Expert Systems with Applications 41 (2014) 3477–3483

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Semantic compared cross impact analysis

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
Compared cross impact analysis
Web mining

ABSTRACT

The aim of cross impact analysis (CIA) is to predict the impact of a first event on a second. For organization's strategic planning, it is helpful to identify the impacts among organization's internal events and to compare these impacts to the corresponding impacts of external events from organization's competitors. For this, literature has introduced compared cross impact analysis (CCIA) that depicts advantages and disadvantages of the relationships between organization's events to the relationships between competitors' events. However, CCIA is restricted to the use of patent data as representative for competitors' events and it applies a knowledge structure based text mining approach that does not allow considering semantic aspects from highly unstructured textual information. In contrast to related work, we propose an internet based environmental scanning procedure to identify textual patterns represent competitors' events. To enable processing of this highly unstructured textual information, the proposed methodology uses latent semantic indexing (LSI) to calculate the compared cross impacts (CCI) for an organization. A latent semantic subspace is built that consists of semantic textual patterns. These patterns are selected that represent organization's events. A web mining approach is used for crawling textual information from the internet based on keywords extracted from each selected pattern. This textual information is projected into the same latent semantic subspace. Based on the relationships between the semantic textual patterns in the subspace, CCI is calculated for different events of an organization. A case study shows that the proposed approach successfully calculates the CCI for technologies processed by a governmental organization. This enables decision makers to direct their investments more targeted.

1. Introduction

Cross impact analysis (CIA) is a well-known set of related methodologies that enable to analyze events e.g. the occurrence probabilities of events and the conditional probability of one event given another (Blanning & Reinig, 1999; Schuler, Thompson, Vertinsky, & Ziv, 1991). Literature proposes many qualitative approaches where estimations of human experts are collected in workshops or surveys. This information is used to calculate both kinds of probabilities (Banuls, Turoff, & Hiltz, in press; Mitchell, Tydeman, & Curnow, 1977). However, using estimation of human experts is time- and cost expensive. Thus, literature also proposes quantitative approaches that use accessible data to calculate the probabilities (Kim, Lee, Seol, & Lee, 2011). While nearly 80% of all data available in the internet are textual data, some quantitative approaches focus on text mining instead of data mining techniques. They can be distinguish in knowledge structure based approaches (Jeong & Kim, 1997) and semantic approaches (Thorleuchter & Van den Poel, 2014) that both are successfully evaluated for calculating cross-impacts (CI).

Compared cross impact analysis (CCIA) can be used to support the strategic planning of an organization (Thorleuchter, Van den Poel, & Prinzie, 2010). CCIA distinguishes between internal and external events. Internal events occur within the organization. Decision makers of the organization have to select them based on their relevance for a strategic decision problem. CIA is used to calculate the occurrence probabilities of these internal events and the conditional probabilities among them. External events occur at organization's competitors (Pillania, 2011). External events are selected that are equal to internal events. Based on the relationships among external events, CIA is also used to calculate occurrence and conditional probabilities of these events. Thus, the internal CIA depicts organization's events and the external CIA depicts the corresponding events from the competitors. CCIA compares the internal CIA to the external CIA. This enables to identify strengths and weaknesses of organization's own events related to its competitors and it offers the possibility for an organization to learn from the competitors (Trumbach, Payne, & Kongthon, 2006; Woon, Zeineldin, & Madnick, 2011).

An example for an internal event is the development of a technology in the organization's research department. A research

* 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.be (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.2013.10.051

project normally processes one or several technologies that can be used in one or in several application fields in future. Thus, technologies are related to other technologies because they are processed together in a research project and they are also related to application fields (Yu, Hurley, Kliebenstein, & Orazem, 2012). Literature shows four different relationships between technologies: integrative, substitutive, precursive, and successive relationships (Geschka, 1983). The integrative relationship for example shows that technologies occur together while creating an application, e.g. fuel and lubricant technology for creating a powerplant application. The relationships differ for each application and they also change over time. This is because advances in some technologies might lead to the use of new technologies to create an application or the dis-use of existing technologies (Kauffman, Lobo, & Macready, 2000). CIA can be used to identify these relationships. For this, the technologies standing behind the research projects have to be identified as well as the application fields. The identification has to be done by multi-label classification (Tsoumakas & Katakis, 2007). This enables the assignment of several technologies and several application fields to the corresponding research project. Based on this assignment, CI can be calculated. To obtain CIA results that are statistically significant, the number of research projects has to be large (>100 projects). Further, the number of processed technologies also has to be large (>20 technology). Thus, applying CIA to identify technological relationships requires a large research department and a large technological scope of an organization.

To continue this example, external events are defined as technologies developed by organization's competitors. Information about current technological developments can be found in internet websites and in internet blogs. Further, patent data are a valuable information source for upcoming technologies. Full texts of patents are also accessible in the internet. With an internet based environmental scanning, textual information can be extracted from the internet that is related to the technologies processed in the organization. This information represents external events because using this information by other organizations let them become organization's competitors. The extracted websites, blogs, or patent data contain one or several technologies and it also can be applied to one or several application fields. With multi-label classification, the external CIA can be calculated to identify the technological relationships of organization's competitors. Then, CCIA is used to calculate the relative impact of a technology on a different technology. This is done by comparing the internal CIA to the external CIA.

Quantitative CIA approaches that use textual information as input source can be distinguished in knowledge structure based approaches and semantic approaches. While knowledge structure based approaches focus on the aspects of words e.g. term frequency, semantic approaches also consider the aspects of meaning. Especially textual information in the internet might be written by different persons using different wordings. Two texts describing the same event may contain different words. Further, two texts where each describe a different event may both use a large number of the same terms. Semantic approaches are able to consider the meaning of the texts from the internet and thus, they assign them to the corresponding event with higher accuracy (Tsai, 2012).

Literature shows knowledge structure based approaches as well as semantic approaches for quantitative CIA using textual information as input source. However for CCIA, literature only provides a knowledge structure based approach. In contrast to related work, we provide a semantic approach for CCIA. That bridges then current gap in CCIA research and that improves performance of CCIA especially by using texts from the internet.

The proposed approach is based on latent semantic indexing (LSI) that extracts the hidden meaning of textual information from the occurrences and co-occurrences of terms in documents (Luo, Chen, & Xiong, 2011). Semantic textual patterns standing behind the textual information are identified, a latent semantic subspace is built, and the impact of terms and documents on these patterns are calculated (Kuhn, Ducasse, & Girba, 2007). An existing approach from literature (Thorleuchter & Van den Poel, 2014) is taken over for calculating the internal CIA. To calculate the external CIA, web mining is applied for crawling relevant texts. The data are projected into the same latent semantic subspace. The impacts of the texts on the semantic textual patterns are used to calculate the external CIA. This enables to calculate the CCI.

The proposed approach is applied in a case study where the impact of technologies on other technologies is calculated. Data source for the internal CIA are descriptions of research projects funded in 2007 by the German Ministry of Denfense (GE MoD). For the external CIA, texts from the internet are used to represent events from organization's competitors. Semantic textual patterns in the retrieved results are extracted and these patterns are selected that represent a technology. They are assigned to the technologies processed by GE MoD. This enables to process the internal CIA, the external CIA, and thus, the CCIA. The results are compared to a knowledge structure based approach that is applied in the same case study.

Overall, a new CCIA approach is provided applied by semantic text classification. This enables to consider textual information from the internet for CCI calculation. This helps decision makers to better identify own strengths and weaknesses related to competitors.

2. Background

Our proposed CCIA approach calculates the CCI for events with semantic text classification. It is based on an existing semantic CIA approach, it extends an existing CCIA approach, and it is applied in a case study where the CCI of technologies is calculated. In Section 2.1, we give an overview on text classification. Related approaches from literature are introduced in Section 2.2. The impacts of technologies on other technologies are described in Section 2.3.

2.1. Text classification

Text classification is used to assign text to different classes. In contrast to clustering, the classes have to be pre-defined in advanced (Ko & Seo, 2009; Lin & Hong, 2011). Classes can be defined as events and a text can be assigned to one or several of these events. Based on this assignment, the conditional probability of one event given a second event can be calculated. Classification can be processed manually by human experts if the number of texts and classes are small. Otherwise, classification is processed based on automated approaches that use examples for training and testing and a machine based learning procedure (Finzen, Kintz, & Kaufmann, 2012; Sudhamathy & Jothi Venkateswaran, 2012). Well-known knowledge structure approaches that are commonly used for classification are decision tree models (e.g. C4.5), k nearest neighbor algorithm, simple probabilistic algorithms (e.g. naïve Bayes), and support vector machine algorithms (Buckinx, Moons, Van den Poel, & Wets, 2004; D'Haen, Van den Poel, & Thorleuchter, 2013; Lee & Wang, 2012; Shi & Setchi, 2012).

In contrast to these knowledge structure based approaches, semantic approaches identify the dependencies among terms e.g. by calculating term co-occurrences to consider aspects of meaning in texts (Choi, Kim, Wang, Yeh, & Hong, 2012a). Further, these approaches also consider term occurrences and term distributions. LSI is a well-known semantic approach. It is based on algebra eigenvector techniques (Jiang, Berry, Donato, Ostrouchov, & Grady, 1999). Based on the calculated eigenvectors, semantic textual