not use the number of true negative elements. Therefore, they are in contrast to the accuracy and its related measure, the specificity, that are based on the number of true negative elements. The aim of idea mining is to identify textual patterns representing a new idea from a very large number of textual patterns in texts, because a text pattern around each term (word) in a text is built. Thus, the aim of idea mining is similar to the aim of information retrieval where interesting patterns from the enormous amount of patterns, e.g., on the internet, have to be identified. For an information retrieval scenario, it is hardly possible to calculate the number of true negative elements. Precision and recall are the commonly used measures in information retrieval and thus, this paper uses the best cross-validated F-measure for the selection of the parameters.

Case Study

In a case study, we examine the extraction of ideas from the social behavior domain. The social behavior of Przewalski horses is a very interesting research area, because they are the only horse species in the world that remains truly wild today although many herds live in zoos. Based on behavioral observations in captivity and in the wild, many social behavior ideas have been published in literature in different categories (e.g. herd behavior in the wild, hierarchical structure, role of the lead mare, role of the stallion, ratio of stallions and mares, communication).

For this case study, we use a document in which the social behavior of Przewalski stallions in different enclosures and reservations is described [14]. The document contains about 120 ideas. As context, a document is selected that gives a review on the existing literature in this field and implies the results of observations of

Przewalski horses in the Cologne zoo as well [15]. Both documents are in German language, thus, the idea mining classifier can compare the ideas for extracting new ideas from the first document.

The parameter selection calculates the parameters with the highest performance using unseen data: they are g_1, g_2, g_3, â, and z. These parameters are selected and used as variables in the following grid search approach. The other parameters l, u, v, and g_4 are used as constants.

Before starting a grid search, some issues have to be considered as mentioned in [1]. The parameters g_2 and g_3 are equally important so the values of these two parameters should be equal. The parameter g_1 should be between 20 % and 80 %. For the parameters g_2 and g_3, 10 % to 40 % is recommended. The value of â should be between 20 % and 50 %. The value of z should be between 10 % and 30 %. In addition, the sum of g_1 , g_2, and g_3 equals 1 because g_4 is set to zero based on the results of the parameter selection.

Thus, our study uses the proposed grid search on the parameters g_1, g_2, g_3, â, and z. We define discrete sequences of the five parameters (g_1 = 0.20, 0.40, 0.60, 0.80, g_2 = g_3 = 0.10, 0.20, 0.30, 0.40, â = 0.20, 0.30, 0.40, 0.50, z = 0.10, 0.20, 0.30). All 48 value combinations are considered in an evaluation to get the parameter values results with the highest performance (g_1 = 0.20, g_2 = g_3 = 0.40, â = 0.20, and z = 0.10). This is in contrast to the calculated values in [1] that are optimized for the technological domain (g_1 = 0.40, g_2 = g_3 = 0.20, â = 0.50, and z = 0.30). Comparing the precision and recall values of both sets of parameter values leads to a significant improvement in recall by use of the newly calculated set of parameter values. While precision slightly decreases from 87 % to 82 %, the recall escalates from 12 % to 88 %. Therefore, the F-measure increases from 49 % to 85 %. Also, these results outperform the precision (40 %) and recall (25 %) values as calculated for the extraction of technological ideas [1]. Last, a comparison to the baseline has to be done. Referring to the evaluation in [1], well-known heuristic similarity measures (Jaccard's coefficient, overlap-index, cosine-similarity, and dice-similarity) are used as baseline by replacing the idea mining measure. The results for precision (30 %) and for recall (20 %) using all these different measures one by one are nearly the same. Thus, the results by use of the newly calculated parameter value set outperform the baseline, too.

Results from the Case Study (Examples)

The results can be distinguished in true positive, false negative, and false positive ideas: One true positive result was the correct identification of the appearance of coalitions between a mature stallion and a colt including the distribution of mares. Another one was the idea that aggression seems to be a positive leadership ability which increases the reproductive success of a stallion. Additionally, this approach correctly detected the idea that the testosterone level of a stallion influences his social behavior.

However, this approach did not identify the new idea that the hormone level is regulated by environmental and social factors, so this is a false negative result. Further, two false negative examples are that a) in the presence of a harem stallion,

concentration of testosterone and aggression among colts is subject to social control and that b) alleviated aggressive behavior has been noticed in winter whereas testosterone level and readiness to combat increase in spring, when mating occurs.

Furthermore, this approach identified ideas on social organization among Przewalski horses. These are relevant aspects but not new ideas as they can already be found in the known context. Two examples for these false positive ideas are, firstly, that in the wild horses organize themselves in harem groups consisting of a mature stallion and several mares with their foals and secondly, young colts and fillies leave their birth group at the age of one to three years. The males come together in bachelor groups whereas related fillies often form the kernel of a new harem.

In contrast to this, the extracted idea that inbreeding is a massive problem of rearing Przewalski horses is new but not relevant for the social behavior of these animals. Thus, this idea also represents a false positive example.

Conclusion and Outlook

This work shows that the existing approach for extracting technological ideas from textual information can also be successfully adapted to the domain of social behavior as shown for Przewalski horses. The calculated F-measure outperforms the corresponding value from the technological domain and from the baseline. Further, the use of the optimized parameter values increases the performance of the approach. Thus – in contrast to previous research -, this work transfers idea mining to the social behavior domain.

The different parameter values that are used for the extraction of a social behavior idea in contrast to the extraction of a technological idea lead to two interesting aspects. Firstly, the decreased value of z and the increased value of g_2 and g_3 in the social behavior domain show that in this domain, only a few technical terms exist, which are used very often to describe an idea. In contrast to this, many technical terms can be found in the technological domain. Hence, each individual term does not appear in technological ideas that often. Secondly, the decreased value of g_1 shows that a well-balanced means and ends distribution is not critically important in social behavior ideas. It can be seen that in contrast to many technological ideas, means are often described by few technical terms (e.g. aggression and testosterone level). Thus, this work also shows the differences between technological ideas and social behavior ideas.

Acknowledgment

We thank Dr. Joachim Schulze, Joerg Fenner, and Ruth Herberz for their constructive technical comments.

References

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

2. 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)
3. Thorleuchter, D., Van den Poel, D., Prinzie, A.: Extracting Consumers Needs for New Products. In: Proceedings WKDD 2010, pp. 440--443. IEEE Computer Society, CA: Los Alamitos (2010)
4. Thorleuchter, D., Van den Poel, D.: Companies Website Optimising concerning Consumer's searching for new Products. In: 2011 International Conference on Uncertainty Reasoning and Knowledge Engineering. IEEE Press, New York (2011)
5. Thorleuchter, D., Van den Poel, D., Prinzie, A.: Mining Innovative Ideas to Support new Product Research and Development. In: Locarek-Junge, H., Weihs, C. (eds.) Classification as a Tool for Research, pp. 587--594. Springer, Berlin (2010)
6. Thorleuchter, D., Van den Poel, D.: Semantic Technology Classification. In: 2011 International Conference on Uncertainty Reasoning and Knowledge Engineering. IEEE Press, New York (2011)
7. Van den Poel, D., Buckinx, W.: Predicting Online-Purchasing Behavior, Eur. J. Oper. Res. 166 (2), 557--575 (2005)
8. 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. in press (2012)
9. Dierendonck, M.C. van, Bandi, M., Batdorj, D., Dügerlham, S., Munkhtsog, B.: Behavioural observations of reintroduced takhi or Przewalski horses (Equus ferus przewalskii) in Mongolia. Appl. Anim. Behav. Sci. 50 (2), 95--114 (1996)
10. Thorleuchter, D.: Finding New Technological Ideas and Inventions with Text Mining and Technique Philosophy. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (eds.) Data Analysis, Machine Learning and Applications, pp. 413--420. Springer, Berlin (2008)
11. Kim, Y.S.: Toward a successful CRM: variable selection, sampling and ensemble. Decis. Support Syst. 41 (2), 542--553 (2006)
12. Coussement, C., Van den Poel, D.: Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert Syst. Appl. 34 (1), 313--327 (2008)
13. Hsu, C.-W., Chang, C.-C., Lin, C.-J.: A Practical Guide to Support Vector Classification. Technical report, Department of Computer Science and Information Engineering, National Taiwan University (2004)
14. Kolter, L., Zimmermann, W.: Die Haltung von Junggesellengruppen für das EEP Przewalskipferd - Hengste in Gehegen und Reservaten. Zeitschrift des Kölner Zoo, 44 (3), 135--151 (2001)
15. Herberz, S.: Dominanzverhalten bei Przewalskipferden in seminatürlicher Haltung im Kölner Zoo unter besonderer Berücksichtigung der Körpersprache. Fraunhofer INT edition, Euskirchen (2011)