Expert Systems with Applications 41 (2014) 406–411

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Quantitative cross impact analysis with latent semantic indexing

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

[a] Fraunhofer INT, Appelsgarten 2, D-53879 Euskirchen, Germany
[b] Ghent University, Faculty of Economics and Business Administration, Tweekerkenstraat 2, B-9000 Gent, Belgium

ARTICLE INFO

Keywords:
Cross impact analysis
Latent semantic indexing
Text mining
Conditional probability

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.

1. 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

* Corresponding author. Tel.: +49 2251 18305; fax: +49 2251 18 38 305.
E-mail addresses: dirk.thorleuchter@int.fraunhofer.de (D. Thorleuchter), dirk.vandenpoel@ugent.bez (D. Van den Poel).
URL: http://www.crm.UGent.be (D. Van den Poel).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.07.066

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, Kim, Wang, Yeh, & Hong, 2012; Tsai, 2012). This is the reason why semantic text classification approaches often outperform knowledge structure based text classification approaches (Thorleuchter & Van den Poel, 2012b).

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, 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.

2. 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 et al., 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.

3. Methodology

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, Allan, and Buckley (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 is applied: for each value of k, LSI is applied and the resulting k