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

Quantitative Cross Impact Analysis with Latent Semantic Indexing

Dirk Thorleuchter[1]

Dirk Van den Poel[2]

November 2013

2013/861

[1] Corresponding author: Dr. Dirk Thorleuchter: 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

[2] Prof. Dr. Dirk Van den Poel, Professor of Marketing Analytics/Analytical Customer Relationship Management, Faculty of Economics and Business Administration, E-mail: dirk.vandenpoel@ugent.be; more papers are available from: www.crm.UGent.be

D/2013/7012/32

Quantitative Cross Impact Analysis with Latent Semantic Indexing

Dirk Thorleuchter[a,*], Dirk Van den Poel[b]

Abstract

Cross impact analysis (CIA) consists of a set of related methodologies that predict the occurrence probability of a specific event and that also predict the conditional probability of a first event given a second event. The conditional probability can be interpreted as the impact of the second event on the first. Most of the CIA methodologies are qualitative that means the occurrence and conditional probabilities are calculated based on estimations of human experts. In recent years, an increased number of quantitative methodologies can be seen that use a large number of data from databases and the internet. Nearly 80% of all data available in the internet are textual information and thus, knowledge structure based approaches on textual information for calculating the conditional probabilities are proposed in literature. In contrast to related methodologies, this work proposes a new quantitative CIA methodology to predict the conditional probability based on the semantic structure of given textual information. Latent semantic indexing is used to identify the hidden semantic patterns standing behind an event and to calculate the impact of the patterns on other semantic textual patterns representing a different event. This enables to calculate the conditional probabilities semantically. A case study shows that this semantic approach can be used to predict the conditional probability of a technology on a different technology.

Keywords: Cross Impact Analysis, Latent Semantic Indexing, Text Mining, Conditional Probability.

Introduction

In literature, cross impact analysis (CIA) is often used to predict the probability that a specific event occur (occurrence probability) as well as the impact of this event on different events (conditional probability) (Blanning & Reinig, 1999; Schuler, Thompson, Vertinsky, & Ziv, 1991). A large number of existing approaches are qualitative. They are based on estimations of human experts (Banuls, Turoff, & Hiltz, 2013; Mitchell, Tydeman, & Curnow, 1977). In recent years, the number of quantitative approaches has increased. This is because the large number of accessible information today makes it possible to use the results of automated data mining approaches instead of using the time- and cost expensive estimations by human experts (Kim, Lee, Seol, & Lee, 2011). Quantitative CIA approaches that are based on textual information are knowledge structure based because they apply multi-label text classification approaches based on well-known text similarity measures to identify the impact of one event on a different event (Thorleuchter, Van den Poel, & Prinzie, 2010). However, this is done by considering aspects of words and not by considering semantic aspects in textual information.

An example for an event could be the appearance of a new technology in the technology landscape. The appearance of new technologies and the change of existing technologies over time from past to future is a well-known topic for futurists (Bell, 2002). This enables to predict future technological capabilities for decision-makers (Thorleuchter & Van den Poel, 2013d). The technological landscape is characterized by a large number of technologies that are impacted by a large number of other technologies (Yu, Hurley, Kliebenstein, & Orazem, 2012). Technologies impact other technologies in different ways e.g. in an integrative, substitutive, precursive, and successive way (Geschka, 1983). A short example for the substitutive way is given below: The electrical fuel cell technology used in an energy supply application can be substituted by electrical battery or solar cell technology. This is because all three technologies can be used to realize this application. They replace each other based on their advances. Thus, the three technologies impact each other in a substitutive way. Further, these impacts change very often because current results from technological research and development lead to new technological advances and to the appearance of new substitutive technologies as an ongoing process (Kauffman, Lobo, & Macready, 2000). As a result, using CIA for monitoring these complex technological impacts makes it necessary to use quantitative rather than qualitative approaches.

Several texts that describe a single event are normally written in several writing styles by different persons. Further, these texts possibly are written in different contexts or in different languages. It is not necessary that two texts describing the same event contain even one common word (Thorleuchter & Van den Poel, 2013b). With semantic approaches the relationship between the two texts can be identified because they share a common meaning (Choi et al., 2012; Tsai, 2012). This is the reason why semantic text classification approaches often outperform knowledge structure based text classification approaches (Thorleuchter & Van den Poel, 2012c).

In contrast to existing CIA approaches, we provide a quantitative CIA approach that considers the aspects of meaning in textual information.

Latent semantic indexing (LSI) is a well-known representative for semantic approaches (Jiang, Berry, Donato, Ostrouchov, & Grady, 1999). It identifies the hidden meaning of textual information in documents considering occurrences and co-occurrences of terms (D'Haen, Van den Poel, & Thorleuchter, 2013; Luo, Chen, & Xiong, 2011). Both, terms and documents are mapped to a semantic structure that consists of several semantic textual patterns (Christidis, Mentzas, & Apostolou, 2012; Park, Kim, Choi, & Kim, 2012). The impact of terms and documents on the patterns is calculated (Kuhn, Ducasse, & Girba, 2007). A semantic textual pattern that represents e.g. a technology might contain terms and documents that also have an impact on a different semantic textual pattern representing e.g. another technology (Thorleuchter & Van den Poel, 2013c). This indicates a relationship between the technologies and based on this relationship, the cross-impact between technologies can be calculated.

To extract semantic patterns from the large number of texts describing events, we use a rank-validation procedure that is taken over from literature (Thorleuchter & Van den Poel, 2013a). This procedure enables to identify a maximal number of semantic patterns where each pattern can be used to represent a specific event. The rank-validation procedure is successfully evaluated by using LSI with singular value decomposition (SVD). Beside LSI, modern semantic approaches exist that outperform LSI in several studies. Examples for these modern approaches are probabilistic latent semantic indexing (Hofmann, 1999), non-negative matrix factorization (Lee & Seung, 1999; Lee & Seung, 2001), and latent dirichlet allocation (Blei, Ng, & Jordan, 2003). However, literature has not validated the use of these modern approaches together with the rank-validation procedure until now. Additionally, the modern approaches are of higher computational complexity than LSI (Ramirez, Brena, Magatti, & Stella, 2012). Thus in this paper, LSI is used together with the rank-validation procedure because this combination is already successful evaluated and it is of good computational performance.

In a case study, we predict the impact of technologies on different technologies. The used data are descriptions of research projects funded by the German Ministry of Defense (GE MoD) in 2007. These research projects deal with one or several technologies to create an application. Semantic textual patterns in the descriptions are extracted, the technologies standing behind the patterns are identified, and the cross-impacts between the technologies are calculated. This semantic approach is compared to a knowledge structure based approach that uses the same data for calculating the cross-impacts.

Overall, we propose a quantitative methodology that combines semantic text classification with CIA. The use of a semantic approach for the CIA calculation is in contrast to related work. The semantic methodology calculates the conditional probabilities of events given different events quantitatively. This enables to depict the complex relationships between events with lower manual effort than qualitative approaches and by considering semantic aspects. Thus, it is helpful for decision makers.

Background

The proposed approach calculates conditional cross impact probabilities by use of semantic text classification. Below, we describe how conditional cross impact probabilities can be calculated and how quantitative text-based CIA is processed up to now.

In 1968, CIA was proposed (Gordon & Haywood, 1968) to calculate the occurrence probabilities of an event and to calculate the conditional probabilities of one event given another. The approach is based on subjective estimations by human experts. The occurrence probability of an event A was simply defined as P(A) and calculated by the number of these human experts who predict the occurrence of A over the number of all human experts. The conditional probability of event B given event A was defined as P(B|A) and calculated by the number of experts who predict both, the occurrence of A and B over the number of all experts who predict the occurrence of A (Dalkey, 1972; Enzer, 1972).

This approach was improved many times and nowadays, most of the new improved approaches focus on a more quantitative way to calculate the probabilities. Examples are the use of cumulative sale probabilities over time by (Caselles-Moncho, 1986) and the use of patent data (Choi, Kim, & Park, 2007). These quantitative approaches start with a multi-label data classification step where the data is assigned to different events (classes). Based on this assignment, the calculation of the probabilities is done in a second step.

About 80% of all data available today are textual data. Thus, modern approaches use the large number of textual data e.g. available in the internet for CIA. Examples are the use of linguistic expressions in technology descriptions (Jeong & Kim, 1997) and the use of terms from technology taxonomies (Thorleuchter, Van den Poel, & Prinzie, 2010). From text classification point of view, these approaches are knowledge-based and they use instance-based learning algorithms where semantic aspects of the textual data are not considered. This is in contrast to the approach presented here where a new methodology is provided that uses a semantic approach (LSI) for calculating the conditional probabilities from texts.

Methodology

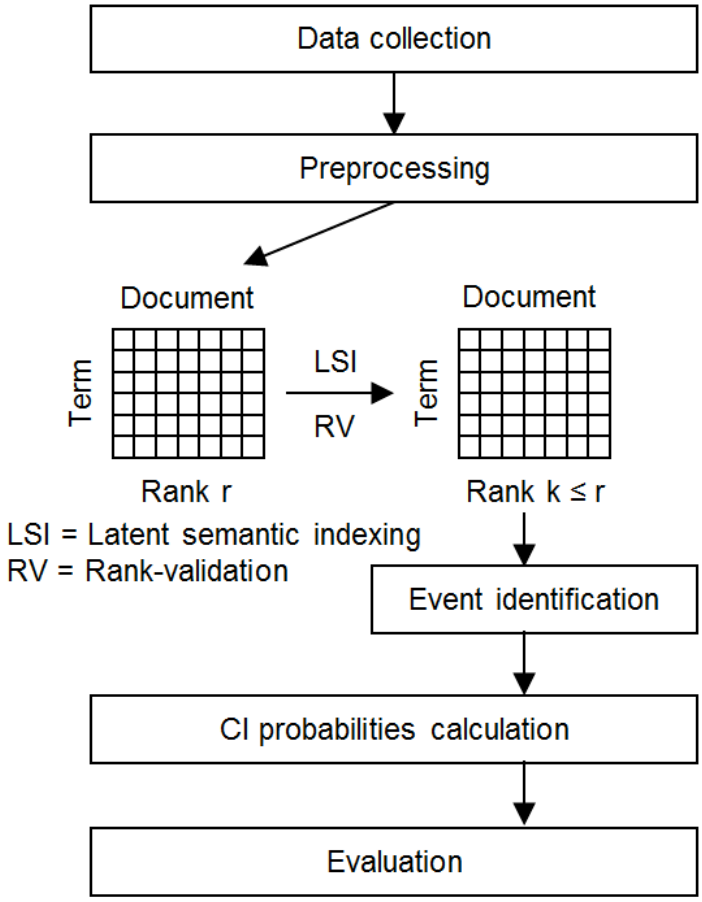

Fig. 1 shows the processing of the methodology in different steps.

The methodology (see Fig. 1) starts with a data collection step. Events are defined and a set of documents are used as input. The documents should consist of textual information describing one or several events. As an example, the case study defines an event as a technology and thus, each document contains a description of a research project where one or several technologies occur.

In a preprocessing step, specific elements (e.g. scripting code, tags, and images) are removed. The text is split in terms and each term is checked for typographical errors by use of a dictionary. The large number of different terms is reduced by applying term filtering methods e.g. stop word filtering, part-of-speech tagging, and stemming. Further, Zipf's law (Zeng, Duan, Cao, & Wu, 2012; Zipf, 1949) is applied where many low frequent terms can be discarded. Each document is represented by a term vector based on vector space model. The size of a vector is based on the reduced number of terms (Thorleuchter & Van den Poel, 2012a). Vector components are represented by weighted frequencies as calculated in accordance to Salton et al. (1994). The frequency of the corresponding term in a specific document is multiplied by its inverse document frequency and it is divided by a length normalization factor.

The term vectors are used to create a term-by-document matrix with rank r. The rank of the matrix is reduced from r to k by LSI. For the selection of on optimal value of k, a rank-validation procedure as introduced in Sect. 2.3 is applied: for each value of k, LSI is applied and the resulting k dimensions are compared to the descriptions of the events. In the case study, this step is done manually by human experts; however, it could be realized automatically by applying a text similarity measure on patterns and events. As a result of this step, the number of semantic textual patterns with a one-to-one correspondence (an exact pairing) to an event is calculated.

A small number of k leads to a small number of semantic textual patterns where most of the terms with high impact on one of these patterns stem from different events. Further, a large number of k leads to a large number of semantic textual patterns where a single event is represented by several patterns. A maximal number of one-to-one correspondences can be obtained by varying the value of k and by calculating the number of identified one-to-one correspondences for each k (rank-validation procedure). As a result, k is selected by applying the rank-validation procedure and j<k semantic textual patterns can be identified with one-to-one correspondences in the step event identification.

After selecting the value of k, LSI uses singular value decomposition to split the term-by-document matrix in a product of the matrices U, Σ, and V^t

$$A = U \Sigma V^t$$

Then, three further matrices U_k, Σ_k and V_k are calculated by discarding the columns of U, Σ, and V from k+1 on. The components of matrix U_k contain values for the impact of each term on each of the k semantic textual patterns. The impact of each document on each of the k semantic textual patterns can be found in matrix V_k (Thorleuchter, Van den Poel, & Prinzie, 2012b). We use matrix V_k to calculate the conditional cross impact probability of an event B given an event A if both events are represented by a specific semantic textual pattern. This is calculated by the number of documents that are assigned to both events A and B divided by the number of documents that are only assigned to event A. While the impact of a document on a semantic textual pattern is a vale in [-1,..,1], a specific threshold q is used to distinguish between documents that are related to a semantic textual pattern and documents that are not. The cross-impact of A on B is calculated by

$$CI(A,B) = P(B|A) = N(A \cap B) / N(A)$$

The number of documents in matrix V_k where the impact on a specific event A is above q is used to calculate N(A). N(A ∩ B) is the number of documents in matrix V_k where the impact on event A and B is above q in both cases.

To evaluate the CI(A,B) score, precision and recall can be used. The CI(A,B) scores that are in [0,..,1] have to be transformed to Boolean variables [false, true] by use of a further threshold. This enables to identify whether event A impacts event B or not. Precision and recall indicators are well-known performance measures in binary classification. For applying precision and recall, the ground truth has to be determined, too.

Case Study

In a case study, we define defense-based technology areas as events. They are taken over from the technology taxonomy of the European Defense Agency (EDA) where 32 technology areas are selected.

For the documents, we use research projects funded by the German Ministry of Defence (GE MoD) in 2007. Descriptions of 985 projects have been identified and stored in documents separately. Some of the research projects examine a specific defense-based technology while other combine several technologies to create new approaches.

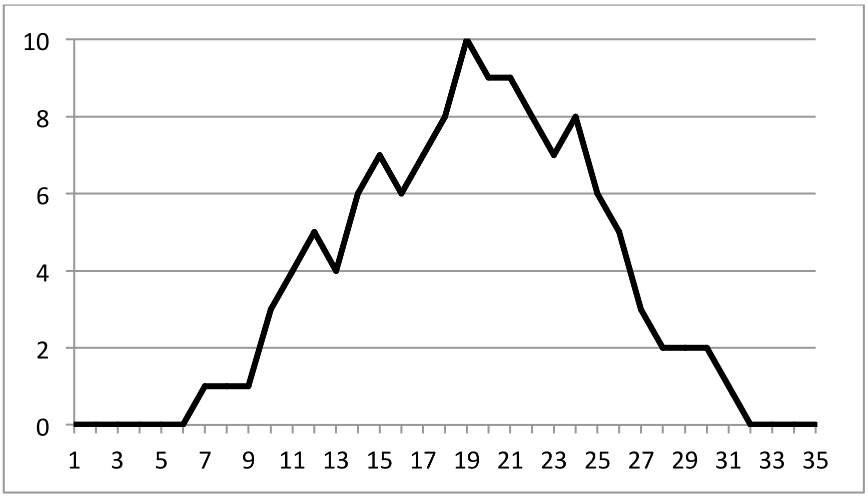

Fig. 2: Number of one to on correspondences (y-axis) based on the value of k (x-axis)

After pre-processing, the term-by-document matrix is built. LSI is applied together with the rank-validation procedure from k = 2 to k = 35. For each k, singular value decomposition is processed and the created k semantic textual patterns are assigned to the 32 technology areas by human experts. Some patterns do not fit to a technology area while others are assigned to several technology areas. The number of semantic textual pattern that are assigned to one and only one technology area (one-to-one correspondences) is depicted in Fig. 2. It shows that up to k = 18, the number of identified one-to-one correspondences is smaller than 9 and from k = 20 on, the number is smaller than 10. Selecting k = 19 leads to the identification of 10 one-to-one correspondences and thus, 10 technology areas. For further processing, k = 19 is selected.

The identified 10 technology areas lead to the calculation of 90 conditional cross impact probabilities as calculated by two times the binomial coefficient 10 choose 2. The identified 10 technology areas are presented in Table 1.

Table 1: Identified technology areas from EDA taxonomy

A02	Signature Related Materials
A03	Electronic Materials Technology
A04	Photonic/Optical Materials & Device Technology
A05	Electronic, Electrical & Electromechanical Device Technology
A08	Computing Technologies & Mathematical Techniques
B02	Propulsion and Powerplants
B04	Electronic Warfare and Directed Energy Technologies
B05	Signature Control and Signature Reduction

B06	Sensor Systems
B08	Simulators, Trainers and Synthetic Environments

The impact of a document on a technology area in matrix V_k is a value in [-1,..,1]. The threshold q is selected as suggested in literature (Thorleuchter & Van den Poel, 2013a). Based on the component values in matrix V_k, the 90 conditional cross impact probabilities are calculated based on the value of q = 0.4. The results are depicted in Table 2 colored in five different grayscales from bright to dark concerning the five cases:
No cross impact: CI(A,B) = 0;
Low cross impact: 0 < CI(A,B) ≤ 0.25;
Medium cross impact: 0.25 < CI(A,B) ≤ 0.50;
High cross impact: 0.50 < CI(A,B) ≤ 0.75;
Very high cross impact: CI(A,B) > 0.75.

Table 2: Result matrix of the calculated CI(A,B) e.g. CI(A02, A03) = 0.07

	A02	*A03*	*A04*	*A05*	*A08*	*B02*	*B04*	*B05*	*B06*	*B08*
A02	-	0.07	0.04	0	0	0	0	0.89	0.04	0
A03	0.05	-	0.13	0.82	0	0.18	0.24	0	0.21	0
A04	0.02	0.12	-	0.19	0	0	0.26	0.16	0.30	0
A05	0	0.35	0.09	-	0	0.27	0.25	0.02	0.24	0.05
A08	0	0	0	0	-	0	0.08	0	0.20	0.29
B02	0	0.16	0	0.53	0	-	0	0.07	0	0
B04	0	0.25	0.31	0.61	0.17	0	-	0.03	0.06	0
B05	0.66	0	0.18	0.05	0	0.08	0.03	-	0.03	0
B06	0.01	0.11	0.17	0.28	0.20	0	0.03	0.01	-	0
B08	0	0	0	0.13	0.58	0	0	0	0	-

Evaluation

For the evaluation of the results, we use a further study (furthermore named comparative study) from literature (Thorleuchter, Van den Poel, & Prinzie, 2010) that has calculated the conditional cross impact probabilities from the same input data. The comparative study uses a knowledge structure based classification approach based on centroid vectors to calculate the impacts. The conditional cross impact probabilities calculated by the comparative study are assigned to a positive and a negative class concerning a specific threshold r. This value (r = 0.25) is also used to evaluate the results of our proposed approach.

Creating a ground truth for calculating precision and recall is mandatory. This can be done manually by human experts. They have to decide whether a technology has an impact on a second technology above the specific threshold or not. The decision that an impact is above a specific threshold is a very subjective task for human experts and it might be that two experts make different decisions.

In the comparative study, human experts have analyzed the calculated results and they are able to give a heuristic explanation that confirms each single result. They have not identified any misclassification because the project descriptions use a technical language where terms are more strictly defined than terms from the colloquial language. This enables very good classification results - in this case 100 % precision at 100 % recall. Thus, the results from the comparative study can be used as ground truth to evaluate our proposed approach.

A selection of 11 from the 90 conditional cross impact probabilities is presented in Table 3. 'Techn. area A' represents the influencing technology area and 'Techn. area B' is the influenced technology area. Both technology areas stem from the 10 technologies as depicted in Table 1. The conditional probability of technology area B given technology area A is CI(A,B) as calculated by our proposed approach. The Boolean cross impact score BCI(A,B) is true if CI(A,B) is above threshold r = 0.25 otherwise it is false. CI_{com}(A,B) and BCI_{com}(A,B) are the corresponding values as calculated from the comparative study. Each row represents two different technology areas from Table 1 ordered by the CI_{com}(A,B) score where BCI_{com}(A,B) is true. Res_{diff}(A,B) is the difference between the residual from CI(A,B) to the residual from CI_{com}(A,B).

Table 3: Technology area pairs with BCI_{com}(A,B) is true ordered by CI_{com}(A,B) in 2007

Techn. area A	*Techn. area B*	*CI (A,B)*	*BCI (A,B)*	*CIcom (A,B)*	*BCIcom (A,B)*	*Resdiff (A,B)*
A02	B05	0.89	True	0.92	True	-0.01
A03	A05	0.82	True	0.86	True	-0.02
B05	A02	0.66	True	0.62	True	0.06
B04	A05	0.61	True	0.61	True	0.02
B02	A05	0.53	True	0.54	True	0.01
B08	A08	0.58	True	0.53	True	0.07
A05	A03	0.35	True	0.32	True	0.05
A08	B08	0.29	True	0.31	True	0.00
A05	B02	0.27	True	0.29	True	0.00
A05	B06	0.24	False	0.27	True	-0.01
A05	B04	0.25	False	0.26	True	0.01

Based on the 90 calculated cross-impact probabilities, the comparative study has shown that in 11 cases, BCI_{com}(A,B) is true and thus, an impact above the threshold can be seen. In 79 cases,

$BCI_{com}(A,B)$ is false. This leads to a frequent baseline of about 12%. The calculated cross-impact probabilities from the proposed approach lead to BCI(A,B) = true in 13 cases and to BCI(A,B) = false in 77 cases. In 9 of the 13 positive cases, $BCI_{com}(A,B)$ is also true while in 4 cases BCI(A,B) is true and $BCI_{com}(A,B)$ is false. These results are depicted in Table 4. Thus, precision is calculated as 9 / 13 = 69 % and recall is calculated as 9 / 11 = 82 %. This outperforms frequent baseline of 12 % precision at 82 % recall. The differences between the residuals are small. This also shows that the results are similar to those of the ground truth.

Table 4: Confusion matrix

		Predictive Class	
		Yes	No
Actual class	Yes	9	2
	No	4	75

To present a detailed example, we discuss the cross-impact among technology area B02 (Propulsion and Powerplants) and A05 (Electronic, Electrical & Electromechanical Device Technology). In 2007, a well-known trend in propulsion and powerplant technology was the creation of a more electric engine. The corresponding research projects that have been processed during that time can be assigned to both technology areas. The comparative study has shown that 54% of all research projects from technology area B02 are also assigned to A05 and that 29% of all research projects from technology area A05 are also assigned to B02. This is because 37 research projects are assigned to B02, 69 are assigned to A05, and 20 are assigned to both. In contrast to this, the proposed approach assigns 34 research projects to B02, 67 are assigned to A05, and 18 are assigned to both. The differences are evaluated manually by human experts. They could not identify concrete hints for a misclassification because the assignment of the corresponding research projects is very subjective. An example is a research project that develops new kinds of lubricants. The research results can be used for improving propulsion and powerplants but also for many further applications. The question whether this project is related to B02 or not is very subjective.

Overall, the proposed approach outperforms the baseline that proves its feasibility. Further, only small differences between CI(A,B) and $CI_{com}(A,B)$ can be seen in the 90 cross-impacts. This also shows the feasibility of the proposed approach. The 2 cases where BCI(A,B) is false and $BCI_{com}(A,B)$ is true as well as the 4 cases where BCI(A,B) is true and $BCI_{com}(A,B)$ is false are resulted by the selection of the threshold r. Despite of the small differences, some CI(A,B) and $CI_{com}(A,B)$ are assigned to different classes because their value is about the value of the threshold. Thus, these 6 cases are not significant.

Conclusion

We propose a new approach that calculates conditional cross-impact probabilities. In contrast to previous work, this work uses semantic classification by applying LSI together with a rank validation procedure. While knowledge structure based approaches are used for quantitative CIA in literature, the aim of this work is to show that semantic approaches also can be used.

In a case study the proposed approach is applied to identify conditional cross-impact probabilities between technologies. The evaluation is based on a further study where an extensive evaluation on the same data was already processed. As a result, the evaluation shows that the proposed approach outperforms the frequent baseline. Thus, it can be successful applied for quantitative CIA.

Comparing the proposed approach to a knowledge structure based approach fails because the assignment of texts to classes in the case study is a very subjective task. Literature has shown that in contrast to knowledge structure based approaches, semantic approaches have advances by processing colloquial texts (e.g. internet blogs) rather than highly structured texts where each term is pre-defined in a way that synonym and homonym problems normally do not occur. Technological descriptions are rather structured texts than colloquial texts because several technical terms have well-known meanings. Thus, the results of the case study from both approaches are similar. Future work could be compared both kinds of approaches by use of colloquial texts, e.g. by including documents from the internet written in different languages and in different writing styles. We expect different results from both approaches so that a comparison possibly will show advances of the semantic approach.

Normally, LSI is a clustering approach. We used it together with a rank validation procedure for classification. In cases where the events are not pre-defined, LSI can be used as clustering approach without the rank validation procedure. This improves performance of the approach on one hand but probably the automatically created events are not comprehensible for the users on the other hand. This might be a further avenue of future research.

Future work also should focus on are the implementation of the compared cross impact (CCI) approach with LSI. CCI analysis extends the CIA and up to now, it is only processed quantitatively by knowledge structure based approaches.

Literature

Banuls, V. A., Turoff, M., & Hiltz, S.R. (2013). Collaborative scenario modeling in emergency management through cross-impact. *Technological Forecasting and Social Change*, in press.

Bell, W. (2002). A community of futurists and the state of the futures field. *Futures*, 34(3-4), 235-247.

Blanning, R. W., & Reinig, B. A. (1999). Cross-impact analysis using group decision support systems: an application to the future of Hong Kong. *Futures*, 31(1), 39-56.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5), 993-1022.

Caselles-Moncho, A. (1986). An empirical comparison of cross-impact models for forecasting sales. *International Journal of Forecasting*, 2(3), 295-303.

Choi, C., Kim, S., & Park, Y. (2007). A patent-based cross impact analysis for quantitative estimation of technological impact: the case of information and communication technology. *Technological Forecasting and Social Change*, 74(2007), 1296-1314.

Choi, O., Kim, K., Wang, D., Yeh, H., & Hong, M. (2012). Personalized Mobile Information Retrieval System. *International Journal of Advanced Robotic Systems*, doi: 10.5772/50910.

Christidis, K., Mentzas, G., & Apostolou, D. (2012). Using latent topics to enhance search and recommendation in Enterprise Social Software. *Expert Systems with Applications*, 39(10), 9297-9307.

Dalkey, N. C. (1972). An elementary cross-impact model, *Technological Forecasting and Social Change*, 3, 341-351.

D'Haen, J., Van den Poel, D., & Thorleuchter, D. (2013). Predicting customer profitability during acquisition: Finding the optimal combination of data source and data mining technique. *Expert Systems with Applications*, 40(6), 2007-2012.

Enzer, S. (1972). Cross-impact techniques in technology assessment. *Futures*, 4(1), 30-51.

Geschka, H. (1983). Creativity techniques in product planning and development: A view from West Germany. *R&D Management*, 13(3), 169-183.

Gordon, T., & Haywood, H. (1968). Initial experiments with the cross impact matrix method of forecasting, *Futures*, 1(2), 100-116.

Hofmann, T. (1999). Probabilistic Latent Semantic Indexing. In: Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99).

Jeong, G. H., & Kim, S. H. (1997). Aqualitative cross-impact approach to find the key technology. *Technological Forecasting and Social Change*, 55(3), 203-214.

Jiang, J., Berry, M. W., Donato, J. M., Ostrouchov, G., & Grady, N. W. (1999). Mining consumer product data via latent semantic indexing. *Intelligent Data Analysis*, 3(5), 377-398.

Kauffman, S., Lobo, J., & Macready, W. G. (2000). Optimal search on a technology landscape. *Journal of Economic Behavior & Organization*, 43(2), 141-166.

Kim, C., Lee, H., Seol, H., & Lee, C. (2011). Identifying core technologies based on technological cross-impacts: An association rule mining (ARM) and analytic network process (ANP) approach. *Expert Systems with Applications*, 38(10), 12559-12564.

Kuhn, A., Ducasse, S., & Girba, T. (2007). Semantic clustering: Identifying topics in source code. *Information and Software Technology*, 49(3), 230-243.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755), 788-791.

Lee, D. D., & Seung, H. S. (2001). Algorithms for Non-negative Matrix Factorization. In: Advances in Neural Information Processing Systems 13. Proceedings of the 2000 Conference. MIT Press. pp. 556-562.

Luo, Q., Chen, E., & Xiong, H. (2011). A semantic term weighting scheme for text categorization. *Expert Systems with Applications*, 38(10), 12708-12716.

Mitchell, R. B., Tydeman, J., & Curnow, R. (1977). Scenario generation: limitations and developments in cross-impact analysis. *Futures*, 9(3), 205-215.

Park, D, H., Kim, H. K., Choi, I. Y., & Kim, J. K. (2012). A literature review and classification of recommender systems research. *Expert Systems with Applications*, 39(11), 10059-10072.

Ramirez, E. H., Brena, R. F., Magatti, D., & Stella, F. (2012). Topic model validation. Neurocomputing 76(1), 125-133.

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

Schuler, A., Thompson, W. A., Vertinsky, I., & Ziv, Y. (1991). Cross impact analysis of technological innovation and development in the softwood lumber industry in Canada: a structural modeling approach. *IEEE Transition of Engineering Management*, 38(3), 224-236.

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010). 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. (2012a). Predicting e-commerce company success by mining the text of its publicly-accessible website. *Expert Systems with Applications*, 39(17), 13026-13034.

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2012b). Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing. *Expert Systems with Applications*, 39(3), 2597-2605.

Thorleuchter, D., & Van den Poel, D. (2012c). Improved multilevel security with latent semantic indexing. *Expert Systems with Applications*, 38(18), 13462-13471.

Thorleuchter, D., & Van den Poel, D. (2013a). Weak signal identification with semantic web mining. *Expert Systems with Applications*, 40(12), 4978-4985.

Thorleuchter, D., & Van den Poel, D. (2013b). Protecting Research and Technology from Espionage. *Expert Systems with Applications*, 40(9), 3432-3440.

Thorleuchter, D., & Van den Poel, D. (2013c). Technology classification with latent semantic indexing. *Expert Systems with Applications*, 40(5), 1786-1795.

Thorleuchter, D., & Van den Poel, D. (2013d). Web Mining based Extraction of Problem Solution Ideas. *Expert Systems with Applications*, 40(10), 3961-3969.

Tsai, H.H. (2012). Global data mining: An empirical study of current trends, future forecasts and technology diffusions. *Expert Systems with Applications*, 39(9), 8172-8181.

Yu, L., Hurley, T., Kliebenstein, J., & Orazem, P. (2012). A test for complementarities among multiple technologies that avoids the curse of dimensionality. *Economics Letters*, 116(3), 354-357.

Zeng, J., Duan, J., Cao, W., & Wu, C. (2012). Topics modeling based on selective Zipf distribution. *Expert Systems with Applications*, 39(7), 6541-6546.

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort.* Cambridge: Addison-Wesley.