UNIVERSITEIT GENT

FACULTY OF ECONOMICS AND BUSINESS ADMINISTRATION

ESSAYS ON TEXT MINING
FOR IMPROVED DECISION MAKING

Dirk Thorleuchter

2011

Dissertation submitted to the Faculty of Economics and Business Administration of Ghent University in fulfillment of the requirements for the degree of Doctor in Applied Economic Sciences

Supervisors:

Prof. Dr. Dirk Van den Poel

Prof. Dr. Anita Prinzie

Doctoral Exam Committee

Prof. Dr. Dirk Van den Poel
Advisor, Ghent University

Prof. Dr. Anita Prinzie
Co-Advisor, Ghent University

Prof. Dr. Aimé Heene
Ghent University

Prof. Dr. Mario Vanhoucke
Ghent University

Prof. Dr. Marc De Clercq
Ghent University

Prof. Dr. Patrick Van Kenhove
Ghent University

Prof. Dr. Bart Baesens
Katholieke Universiteit Leuven, Belgium & University of Southampton, UK

Prof. Dr. Horst Geschka
University of Darmstadt, Germany

Table of Contents

1. Acknowledgments

An undertaking such as a dissertation is not completed without the support of many people. My first debt of gratitude must go to my advisor, Prof. Dr. Dirk van den Poel. He patiently provided the vision, encouragement and advice necessary for me to precede through the doctorial program and complete my dissertation.

I'd also like to give special thanks to Prof. Dr. Anita Prinzie for her support, guidance and good ideas.

Special thanks to my doctoral committee, Prof. Dr. Aimé Heene, Prof. Dr. Mario Vanhoucke, Prof. Dr. Marc De Clercq, Prof. Dr. Patrick Van Kenhove, Prof. Dr. Bart Baesens, and Prof. Dr. Horst Geschka for their helpful suggestions. I appreciate the members of the doctoral committee who took time from their busy schedules to complete this survey. Without their help, this dissertation would not have been possible.

This undertaking is realized beside my (full-time) job at the Fraunhofer research institute INT. I'd like to give a special thanks to my department chief and deputy director Dr. Joachim Schulze for his helpful suggestions and his support. Further, thanks to the institute director Prof. Dr. Uwe Wiemken. Additionally, I thank all my colleagues especially Dr. Wilfried Gericke and Jörg Fenner for their support and good advise.

Finally, I give special thanks to family and friends for their supporting during this work-intensive time.

2. Nederlandstalige Samenvatting

In principe verstaat men onder “Data Mining” het systematisch gebruik methoden op een dataset met als doel patroonherkenning. De database kan worden onderverdeeld in gestructureerde gegevens (bv. numerieke gegevens in databases, afbeeldingen, enz.), semi-gestructureerde gegevens (zoals vrije tekstinformatie in een database) en ongestructureerde gegevens (zoals vrije tekstdocumenten). De meeste analyses met behulp van “data mining” worden uitgevoerd op gestructureerde gegevens. Het onderzoek van semi-gestructureerde en in het bijzonder van ongestructureerde data en het ontwikkelen van nieuwe methoden hiervoor vormt een snelgroeiend onderzoeksgebied in de afgelopen jaren. Hieruit zijn “text mining” en “web content mining” ontstaan in een streven naar het extraheren van hoogwaardige informatie uit ongestructureerde teksten respectievelijk internetsites.

Dit proefschrift is samengesteld uit verschillende studies. Meer in het bijzonder worden methoden van “text mining” en “web content mining” gebruikt om marketeers/personeel te assisteren in verschillende deeldomeinen van hun besluitvorming. Belangrijkste bronnen van informatie zijn hier de tekstinformatie uit boeken, tijdschriftartikelen, wetenschappelijke publicaties, patenten en websites.

“Latent semantic indexing” (LSI) is een bekende methode, die kan worden gebruikt voor semantische analyse van teksten, alsook voor classificatie. Door het combineren van deze methode met verschillende methoden uit “text mining” en webmining wordt in de eerste studie aangetoond, dat door de geautomatiseerde analyse van de webpagina's van de bestaande klanten nieuwe zakelijke klanten voor een bepaald bedrijf kunnen worden geïdentificeerd. Bovendien kunnen deze nieuwe potentiële klanten ook worden geëvalueerd op basis van hun verwachte winstgevendheid voor het gegeven bedrijf. Dit maakt het klantacquisitieproces effectiever.

Een tweede studie toont aan op welke wijze een classificatiemodel kan worden gebouwd (met behulp van LSI en “web content mining”), dat succesvolle e-commerce bedrijven van minder succesvolle ondernemingen onderscheidt, uitsluitend op basis van de analyse van de websites van deze respectieve bedrijven. Dit kan marketingmedewerkers inzichten verschaffen in het ontwerp en de optimalisatie van hun website van het bedrijf.

Een belangrijke succesfactor voor grote bedrijven is een goede samenwerking tussen marketing en de afdeling voor onderzoek en ontwikkeling (O & O), met name om tot succesvolle nieuwe producten te komen. Hier worden medewerkers van de marketingafdeling geconfronteerd met de uitdaging van de ondersteuning van het hele proces vanaf de ideegeneratie, selectie van nieuwe ideëen voor nieuwe producten over de fasen van onderzoek, ontwikkeling en productie tot het op de markt brengen van de nieuwe producten. In verschillende studies laten we zien hoe activiteiten van marketingmedewerkers kunnen worden ondersteund door "text mining"-methoden in hun besluitvorming, meer in het bijzonder voor de O & O-interface.

Een belangrijke taak in de ontwikkeling van nieuwe producten is de selectie van ideeën voor nieuwe producten, die als uitgangspunt dienen voor de daaropvolgende onderzoeks- en ontwikkelingsactiviteiten. Gebaseerd op de bekende "technology push" en "market pull"-effecten zijn hiervoor aan de ene kant nieuwe technologische ideeën nodig (b.v. vanuit de wetenschap), maar ook ideeën, voortgebracht vanuit de behoeften van de consument. De methodologie, gepresenteerd in de studies 3 tot en met 6 moeten marketeers ondersteunen bij het opsporen van dergelijke ideeën.

De derde en vierde studie behandelt de automatische identificatie van nieuwe technologische ideeën op basis van technologische teksten, zoals uit onderzoekspublicaties en patenten. Dit wordt gerealiseerd in de derde studie vanuit de filosofie dat door het gebruik van collocaties (het veelvuldig samen voorkomen van woorden) een automatische detectie van technologische ideeën mogelijk is.

De vierde studie bouwt voor op de derde studie. Hierin wordt namelijk een nieuwe similariteitsmaat voorgesteld alsook een methodologie die de geautomatiseerde ontdekking van technologische ideeën mogelijk maakt via een webgebaseerde applicatie. Het succes van deze methode wordt aangetoond door vergelijking met alternatieve bestaande heuristische gelijkenismaten.

De ideeën voor productontwikkeling in de vierde studie moesten innovatief zijn. Dit verhoogt de kans dat het uiteindelijke product een commercieel succes wordt. In de vijfde studie wordt een methode gepresenteerd die het potentieel inschat van nieuwe technologische ideeën. Om dit doel te bereiken wordt het abstracte begrip "innovatie" in een concreet toepasbare methode voor "text mining" omgezet. Het succes van deze aanpak blijkt op basis van een benchmark gebaseerd op frequentie.

De zesde studie onderzoekt de automatische identificatie van ideeën, die consumentenbehoeften voorstellen. Suggesties, wensen en ideeën van consumenten over de producten uit Internetblogs worden geanalyseerd in termen van het feit hoe "nieuw" deze ideeën zijn. Ten einde dit doel te bereiken, wordt de methodologie van de vierde studie aangepast en uitgebreid met elementen uit “Web Content Mining”. Hierin wordt aangetoond dat een automatische identificatie van de consumentenideeën succesvol kan zijn ondanks het gebruik van weinig gestandaardiseerde omgangstaal.

Een belangrijke taak in het kader van de ontwikkeling van nieuwe producten is het creëren van een technologisch "Marktoverzicht", waarbij technologie-invloeden en trends van de eigen organisatie worden geplaatst tegenover deze van concurrenten. Een nieuwe aanpak om deze in grotere organisaties, zoals overheidsorganisaties uit te voeren, wordt gepresenteerd in studie 7. De methodologie achter deze aanpak combineert de “cross-impact” methode om technologische verbanden te identificeren, met “text mining”-methoden om beschrijvingen van in de eigen onderneming ontwikkelde O & O-projecten met deze van patenten van concurrenten te kunnen vergelijken. Als gevolg hiervan komen de relatieve technologische sterkten en zwakten van de organisatie aan het licht, wat op zich opnieuw het markoverzicht verduidelijkt.

3. Deutsche Zusammenfassung

Grundsätzlich versteht man unter Data Mining die systematische Anwendung von Methoden auf einen Datenbestand mit dem Ziel der Mustererkennung. Der Datenbestand lässt sich unterteilen in strukturierte Daten (z.B. numerische Daten in Datenbanken, Bilder etc.), semi-strukturierte Daten (z.B. freie Textinformationen sind in einer Datenbank) und unstrukturierte Daten (z.B. freie Textdokumente). Die meisten Methoden des Dataminings werden auf einen strukturierten Datenbestand eingesetzt. Die Untersuchung von semi-strukturierten und im Besonderen von unstrukturierten Daten sowie die Erzeugung neuer Methoden hierfür stellt in den letzten Jahren ein neu aufkommendes und stark wachsendes Forschungsgebiet dar. Hieraus entwickelte sich das Textmining und das Web Content Mining, als Prozess zur Ermittlung qualitativ hochwertiger Informationen aus unstrukturierten Texten bzw. aus Internetseiten.

Diese Dissertation setzt sich aus verschiedenen Studien zusammen. Hier werden Methoden des Textminings und des Web Content Minings genutzt, um Mitarbeiter im Marketing in unterschiedlichen Bereichen bei ihrer Entscheidungsfindung zu unterstützen. Wesentliche Informationsquellen sind hierbei Textinformationen aus Büchern, Zeitschriftenartikeln, Forschungspublikationen, Patente und Internetseiten.

Latent semantic indexing (LSI) ist eine bekannte Methode, die u.a. zur semantischen Analyse und Klassifikation von Texten verwendet werden kann. Durch Kombination dieser Methode mit verschiedenen Methoden aus dem Bereich des Text- und Webminings wird in der ersten Studie aufgezeigt, dass sich durch Analyse von Internetseiten bestehender Kunden neue Unternehmenskunden für ein vorgegebenes Unternehmen automatisiert identifizieren lassen. Zudem können diese neuen potentiellen Kunden auch hinsichtlich ihrer zu erwartenden Profitabilität für das vorgegebene Unternehmen bewertet werden. Damit können Mitarbeiter im Bereich der Akquise wirkungsvoll unterstützt werden.

In einer zweiten Studie wird gezeigt, wie durch LSI und Web Content Mining ein Klassifikationsmodell aufgebaut werden kann, dass erfolgreiche e-Commerce Unternehmen von weniger Erfolgreichen unterscheidet, allein aufgrund der Analyse der Internetseiten des Unternehmens. Dies unterstützt Mitarbeiter des Marketings bei der Gestaltung und Optimierung der Internetseite ihres Unternehmens.

Ein wesentlicher Erfolgsfaktor für große Unternehmen ist eine gute Zusammenarbeit zwischen Marketing und der Forschungs- und Entwicklung (F&E), insbesondere zur erfolgreichen Entwicklung neuer Produkte. Hier stehen Mitarbeiter des Marketings den Hausforderungen gegenüber, den gesamten Prozess von der Auswahl neuer Produktideen, über die Phasen der Erforschung, Entwicklung und Produktion bis hin zur Markteinführung begleitend zu unterstützen. In mehreren Studien wird aufgezeigt, wie mit Methoden des Textminings die Mitarbeiter des Marketings bei ihrer Entscheidungsfindung speziell bei ihrer Arbeit an der Marketing – F&E – Schnittstelle unterstützt werden können.

Eine wichtige Aufgabe im Rahmen der Entwicklung neuer Produkte ist die Auswahl neuer Produktideen, die als Startpunkt für dann folgende Forschungs- und Entwicklungsaktivitäten dienen. Aufgrund des bekannten „Technology push“ und „Market pull“ Effekts werden hierfür auf der einen Seite neue technologische Ideen z.B. aus Wissenschaft und Forschung benötigt, auf der anderen Seite aber auch Ideen, die Konsumentenbedürfnisse darstellen. Die in den Studien 3 bis 6 vorgestellte Methodik unterstützt Mitarbeiter des Marketings bei der Auffindung dieser Ideen.

Die dritte und vierte Studie befasst sich mit der automatischen Identifizierung neuer technologischer Ideen aus technologischen Texten z.B. aus Forschungspublikationen und Patente. Hierzu wird in der dritten Studie der aus der Technikphilosophie stammende abstrakte Begriff der „Technologischen Idee“ in eine konkrete Form überführt, die für eine Berechnung mit Textmining Methoden anwendbar ist. Es wird theoretisch nachgewiesen, dass durch Verwendung von Kollokationen (das gehäufte benachbarte Auftreten von Wörtern) ein automatisches Auffinden von technologischen Ideen möglich ist.

In der vierten Studie, wird auf Grundlage der theoretischen Untersuchungen der dritten Studie ein neues Ähnlichkeitsmaß definiert sowie darauf aufbauend eine Methodik vorgestellt, die das automatisierte Auffinden von technologischen Ideen durch eine web basierte Anwendung ermöglicht. Der Erfolg dieser Methodik ist nachgewiesen durch Vergleich mit bekannten heuristischen Ähnlichkeitsmaßen.

Die in der vierten Studie ermittelten Startideen für die Produktentwicklung sollten innovativ sein. Damit wird die Wahrscheinlichkeit erhöht, dass das spätere Produkt auch wirtschaftlich erfolgreich ist. Daher wird in der fünften Studie eine Methodik vorgestellt, die das Innovationspotential von technologischen Ideen schätzt. Hierzu wird der aus dem Bereich der Innovationsforschung stammende abstrakte Begriff der „Innovation“ in eine konkrete, für

Textmining anwendbare, Methodik überführt. Der Erfolg dieses Ansatzes wird durch Vergleich mit der „frequency baseline“ belegt.

Die sechste Studie befasst sich mit der automatischen Identifikation von Ideen, die Konsumentenbedürfnisse darstellen. Anregungen, Wünsche und Vorstellungen, die Konsumenten über Produkte in Internet Blogs veröffentlichen, werden hinsichtlich neuer Ideen analysiert. Hierzu wird die Methodik aus der vierten Studie modifiziert sowie durch Elemente aus dem Web Content Mining erweitert. Es wird nachgewiesen, dass trotz Verwendung der wenig genormten Umgangssprache, eine automatische Identifikation von Konsumentenideen erfolgreich ist.

Eine wichtige Aufgabe im Rahmen der Entwicklung neuer Produkte ist die Erstellung einer technologischen „Marktübersicht“, bei der technologische Einflüsse und Trends der eigenen Organisation denen der Konkurrenten gegenüber gestellt werden. Einen neuen Ansatz, um dies bei größeren Organisationen z.B. Regierungsorganisationen umzusetzen, ist in Studie 7 präsentiert. Die Methodik hinter diesem Ansatz verbindet die Cross-Impact Methode, um technologische Beziehungen aufzuzeigen, mit Methoden des Textminings, um Beschreibungen organisationseigener F&E-Projekte mit Patente der Konkurrenten vergleichen und zu Technologien zuordnen zu können. Im Ergebnis können die relativen technologischen Stärken und Schwächen der Organisation aufgezeigt werden und Mitarbeiter des Marketings bei der Erstellung der technologischen „Marktübersicht“ unterstützt werden.

4. Summary

Data mining is defined as the process of extracting patterns from data. Thus, it is used to transform this data into information [54]. A specific area of data mining is text mining or text data mining as the process of deriving high quality information from texts (unstructured data) [61]. Text mining structures the input text in a first step, it identifies new and unseen patterns within the structured textual data in a second step, and in a third step, it evaluates and interprets the results. Closely related to text mining is web content mining [56] as process of deriving high quality information from the content of web pages.

In total, this dissertation shows how methods from text mining and from web content mining can be used to support marketing professionals by improving marketing decision-making. The methodologies are presented in several studies where relevant textual information is analyzed that can be found in different sources (e.g. in web pages, documents, research papers, articles in technical periodicals, reports, etc.) [53].

This dissertation consists of seven studies. In two studies, this dissertation supports marketing professionals by identifying profitable customers and companies. This dissertation contains further five studies, which contribute to methods from text mining into the marketing - R&D interface of the new product development process. Table 1 gives an overview of the different studies.

	Title	Published
Study 1	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 (in review)
Study 2	Predicting E-Commerce Company Success by Mining the Text of Its Publicly-Accessible Website	The Journal of Applied Research and Technology (in review)
Study 3	Finding New Technological Ideas and Inventions with Text Mining and Technique	Schmidt-Thieme, L. (editor): Data Analysis, Machine Learning, and Applications, Proceedings of the 31st Annual Conference of

	Philosophy	the Gesellschaft für Klassifikation e.V., Springer, Berlin-Heidelberg-New York, pp. 413-420, 2008.
Study 4	Mining Ideas from Textual Information	Expert Systems with Applications 37 (10), 2010, 7182-7188.
Study 5	Mining Innovative Ideas to Support new Product Research and Development	Hermann Locarek-Junge, Claus Weihs (editors): Classification as a Tool for Research, Proceedings of the 11th IFCS Biennial Conference and 33rd Annual Conference of the Gesellschaft für Klassifikation e.V., Springer, Berlin-Heidelberg-New York, pp. 587-594, 2010.
Study 6	Extracting Consumers Needs for New Products - A Web Mining Approach	Proceedings of WKDD 2010, IEEE Computer Society, Los Alamitos, CA, pp. 440-443, 2010.
Study 7	A compared R&D-based and patent-based cross impact analysis for identifying relationships between technologies	Technological Forecasting and Social Change 77 (7), 2010, 1037-1050.

Table 1: Overview of the different studies.

Sect. 5 describes the research objectives of the studies as well as the methods from text mining and from web content mining that are used in this doctoral dissertation. Sect. 6 summarizes the main findings of the different studies, while Sect. 7 relates to the limitations and suggestions for further research.

5. Research Objectives

5.1. Overview

In the first two studies, a methodology is provided to support people from marketing department by the identification of profitable business-to-business customers and profitable e-commerce companies. The methodology is based on a combination of text classification (latent semantic indexing) [9] and web content mining [31, 52, 56, 63]. Textual information from customers' and companies' websites is used to support acquisition manager and marketing professionals by identifying profitable customers and companies.

Additionally, it is a well-known truth that good relationships between marketing departments and the R&D (research and development) are crucial for successful new product developments [23]. Here, marketing professionals are facing many challenges to support the complete process from selecting a new idea as starting point via processing of research, development, and production through to the introduction in market [47]. With text mining, one can support marketing professionals by their decisions concerning several of these challenges. Thus, this doctoral dissertation introduces the reader to the marketing - R&D interface [49] in the field of new product development [20] and focuses on how the marketing professional is able to improve marketing decision making by means of text mining approaches.

One important part of new product developments is the selecting of new and innovative ideas as starting point for a research or a development project [22,24]. Research projects have the aim to find new solutions for existing technological problems [43]. Therefore, new technological problem solution ideas [47] are needed. Alternatively, if new ideas are used as starting point for a development project then mainly, product ideas from the consumers are needed [45]. Three studies support people from R&D and from marketing department by the automatic identification of new product ideas from the consumers [32] as well as by the automatic identification of new technological ideas from the scientists and technologists [5, 57].

Consumers notices products in market in a direct way but they notices technologies only in an indirect way. This is because each product based on one (or probably several) technologies and each technology is the basis for several different products [2]. Additionally, life cycles of technologies are not correlated to life cycles of products [40]. However, in reality, consumers normally see no difference between products and the technologies behind the products. Therefore, they talk about product ideas if they mean technological ideas and vice versa [7]. Additionally, a technological idea from applied scientists and technologists sometimes also can be assigned to product ideas [47]. This means, product ideas and technological ideas based on a general rationale. This rationale is used to identify the innovative potential of the extracted ideas in a further study.

Identifying innovative product and technological ideas probably leads to new applied science research areas. A well-known task in marketing is to provide an overview and to generate useful insights of existing markets for a new product of a company [3]. However here, it is also necessary to provide an overview and to generate useful insights of the current technological landscape concerning these new research areas of a company [58]. This is helpful to support companies' research planning. A new approach for generating these insights specifically for large organizations is presented in a further study.

5.2. Study 1: Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing

This study investigates the issue of predicting new customers as profitable based on information about existing customers in a business-to-business environment. In particular, it is shown how latent semantic concepts [9] from textual information of existing customers' websites can be used to uncover characteristics of websites of companies that will turn into profitable customers. Hence, the use of predictive analytics will help to identify new potential acquisition targets. Additionally, it is shown that a regression model [4, 10, 19] based on these concepts is successful in the profitability prediction of new customers. The study contains a case study where the acquisition process of a mail order company is supported by creating a prioritized list of new customers generated by this approach. It is shown that the density of profitable customers in this list outperforms the density of profitable customers in traditional generated address lists (e.g. from list brokers).

From a managerial point of view, this approach supports the identification of new business customers and helps to estimate the future profitability of these customers in a company. Consequently, the customer acquisition process can be targeted more effectively and efficiently. This leads to a competitive advantage for B2b companies and improves the traditional acquisition process that is time- and cost-consuming with traditionally low conversion rates.

5.3. Study 2: Predicting E-Commerce Company Success by Mining the Text of Its Publicly-Accessible Website

The second study analyzes the impact of textual information from e-commerce companies' web sites on their commercial success. The textual information is extracted from web content of e-commerce companies divided into the top 100 worldwide most successful companies and into the top 101 to 500 worldwide most successful companies. It is shown that latent semantic concepts [9] extracted from the analysis of textual information can be adopted as success measures for a top 100 e-commerce company classification. This contributes to the existing literature concerning web site success measures for e-commerce [65]. As evaluation, a regression model [4, 10, 19] based on these concepts is build that is successful in predicting commercial success of the top 100 companies. These findings are valuable e.g. for e-commerce web sites creation.

5.4. Study 3: Finding New Technological Ideas and Inventions with Text Mining and Technique Philosophy

The main objective of study 3 is to introduce a rationale for finding textual patterns representing new technological ideas in unstructured technological texts. This rationale follows the statements of technique philosophy [48]. Therefore, a technological idea or invention represents not only a new mean, but a new purpose and mean combination [2]. By systematic identification of the purposes, means, and purpose-mean combinations in unstructured technological texts compared to specialized reference collections, a semiautomatic finding of ideas and inventions is realized. For this, collocations [37] are used as known method in corpus linguistics. Collocations are defined as a combination of terms, which co-occur more frequently than it would be expected by chance [30]. Additionally, characteristics (comprehensibility, novelty, and usefulness [28]) that are used to measure the quality of these patterns found in technological texts are examined.

5.5. Study 4: Mining Ideas from Textual Information

The purpose of study 4 is to use the rationale of study 3 and extend it to a text mining approach that extracts new and useful ideas from unstructured text automatically. Therefore, the study also uses the idea definition from technique philosophy [48] as described above and it focuses on ideas that can be used to solve technological problems [47]. This approach follows how persons create ideas [44, 60].

To realize the processing, methods from text mining and text classification (tokenization [14], term filtering methods [27, 36], Euclidean distance measure [62], alpha cut method [1] etc.) are used and are combined with a newly created heuristic measure for mining ideas. This “idea mining measure” is contributed to the scientific community.

As a result, this idea mining approach extracts automatically new and useful ideas from a user given text. The problem solution ideas are presented in a comprehensible way to the users [16, 21, 34]. This approach is evaluated with patent data and it is realized as web-based application, named "Technological Idea Miner" that can be used for further testing and evaluation.

5.6. Study 5: Mining Innovative Ideas to Support new Product Research and Development

For new product development, a new idea with innovative potential as starting point is needed [41]. Unfortunately, a high percentage of innovations fail, which means many selected ideas do not have the potential to become an innovation in future [6]. Additionally, the process from a new idea as starting point via research, development, and production activities through to an innovative product is very cost- and time-consuming [12]. Thus, the main objective of study 5 is to identify the innovative potential of new technological ideas to improve the performance of the innovation process.

The study 5 uses new technological ideas as inputs that are extracted from provided textual information by the application from study 4. It identifies innovative technology fields based on an approach from [46] by analyzing relationships among technologies [17]. All identified ideas are assigned to innovative technology fields by using text mining and text classification methods (tokenization [14], term filtering methods [27, 36], Jaccard's coefficient [15]).

For this, the application from study 4 is extended with these innovation-related aspects. With this new application, technological ideas in innovative fields are presented to the user as innovative ideas and might be used as starting point for new product research and development divisions.

5.7. Study 6: Extracting Consumers Needs for New Products - A Web Mining Approach

The purpose of study 6 is to introduce a web mining approach for automatically identifying new product ideas extracted from web logs. A web log - also known as blog - is a web site that provides commentary, news, and further information on a subject written by individual persons [64]. We can find a large amount of web logs for nearly each topic [25] where consumers present their needs for new products [51]. Finding these new product ideas is a well-known task in marketing [32, 35]. Therefore, with the automatically approach in study 6, we support marketing professionals by extracting new and useful product ideas from textual information in web logs.

The approach in study 6 is based on the approach in study 4. However, we consider peculiarities like the usage of text-mining in the colloquial language field [26, 38]. This approach is implemented by a web-based application named "Product Idea Web Log Miner" where marketing professionals provide descriptions of existing products. As a result, new product ideas are extracted from the web logs and presented to the users.

5.8. Study 7: A compared R&D-based and patent-based cross impact analysis for identifying relationships between technologies

The planning of technological research and development (R&D) is demanding in areas with many relationships between technologies [8]. To support decision makers of a government organization with R&D planning [39, 42] in these areas, a methodology to make the technology impact more transparent is introduced. The method shows current technology impact and impact trends from the R&D of an organization's competitors and compares these to the technology impact and impact trends from the organization's own R&D. This way, relative strength, relative weakness, plus parity of the organization's R&D activities in technology pairs can be identified.

A quantitative cross impact analysis (CIA) approach [13, 29, 50] is used to estimate the impact across technologies. The quantitative CIA approach contrasts to standard qualitative CIA approaches [18] that estimate technology impact by means of literature surveys and expert interviews. In this study, the impact is computed based on the R&D information regarding the respective organization on one hand, and based on patent data [11, 59] representative regarding R&D information of the organization's competitors on the other hand. As an illustration, the application field 'defense' is used, where many interrelations and interdependencies between defense-based technologies occur [33, 55]. Firstly, an R&D-based and patent-based Compared Cross Impact (CCI) among technologies is computed. Secondly, characteristics of the CCI are identified. Thirdly, the CCI data is presented as a network to show the overall structure and the complex relationships between the technologies. Finally, changes of the CCI are analyzed over time. The results show that the proposed methodology generates useful insights for government organizations to direct technology investments.

6. Main Findings

Based on the research objectives postulated, an overview of the main findings is given per study.

Study 1 contributes to previous research in multiple ways. Firstly, the main contribution of the proposed approach is to show the ability of latent semantic concepts from textual information of existing customers' websites to predict the profitability of new customers. Secondly, a new web structure/content mining approach is presented to extract relevant information from the websites of existing customers. Thirdly, a new combined (clustering / web mining) approach is contributed that shows how clustering of websites based on latent semantic concepts can be used to identify prevalent terms and how these terms can be used in a web content mining approach to identify addresses of new potential customers. Finally, it is shown that using these new addresses in an acquisition process pre-dominates the standard acquisition process e.g. by using addresses of list brokers. Overall, the crawling of new customers using the internet leads to a competitive advantage for B2b companies. Thus, the results contribute to the customer B2B acquisition literature and they testify to the ability of this website-based profitable-customer prediction approach to improve the acquisition process of companies while reducing costs.

Study 2 has analyzed the impact of textual information from e-commerce companies' web sites on their commercial success. It is shown that a regression model based on latent semantic concepts from this textual information is successful in predicting the most successful top 100 e-commerce companies. A case study shows that internet vendor trust, human computer interaction, and internet customer relation by rating and providing services are successful measures in predicting the top 100 e-commerce companies and that money-back policy, trusted order delivery, and internet customer relation by use of newsletters are successful measures in predicting the top 101 to top 500 e-commerce companies. This contributes to the existing literature concerning web site success measures for e-commerce and these findings are valuable for e-commerce web sites creation.

Study 3 shows that a semi-automatically text mining approach for identifying new ideas from provided textual information is generally feasible. This is proved by getting an acceptable precision and recall value using a semi-automatically approach (e.g. using a domain specific stop word list). The main finding here is to redefine an abstract term (an idea) in a concrete way that it can be used for computing with text mining methods. In detail, it is shown that a technological idea represents a combination of a purpose and a mean and that purposes and means are defined by collocations. Additionally, characteristics (comprehensibility, novelty and usefulness) for ideas are defined in a specific way that they can be computed by text mining methods.

Study 4 shows the success of an automatically approach for finding new ideas from provided textual information. For this, the study transforms creativity approaches from psychology and cognitive science to text mining approaches. Additionally, it is shown that problems and problem solution ideas can be represented as term vectors in vector space model. For this, the study contributes a new idea mining measure to the text classification measures. This measure identifies new ideas by comparing vectors that represent a problem to vectors that represent a problem solution idea. Last, approaches from comprehensibility research are adopted to this approach to present the new ideas in a comprehensible way to the user. As further main finding, it is demonstrated that this theoretically approach can be realized by a web-based application. The success of the idea mining measure is proved by comparing it to further heuristic measures (overlap-index, cosine-similarity and dice-similarity).

By comparing the evaluation results (precision and recall values) of study 3 to the evaluation results of study 4, we see that the use of a domain specific stop word list is far more successful than the use of a general stop word list. In the evaluations of the studies 4 - 7, we use a standard stop word list. This is because if an evaluation is successful by use of a standard stop word list then an evaluation is successful by use of a domain specific stop word list, too.

Study 5 shows a successful evaluated approach for identifying the innovative potential of new ideas. We redefine an abstract definition of innovation [46] to a concrete definition that can be computed by use of standard methods from text mining. The success of this approach is proved by comparing it to the frequency baseline.

Study 6 shows that the approach from study 4 (after modification) is successful by finding new product ideas from consumers where text mining methods operate on textual information from the colloquial language. A successful comparing to the frequency baseline proves the feasibility of this approach.

Study 7 contributes to previous research in multiple ways. The main contribution of the proposed approach is the new CCI index that identifies relative strength, relative weakness, plus parity of the organization's R&D activities in technology pairs. The second contribution is a method to determine the characteristics of relationships and to show whether two technologies are equally influencing one another (symmetry) or whether the impact of the first technology on the second is different from the impact of the second technology on the first (asymmetry). A third contribution is the presentation of a CCI network graph that shows the overall structure and the complex CCI relationships between several technologies. Finally, changes of the CCI are analyzed over time to discover trends regarding how the technology impact changes over time. They show which technology should receive more or less development and investment. Overall, the results testify to the ability of CCI to generate potentially useful insights for R&D decision makers of organizations.

7. Shortcomings & Further Research

Future work concerning study 1 should focus on improving the prediction by adding further unstructured information from existing customers (e.g. e-mails) to the prediction model.

Study 2 is evaluated concerning e-commerce companies. It can be enlarged by selecting further line of businesses e.g. automotive supplier companies.

The parameters of study 3 can be optimized concerning the precision and recall values of the evaluation.

Directions for future research of study 4 are given by the fact that nowadays there is a large amount of textual information available on the internet and this information probably contains many new technological ideas. Enlarging this approach to a web idea mining approach that automatically identifies problem solution ideas from the internet is an interesting topic for further research.

Further work of study 5 should aim at enlarging and optimizing this approach e.g. by identifying further properties of innovative ideas. A second avenue of further research could take the granularity of the context information into account e.g. by using technologies rather than scientific categories. This also probably leads to an increasing precision and recall.

Study 6 can be enlarged by optimizing the parameters concerning the precision and recall values of the evaluation.

Future research for study 7 should aim at assigning R&D projects to technologies rather than technology areas. In the case study, internal R&D projects and patent data should be assigned to the 200 defense-based technologies from EDA taxonomy. Then, a more detailed view on the technological landscape in the ’defense’ application field could be provided. A second avenue of further research could take the occurrence of new technologies into account. This research focuses on computing the impacts between technologies or technology areas. It does not consider the computation of the occurrence probability of new technologies or technology areas. This could be an interesting topic for future research.

8. Literature

[1] Abebe, A. J., Guinot, V., Solomatine, D. P. (2000). Fuzzy alpha-cut vs. Monte Carlo techniques in assessing uncertainty in model parameters. In: Proc. 4-th International Conference on Hydroinformatics, Iowa City.

[2] Albers, S., Gassmann, O. (2005). Handbuch Technologie- und Innovationsmanagement: Strategie- Umsetzung- Controlling. p. 196, Gabler Verlag, Wiesbaden.

[3] Allgeier, H., Albrecht, K. (2003). Campus Management. p. 394, Campus Verlag, Frankfurt.

[4] Allison, P.D. (1999). Logistic Regression using the SAS System: Theory and Application. SAS Institute Inc., Cary, NC.

[5] Annacchino, M.A. (2003). New Product Development: From Initial Idea to Product Management. p. 88 f., Butterworth-Heinemann, Oxford.

[6] Berth R. (1997). Der große Innovations-Test: das Arbeitsbuch für Entscheider: Chancen erkennen, Flops vermeiden - Theorie und Praxis des Management of Change. Econ, Düsseldorf.

[7] Bürgel, H.D., Haller, C., Binder, M. (1996). F&E-Management. p. 85., Vahlen, München.

[8] 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. Technol. Forecast. Soc. Change 74, 1296-1314.

[9] Coussement, K., Van den Poel, D. (2008). Information & Management 45, 164-174.

[10] DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44 (3), 837-845.

[11] Dernis, H., Guellec, D. (2002). Using patent counts for cross-country comparisons of technology output. Special Issue on New Science and Technology Indicators, STI Review 27, 129-146.

[12] Disselkamp, M. (2005). Innovationsmanagement: Instrumente und Methoden zur Umsetzung im Unternehmen. p. 179. Gabler Verlag, Wiesbaden.

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

[14] Feldman, R., Sanger, J. (2007). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. p. 318, Cambridge University Press, Cambridge.

[15] Ferber, R. (2003). Information Retrieval. p. 78, dpunkt.verlag, Heidelberg.

[16] Flesch, R. (1978). A new readability yardstick. Journal of Applied Psychology 32, 221-233.

[17] Geschka, H., Schauffele, J., Zimmer, C. (2005). Explorative Technologie-Roadmaps - Eine Methodik zur Erkundung technologischer Entwicklugslinien und Potenziale. In: Möhrle, M.G., Isenmann, R. (Eds.). Technologie-Roadmapping. p. 165, Springer, Berlin, Heidelberg.

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

[19] Greiff, W.R. (1998). A theory of term weighting based on exploratory data analysis. In: Croft, W.B., Moffat, A., van Rijsbergen, C.J., Wilkinson, R., Zobel, J. (Eds.). Proceedings of the 21st SIGIR Conference, pp. 11-19, ACM, New York.

[20] Grimm, S. (2004). Marketing für High-Tech-Unternehmen: Wie sie Markt- und Technologiezyklen strategisch nutzen und beeinflussen. p. 83, Gabler Verlag, Wiesbaden.

[21] Groeben, N. (1982). Leserpsychologie: Textverständnis - Textverständlichkeit. Aschendorff, Münster.

[22] Gupta, A.K., Raj, S.P., Wilemon, D. (1986). A model for studying R&D-marketing interface in the product innovation process. Journal of Marketing 50, 7-17.

[23] Gupta A.K., Raj S.P., Wilemon D. (1985). The R&D-Marketing Interface in High-Technology Firms. Journal of Product Innovation Management 2/1 (13), 12-24.

[24] Gupta, A.K., Raj, S.P., Wilemon, D. (1987). Managing the marketing-R&D interface. Research Management March/April, 38-43.

[25] Herring, S.C., Scheidt, L.A., Bonus, S., Wright, E. (2004). Bridging the gap: a genre analysis of Weblogs. In: Proceedings of the 37th Annual Hawaii International Conference on System Sciences, Hawaii.

[26] Hoffmann, L., Kalverkämper, H., Wiegand, H.E. (1998). Fachsprachen - Languages for Special purposes: Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft - an international Handbook of Special-language and Terminology Research. p. 1602, Walter de Gruyter, Berlin.

[27] Hotho, A., Nürnberger, A., Paaß, G. (2005). A Brief Survey of Text Mining. LDV Forum 20(1), 19-26.

[28] Hotho, A. (2004). Clustern mit Hintergrundwissen. Univ. Diss., p. 29, Karlsruhe.

[29] Jeong, G.H., Kim, S.H. (1997). Aqualitative Cross-Impact Approach to find the Key Technology. Technol. Forecast. Soc. Change 55 (3), 203-214.

[30] Kamphusmann, T. (2002). Text-Mining. p. 28, Symposion Publishing, Düsseldorf.

[31] Kosala, R., Blockeel, H. (2000). Web mining research: A survey. SIGKDD Explorations 2 (1).

[32] Kuß, A. (2006). Marketing-Einführung: Grundlagen - Überblick - Beispiele. p. 189, Springer, Berlin.

[33] Lange, H.J, Ohly, H.P., Reichertz, J. (2008). Auf der Suche nach neuer Sicherheit: Fakten, Theorien und Folgen. p. 11, VS Verlag, Wiesbaden.

[34] Langer, I., Schulz v. Thun, F., Tausch, R. (1974). Verständlichkeit in Schule und Verwaltung. Ernst Reinhardt, München.

[35] Lawton, L., Parasuraman, A. (1980). The impact of the marketing concept on new product planning. Journal of Marketing 44 (19).

[36] Lustig, G. (1986). Automatische Indexierung zwischen Forschung und Anwendung. p. 92, Georg Olms Verlag, Hildesheim.

[37] Manning, C. D., Schütze, H. (1999). Foundations of Statistical Natural Language Processing. p. 35, The MIT Press, Cambridge.

[38] Martin-Bautista, M.J., Sanches, D., Serrano, J.M., Vila M.A. (2004). Text Mining using Fuzzy Association Rules. In: Loia, V., Nikravesh, M., Zadeh, L.A. (Eds.). Fuzzy Logic and the Internet. p. 173, Springer, Berlin, Heidelberg.

[39] Mogee, M.E., Kolar, R.G. (1994). International patent analysis as a tool for corporate technology analysis and planning. Technol. Anal. Strat. Manag. 6 (4), 485-503.

[40] Morris, M.H., Pitt, L.F., Honeycutt, E.D. (2001). Business-to-business Marketing: A Strategic Approach. p. 216 f., Sage Pubn Inc, London.

[41] Möslein, K.M., Matthaei, E.E. (2008). Strategies for Innovators: A Case Book of the HHL Open School Initiative. p. 13, Gabler Verlag, Wiesbaden.

[42] Narin, F., Noma, E. (1987). Patents as indicators of corporate technological strength. Res. Policy 16 (2/4), 143-155.

[43] Nolden, R.G., Körner, P., Bizer, E. (2004). Industriebetriebslehre: Management betrieblicher Prozesse. p. 142, Bildungsverlag Eins, Troisdorf.

[44] Osborn, A.-F. (1948). Your Creative Power. C. Scribner's sons, New York.

[45] O'Shaughnessy, J. (1995). Competitive Marketing: A Strategic Approach. p. 360, Routledge, New York.

[46] Reiß, T. (2006). Innovationssysteme im Wandel - Herausforderungen für die Innovationspolitik. In: Müller, B., Glutsch, U. (Eds.). Fraunhofer-Institut für System- und Innovationsforschung - Jahresbericht 2006, p. 10. Karlsruhe.

[47] Ripke, M., Stöber, G. (1972). Probleme und Methoden der Identifizierung potentieller Objekte der Forschungsförderung. In: Paschen, H., and Krauch, H. (Eds.). Methoden und Probleme der Forschungs- und Entwicklungsplanung. p. 47, Oldenbourg, München.

[48] Rohpohl, G. (1996). Das Ende der Natur. In: Schäfer, L., Sträker, E. (Eds.). Naturauffassungen in Philosophie, Wissenschaft und Technik. Bd. 4, pp. 143-63, Alber, Freiburg, München.

[49] Schneider, D.J.G. (2002). Einführung in das Technologie-Marketing. Oldenbourg, München.

[50] 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 Trans. Eng. Manage. 38 (3), 224-236.

[51] Soll, J.H., Strauch, S. (2006). Ideengenerierung mit Konsumenten im Internet. Springer, Berlin, Heidelberg.

[52] Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N. (2000). Web usage mining: Discovery and application of usage patterns from Web data. SIGKDD, Explorations 1 (2), 12-23.

[53] Thorleuchter, D. (2008). Finding Technological Ideas and Inventions with Text Mining and Technique Philosophy. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker R. (Eds.). Data Analysis, Machine Learning, and Applications. pp. 413-420. Springer, Berlin, Heidelberg.

[54] Thorleuchter, D., Gericke, W., Weck, G., Reiländer, F., Loß, D. (2008). Vertrauliche Verarbeitung staatlich eingestufter Information - die Informationstechnologie im Geheimschutz. Informatik Spektrum 32 (2), 102-109.

[55] Thorleuchter, D., Van den Poel, D., Prinzie, A. (2010). Mining Innovative Ideas to Support new Product Research and Development. In: Locarek-Junge, H., Weihs, C. (Eds.). Classification as a Tool for Research. pp. 587-594, Springer, Berlin, Heidelberg.

[56] Thorleuchter, D., Van den Poel, D., Prinzie, A. (2010). Extracting Consumers' Needs to support New Product Development. Proceedings of WKDD 2010, pp. 440-443, IEEE Computer Society, Los Alamitos, CA.

[57] Thorleuchter, D., Van den Poel, D., Prinzie, A. (2010). Mining Ideas from Textual Information. Expert Systems with Applications 37 (10), 7182-7188.

[58] 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.

[59] Trajtenberg, M. (2002). A penny for your quotes: patent citations and the value of innovations. In: Jaffe, A., Trajtenberg, M. (Eds.). Patents, Citations and Innovations. MIT Press, Cambridge, MA.

[60] Wallas, G. (1926). The Art of Thought. Harcourt Brace, New York.

[61] Wang, J. (2008). Encyclopedia of Data Warehousing and Mining. p. 28, Idea Group Inc., Calgary.

[62] Wrobel, S., Morik, K., Joachims, T. (2003). Maschinelles Lernen und Data Mining. In: Görz, G., Rollinger, C.R., Schneeberger, J. (Eds.). Handbuch der Künstlichen Intelligenz, p. 537, 4. Auflage, Oldenbourg, München.

[63] Zaiane, O.R. (1998). From resource discovery to knowledge discovery on the internet. Technical Report TR 1998-13, Simon Fraser University, Burnaby.

[64] Zsunyi, C. (2007). Weblogs/ Blogs: Stand der Technik und Zukunftspotentiale. p. 4 f., GRIN Verlag, München.

[65] Zvirana, M., Glezerb C., Avnia, I. (2006). User satisfaction from commercial web sites: The effect of design and use. Information & Management 43 (2), 157-178.

Chapter I

Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing

This chapter is based on Thorleuchter, D., Van den Poel, D., Prinzie, A. Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing. Expert Systems with Applications (in review).

Table of Contents

Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing

Dirk Thorleuchter[a], Dirk Van den Poel[b], and Anita Prinzie[b]

[a] Fraunhofer INT, D-53879 Euskirchen, Appelsgarten 2, Germany, & PhD Candidate, Ghent University, Belgium, dirk.thorleuchter@int.fraunhofer.de
[b] Ghent University, Faculty of Economics and Business Administration, B-9000 Gent, Tweekerkenstraat 2, Belgium, dirk.vandenpoel@ugent.be, anita.prinzie@ugent.be

Abstract

We investigate the issue of predicting new customers as profitable based on information about existing customers in a business-to-business environment. In particular, we show how latent semantic concepts from textual information of existing customers' websites can be used to uncover characteristics of websites of companies that will turn into profitable customers. Hence, the use of predictive analytics will help to identify new potential acquisition targets. Additionally, we show that a regression model based on these concepts is successful in the profitability prediction of new customers. In a case study, the acquisition process of a mail-order company is supported by creating a prioritized list of new customers generated by this approach. It is shown that the density of profitable customers in this list outperforms the density of profitable customers in traditional generated address lists (e. g. from list brokers).

From a managerial point of view, this approach supports the identification of new business customers and helps to estimate the future profitability of these customers in a company. Consequently, the customer acquisition process can be targeted more effectively and efficiently. This leads to a competitive advantage for B2B companies and improves the acquisition process that is time- and cost-consuming with traditionally low conversion rates.

Keywords

B-to-B marketing, Text Mining, Web Mining, Acquisition, SVD

1 Introduction

While products and services are sold by companies with little knowledge or strategy concerning the customers who bought the products in the past a change from this product-centered to a customer-centered environment can be seen today (Coussement & Van den Poel, 2008). This is because for companies it is important to capture and enhance market share while reducing costs. Therefore, they must reconsider the business relationships with their existing customers (Pan & Lee, 2003).

One important aspect is to improve the acquisition of new customers. Normally, this is time- and cost-consuming because it is easier to keep and satisfy existing customers than to attract new ones with a high attrition rate (Reinartz & Kumar, 2003). Therefore, new customers have to be identified who are interested in companies' products and services. This probably can be done in several different ways (e.g. by presenting products and services on fairs, buying addresses from list brokers etc.). However, only a small percentage of potential customers become profitable customers in the future.

In this paper, we propose a new approach to identify new business customers and to predict them as profitable using information of existing customers. For this, existing customers' information is collected from customer relationship management (CRM) systems where customers probably can be divided into different classes e.g. concerning their sales volume. Then existing customers can be classified as profitable customers, if their sales volume over a specific period of time is greater than a specific threshold (Menon, Homburg, & Beutin, 2005).

If we specifically consider on existing customers in a business-to-business environment then we also find information in the CRM system about customers' companies. Nowadays, companies normally present information on internet websites because of the rapid development of IT and the Internet. Many firms rely on Internet websites to provide product information for their customers. Information on existing customers' websites can be crawled and analyzed by use of web structure and content mining approaches. However, the received information consists of masses of unstructured textual information (Coussement & Van den Poel, 2009) and decision makers normally do not use it for acquisition purposes. This is because the information is not directly usable in a traditional acquisition context and there is often a lack of in-house knowledge on how to analyze this unstructured information for acquisition purposes. Additionally, ready-to-use frameworks are also not available to integrate this information in the acquisition process.

In this approach, textual information of existing customers' websites is analyzed by latent semantic indexing (LSI) to identify specific textual features (concepts). An expectation-maximization (EM) algorithm is used to cluster customers' websites based on the concepts to select prevalent terms from those concepts that mainly occur on the websites of profitable business customers and that seldomly occur on the websites of non-profitable customers. Then, these terms are used as query for web content mining to create a list of further companies with similar concepts on their website. A logistic regression model is built based on the concepts of existing customers' websites to predict the profitability of the new customers from the created list. Comparing this list of potential customers to the traditional acquisition process – e.g. where list brokers' lists of potential customers are used that are expensive - shows that this new approach improves the identification and prediction of new profitable business customers while reducing costs.

This paper contributes to previous research in multiple ways. Firstly, the main contribution of the proposed approach is to show the ability of latent semantic concepts from textual information of existing customers' websites to predict the profitability of new customers (see Sect. 3.5). Secondly, a new web structure/content mining approach is presented to extract relevant information from the websites of existing customers (see Sect. 3.1). Thirdly, a new combined (clustering / web mining) approach is contributed that shows how clustering of websites based on latent semantic concepts can be used to identify prevalent terms and how these terms can be used in a web content mining approach to identify addresses of new potential customers (see Sect. 3.4). Finally, it is shown that using these new addresses in an acquisition process pre-dominates the standard acquisition process e.g. by using addresses of list brokers. Overall, the crawling of new customers using the internet leads to a competitive advantage for B2B companies. Thus, the results contribute to the customer B2B acquisition literature and they testify to the ability of this website-based profitable-customer prediction approach to improve the acquisition process of companies while reducing costs.

2 Related Work

In marketing, we distinguish between transactional and relational approaches. Transactional marketing (Coviello, Brodie, & Munro, 1997) can be defined as an impersonal approach with focus on single point of sale transactions. It describes a company-centric model with an active company and its passive customers, a homogeneous marketplace, and mainly unidirectional information flow from the company to the marketplace / to its customers. In the other direction, little feedback from company's customers to the company can be seen.

In contrast to this, relational marketing focuses on customer retention and satisfaction, rather than single point-of-sale transactions (Kim, 2006; Neslin et al., 2006). Based on relational marketing, information exchange is the main principle in the acquisition of business customers. Its fundamental effect on market growth and structure as well as on new customer acquisition is shown (Naude & Holland, 1996; Verhoef et al., 2010). Further work examines the impact of e-commerce as a new information exchange technology on the acquisition of new business customers (Archer & Yuan, 2000; Baecke & Van den Poel, 2010a; Baecke & Van den Poel, 2010b; De Bock & Van den Poel, 2009; Van den Poel & Buckinx, 2005). Moreover, the impact of word-of-mouth referrals as a traditional information exchange approach is shown on the acquisition of new business customers (Wangenheim & Bayon, 2007).

Related work in the field of web mining focuses on the identification of customer's behaviors in the internet (Bose & Mahapatra, 2001; Bucklin & Gupta, 1992; Lee & Chung, 2003; Park & Chang, 2009) and in the identification of collaborative partners (Engler & Kusiak, 2010).

In contrast to previous work on customer acquisition in a relational B2B context and on web mining, this approach examines latent semantic indexing and web mining as text mining / information retrieval technology for improving the information flow from a company's customers to the company. As a result, the impact of web mining on the acquisition of new business customers can be shown as contribution to the customer B2B acquisition literature.

3 Methodology

Textual information from customer's websites is collected and is transformed in a pre-processing phase to a term-website matrix. After dimension reduction, latent semantic concepts are identified and clustered. Class labels mainly representing profitable customers' websites are used to identify websites of new potential customers. Textual information from these websites is projected into the dimension-reduced latent semantic concept space. A prediction model is built on this concept-space matrix to show that latent semantic concepts from existing customers' websites can be used to predict the profitability of new customers. Fig. 1 shows the methodology of this approach.

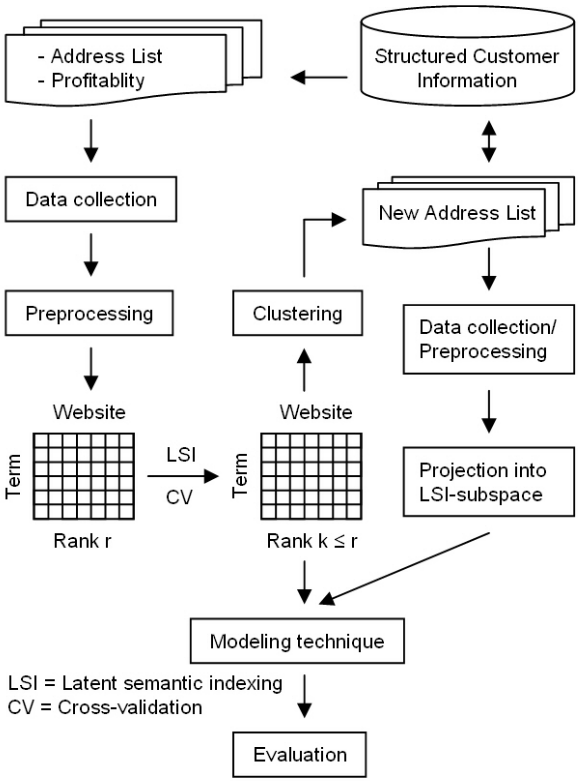


Figure 1: Methodology of the approach

3.1 Data collection

For the data collection phase, structured customer information is needed to collect unstructured content information from customers' websites. Structured customer information can be extracted from CRM systems of a company in which sales volume as well as e-mail addresses or website addresses for each customer is stored. Fig. 2 shows different steps in the data collection phase.

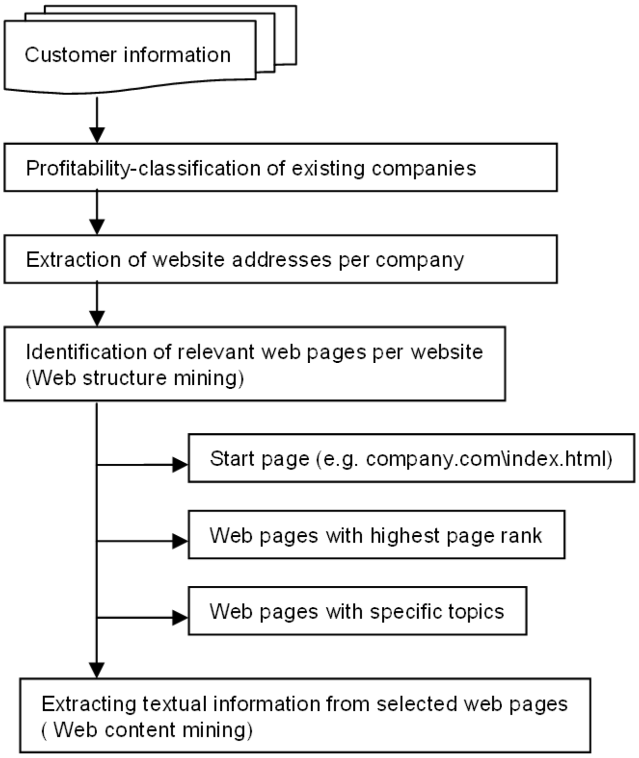

Customer information
Profitability-classification of existing companies
Extraction of website addresses per company
Identification of relevant web pages per website
(Web structure mining)
Start page (e.g. company.com\index.html)
Web pages with highest page rank
Web pages with specific topics
Extracting textual information from selected web pages
(Web content mining)

Fig. 2: Different steps in data collection phase.

Information about existing customers' sales volume is used to classify companies as profitable. An aggregation of the sales volume that belongs to the same company is done because several customers probably belong to the same company. Then, companies are assigned to a sales volume and they are defined as profitable if their volume exceeds a specific threshold.

To identify customers' company websites, the website addresses are used or if unavailable, e-mail addresses are converted to website addresses. In general, an e-mail address from a business customer based on the corresponding company website e.g. miller@company-name.com. Therefore, it is often possible to identify the corresponding company website for each customer. If a customer's e-mail address is based on an ISP (internet-service provider) or e-mail provider (e.g. hotmail.com) then his company's website is identified manually, otherwise information about this customer is discarded for further processing.

A website consists of several web pages. To extract information from customers' websites, relevant web pages have to be identified first. This avoids crawling trivial web pages e.g. 'disclaimer', 'privacy / data protection policy', 'sitemap', 'about', 'contact formulary' etc. In general, the starting page of an internet website is relevant. To identify further relevant web pages, the corpus of an internet search engine is used by access to web services. A web service is a software system that is designed to support interoperable machine-to-machine interaction over a network. Frequently, web services are just web-based advanced programming interfaces (APIs). Access to these interfaces is possible over the internet. Then, the requested service is executed, resulting data is ordered by page rank, and it is transferred back to an application that requested the service (Thorleuchter, Van den Poel, & Prinzie, 2010c). A lot of internet search engines offer web services. In this approach, Google is used as well-known internet search engine because its page rank is of high quality and we suppose that Google indexes most commercial websites. For each website, we build a query that is restricted to web pages of the corresponding website and that is additionally restricted on a specific language in a first step. This language restriction is necessary for comparing terms from different web pages and different websites. The query results consist of all indexed web pages ordered by the page rank. For further processing, the starting web page and three further web pages with the highest page rank are selected. Additionally, we suppose that information about a company's history might be relevant for identifying the company as profitable customer. To identify web pages containing this information, we build a query that is restricted to web pages of the corresponding website and in addition contain of specific search terms in a second step. Examples for these search terms are 'founded, history' etc. that have to be translated to the selected language. The resulting web page with the highest page rank is selected for further processing if not already selected in the first step.

With web content mining, information from the selected customers' web pages is extracted. In contrast to the structured information about sales volume for each customer, extracting textual information from customers' websites is highly unstructured. Thus, text pre-processing is necessary to capture the relevant details from this information for integration in the acquisition process.

3.2 Pre-processing

The extracted content information has to be converted into a structured representation as term vectors in a vector-space model. Thus, each web page is represented as a vector of weighted frequencies of designated words (Thorleuchter, Van den Poel, & Prinzie, 2010b). The size of the vector is defined by the number of distinct terms in the dictionary. The importance of a term - with respect to the semantics - is reflected by each corresponding vector component. A vector component is set to its weight if the corresponding term is used in the web page and to zero if the term is not. A collection of these term vectors is used to build a term-by-web page matrix firstly and a term-by-website matrix secondly. The process of converting web pages to a term-by-website matrix is depicted in Fig. 3.

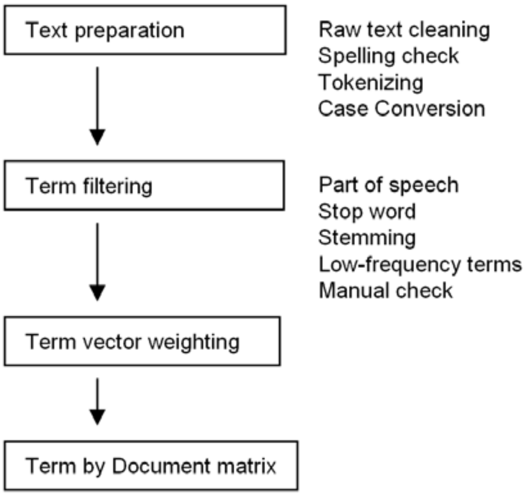


Fig. 3: Different steps in the pre-processing phase.

3.2.1 Text preparation

In the text preparation phase, raw text cleaning is done. For this, images, html-, xml-tags as well as scripting code (e.g. javascript) from the web pages are removed. Additionally, specific characters and punctuation are deleted and typographical errors are corrected by use of a dictionary. With tokenization, all words that are used in the web page can be identified which means texts are separated in terms whereby the term unit is a word. All terms are converted to lower case whereby the first sign is capitalized (case conversion).

3.2.2 Term filtering

The set of different terms in a web page can be reduced by using filtering methods (Thorleuchter, Van den Poel, & Prinzie, 2010a). For further processing, informative terms are selected that belong to a specific syntactic category (nouns, verbs, adjectives and adverbs) by use of part-of-speech tagging. Other (non-informative) terms are discarded. Stop word filtering is the standard filtering method in text mining applications. It is used to remove words that bear little or no content information, like articles, conjunctions, prepositions, etc (Thorleuchter, Van den Poel, & Prinzie, 2010d). Further filtering methods are lemmatization and stemming. A stemmer transforms words to their basic forms named stem by stripping the plural 's' from nouns, the 'ing' from verbs etc. Related words map to the same stem. Stemming is closely related to lemmatization. The difference is that lemmatization uses knowledge of the context to discriminate between words that have different meanings depending on part of speech. Unfortunately, lemmatization is time consuming and still error prone (Thorleuchter, 2008). Therefore, a dictionary-based stemmer is used combined with a set of production rules to give each term a correct stem. The production rules are used when a term is unrecognizable in the dictionary. The term frequencies in textual information follow a Zipf distribution (Zipf, 1949). Half of them appear only once or twice. Thus, those rare terms under these thresholds are deleted that often yield great savings. The last step in term filtering is to check the selected terms manually (Gericke et al., 2009).

3.2.3 Term vector weighting

The selected terms are used to construct a vector of weighted frequencies for each web page. In contrast to term vectors where the component values are raw frequencies of appearance for a term in a web page, the use of term weighting schemes leads to significant improvements in retrieval performance (Sparck Jones, 1972). The weights reflect the importance of a term in a specific web page of the considered web page collection. Large weights are assigned to terms that are used frequently in selected web pages but rarely in the whole web page collection (Salton & Buckley, 1988). Thus a weight $w_{i,j}$ for a term i in web page j is computed by term frequency $tf_{i,j}$ times inverse web page frequency idf_i, which describes the term specificity within the web page collection. In (Salton, Allan, & Buckley, 1994) a weighting scheme was proposed that has meanwhile proven its usability in practice. Besides term frequency – defined as the absolute frequency of term i in web page j - and inverse document frequency - defined as $idf_i := \log(n/ df_i)$ -, a length normalization factor is used to ensure that all documents have equal chances of being retrieved independent of their lengths:

$$w_{i,j} = \frac{tf_{i,j} \cdot \log(n / df_i)}{\sqrt{\sum_{p=1}^{m} tf_{i,j_p}{}^2 \cdot (\log(n / df_{i_p}))^2}} \qquad (1)$$

where n is the size of the web page collection D where web pages are represented by term vectors in m-dimensional space, and df_i is the number of web pages in D that contain term i (Chen, Chiu, & Chang, 2005).

3.2.4 Term vector aggregation

As a result, a high-dimensional, weighted term-by-web page matrix is created. However, from the managerial point of view, a prediction is done per customer's company website.

Thus, an aggregation of the web pages that belong to the same customer's company website is needed. The aggregated weight of term i for all web pages belonging to a customer's company website j (Coussement & Van den Poel, 2009) is

$$Aw_{i,j} = \sum_{k=1}^{r} w_{i,k} \qquad (2)$$

with $w_{i,k}$ equal to the weight of term i in web page k and r equal to the number of web pages belonging to the same customer's company website.

3.3 Concept identification with LSI and singular value decomposition

Using each distinct term as a feature would lead to an unmanageable high dimensionality of the feature space. Additionally, most weights are zero for a customer's company. To reduce the dimension of the feature space LSI is used. LSI groups together related terms (Deerwester et al., 1990) and together with singular value decomposition (SVD) it forms semantic generalizations due to the fact that relationships between terms are recognized by the appearance of terms in similar documents (e.g. web pages). SVD transforms web pages from the high-dimensional feature space to an orthonormal, semantic, latent subspace. Similar terms (keywords) are grouped into concepts. Each concept has a high discriminatory power to other concepts in the reduced feature space.

3.3.1 Feature space dimension reduction

The SVD of a term-by-website (m x n) matrix A with rank r (r ≤ min(m,n)) is a transformation into a product of three matrices in form of

$$A = U \Sigma V^t \qquad (3)$$

with Σ equal to a diagonal (r x r) matrix containing positive singular values of matrix A where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r$, U equal to the term-concept similarity (m x r) matrix, and V equal to the concept-company similarity (n x r) matrix. The columns of U and V are orthonormal in the Euclidean sense (Yen & Lina, 2010). The weights of the matrix A depended on the latent concepts by

$$w_{i,j} = \sum_{x=1}^{r} U_{i,x} \cdot \Sigma_x \cdot d_{j,x} \qquad (4)$$

If k ≤ r and the singular values $\lambda_{k+1}, \ldots, \lambda_r$ are small compared to $\lambda_1, \ldots, \lambda_k$ then LSI based on SVD allows a good approximation of A_r with rank r by A_k with rank k. Therefore, LSI dropped the smaller lambda values in Σ by retaining only the first predetermined singular values equal to or greater than k while only the first k columns of U and V were retained.

$$A_k = U_k \, \Sigma_k \, V_k^t \tag{5}$$

with U_k, Σ_k and V_k were equal to the k-rank approximation of U, Σ and V, respectively.

The approximated k-rank concept-website similarity matrix V_k contained information on how well a certain website loads on the different k concepts, which reflect the hidden (latent semantic) patterns in the textual information.

3.3.2 Concept dimension selection

The choice of k – the number of concepts - is critical for optimal predictive performance by using SVD. If k is too large then too many irrelevant or unimportant concepts are used for prediction. Otherwise, if k is too small then relevant concepts probably are not considered. The calculation of an optimal number of concepts k can be done using an operational criterion, i.e. a value of k that yields good performance (Chen et al., 2010). In this paper, we use a parameter-selection procedure by constructing several rank-k models, by using a fivefold cross-validation on the training set for each rank-k model, and by selecting the most favorable rank-k model (based on the cross-validated performance) for further analysis.
The cross-validation performance is determined by the results of the prediction model (see Sect. 3.5).

3.3.3 Projection of test examples into the LSI-subspace

The meaning of the concepts during testing should stay the same as during training. Consequently, test examples are transformed to term vectors by using the different steps in the pre-processing phase (see Sect. 3.2). Additionally, the projection of the test examples is done into the same semantic latent subspace as created during training (Zhong & Li, 2010). As a result, the term vector A_d is created for each test example and the new concept-website vector can be calculated by

$$V_d = A'_d \cdot U_k \cdot \Sigma_k^{-1} \tag{6}$$

with U_k the k-rank concept-term similarity matrix and Σ_k the diagonal singular value matrix in rank k, both of the original SVD. The new concept-website vector V_d is comparable to the concept-website vectors of the matrix V_k.

3.4 Website clustering

Concepts of the dimension reduced new concept-website matrix reflect the hidden, latent semantic patterns in the textual information from companies' websites that have a high discriminatory power to other concepts. For applying these concepts for acquisition (e.g. to create a list of new potential customers), we have to identify prevalent terms mainly representing concepts from profitable companies' websites and least of all from non-profitable companies' websites. Websites are clustered to identify these terms. An expectation-maximization (EM) algorithm is used for finding maximum likelihood estimates of parameters in a probabilistic model, where the model depends on the dimension-reduced SVD concepts.

As a result, classes contain profitable companies' websites as well as non-profitable company's websites. Class labels represent a number of prevalent terms from the websites assigned to a class. Labels are selected from classes that are mainly assigned to profitable customers' websites. Terms from these labels are used as search query in a web mining approach. Thus, companies can be identified where the selected terms occur on their websites and where the company itself does not occur in the training examples. It can be supposed that their websites contain similar concepts as concepts from the corresponding class and therefore, they probably are profitable customers, too. To evaluate their profitability, textual information from their websites is collected (see Sect. 3.1), pre-processed (see Sect. 3.2), projected into the LSI-subspace (see Sect. 3.3.3), and used as test examples in a prediction modeling approach (see Sect. 3.5).

3.5 Prediction Modeling

As modeling technique, logistic regression is used by producing a maximum likelihood function and by maximizing it in order to become an appropriate fit to the data (Allison, 1999; Inagaki, 2010). Logistic regression is conceptually simple (DeLong, DeLong, & Clarke-Pearson, 1988), a closed-form solution for the posterior probabilities is available and it provides quick and robust results in a prediction context (Greiff, 1998). Therefore with a training set of $T = \{(x_i, y_i)\}$ and $i = \{1,2,\ldots,N\}$ and input data $x_i \in R^n$ and corresponding binary target labels $y_i \in \{0,1\}$ (non-profitable, profitable), logistic regression is used to estimate the probability $P(y = 1 \mid x)$ given by

$$P(y = 1 \mid x) = \frac{1}{1 + exp(-(w_0 + wx))} \tag{7}$$

with $x \in R^n$ an n-dimensional input vector (a concept-website vector) as representative for companies' websites load on the concepts, w the parameter vector and w_0 the intercept.

3.6 Evaluation criteria

This evaluation focuses on examining the performance of the prediction model to show that latent semantic concepts from existing customers' websites can be used to predict the profitability of new customers and to show that newly created address lists of potential customers contain more profitable customer addresses than lists from list brokers. This is done with the commonly used criteria: lift, precision, recall, area under the receiver operating characteristics curve (AUC), sensitivity, and specificity.

To evaluate the performance of classification models, lift is the most commonly used performance measure for business applications. It measures the increase in density of the number of profitable new customers relative to the density of new customers in total. For an acquisition process, it is interesting to increase the density of profitable customers, especially in the top 30 percentile of a potential customer list because a new customer acquisition is time- and cost-consuming and budgets / personnel resources for acquisition are often limited. Thus, acquisition managers often focus on a subset of new customers. Practically, all new customers are sorted from most profitable to least profitable by the model. Afterwards, the density of profitable customers from the top 30 percentile can be computed.

Based on the number of positives that are correctly identified (TP), the number of negatives that are classified as positives (FP), the number of positive cases that are identified as negatives (FN), and the number of negative cases that are classified as negatives (TN), we use the sensitivity (TP/(TP+FN)) as the proportion of positive cases that are predicted to be positive, the specificity (TN/(TN + FP)) as the proportion of negative cases that are predicted to be negative, the precision (TP/(TP+FP)) as a measure of exactness or fidelity, and the recall (TP/(TP+TN)) as a measure of completeness. These vary when the threshold value is varied. The receiver operating characteristic curve (ROC) is a two dimensional plot of sensitivity versus (1-specificity). In order to compare the performance of two or more classification models, the AUC is calculated. This measure is used to evaluate the performance of a binary classification system (Hanley & McNeil, 1982). The optimal reduced number of concepts is obtained by optimizing the performance of the predictive model as reflected by a cross-validated AUC.

4 Empirical verification

4.1 Research data

In this study, we forecast the profitability of new customers and we support the identification of new profitable customers for a large German business-to-business mail-order company. The company has a structured, marketing database where information is stored about existing customers and their sales volume, as well as their e-mail addresses.

Based on the information from the structured, marketing database, the company identifies 150,000 customers. An aggregation of customers' affiliation is done because several customers probably belong to the same company. As a result, about 60,000 companies can be identified. This number is reduced to about 35,568 companies for which a corresponding website in German language can be identified. Additionally, the sales volume is calculated summing the incoming orders in the recent year to ensure that companies are currently profitable for the mail-order company. Then, profitable companies are defined with a sales volume exceeding a specific threshold determined by the mail-order company.

The data characteristics are shown in Table 1 for the randomly split training, validation and test set. The training and validation set was used to obtain the optimal SVD dimension and the model estimates, while the test set is used to validate and compare the different models.

	Number of customer groups	Relative percentage
Training set (including validation set):		
Non-profitable customer group website addresses	11,344	45.56
Profitable customer group website addresses	13.,553	54.44
Total	24,897	
Test set:		
Non-profitable customer group website addresses	4,793	44.92
Profitable customer group website addresses	5,878	55.08
Total	10,671	

Table 1: Overview of the website characteristics

4.2 Optimal dimension selection

After the pre-processing phase, a high-dimensional term-by-website matrix was created. To obtain its optimal reduced rank, a cross-validation procedure was applied on the training data (see Fig. 4). The x-axis represents the number of concepts and the y-axis represents the cross-validated AUC under the ROC curve. In the range of 1–50 concepts, the cross-validated AUC was increasing rapidly. From 50 concepts on, it was increasing less rapidly, while in the region around 150 concepts, the cross-validated performance was stabilizing. Including more than 150 concepts resulted in a more complex prediction model, while the AUC hardly increased. Thus, 150 concepts were chosen as the optimal number for representing the textual information in our study. At this point, a good balance was achieved between the number of concepts and the predictive performance.

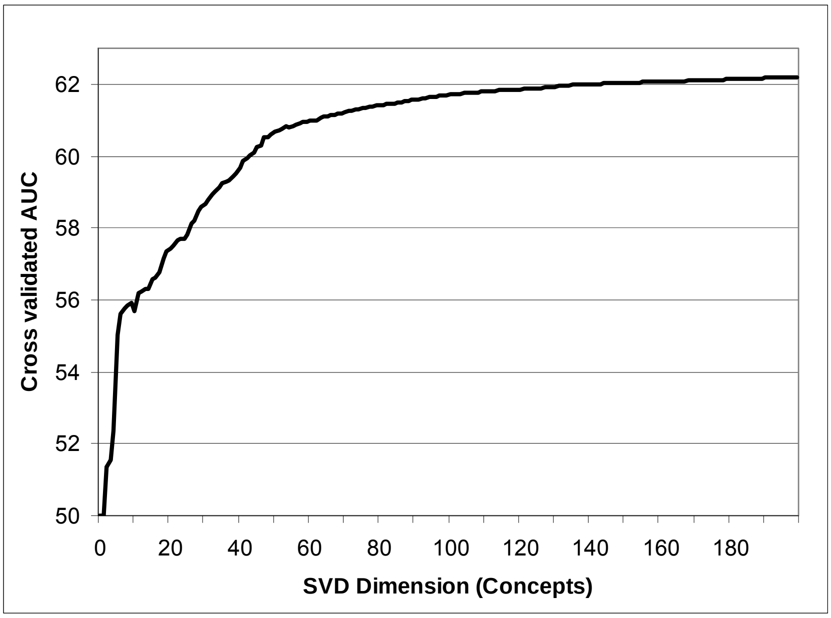


Figure 4: SVD Dimension

Each calculated latent semantic concept shows that the above-chance frequent occurrence of a group of several terms together with the non-occurrence of a further group of several terms on a customer's website can be used to classify this customer as profitable. The terms represent words in stemmed form and they are in German language because only German language websites are considered. Two examples for the interpretation of single SVD dimensions are presented below where the terms are translated to the English language.

Develop (including development, developer etc.) and System (including systems etc.) are two terms that frequently occur together on profitable customers' websites together with the frequent occurrence of following terms (also in stemmed form): Planning, Material, Technique, Build, Product, Machine, Protection, Industry, and Workshop. Further, the following terms should not occur frequently in this context to increase probability of a profitable customer website: Section, History, Experience, Business, Insurance, Energy, Quality and Mobile.

Service (including services, serviced, servicing etc.) and Project (including projects etc.) are two terms that frequently occur together on profitable customers’ websites together with the frequent occurrence of following terms (also in stemmed form): Conference, Consulting, Law, Information, Data, Management, Meeting, Union, Contract, Partner, and Staff. Further, the following terms should not occur frequently in this context to increase probability of a profitable customer website: Price, Customer, Offer, Payment, Market, and Tax

The first example could be interpreted as a customer who is interested in workshop equipment and furniture for his production process and the second examples probably shows a customer who is interested in office equipment and furniture. However, it is hard to interpret intuitively why some specific terms should not occur frequently in this context.

4.3 Creating and comparing address lists

In the clustering phase, the EM algorithm identifies seven clusters as well as terms representing cluster labels. Precision (the number of profitable customers’ websites over the number of all websites in a cluster) and recall (the number of profitable customers’ websites in a cluster over the number of all profitable customers’ websites) are computed and clusters are selected with the highest precision values at a recall value over a specific threshold. Terms from the selected cluster labels are used for further processing. As a result, one cluster can be identified with the highest precision and recall value (e.g. 58% precision at 37% recall). Ten terms are extracted from the cluster label (Arbeit, Unternehmen, System, Mitarbeiter, Bereich, Bauen, Technisch, Inhalt, Produkt, Kunde). Heuristically, we are searching for websites that contain at least four of these ten terms by a web search engine API. Companies behind the resulting addresses are manually identified and are added to a list of new potential customers for the mail-order company if they do not occur in the training or validation set. As a result, 160 companies are identified. Comparing these company addresses to addresses from profitable companies from the test set shows that 29 of them (about 18 %) can be classified as profitable (see Table 2). Additionally, 5 of them can be classified as non-profitable. The remaining 127 addresses are used in the acquisition process of the mail-order company. Regardless of the acquisition results - whether further addresses can be classified as profitable or not – a success rate of about 18 % is a good value for this automatically generated list.

	Number of addresses	Relative percentage
Test set A: Addresses generated by this approach		
Non-profitable and non-classified website addresses	131	81.87
Profitable customer group website addresses	29	18.13
Total	160	

Table 2: Overview of the address list characteristics

For evaluation purposes (to determine the frequent baseline), it is critical that the success rate of the traditional acquisition process - by using lists of customers from list brokers - can be estimated. In the year 2008, 3200 company addresses are received from list brokers. The acquisition process leads to 160 profitable customers. The probability that a new customer address leads to a profitable customer, therefore, can be estimated based on an acquisition process as follows: P(A/B) = 160/3200 = about 5 %. According to the acquisition manager, 5 % seems to be a representative value for the success rate of those lists. This low success rate shows the problem for acquisition manager because they get a large amount of addresses but only a few of them lead to profitable acquisitions.

As seen from this example, the density of the number of profitable customers from the address list generated by the approach presented in this paper is about three times larger than the density from list brokers’ lists, which companies even have to pay for. Thus, the use of the new potential customer list created by our approach outperforms lists from list brokers and it improves the identification of new profitable business customers while reducing costs.

4.4 Comparing predictive performance

The test set is built on a sample of customers who ordered at least once in the last three years. In contrast to this, the new address list (test set A) additionally includes those who never ordered anything. Thus, it is important to use both, the test set and the test set A to measure the predictive performance.

Overall, Fig. 5 and Fig. 6 show that the predictive performance of the regression model significantly outperforms the baseline because curves from the test sets are situated above the baseline. Fig. 7 also reveals that test sets outperform the baseline at a recall greater than a specific threshold.

Firstly, the cumulative lift curve of the test set and the test set A are above the baseline. Thus, the test sets are able to identify more profitable customers than the baseline within a specific percentile, e.g. the lift value in the top 30 percentile increases from one to 1.21 (test set) and from one to 1.11 (test set A). Secondly, the ROC curve of the test sets lay above the random baseline. Thus, the AUC of the test set (0.6116) and test set A (0.6352) is larger than the baseline (0.5000). This improvement is significant (χ^2=0.02 , d.f.=1, p<0.001). This shows that the model is able to better distinguish profitable from non-profitable customers than the baseline. Thirdly, the precision and recall diagram shows the test sets outperform the baseline at a recall greater than 32 % (test set) and 76 % (test set A). Especially this precision and recall diagram additionally shows that this approach should not be used alone as predictive model but it should be used in addition to conventional acquisition information to transform the acquisition process into a more targeted approach.

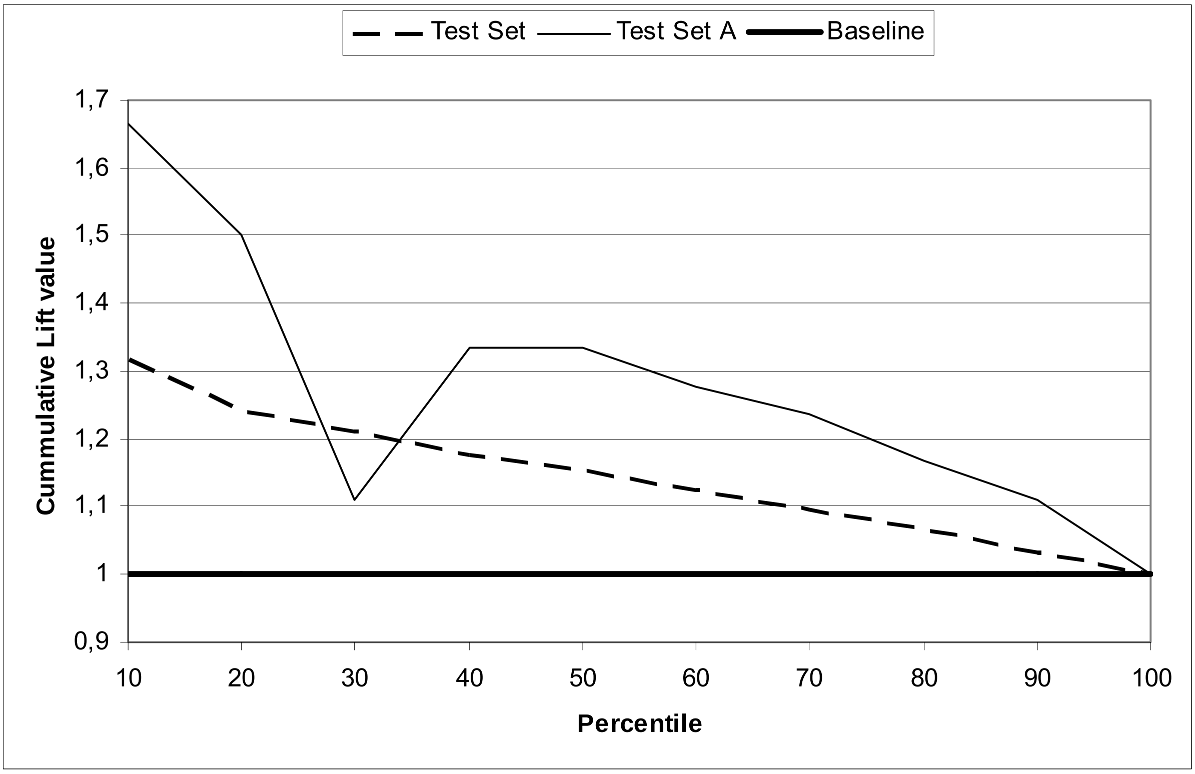


Figure 5: Test sets and baseline lift for the logistic regression model

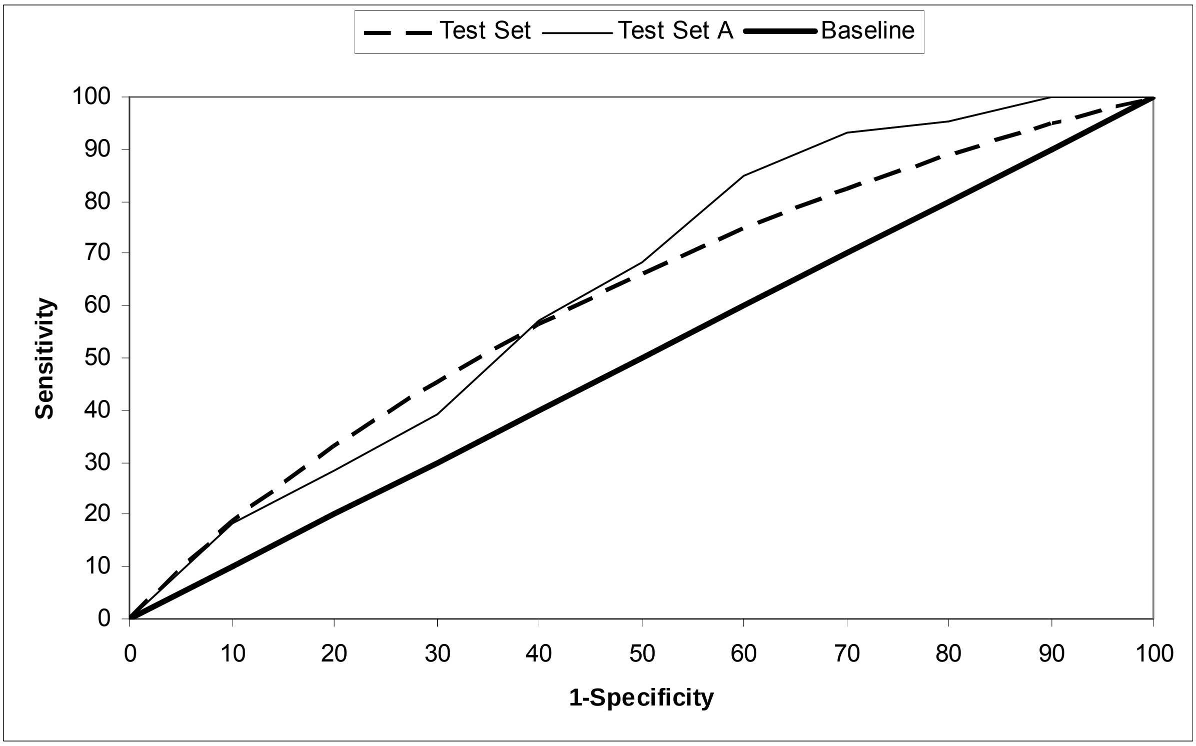


Figure 6: Sensitivity / Specificity Diagram

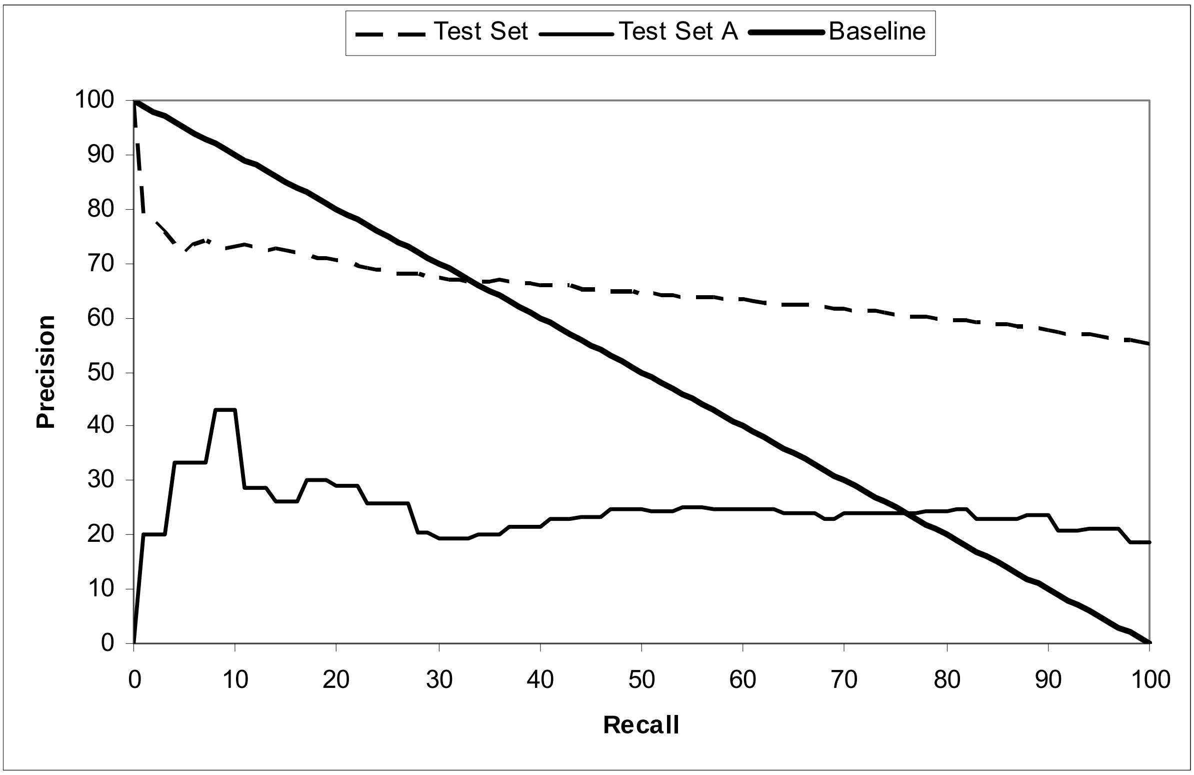


Figure 7: Precision / Recall Diagram

5 Conclusion

In this paper, we demonstrate that using information of existing customers' websites for acquisition purposes helps a B-to-B acquisition manager to identify profitable customers with a higher precision. Consequently, the acquisition process can become more targeted by additionally integrating this textual information. Specific data collection, pre-processing, and dimension reduction steps are required to convert the unstructured textual information into a structured form suitable for profitability prediction. A clustering of websites based on latent semantic concepts leads to the identification of further potential customers that outperforms customers acquired from list brokers by a wide margin. Future work should focus on improving the prediction by adding further unstructured information from existing customers (e.g. e-mails) to the prediction model.

Acknowledgments

We would like to thank the anonymous German mail-order company for their efforts in providing us their data. This project was realized using SAS v9.1.3, SAS Text Miner v5.2, Fraunhofer Idea Web Miner v1.0, and Matlab v7.0.4. For data collection, a self-developed program for web content mining was used. The second author acknowledges the IAP research network grant nr. P6/03 of the Belgian government (Belgian Science Policy).

Bibliography

Allison, P. D. (1999). *Logistic Regression using the SAS System: Theory and Application.* Cary: SAS Institute Inc.

Archer, N., & Yuan, Y. (2000). Managing business-to-business relationships throughout the e-commerce procurement life cycle. *Internet Research: Electronic Networking Applications and Policy,* 10(5), 385-395.

Baecke, P. H., & Van den Poel, D. (2010a). Improving purchasing behavior predictions by data augmentation with situational variables. *International Journal of Information Technology and Decision Making,* Forthcoming.

Baecke, P. H., & Van den Poel, D. (2010b). Data Augmentation by Predicting Spending Pleasure Using Commercially Available External Data. *Journal of Intelligent Information Systems,* Forthcoming.

Bose, I., & Mahapatra, R. K. (2001). Business data mining - a machine learning approach. *Information and Management,* 39(3), 211–225.

Bucklin, R. E., & Gupta, S. (1992). Brand choice, purchase incidence and segmentation: an integrated modeling approach. *Journal of Marketing Research*, 29(2), 201–215.

Chen, M.-C., Chiu, A..L., & Chang, H.H. (2005). Mining changes in customer behavior in retail marketing. *Expert Systems with Applications,* 28(4), 773-781.

Chen, M.-Y., Chu, H.-C. & Chen, Y.-M. (2010). Developing a semantic-enable information retrieval mechanism. Expert Systems with Applications, 37(1), 322-340

Coussement, K., & Van den Poel, D. (2009). Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. *Expert Systems with Applications,* 36, 6127-6134.

Coussement, K., & Van den Poel, D. (2008). Integrating the voice of customers through call center emails into a decision support system for churn prediction. *Information & Management,* 45, 164-174.

Coviello, N., Brodie, R.J., & Munro, H. (1997). Understanding contemporary marketing: Development of a classification scheme. *Journal of Marketing Management,* 13(6), 501-522.

De Bock, K. W., & Van den Poel, D. (2009). Predicting website audience demographics for web advertising targeting using multi-website clickstream data. *Fundamenta Informaticae,* 97, 1-19.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science,* 41(6), 391-407.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics,* 44(3), 837–845.

Engler, J., & Kusiak, A. (2010). Mining Authoritativeness of Collaborative Innovation Partners. *International Journal of Computers, Communications & Control,* V(1), 42-51.

Gericke, W., Thorleuchter, D., Weck, G., Reiländer F., & Loß, D. (2009). Vertrauliche Verarbeitung staatlich eingestufter Information - die Informationstechnologie im Geheimschutz. *Informatik Spektrum,* 32(2), 102-109.

Greiff, W. R. (1998). A theory of term weighting based on exploratory data analysis. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), Proceedings of the 21st *SIGIR Conference*. New York: ACM, pp. 11-19.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology,* 143(1), 29-36.

Inagaki, S. (2010). The Effects of Proposals for Basic Pension Reform on the Income Distribution of the Elderly in Japan. *The Review of Socionetwork Strategies*, 4(1), 1-16.

Kim, Y. S. (2006). Toward a successful CRM: variable selection, sampling and ensemble. *Decision Support Systems,* 41(2), 542–553.

Lee, K.-C., & Chung, N. (2003). Identification of Customer Segmentation Strategies by Using Machine Learning-Oriented Web-mining Technique. *IE Interfaces,* 16(1), 54-62.

Menon, A., Homburg, C., & Beutin, N. (2005). Understanding customer value in business-to-business relationships. *Journal of business-to-business mark*eting*,* 12(2), 1-38.

Naude, P., & Holland, C. (1996). *Relationship Marketing.* London: Paul Chapman Publishing. pp. 40-54.

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. (2006). Defection detection: measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research,* 43(2), 204–211.

Pan, S. L., & Lee, J. N. (2003). Using e-CRM for a unified view of the customer. *Communications of ACM,* 46(4), 95–99.

Park, Y.-J., & Chang, K.-N. (2009). Individual and group behavior-based customer profile model for personalized product recommendation. *Expert Systems with Applications,* 36(2), 1932-1939.

Reinartz, W., & Kumar, V. (2003). The impact of customer relationship characteristics on profitable lifetime duration. *Journal of Marketing,* 67(1), 77–99.

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

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management,* 24(5), 513–523.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation,* 28(1), 11-21.

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010a). Mining Ideas from Textual Information. *Expert Systems with Applications,* 37(10), 7182-7188.

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010b). 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., & Prinzie, A. (2010c). Extracting consumers needs for new products - A web mining approach. In *Proceedings WKDD 2010* (p. 441), Los Alamitos: IEEE Computer Society.

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010d). Mining innovative ideas to support new product research and development. In H. Locarek-Junge, & C. Weihs (Eds.), *Classification as a Tool for Research* (pp. 587-594). Berlin: Springer-Verlag.

Thorleuchter, D. (2008). Finding Technological Ideas and Inventions with Text Mining and Technique Philosophy. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data Analysis, Machine Learning, and Applications* (pp. 413-420). Berlin: Springer-Verlag.

Van den Poel, D., & Buckinx, W. (2005). Predicting Online-Purchasing Behavior. *European Journal of Operational Research,* 166(2), 557-575.

Verhoef, P. C., Venkatesan, R., McAlister, L., Malthouse, E. C., Krafft, M., & Ganesan, S. (2010). CRM in Data-Rich Multichannel Retailing Environments: A Review and Future Research Directions. *Journal of Interactive Marketing,* 24(2), 121-137.

Wangenheim, F., & Bayon, T. (2007). The chain from customer satisfaction via word-of-mouth referrals to new customer acquisition. Journal *of the Academy of Marketing Science,* 35, 233-249.

Yen, E., & Lina L.-H. (2010). Rubik's cube watermark technology for grayscale images. Expert Systems with Applications, 37(6), 4033-4039.

Zhong, J., & Li, X. (2010). Unified collaborative filtering model based on combination of latent features. Expert Systems with Applications, 37(8), 5666-5672.

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

Chapter II

Predicting E-Commerce Company Success by Mining the Text of Its Publicly-Accessible Website

This chapter is based on Thorleuchter, D., Van den Poel, D., Predicting E-Commerce Company Success by Mining the Text of Its Publicly-Accessible Website. The Journal of Applied Research and Technology (in review).

Table of Contents

Predicting E-Commerce Company Success by Mining the Text of Its Publicly-Accessible Website

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

[a] Fraunhofer INT, D-53879 Euskirchen, Appelsgarten 2, Germany & PhD Candidate, Ghent University, Belgium, dirk.thorleuchter@int.fraunhofer.de
[b] Ghent University, Faculty of Economics and Business Administration, B-9000 Gent, Tweekerkenstraat 2, Belgium, dirk.vandenpoel@ugent.be

Abstract

We analyze the impact of textual information from e-commerce companies' web sites on their commercial success. The textual information is extracted from web content of e-commerce companies divided into the top 100 worldwide most successful companies and into the top 101 to 500 worldwide most successful companies. It is shown that latent semantic concepts extracted from the analysis of textual information can be adopted as success measures for a top 100 e-commerce company classification. This contributes to the existing literature concerning web site success measures for e-commerce. As evaluation, a regression model based on these concepts is built that is successful in predicting the commercial success of the top 100 companies. These findings are valuable for e-commerce web sites creation.

Key Words

Success Measure, E-Commerce, SVD, Classification, Textmining, Website

1 Introduction

In literature, web site success measures for e-commerce companies are used to predict successful e-commerce companies in contrast to none or less successful e-commerce companies (Zvirana et al. 2006). These include e.g. the usability of the web page, human computer interaction, well-known brand, price reduction, money-back guarantee, etc. In this paper, we take a different approach: Success measures are extracted by mining the text published on their e-commerce company's website. The dimensions are evaluated in terms of their usefulness for predicting most successful e-commerce companies (Top 100) in contrast to successful e-commerce companies (Top 101 to 500). Thus, existing success measures for predicting successful e-commerce companies from literature are analyzed to show their (non-) success in predicting most successful e-commerce companies. These results contribute to the existing e-commerce success measure literature.

Web mining (Thorleuchter et al. 2010c) is used for crawling textual information from the top 500 e-commerce companies' websites and latent semantic indexing is used to analyze their impact on the most successful e-commerce companies. Further, a regression model is built that is based on the latent semantic concepts (Coussement and Van den Poel 2008). It can be shown that the model is successful in predicting the top 100 e-commerce companies. These findings are valuable for e-commerce web sites creation.

2 Related Work

Measuring the success of web sites is a well-known topic in literature (e.g. Baecke and Van den Poel 2010a, Baecke and Van den Poel 2010b, Delone and McLean 1992, DeBock and Van den Poel 2009, Lopeza and Ruiz 2010, Lu et al. 2010, Serrano-Cinca et al. 2010, Van den Poel and Buckinx 2005, Verhoef et al. 2010). Many relevant practical measures are described e.g. the usability, the user acceptance, the user participation, the user interaction, and the attitude (Barki and Hardwick 1994).

However, for e-commerce companies the success of their web sites can only be measured indirectly (Galletta and Lederer 1989). Thus, different success measures for e-commerce web sites are relevant for identifying successful e-commerce companies.

Offering a money-back guarantee, a well-known brand, and a price reduction are mentioned by Robins et al. 2002 as an important success measure. Further research identifies the usability of the web page, and a human computer interaction (HCI) as additional success measure (Heldal et al. 2004).

Several success measures are introduced by Chang et al. 2004: internet product choice, online payment, internet vendor trust, shopping travel, internet shopping convenience, internet ecology, internet customer relation, and internet product value.

The impact of order delivery on the success of e-commerce companies is measured by Van den Poel and Leunis 1999. Further success measures are design features, information and web site quality, and user characteristics (Zvirana et al. 2006).

3 Methodology

3.1 Overview

In this paper, we use textual information from existing e-commerce companies' websites. Lists of the top 100 and top 500 successful e-commerce companies are used and the web sites behind the companies are identified. For data collection, textual information from these web sites is crawled by use of methods from web mining and is stored in documents. Documents are divided in training set and test set and they are also pre-processed by use of text mining methods. A term-website matrix based on the training set is created that is used to identify the latent semantic patterns of the training documents. The test documents are projected into the same latent semantic concept-space. A logistic regression model is built on this concept-space matrix to show that this approach is successful in predicting the most successful top 100 e-commerce companies. Fig. 1 shows the methodology of this approach.

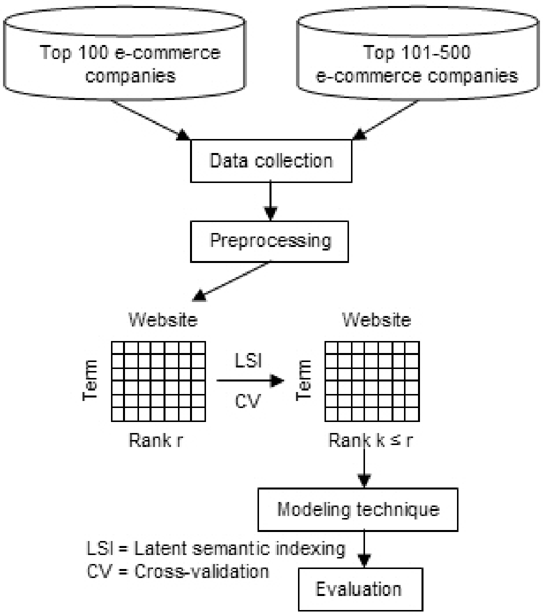


Figure 1: Different steps of the approach

3.2 Data collection

The unstructured content information from e-commerce companies' websites is collected by use of web mining methods. After identifying companies' web sites, it is considered that a website consists of several web pages. Crawling all textual information from all web pages (e.g. 'sitemap', disclaimer', 'data protection policy / privacy') leads to a huge amount of information. Thus, only relevant information from companies' websites is extracted by limiting the number of web pages per company to five. To identify the five most relevant web pages from a company's website, the starting page is selected. Additionally, three (sub) web pages are selected with the highest page rank returned by an internet search engine. Further relevant information might be descriptions about a company's history. Thus, web pages that contain specific terms (e.g. history, founded) are identified and the web page with the highest page rank from these identified web pages is selected if not already selected before.

To identify the relevance of a web page concerning its page rank, Google is used as internet search engine because of the high quality of its page rank algorithm and because of the fact that all top 500 e-commerce companies' websites can be found in the Google index. For each company, search queries are restricted to web pages of the respective company. They are automatically executed by web services (Carl 2008) (web based advanced programming interfaces). The result data contain web pages from the company ordered by the page rank.

By use of this web mining approach, a large amount of highly unstructured information is extracted from the companies' web sites. This information has to be pre-processed by use of text mining approaches to discover relevant features.

3.3 Pre-processing

To represent the extracted textual information as term vector of weighted frequencies, several methods from text mining are applied based on

- a text preparation step,
- a term filtering step,
- a vector weighting step, and
- a term vector aggregation step.

In a text preparation step, the raw text is cleaned (e.g. by deleting images, html-, or xml-tags, specific characters as well as scripting code). The punctuation is removed and a dictionary is used to correct typographical errors. Then, tokenization (Thorleuchter et al. 2010a) where the term unit is word and case conversion (converting terms in lower case and capitalizing the first character) is applied.

In a term filtering step, several filtering methods (Thorleuchter et al. 2010b) are used. Part-of-speech tagging is used to identify the syntactic category of a term. Stop word filtering is also used to identify terms with little or no content information (Thorleuchter et al. 2010d). With dictionary-based stemming, the basic form of words - where the same stem represents related words - is identified. Additionally, Zipf distribution (Zipf 1949) is used to reduce the number of terms by deleting rare terms. After this, the selected terms are checked manually (Gericke et al. 2009).

Then, a term vector is built. The component values of a term vector are weighted frequencies instead of using raw frequencies (the number of term-appearance in a web page) because the use of weighted frequencies significantly improves retrieval performance (Jones 1972). Terms with large weights frequently occur in a small number of web pages but they do not occur frequently in all web pages (Salton and Buckley 1988).

A well-known term weighting scheme is used (Salton et al. 1994) and described in formula (1). The term frequency $tf_{i,j}$ equals the absolute frequency of term i in web page j, the inverse document frequency idf_i equals $\log(n/ df_i)$, the square root represents a length normalization factor, n equals the number of web pages, and m equals the dimension of the term vectors (Hotho 2005).

$$w_{i,j} = \frac{tf_{i,j} \cdot \log(n/df_i)}{\sqrt{\sum_{p=1}^{m} {tf_{i,j_p}}^2 \cdot (\log(n/df_{i_p}))^2}} \qquad (1)$$

In a term vector aggregation step, all vectors representing web pages from a specific company are aggregated to build one term vector for each company's website j (Coussement and Van den Poel 2008). This is calculated by

$$Aw_{i,j} = \sum_{k=1}^{r} w_{i,k} \qquad (2)$$

where the weight of term i in web page k is represented by $w_{i,k}$ and where r equals five (the number of web pages per company). Then, a term-by-website matrix with weighted frequencies is created.

3.4 Concept identification with LSI and singular value decomposition

Normally, the term-by-website matrix is high dimensional and most of its weights are zero. To reduce the dimensionality, latent semantic indexing (LSI) combined with singular value decomposition (SVD) is used. This method groups terms into concepts by forming semantic generalizations. If A is the term-by-website (m x n) matrix with rank r (r ≤ min(m,n)) then SVD of A is a transformation into a product of three matrices, the term-concept similarity (m x r) matrix U, the concept-website similarity (n x r) matrix V, and a diagonal (r x r) matrix Σ containing positive singular values of matrix A.

$$A = U \Sigma V^t \tag{3}$$

To reduce the rank r of A to k (k ≤ r), LSI considers the first k singular values in Σ by retaining only the first k columns of U and V. Further singular values are discarded.

An important decision to be taken is the choice of k. This critical parameter influences the predictive performance. To determine the value of the parameter k, several rank k-models are constructed, are evaluated concerning their predictive performance, and the most favorable rank-k model is selected. To calculate the predictive performance, a prediction model is used (see Sect. 3.5).

The selected rank k-model is built on the training examples. The test examples are integrated into the same semantic subspace as created by the training examples (Deerwester 1990).

3.5 Prediction modeling

The predictive performance is measured by logistic regression as modeling technique where a maximum likelihood function is maximized (Allison 1999, Inagaki 2010). Advantages for the use of logistic regression are the simplicity (DeLong et al. 1988), computational speed and robustness (Greiff 1998). It can be calculated by

$$P(y=1 \mid x) = \frac{1}{1+\exp(-(w_0+wx))} \tag{4}$$

with $T = \{(x_i, y_i)\}$ the training set, $i = \{1,2,...,N\}$, $x \in R^n$ the n-dimensional input vector (a concept-website vector), w the parameter vector, w_0 the intercept, and $y_i \in \{0,1\}$ the corresponding binary target labels (company in top 101 to 500, company in top 100).

3.6 Evaluation criteria

This evaluation is done with the commonly used criteria: cumulative lift, precision, recall, area under the receiver operating characteristics curve (AUC), sensitivity, and specificity. Cumulative lift measures the increase in density concerning the number of top 100 companies relative to the density of the companies in total. The precision measures the fidelity or exactness and the recall measures the completeness of the predicted results. The proportion of positive cases predicted to be positive is named sensitivity and the proportion of negative cases predicted to be negative is named specificity. The two-dimensional plot of the sensitivity versus (1-specificity) is named receiver operating characteristics curve (ROC). The AUC is the area under the ROC curve. It can be used as performance measure for binary classification (Hanley and McNeil 1982).

4 Case Study

4.1 Research data

In this study, we use lists of the top 100 and top 500 successful e-commerce companies as published on the internet (www.welt.de and www.internetretailer.com). The corresponding websites behind these companies are manually identified. Normally, successful companies offer web sites in several languages in the internet. However here, the selected web sites are restricted to the English language to prevent the language translation problem. As a result, all 500 companies offer websites in English language. Thus, 500 textual documents are created that contain content information from the most relevant web sites of each company in English.

Table 1 provides summary information of the (randomly-selected) training and test set. The optimal SVD dimension is calculated using the training set and a regression model is estimated. The test set is used to show the success of the regression model compared to the frequent baseline as calculated from the relative percentage in Table 1.

	Number of customer groups	*Relative percentage*
Training set:		
Top 100 companies' websites	50	20
Top 101 to 500 companies' websites	200	80
Total	250	
Test set:		
Top 100 companies' websites	50	20
Top 101 to 500 companies' websites	200	80
Total	250	

Table 1: Overview of website characteristics

4.2 Optimal dimension selection

The result of the pre-processing step is a term-by-website matrix with high dimensionality. The training set is used to calculate an optimal SVD dimension with a cross-validation procedure (see Fig. 2). The number of concepts is represented by the x-axis and the cross-validated AUC is represented by the y-axis. The cross-validated AUC increases in the range of 2-17 concepts, it reaches a maximum at 18 concepts, and from 19 concepts on, it decreases. Thus, 18 concepts were selected as the optimal number for the SVD dimension.

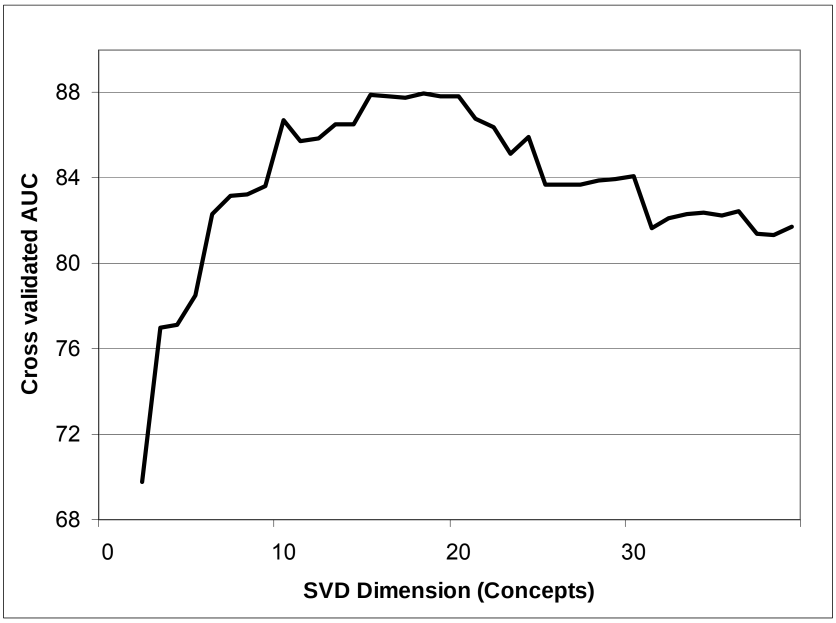


Figure 2: Calculating an optimal SVD dimension

4.3 Comparing predictive performance

The predictive performance of the regression model is compared to the baseline by use of the following criteria: Cumulative lift curve, ROC curve, and precision/recall diagram. Fig. 3, Fig. 4, and Fig. 5 show the general success of the regression model compared to the baseline. Additionally, a three-fold cross validation is used to prevent overfitting.

Fig. 3 shows that the cumulative lift curve lies above the baseline. Thus, the density concerning the number of top 100 companies in each percentile is greater than the density from the baseline. The ROC curve of the test sets also lies above the random baseline. Thus, the AUC of the test set (87,96) is larger than that of the baseline (50,00) with a significant improvement ($\chi^2=0.02$, d.f.=1, $p<0.001$). Additionally, the precision and recall diagram lies over the baseline at all recall values. These criteria show that the model is able to better distinguish top 100 from top 101 to 500 companies than the baseline.

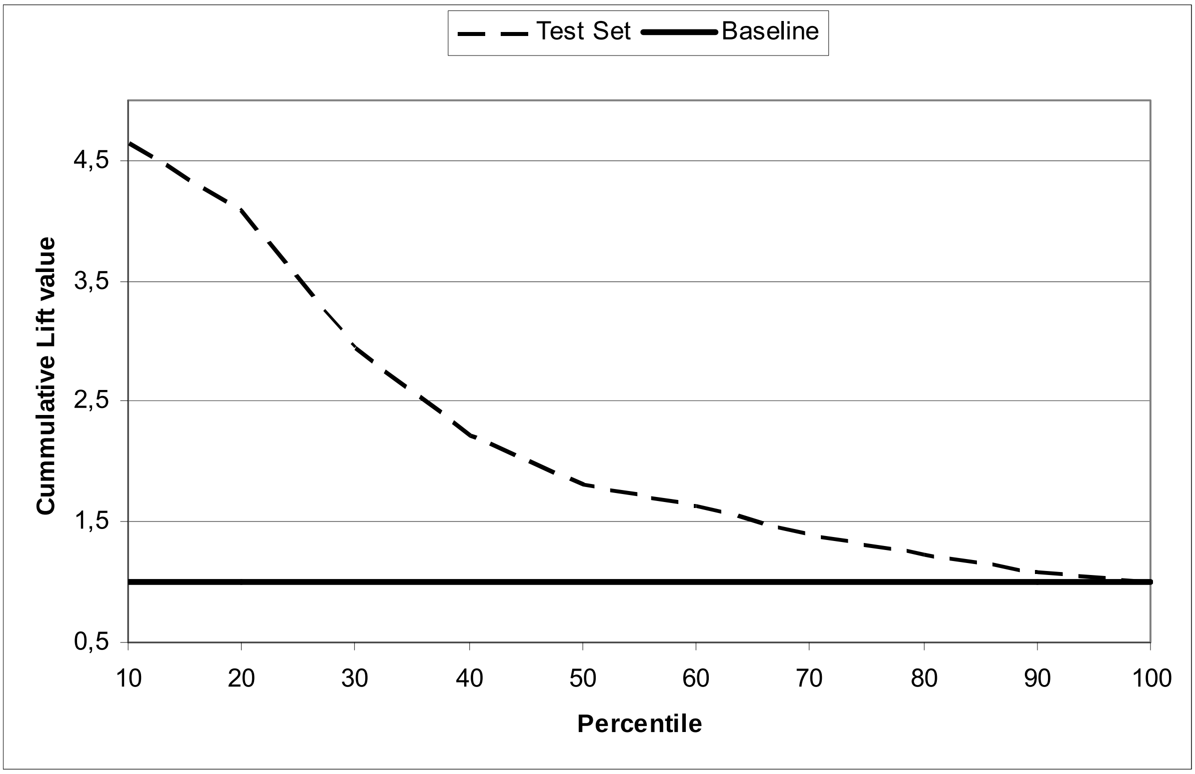


Figure 3: Cumulative lift value of the test set and of the baseline

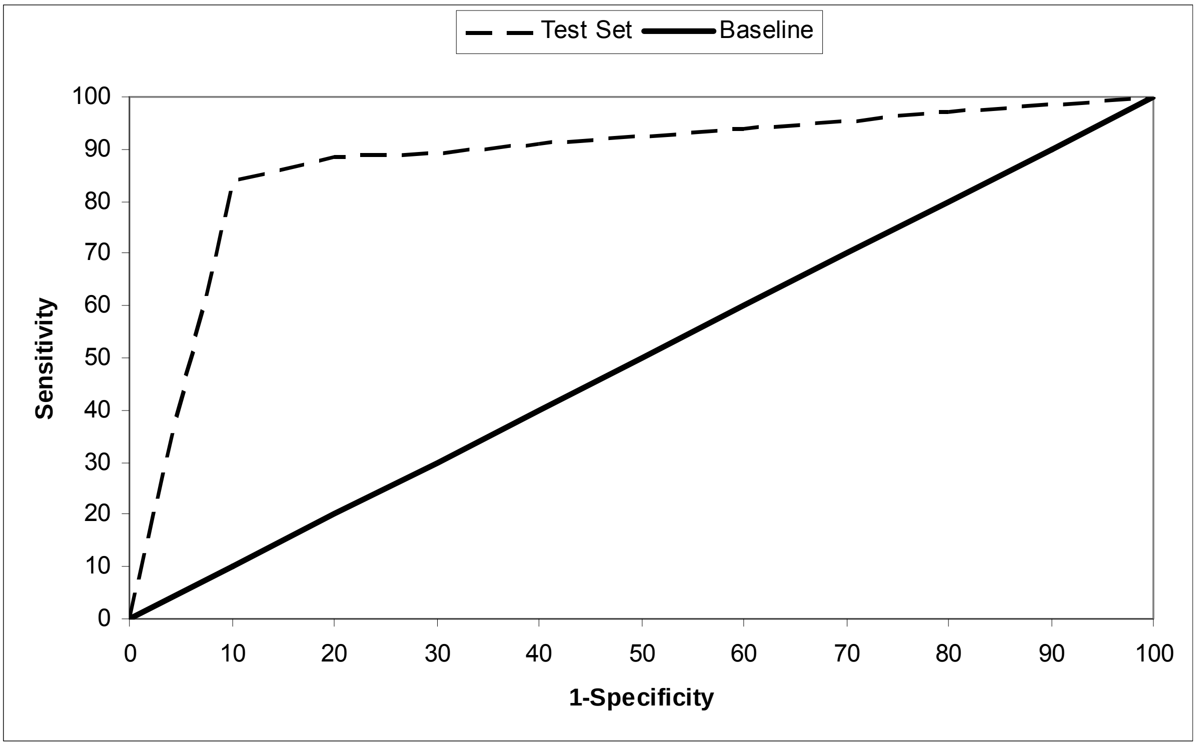


Figure 4: Sensitivity - specificity diagram of test set and baseline

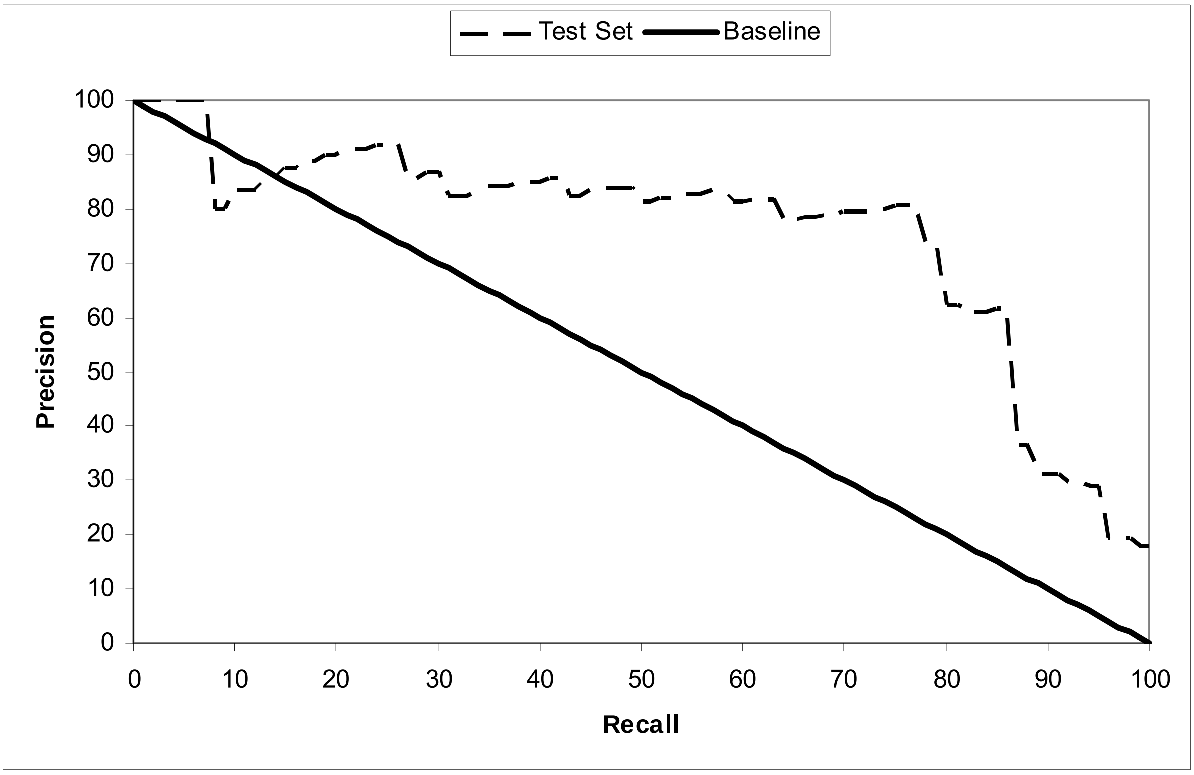


Figure 5: Precision - recall diagram of test set and baseline

4.4 Results

Based on the calculated latent semantic concepts, groups of several terms can be identified as representative for the positive examples. A group of terms is defined by the above-chance frequent occurrence of several terms in a text pattern from the positive examples. The term unit is word in stemmed form and a text pattern is defined as a number of l terms alongside each other in textual information (Thorleuchter 2008). For this case study, l is set to 10. As a result, the identified groups of terms occur frequently in the positive examples but never in the negative examples. Additionally groups of terms are identified that are representative for the negative examples. These groups occur frequently in the negative examples but not in the positive examples.

We examine existing measures from Sect. 2 concerning their impact on predicting successful e-commerce companies: Money-back guarantee, human computer interaction, internet vendor trust, and internet customer relation. Additionally, we also use trusted order delivery (Khouja 2001).

Three objectives can be shown that are representative for the positive examples. The occurrence of textual information on companies' web sites describing internet customer relation by rating and providing services (Chang et al. 2004) is a good predictor for the top 100 e-commerce companies. A second predictor is the internet vendor trust (Chang et al. 2004) based on textual information describing trusted financing possibilities for products. The third predictor is the human computer interaction (Heldal et al. 2004) based on textual information about an automatic recommendation for products, features, and services concerning user-given information.

Additionally, three objectives can be shown as predictor for the negative examples. The refund of money (money-back policy / guarantee) is mainly described on web pages of the top 101 to 500 e-commerce company list. A second predictor is a trusted order delivery that can be monitored by the user. A third predictor is to improve internet customer relation by use of newsletters.

To show the important results in detail, groups of terms that are representative for the positive examples are presented below:

A1. Service (including services, serviced, servicing etc.) and company (including companies etc.) are two terms that occur frequently in text patterns of the positive examples together with the following terms in stemmed form: Rate, provide, product, offer, user, exclusive, ecommerce, and market. This group of terms does not occur in text patterns of the negative examples. The terms describe the rating and providing of services for products in an e-commerce market to users. This is important to quality for sale and after-sale service, it leads to an improved internet customer relation, and it confirms the corresponding success measure as mentioned in Chang et al. 2004.

A2. Product and finance are two terms that occur frequently in text patterns of the positive examples together with the following terms in stemmed form: Individual, offer, download, business, account, resource, asset, information, service, institute, call, and competition. This group of terms does not occur in text patterns of the negative examples. Providing trusted financing possibilities for products is described by these terms. This is important to improve vendor legitimacy and it leads to improved internet vendor trust. Thus, it confirms the corresponding success measure mentioned in Chang et al. 2004.

A3. Search and recommendation are two terms that occur frequently in text patterns of the positive examples together with the following terms in stemmed form: System, feature, product, service, automatic, user, query, and multiple. This group of terms does not occur in text patterns of the negative examples. The terms describe an automatic recommendation of products, features, and services based on user-given search queries. This is important to improve human computer interaction. Thus, it confirms the corresponding success measure mentioned in Heldal et al. 2004.

Furthermore, groups of terms that are representative for the negative examples are presented below:

B1. Money and refund are two terms that occur frequently in text patterns of the negative examples together with the following terms in stemmed form: Price, order, replace, purchase, address, cancel, product, send, back, and simplify. This group of terms does not occur in text patterns of the positive examples. The terms describe a refund of money (money-back policy of a company) as predictor for the negative examples. This means if a money-back policy of a company is described on the company’s web page then this company is not in the top 100 e-commerce companies. This contradicts research results of Van den Poel and Leunis 1999 where money-back guarantee is a good predictor for successful e-commerce companies. A potential reason for this could be that times have changed since 1999, money-back guarantee is now quite natural, and very successful companies do not need to mention it separately.

B2. Order and delivery are two terms that occur frequently in text patterns of the negative examples together with the following terms in stemmed form: Inbox, accurate, home, select, week, view, depart, monitor, and product. This group of terms does not occur in text patterns of the positive examples. Terms describe a trusted order delivery that can be monitored by the user. This means if a trusted order delivery of a company is described on the company's web page then this company is not in the top 100 e-commerce companies. This contrasts order delivery as success measure as described in Khouja 2001 because for a top 100 e-commerce company, a trusted delivery and a monitoring is also quite natural and it is also not necessary to mention it on the web site.

B3: Custom and newsletter are two terms that occur frequently in text patterns of the negative examples together with the following terms in stemmed form: Mail, cost, store, product, sale, subscribe, price, registration, and coupon. This group of terms does not occur in text patterns of the positive examples. These terms describe the internet customer relation by use of newsletters. This one-directional newsletter communication is probably dated because textual information about this internet customer relation only can be found on websites of the top 101 to top 500 e-commerce companies but not in the top 100 e-commerce companies.

5 Conclusions

This work has analyzed the impact of textual information from e-commerce companies' web sites on their commercial success. It is shown that a logistic regression model based on latent semantic concepts from this textual information is successful in predicting the most successful top 100 e-commerce companies. The case study shows that internet vendor trust, human computer interaction, and internet customer relation by rating and providing services are successful measures in predicting the top 100 e-commerce companies and that money-back policy, trusted order delivery, and internet customer relation by use of newsletters are successful measures in predicting the top 101 to top 500 e-commerce companies.

This contributes to the existing literature concerning web site success measures for e-commerce and these findings are valuable for e-commerce web sites creation.

Bibliography

Allison, P. D. *Logistic Regression using the SAS System: Theory and Application*, SAS Institute Inc., Cary, NC, 1999.

Baecke, P. H., and Van den Poel, D. Improving purchasing behavior predictions by data augmentation with situational variables. *International Journal of Information Technology and Decision Making,* Forthcoming, 2010a.

Baecke, P. H., & Van den Poel, D. Data Augmentation by Predicting Spending Pleasure Using Commercially Available External Data. *Journal of Intelligent Information Systems,* Forthcoming, 2010b.

Barki, H., Hardwick, J. Measuring user participation, user involvement, and user attitude. *MIS Quarterly,* 18, 1, 1994, 59-79.

Carl, D., Clausen, J., Hassler, M., and Zund, A. *Mashups programmieren.* O'Reilly, Köln, Germany, 2008, 51-53.

Chang, J. C.-J., Torkzadeh, G., and Dhillon, G. Re-examining the measurement models of success for Internet commerce. *Information & Management,* 41, 2004, 577-584.

Coussement, K., and Van den Poel, D. Integrating the voice of customers through call center emails into a decision support system for churn prediction. *Information & Management,* 45, 2008, 164-174.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science,* 41, 6, 1990, 391-407.

DeBock, K. W., and Van den Poel, D. Predicting website audience demographics for web advertising targeting using multi-website clickstream data. *Fundamenta Informaticae*, 97, 2009, 1-19.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics,* 44, 3, 1988, 837-845.

Delone, W.H., and McLean, E.R. Information systems success: the quest for the dependent variable. *Information Systems Research,* 3, 1, 1992, 60-95.

Galletta, D.F., and Lederer, A.L. Some cautions on the measurement of user information satisfaction. *Decision Sciences*, 20, 1989, 419-438.

Gericke, W., Thorleuchter, D., Weck, G., Reiländer F., and Loß, D. Vertrauliche Verarbeitung staatlich eingestufter Information - die Informationstechnologie im Geheimschutz. *Informatik Spektrum,* 32, 2, 2009, 102-109.

Greiff, W. R. A theory of term weighting based on exploratory data analysis. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel (eds.), *Proceedings of the 21st SIGIR Conference*, ACM, New York, NY, 1998, 11-19.

Hanley, J. A., and McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology,* 143, 1, 1982, 29-36.

Heldal, F., Sjøvold, E., and Heldal, A.F. Success on the Internet—optimizing relationships through the corporate site. *International Journal of Information Management,* 24, 2, 2004, 115-129.

Hotho, A., Nürnberger, A., and Paaß, G. A Brief Survey of Text Mining. *LDV Forum*, 20, 1, 2005, 19-26.

Inagaki, S. The Effects of Proposals for Basic Pension Reform on the Income Distribution of the Elderly in Japan. *The Review of Socionetwork Strategies*, 4, 1, 2010, 1-16.

Jones, K. S. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28, 1, 1972, 11-21.

Khouja, M. The evaluation of drop shipping option for e-commerce retailers. *Computers & Industrial Engineering*, 41, 2, 2001, 109-126.

Lopeza, I., and Ruiz, S. Explaining website effectiveness: The hedonic–utilitarian dual mediation hypothesis. *Electronic Commerce Research and Applications*, doi:10.1016/j.elerap.2010.04.003.

Lu, Y., Zhao, L., and Wang, B. From virtual community members to C2C e-commerce buyers: Trust in virtual communities and its effect on consumers' purchase intention. *Electronic Commerce Research and Applications*, 9, 4, 2010, 346-360.

Robins, D., and Kelsey, S. Analysis of Web-based information architecture in a university library: navigating for known items. *Information Technology and Libraries*, 21, 4, 2002, 158-169.

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

Salton, G., and Buckley, C. Term-weighting approaches in automatic text retrieval. *Information Processing & Management,* 24, 5, 1988, 513–523.

Serrano-Cinca, C., Fuertes-Callén, Y., and Gutiérrez-Nieto, B. Internet positioning and performance of e-tailers: An empirical analysis. *Electronic Commerce Research and Applications*, 9, 3, 2010, 237-248.

Thorleuchter, D., Van den Poel, D., and Prinzie, A. Mining Ideas from Textual Information. *Expert Systems with Applications,* 37, 10, 2010a, 7182-7188.

Thorleuchter, D., Van den Poel, D., and Prinzie, A. A compared R&D-based and patent-based cross impact analysis for identifying relationships between technologies. *Technological Forecasting and Social Change,* 77, 7, 2010b, 1037-1050.

Thorleuchter, D., Van den Poel, D., and Prinzie, A. Extracting consumers needs for new products - A web mining approach. In *Proceedings WKDD 2010*, IEEE Computer Society, Los Alamitos, CA, 2010c, 441.

Thorleuchter, D., Van den Poel, D., and Prinzie, A. Mining innovative ideas to support new product research and development. In H. Locarek-Junge, and C. Weihs (eds.), *Classification as a Tool for Research,* Springer-Verlag, Berlin, Germany, 2010d.

Thorleuchter, D. Finding Technological Ideas and Inventions with Text Mining and Technique Philosophy. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker (Eds.), *Data Analysis, Machine Learning, and Applications*. Springer-Verlag, Berlin, Germany, 2008, 413-420.

Van den Poel, D., and Buckinx, W. Predicting Online-Purchasing Behavior. *European Journal of Operational Research,* 166, 2, 2005, 557-575.

Van den Poel, D., and Leunis, J. Consumer Acceptance of the Internet as a Channel of Distribution. *Journal of Business Research,* 45, 3, 1999, 249-256.

Verhoef, P. C., Venkatesan, R., McAlister, L., Malthouse, E. C., Krafft, M., and Ganesan, S. CRM in Data-Rich Multichannel Retailing Environments: A Review and Future Research Directions. *Journal of Interactive Marketing,* 24, 2, 2010, 121-137.

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

Zvirana, M., Glezerb, C., and Avnia, I. User satisfaction from commercial web sites: The effect of design and use. *Information & Management,* 43, 2, 2006, 157-178.

Chapter III

Finding new technological ideas and inventions with text mining and technique philosophy

This chapter is based on Thorleuchter, D. (2008). Finding Technological Ideas and Inventions with Text Mining and Technique Philosophy. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker R. (Eds.). Data Analysis, Machine Learning, and Applications. pp. 413-420. Springer, Berlin, Heidelberg.

Table of Contents

Finding new technological ideas and inventions with text mining and technique philosophy

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

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

[b] Ghent University, Faculty of Economics and Business Administration, B-9000 Gent, Tweekerkenstraat 2, Belgium, dirk.vandenpoel@ugent.be

Abstract

Text mining refers generally to the process of deriving high quality information from unstructured texts. Unstructured texts come in many shapes and sizes. It may be stored in research papers, articles in technical periodicals, reports, documents, web pages etc. Here we introduce a new approach for finding textual patterns representing new technological ideas and inventions in unstructured technological texts.

This text mining approach follows the statements of technique philosophy. Therefore, a technological idea or invention represents not only a new mean, but also a new purpose and mean combination. By systematic identification of the purposes, means and purpose-mean combinations in unstructured technological texts compared to specialized reference collections, a (semi-) automatic finding of ideas and inventions can be realized. Characteristics that are used to measure the quality of these patterns found in technological texts are comprehensibility and novelty to humans and usefulness for an application.

Keywords

Idea Mining, Text Mining, Text Classification, Technology

1. Introduction

The planning of technological and scientifically research and development (R&D-) programs is a very demanding task, e.g. in the R&D-program of the German ministry of defense there are at least over 1000 different R&D-projects running simultaneously. They all deal about 100 different technologies in the context of security and defense. There is always a lot of change in these programs – a lot of projects starting new and a lot of projects running out. New ideas or new inventions are a basis for a new R&D-project. That means for planning an R&D-program it is necessary to identify a lot of new technological ideas and inventions from the scientific community as a basis for future R&D-projects [7]. Up to now, the identification of new ideas and invention in unstructured texts is done manually (that means by humans) without the support of text mining. Therefore, in this paper, we will describe the theoretical background of the text mining approach to discover (semi-) automatically new ideas and inventions in unstructured texts.

Characteristics that are used to measure the quality of these patterns as described by [3] are comprehensibility and novelty for the program planers and the researchers - in further they will be called "users" and usefulness for identifying new R&D-projects inside the R&D-program.

Hotho [3] described the characteristics that are used to measure the quality of these textual patterns extracted by knowledge discovery tasks. They are comprehensibility and novelty to the users and usefulness for a task. In this paper, the users are the program planers and the researchers and the task is to find ideas and inventions which can be used as basis for new R&D-projects.

It is known from the cognition research that analysis and evaluation of textual information requires the knowledge of a context [9]. The selection of the context depends on the user and the tasks. Referring to our task, we have on one hand textual information about word wide available technological R&D-projects (in further this will be called “raw information”). This information contains a lot of new technological ideas and inventions. New means, that ideas and inventions are unknown to the user [4]. On the other hand we have descriptions about own R&D-projects. This represents our “context information”. Ideas and inventions in the context information are already known to the user.

Raw information and context information contain different structures and formats. To create a text mining approach for finding new ideas and inventions inside the raw information we have to create a common structure for raw and context information first. This is necessary for the comparison between raw and context information e.g. to distinguish new (that means unknown) ideas and inventions from known ideas and inventions.

In brief, we have to do 2 steps:
1. Create a common structure for raw and context information.
2. Create a text mining approach for finding new ideas and inventions inside the raw information.
Below we describe step 1 and 2 in detail.

2. A common structure for raw and context information

Raw information is stored in research papers, articles in technical periodicals, reports, documents, databases, web pages etc. That means raw information contains a lot of different structures and formats. Converting all structures and formats to a common structure and format for raw and context information costs a lot of work. Therefore our structure approach is to convert all information into plain text. That means we destroy all existing structures first and build up a new common structure second.

The new structure should appoint to the relationship between terms or term-combinations [5]. In this paper we realize this by creating sets of domain specific terms and sets of domain specific term-combinations occurring in the context of a term or a combination of terms. For the structure formulation we define the term unit as word.

First we create a set of domain specific terms.

Definition 1 Let (a text) $\mathrm{T} = [w_1,..,w_n]$ be a list of terms (words) w_i in order of appearance and let $n \in N$ be the number of terms in T and $i \in [1,..,n]$. Let $\Sigma = \{\tilde{w}_1,..,\tilde{w}_m\}$ be a set of domain specific stop terms [6] and let $m \in N$ be the number of terms in Σ.
The set of domain specific terms in text T is the relative complement "T without Σ". Therefore:
$\Omega = \mathrm{T} \setminus \Sigma$

Then, we create a set of domain specific term-combinations.

For each $w_i \in \mathrm{T}$ we create a set of domain specific terms occurring in the context of term w_i.

Definition 2 Let $l \in N$ be a context length that means the maximum distance (number of words) between w_i and w_j in a context of term w_i in text T. Let $j \in \{1,..,n\}$. Φ_i is defined as the set of domain specific terms in text T occurring in the l-length context of term w_i:

$$\mathbf{\Phi}_i := \{ w_j | (w_i, w_j \in \mathbf{\Omega}) \wedge (|i\text{-}j| \leq l) \wedge (i \neq j)\}$$

For each term–combination $\delta_1,..,\delta_\mu$ we create a set of domain specific terms occurring in the context of the term–combination $\delta_1,..,\delta_\mu$.

Definition 3 Let $\delta_p \in \Omega$ be a **domain specific** term in a term – combination with number $p \in [1,..,\mu]$. Let $\delta_1,..,\delta_\mu$ be a list of terms - in further this will be called term-combination - with $\delta_p \neq \delta_q \forall p \neq q \in [1,..,\mu]$ occur together in an l-length context in text T. Let $\mu \in N$ be the number of terms in the term–combination $\delta_1,..,\delta_\mu$. $\Xi_{\delta_1,..,\delta_\mu}$ is defined as the set of domain specific terms occur in an l-length context of the term–combination $\delta_1,..,\delta_\mu$ in text T:

$$\Xi_{\delta_1,..,\delta_\mu} := \bigcup \Phi_i \setminus \bigcup_{p=2}^{\mu} \delta_p \left| \delta_1 = w_i \wedge \bigcup_{p=2}^{\mu} \delta_p \subset \Phi_i \right.$$

For each term–combination $\delta_1,..,\delta_\mu$ we will create a set of domain specific terms occurring in the context of the term–combination $\delta_1,..,\delta_\mu$.

The set T could be a) the textual raw information or b) the textual context information. As result we get in case of a) $\Xi^{raw}_{\delta_1,..,\delta_\mu}$ and in case of b) $\Xi^{context}_{\delta_1,..,\delta_\mu}$.

Definition 4 To identify terms or term-combinations in the raw information also occur in the context information – that means the term-combinations are known to the user - we define $\Xi^{known}_{\delta_1,..,\delta_\mu}$ as the set of terms occur in $\Xi^{raw}_{\delta_1,..,\delta_\mu}$ and $\Xi^{context}_{\delta_1,..,\delta_\mu}$:

$$\Xi^{known}_{\delta_1,..,\delta_\mu} = \Xi^{raw}_{\delta_1,..,\delta_\mu} \cap \Xi^{context}_{\delta_1,..,\delta_\mu}$$

3. Relevant aspects for the text mining approach from technique philosophy

Our text mining approach follows the statements of technique philosophy [8]. Below we describe some relevant aspects of the statements and some conclusions special for our text mining approach.

a) A technological idea or invention represents not only a new mean, but a new purpose and mean combination. That means to find an idea or invention it is necessary to identify a mean and an appertaining purpose in the raw information. Appertaining means that purpose and mean shall occur together in an l-length context. Therefore, for our text mining approach, we firstly want to identify a mean and secondly we want to identify an appertaining purpose or vice versa.

b) Purposes and means can be exchanged. That means a purpose can become a mean in a specific context and vice versa. Example: A raw material (mean) is used to create an intermediate product (purpose). The intermediate product (mean) is then used to produce a product (purpose). In this example, the intermediate product changes from purpose to mean because of the different context. Therefore, for our text mining approach it is possible to identify textual patterns representing means or purposes. However, it is not possible to distinguish between means and purposes without the knowledge of the specific context.

c) A purpose or a mean is represented by a technical term or by several technical terms. That means a mean (a purpose) respectively can be represented by a combination of domain specific terms ($\delta_1,..,\delta_\mu$) occur together with an appertaining purpose (mean) in an l-length context. The purpose-mean combination therefore is a combination of 2 term-combinations and it also occurs in an l-length context like described in 3 a). For the formulation a term combination $\delta_1,..,\delta_\mu$ represents a mean (a purpose) only if $\Xi^{raw}_{\delta_1,..,\delta_\mu} \neq \varnothing$, which means there are further domain-specific terms representing a purpose (a mean) occurring in an l-length context together with the term-combination $\delta_1,..,\delta_\mu$ in the raw information.

d) To find an idea or invention that is really new to the user, the purpose-mean combination must be unknown to the user. That means a mean and an appertaining purpose in the raw information must not occur as a mean and an appertaining purpose in the context information. For the formulation the term combination $\delta_1,..,\delta_\mu$ represents a mean (a purpose) in a new idea or invention only if $\Xi^{known}_{\delta_1,..,\delta_\mu} = \varnothing$, which means there are no further domain-specific terms occurring in an l-length context together with the term-combination $\delta_1,..,\delta_\mu$ in the raw and in the context information.

e) To find an idea or invention that is comprehensible to the user, either the purpose or the mean must be known to the user. That means one part (a purpose or a mean) of the new idea or invention is known to the user and the other part is unknown. The user understand the known part because it is also a part of a known idea or invention that occurs in the context information and therefore, he gets an access to the new idea or invention in the raw information.

That means the terms representing either the purpose or the mean in the raw information must occurred as purpose or mean in the context information. For the formulation the term combination $\delta_1,..,\delta_\mu$ represents a mean (a purpose) in a comprehensible idea or invention only if $\Xi^{context}_{\delta_1,..,\delta_\mu} \neq \varnothing$, which means $\delta_1,..,\delta_\mu$ is known to the user and there are further unknown domain-specific terms representing a purpose (a mean) occurring in an l-length context together with the term-combination $\delta_1,..,\delta_\mu$ in the context information.

f) Normally an idea or an invention is useful for a special task. Transferring an idea or an invention to a different task makes it sometimes necessary that the idea or invention has to be changed to become useful for the new task. To change an idea or invention you have to change either the purpose or the mean. That is because the term combination $\delta_1,..,\delta_\mu$ must not change otherwise it will become unknown to the user and then the idea or invention is not comprehensible for the user like described in 3 e).

g) After some evaluation we get the experience that for finding new ideas and inventions the number of known terms (e.g. representing a mean) and the number of unknown terms (e.g. representing the appertaining purposes) shall be well balanced. Example: one unknown term among many known terms often indicates that an old idea got a new name. Therefore the unknown term is probably not a mean or a purpose. That means the probability that $\delta_1,..,\delta_\mu$ is a mean or a purpose increases when μ is close to the cardinality of $\Xi^{raw}_{\delta_1,..,\delta_\mu}$.

h) There are often domain specific stop words (like better, higher, quicker, integrated, minimized etc.) occur with the new ideas and inventions. They point to a changing purpose or a changing mean and can be an indicator for new ideas and inventions.

i) An identified new idea or invention can be a basis for further new ideas and inventions. All ideas and inventions that are similar to the identified new idea and invention are also possible new ideas and inventions.

4. A text mining approach for finding new ideas and inventions

In this paper we want to create a text mining approach by realizing point 3. a) to 3. g). Further we want to proof the feasibility of our text mining approach.

Firstly, we want to identify a mean and secondly we want to identify an appertaining purpose below as described in 3. a). The other case - identify a purpose first and identify an appertaining mean second – is trivial because of the purpose-mean dualism described in 3. b).

Definition 5 We define p($\Xi^{raw}_{\delta_1,..,\delta_\mu}$) as the probability that the term-combination $\delta_1,..,\delta_\mu$ in the raw information is a mean. That means whether μ is close to the cardinality of $\Xi^{raw}_{\delta_1,..,\delta_\mu}$ or not like described in 3 g):

$$p(\Xi^{raw}_{\delta_1,..,\delta_\mu}) = \begin{cases} \dfrac{\left|\Xi^{raw}_{\delta_1,..,\delta_\mu}\right|}{\mu} & \mu \geq \Xi^{raw}_{\delta_1,..,\delta_\mu} \\ \dfrac{\mu}{\left|\Xi^{raw}_{\delta_1,..,\delta_\mu}\right|} & \mu < \Xi^{raw}_{\delta_1,..,\delta_\mu} \end{cases}$$

The user commits to a minimum probability p_{min} .

For the text mining approach the term-combinations $\delta_1,..,\delta_\mu$ are means only if

a) $\Xi^{raw}_{\delta_1,..,\delta_\mu} \neq \varnothing$ as described in 3 c)

b) $\Xi^{context}_{\delta_1,..,\delta_\mu} \neq \varnothing$ as described in 3 e) to get a comprehensible idea or invention

c) $\Xi_{\delta_1,..,\delta_\mu}^{known} = \emptyset$ as described in 3 d) to get a new idea or invention

d) $p(\Xi_{\delta_1,..,\delta_\mu}^{raw}) \geq p_{\min}$ as described in 3 g)

For each of these term-combinations we collect all appertaining purposes (that means the combinations of all further terms), which occur in an l-length context together with $\delta_1,..,\delta_\mu$ in the raw information.

We present each $\delta_1,..,\delta_\mu$ as a known mean and all appertaining unknown purposes to the user. The user selects the suited purposes for his task or he combines some purposes to a new purpose. That means he changes the purpose to become useful for his task like described in 3. f).

Additionally it is possible that the user changes known means to known purposes and appertaining purposes to appertaining means like described in 3 b) because at this point the user gets the knowledge of the special context.

With this selection, the user gets the purpose-mean combination that means he gets a new idea or invention. This idea or invention is comprehensible for him because of 3 d) and it is novel for him because of 3. c). Further it is useful for his application because the user selects the suited purposes for his task.

5. Evaluation and Outlook

We have done a first evaluation with a text about R&D-projects from the USA as raw information [2], a text about own R&D-projects as context information [10], a stop word list created for the raw information and the parameter values l=8 and $p_{\min}$=50%. The aim is to find new, comprehensible, and useful ideas and inventions in the raw information. According to human experts the number of these relevant elements - the so-called "ground truth" for the evaluation - is eighteen. That means eighteen ideas or inventions can be used as basis for new R&D-areas. With the text mining approach, we extracted about fifty patterns (retrieved elements) from the raw information. The experts have evaluated the patterns. Thirteen patterns are new, comprehensible, and useful ideas or inventions that means thirteen from fifty patterns are relevant elements. Five new, comprehensible, and useful ideas or inventions are not found by the text mining approach. Therefore, as result we get a precision value of about 26 % and a recall value of about 72 %. This is not representative because of the small number of relevant elements but we think this is above chance and it is sufficient to prove the feasibility of the approach.

For future work, firstly we will enlarge the stop word list to a general stop word list for technological texts and optimize the parameters concerning the precision and recall value. Secondly, we will enlarge the text mining approach with further thoughts e.g. the two thoughts described in 3 h) and 3 i). The aim of this work shall be to get better results for the precision and recall value. Thirdly, we will implement the text mining approach to a web based application. That will help the users to find new, comprehensible, and useful ideas and inventions with this text mining approach. Additionally with this application, it will be easier for us to do a representative evaluation.

Acknowledge

This work was supported by the German Ministry of Defense. We thank Joachim Schulze for his constructive technical comments and Jörg Fenner for helping collect the raw and context information and evaluate the text mining approach.

Bibliography

[1] Feldman, R. and Dagan, I. (1995). Kdt - knowledge discovery in texts. In: Proceedings of the First International Conference on Knowledge Discovery (KDD). Montreal, 112–113.

[2] Fenner, J. and Thorleuchter, D. (2006). Strukturen und Themengebiete der mittelstandsorientierten Forschungsprogramme in den USA. Fraunhofer INT's edition, Euskirchen, 2.

[3] Hotho, A. (2004). Clustern mit Hintergrundwissen. Univ. Diss., Karlsruhe, 29.

[4] Ipsen, C. (2002). F&E-Programmplanung bei variabler Entwicklungsdauer. Verlag Dr. Kovac, Hamburg, 10.

[5] Kamphusmann, T. (2002). Text-Mining. Symposion Publishing, Düsseldorf, 28.

[6] Lustig, G. (1986). Automatische Indexierung zwischen Forschung und Anwendung. Georg Olms Verlag, Hildesheim, 92.

[7] Ripke, M. and Stöber, G. (1972). Probleme und Methoden der Identifizierung potentieller Objekte der Forschungsförderung. In: Paschen, H., Krauch, H. (Eds.). Methoden und Probleme der Forschungs- und Entwicklungsplanung. Oldenbourg, München, 47.

[8] Rohpohl, G. (1996). Das Ende der Natur. In: Schäfer L., Sträker E. (Eds.). Naturauffassungen in Philosophie, Wissenschaft und Technik. Bd. 4, Freiburg, München, 151.

[9] Strube, G. (2003). Menschliche Informationsverarbeitung. In: Görz, G., Rollinger C.-R., Schneeberger, J. (Eds.). Handbuch der Künstlichen Intelligenz. 4. Auflage, Oldenbourg, München, 23–28.

[10] Thorleuchter, D. (2007). Überblick über F&T-Vorhaben und ihre Ansprechpartner im Bereich BMVg. Fraunhofer Publica, Euskirchen, 2–88.

Chapter IV

Mining Ideas from Textual Information

This chapter is based on Thorleuchter, D., Van den Poel, D., Prinzie, A. Mining Ideas from Textual Information. Expert Systems with Applications 37 (10), 2010, 7182-7188.

Table of Contents

Mining Ideas from Textual Information

Dirk Thorleuchter[a], Dirk Van den Poel[b], and Anita Prinzie[b]

[a] Fraunhofer INT, D-53879 Euskirchen, Appelsgarten 2, Germany & PhD Candidate, Ghent University, dirk.thorleuchter@int.fraunhofer.de
[b] Ghent University, Faculty of Economics and Business Administration, B-9000 Gent, Tweekerkenstraat 2, Belgium, dirk.vandenpoel@ugent.be, anita.prinzie@ugent.be

Abstract

This approach introduces idea mining as process of extracting new and useful ideas from unstructured text. We use an idea definition from technique philosophy and we focus on ideas that can be used to solve technological problems.

The rationale for the idea mining approach is taken over from psychology and cognitive science and follows how persons create ideas. To realize the processing, we use methods from text mining and text classification (tokenization, term filtering methods, Euclidean distance measure etc.) and combine them with a new heuristic measure for mining ideas.

As a result, the idea mining approach extracts automatically new and useful ideas from a user given text. We present these problem solution ideas in a comprehensible way to support users in problem solving. This approach is evaluated with patent data and it is realized as a web-based application, named 'Technological Idea Miner' that can be used for further testing and evaluation.

Keywords

Idea Mining, Text Mining, Text Classification, Technology

1. Introduction

1.1. Overview

An idea is an image existing or formed in the mind but it can be written down as textual information. In the last years, we see a continually increasing amount of information. About 80 % off all this information is stored in textual form [9]. Examples are research papers, articles in technical periodicals, reports, documents, web pages etc. These texts possibly contain many new ideas. A new idea is often needed to discover unconventional approaches e.g. to create a technological breakthrough. However, a manual extraction of new ideas from these masses of texts is time consuming and costly. Therefore, it is useful to search for new problem solution ideas automatically.

Text mining or knowledge discovery from texts refers generally to the process of extracting interesting information and knowledge from unstructured text [13]. Referring to this, we introduce idea mining as an automatic process of extracting new and useful ideas from unstructured text using text-mining methods.

Creating ideas is a well-known topic that is related to psychology and cognitive science. There, we find many approaches dealing with how persons create ideas especially for problem solution. Therefore, in Sect. 2 we focus on a general process of creating problem solution ideas and use it as rationale for the idea mining approach.

In recent years, data and text mining techniques explore and analyze
huge amounts of available textual data [4]. Idea mining uses known methods from these techniques and combine them with a new method to create text patterns and a new heuristic measure for mining ideas to realise the rationale. Therefore, we present the processing of the idea mining approach in Sect. 3 and we introduce this new idea mining measure in Sect. 4.

A further task of idea mining is to present the extracted ideas in a comprehensible way to the user. Therefore, we focus on results of comprehensibility research and their relations to our task (see Sect. 5). Additionally, we provide an extensive evaluation to show the success of the idea mining approach and specifically the heuristic idea mining measure (see Sect. 6).

1.2. Idea Definition

We limit our approach to the technological language because of two reasons. Firstly, the technological language is much more standardized than the colloquial language [11,16]. Therefore, we get better results by analyzing technological texts with text mining approaches. Secondly, our idea definition is taken over from technique philosophy [25]. There, an idea is defined as a combination of two things: a mean and an appertaining purpose. An example for an idea is a transistor. A transistor is a semiconductor device. It can be used to amplify or switch electronic signals. Here, we have a mean (a semiconductor device) and an appertaining purpose (to amplify or switch electronic signals).

In general, we talk about a new idea if a known mean is related to an unknown purpose or if a known purpose is related to an unknown mean [1]. Then, a new idea is a nanomagnet because a nanomagnet is a miniaturized magnet that also can be used to amplify or switch electronic signals. Here we have an unknown mean (a miniaturized magnet) appearing together with a known purpose. This new idea could be useful to humans who are working in the field of electronic signals because in future nanomagnetic technology possibly could replace transistor technology.

Therefore, we define a new and probably useful idea as a text phrase. This text phrase consists of domain specific terms that occur together in textual information. These terms can be divided up into two subsets. The first subset should represent a known mean (or a known purpose) and the second subset should represent an unknown purpose (or an unknown mean). Additionally, all terms in the first subset should occur together in a text phrase of the technological problem description.

2. Rationale behind Idea Mining

Creating ideas is a well-known topic that is related to creativity in psychology and cognitive science. One of the first descriptions of the creative process was published by Wallas [27]. His stage model explains creative insights and illuminations for finding a problem solution. This model consists of a four stages process. In stage one 'preparation', the problem is analyzed so that a person recognizes the problem's dimensions. The stage two 'incubation / intimation' and the stage three 'illumination' transfer the problem from the conscious to the unconscious mind. The unconscious mind works on the problem continuously and it probably finds a solution by creative insights and illuminations. This solution is transferred to the conscious mind, which means after some time the person suddenly gets an idea that is new for him and that probably solves the problem. In the last stage 'verification', the idea is tested for novelty and usefulness.

One of the best-known pragmatic approaches of using practical creativity is brainstorming from Osborn [21]. The first step in brainstorming is to define the problem e.g. by creating descriptions of the problem. Then, persons generate new ideas using creativity methods like idea association etc. The last step in the brainstorming process is to cluster the generated ideas and to evaluate it for novelty and usefulness.

Beside this, there are several further approaches dealing with the creation of new ideas. We can learn from all these approaches that for creating ideas three steps are necessary. The first step is to focus on a problem, the second step is to generate some new ideas specific for this problem with creative methods and the third step is to evaluate the generated ideas for novelty and usefulness concerning the problem.

Referring to these approaches, we build an adequate rationale for the idea mining process. Therefore, idea mining also consists of three steps. In the first step, we focus on the problem. Here, the user of our idea mining approach has to provide textual information where he describes his specific problem (a problem description). In the second step, the user has to provide further textual information where he supposes the existence of new and useful ideas (a new text) that probably can solve his problem [24]. Ideas are contained in text phrases inside this new text as described in Sect. 1.2. Therefore, with an automatic process, we extract a very large number of overlapping text phrases from the new text. In the remainder of this paper, text phrases will be named text patterns. In the third step, all extracted text patterns are evaluated for novelty and usefulness. This means, they are compared to the problem description by using a specific idea mining measure. With this measure, text patterns can be classified as new and useful idea. Therefore, idea mining identifies new and useful ideas in three steps:

1. Preparation of a problem description
2. Extraction of text patterns from a new text and
3. Evaluation of text patterns for novelty and usefulness concerning problem description.

3. Idea Mining Process

Fig. 1 shows the processing of the idea mining approach in different steps based on the rationale for the idea mining process (see Sect. 2).

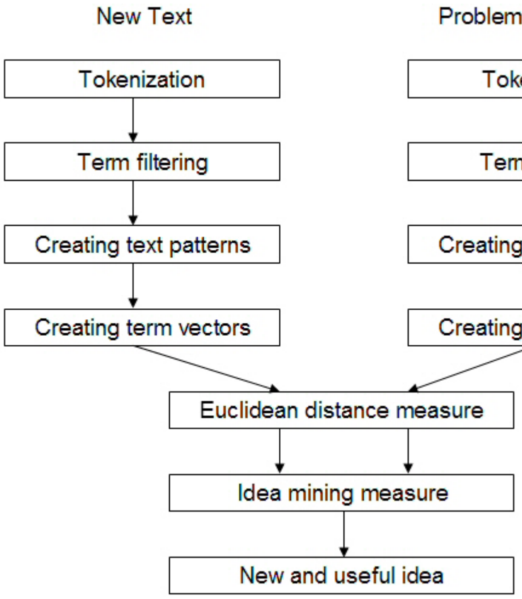


Figure 1: Processing of our idea mining approach in different steps: After tokenization and term filtering, text patterns are created and term vectors are built representing these text patterns. Term vectors from the new text are compared to term vectors from the problem description using the Euclidean distance measure. Then, term vectors from the new text are compared to their most similar term vectors from the problem description using the idea mining measure. As a result, we get term vectors from the new text that represent new and useful ideas.

With tokenization [3], texts are separated in terms and the term unit is word. The set of different terms in a text is reduced by using stop word filtering methods and stemming [13]. For this, a general list of stop words is used as well as the well-known Porter stemming algorithm [22].

A related problem to the use of stemming is to identify synonyms and homonyms. Synonyms are different words with identical or at least similar meanings. Homonyms are groups of words with the same spelling but with different meanings. With stemming synonyms and homonyms cannot be identified because stemming does not use knowledge of the context of a term. In this idea mining approach, we do not identify synonyms and homonyms. This is because the approach always considers the context of a term by working on text patterns containing several co-occurring terms as described below.

Here, we show how to create these text patterns automatically. Around each appearance of each term in the new text, we create a text pattern containing the selected term and all terms, which occur in the left and right context of the selected term. To reduce the number of text patterns, we only create text patterns around non-stop words and around terms that occur both in the new text and in the problem description.

One important decision to be taken is to determine the length of a text pattern. Text patterns should not be too small so that they contain all terms representing a new idea. Further text patterns should not be too large so that only terms occur in the text patterns that are related to the new idea. For example if we set the length of the text patterns to l then a text pattern contains the selected term, l terms from its left context and also l terms from its right context. The cardinality of the set of stop word filtered and stemmed terms from this pattern is normally smaller than $2 * l + 1$because some terms are stop words, some terms occur twice and some terms have the same stem.

In this paper, we do not use a constant length l for all patterns but a variable length of text patterns based on a dynamic adaptation of its context. This is realized by using a term weighting scheme based on the difference between stop words and non-stop words because the importance of a stop word in a text pattern is not as high as the importance of a non-stop word. If an author formulates an idea very briefly by joining catchwords together then he normally does not use many stop words and the text pattern length can be small. If an author formulates an idea in a flowery style that means his writing is not expressed in a clear and simple way then he normally uses more stop words and the text pattern length has to be larger. In the idea mining application the value of text pattern length l and the percentage of the importance of stop words u and of non-stop words v can be provided by the user.

To compute the variable length of a text pattern, we firstly define the term weighting scheme.

Definition 1. Let (a text) $T = [w_1,..,w_n]$ be a list of terms (words) w_i in order of appearance and let $n \in N$ be the number of terms in T and $i \in [1,..,n]$. Let $\Sigma = [\tilde{w}_1,..,\tilde{w}_m]$ be a set of domain specific stop terms [18] and let $m \in N$ be the number of terms in Σ. Let the percentage u be a term weighting coefficient for stop words. Let the percentage v be a term weighting coefficient for non-stop words. Then, we define $f_g(w_i) \in N$ as term weighting scheme:

$$f_g(w_i) = \begin{cases} u \,|\, w_i \in \Sigma \\ v \,|\, w_i \notin \Sigma \end{cases} \qquad (\forall i \in \{1,..,n\}) \tag{1}$$

We give an example for this. The text pattern 'components for frequency conversion of infrared lasers' is built around the word 'conversion'. It contains the word conversion itself, three terms from its left context (components for frequency), and three terms from its right context (of infrared lasers). Here, we use a constant length $l = 3$ and a term weighting scheme with $\alpha = \beta =$ 100 %. This means the importance of a stop word is equal to the importance of a non-stop word. The next text pattern is an example for a variable length: 'In a 1st phase, known but so far not available materials and technologies such as layer systems and crystals'. This text pattern is built around the word 'technologies'. Here we use a constant length $l = 3$ and a term weighting scheme with $u =$ 10 % and $v = 100$ %. As a result, this text pattern contains six terms from the right context and eleven terms from the left context of the term 'technologies'. In this example, non-stop words are phase, materials, technologies, layer, systems, and crystal. We compute the number of terms from the left and right context as described below:

Definition 2. Let $l \in N$ be a constant length of text patterns. Let l_i^{left} be the number of terms from the left context of a text pattern that is built around the term w_i. Let l_i^{right} be the number of terms from the right context of a text pattern that is built around the term w_i. Then, we define $l_i^{left} \in N$ and $l_i^{right} \in N$ as:

$$l_i^{right} = \min_j \left| (\sum_{k=1}^{j} f_g(w_{i+k}) \geq l) \vee (i + j = n) \right. \qquad \forall i \in \{1,..,n\} \tag{2}$$

$$l_i^{left} = \min_j \left| (\sum_{k=1}^{j} f_g(w_{i-k}) \geq l) \vee (i - j = 1) \right. \qquad \forall i \in \{1,..,n\} \tag{3}$$

After computing l_i^{left} and l_i^{right}, we can build a text pattern T_i around the term w_i from the text $T = [w_1,..,w_n]$.

$$T = [w_{i-l_i^{left}}, .., w_i, ..., w_{i+l_i^{right}}] \tag{4}$$

For each text pattern from the new text, we create a term vector in vector space model. The size of the vector is defined by the number of different stemmed and stop word filtered terms in the new text. For text pattern encoding, we use binary term vectors that means a vector element is set to one if the corresponding unstemmed term is used in the text pattern and to zero if the term is not. We also build text patterns from the problem description and create term vectors as described above.

To identify new and useful ideas, we create a specific idea mining measure. This idea mining measure is described in Sect. 4. By comparing a vector from the new text to one from the problem description, we can compute a result value always between 0 % and 100 % using this measure. The greater the result value the more is the probability that the vector from the new text represents a new and useful idea concerning a vector from the problem description.

We use this measure for comparing vectors from the new text to their most similar vectors from the problem description but not to all vectors. This is because result values from comparing a vector to its most similar vectors predominate result values from comparing a vector to its further vectors. For example if a vector from the new text is similar to one from the problem description then the idea is not new to the user regardless whether result values from comparing this vector to further vectors from the problem description are greater than zero. Therefore, we can be sure that a vector represents a new and useful idea only if it gets a great result value from idea mining measure concerning one of its most similar vectors. Further, the computing of the idea mining measure is time consuming. Therefore, it is necessary to limit the number of comparisons with idea mining measure for implementing an idea mining application.

We choose a two-step classification way. In the first step, we compare each vector from the new text to all vectors from the problem description by using the well-known Euclidean distance measure. Fortunately, the computing of the Euclidean distance measure is not time consuming so that it is suited for implementing in an idea mining application. In detail, for each vector from the new text, we identify all vectors from the problem description where the Euclidean distance result value is the lowest that means we identify the most similar vectors. In the second step, we compare each vector from the new text to its most similar vectors using the idea mining measure.

Each vector from the new text - that is compared to several similar vectors - gets the highest result value from idea mining measure as result value. To identify a new and useful idea we use alpha-cut method. An alpha-cut of the idea mining measure result value is the set of all vectors from the new text such that the appertaining result value is greater than or equal to alpha ($\tilde{\alpha}$). In the idea mining application, the user can provide the value of $\tilde{\alpha}$.

4. Idea Mining Measure

With the idea mining measure, we compare a vector that represents a text pattern from the new text to its most similar vectors from the problem description to identify a new and useful idea inside the text pattern from the new text. In detail, we have to find text pattern from the new text where all terms representing a mean (purpose) and no terms representing a purpose (mean) occur in a text pattern from the problem description.

If all terms in the text pattern from the new text are known, which means all terms also occur in a text pattern from the problem description then the idea is not new to the user. Furthermore, the idea is not useful if all terms in the text pattern from the new text are unknown because there is no relation to the problem. It is shown in [26] that to find new and useful ideas the number of known terms (e.g. representing a mean) and the number of unknown terms (e.g. representing an appertaining purpose) shall be well balanced.

Definition 3. Let α_i be a set of stemmed and stop word filtered terms representing a text pattern with number i from the new text. Let β_j be a set of stemmed and stop word filtered terms representing a text pattern with number j from the problem description. Let γ be the set of all stemmed and stop word filtered terms from the new text. Let $x = |\gamma|$ be the cardinality of γ. Let $\omega_i \in \{0,1\}^x$ be a term vector in vector space model concerning α_i. Let $\rho_j \in \{0,1\}^x$ be a term vector in vector space model concerning β_j. Let $p = |\alpha_i| = \sum_{k=1}^{x} \omega_{i,k}$ be the number of all (known and unknown) terms in text pattern with number i. Let $q = |\alpha_i \cap \beta_j| = \sum_{k=1}^{x} \omega_{i,k} \bullet \rho_{j,k}$ be the number of known terms in text pattern with number i concerning a text pattern with number j from the problem description. Then, we define m_1 as measure for well-balanced known and unknown term distribution.

$$m_1 = \begin{cases} \frac{2 \cdot (p-q)}{p} & (q \geq \frac{p}{2}) \\ \frac{2 \cdot q}{p} & (q < \frac{p}{2}) \end{cases} \tag{5}$$

The known terms in the text pattern from the new text should occur in the problem description more frequently than other terms. This is because they represent a known mean or a known purpose that is a central part of the problem. In the problem description, terms that represent the problem occur more frequently than other terms. For this, we define these frequent terms by using a percentage z as parameter and we compute m_2 as the number of known and frequent terms over the number of all known terms.

Definition 4. Let z be a percentage. Let δ be a set of z % most frequently stemmed and stop word filtered terms in the problem description. Let $\xi \in \{0,1\}^x$ be a term vector in vector space model concerning δ. Let $r = |\alpha_i \cap \beta_j \cap \delta| = \sum_{k=1}^{x} \omega_{i,k} \bullet \rho_{j,k} \bullet \xi_k$ be the number of known terms, which occur frequently in the problem description. We define m_2 as measure for frequently occurrence of known terms in the problem description.

$$m_2 = \frac{r}{q} \tag{6}$$

The unknown terms in the text pattern from the new text represent a new approach (an unknown mean or purpose), which is a central part of the new idea. These terms normally occur more frequently than other terms in the new text because this text deals about the new idea. For this, we also define these frequent terms by using a percentage z as parameter and we compute m_3 as the number of unknown and frequent terms over the number of all unknown terms.

Definition 5. Let φ be a set of z % most frequently stemmed and stop word filtered terms in the new text. Let $\tau \in \{0,1\}^x$ be a term vector in vector space model concerning φ. Let $s = |\alpha_i \cap \overline{\beta}_j \cap \varphi| = \sum_{k=1}^{x} \omega_{i,k} \bullet \tau_k - \sum_{k=1}^{x} \omega_{i,k} \bullet \rho_{j,k} \bullet \tau_k$ be the number of unknown terms, which occur frequently in the new text. We define m_3 as measure for frequently occurrence of unknown terms in the new text.

$$m_3 = \frac{s}{p-q} \tag{7}$$

There are often characteristic terms (higher, quicker, integrated, minimized etc.) that occur together with new ideas. They point to a changing purpose or a changing mean and can be an indicator for new ideas.

Definition 6. Let λ be a set of these characteristic terms (stemmed and stop word filtered). Let $\theta \in \{0,1\}^x$ be a term vector in vector space model concerning λ. Let $t = |\alpha_i \cap \lambda| = \sum_{k=1}^{x} \omega_{i,k} \bullet \theta_k$ be the number of these characteristic terms in text pattern with number i. We define m_4 as measure for changing means and purposes.

$$m_4 = \begin{cases} 1 & (t > 0) \\ 0 & (t = 0) \end{cases} \tag{8}$$

The idea mining measure bases on all four heuristic sub measures.

Definition 7. Let $h \in \{1,..,4\}$ and let $g_h \geq 0$ be weighting factors with $\sum_{h=1}^{4} g_h = 1$. Let the idea mining measure be the sum of all four sub measures multiplied by weighting factors g_h in case of $p \neq q$.

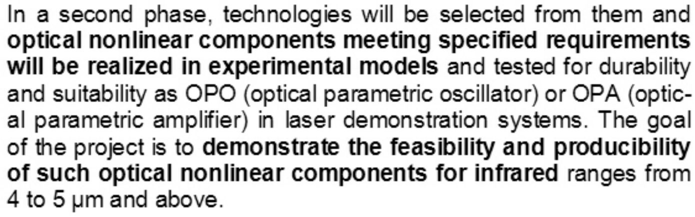
$$m = \begin{cases} g_1 m_1 + g_2 m_2 + g_3 m_3 + g_4 m_4 & (p \neq q) \\ 0 & (p = q) \end{cases} \tag{9}$$

5. Idea Mining and Comprehensibility Research

The aim of idea mining is to find new and useful ideas but also to present these ideas in a comprehensible way to the user. To realize this, we focus on comprehensibility research.

Up to the 1960s comprehensibility was a property of the text. It was measured in an objective way by analysing text parameters like word length, sentence length, word-usability, relationship between number of different words and number of words. The well-known approach in this time was the 'Reading Ease'-formula from Flesch [8].

Later research in this field focuses on cognitive effects by doing textual production and reception. The results of this research are presented by two approaches: the 'Hamburger Verständlichkeitsmodell' [17] and the 'Groebener Modell' [11]. Both approaches describe four dimensions of comprehensibility: simplicity, structure-organization, brevity-shortness and interest-liveliness.



In a second phase, technologies will be selected from them and **optical nonlinear components meeting specified requirements will be realized in experimental models** and tested for durability and suitability as OPO (optical parametric oscillator) or OPA (optical parametric amplifier) in laser demonstration systems. The goal of the project is to **demonstrate the feasibility and producibility of such optical nonlinear components for infrared** ranges from 4 to 5 µm and above.

Figure 2: We present the new text back to the user with text patterns in bold print that represent new and useful ideas.

A further approach from cognition research is named text excerption. If a human expert finds new and useful ideas in texts he highlights all corresponding text phrases e.g. with text marking. This behaviour is described by Puppe et al. [23].

In the idea mining application, text excerption is used to present the extracted ideas to the user (Fig. 2 shows an example). For the 'Groebener Modell' marking text pattern is important for structure-organization and this leads directly to comprehensibility. In this point, there are differences between the 'Groebener Modell' and the 'Hamburger Verständlichkeitsmodell' in which structure-organization is not so important for comprehensibility.

As a result, the presentation of ideas in the idea mining application based on text excerption. It is comprehensible after the 'Groebener Modell' and it is less comprehensible after the 'Hamburger Verständlichkeitsmodell'.

6. Results and Discussions

In a study for the German Ministry of Defence (MoD), we use this approach to identify new technological ideas for the German defence research program. In detail, we have to identify new solution ideas to solve current problems in German defence based research projects. We extract new ideas from 300 descriptions of research projects granted in 2006 by the National Institute of Standards and Technology (NIST) in the United States Small Business Innovation Research (SBIR) Program. We use textual information from current defence based research projects of the German MoD as problem description. As a result, we extract several new ideas that are useful for German defence research planners and that now are used as starting point for collaboration projects or for new defence based research projects. A proper selection of these ideas is a strategic issue and - together with the weapon selection problem [5] - it has significant impacts to the efficiency of future defence systems. The results are published in [6]. Here, we show some successful examples:

A modified focal plane array technology is identified that can be used to create a detector for the far ultraviolet spectrum. It leads to an improvement of military reconnaissance. This idea is new because up to now focal plane array technology is only used in the infrared, visual and near ultraviolet area.

Further, the approach identifies personnel ultrasonic locating equipment that was originally developed to make orientation possible for fire fighters in dense smoke. It also can be used to improve the location and navigation of soldiers in urban warfare (e.g. in buildings).

Additionally, the approach shows that the use of avalanche photodiode (APD) technology can improve the internal gain and the dark current of infrared detectors. This also leads to an improvement of military reconnaissance.

This study shows that some of the automatically extracted ideas are useful for technological research planners from the German MoD. Unfortunately, the used problem description (textual information about current defence based research projects) is classified as German restricted (Verschlusssache - Nur für den Dienstgebrauch) that means it is not allowed to distribute it to the scientific community. Therefore, we cannot use the results of this study to evaluate this idea mining approach. However, a separate evaluation (see Sect. 6) is done using (unclassified) patent data that allows re-computing of the evaluation.

7. Evaluation

The idea mining measure as central point in the idea mining approach consists of four heuristic sub-measures that are not theoretically founded. Therefore, it is crucial to provide an extensive evaluation to show their success. We compare this approach to a baseline because we are not aware of other approaches for idea mining. As measure for the baseline, we use Jaccard's coefficient [7] as well-known heuristic similarity measure.

The idea mining approach is evaluated by using our idea mining application (see Sect. 8). There the web based application and all texts that are used for evaluation are presented. Additionally, we create an alternative idea mining application, based on Jaccard's coefficient instead of the idea mining measure for the sole purpose of comparison to the baseline.

For evaluation, we use patent data because in patent descriptions, we normally can find new ideas, which include a considerable part of scientific and technological knowledge [16]. We use the abstract of a patent as new text. A patent often bases on further patents. We aggregate abstracts of theses references as problem description. Then we identify new and useful ideas from this patent concerning its patent references using the idea mining applications.

We use abstracts from 40 randomly selected patents and from their references, a general stop word list and Porter stemmer for evaluation. Then we determine the parameters of the idea mining measure (g_1, g_2, g_3, g_4, $\tilde{\alpha}$, and z) as well as the parameters for the length of the text patterns (l, u, and v).

For this, we use further patent data and their references as new text and as problem description. The results are evaluated by a human expert and compared to each single sub measure m_1, m_2, m_3 and m_4 alone. We find out that using the first sub measure alone is successful. If this sub measure is small then the corresponding text pattern normally does not contain a new and useful idea. If this sub measure is large then the probability that the text pattern contains a new idea is also high. We also find out that using the further sub measures alone is not successful. This means, they are successful only if the result value of the first sub measure is medium to high. Therefore, they only can be used in addition to the first sub measure.

The results of the second and third sub measures depend on the parameter z. This parameter is used to define frequent terms by building a set of z % most frequently stemmed and stop word filtered terms. We heuristically think that this parameter should be between 10 % and 30 % to get good sub measures. This is because if z is greater than 30 % then we probably classify several terms, which only occur once as frequent terms. If z is smaller than 10 % then we only identify high frequently terms for the set. In this case, the result values of the second and third sub measures are small regardless weather known terms occur frequently in the problem description or unknown terms occur frequently in the new text. Therefore, we determine z to the mean value (20 %). Additionally, we see that the second and third sub measure is nearly equally successful and that the fourth sub measure is less successful. Therefore, we heuristically determine the parameters of g_1 to 50 %, g_2 to 20 %, g_3 to 20 % and g_4 to 10 %.

We also have used other values to optimize the combination of these four sub measures. However, we do not find a combination that is generally superior to the selected combination. This is because the success of these value combinations depends on the quality of the user given textual information.

Then, we determine the alpha cut value $\tilde{\alpha}$ of the idea mining measure m. If the percentage $\tilde{\alpha}$ is small then we get many result items. This leads to a small precision value because many extracted text patterns do not contain a new and useful idea. If $\tilde{\alpha}$ is large then we only get a very small number of results and probably our recall value is small because we do not find most of the new and useful ideas in the new text. A human expert checks the results of several patent descriptions for an optimal value of $\tilde{\alpha}$. He gets the experience that 60 % is a good compromise. Therefore, we set $\tilde{\alpha}$ to 60 %. We also determine the alpha cut value of Jaccard's coefficient as measure for the baseline to 20 % by using the same way of evaluation as described above.

After this, we determine the length of the text patterns. The length depends on the parameter l and on $f_g(w_i)$, a term weighting scheme that is based on the difference between stop words and non-stop words (see Sect. 3). Text patterns should not be too small so that they contain all terms representing a new and useful idea. Additionally, text patterns should not be too large so that further terms occur in the text patterns that are not related to the new and useful idea. To find out an optimal size of text patterns, we create text patterns from several patent descriptions by using different values for l and for the percentages u and v. A human expert checks the different length of these text patterns for an optimal size. He gets the best results by setting the value of text pattern length l to 7 terms and the percentage u to 50 % and v to 100 % .

Then, the approach extracts automatically about 200 new ideas from the 40 randomly selected patents. To cluster these results, means and purposes are assigned to scientific categories in the science citation index and examples are presented below. Several ideas are identified that uses methods from 'Artificial Intelligence' (mean) for applications in 'Health Care Sciences and Services' (purpose). We also identify new ideas using 'Imaging Science and Photographic Technology' (mean) for 'Medical Informatics' purposes. Further ideas use techniques from 'Remote Sensing' (mean) in the field of 'Tropical Medicine' (purpose). Additionally, several ideas use 'Computer Science, Theory and Methods' (mean) for applications in 'Psychiatry' (purpose). Furthermore, methods from 'Artificial Intelligence' (mean) are used for 'Automation and Control Systems' purposes.

To evaluate these results, we use precision and recall measures commonly used in information retrieval based on true positives, false positives and false negatives. For this, we have to define the ground truth for our evaluation. Therefore, a human expert also identifies new and useful ideas from these patents manually that means without using our idea mining approach. He uses the idea definition in Sect. 1.2. This means, he checks each text pattern for finding terms representing a known mean (purpose) and terms representing an unknown purpose (mean). These results are the ground truth for the evaluation.

For each patent, we compute its precision and recall values by using the idea mining measure and by using the Jaccard's coefficient. Then, we compute the average precision and recall values. As a result, we get a precision value of 40 % and a recall value of 25 % by using the idea mining approach with the idea mining measure. A precision value of 40 % means that if the idea mining approach extracts ten text patterns then four of them represent a new and useful idea. A recall value of 25 % means that if there are four new and useful ideas in the new text then the idea mining approach extracts only one of them. In contrast to this, we get a precision value of 30 % and a recall value of 20 % by using Jaccard's coefficient. This is because in some texts Jaccard's coefficient extracts text patterns from the new text that are similar to text patterns from the problem description. This represents probably a known idea but not a new idea.

Beside Jaccard's coefficient, we also test other well-known heuristic measures like overlap-index, cosine-similarity and dice-similarity [7] as baseline. However, we get nearly the same results for the precision (30 %) and for the recall (20 %) value.

8. The Idea Mining Application

The idea mining application focus on users without extensive knowledge in the text mining field as well as on text mining experts. We give them the possibility to extract specifically problem solution ideas for their own needs using this idea mining approach. They can access to the web-based application via the internet. It is available under http://www.text-mining.info and it is programmed in perl and ruby.

A user has to provide two textual files, a problem description and a new text that probably consists of problem solution ideas. These files can be formatted in various ways e.g. as plain text, html, xml etc. However, scripting code, (html- or xml-) tags, and images are discarded that means the application extracts plain text from the provided files. Then, the user has to select the language of these texts to integrate a general stop word list of this language. The application offers general stop word lists in English, German, Dutch, Spain and French. After determining the parameters of the application the automatically extraction of new and useful ideas from the new text starts as described in the idea mining process (see Sect. 3 and Sect. 4). As a result, new ideas are presented as described in Sect. 5.

9. Conclusions and Future Research

This study shows the success of an automatic approach for finding new ideas from textual information. For this, the study transforms creativity approaches from psychology and cognitive science to text mining approaches. One main finding here is to redefine an abstract term (an idea) in a concrete way that it can be used for computing with text mining methods. In detail, it is shown that a technological idea represents a combination of a purpose and a mean and that purposes and means are defined by a combination of terms, which co-occur.

Additionally, it is shown that problems and problem solution ideas can be represented as term vectors in vector space model. For this, the study contributes a new (idea mining) measure. This measure identifies new ideas by comparing vectors that represent a problem to vectors that represent a problem solution idea. Last, it is shown that approaches from comprehensibility research can be adopted to this approach to present the new ideas in a comprehensible way to the user. As further main finding, it is demonstrated that this theoretical approach can be realized by a web-based application. The success of the idea mining measure is proved by comparing it to further heuristic measures (overlap-index, cosine-similarity and dice-similarity).

Directions for future research are given by the fact that nowadays there is a large amount of textual information available on the internet and this information probably contains many new technological ideas. Enlarging this approach to a web idea mining approach that automatically identifies problem solution ideas from the internet is an interesting topic for further research.

Additionally, the parameters of the approach can be optimized and the idea mining measure can probably be enlarged with further aspects to improve its quality that means to get better results for the precision and recall values.

A further aspect is to transform this idea mining approach to the colloquial language. For this, it is necessary that the idea definition also contains new product ideas from the consumers. Then, new product ideas can be identified to support marketing activities.

Last, the approach can be extended with innovation-related aspects. Then, extracted ideas can be classified as innovative ideas and might be used as starting point for the new product development.

Acknowledge

We thank Joachim Schulze and Jörg Fenner for constructive technical comments.

Bibliography

[1] Albers, S., & Gassmann, O. (2005). *Handbuch Technologie- und Innovationsmanagement: Strategie- Umsetzung- Controlling* (p.196). Wiesbaden: Gabler Verlag.

[2] Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. New York: ACM Press.

[3] Coussement, K., & Van den Poel, D. (2008). Integrating the voice of customers through call center emails into a decision support system for churn prediction. *Information & Management* 45, 165.

[4] Coussement, K., & Van den Poel, D. (2009). Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. *Expert Systems with Applications* 36, 6127-6134.

[5] Dagdeviren, M., Yavuz, S., & Kilinc, N. (2009). Weapon selection using the AHP and TOPSIS methods under fuzzy environment. *Expert Systems with Applications* 36, 8150.

[6] Fenner, J., & Thorleuchter, D. (2009). *Textmining-Analyse von Forschungsvorhaben des National Institute of Standards and Technology*. Euskirchen: Fraunhofer INT Edition.

[7] Ferber, R. (2003). *Information Retrieval* (p. 74-80). Heidelberg: dpunkt.verlag.

[8] Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology* 32, 221-233.

[9] Gentsch, P., & Hänlein, M. (1999). Text Mining. *WISU* 12, 1646.

[10] Goh, D., & Foo, S. (2008). *Social Information Retrieval Systems: Emerging Technologies and Applications for Searching the Web Effectively* (p. 53). Idea Group Inc (IGI).

[11] Groeben, N. (1982). *Leserpsychologie: Textverständnis - Textverständlichkeit*. Münster: Aschendorff.

[12] Hoffmann, L., Kalverkämper, H., Wiegand, H.E. (1998). *Fachsprachen - Languages for Special purposes: Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft - an international Handbook of Special-language and Terminology Research* (p. 1602). Berlin: Walter de Gruyter.

[13] Hotho, A., Nürnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining. *LDV Forum* 20(1), 19-26.

[14] Kamphusmann, T. (2002). *Text-Mining.* (p. 28). Düsseldorf: Symposion Publishing.

[15] Kao, A., & Poteet, S.R. (2006). Overview. In: Kao, A., Poteet, S.R. (Eds.), *Natural Language Processing and Text Mining* (p. 6). Berlin, Heidelberg: Springer.

[16] Li, Y.R., Wang, L.H., & Hong, C.F. (2009). Extracting the significant-rare keywords for patent analysis. *Expert Systems with Applications* 36, 5200-5204.

[17] Langer, I., Schulz v. Thun, F., & Tausch, R. (1974). *Verständlichkeit in Schule und Verwaltung*. München: Ernst Reinhardt.

[18] Lustig, G. (1986). *Automatische Indexierung zwischen Forschung und Anwendung* (p. 92). Hildesheim: Georg Olms Verlag.

[19] Manning, C. D., & Schütze, H. (1999). Foundations of Statistical Natural Language Processing (p. 35). The MIT Press.

[20] Martin-Bautista, M.J., Sanches, D., Serrano, J.M., & Vila M.A. (2004). Text Mining using Fuzzy Association Rules. In: V. Loia, M. Nikravesh, & L.A. Zadeh (Eds.), *Fuzzy Logic and the Internet* (p. 173). Berlin: Springer-Verlag.

[21] Osborn, A.-F. (1948). *Your Creative Power*. New York: C. Scribner's sons.

[22] Porter, M.F. (1980). An algorithm for suffix stripping. *Program* 14 (3), 130-137.

[23] Puppe, F., Stoyan, H., & Studer, R. (2003). Knowledge Engineering. In: G. Görz, C.R. Rollinger, & J. Schneeberger (Eds.), *Handbuch der Künstlichen Intelligenz* (p. 611). München: Oldenbourg.

[24] Ripke, M., Stöber, G. (1972). Probleme und Methoden der Identifizierung potentieller Objekte der Forschungsförderung. In: H. Paschen & H. Krauch (Eds.), *Methoden und Probleme der Forschungs- und Entwicklungsplanung* (p. 47). München: Oldenbourg.

[25] Rohpohl, G. (1996). Das Ende der Natur. In L. Schäfer, & E. Sträker (Eds.), *Naturauffassungen in Philosophie, Wissenschaft und Technik* (pp. 143-163). Freiburg, München: Alber.

[26] Thorleuchter, D. (2008). Finding Technological Ideas and Inventions with Text Mining and Technique Philosophy. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data Analysis, Machine Learning, and Applications* (pp. 413-420). Berlin: Springer-Verlag.

[27] Wallas, G. (1926). *The Art of Thought*. New York: Harcourt Brace.

Chapter V

Mining Innovative Ideas to Support New Product Research and Development

This chapter is based on Thorleuchter, D., Van den Poel, D., Prinzie, A. (2010). Mining Innovative Ideas to Support new Product Research and Development, in: H. Locarek-Junge, C. Weihs (Eds.). Classification as a Tool for Research. pp. 587-594. Springer, Berlin, Heidelberg.

Table of Contents

Mining Innovative Ideas to Support New Product Research and Development

Dirk Thorleuchter[a], Dirk Van den Poel[b], and Anita Prinzie[b]

[a] Fraunhofer INT, D-53879 Euskirchen, Appelsgarten 2, Germany & PhD Candidate, Ghent University, dirk.thorleuchter@int.fraunhofer.de
[b] Ghent University, Faculty of Economics and Business Administration, B-9000 Gent, Tweekerkenstraat 2, Belgium, dirk.vandenpoel@ugent.be, anita.prinzie@ugent.be

Abstract

Here, we present an approach for automatically identifying the innovative potential of new technological ideas extracted from textual information. The starting point of each innovation is a good and new idea. Unfortunately, a high percentage of innovations fail, which means many ideas do not have the potential to become an innovation in future. The innovation process from a new idea as starting point via research, development, and production activities through to an innovative product is very cost- and time-consuming. Thus, the objective of our work is to identify the innovative potential of new technological ideas to improve the performance of the innovation process.

We extract new technological ideas from provided textual information. We also identify innovative technology fields by analysing relationships among technologies. All identified ideas are assigned to innovative technology fields by using text mining and text classification methods. Technological ideas in these fields are presented to the user as innovative ideas.

Key Words
Idea Mining, Text Mining, Text Classification, Innovation

1. Introduction

The word innovation refers to the latin terms novus (that means new) and innovatio (that means something is newly created). An innovation includes a new idea [9] as well as its realization e.g. as innovative product that is successful in market. Thus in economical sense, we talk about innovations if the newly created object increases producer or customer value [13].

To create an innovation, an innovation process can be used. It has the aim to lead a new idea to an innovative product. Thus, the starting point of the innovation process is a new technological idea [14]. Based on this idea, a research process starts. The result is probably a prototype that is developed further in a developing process. After this developing process a production process starts and it leads to a product [4]. If this product is successful in market that means it increases producer or customer value then it is an innovative product and the idea standing behind this innovation can be defined as innovative idea. However, by use of this economical definition, we only can identify innovative ideas subsequent to the innovation process that means after they become successful products in market.

Unfortunately, the innovation process is very cost- and time-consuming [5] and a high percentage of innovations fail. Thus, the objective of our work is to identify the innovative potential of new technological ideas before selecting them as starting ideas. This probably can improve the performance of the innovation process.

2. Background

Our definition of a technological innovation is based on bibliometrical analyses as described in [15]. There, it is shown that innovations normally do not occur alone but together with several further innovations. These groups of innovations are based on a common innovation field. Innovation fields are newly appeared technologies or scientific disciplines that occur on the borders of established technologies or scientific disciplines. This means they occur between at least two technologies or scientific disciplines that are not related. A definition of possible relationships is given in Sect. 6. Thus, innovations can be classified as interdisciplinary products. The (innovative) ideas behind these innovations also are of an interdisciplinary nature and they also occur together in an innovation field.

Our idea definition derived from technique philosophy [17]. There, a technological idea consists of two things: a means and an appertaining purpose [2]. Thus, we define an idea as a text phrase. This text phrase consists of domain specific terms that occur together in textual information. These terms can be divided up into two subsets. The first subset should represent a means and the second subset should represent a purpose. An example for an idea is a nanomagnet (the means) that can be used to switch electronic signals (the appertaining purpose). This definition is used to identify interdisciplinary ideas by assigning means and purpose of an idea to different non-related, established technologies or scientific disciplines.

To classify ideas as innovative, we have to identify several interdisciplinary ideas that occur together in an innovation field. For this, we firstly have to provide technological context information containing descriptions of established technologies or scientific disciplines and we have to define their relations.

Secondly, we have to classify ideas as interdisciplinary by assigning means and purposes to established technologies or scientific disciplines that are not related. For example, if a means from a bionic idea can be assigned to biology and the appertaining purpose can be assigned to technological engineering then the bionic idea is interdisciplinary. This gives a hint that the combination of biology and technological engineering is probably an innovation field.

To be sure that it is really an innovation field, we thirdly have to find several further interdisciplinary ideas that can be assigned to the same non-related technologies or scientific disciplines combination and classify all the interdisciplinary ideas in this field as innovative ideas.

3. Process of Mining Innovative Ideas

This approach uses an existing idea mining approach [18] that supports users to identify means and purposes in text phrases (see Sect. 4). Then, we provide descriptions of scientific categories as context information (see Sect. 5). Both the means and the purpose of extracted new and useful ideas are assigned to several scientific categories by use of multi-label classification (see Sect. 7). After this, we compare each scientific category from means to each scientific category from purpose to find out relationships between them (see Sect. 6). Fig. 1 shows an example for the processing of this approach.

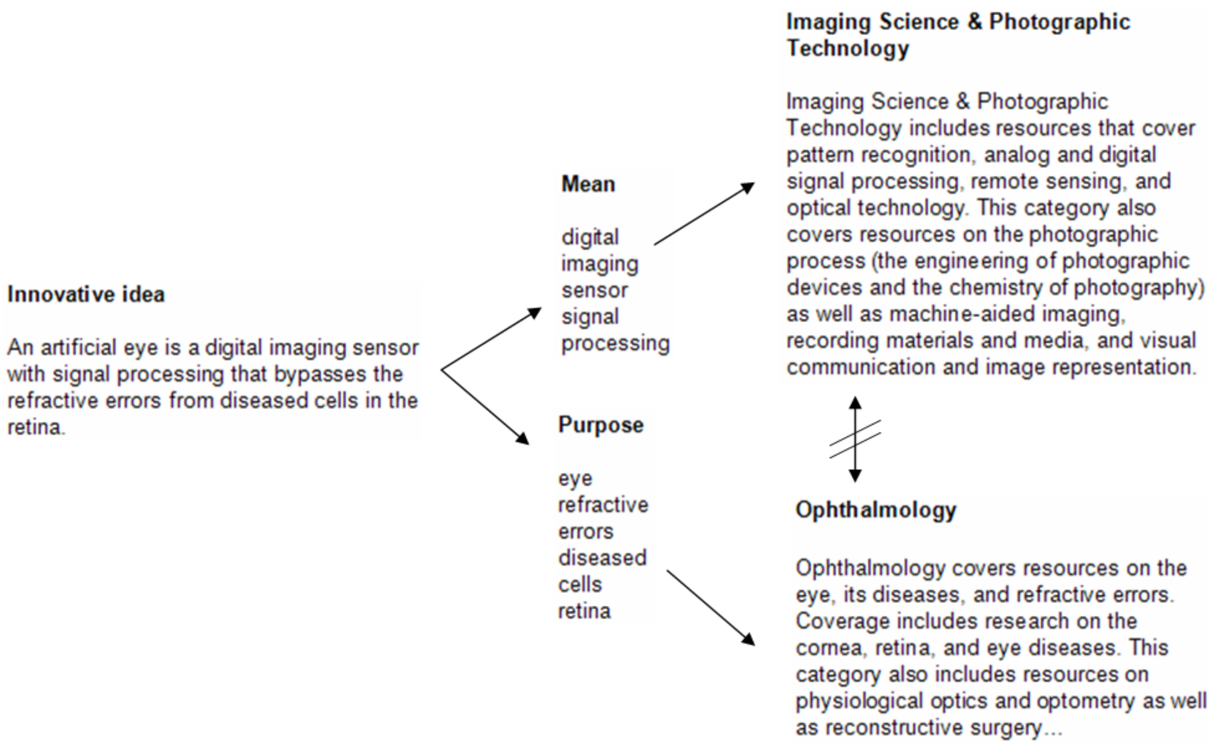

Innovative idea
An artificial eye is a digital imaging sensor with signal processing that bypasses the refractive errors from diseased cells in the retina.
Mean
digital imaging sensor signal processing
Purpose
eye refractive errors diseased cells retina
Imaging Science & Photographic Technology
Imaging Science & Photographic Technology includes resources that cover pattern recognition, analog and digital signal processing, remote sensing, and optical technology. This category also covers resources on the photographic process (the engineering of photographic devices and the chemistry of photography) as well as machine-aided imaging, recording materials and media, and visual communication and image representation.
Ophthalmology
Ophthalmology covers resources on the eye, its diseases, and refractive errors. Coverage includes research on the cornea, retina, and eye diseases. This category also includes resources on physiological optics and optometry as well as reconstructive surgery…

Fig. 1. Means and purpose are extracted from an idea and are assigned to different scientific categories. If the categories that are assigned by the means are not related to categories that are assigned by the purpose then the idea is of an interdisciplinary nature. If several ideas also are of an interdisciplinary nature concerning these categories then the combination of both categories is defined as innovation field and ideas from this field are presented as innovative ideas.

If a means is assigned to several scientific categories and the appertaining purpose is assigned to further scientific categories that are not related to any scientific category of the means then the corresponding idea is of an interdisciplinary nature. If several further ideas also are interdisciplinary concerning at least two of the above mentioned scientific categories then we define the combination of these scientific categories as innovation field. For this, the user provides the smallest number of interdisciplinary ideas that is sufficient to define such an innovation field. Ideas from these innovation fields are classified as innovative ideas.

4. Acquisition of Ideas

Our approach based on technological ideas. The user extracts them from provided textual information e.g. patent data. He is supported by a further approach that automatically extracts new and useful ideas from textual information as presented in [18]. This approach extracts textual phrases that represent new and useful ideas. Additionally for each idea, it identifies terms that represent a means as well as terms that represent the appertaining purpose. This information is used as input for our approach.

5. Acquisition of Technological Context Information

To provide technological context information, we focus on scientific categories. We can find an overview of current scientific categories in the science citation index (SCI). This index is built on bibliographic information, author abstracts, and cited references from about 3,700 science and technical journals. The content of these highly cited journals is assigned to 172 scientific categories. The official description of all categories in the SCI is available in scope notes [11] that is manually created, of good quality, and up to date. We use this description as technological context information for our approach.

6. Relationship among Scientific Categories

After providing descriptions of scientific categories that represent technological context information, the next step is to identify relationships among these scientific categories.

In general, we identify two different kinds of relationships [8]. One kind of relationship is that technologies can be similar to other technologies. They deal with the same technology field but have a different focus. The descriptions of two similar technologies also are similar because they both contain the same domain specific terms by describing the technological field.

A further kind of relationship is that technologies are related in a substitutive, integrative, predecessor or successor way. If technologies are related in this way then they deal with the same application field. Their descriptions also are similar because they both contain the same domain specific terms representing the application field.

The descriptions of the scientific categories in scope notes contain terms representing the technological field as well as terms representing potential application fields. If we identify similar terms in descriptions of two different scientific categories then both categories are related according to at least one kind of relationship. Therefore, related categories are identified by comparing category descriptions among each other.

Comparing is done by transforming category description to term vectors in vector space model. For this, terms in the descriptions are tokenized [6] by using the term unit as word, stop word filtered by using a standard stop word list [12], and stemmed [10] using a dictionary-based stemmer combined with a set of production rules [16] to give each term a correct stem. The production rules are used when a term is unrecognizable in the dictionary. Vectors representing scientific categories can be compared using similarity measures in combination with the alpha cut method [1] and two categories are classified as related if the corresponding similarity measure result value is greater than or equal to alpha. For comparing, we prefer the well-known Jaccard's coefficient measure [7] because it considers well the different sizes of both vectors.

7. Classification of Ideas

Each selected idea consists of a set of terms that represents a means and of a set of terms, which represents an appertaining purpose. To identify an interdisciplinary technological idea we have to assign both sets to scientific categories. Both sets of terms are stop word filtered and stemmed as described in Sect. 6. For multi-label classification, we transform these sets to term vectors in vector space model and compare them with term vectors from each scientific category. For comparing, we also use Jaccard's coefficient measure in combination with the alpha cut method. As a result, means and purposes are assigned to scientific categories only if the appertaining Jaccard's coefficient result value is greater than or equal to alpha.

Each means and each purpose of a new idea is probably assigned to several scientific categories. To identify relations, we compare each scientific category from means to every single scientific category from purpose as described in Sect. 6. If we cannot find any relationships then the new idea is of interdisciplinary nature and each of these scientific category combinations from means and purpose is probably an innovation field. If we identify at least n interdisciplinary ideas that can be assigned to one specific scientific category combination then we define an innovation field on this basis. The user provides the smallest number n of interdisciplinary ideas that are sufficient to define such an innovation field.

8. Results and Evaluation

We present a heuristic approach for automatically identifying the innovative potential of new technological ideas. The extraction of ideas and the identification of terms that represent means and purposes is already evaluated in [18]. Thus, the evaluation is limited to the further steps of our approach and it is based on current context information. For this, scientific categories in the science citation index as current technological information described in scope notes [11] are used.

The approach extracts 1000 new ideas from randomly selected patents because patent descriptions consist of new ideas that also are innovative. However, not all new ideas are innovative in terms of the technological innovation definition in Sect. 2. 500 ideas are used as training examples to obtain the optimal parameter values and 500 ideas are used as test set to validate and compare the model. To evaluate the results of the approach, we use precision and recall measures commonly used in information retrieval based on true positives, false positives, and false negatives. For this, the ground truth for our evaluation is defined. Therefore, a human expert classifies the 1000 new ideas as innovative or as non-innovative.

The approach depends on three parameters (n, α_1, α_2). The smallest number (n) of interdisciplinary ideas that is sufficient to define an innovation field gives a hint concerning the innovative potential of the new idea. If the number n is large then we only obtain ideas as result items that probably consist of a very high innovative potential. This is because we identify many ideas that are classified concerning a specific non-related combination of scientific categories. Here, we have a high probability that this category combination represents an innovation field. If the number n is small e.g. it equals one then we get all interdisciplinary ideas as result items regardless weather they consists of high or low innovative potential. This is because every idea - that is classified concerning a specific non-related combination of scientific categories - is presented as innovative idea. We estimate that an optimal value of n is between $4 \le n \le 8$.

After this, the alpha cut of Jaccard's coefficient results are estimated. The first alpha cut is the set of all terms that represents a means or a purpose such that the corresponding result value by comparing this set to a scientific category is greater than or equal to α_1. With the second alpha cut we identify two related scientific categories only if the appertaining Jaccard's coefficient result value is greater than or equal to α_2. If α_1 is too small or too large then means and purposes are not classified correctly. If α_2 is too small or too large then the identification of relationships among scientific categories fails. This leads both to a small precision and to a small recall value. An optimal value of α_1 and α_2 is estimated between
$5\% \le \alpha_1, \alpha_2 \le 20\%$.

To investigate the dependency of the approach on the parameters, we explicitly check if the parameter values are identifiable on the training set. These values are used to compute precision and recall on the test set. For this, we use the estimations for $n \in \{4, 5, ..., 8\}$ and the percentages $\alpha_1 \in \{5\%, 6\%, ..., 20\%\}$ and $\alpha_2 \in \{5\%, 6\%, ..., 20\%\}$. We identify 5 •16 • 16 = 1280 different parameter combinations of (n, α_1, α_2). The training set is used to compute average precision and recall for each parameter combination to identify the optimal parameter values with a maximal F-measure. The F-measure is used because precision and recall are equally important. As a result, parameter values $n = 5$, $\alpha_1 = 14\%$, and $\alpha_2 = 16\%$ are identified. These parameter values are used to compute precision and recall for each test example and the average precision and recall values for all test examples. We get a precision value of 38% and a recall value of 30%. A precision value of 38% means that if this approach predicts 100 ideas as innovative ideas then 38 of them are innovative. A recall value of 30% means that if there are 10 innovative ideas in the provided text then this approach identifies three of them.

We compare this approach to a baseline model because we are not aware of other approaches for identifying the innovative potential of ideas at the present time. A positive class probability of 5% is already calculated by human experts. This leads to a 5% precision at 30% recall for a random prediction and it shows that this approach is much better than random. We think that the results are sufficient to proof the feasibility of our approach.

Using the 500 new ideas from the test set, the approach automatically computes several innovation fields. We present examples for these innovation fields. They can be found between 'Health Care Sciences and Services' and 'Computer Science, Artificial Intelligence' (e.g. the use of methods from artificial intelligence for health care applications), between 'Imaging Science and Photographic Technology' and 'Medical Informatics', between 'Remote Sensing' and 'Tropical Medicine', and between 'Computer Science, Theory and Methods' and 'Psychiatry'. Then, the approach identifies ideas from these innovation fields as innovative ideas.

This approach can be re-evaluated by using our application for mining innovative ideas (see http://www.text-mining.info). There, the web based application that is programmed in perl/ruby and all texts that are used for evaluation are presented. The application extracts ideas from a provided text, creates terms representing means and purposes, identifies innovation fields, and classifies the ideas as (non-) innovative ideas.

9. Outlook

This work shows that the automatic identification of the innovative potential of new technological ideas is feasible using text classification and specific technological definitions. Further work should aim at enlarging and optimizing this approach e.g. by identifying further properties of innovative ideas. A second avenue of further research could take the granularity of the context information into account e.g. by using technologies rather than scientific categories. This also probably leads to an increasing precision and recall.

Bibliography

1. Abebe, A. J., Guinot, V., Solomatine, D. P. (2000). Fuzzy alpha-cut vs. Monte Carlo techniques in assessing uncertainty in model parameters. In: Proc. 4-th International Conference on Hydroinformatics. Iowa City, USA.
2. Albers, S., Gassmann, O. (2005). Handbuch Technologie- und Innovationsmanagement: Strategie- Umsetzung- Controlling. p. 196. Gabler Verlag.
3. Berth, R. (1997). Der große Innovations-Test: das Arbeitsbuch für Entscheider: Chancen erkennen, Flops vermeiden - Theorie und Praxis des Management of Change. Econ, Düsseldorf.
4. Bürgel, H.D., Haller, C., Binder, M. (1996). F&E-Management. p. 85. Vahlen, München.
5. Disselkamp, M. (2005). Innovationsmanagement: Instrumente und Methoden zur Umsetzung im Unternehmen. p. 179. Gabler Verlag.
6. Feldman, R., Sanger, J. (2007). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. p. 318. Cambridge University Press.
7. Ferber, R. (2003). Information Retrieval. p. 78. dpunkt.verlag, Heidelberg.
8. Geschka, H., Schauffele, J., Zimmer, C. (2005). Explorative Technologie-Roadmaps - Eine Methodik zur Erkundung technologischer Entwicklugslinien und Potenziale. In Möhrle, M.G., Isenmann, R. (eds.) Technologie-Roadmapping, p. 165. Springer, Berlin, Heidelberg.
9. Guiltinan, J.P., Paul, G.W. (1991). Marketing Management: Strategies and Programs. p. 196. McGraw-Hill.
10. Hotho, A., Nürnberger, A., Paaß, G. (2005). A Brief Survey of Text Mining. LDV Forum 20 (1), 19-26.
11. Institute for Scientific Information ISI (eds.) (1997) SCI Journal Citation Reports.
12. Lustig, G. (1986). Automatische Indexierung zwischen Forschung und Anwendung. p. 92. Georg Olms Verlag, Hildesheim.

13. Mckeown M. (2008). The Truth About Innovation. Pearson Education, Harlow.

14. Möslein, K.M., Matthaei, E.E. (2008). Strategies for Innovators: A Case Book of the HHL Open School Initiative. p. 13. Gabler Verlag.

15. Reiß, T. (2006). Innovationssysteme im Wandel - Herausforderungen für die Innovationspolitik. In Müller, B., Glutsch, U. (eds.) Fraunhofer-Institut für System- und Innovationsforschung - Jahresbericht 2006, p. 10. Karlsruhe.

16. Porter, M.F. (1980). An algorithm for suffix stripping. Program, 14 (3), 130–137.

17. Rohpohl, G. (1996). Das Ende der Natur. In: Schäfer, L., Sträker, E. (eds.) Naturauffassungen in Philosophie, Wissenschaft und Technik, Bd. 4, pp. 143-163. Alber, Freiburg, München.

18. Thorleuchter, D. (2008). Finding Technological Ideas and Inventions with Text Mining and Technique Philosophy. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker R. (eds.) Data Analysis, Machine Learning, and Applications, pp. 413-420. Springer, Berlin, Heidelberg.

Chapter VI

Extracting Consumers Needs for New Products
A Web Mining Approach

This chapter is based on Thorleuchter, D. Van den Poel, D., Prinzie, A. (2010). Extracting Consumers Needs for New Products - A Web Mining Approach. in: Proc. WKDD 2010, IEEE Computer Society, Los Alamitos, CA, p. 441 ff.

Table of Contents

Extracting Consumers Needs for New Products
A Web Mining Approach

Dirk Thorleuchter[a], Dirk Van den Poel[b], and Anita Prinzie[b]

[a] Fraunhofer INT, D-53879 Euskirchen, Appelsgarten 2, Germany & PhD Candidate, Ghent University, dirk.thorleuchter@int.fraunhofer.de
[b] Ghent University, Faculty of Economics and Business Administration, B-9000 Gent, Tweekerkenstraat 2, Belgium, dirk.vandenpoel@ugent.be, anita.prinzie@ugent.be

Abstract

Here we introduce a web mining approach for automatically identifying new product ideas extracted from web logs. A web log - also known as blog - is a web site that provides commentary, news, and further information on a subject written by individual persons. We can find a large amount of web logs for nearly each topic where consumers present their needs for new products. These new product ideas probably are valuable for producers as well as for researchers and developers. This is because they can lead to a new product development process. Finding these new product ideas is a well-known task in marketing. Therefore, with this automatic approach we support marketing activities by extracting new and useful product ideas from textual information in internet logs. This approach is implemented by a web-based application named Product Idea Web Log Miner where users from the marketing department provide descriptions of existing products. As a result, new product ideas are extracted from the web logs and presented to the users.

Key Words

Web Mining; Text Classification; New Product Development, Knowledge Discovery

1. Introduction

Web logs are web pages that consist of textual and non-textual information combined with links to further web logs, web pages etc. Individual persons normally write them [1]. Many web logs offer the possibility for readers to leave comments in an interactive format. We can see an increasing amount of these web logs for nearly each topic [2]. Most web logs consist of textual information. To analyze textual information tools and methods from text mining can be used. Web logs have been analyzed for marketing aspects by these text-mining tools in three different ways. Firstly, we can find out brand reputation, secondly we can measure effects of advertisements, and thirdly we can focus on consumers needs for new products [3].

In this paper, we focus on analyzing consumers' needs for new products. Many web logs deal with existing products. There, enterprises publish descriptions of existing products in the internet and offer the possibility for readers to leave comments. In general, consumers use this possibility to describe their experiences and problems with the product. However, sometimes, suggestions for new product features or even for completely new products are published.

These new product ideas are of particular interest for this approach. This is because they probably can lead to a change in the product development process. We identify product ideas that have two properties. Firstly, they have to be new. Then, an existing product does not yet contain these new ideas and new ideas are not part of product development activities. Secondly, product ideas have to be useful. They should have a relation to existing products or actual product development activities otherwise it is not possible to integrate them in the product development process. An example for this is that a manufacturer who produces coffee machines can start a new product development process for espresso machines. This is because coffee machines are related to espresso machines. However, he normally is not able to produce completely different products in the (near) future (computer, furniture etc.).

Identifying new and useful product ideas is a well-known activity in marketing [4,5]. Therefore, with an automatically process we support these marketing activities by extracting these new product ideas from web logs.

One important aspect in this approach is that consumers write new product ideas in web logs by use of a colloquial language [6]. In contrast to the technical language, we see that terms are not defined exactly and that many homonym and synonym problems occur by evaluating this textual information with text mining methods [7]. However sometimes, consumers also use technical terms if their need for new products refers to a technological product.

2. Rationale Behind this Approach

Finding these new product ideas is a well-known task in marketing. Below, we describe how marketing professionals identify these new product ideas in the internet.

A person - usually from marketing department - analyzes the situation of an enterprise so that (s)he recognizes all existing products and all product developments. We assume, that these products and product developments are already described in form of textual information. Then, the person searches the internet for new product ideas that are published by consumers. For this, (s)he specially searches web logs because there, consumers present their needs for new products. If a web log system is installed on the enterprise website then the person checks these web log comments for new product ideas. After this, he searches external web log sites where consumers provide comments concerning these products or concerning similar products e.g. competing products on the enterprise website of competitive firms.

Searching in web logs can be done by using an internet search engine and by limiting the query results to textual information from web logs. Therefore, a search query consists of several domain-specific terms. For this, the person uses terms from the description of the products and the product developments. These search queries are executed by an internet search engine.

Each retrieved document from a query result consists of a title, a short description, and an internet link. The short description contains search terms from the query in bold print and the internet link leads directly to the full text of the retrieved web log site. The person checks the title and the short description for new and useful product ideas. This normally is done in an intuitive way. If query results have promise to the person then (s)he focuses on the full text by using the query result link. There (s)he probably extracts the new product idea.

3. Methodology

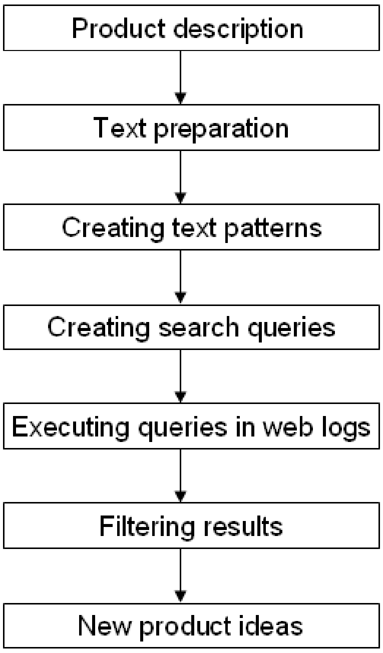


Figure 1. Processing of the web log mining approach in different steps.

We use the methodology in Fig. 1 to realize the rationale as described in Sect. 2. Therefore, this web log mining approach has the aim to support users by finding new product ideas from web logs. Below, we describe how to prepare textual information from a provided product description (see Sect. 4), how to create text patterns, and how to build search queries based on the created text patterns (see Sect. 5). Additionally, we describe how to execute search queries by use of a web log search engine and how to filter the query results to get new product ideas that are presented to the user (see Sect. 6).

4. Text Acquisition and Preparation

For this web log mining approach, a user has to provide a product description. Normally descriptions of existing products are available because persons from marketing department of an enterprise use them for marketing/public relations purposes. However sometimes, descriptions of future products that are in a product development process are not existent. Then, to extract new product ideas concerning these future products, the user first has to create these descriptions.

In a pre-processing phase, the provided product description is tokenized [8] by using the term unit as word. Additionally, we use a standard word stop list [9]. Normally in text mining applications, stop words are deleted and all other terms are used for further processing.

However, in this approach it is important to know the position of stop words in the texts for creating text patterns.

5. Text Pattern and Search Query Creating

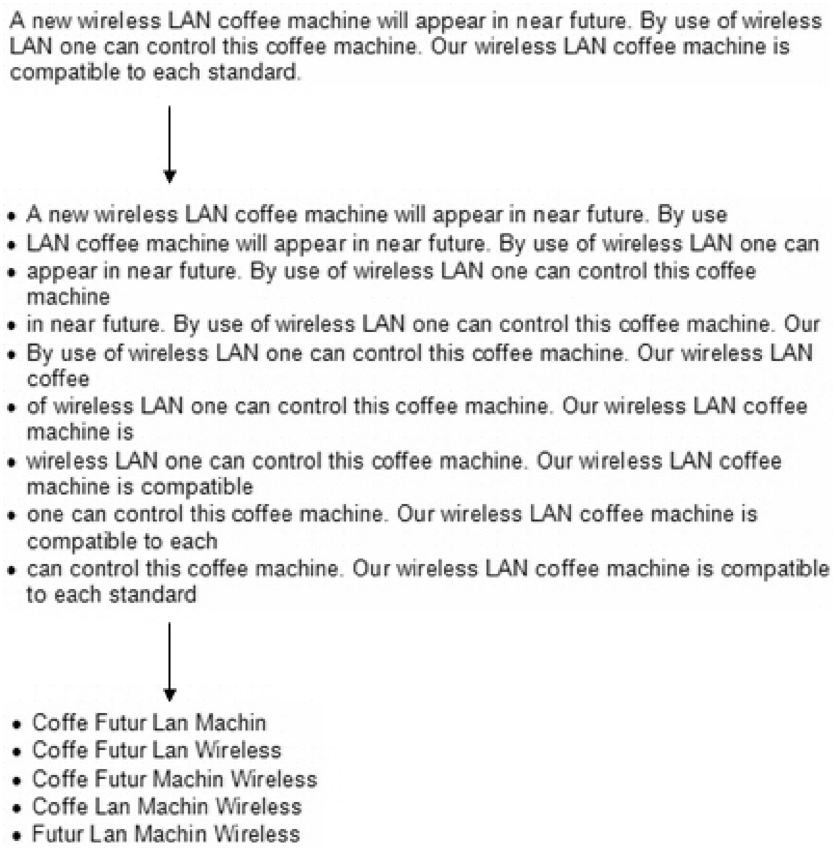
Around each term in the provided product description, we build a text pattern if the selected term is not a stop word. We compute the length of text patterns by a provided length from the user and by a user-given term weighting schema. The schema distinguishes between stop words and non-stop words because they are not equally important.

Then, we build search queries from the created text patterns. As described in the rationale, a person uses several terms from the product description to build a search query. Heuristically, we estimate the number of terms in a search query at four. This is, because if the number of terms in a search query is too large then the query is too specific and we probably do not find the relevant web log comments. If the number of terms in a search query is too small, then we probably get homonym and synonym problems, because consumers normally use colloquial language. Additionally, we get too much query results. This causes performance problems. Therefore, we think that four terms is a good compromise.

We use stemmed terms [10] to build the queries. This is because searching in a web search engine with a stemmed term leads to query results that contain several different terms, which all have the same stem. Further, we do not use stop words in the search query because normally search engines delete these terms. Therefore, we delete the stop words and all further terms are stemmed using the well-known Porter stemmer [11]. Then, we build search queries that consist of four different stemmed terms from one text pattern.

In Fig. 2 we show an example for this processing. Here we start the processing with a user-given product description. We build text patterns around each term that is not a stop word. We identify these sixteen terms in the following order of appearance: wireless, LAN, coffee, machine, future, wireless, LAN, control, coffee, machine, wireless, LAN, coffee, machine, compatible, and standard. In this example, we set the length of a text pattern to seven. Then, a text pattern consists of the selected term and seven terms from its left context as well as seven terms from its right context.

Therefore, in the first step, we create these sixteen text patterns. Text patterns that are built around the first appearance of the terms wireless, LAN, and coffee as well as text patterns that are built around the last appearance of the terms coffee, machine, compatible, and standard are contained in further text patterns. This is because text patterns at the beginning or at the end of a text are smaller than these further text patterns. They do not contain additional information. Therefore, these text patterns are discarded.



A new wireless LAN coffee machine will appear in near future. By use of wireless LAN one can control this coffee machine. Our wireless LAN coffee machine is compatible to each standard.

- A new wireless LAN coffee machine will appear in near future. By use
- LAN coffee machine will appear in near future. By use of wireless LAN one can
- appear in near future. By use of wireless LAN one can control this coffee machine
- in near future. By use of wireless LAN one can control this coffee machine. Our
- By use of wireless LAN one can control this coffee machine. Our wireless LAN coffee
- of wireless LAN one can control this coffee machine. Our wireless LAN coffee machine is
- wireless LAN one can control this coffee machine. Our wireless LAN coffee machine is compatible
- one can control this coffee machine. Our wireless LAN coffee machine is compatible to each
- can control this coffee machine. Our wireless LAN coffee machine is compatible to each standard

- Coffe Futur Lan Machin
- Coffe Futur Lan Wireless
- Coffe Futur Machin Wireless
- Coffe Lan Machin Wireless
- Futur Lan Machin Wireless

Figure 2. This example shows how text patterns are extracted from a user-provided product description.

Then we build search queries that consist of four stemmed and stop-word filtered terms, which occur together in a text pattern. To reduce the number of created search queries, we only build a search query with terms that occur frequently in the product description. For this, we compute the z % most frequently stemmed and stop-word filtered terms from the product description. In this example, we can identify eight stemmed and stop-word filtered terms. The terms wireless, LAN, coffee, and machine occur three times in the text. The terms future, control, compatible, and standard occur once. We set the parameter z to 70 % that means, we identify the five most frequent terms. For this, we select the four terms: wireless, LAN, coffee, and machine. Additionally, we also select further non-frequent terms in order of appearance. Then, the five most frequent terms are wireless, LAN, coffee, machine, and future.

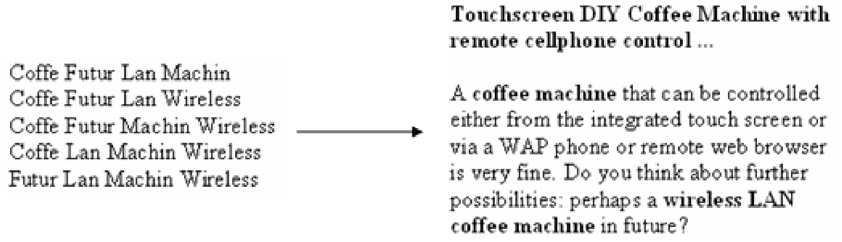
All these five terms appear together in one text pattern e.g. in the first text pattern in Fig. 2. Therefore, five search queries are built that consist of each combination of four terms.

The parameter z can be selected by the user to reduce or increase the number of created search queries. This is because large texts lead to a large number of search queries, which causes performance problems. Additionally, small texts lead to a small number of queries. Then probably, new product ideas cannot be identified.

6. Search Query Executing and Result Filtering

To execute the created search queries we use web services. A web service is a software system that is designed to support interoperable machine-to-machine interaction over a network. Frequently web services are just web based advanced programming interfaces. Access to these interfaces is possible over the internet. Then the requested service is executed and resulting data is transferred back to an application that requested the service [12]. A lot of internet search engines offer web services. By use of these web services search queries can be limited on information from web logs [13]. Therefore, we use them in our web-based application to execute queries automatically and to get the query results.

The query results consist of a title, a short description that contains terms from the search query in bold print, and a hyperlink that leads to the full text (see Fig. 3). In this approach, we identify new product ideas from the short description text pattern. However, a short description from a query result probably consists of several text patterns that are separated by several dots. In this case, we discard this query result.



Coffe Futur Lan Machin
Coffe Futur Lan Wireless
Coffe Futur Machin Wireless
Coffe Lan Machin Wireless
Futur Lan Machin Wireless

Touchscreen DIY Coffee Machine with remote cellphone control ...

A **coffee machine** that can be controlled either from the integrated touch screen or via a WAP phone or remote web browser is very fine. Do you think about further possibilities: perhaps a **wireless LAN coffee machine** in future?

Figure 3. In this example, we identify a consumer’s need for a new product: This idea deals about integrating touch screens, WAP phone interfaces, or remote web browser interfaces in the future product (the wireless LAN coffee machine) of the enterprise.

As described in Sect. 1 the identified product ideas should have two properties: novelty and usefulness [14]. Novelty means that the short description consists of information, which must not be in the product description. This means, several stop-word filtered terms from the short description should not occur in the product description. For this, we compare the stemmed and stop-word filtered terms from short descriptions to all extracted text patterns from the product description. This comparison has the aim to identify several terms from a short description that does not occur together in a text pattern from the product description. We compare the short description to a text pattern and not to the complete product description because the product description probably consists of textual information about several existing or future products from an enterprise. Then, product description consists of a large size and probably many domain-specific terms are contained in this description. In this case, a new product idea that combines several features from different products will not be assigned as a new idea.

Usefulness means that the short description consists of information, which must be in the product description because it should have a relation to the problem. This means, several stop-word filtered terms from the short description should occur in the product description. We already have identified terms from the search query as characteristic (most frequent) terms from the product descriptions. Therefore, we check for the appearance of all four terms from the corresponding search query in the short description.
Then we present the selected short description to the user as new and useful product ideas.

7. Evaluation

One important aspect in this approach is that consumers write new product ideas in web logs by use of colloquial language. In contrast to the technical language, we see that terms are not defined exactly and that many homonym and synonym problems occur by evaluating this textual information with text mining methods.

We compare this web log mining approach to a baseline model because we are not aware of other approaches for identifying new product ideas from web logs at the present time. For this, we do not use the chance baseline, which assigns a classification randomly because a high percentage of extracted query results do not represent a new product idea. Therefore, we use the frequency baseline. Here we have two classes (A means a query result represents a new product idea, B means a query result does not represent a product idea) in our data, and we classify each instance (query results) with a specific percentage as either A or B.

Additionally, consumers use colloquial language by providing comments in web logs. Therefore, we also compare this approach to an approach that identifies new ideas from the technical language.

For evaluation, we use product descriptions from several randomly selected products that are published in web logs. We create a web log mining application that realizes this approach. It is available at http://www.text-mining.info. There, the web-based application and all texts that are used for evaluation are presented. The application automatically extracts new product ideas from web logs and presents them to the user.

To evaluate the results of our approach, we use precision and recall measures commonly used in information retrieval based on true positives, false positives and false negatives. For this, we have to define the ground truth for our evaluation. Therefore, a human expert uses descriptions of 40 products. For each product, he manually identifies problem solution ideas from the internet. Additionally, he checks the results of this approach to find further new and useful ideas by using the web log mining application.

To compute the percentage for the frequency baseline, we use the average percentage of all 40 products, which is computed by the number of new product ideas - as computed by the human expert - divided by the number of query results.

We use the web mining approach to identify the number of queries from each patent. For each query, we focus on the first ten query results. Therefore, we multiply the number of queries by ten to get the number of query results for each patent. Then, for each patent, we divide the number of new and useful ideas as computed above by the number of query results. After this, we get a percentage x. It says that x % of all query results represent a new and useful problem solution idea. Then we compute the average percentage for all 40 patents. As a result, 3 % of all query results represents a new product idea. Therefore, we set the frequency baseline to 3 %.

For each product, we compute the values of true positives, false positives and false negatives using the web mining application. Then, we compute the precision and recall values. After this, we compute the average precision and recall values for all products.

As a result, we get a precision value of 30 % and a recall value of 50 % (see Table 1). A precision value of 30 % means that if this approach predicts ten new and useful ideas then three of them are new and useful ideas. A recall value of 50 % means that if there are 10 new and useful ideas in the internet then this approach identifies five of them.

	Precision	Recall
Identify new ideas from the colloquial language	30%	50%
Identify new ideas from the technical language	40%	50%
Frequency baseline	3%	50%

Table 1. Results of Precision and Recall

To see whether these results are good or bad we compare them to a further approach that has the aim to identify new ideas from the technical language [14]. Here, we get a precision value of 40 % at a recall value of 50 %. This is because in the technical language, terms are defined more exactly and homonym and synonym problems do not occur so often.

Additionally, we compare the results to the frequency baseline. Here, we get a precision value of 3 % at a recall value of 50 %. Therefore, we think that this web mining approach can be used to support persons from marketing department by finding new product ideas from web logs.

8. Conclusion

This study has shown that a web mining approach that automatically identifies product ideas extracted from web logs outperforms the frequent baseline. Thus, it can be used to support marketing professionals by extracting new and useful product ideas. Further, it is shown that the use of a colloquial language by consumers leads to a decreased performance in contrast to the technical language, where terms are defined more exactly. Future research should aim at extracting product ideas from social networks e.g. from Facebook or Twitter accounts. A further avenue for future research is to extract product ideas in a multi-language environment. This means, a product description e.g. in English is used to find new product ideas e.g. in German language.

Bibliography

[1] C. Zsunyi, "Weblogs/ Blogs: Stand der Technik und Zukunftspotentiale," GRIN Verlag, pp. 4-6, 2007.

[2] S.C. Herring, L.A. Scheidt, S. Bonus, E. Wright, "Bridging the gap: a genre analysis of Weblogs," Proc. 37th Annual Hawaii International Conference on System Sciences, Hawaii, 2004.

[3] J.H. Soll, S. Strauch, "Ideengenerierung mit Konsumenten im Internet," Springer, Berlin, Heidelberg, 2006.

[4] L. Lawton, A. Parasuraman, "The impact of the marketing concept on new product planning," Journal of Marketing, vol. 44:19, 1980.

[5] A. Kuss "Marketing-einfuehrung: Grundlagen- Ueberblick- Beispiele," Springer, p. 189, 2006.

[6] L. Hoffmann, H. Kalverkaemper, H.E. Wiegand, "Languages for Special purposes," Walter de Gruyter, p. 1602, 1998.

[7] M.J. Martin-Bautista, D. Sanches, J.M. Serrano, M.A. Vila "Text Mining using Fuzzy Association Rules," V. Loia, M. Nikravesh, L.A. Zadeh, "Fuzzy Logic and the Internet," Springer, Berlin, p. 173, 2004.

[8] R. Feldman, J. Sanger, "The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data," Cambridge University Press, p. 318, 2007.

[9] G. Lustig, "Automatische Indexierung zwischen Forschung und Anwendung," Georg Olms Verlag, Hildesheim, p. 92, 1986.

[10] A. Hotho, A. Nuernberger, G. Paass, "A Brief Survey of Text Mining," LDV Forum, vol. 20(1), pp. 19-26, 2005.

[11] M.F. Porter, "An algorithm for suffix stripping," Program, vol. 14(3), pp. 130-137, 1980

[12] D. Carl, J. Clausen, M. Hassler, A. Zund, "Mashups programmieren," O'Reilly Germany, pp. 51-53, 2008.

[13] P. Mayr, F. Tosques, "Webometrische Analysen mit Hilfe der Google Web APIs," Information Wissenschaft und Praxis, vol. 56(1), pp. 41-48, 2005.

[14] A. Hotho, "Clustern mit Hintergrundwissen," Diss., Uni Karlsruhe, p. 29, 2004.

[15] D. Thorleuchter, „Finding Technological Ideas and Inventions with Text Mining and Technique Philosophy," C. Preisach, H. Burkhardt, L. Schmidt-Thieme, R. Decker Eds. "Data Analysis, Machine Learning, and Applications," Springer, Berlin, pp. 413-420, 2008.

Chapter VII

A compared R&D-based and patent-based cross impact analysis for identifying relationships between technologies

This chapter is based on 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), 2010, 1037-1050.

Table of Contents

A compared R&D-based and patent-based cross impact analysis for identifying relationships between technologies

Dirk Thorleuchter[a], Dirk Van den Poel[b], and Anita Prinzie[b]

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

[b] Ghent University, Faculty of Economics and Business Administration, B-9000 Gent, Tweekerkenstraat 2, Belgium, dirk.vandenpoel@ugent.be, anita.prinzie@ugent.be

Abstract

The planning of technological research and development (R&D) is demanding in areas with many relationships between technologies. To support decision makers of a government organization with R&D planning in these areas, a methodology to make the technology impact more transparent is introduced. The method shows current technology impact and impact trends from the R&D of an organization's competitors and compares these to the technology impact and impact trends from the organization's own R&D. This way, relative strength, relative weakness, plus parity of the organization's R&D activities in technology pairs can be identified.

A quantitative cross impact analysis (CIA) approach is used to estimate the impact across technologies. Our quantitative CIA approach contrasts to standard qualitative CIA approaches that estimate technology impact by means of literature surveys and expert interviews. In this paper, the impact is computed based on the R&D information regarding the respective organization on one hand, and based on patent data representative regarding R&D information of the organization's competitors on the other hand. As an illustration, the application field 'defence' is used, where many interrelations and interdependencies between defence-based technologies occur. Firstly, an R&D-based and patent-based Compared Cross Impact (CCI) among technologies is computed. Secondly, characteristics of the CCI are identified. Thirdly, the CCI data is presented as a network to show the overall structure and the complex relationships between the technologies. Finally, changes of the CCI are analyzed over time. The results show that the proposed methodology generates useful insights for government organizations to direct technology investments.

Key Words

Compared cross impact, Cross impact analysis, Technological impact analysis, R&D, Patent analysis, Defence Taxonomy, Centroid Vector, Machine Learning, Multi Label Classification

1. Introduction

The planning of research and development (R&D) requires technological trend analysis to ensure an effective investment of limited R&D budgets within organizations [1]. However, trend analysis is a very demanding task in areas where many interrelations and interdependencies between technologies occur because the impact of all related technologies has to be considered. Therefore, analyzing the impact across technologies is helpful for R&D planning and also to develop R&D strategies in these areas.

To support an organization's strategy and R&D planning, the technology impact analysis should be done both for the organization's own R&D activities (from now on referred to as 'internal R&D') as well as for the competitors' R&D activities (from now on referred to as 'external R&D'). By comparing the technology impact from internal R&D to the impact from external R&D, one can portray the advantages and the disadvantages of the internal R&D to competitors' external R&D. This improves the planning of R&D activities [2,3], the systematic identification of R&D priorities [4], the discovery of current technological vacuums [5], and the analysis of technological trends and opportunities [6,7] for the organization at hand.

The internal R&D technology impact analysis focuses on the relationships between technologies of the many simultaneously run R&D projects in the organization's R&D department. Typically, one R&D project deals with several different technologies. Therefore, each internal R&D project is assigned to one or several technologies from a specific technology list or taxonomy [8] by multi-label classification [9]. Analyzing this multiple classification shows which R&D projects are frequently assigned to specific technologies. This enables to calculate the cross impact index estimating the impact across these technologies developed by the organization. A proper calculation of this cross impact index requires a large number of internal R&D projects working on many different technologies. Companies normally do not have a large number of internal R&D projects or they are limited to a small number of technologies. Therefore, our approach focuses specifically on government organizations with a large number of R&D projects (> 100 projects) and a large technological scope (> 20 technologies).

The estimation of the technology impact from internal R&D should be augmented with the analysis of the relationships between technologies of external R&D. After all, no organization is so large that it has enough resources to excel in all technological areas or that it could not benefit from the advice of others [10]. For instance, organizations could learn from small firms, which are often more innovative. Therefore, it is necessary to consider R&D activities related to the internal R&D technologies from other organizations, i.e. external R&D. Patent data are used as representative for external R&D (see Sect. 2.3) because patents normally represent results of R&D projects. If this external R&D is also assigned to several technologies from the above mentioned technology list or taxonomy using multi-label classification then the impact across these technologies can also be estimated for the R&D activities of the organization's competitors.

This paper uses cross impact analysis (CIA) to estimate the impact of each technology on other technologies in a quantitative way as opposed to the more common qualitative approach by means of literature surveys and expert interviews. Our focus on large application fields characterized by a large number of corresponding technologies makes traditional qualitative CIA inappropriate (where a cross impact matrix is constructed by technology experts estimating the initial impact probabilities of each technology and the conditional impact probabilities of each technology pair [11,12,13]). However, in large application fields, a large number of corresponding technologies exists e.g. in the 'defence' application field the European Defence Agency (EDA) taxonomy of technologies consists of more than 200 technologies. To construct a 200-by-200 cross impact matrix n * (n-1) = 200 * 199 = 39.800 estimations are required by human experts. As can be seen from this example, a qualitative CIA approach in large application fields seems infeasible.

In this paper, a quantitative CIA approach is used to compute technology impact estimates that incorporate both internal and external R&D. In contrast to other quantitative CIA approaches which estimate the absolute impact of technologies (see Sect. 2.1), we first focus on technologies from an application field (e.g. 'defence') by assigning both internal R&D from an organization as well as external R&D to these technologies by multi-label classification. Then, we evaluate the relative impact of technologies by comparing the impact from internal R&D to the impact from external R&D, as captured by a new index we developed called the Compared Cross Impact (CCI) index (see Sect. 3). This relative impact shows how a government organization with many R&D projects can profit from the R&D of others (see Sect. 4 and 5).

This paper contributes to previous research in multiple ways. The main contribution of the proposed approach is the new CCI index that identifies relative strength, relative weakness, plus parity of the organization's R&D activities in technology pairs. The second contribution is a method to determine the characteristics of relationships and to show whether two technologies are equally influencing one another (symmetry) or whether the impact of the first technology on the second is different from the impact of the second technology on the first (asymmetry). A third contribution is the presentation of a CCI network graph that shows the overall structure and the complex CCI relationships between several technologies. Finally, changes of the CCI are analyzed over time to discover trends regarding how the technology impact changes over time. They show which technology should receive more or less development and investment. Overall, the results testify to the ability of CCI to generate useful insights for R&D decision makers of organizations.

2. Background

This approach combines methods from CIA and text classification and it applies them on patent data. The following paragraphs give an overview on existing CIA and text classification methods and on the (dis-) advantages of patent data.

2.1. Cross impact analysis

The use of CIA was first mentioned in 1968 [16] and consists of five steps. Firstly, events (e.g. technologies) are defined. Then, the occurrence probabilities and the conditional probabilities between events are estimated in the second and third step. Fourthly, a calibration run is performed to access the consistency / stability of the probabilities and last, the results are evaluated.

In literature, many improved CIA approaches have been introduced. Most of these necessitate the involvement of human experts and are therefore more subjective. The approaches are applied to different areas. Dalkey presents conditions for computing the occurrence probabilities of the first- and second-order [17]. To compute the higher-order probabilities, Duperring and Godet suggest a quadratic programming method [18] and Mitchell provides a linear programming method [19]. Enzer uses CIA to forecast future technologies based on a Delphi survey. Blanning and Reinig use the ratio of experts to define the occurrence probability P(A) (the percentage of all experts who predict the occurrence of A) and the conditional probability P(B|A) (the number of all experts who predict the occurrence of both A and B divided by the number of all experts who predict the occurrence of A) [20].

Additionally, more objective CIA approaches have also been introduced. Caselles-Moncho uses cumulative sales probabilities over time to compute the occurrence probabilities [21]. Jeong and Kim create inference algorithms based on linguistic values and the time lag as fuzzy numbers to compute the conditional probabilities between technologies [11]. A patent based CIA is presented in [1]. The standard assignment of US patents to the United States Patent Classification [22] is used to assign patents to several patent classification codes (PCC). A PCC impact index Impact(A,B) = P(B|A) is proposed to compute the impact of PCC A on PCC B.

2.2. Text classification methods

Text classification aims at assigning pre-defined classes (e.g. technology areas) to text documents (e.g. patent descriptions). The most frequently used data mining methods for text classification (categorization) are described in [26]: Naive Bayes is a probabilistic classifier simplifying Bayes' Theorem by naively assuming class conditional independence. The k nearest neighbor (k-NN) classification as instance-based learning algorithm selects documents from the training data which are 'similar' to the target document. Subsequently the class of the target document can be inferred from the class labels of these similar documents. Decision trees [27] are non-parametric classifiers recursively partitioning the observations (patent documents) into subgroups with a more homogeneous response (technology area). C4.5 is a well-known decision tree algorithm. A Support Vector Machine (SVM) is a supervised classification algorithm that determines a hyperplane, which separates the positive examples from the negative examples of the training data. A small number of training examples (support vectors) determine the actual location of the hyperplane. Then, target documents are assigned to one side of this hyperplane. The centroid-based approach [28] describes classes by a centroid vector that summarizes the characteristics of each class, but not by a number of training examples like k-NN and SVM. The assumption of a centroid classifier is that a target document should be assigned a particular class if the similarity of the document vector to the centroid vector of the class is the largest.

2.3. Patent data

Patent data are a valuable source of information concerning R&D. The data are useful to researchers for technological decision-making as well as to technology planners for R&D strategy making. Nevertheless, there are some limitations to use patent data because not all inventions are patented [14], the interpretations of patent analyses are not consistent across technology fields [15], and changes in patent law make it difficult to analyze trends over time [14]. However, patents are often used in analyses on technological innovation.

In patent research, statistical data are normally used (e.g. number of patents, application year, registration country, citation information). On the contrary, this research focuses on patent classification data by multiple assignment of patents to technologies and by computing the impact across these technologies. Patent data are used as representative for external R&D. Comparing the external R&D impact to the impact of internal R&D activities from a large organization leads to interesting knowledge for planning and managing R&D activities in this organization.

3. Methods: A compared R&D-based and patent-based CIA

3.1. Overview of proposed CIA

Our proposed quantitative CIA approach to estimate the impact between technologies for organizations with many R&D projects consists of multiple steps as depicted in Fig. 1. In a pre-processing step, internal R&D and external R&D are assigned to specific technologies based on internal R&D project information and patent data respectively. In a second step, the cross impact indexes $CI_{int}(A,B)$ and $CI_{ext}(A,B)$ for each technology pair are calculated. Next, the cross impact indexes $CI_{int}(A,B)$ and $CI_{ext}(A,B)$ are rounded and recoded to boolean cross impact indexes $BCI_{int}(A,B)$ and $BCI_{ext}(A,B)$. In the fourth step, a CCI index CCI(A,B) for each technology pair is calculated and characterized. These CCI scores already provide insights into the organization’s relative strength and relative weakness. In the fifth step, a CCI network graph is created visualising the CCI of technologies thereby facilitating the identification of relative strength and relative weakness even more. Steps one to four are discussed in Sect. 3.2 and Sect. 3.3 below. Sect. 3.4 elaborates on step 5. Finally, Sect. 3.5 documents on how the entire five-step approach can be applied on longitudinal data to infer evolution in technology impacts.

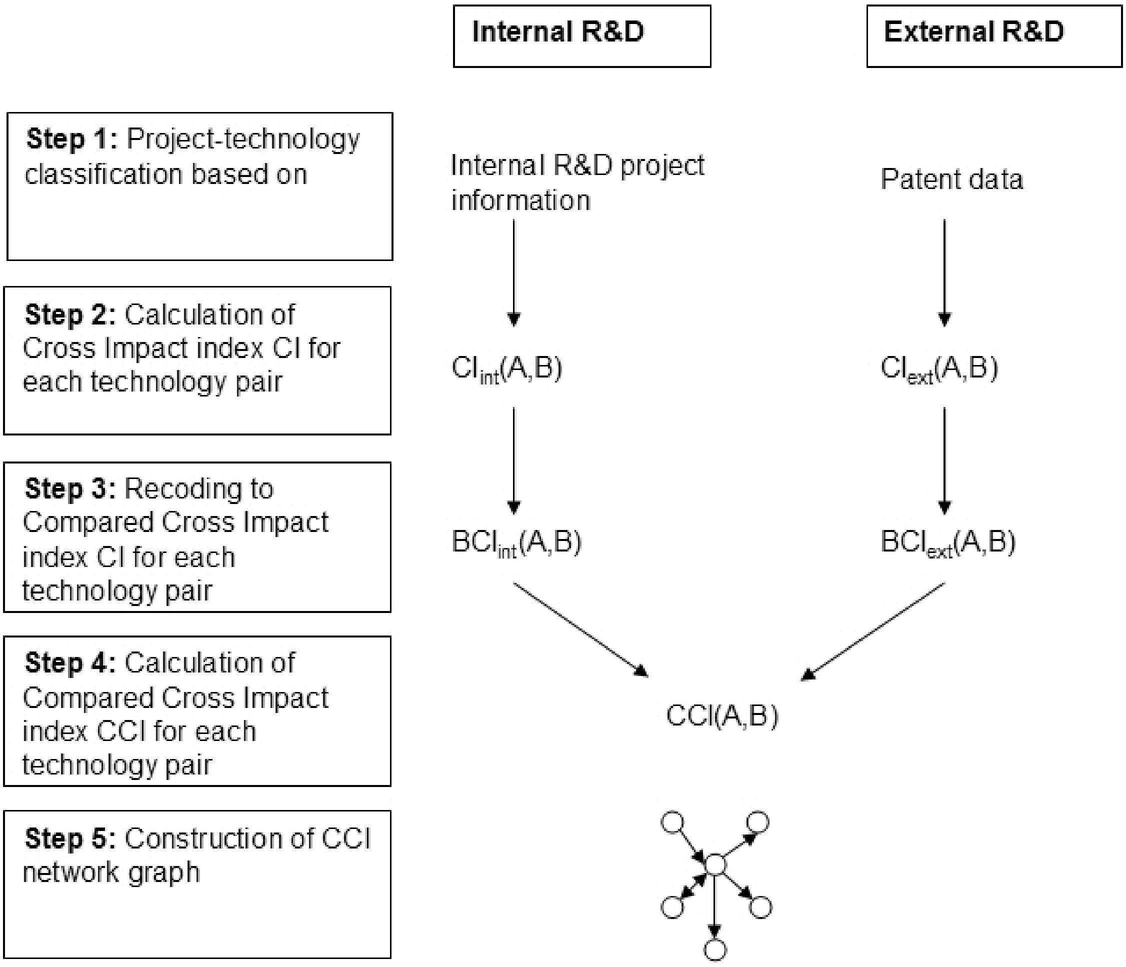

Internal R&D
External R&D
Step 1: Project-technology classification based on
Internal R&D project information
Patent data
Step 2: Calculation of Cross Impact index CI for each technology pair
CI$_{int}$(A,B)
CI$_{ext}$(A,B)
Step 3: Recoding to Compared Cross Impact index CI for each technology pair
BCI$_{int}$(A,B)
BCI$_{ext}$(A,B)
Step 4: Calculation of Compared Cross Impact index CCI for each technology pair
CCI(A,B)
Step 5: Construction of CCI network graph

Figure 1: Overview of quantitative CIA-approach.

3.2. Estimation of the new compared cross impact index

We adapt the PCC impact index from Sect. 2.1 to a) measure the cross-technology impact of external R&D as reflected by patents and b) measure the cross-technology impact of internal R&D. These modified indices are defined below:

Definition 1. Let Next(A) be the number of patents (as representative for external R&D) that are associated with technology A and let $Next(A \cap B)$ be the number of patents associated with both, technology A and B. Then, the cross impact index for external R&D CIext(A,B) is defined as the conditional probability between technology A and technology B considering patent data.

$$CIext(A,B) = Pext(B|A) = Next(A \cap B) / Next(A) \qquad (1)$$

In a similar way the cross impact index for external R&D $CI_{ext}(B,A)$ is defined as the conditional probability between technology B and technology A considering patent data.

Let $N_{int}(A)$ be the number of R&D projects (as representative for internal R&D) that are associated with the technology A and let $N_{int}(A \cap B)$ be the number of R&D projects associated with both, technology A and B. Then, the cross impact index for internal R&D $CI_{int}(A,B)$ is defined as the conditional probability between technology A and technology B considering internal R&D projects.

$CI_{int}(A,B) = P_{int}(B|A) = N_{int}(A \cap B) / N_{int}(A)$ (2)

Likewise, the cross impact index for internal R&D $CI_{int}(B,A)$ is defined as the conditional probability between technology B and technology A considering internal R&D projects.

Result values of $CI_{ext}(A,B)$, $CI_{ext}(B,A)$, $CI_{int}(A,B)$, and $CI_{int}(B,A)$ are between 0 and 1. A result value of one means that the first technology has a strong impact on the second technology and a result value of zero means that there is no impact. Two examples to illustrate the meaning of the cross impact index for internal R&D and external R&D are presented. A $CI_{int}(A,B)$ of 0.25 means that 25% of all internal R&D projects adopting technology A also employ technology B. A $CI_{ext}(A,B)$ of 0.20 means that 20% of all patents related to technology A also refer to technology B.

The estimation of cross impact between technologies is done in two different ways.

Firstly, relationships between technologies are estimated using data regarding R&D activities from an organization. Internal R&D projects are assigned to technologies from a specific technology list or taxonomy (that is normally used in the organization for technology classification). This multiple assignment can be used to compute $CI_{int}(A,B)$. A proper computation of the cross impact index requires that each technology is associated with many R&D projects from the organization (see Sect. 1). The calculation of the $CI_{int}(A,B)$ provides organization researchers and research planners with an internal view of the relationships between technologies. However, this internal view does not consider relationships between technologies as apparent from external R&D.

Next, the R&D-technologies multiple assignment and calculation of the cross impact index is repeated for external R&D using patent data instead of internal R&D information. The patent data are assigned to the technologies from the above described technology list or taxonomy.

For this, methods from text classification can be used (see Sect. 2.2). This means those patents are considered that are related to at least one technology. The advantages of this patent-based CIA for researchers and technology planners are described in [1]. The disadvantage of patent-based CIA is that it neglects the technological relationships of the internal R&D when assessing the cross-technology impact.

Therefore, a compared R&D-based and patent-based CIA is proposed. Hence, we compute $CI_{int}(A,B)$ and $CI_{ext}(A,B)$. Then, boolean cross impact indexes and cutoff values are defined to decide whether there is an impact of technology A on technology B taking both internal R&D as well as external R&D into account.

Definition 2. Let c_{int} and c_{ext} be the internal and external cutoff percentages respectively. The boolean cross impact index $BCI_{int}(A,B)$ for internal R&D and the boolean cross impact index $BCI_{ext}(A,B)$ for external R&D are defined as follows:

$$BCI_{int}(A,B) = \begin{cases} 1 & (CI_{int}(A,B) \geq c_{int}) \\ 0 & (CI_{int}(A,B) < c_{int}) \end{cases} \tag{3}$$

$$BCI_{ext}(A,B) = \begin{cases} 1 & (CI_{ext}(A,B) \geq c_{ext}) \\ 0 & (CI_{ext}(A,B) < c_{ext}) \end{cases} \tag{4}$$

The cutoff percentage is separately defined for internal and external R&D. This is because the number of internal R&D projects is much smaller than the number of patents. As an example, if $N_{int}(A)$ equals five and $N_{int}(A \cap B)$ equals one then $CI_{int}(A,B)$ equals 0.20. However, this high value does not mean that this technology pair is a focal point in the R&D of the organization and that technology A has an impact on technology B. In contrast to this, a $CI_{ext}(A,B)$ of 0.20 means that 20% of all patents in technology A are also in technology B. Therefore, technology A has an impact on technology B. As seen from this example, it is necessary that cutoff values are separately defined for internal and external R&D e.g. for the case study in Sect. 4, the cutoff percentage for internal R&D c_{int} is set to 0.25 whereas the cutoff percentage for external R&D c_{ext} is set to 0.20.

Definition 3. Starting from the boolean cross impact indexes we define a CCI index CCI(A,B) as the difference between the internal and external boolean cross impact index.

$$CCI(A,B) = BCI_{int}(A,B) - BCI_{ext}(A,B) \tag{5}$$

Depending on whether $BCI_{int}(A,B)$ and $BCI_{ext}(A,B)$ are zero or one, the result value of CCI(A,B) is negative one, zero, or positive one (see Table 1). If CCI(A,B) equals negative one then a relative weakness in this area is observed for the organization. Technology A has an impact on technology B in the external R&D but not in the internal R&D. The internal R&D does not exploit this technology pair intensively. A potential strategic decision could be to increase R&D activities in this area. Alternatively, to gain strength in this area, the organization could outsource these R&D activities (to buy external R&D know how).

If CCI(A,B) equals positive one then this area can be considered a strength. This occurs, when technology A has an impact on technology B, in the internal R&D but this impact is absent from the competitors' R&D. A potential strategic decision based on this information is presented below: R&D in this area that does not increase value (e.g., it is old-fashioned or no consumers can be identified that are interested in future products from this area) leads to a strategic decision that decreases R&D activities in this area.

A CCI(A,B) of zero leads to two different cases. Firstly, if $BCI_{ext}(A,B)$ and $BCI_{int}(A,B)$ equal positive one then technology A has an impact on technology B both in the internal R&D and the external R&D. The R&D activities in this area can be classified as parity. If both cross impact values ($BCI_{ext}(A,B)$ and $BCI_{int}(A,B)$) equal zero then there is no impact of technology A on technology B because R&D activities in this area do not intensively occur. Then, the strategic decision to start new internal R&D activities in this area might lead to a relative strength in future.

Using a Boolean cross impact index leads to information loss. However, this is more appropriate than using a ratio scale because cutoff values can be determined intuitively (to decide whether there is an impact of technology A on technology B) at an early step and the results are easy to interpret (e.g. a CCI(A,B) of positive one means a relative strength). This makes the approach more transparent to the decision makers. Using a ratio scale instead leads firstly to a normalization of $CI_{ext}(A,B)$ and $CI_{int}(A,B)$ concerning the cutoff values and secondly to a ratio CCI(A,B) score between [-1,..,1]. The higher the CCI(A,B) score the more is the relative strength and the less is the relative weakness. Additionally, the closer the CCI(A,B) is to zero the more is the parity or the probability that there is no impact. Normally, decision makers of organizations preferred results that are easy to interpret created by transparent approaches. Thus, the use of Boolean cross impact indices is preferred in this approach.

Table 1: Result values of CCI(A,B)

$BCI_{int}(A,B)$	*$BCI_{ext}(A,B)$*	*CCI(A,B)*
0	0	0 (No impact)
0	1	-1 (Relative weakness)
1	0	1 (Relative strength)
1	1	0 (Parity)

3.3. Characteristics of the CCI between technology pairs

The CCI between two technologies can be classified as symmetrical, asymmetrical, or nonexistent. The impact between technology A and B is nonexistent if all four boolean cross impact indexes $BCI_{ext}(A,B)$, $BCI_{ext}(B,A)$, $BCI_{int}(A,B)$ and $BCI_{int}(B,A)$ equal zero.
Otherwise, if $BCI_{ext}(A,B)$ equals $BCI_{ext}(B,A)$ and $BCI_{int}(A,B)$ equals $BCI_{int}(B,A)$ then there is an impact of technology A on technology B and a similar impact of technology B on technology A. In this case, the CCI is classified as symmetrical. In the other case, the CCI between two technologies can be classified as having a asymmetrical impact. An example for this is a relative strength concerning CCI(A,B) and a relative weakness concerning CCI(B,A). These characteristics are used to build a CCI network graph (see Sect. 3.4).

3.4. CCI network graph

The CCI calculates the relationship between two technologies considering both internal and external R&D. However, each technology can affect two or more technologies and vice versa. Therefore, it is useful to identify the complex relation among three or more technologies. To visually express the relationships between several technologies network analysis - as well-known technique from graph theory [23] - is used. In this graph, each node represents a technology and each edge represents the CCI between two technologies. The direction of the edge shows the direction of the asymmetrical or symmetrical impact.

With the network graph, influencing and influenced technologies can be identified. For example, a technology might influence several other technologies or may be influence by several technologies. For a technology, that influences a large number of related technologies, an increased development and investment also probably increases strength in the related technologies. Additionally, forecasting future trends is easier in technologies that are influenced by a small number of other technologies.

A sequential impact between several technologies (where technology A has an impact on technology B and technology B has an impact on technology C) also can be found in the network graph. Then, the strategic decision to start new internal R&D activities in technology A might lead to an increased strength in technology C.

As an example, Fig. 2 shows a symmetrical relative strength between A and B and it also shows an asymmetrical relationship between A and C as well as between B and C. The impact of C on B represents a parity and the impact of B on C represents a relative strength. Additionally, a relative weakness is seen concerning the impact of C on A and no impact is seen of A on C. Further, a 3-element long sequential relative strength A → B → C can be seen. Last, technology A influences B and is influenced by B and C.

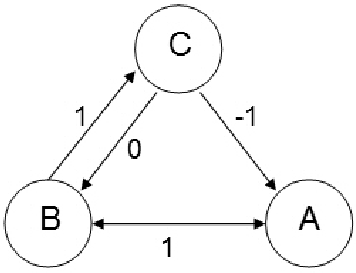


Figure 2: Example for a CCI network graph.

3.5. Changes of CCI

The CCI constantly changes over time because it is based on the cross impact with regard to internal R&D and external R&D. It is characteristic for R&D activities of organizations that many new R&D projects start and many existing R&D projects are completed every year. A new R&D project often focuses on a different technology combination and therefore, the impact across technologies changes over time. It is also characteristic for patent data that the impact across technologies changes because of the change in customer needs and the occurrence of new technologies.

The change of the cross impact between technologies concerning internal R&D can be computed by using information from the R&D program of the organization in a specific year. An R&D program is the collection of all active R&D projects. Using this yearly internal R&D information, the cross impact between technologies in a specific year can be identified and used in the CIA approach. Additionally, by collecting the patents that are registered in a specific year the cross impact between technologies in a specific year concerning patent data can be identified.

Then, the CCI and the degree of change can be computed using the proposed compared CIA approach. As an example, CCI(A,B) for 2006 equals positive one and CCI(A,B) for 2007 equals zero with $BCI_{int}(A,B)$ and $BCI_{ext}(A,B)$ both being positive one. Then, the relative strength in the impact of technology A on technology B has become parity. For strategic decision making, this information could be interesting because it shows the impact trend and therefore, it shows which technology should receive more or less development and investment in the future.

4. Case study ’defence’ - data collection and text classification

4.1. Application field

In the last years, the rising asymmetrical threat is causing governments to pay more attention to defence, especially in technological areas. New and ever more complex tasks in areas concerned with defence against these new types of threats require additional R&D of new techniques. For this reason, European governments and the European Union are increasingly funding defence-based technological R&D. For example, the EDA was established in 2004 and coordinates defence-based R&D between State Members of the European Union. Because of growing budgets in the field of defence-based R&D, one can monitor an increasing number of research projects and an increasing collaboration especially between defence-based R&D and (civil) security-based R&D. This leads to a continuous change of the defence-related technological landscape: the appearance of many new technologies and new interrelations and interdependencies between technologies [24]. This is partly due to applied science R&D projects often using several technologies to create a defence application [25].

4.2. Technology collection

For this case study, technologies from the application field ’defence’ are needed. A well-known European technology taxonomy in this field is from the EDA. The EDA taxonomy of technologies (CAPTECH) contains about 200 defence-based technologies that are assigned to 32 technology areas. Additionally, EDA provides detailed descriptions for each technology. For this case study, we use all 32 technology areas from this taxonomy as described in Table 2.

Table 2: List of technology areas from EDA taxonomy of defence-based technologies

Number	*Technology area*
A01	Structural & Smart Materials & Structural Mechanics
A02	Signature Related Materials
A03	Electronic Materials Technology
A04	Photonic/Optical Materials & Device Technology
A05	Electronic, Electrical & Electromechanical Device Technology
A06	Energetic Materials and Plasma Technology
A07	Chemical, Biological & Medical Materials
A08	Computing Technologies & Mathematical Techniques
A09	Information and Signal Processing Technology
A10	Human Sciences
A11	Operating Environment Technology
A12	Mechanical, Thermal & Fluid Related Technologies & Devices
B01	Lethality & Platform Protection
B02	Propulsion and Powerplants
B03	Design Technologies for Platforms and Weapons
B04	Electronic Warfare and Directed Energy Technologies
B05	Signature Control and Signature Reduction
B06	Sensor Systems
B07	Guidance and Control systems for Weapons and Platforms
B08	Simulators, Trainers and Synthetic Environments
B09	Integrated Systems Technology
B10	Communications and CIS-related Technologies
B11	Personnel Protection Systems
B12	Manufacturing Processes/Design Tools/Techniques
C01	Defence Analysis
C02	Integrated Platforms
C03	Weapons
C04	Installations and Facilities
C05	Equipped Personnel
C06	Miscellaneous Defence Functions and Policy Support
C07	Battlespace Information
C08	Business Process

4.3. Collection of internal R&D

We use R&D projects from the German Ministry of Defence (GE MoD) as internal R&D information. 985 R&D projects from the GE MoD have been identified. The projects are already manually assigned to technologies and therefore also to the technology areas of the EDA taxonomy by use of multi-label classification. This means, each project is assigned to one or several technology areas from the EDA taxonomy.

4.4. Collection of external R&D

Patent data are collected from the United States Patent and Trademark Office (USPTO). We use the Derwent Innovations Index to extract patent numbers, titles, and abstracts from the 182,928 patents from the year 2007. Patents from the GE MoD are not collected as well as patents from other organizations and companies where the R&D behind this patent is funded by the GE MoD. Then, patents are assigned to none, one or several technology areas of the EDA taxonomy by use of multi-label text classification.

4.5. Centroid-based patent classification

In this case study, we opt for centroid-based patent classification. Below, we substantiate this methodological choice. Centroid-based classifiers have been widely used in many web applications and previous work [29] has shown that the prediction accuracy of centroid-based classifiers is significantly lower than other approaches (e.g., SVM). However, two advantages are important in practice. Firstly, the centroid-based algorithm has a very intuitive meaning [30], which is important because classification results are used as decision support for managers and decision makers of the GE MoD. Secondly, the computational complexity of this centroid-based approach is important given the large number of patents (182,928 patents from the year 2007) and the large number of classes / training examples in the application field ’defence’ (32 technology areas / 200 technology descriptions). In the training phase, the centroid-based algorithm has a linear-time complexity that depends on the number of training examples. We also observe a linear complexity in the classification phase that depends on the number of classes. Hence, the overall computational complexity of the centroid-based algorithm is very low.

Each technology area consists of several technologies (see Sect. 4.2). To acquire training examples for each technology area, we use the descriptions of the respective defence-based technologies from each technology area as reflected in the EDA taxonomy of technologies. Then, terms and term frequencies are extracted and term vectors in a vector space model [31] are built for each training example. For each technology area, we build the centroid vector of all term vectors that are assigned to the technology area. For this, we use tokenization [32], stop word filtering (by use of domain specific stop word list), stemming (by use of Porter stemmer [33]), and manual extraction of prevalent features [34] that are characteristic for a technology area. This centroid vector is used to describe the corresponding technology area.

For classification, patents are used as test examples (see Sect. 4.4). Patent descriptions of these examples are prepared and terms and term frequencies are extracted for each patent. Then, these terms are used to create term vectors. Each term vector from the test examples is compared to each centroid vector using a similarity measure. Here, Jaccard’s coefficient measure [31] is selected because it handles well vectors of different length; e.g. the term vector might have a different length than the centroid vector to which it is compared.

To identify whether a term vector from a test example (patent) is similar to a centroid vector representing a technology area a maximal distance to the centroid vector is determined. A term vector from a test example is defined as similar to a centroid vector if the corresponding Jaccard’s coefficient measure is greater than or equal to a user-set α (alpha-cut method [35]). A term vector from a test example (patent) is simply assigned to all classes of its similar centroids. As a result, one can identify none, one or several corresponding technology areas for a given patent. For the case study ’defence’ α is set to 0.15 to balance the type I and type II error. If the percentage of α is too small then probably patents are falsely assigned to technology areas (type I error). If the percentage of α is too large then patents are probably not assigned to the technology areas they belong to (type II error).

5. Results and Discussions

5.1. CCI between technology areas

Table 3 shows the results of the case study ’defence’. The technology area pairs are ordered by the CCI score. Within CCI score the technology area pairs are ordered by R&D-based cross impact score $CI_{int}(A,B)$ if CCI(A,B) equals positive one or zero, otherwise they are ordered by the patent-based cross impact score $CI_{ext}(A,B)$. The influencing technology area is represented by ’Techn. area A’ and the influenced technology area is represented by ’Techn. area B’. Table 3 does not consider technology area pairs with no impact. For each technology area pair, R&D-based and patent-based cross impacts are computed and rounded, i.e. $CI_{int}(A,B)$ and $CI_{ext}(A,B)$ respectively. R&D-based cross impacts scores $CI_{int}(A,B)$ exceeding the 0.25 threshold are indicated in bold face and patent-based cross impact scores $CI_{ext}(A,B)$ exceeding the 0.20 threshold are indicated in italics. Next, the boolean cross impact scores $BCI_{int}(A,B)$ and $BCI_{ext}(A,B)$ are calculated. The $BCI_{int}(A,B)$ and $BCI_{ext}(A,B)$ are positive one if the $CI_{int}(A,B)$ and $CI_{ext}(A,B)$ are at least 0.25 and 0.20 respectively as described in Sect. 3.2. Finally, the CCI scores are computed. The last column shows that CCI(A,B) is classified as symmetrical or asymmetrical as described in Sect. 3.3.

Table 3: Technology area pairs with high cross impact in 2007

Techn. area A	*Techn. area B*	CI_{int} *(A,B)*	BCI_{int} *(A,B)*	CI_{ext} *(A,B)*	BCI_{ext} *(A,B)*	*CCI (A,B)*	*Sym. Asym.*
B02	A05	**0.54**	1	0.07	0	1	S
B07	C03	**0.39**	1	0.02	0	1	A
A04	B07	**0.34**	1	0.13	0	1	A
C05	B11	**0.32**	1	0.08	0	1	A
A05	B02	**0.29**	1	0.01	0	1	S
A07	A04	**0.25**	1	0.08	0	1	A
B10	B07	0.03	0	*0.26*	1	-1	S
A05	C05	0.16	0	*0.23*	1	-1	A
A05	B10	0.07	0	*0.21*	1	-1	A
B07	B10	0.11	0	*0.20*	1	-1	S
A02	B05	**0.92**	1	*0.58*	1	0	S
A03	A05	**0.86**	1	*0.30*	1	0	S
B05	A02	**0.62**	1	*0.46*	1	0	S
B04	A05	**0.61**	1	*0.35*	1	0	S
A12	B02	**0.58**	1	*0.22*	1	0	A
B08	A08	**0.53**	1	*0.26*	1	0	S
B01	A01	**0.42**	1	*0.27*	1	0	A
B06	A09	**0.38**	1	*0.23*	1	0	A
B07	C02	**0.34**	1	*0.23*	1	0	A
A05	A03	**0.32**	1	*0.26*	1	0	S
A08	B08	**0.31**	1	*0.22*	1	0	S
A05	B06	**0.27**	1	*0.20*	1	0	A
A05	B04	**0.26**	1	*0.21*	1	0	S

For example, let us consider row 1 in Table 3. The number of R&D projects / patents in the technology area B02 'Propulsion and Powerplants' is 37 for R&D projects and 563 for patents. The number of R&D projects and patents included both in technology area B02 and A05 'Electronic, Electrical & Electromechanical Device Technology' is 20 for R&D projects and 39 for patents. Table 4 explains the calculation of the R&D-based cross impact score $CI_{int}(A,B)$, the patent-based cross impact score $CI_{ext}(A,B)$, the boolean cross impact scores $BCI_{int}(A,B)$ and $BCI_{ext}(A,B)$ using a cutoff of 0.25 and 0.20 respectively and finally the CCI score CCI(A,B). CCI(A,B) is classified as symmetrical because $BCI_{int}(A,B)$ equals $BCI_{int}(B,A)$ and $BCI_{ext}(A,B)$ equals $BCI_{ext}(B,A)$.

Table 4: Explanation of calculation of cross impact scores and CCI score for row 1 of Table 3

$CI_{int}(A,B)$	$= N_{int}(A \cap B) / N_{int}(A)$
	$= 20 / 37$
	$= 0.54$
$CI_{ext}(A,B)$	$= N_{ext}(A \cap B) / N_{ext}(A)$
	$= 39 / 563$
	$= 0.07$
$BCI_{int}(A,B)$	$= 1$
$BCI_{ext}(A,B)$	$= 0$
$CCI(A,B)$	$= BCI_{int}(A,B) - BCI_{ext}(A,B)$
	$= 1 - 0$
	$= 1$

5.1.1. Relative strength

In the case study, a relative strength for the R&D of the GE MoD can be seen in various technology area pairs where the CCI(A,B) equals positive one (see Table 3). Here, the R&D-based cross impact score $CI_{int}(A,B)$ is greater than or equal to the internal cutoff value and the patent-based cross impact score $CI_{ext}(A,B)$ is smaller than the external cutoff value. Below, we describe these technology area pairs.

A focal point of the GE MoD is the R&D to create a MEE (More Electric Engine). 54% R&D projects in the technology area B02 (Propulsion and Powerplants) are also assigned to A05 (Electronic, Electrical & Electromechanical Device Technology) and 29% vice versa. The external R&D is not focused on the combination of these two technology areas B02 and A05. A further core theme is the R&D in fibre optic gyroscope technology for navigation. 34% of all R&D projects from ’Photonic/Optical Materials & Device Technology’ (A04) also are assigned to B07 (Guidance and Control systems for Weapons and Platforms). An impact of technology area C05 on technology area B11 can be seen. This is because research in the technology area C05 ’Personnel Equipment’ is mainly focused on the technology area B11 ’Personnel Protection Systems’ e.g. to provide significant survivability to the German infantryman. Therefore, 32% of all R&D projects in technology area C05 are also assigned to technology area B11. Additionally, the intensive R&D in guidance and control systems for weapons to reduce collateral damage leads to an impact of technology area B07 on technology area C03 and the intensive R&D for a chemical oxygen iodine laser leads to an impact of technology area A07 on technology area A04. Together with expert knowledge (e.g. the fact that R&D in chemical oxygen iodine lasers probably does not increase value because it might be old-fashioned concerning fibre lasers), an advise can be given to decrease these R&D activities.

These results show that the GE MoD has strength in several technology area pairs and that other organizations (e.g. competitors) do not have strength in these technology area pairs as apparent from the small patent-based cross impact scores $CI_{ext}(A,B)$. An organization should aim to build on its relative strength specifically when R&D in these technology area pairs increases value. As such, knowledge about own relative strength and its competitors’ relative weakness can be used for R&D planning and strategic decision-making.

5.1.2. Relative weakness

Besides relative strength, Table 3 also portrays a relative weakness for the R&D of the GE MoD in technology area pairs where the CCI score equals negative one. This is the case when the R&D-based cross impact score $CI_{int}(A,B)$ is smaller than the internal cutoff value of 0.25 and the patent-based cross impact score $CI_{ext}(A,B)$ is greater than or equal to the external cutoff value of 0.20. Below, we describe these relative weakness technology area pairs.

In patent data, a symmetrical impact of navigation technology on communication technology can be found as described in [1]. Here in this case study, we also identify a symmetrical patent-based impact of B07 ’Guidance and Control systems for Weapons and Platforms’ (that includes e.g. navigation technology) and B10 ’Communications and CIS-related

Technologies’. However, only a small number of the GE MoD’s R&D projects that are assigned to technology area B07 are also assigned to the technology area B10 and vice versa. Further results of the case study are the patent-based impact of A05 (Electronic, Electrical & Electromechanical Device Technology) on C05 (Equipped Personnel) and on B10 (Communications and CIS-related Technologies). Here, it can also be seen that only a small number of internal R&D projects from A05 are assigned to C05 or B10.

These results show that the GE MoD does not have strength in several technology area pairs. However, other organizations often do have R&D projects in these technology area pairs as reflected by the patent-based cross impact score $CI_{ext}(A,B)$ exceeding the 0.20 threshold. An organization should aim to reduce its relative weaknesses specifically when R&D in these technology area pairs increase value. If the GE MoD has strength in a technology area like B07 then it can easily gain strength in a technology area like B10 in which it has relative weakness e.g. by R&D outsourcing. From this ’defence’ application it is clear that the knowledge about these relative weaknesses and about the possibilities to bridge these gaps can be used for R&D planning and strategic decision making.

5.1.3. Parity technology area pairs

The case study also identifies R&D technology area pairs being both focal to the GE MoD as well as to other organizations. These technology area pairs appear as third group in Table 3 where both the R&D-based cross impact score $CI_{int}(A,B)$ as well as the patent-based cross impact score $CI_{ext}(A,B)$ is greater than or equal to the internal or external cutoff value, respectively. Some technology area pairs have a large R&D-based and a large patent-based cross impact e.g. many R&D projects in A02 ’Signature Related Materials’ are also assigned to B05 ’Signature Control and Signature Reduction’. This is because the centroid vectors of A02 and B05 contain similar features. Then, the R&D-based and the patent-based cross impact score are both high and the CCI score equals zero. Further examples for centroid vectors with similar features are the technology area pair A03 ’Electronic Materials Technology’ and A05 ’Electronic, Electrical & Electromechanical Device Technology’ as well as the impact of technology area A12 ’Mechanical, Thermal & Fluid Related Technologies & Devices’ on B02 ’Propulsion and Powerplants’.

Another core theme of GE MoD is R&D for an intelligent smart sensor. Therefore, many R&D projects from technology area B06 ’Sensor Systems’ are also assigned to technology area A09 ’Information and Signal Processing Technology’. The R&D activities can be classified as parity because a patent-based impact of B06 on A09 is also observed. These results show that the GE MoD has strength in several technology area pairs in which other organizations also have strength. A strategic decision to decrease development and investment in a parity technology area pair probably leads to a relative weakness in the future. Therefore, this information can be used for R&D program planning and strategic decision making.

5.1.4. Technology area pairs with no impact

Technology area pairs with no impact are not listed in Table 3 because the R&D-based and patent-based cross impact scores are smaller than the corresponding cutoff values. However, they represent potential future strengths if they receive more development and investment from the GE MoD in the future. An example for using these technology area pairs in R&D planning is given in Sect. 5.3

5.2. Characteristics of the CCI between technology area pairs

Table 3 presents examples of symmetrical (S) and asymmetrical (A) impacts. The technology area impact between A05 and B02 is symmetrical. This means that the GE MoD portrays relative strength both for the (A05, B02) technology area pair as for the (B02, A05) pair. If the CCI score is negative one then a symmetrical cross impact can be observed between technology areas B07 and B10. Hence, the GE MoD has a relative weakness in the technology area pair (B07, B10) as well as and in the technology area pair (B10, B07). Additionally, (A03, A05) and (A05, B04) are examples of symmetrical parity cross impacts where the corresponding CCI score is zero.

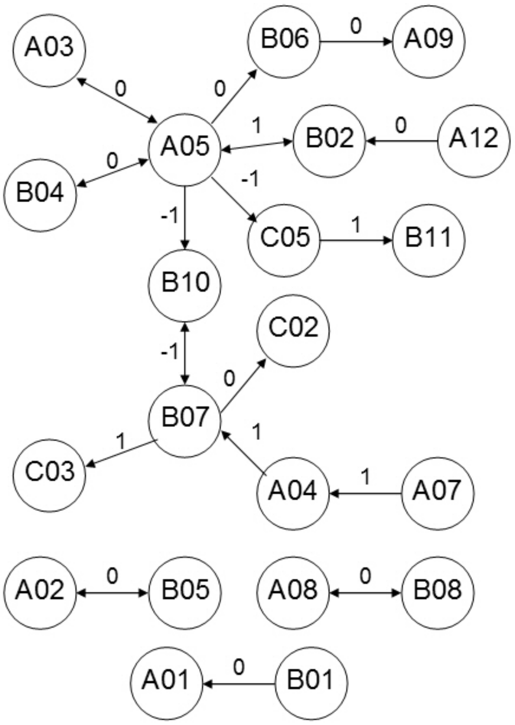


Figure 3: CCI network graph of EDA-technology areas in 2007.

5.3. CCI network graph

Based on the results of the case study, a CCI network graph of EDA technology areas is presented in Fig. 3 showing technology area impacts in 2007. With this CCI network graph, the overall structure and the complex relationships between several EDA technology areas can be shown. Each node represents an EDA technology area. Every edge is annotated with its corresponding CCI score that classifies the impact between two technology areas as an R&D-based cross impact; i.e. relative strength (1), a patent-based cross impact; i.e. relative weakness (-1) or an R&D-based and patent-based cross impact; i.e. parity (0). Additionally, the direction of each arrow characterizes the technology area impact as symmetrical or asymmetrical.

The CCI network graph shows the impact of two or more technology areas on a specific technology area. For example the impacts of A03, B02, and B04 on A05 are all symmetrical but differ in whether they display relative strength: CCI(B02,A05) = 1, or parity of the GE MoD: CCI(A03,A05) = 0 and CCI(B04,A05) = 0.

The CCI network graph also reveals the direction of technology area impacts. This way the influencing technology areas and the influenced technology areas can be identified. As an example, the technology area A05 influences six technology areas (A03, B02, B04, B06, B10, and C05). However, it is also influenced by three technology areas (A03, B02, and B04). Non-influenced technology areas are A07, A12, and B01. Each of them only influences one technology area. Additionally, A01, A09, B11, C02, and C03 are examples for non-influencing technology areas.

Three islands can be found in the CCI network graph. The two symmetrical parity technology area pairs are A02 and B05 as well as A08 and B08. The third island represents the asymmetrical impact of B01 on A01.

Sequential impacts between technology areas can be detected as well. For example, a sequential impact starts with A07 via A04 via B07 and it ends with C03. All these impacts are asymmetrical and every corresponding CCI score equals one. This means that the sequential impact represents a relative strength. A further sequential impact with different CCI scores is A12, B02, A05, B10, B07, and C03. Examples for a symmetrical sequential impact are B02, A05, and B04 as well as B02, A05, and A03.

As such, the CCI network graph facilitates the detection of asymmetrical / symmetrical relative strengths or relative weaknesses by showing the structure and the complex relationship between several technology areas. This is helpful information for research planning and strategic decision making. Searching for the edges annotated with -1 immediately indicates for which technology area pairs the GE MoD has relative weakness: in the technology area pairs (A05, C05), (A05, B10), (B10, B07). Likewise scanning for the edges annotated with +1 points out in which technology area pairs GE MoD excels: in (B02, A05), (C05, B11), (B07, C03), (B07, A04) and (A04, A07). From the CCI network graph it is apparent that the GE MoD's relative strengths are located along the B07 star whereas its relative weaknesses are mainly located along the A05 star. In general, an organization should aim a) to build on its relative strengths and b) to reduce its relative weaknesses. As to the former, the GE MoD should investigate whether it could extend the sequential impact A07 → A04 → B07 → C03. New relative strengths could be (*, A07), (A07, *), (*, A04), (A04, *), (*, B07), (B07,*), (*, C03), or (C03, *) with * referring to any technology area being part of the technology area pair with no impact (see Sect. 5.1.4). The advantage of building upon existing relative strengths stems from the fact that the organization already has experience with one of the technology areas belonging to the new relative strength technology area pair.

Besides building on its existing relative strengths, GE MoD should equally investigate whether it could connect its relative strengths. For the GE MoD turning one of the technology area pairs with no impact (C03, A05), (C03, B02), (C03, C05), (A05, A07), (B02, A07), and (B11, A07) in a relative strength would build on its sequential relative strength at the same time. Given that the GE MoD would gain strength in the technology area pair (C03, C05), the sequential relative strength A07 → A04 → B07 → C03 could be extended with C03 → C05 → B11 to form a 6-element long sequential relative strength A07 → A04 → B07 → C03 → C05 → B11. As such, the GE MoD should initially focus on turning specifically technology area pairs with no impact in a relative strength by increasing development and investment. If it is not possible to gain strength in the technology area pair e.g. (C03, C05) then turning two technology area pairs with no impact (C03, x) and (x, C05) into a relative strength also builds on its sequential relative strength, e.g. x could be technology area A06. This would establish the 7-element long sequential relative strength A07 → A04 → B07 → C03 → A06 → C05 → B11. As such, the GE MoD should initially focus on turning the relative weaknesses in parity technology area pairs and the technology area pairs with no impact in relative strengths.

In summary, the above illustrates how the CCI network graph allows guiding research planning and strategic decision making.

5.4. Changes of the CCI

In Table 3 the CCI is computed by use of R&D information and patent data from year 2007. However, technology areas / technologies change and therefore, the CCI as well as the R&D-based and patent-based cross impact also change. To analyze this change over time, two technology area pairs have been tracked for years 2004 to 2008.

The technology area B06 (Sensor System) has an impact on technology area A09 (Information and Signal Processing Technology) because of R&D for smart (intelligent) sensors. The patent-based cross impact shows a nearly increasing trend from 2004 to 2008 (see Table 5). In the R&D of the GE MoD smart sensor activities become a focal point since 2006. Given that the internal and external cutoff values were set to 0.25 and to 0.20 then no impact of B06 on A09 can be seen in 2004. There is a relative weakness in 2005 because the CI_{ext}(B06,A09) is exceeding the threshold (printed in italics). This has led to an increased development and investment by the GE MoD and since 2006 the smart sensor R&D activities can be classified as being at parity because the CI_{int}(B06,A09) is exceeding threshold (in bold print). An advice for 2008 probably can be that the GE MoD should cut back investment a little bit in this technology area pair to keep the parity with a smaller investment.

A further example is the R&D to create a MEE, which is a focal point in the R&D of the GE MoD since 2005. Patents that deal with electronic, electrical or electromechanical device technology (A05) are normally assigned to other applications (communication, computer systems etc.) but not to propulsions and powerplants (B02). Therefore, a small patent-based cross impact CI_{ext}(A05,B02) can be seen from Table 6. Given that the internal and external cutoff values were set to 0.25 and to 0.20 there is no impact of A05 on B02 in 2004, but since 2005 the R&D activities combining A05 and B02 can be classified as a relative strength. This example shows how an increased development and investment in 2005 turn a technology area pair with no impact into a relative strength and it also shows that the value of CI_{int}(A05, B02) in 2008 is much larger than 0.25. An advice for 2008 probably can be that the GE MoD should cut back investment a little bit by reducing the number of R&D projects in this technology area pair to keep the relative strength with a smaller investment.

Table 5: Change of the (compared) cross impact of technology area B06 on technology area A09 from years 2004 to 2008

	2004	*2005*	*2006*	*2007*	*2008*
CI_{int}(B06, A09)	0.14	0.14	**0.32**	**0.38**	**0.48**
CI_{ext}(B06, A09)	0.19	*0.20*	*0.24*	*0.23*	*0.24*
BCI_{int}(B06, A09)	0	0	1	1	1
BCI_{ext}(B06, A09)	0	1	1	1	1
CCI(B06, A09)	0	-1	0	0	0

Table 6: Change of the (compared) cross impact of technology area A05 on technology area B02 from years 2004 to 2008

	2004	*2005*	*2006*	*2007*	*2008*
CI_{int}(A05, B02)	0.12	**0.28**	**0.28**	**0.29**	**0.36**
CI_{ext}(A05, B02)	<0.01	0.01	<0.01	0.01	0.01
BCI_{int}(A05, B02)	0	1	1	1	1
BCI_{ext}(A05, B02)	0	0	0	0	0
CCI(A05, B02)	0	1	1	1	1

6. Summary and conclusions

This paper introduced an analytical Cross Impact Analysis (CIA) approach to support strategy making and R&D planning for organizations with many R&D projects in areas with many relationships between technologies. Unlike traditional qualitative CIA approaches the newly proposed quantitative CIA approach is able to show relative technology impacts and trends for a large number of R&D projects. The quantitative CIA analyzes the cross impacts between selected technologies based on R&D information of a organization. Additionally, the cross impacts between these technologies based on patent data are computed. Both internal and external cross impacts are compared to compute the relative impact between technology pairs as measured by the newly introduced Compared Cross Impact (CCI) index. CCI indices of positive one point out in which technology pairs the organization excels whereas CCI indices of negative one signify technology pairs in which the organization has relative weakness. Comparing CCI(A,B) to CCI(B,A) indicates whether two technologies are equally influencing one another (symmetrical) or whether the impact between two technologies is different (asymmetrical). As such, symmetrical / asymmetrical relative strengths and relative weaknesses are identified for the organization by inspecting the CCI values. However, to facilitate the detection of symmetrical / asymmetrical relative strengths and relative weaknesses a CCI network graph is introduced as an exploratory management tool supporting the organization’s strategy making and R&D planning. The CCI network graph visualizes the overall structure and the complex relationships between several technologies from the organization’s perspective. In a glance, managers can detect (sequential) relative strengths and relative weaknesses from the CCI network graph. Finally, the analysis of changes in the CCI values for technology pairs over time reveals trends in technology impacts thereby signaling which technologies should receive more or less development and investment. Overall, the quantitative CIA approach shows that the CCI supports strategy making and R&D planning for organizations with many R&D projects in areas with many relationships between technologies.

The results of the case study show technology impacts and current trends from the application field ’defence’. The selected R&D information from the German Ministry of Defence (GE MoD) is manually assigned to technology areas from the European Defence Agency (EDA) taxonomy of technologies. Patent data are assigned to these technology areas by use of a centroid-based multi-label text classification approach. The R&D-based cross impact CI_{int}(A,B) is compared to the patent-based cross impact CI_{ext}(A,B) and summarized in the new CCI index CCI(A,B). The CCI between technology area pairs can be used by the GE MoD for research planning and strategy making. For example, the GE MoD has a very strong relative strength in the ’electronic, electrical & electromechanical device technology’ (A05) and the ’propulsion and powerplants’ (B02) technology area pair. Regardless of the GE MoD’s experience with the ’electronic, electrical & electromechanical device technology’ (A05), it has a serious relative weakness in the technology area pair ’electronic, electrical & electromechanical device technology’ (A05) and ’communications and CIS-related technologies (B10). The construction of the CCI network graph suggested several ways to extend the GE MoD relative strengths as pinpointed technology area pairs with no impact to turn into relative strengths. The analysis of the change in CCI showed that the GE MoD excels in the ’electronic, electrical & electromechanical device technology’ (A05) and the ’propulsion and powerplants’ (B02) technology areas since 2005. Overall, the ’defence’ application illustrates how the compared R&D-based and patent-based cross impact analysis can support an organization’s strategy making and R&D planning.

This paper contributed to previous technology impact research in four ways: 1) the introduction of a CCI measure, 2) the characterization of technology impacts as symmetrical or asymmetrical, 3) the presentation of the CCI network graph as exploratory management tool, and 4) the analysis of changes in CCI to discover trends in technology impact. Still there are at least two avenues for future research. The most important avenue of research relates to granularity. The case study focuses on the impact between 32 technology areas. However, a more detailed view at the technology level rather than at the technology *area* level could lead to better R&D planning support and better strategic decision making. Therefore, future research should aim at assigning R&D projects to technologies rather than technology areas. In the case study, internal R&D projects and patent data should be assigned to the 200 defence-based technologies from EDA taxonomy. Then, a more detailed view on the technological landscape in the 'defence' application field could be provided. A second avenue of further research could take the occurrence of new technologies into account. This research focuses on computing the impacts between technologies or technology areas. It does not consider the computation of the occurrence probability of new technologies or technology areas. This could be an interesting topic for future research.

Bibliography

[1] C. Choi, S. Kim, Y. Park, A patent-based cross impact analysis for quantitative estimation of technological impact: The case of information and communication technology, Technol. Forecast. Soc. Change 74 (2007) 1296-1314.

[2] F. Narin, E. Noma, Patents as indicators of corporate technological strength, Res. Policy 16 (2/4) (1987) 143-155.

[3] M.E. Mogee, R.G. Kolar, International patent analysis as a tool for corporate technology analysis and planning, Technol. Anal. Strat. Manag. 6 (4) (1994) 485-503.

[4] M. Hirschey, V.J. Richardson, Are scientific indicators of patent quality useful to investors? J. Empir. Finance 11 (1) (2004) 91-107.

[5] B. Yoon, C. Yoon, Y. Park, On the development and application of a self-organizing feature map-based patent map, R&D Manage. 32 (4) (2002) 291-300.

[6] B. Yoon, Y. Park, A text-mining-based patent network: analytical tool for high-technology trend, J. High Technol. Managem. Res. 15 (1) (2004) 37-50.

[7] B. Yoon, Y. Park, A systematic approach for identifying technology opportunities: keyword-based morphology analysis, Technol. Forecast. Soc. Change 72 (2) (2004) 145-160.

[8] D. Thorleuchter, Finding Technological Ideas and Inventions with Text Mining and Technique Philosophy, in: C. Preisach, H. Burkhardt, L. Schmidt-Thieme, R. Decker (Eds.), Data Analysis, Machine Learning, and Applications, Springer, Berlin, Heidelberg, 2008, pp. 413-420.

[9] G. Tsoumakas, I. Katakis, Multi Label Classification: An Overview, International Journal of Data Warehousing and Mining 3 (3) (2007) 1-13.

[10] C. Trumbach, D. Payne, A. Kongthon, Technology mining for small firms: Knowledge prospecting for competitive advantage, Technol. Forecast. Soc. Change 73 (2006) 937-949.

[11] G.H. Jeong, S.H. Kim, Aqualitative cross-impact approach to find the key technology, Technol. Forecast. Soc. Change 55 (3) (1997) 203-214.

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

[13] A. Schuler, W.A. Thompson, I. Vertinsky, Y. Ziv, Cross impact analysis of technological innovation and development in the softwood lumber industry in Canada: a structural modeling approach, IEEE Trans. Eng. Manage. 38 (3) (1991) 224-236.

[14] H. Dernis, D. Guellec, Using patent counts for cross-country comparisons of technology output, Special Issue on New Science and Technology Indicators, STI Review, vol. 27, OECD, Paris, 2002, 129-146.

[15] M. Trajtenberg, A penny for your quotes: patent citations and the value of innovations, in: A. Jaffe, M. Trajtenberg (Eds.), Patents, Citations and Innovations, MIT Press, Cambridge, MA, 2002.

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

[17] N.C. Dalkey, An elementary cross-impact model, Technol. Forecast. Soc. Change 3 (1972) 341-351.

[18] J.C. Duperrin, M. Godet, SMIC 74 - a method for constructing and ranking scenarios, Futures 7 (4) (1975) 302-312.

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

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

[21] A. Caselles-Moncho, An empirical comparison of cross-impact models for forecasting sales, Int. J. Forecast. 2 (3) (1986) 295-303.

[22] USPTO, Overview of the classification system, USPTO (2009), Available at, http://www.uspto.gov/.

[23] D. Knoke, J. Kuklinski, Network Analysis, Sage, London, 1982.

[24] H.J. Lange, H.P. Ohly, J. Reichertz, Auf der Suche nach neuer Sicherheit: Fakten, Theorien und Folgen, VS Verlag, 2008, p. 11.

[25] D. Thorleuchter, D. Van den Poel, A. Prinzie, Mining Innovative Ideas to Support new Product Research and Development, in: H. Locarek-Junge, C. Weihs (Eds.), Classification as a Tool for Research, Springer, Berlin, Heidelberg, 2010.

[26] A. Hotho, A. Nürnberger, G. Paaß, A Brief Survey of Text Mining, LDV Forum 20 (1) (2005) 19-26.

[27] L. Breiman, Classification and regression trees, in: L. Breiman, J.H. Friedman, R. A. Olshen, C. J. Stone (Eds.), The Wadsworth statistics/probability series, Wadsworth International Group, Belmont, CA, 1984, p. 358.

[28] E.S. Han, G. Karypis, Centroid-Based Document Classification: Analysis and Experimental Results, in: Principles of Data Mining and Knowledge Discovery, Springer, Berlin, Heidelberg, 2000, pp. 116-123.

[29] V. Tam, A. Santoso, R. Setiono, A comparative study of centroid-based, neighborhood-based and statistical approaches for effective document categorization, 16th International Conference on Pattern Recogniton, Iv:235-238, 2002.

[30] S.M. Weiss, N. Indurkhya, T. Zhang, F.J. Damerau, Text Mining, Springer, Berlin, 2005, p. 113.

[31] R. Ferber, Information Retrieval, dpunkt.verlag, Heidelberg, 2003, p. 78.

[32] D. Thorleuchter, D. Van den Poel, A. Prinzie, Extracting Consumers Needs for New Products – A Web Mining Approach, in: Proc. WKDD 2010, IEEE Computer Society, Los Alamitos, CA, 2010, p. 441.

[33] M.F. Porter, An algorithm for suffx stripping, Program 14 (3) (1980) 130-137.

[34] K. Coussement, D. Van den Poel, Integrating the voice of customers through call center emails into a decision support system for churn prediction, Information & Management 45 (2008) 165.

[35] A. J. Abebe, V. Guinot, D. P. Solomatine, Fuzzy alpha-cut vs. Monte Carlo techniques in assessing uncertainty in model parameters, in: Proc. 4-th International Conference on Hydroinformatics, Iowa City, USA, 2000.