Semantic Compared Cross Impact Analysis

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

[a] Fraunhofer INT, D-53879 Euskirchen, Appelsgarten 2, Germany, dirk.thorleuchter@int.fraunhofer.de

[b] Ghent University, Faculty of Economics and Business Administration, B-9000 Gent, Tweekerkenstraat 2, Belgium, dirk.vandenpoel@ugent.be URL: http://www.crm.UGent.be

[*] *Corresponding author at: 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 (D. Thorleuchter).

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.

Key Words: Cross Impact Analysis, Latent Semantic Indexing, Text Mining, Compared Cross Impact Analysis, Web Mining.

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, 2013; 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, 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 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, 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 Sect. 2.1, we give an overview on text classification. Related approaches from literature are introduced in Sect. 2.2. The impacts of technologies on other technologies are described in Sect. 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 (Sudhamathy & Jothi Venkateswaran, 2012; Finzen, Kintz, & Kaufmann, 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, 2012). 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 patterns can be selected that occur in several documents from a document collection (Park, Kim, Choi, & Kim, 2012). These semantic textual patterns consist of a list of terms that are semantically related. The aspect of meanings stated in each semantic text pattern can be found in several of the documents (Christidis, Mentzas, & Apostolou, 2012).

Besides LSI, new semantic approaches have been introduced in scientific community with an improved performance. Examples for these new approaches are PLSI (Hofmann, 1999), NMF (Lee & Seung, 1999; Lee & Seung, 2001), and LDA (Blei, Ng, & Jordan, 2003). They are of better performance because the weak point of LSI is the manual selection of a parameter k. The parameter k is the rank of the reduced term-document matrix and it determines the number of semantic textual patterns created by the LSI approach. It has to be selected carefully to obtain an optimized performance. In literature (Thorleuchter & Van den Poel, 2012a), a cross validation procedure based on prediction modeling is often used for a good estimation of parameter k.

However, literature (Thorleuchter & Van den Poel, 2013a) also shows that the selection of k can be used to adapt the semantic textual patterns to a specific event (rank-validation procedure). This adaption decreases performance of the approach on one hand but on the other hand, it leads to many one-to-one correspondences between semantic textual patterns and events. These correspondences can be used to calculate the cross-impacts between events as represented by semantic textual patterns. To identify an optimal value of k where the highest number of one-to-one correspondences is obtained, LSI has to be processed several times for each potential value of k. This is possible, because LSI uses singular value decomposition that is of low computational complexity. LSI is better suited for this task than PLSI, NMF, and LDA because the computational complexity of these approaches is much higher than LSI. Thus, LSI is used in this approach. While the proposed approach is based on a non-optimized LSI approach, the results have to be evaluated carefully to prove the feasibility of this approach.

2.2 Related work

A semantic CIA approach is proposed by Thorleuchter & Van den Poel (2014). This quantitative approach combines CIA with semantic text classification. It discovers the semantic structure of given textual information based on LSI with singular value decomposition. A specific rank-validation procedure is proposed that identifies events from the discovered structure. CIA is applied on the identified events to predict the cross impacts among these events semantically. This approach applies CIA semantically however; it does not apply CCIA semantically.

A knowledge structure based CCIA approach is proposed by Thorleuchter, Van den Poel, & Prinzie (2010). It supports the planning of research and development (R&D) for organizations with a wide technological scope. The relative impacts of technologies on other technologies are calculated in three steps. In a first step, a CIA is applied on technologies processed within the organization. In a second step, patent data are used as representative for technologies processed by organization's competitors and CIA is applied on these technologies. In a last step, the CIA results of the first step are compared to the corresponding CIA results of the second step. As a result, a CCI index is built that indicates areas between two technologies where the organization excels or where the organization has relative weaknesses. This approach applies CCIA however; it does not apply CCIA semantically.

2.3 Impacts of technologies on other technologies

In the case study, the proposed approach is applied to identify technological impacts based on textual descriptions. Literature has shown that many of these impacts exist (Choi et al., 2012; Subramanian & Soh, 2010; Radder, 2009; Jiménez, Garrido-Vega, Díez de los Ríos, & González, 2011; Herstatt & Geschka, 2002). Five of these impacts that are relevant to this study are presented below.

A technology has an impact on a similar technology. Technologies are similar to other technologies if they stem from the same technology field but focus on different aspects within the technology field. An example is radar technology that can be divided into active and passive radar technology. Advances in active radar technology can lead to advances in passive radar technology because it is often possible to take over new findings to a similar technology. From text classification point of view, textual descriptions of two similar technologies are also similar because they both contain the same technical terms from the core area of the technology field. However, term distributions and term co-occurrences differ because each description has its own focal point. Thus, similar technologies can be identified by text classification considering term appearances, distributions, and co-occurrences.

A second kind of impact is that technologies substitute each other e.g. semiconductor technology and vacuum tube technology. The technologies can be used to create the same application. Advances in a technology might lead to the dis-use of a substitutive technology e.g. transistors have replaced vacuum tubes in many applications. From text classification point of view, the descriptions of substitutive technologies contain the same terms describing

the application field. They also contain different terms to describe the corresponding technology field. Thus, the identification of substitutive technologies is possible with text classification.

A technology impacts a complementary technology if both technologies are used to create an application. An example is fuel technology and lubricants technology. For creating a power plant application both technologies are used. However, the performance of the power plant application is probably limited by one of the used technologies e.g. using high quality fuel and low quality lubricants or vice versa. This is the well-known bottleneck phenomenon. Advances in one technology forces researches and research planners to focus on possible advances in its complementary technology to avoid this phenomenon. From text classification point of view, descriptions of these technologies contain the same terms from the application field and they also use different terms describing the technology field. This is similar to substitutive technologies as mentioned above.

A further impact is based on predecessor or successor technologies. A technology impacts a second technology if it precedes the second technology during the process of creating an application. Creating a laser system is often done by using adaptive optics mirrors technology. This is a successor technology because it is based on the liquid crystal technology used for creating mirrors. It is trivial that advances in a predecessor technology directly influence its successor technologies. From text classification point of view, textual descriptions from both technologies contain the same terms e.g. describing the mirror aspects and they also contain different terms. This is because predecessor technologies use highly detailed information from a very small part of a technology area while successor technologies use a broader view on the technology.

As a result, the different characteristics of texts (term occurrences, distribution, co-occurrences etc.) can be used to identify that a specific technology is mentioned in a given text (e.g. an internet website) and that this technology is impacted by other technologies (Thorleuchter & Van den Poel, 2013c). While texts extracted from the internet are normally written by different persons, semantic text classification approaches are normally better suited for this task than knowledge structure based text classification approaches. This is because they consider the aspect of meaning in texts.

3 Methodology

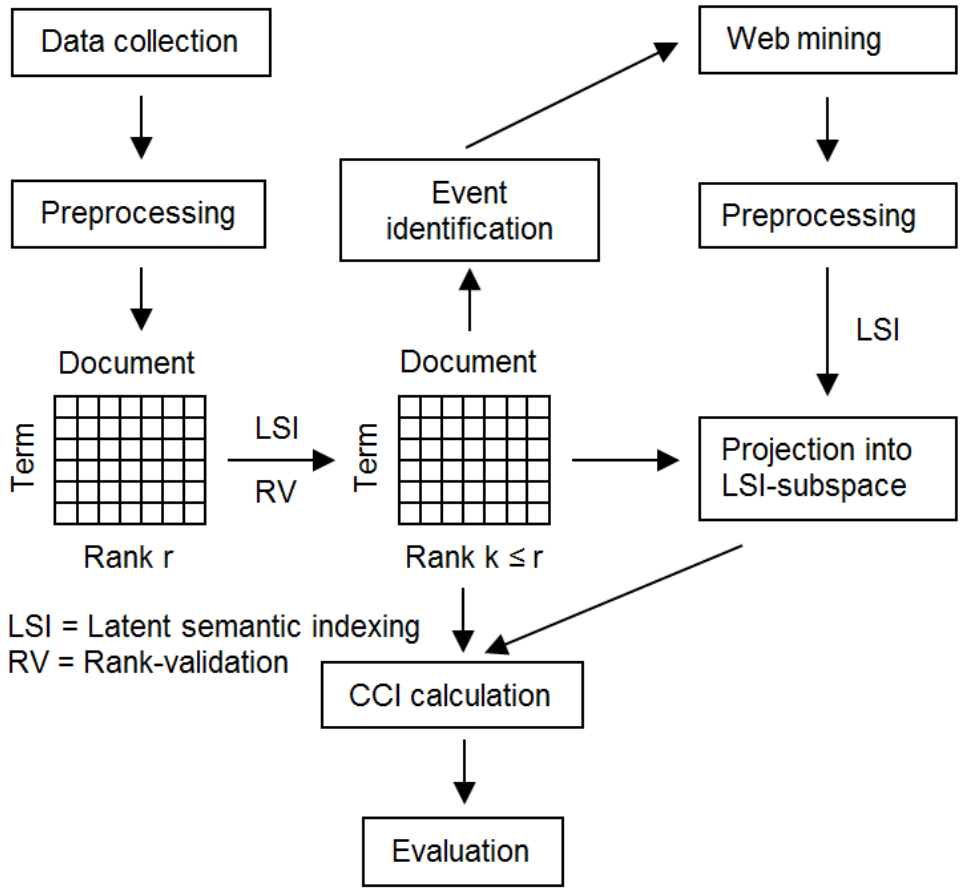

Fig. 1 shows the processing of the semantic CCI methodology in different steps.

The proposed semantic CCI methodology is depicted in Fig. 1. In a data collection step, events have to be defined with relevance to an existing strategic decision problem. A collection of textual documents have to be provided. They should stem from the organization and their content should contain the description of one or several events.

The collection of textual documents is preprocessed in the next step. Existing methods and tools from text mining are used to remove specific elements, to split text in terms, and to check for typographical errors. Term filtering methods (e.g. stop word filtering, part-of-speech tagging, and stemming) are also applied to reduce the number of different terms. A further reduction of terms is done by applying Zipf's law (Zeng, Duan, Cao, & Wu, 2012; Zipf, 1949). As a result of the preprocessing step, a term vector based on vector space model and on the reduced number of terms is built for each document from the collection (Thorleuchter & Van den Poel, 2013b). The components of the vectors are weighted frequencies based on the proposed term weighting scheme from Salton et al. (1994).

A term-by-document matrix is created based on the term vectors from the document collection. The rank of this matrix is unmanageable high because of the large number of different terms that have to be considered despite reducing the number with filtering methods (Thorleuchter & Van den Poel, 2012c). Thus, LSI is applied together with singular value decomposition to reduce the rank of the term-by-document matrix from r to k. This reduction is equivalent to the identification of k semantic textual patterns that can be found as latent patterns in the data structure.

The selection of k is done by a rank-validation procedure (Thorleuchter & Van den Poel, 2013a). LSI is applied several times for different values of k and for each time, the resulting k semantic textual patterns are compared to the defined events from the data collection step. The aim of this comparison is to identify one-to-one correspondences (an exact pairing) between the k semantic textual patterns and the defined events. This is done manually by human experts or alternatively by use of text similarity measures and by considering a specific threshold. Let n_k be the number of one-to-one correspondences for each k. The variable k is set to a value where n_k is at its maximum. As a result, the rank-validation procedure leads to the identification of n_k events that occur semantically in the textual collection.

After selecting k by the rank-validation procedure, singular value decomposition is applied to split the original term-by-document matrix in three matrices U, Σ, and V^t.

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

The diagonal matrix Σ contains the singular values ordered by size. LSI reduces the rank of the three matrices to k by discarding the corresponding columns from k+1 on. This keeps the large singular values stored in Σ and discards the smaller ones.

$$A_k = U_k \Sigma_k V^t_k \quad (2)$$

Matrix U_k represents the impact of the reduced number of terms from the preprocessing step (rows) on the k semantic textual patterns (columns) as defined by the rank validation procedure. Matrix V_k represents the impact of each document from the collection (rows) on the k semantic textual patterns (columns) (Thorleuchter, Van den Poel, & Prinzie, 2012b).

Thus, matrix U_k is used to identify events from the content of the semantic textual patterns and matrix V_k is used to calculate the CI between these semantic textual patterns that represent events. In both matrices, the matrix components show the corresponding impacts that is a value in [-1,..,1].

Based on the $n \leq k$ semantic textual patterns that represent events, an environmental scanning procedure is applied to collect information about the identified events. This is done by web mining, where documents in the internet are retrieved that contain at least one of the n semantic textual patterns. The starting point of web mining is always a set of search queries. This is built separately for each of the n semantic textual patterns. Terms from matrix U_k with the highest impact on a semantic textual pattern (the corresponding matrix component exceeds a specific threshold) are used to build the search queries based on their co-occurrences in the document collection. The processing is taken over from Thorleuchter & Van den Poel (2013d).

A web search advanced programming interface (e.g. from search engine Google) is used to execute the sets of search queries automatically. The results are a collection of website addresses. The full text from these addresses (e.g. webpages and blogs) is crawled and the texts are stored separately in documents. A preprocessing step is applied on these documents as described above. The documents are projected into the latent semantic subspace that was created by LSI before. The matrix V'_k is built that contains the impact of each retrieved document (rows) on each of the k semantic textual patterns (columns) and thus, on each of the n events.

Based on matrix V'_k and on matrix V_k, the CI from the environmental scanning procedure and from the provided document collection as well as the CCI is calculated in accordance to Thorleuchter, Van den Poel, & Prinzie (2010):

Definition 1. Let $N_{ext}(A)$ be the number of retrieved documents that are related to event A. $N_{ext}(A)$ is calculated by the number of documents where the component value of matrix V'_k - from the corresponding document (row) and from the corresponding semantic textual pattern (column) - is above a specific threshold.
Let $N_{ext}(A \cap B)$ be the number of retrieved documents related to both, event A and event B. $N_{ext}(A \cap B)$ is calculated by the number of documents where both component values of matrix V'_k - from the corresponding document (row) and from the two corresponding semantic textual patterns (columns) - are above a specific threshold.

Let $CI_{ext}(A,B)$ be the external conditional probability of event A given event B as calculated by:

$$CI_{ext}(A,B) = P_{ext}(B|A) = N_{ext}(A \cap B) / N_{ext}(A) \quad (3)$$

Definition 2. Let $N_{int}(A)$ be the number of documents from the document collection that are related to event A. $N_{int}(A)$ is calculated by the number of documents where the component value of matrix V_k - from the corresponding document (row) and from the corresponding semantic textual pattern (column) - is above a specific threshold.
Let $N_{int}(A \cap B)$ be the number of documents from the document collection related to both, A and B. $N_{int}(A \cap B)$ is calculated by the number of documents where both component values of matrix V_k - from the corresponding document (row) and from the two corresponding semantic textual patterns (columns) - are above a specific threshold.
Let $CI_{int}(A,B)$ be the internal conditional probability of A given B as calculated by:

$$CI_{int}(A,B) = P_{int}(B|A) = N_{int}(A \cap B) / N_{int}(A) \quad (4)$$

The result values of $CI_{ext}(A,B)$ and $CI_{int}(A,B)$ are in [0,..,1]. A result value of zero means that no impact can be seen of event A on event B and a result value of one means that event A has a strong impact on event B. A value between zero and one indicates the impact strength. The decision whether an impact exists or not can be done using thresholds.

Definition 3. Let c_{ext} be the external threshold. $BCI_{ext}(A,B)$ as the external Boolean cross impact index is defined as follows:

$$BCI_{ext}(A,B) = \begin{cases} true & (CI_{ext}(A,B) > c_{ext}) \\ false & (CI_{ext}(A,B) \le c_{ext}) \end{cases} \quad (5)$$

Definition 4. Let c_{int} be the internal threshold. $BCI_{int}(A,B)$ as the internal Boolean cross impact index is defined as follows:

$$BCI_{int}(A,B) = \begin{cases} true & (CI_{int}(A,B) > c_{int}) \\ false & (CI_{int}(A,B) \le c_{int}) \end{cases} \quad (6)$$

Definition 5. Let CCI(A,B) indicate the difference between the internal and external Boolean cross impact index.

$$CCI(A,B) = \begin{cases} 1 & BCI_{\text{int}}(A,B) = true \wedge BCI_{ext}(A,B) = false \\ -1 & BCI_{ext}(A,B) = true \wedge BCI_{\text{int}}(A,B) = false \\ 0 & else \end{cases} \quad (7)$$

4 Case Study

The proposed approach is applied in a case study. We use the same case study as already applied by the knowledge structure based CCI approach and by the semantic CI approach (see Sect. 2.2). This enables a comparison to the related approaches and it also enables to use existing evaluated results from these approaches. Defense based technology areas are defined as events. We use the 32 technology areas that are listened by the technology taxonomy of the European Defense Agency (EDA). German Ministry of Defence (GE MoD) is selected as organization. A collection of 985 documents is used that describe research projects from the GE MoD in 2007. The documents are in English language. Each project processes one or several technologies (events). Thus, the collection of documents represents organization's internal events.

The first steps - from the data collection step via the processing step up to the event identification step - are taken over from Thorleuchter & Van den Poel (2014): The term-document matrix is built based on the pre-processed textual collection. The rank-validation procedure is applied from k = 2 to k = 35. Thus, LSI with singular value decomposition is processed 34 times. Each time, k semantic textual patterns are created. Human experts compare the patterns to the 32 technology areas manually. They identify n_k one-to-one correspondences. The results show that up to k = 7 and from k = 33 on, n_k equals zero. In the area of $8 \leq k \leq 18$, all values of n_k are smaller than 9 and in the area of $20 \leq k \leq 32$, all values of n_k are smaller than 10. Thus, k = 19 is selected where n_k equals 10. The corresponding 10 semantic textual patterns are selected and depicted in Table 1.

Table 1: Events (technology areas from EDA taxonomy) that are identified by the rank-validation procedure

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 component values of matrix V_k show the impact of terms on the identified 10 semantic textual patterns. Terms with an impact above a specific threshold on the patterns are used to create internet search queries. The search queries are in English language to prevent translation problems. For each of the 10 semantic textual patterns, a set of search queries is created that consists of four terms each. This is done in accordance to Thorleuchter & Van den Poel (2013d). An example is ‘Acoustic +Vibration +Absorbing +Materials’ that is used for the event ‘Signature Related Materials’. Thus, each set of search queries describes the event standing behind the semantic textual pattern.

Google advanced programming interface is used to execute the queries automatically in December 2012. Hyperlinks from the retrieved results are used by a crawler to extract the full text of the documents. The documents are validated e.g. by deleting double occurrences of documents. As a result, 5364 documents are identified as representative for external events. The documents are projected into the same latent semantic subspace as created by the GE MoD document collection. Then, the internal and the external CI are calculated as well as the CCI for the 10 technology areas.

The data characteristics are depicted in Table 2.

Table 2: Data characteristics of the case study

Number of documents from GE MoD	985
Number of identified events	10
Average number of search queries per event	30
Number of search queries in total	303
Number of retrieved documents per search query	20
Number of retrieved documents in total (after validating)	5364

5 Results and Evaluation

The case study identifies 10 events. The impact of a first event on a second is different than the impact of the second on the first. Thus, the results of the case study are 90 CCI indices as calculated by two times the binomial coefficient 10 choose 2. This requires the calculation of 90 internal CI probabilities and the calculation of 90 external CI probabilities. The 90 internal CI probabilities are already calculated and evaluated by the semantic CI approach (see Sect. 2.2). Thus, this evaluation focusses on the 90 external CI probabilities. The knowledge structure based CCI approach (see Sect. 2.2) - furthermore it is named comparative study - also has calculated these 90 external CI probabilities. However, two differences can be seen: The comparative study uses a knowledge structure based approach in contrast to the semantic approach used here. Further, data collection is restricted on patent data collected in 2007 in contrast to the environmental scanning procedure as applied here. Considering current scientific articles, webpages or internet blogs allows our approach to use current information about technologies with a large scope. Thus, the document collection used in our approach for external CI is different to the document collection of the comparative study.

Table 3: Result matrix of the $CI_{ext}(A,B)$ e.g. $CI_{ext}(A02, A03) = 0.08$

	A02	A03	A04	A05	A08	B02	B04	B05	B06	B08
A02	-	0.08	0.08	0.01	0	0	0	0.76	0.07	0
A03	0.03	-	0.15	0.45	0	0.04	0.14	0	0.14	0
A04	0.03	0.16	-	0.14	0	0.05	0.11	0.12	0.37	0
A05	0	0.29	0.08	-	0.05	0.03	0.10	0.01	0.34	0.10
A08	0	0	0	0.08	-	0	0.01	0	0.16	0.36
B02	0	0.10	0.12	0.12	0	-	0	0	0	0
B04	0	0.28	0.21	0.32	0.02	0	-	0.01	0.09	0
B05	0.74	0	0.32	0.04	0	0	0.01	-	0.08	0
B06	0.02	0.09	0.23	0.36	0.10	0	0.03	0.02	-	0.03
B08	0	0	0	0.18	0.36	0	0	0	0.05	-

Because of these two differences, the results of the comparative study should be different in some CI values to the results from our case study. Nevertheless, the results of the

comparative study are used as ground truth for the evaluation and the differences are discussed manually.

Table 3 shows the 90 external CI probabilities in different grayscales from bright to dark considering the five cases: no external cross impact: $CI_{ext}(A,B) = 0$; low cross impact: $0 < CI_{ext}(A,B) \leq 0.25$; medium cross impact: $0.25 < CI_{ext}(A,B) \leq 0.50$; high cross impact: $0.50 < CI_{ext}(A,B) \leq 0.75$; very high cross impact: $CI_{ext}(A,B) > 0.75$.

Table 4: $CI_{ext}(A,B)$ results of selected technology area pairs compared to results of a comparative study

Techn. area A	Techn. area B	CI_{ext} (A,B)	BCI_{ext} (A,B)	CI'_{ext} (A,B)	BCI'_{ext} (A,B)	Res_{ext} (A,B)	Res_{int} (A,B)
A02	B05	0.76	true	0.58	true	0.14	-0.01
A03	A05	0.45	true	0.30	true	0.11	-0.02
B05	A02	0.74	true	0.46	true	0.24	0.06
B04	A05	0.32	true	0.35	true	-0.07	0.02
B02	A05	0.12	false	0.07	false	0.01	0.01
B08	A08	0.36	true	0.26	true	0.06	0.07
A05	A03	0.29	true	0.26	true	-0.01	0.05
A08	B08	0.36	true	0.22	true	0.10	0.00
A05	B02	0.03	false	0.01	false	-0.02	0.00
A05	B06	0.34	true	0.20	false	0.10	-0.01
A05	B04	0.10	false	0.21	true	-0.15	0.01

Table 4 shows these 11 external CI values where the corresponding internal CI values are above a specific threshold and thus, the 11 technology area pairs are of high relevance for the organization. The influencing technology area is labeled with 'Techn. Area A' and the influenced technology area is labeled with 'Techn. Area B'. $CI_{ext}(A,B)$ is the external CI value as calculated by the proposed approach and $CI'_{ext}(A,B)$ is the external CI value as calculated by the comparative study. $BCI_{ext}(A,B)$ is true if $CI_{ext}(A,B)$ is equal to or greater than threshold $r = 0.20$. $BCI'_{ext}(A,B)$ is the corresponding value to $BCI_{ext}(A,B)$ that refers to $CI'_{ext}(A,B)$. The

residuals of $CI_{ext}(A,B)$ and $CI'_{ext}(A,B)$ are $Res_{ext}(A,B)$ and they are compared to the residuals $Res_{int}(A,B)$ of the internal CI values as calculated by the semantic CI approach (see Sect. 2.2).

The residuals of the internal CI values are based on the difference between results of a semantic CI approach and results of a knowledge structure based CIA approach by using the same document collection. These residuals are much smaller than the residuals calculated from the external CI values. Thus, the large differences between $CI_{ext}(A,B)$ and $CI'_{ext}(A,B)$ stem from the fact that the document collection (from the internet) of the case study is different to the document collection (patent data) of the comparative study.

An example for this is presented below. Research in the technology field A02 'Signature Related Materials' examines various materials to improve their absorbing characteristics. It focusses on materials by reducing its radar, infrared, and acoustical radiation. Research in the technology field B05 'Signature Control and Signature Reduction' has the aim to analyze and manipulate the absorbing characteristics of land, air, and maritime vehicles. It focusses on systems by reducing its radar, visible, ultraviolet, infrared, acoustical, electrical, electrochemical, and magnetic radiation.

The external CI of in A02 on B05 is large in the comparative study (0.58) but it is larger in the case study (0.76). This is because patents in the A02 technology area focus on material aspects rather than on system aspects. Thus, the term distribution in these patens is more related to a material sciences than to system sciences. Besides patents, the environmental scanning procedure also identifies further documents from the internet e.g. websites and blogs. These documents are more often related to system related aspects than patents. This explains the differences in the external CI values.

A further explanation could be given for the differences of the external CI of B8 'Simulators, Trainers and Synthetic Environments' on A8 'Computing Technologies & Mathematical Techniques'. Most of all current breakthroughs in simulators or in synthetic environments are based on computing technologies and thus, the number of documents where both technology fields are mentioned increases each year. The use of current documents - instead of using patents from 2007 - influences the external CI of B8 on A8 as well as the external CI of A8 on B8.

Table 5 shows the confusion matrix of the evaluated results where the calculated 90 external CI probabilities are compared to the comparative study. In 8 cases, the value of $CI'_{ext}(A,B)$ is above 0.20 and thus, $BCI'_{ext}(A,B)$ is true. Further, $BCI'_{ext}(A,B)$ is false in 82 cases. The frequent baseline is set to about 9% as calculated by 8 divided by (82 + 8). BCI_{ext} is true in 14

cases and $BCI_{ext}(A,B)$ is false in 76 cases. In 7 seven cases, both, $BCI_{ext}(A,B)$ and $BCI'_{ext}(A,B)$ are true. $BCI_{ext}(A,B)$ is true and $BCI'_{ext}(A,B)$ is false in 7 cases, too. In one case, $BCI_{ext}(A,B)$ is false and $BCI'_{ext}(A,B)$ is true. Thus, the precision of the proposed approach is 7 / 14 = 50% and the recall is 7 / 8 = 87%. The results outperform the frequent baseline as set to 9% precision at 87% recall.

Table 5: Confusion matrix

		Predictive Class	
		Yes	No
Actual class	Yes	7	1
	No	7	75

6 Conclusion

This paper provides a new semantic CCI approach. This is in contrast to the knowledge structure based approaches presented in literature. The strength of a semantic approach is that the aspect of meaning is considered rather than syntactical aspects. This allows the use of highly unstructured data e.g. documents from the internet where different people write texts in different syntactical styles. To use this strength, an environmental scanning procedure is introduced that collects relevant documents from the internet. Thus, CCI can be processed based on a wide information scope. That is also in contrast to existing approaches.

The new approach is applied in the same case study as used for a knowledge structure based approach. Differences between the results of the approaches are discussed and an evaluation shows that the approach outperforms the frequent baseline.

Overall, the new approach supports decision makers by providing current strengths and weaknesses. The time series of the calculated CCI values are probably an interesting field of further research. Strengths and Weaknesses of organizations could be traced over time considering current information from the internet. This possibly allows decision makers to improve strategic planning.

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.

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.

Buckinx, W., Moons, E., Van den Poel, D., & Wets, G. (2004). Customer-adapted coupon targeting using feature selection. *Expert Systems with Applications*, 26(4), 509-518.

Choi, O., Kim, K., Wang, D., Yeh, H., & Hong, M. (2012). Personalized mobile information retrieval system. *International Journal of Advanced Robotic Systems*. http://dx.doi.org/10.5772/50910.

Choi, S., Park, H., Kang, D., Lee, J. Y., & Kim, K. (2012). An SAO-based text mining approach to building a technology tree for technology planning. *Expert Systems with Applications*, 39(13), 11443-11455.

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.

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.

Finzen, J., Kintz, M., & Kaufmann, S. (2012). Aggregating web–based ideation platforms. *International Journal of Technology Intelligence and Planning*, 8(1), 32-46.

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

Herstatt, C., & Geschka, H. (2002). Need assessment in practice - methods, experiences and trends. *International Journal of Entrepreneurship and Innovation Management*, 2(1), 56-68.

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.

Jiménez, C. H. O., Garrido-Vega, P., Díez de los Ríos, J. L. P., & González, S. G. (2011). Manufacturing strategy–technology relationship among auto suppliers. *International Journal of Production Economics*, 133(2), 508-517.

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.

Ko, Y., & Seo, J. (2009). Text classification from unlabeled documents with bootstrapping and feature projection techniques. *Information Processing & Management*, 45(1), 70-83.

Lee, C. H., & Wang S. H. (2012). An information fusion approach to integrate image annotation and text mining methods for geographic knowledge discovery. *Expert Systems with Applications*, 39(10), 8954-8967.

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.

Lin, M. H., & Hong, C. F. (2011). Opportunities for Crossing the Chasm between Early Adopters and the Early Majority through New Uses of Innovative Products. *The Review of Socionetwork Strategies*, 5(2), 27-42.

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.

Pillania, R.K. (2011). The state of research on technological uncertainties, social uncertainties and emerging markets: A multidisciplinary literature review. Technological Forecasting and Social Change, 78(7), 1158-1163.

Radder, H. (2009). Science, Technology and the Science-Technology Relationship. *Philosophy of Technology and Engineering Sciences*, 2009, 65-91.

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.

Shi, L., & Setchi, R. (2012). User-oriented ontology-based clustering of stored memories. *Expert Systems with Applications*, 39(10), 9730-9742.

Subramanian, A. M., & Soh, P. H. (2010). An empirical examination of the science–technology relationship in the biotechnology industry. *Journal of Engineering and Technology Management*, 27(3-4), 160-171.

Sudhamathy, G. & Jothi Venkateswaran, C. (2012). Fuzzy Temporal Clustering Approach for E-Commerce Websites. *International Journal of Engineering and Technology*, 4(3), 119-132.
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.
Thorleuchter, D, & Van den Poel, D. (2014). Quantitative cross impact analysis with latent semantic indexing. *Expert Systems with Applications*, 41(2), 406-411.
Trumbach, C., Payne, D., & Kongthon, A. (2006). Technology mining for small firms: Knowledge prospecting for competitive advantage. Technological Forecasting and Social Change, 73(2006), 937-949.
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.
Tsoumakas, G. & Katakis, I. (2007). Multi Label Classification: An Overview, International Journal of Data Warehousing and Mining, 3(3), 1-13.
Woon, W.L., Zeineldin, H., & Madnick, S. (2011). Bibliometric analysis of distributed generation. Technological Forecasting and Social Change, 78(3), 408-420.
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.