Expert Systems with Applications 42 (2015) 1314–1324

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation

Sérgio Moro [a,*], Paulo Cortez [b], Paulo Rita [c]

[a] Business Research Unit (UNIDE-IUL), Department of Information Science and Technology, ISCTE – University Institute of Lisbon, 1649-026 Lisbon, Portugal
[b] ALGORITMI Research Centre, Department of Information Systems, University of Minho, 4800-058 Guimarães, Portugal
[c] Business Research Unit (BRU-UNIDE), ISCTE – University Institute of Lisbon, 1649-026 Lisbon, Portugal

ARTICLE INFO

Article history:
Available online 29 September 2014

Keywords:
Banking
Business intelligence
Data mining
Text mining
Decision support systems

ABSTRACT

This paper analyzes recent literature in the search for trends in business intelligence applications for the banking industry. Searches were performed in relevant journals resulting in 219 articles published between 2002 and 2013. To analyze such a large number of manuscripts, text mining techniques were used in pursuit for relevant terms on both business intelligence and banking domains. Moreover, the latent Dirichlet allocation modeling was used in order to group articles in several relevant topics. The analysis was conducted using a dictionary of terms belonging to both banking and business intelligence domains. Such procedure allowed for the identification of relationships between terms and topics grouping articles, enabling to emerge hypotheses regarding research directions. To confirm such hypotheses, relevant articles were collected and scrutinized, allowing to validate the text mining procedure. The results show that credit in banking is clearly the main application trend, particularly predicting risk and thus supporting credit approval or denial. There is also a relevant interest in bankruptcy and fraud prediction. Customer retention seems to be associated, although weakly, with targeting, justifying bank offers to reduce churn. In addition, a large number of articles focused more on business intelligence techniques and its applications, using the banking industry just for evaluation, thus, not clearly acclaiming for benefits in the banking business. By identifying these current research topics, this study also highlights opportunities for future research.

1. Introduction

Banking has been a prolific industry for innovation concerning information systems and technologies (Shu & Strassmann, 2005). For example, new technologies have enabled new communication channels which were quickly adopted by banks. Also, advanced data analysis techniques are currently used to evaluate risk in credit approval (Huang, Chen, Hsu, Chen, & Wu, 2004) and fraud detection (Ngai, Hu, Wong, Chen, & Sun, 2011).

Business intelligence (BI) is an umbrella term that includes architectures, tools, databases, applications and methodologies with the goal of analyzing data in order to support decisions of business managers (Turban, Sharda, & Delen, 2011). Banking domains, such as credit evaluation, branches performance, e-banking, customer segmentation and retention, are excellent fields for application of a wide variety of BI concepts and techniques, including data mining (DM), data warehouses and decision support systems (DSS). For bank firms to survive and even excel in today's turbulent business environment, bank managers need to have a continuous focus on solving challenging problems and exploiting opportunities. That demands a need for computerized support of managerial decision making thus implying the need of decision support and business intelligence systems.

There are several surveys/reviews of the banking domain. Wilson, Casu, Girardone, and Molyneux (2010) published a recent literature review covering the impact of the global financial crisis in the banking business. Their results put the risk domain as a subject that deserves a deeper attention in order to achieve a systemic stability. The review of Ngai et al. (2011) devoted attention to financial fraud detection, and classified 49 articles depending on the type of fraud. The findings suggest a lack of research on mortgage fraud, money laundering, and securities and commodities fraud, by contrast to a large number of articles on credit fraud. More related with this paper, Fethi and Pasiouras (2010) presented a survey on bank branches performance based on 196 articles which employ operational research and artificial intelligence

* Corresponding author.
E-mail addresses: scmoro@gmail.com (S. Moro), pcortez@dsi.uminho.pt (P. Cortez), paulo.rita@iscte.pt (P. Rita).

http://dx.doi.org/10.1016/j.eswa.2014.09.024
0957-4174/

techniques, concluding that profit efficiency and capacity efficiency have received limited attention in the studies evaluated.

A large research attention has been given toward credit. In fact, although credit is traditionally related to banking, it has long spread to other industries. Therefore, some recent reviews and surveys are naturally available on the subject. Abdou and Pointon (2011) reviewed 214 articles/books/thesis on credit scoring applications, searching for the statistical techniques used for evaluation and found that there is not an overall best technique for building models. The review of Marqués, García, and Sánchez (2012) reports over the use of evolutionary computation for credit scoring. Another subject of interest is e-banking, specifically customer acceptance toward a new communication channel. Dahlberg, Mallat, Ondrus, and Zmijewska (2008) reviewed publications on mobile payments and found through their framework lacking of research on social and cultural factors impacting mobile payments, as well as traditional payment services.

The enlisted surveys and reviews cover some themes in banking. However, within the authors' knowledge, there is a lack of a recent literature analysis for BI applications in the main subjects related to the banking industry, thus motivating the present research. Furthermore, none of the discussed reviews adopted an automated text analysis, by using Text Mining (TM) techniques such as the ones presented in this study, thus facilitating the analysis of a much larger set of sources.

This paper presents an automated text mining literature analysis, from 2002 to 2013, of BI applications within the banking domain, allowing the identification of current research trends and interesting future applications, thus highlighting opportunities for further research. Although BI has been extensively studied, recent years and particularly the last decade have experienced a huge increase in BI applications through the industry, especially in the banking sector, therefore stimulating research. This article is organized as follows. Section 2 introduces the main concepts related with both banking and BI domains, and presents also other references of literature analyses. Next, Section 3 presents the methods used for analyzing the literature. Then, the results are discussed in Section 4. Finally, conclusions are summarized in Section 5, which also presents future research directions.

2. Background

2.1. Text mining

Data mining (DM) aims to extract useful knowledge (e.g., patterns or trends) from raw data (Witten & Frank, 2005). Text mining (TM) is a particular type of DM that is focused on handling unstructured or semi structured data sets, such as text documents (Fan, Wallace, Rich, & Zhang, 2006). Delen and Crossland (2008) proposed the application of TM for analyzing the literature and identify research trends, thus helping researchers in conducting state of the art reviews on a given research subject. Their research focused on three major journals in management information systems, although they argue that their TM approach can be valuable in virtually any research field.

Within a literature analysis, searching with individual words is often not enough, since many searchable terms can be composed of a sequence of words, such as "data mining" or "decision support systems". Those sequences, which can be made of n words, are called n-grams. When extracted from large texts, n-grams constitute a valuable asset, in particularly when analyzing publications, such as the study of Soper and Turel (2012) showed by analyzing eleven years (from 2000 to 2010) of publications in the Communications of the ACM journal.

When conducting TM over text documents, relevant words and terms are often extracted in order to produce a categorization that can help building a body of knowledge over the literature considered (Delen & Crossland, 2008). An interesting approach is modeling a certain number of distinct topics defined according to the number and distribution of terms across the documents, which can be achieved through the latent Dirichlet allocation (LDA) model (Blei, Ng, & Jordan, 2003). For each document, it is determined the probability of belonging to each of the topics, allowing to group documents to the more likely matching topics. This organization structure can help identifying which topics are capturing more attention from researchers and also to find gaps for future research. TM can be used indiscriminately, by looking for the most overall referred words, or through the use of specific dictionary words. Since this work is about a focused literature analysis, a dictionary of terms in both banking and BI domains is used.

2.2. Banking

Banks are institutions that operate in the financial business domain, concerning activities such as loaning, deposits management and investments in capital markets, among others. The banking industry is crucial for the economy and thus it is a subject of great interest for researchers in a widespread of different domains, such as management science, marketing, finance and information technologies. Berger (2003) found evidence of a relation between technological progress and productivity in banking. The same author also emphasizes that banks employ statistical models based on their financial data for different purposes, such as credit scoring and risk evaluation.

Financial sector reforms allowed an increase in competition, turning bank lending an important source of funding. Credit risk evaluation is by its own a vast domain, encompassing a large number of research publications within banking and spread through the last twelve years (e.g., Marqués et al. (2012)). Other banking related subject where research has been active is fraud prevention and detection in traditional banking services (e.g., Abbasi, Albrecht, Vance, & Hansen (2012)) and in new communication channels that support e-banking services (e.g., Shuaibu, Norwawi, Selamat, & Al-Alwani (2013)), from which electronic mail spamming in order to illicitly obtain private financial information is a specific case of interest (e.g., Amayri & Bouguila (2010)). E-banking is also subject of another research domain related to technology acceptance regarding new communication channels adopted by banks (e.g., Vatanasombut, Igbaria, Stylianou, & Rodgers (2008), Lin (2011)). A not so recent theme that however has boomed in research, driven by the global financial crisis, is bankruptcy and related subjects such as systemic risk and contagion (e.g., Hu, Zhao, Hua, & Wong (2012)). Competition had also an effect on client related areas, with banks increasing investment in customer retention, customer relationship management (CRM) and targeting (e.g., Karakostas, Kardaras, & Papathanassiou (2005)).

Research in banking is currently an interesting domain of research. Due to advances in information technology, virtually all banking operations and procedures are automated, generating large amounts of data. Therefore, all the subjects mentioned above can potentially benefit from BI solutions.

2.3. Business intelligence

BI involves several distinct areas and technologies that converge in the common goal of having access to data in order to help businesses by facilitating knowledge and supporting better management decisions. One way to accomplish this is by predicting a certain behavior or result based on data-driven models, in what is known as DM or predictive analytics, thus providing the most likely outcomes to managers (Han, Kamber, & Pei, 2006; Turban et al., 2011; Witten & Frank, 2005).

Intersecting several fields of research, such as artificial intelligence, statistics and databases, several supervised learning DM algorithms have been proposed for building data-driven models. These predictive DM models are classified into two main types: classification, if the output target is a categorical value, and regression, if the target variable is a numeric value. Examples of popular DM models that can be applied to both classification and regression are decision trees, neural networks and support vector machines (Witten & Frank, 2005). There are also other DM goals, such as clustering, which uses unsupervised learning to group similar items. Self-organizing maps is an example of a popular clustering technique. Data warehouses (DW) are another popular BI concept that consists in data repositories for accessing data from different sources, organized in a unique schema and place in order to facilitate information extraction to produce knowledge.

A DSS is an information system that provides assistance in supporting business decision making (Turban et al., 2011). While often used as a synonym of BI, DSS can also use expert knowledge rather than data-driven models (e.g., group DSS). New concepts are emerging related to DSS and BI, such as the adaptive business intelligence, which aims to reduce the gap between supporting and making the decision by adding adaptive prediction and optimization modules to classical BI systems (Michalewicz et al., 2008).

2.4. Literature analysis

A literature review of a set of articles enables to analyze a given subject and identify trends of research and possible gaps that can lead to new studies and discoveries (Levy & Ellis, 2006). In fact, it is considered a critical step and a baseline to unveil new insights on a research subject, thus an enabler and driver of new findings. Such relevance is expressed through the numerous publications on conducting literature reviews across the different sciences (Cronin, Ryan, & Coughlan, 2008; Jesson & Lacey, 2006).

Traditionally, exhaustive literature analyses demand considerable amount of efforts from researchers, in pursuit for the state of the art on a given subject which may serve as a driver on new research. New technologies enabled online library databases, easy to access from any location, offering researchers an enormous amount of available articles. The usage of search queries provided by such libraries facilitates the retrieval of articles on a given subject; however, the high volumes of articles returned present the challenging task of reading the contents of each paper, even though smaller parts of the articles (e.g., title, abstract, keywords) may provide a lead on the research conducted. To address this difficulty, a few TM approaches have been proposed recently for analyzing literature.

Table 1 summarizes four frameworks for literature analysis that use different techniques. The first (Jourdan, Rainer, & Marshall, 2008) uses a traditional human effort approach, while the remaining three conducted TM literature analyses. Finally, in the last row, the characteristics of the present approach are also displayed, to allow a straightforward comparison. The four frameworks were chosen to represent different and recent literature analysis methodologies on research areas closely related to BI, which is the focus of the present research, here applied to the banking industry. It was also taken into account that each of those frameworks should mention the criteria and methods of research, expressed in the columns of Table 1, to enable comparing different approaches with the proposed method.

The work of Jourdan et al. (2008) provides a general review on BI and requires that at least two humans (sometimes three, in cases of different opinion from the two authors) manually read every of the 167 articles. One main advantage of such approach is the fact that a human reader can readily understand the meaning of a word by the context of the remaining text (e.g., "senior" may refer to older people, or to senior professionals, which may not be so old), while an automated approach cannot. However, the time needed to conduct such a manual analysis prohibits it from being applied to a large number of manuscripts.

The remaining three frameworks use TM approaches, analyzing a number of articles greater than a thousand. The work of Sunikka and Bragge (2012) focus on two subjects, still it performs a separate analysis of both results, while the remaining two focus

Table 1
Examples of relevant frameworks for literature analysis and the proposed approach.

Reference	Areas of research	Nr. articles	Nr. journals	Search period	Search query	Description of the techniques used
Jourdan et al. (2008)	Business intelligence	167	10	1997–2006	Business analytics OR business intelligence OR data mining OR data warehousing	Classification by research strategy by 2–3 human coders; Classification of articles by topic using brainstorming and discussions
Delen and Crossland (2008)	Management information systems	1123	3	1994–2005	All articles from the 3 journals	TM on title and abstract of articles, using singular value decomposition to reduce the size of the document term matrix, and then clustering using an expectation–maximization algorithm
Sunikka and Bragge (2012)	Customization and personalization	883 + 1544 (customization + personalization)	457 + 664 (customization + personalization)	1986–2009	Customization OR personalization (two separate searches)	TM (tool: VantagePoint) on articles keywords, using Aduna cluster map of the keywords used; Autocorrelation map of authors with some selected keywords
Bragge et al. (2012)	Multiple criteria decision making	15198	Usage of the Web of Science database (not mentioned the different publication titles found)	1970–2007	Multiple criteria OR multiple attribute OR multiple objective OR goal programming OR vector optimization	TM on articles keywords, using auto-correlation maps based on the 60 most cited authors per decade
Proposed approach	BI in banking	219	14	2002–2013	Described in Section 3.2	TM, using dictionaries to reduce the size of search-space, and then the LDA to group articles

on just a subject of analysis. As a result, the present work, which analyzes BI applications in banking, is the only one from Table 1 using a search query with a conjunction (“AND”) element (explained further ahead in Section 3.2). This justifies the significantly smaller number of articles analyzed in the present article, even though the procedure presented is scalable.

Another difference between the TM approaches is the procedure used to reduce the search space to a manageable number of terms: Delen and Crossland (2008) analyzed the abstract, discarding the keywords, and used a singular value decomposition, while the remaining two frameworks considered only the keywords. The former authors argued that the keywords are generally mentioned in the abstract, and even that some authors select keywords that they would like to be associated with their work. However, it can be argued that the approach of Delen and Crossland (2008) discards relevant terms composed of more than one word such as “data mining” or “decision support systems”, which are included in the present work through the usage of a specific domain dictionary, overcoming both this limitation (Han, Wang, & El-Kishky, 2014) and the one associated with the usage of just the keywords mentioned above. It should be noted also that while all the TM analysis in Table 1 perform clustering analyses, none of the mentioned works used the LDA clustering algorithm. Also no evidence was found of literature analysis using this technique. Its computational complexity of probabilistic inference for finding a large number of topics is considered NP-hard, while unsupervised learning using expectation–maximization requires the repeated computation of marginal probabilities of what topics are present in the documents. LDA model is considered one of the most important probabilistic models in widespread use today (Sontag & Roy, 2011).

3. Materials and methods

3.1. Journal selection

Given the emphasis on technology aspects of BI applications to the banking industry, the articles were chosen from journals more related with technology rather than management. Nevertheless, with the popularity increase of BI (in both industry and research), the corresponding publications have boomed, making a literature review in this domain a challenging task. To select the relevant publications where to search, the focus was set on finding the most influential peer-reviewed journals on BI applications to business and management, within a recent time frame that includes around one decade (last twelve years, 2002–2013).

Instead of defining one specific publication metric criterion (e.g., by using impact factor or number of citations) for selecting journals, the choices were based on literature reviews and publication analysis. It should be noted that there are studies that criticize impact factor rankings accuracy, such as Andersen, Belmont, and Cho (2006) that analyzed the impact factor of the journal citation reports (JCR) published by the Institute of Scientific Information (ISI). The value of survey and review studies on the subjects in analysis is that the journals selected through those were already validated through a deeper analysis of publications rather than just citations considerations. Few articles evaluate the influence of journals on the information systems (IS) domain. To assist in the selection of journals, three review articles were chosen, one from each third of the time frame (i.e., 2002–2005, 2006–2009 and 2010–2013). The criteria for such review article selection included: consider only journal articles but with no restriction regarding journal title; consider reviews on related areas to this study (i.e., BI and banking); consider articles with a list of journals used in their review and the number of articles retrieved for each journal in such list.

The oldest of them (Huang & Hsu, 2005) analyzed publication productivity in IS from 1999 to 2003. Their study also used three other journal reviews as a base of work, and selected 12 reference journals on the field of IS. Ngai, Xiu, and Chau (2009) analyzed literature from 2000 to 2006 on a more specific field related to the research presented here, DM and its applications to CRM, by reviewing 87 articles from 28 different journals. Finally, the more recent study of Chen, Chiang, and Storey (2012) focused in BI and analytics and its impact to business through a literature review on those subjects in the past decade (2000–2011). Those three studies were also selected in order to be complementary in terms of the domains of IS, CRM and BI, thus providing with a vaster choice of journals.

The criteria for selection of journals is to include every journal used in at least one of the three reviews mentioned, except for:

- non technical journals, which are more related to business and management, were excluded, such as the MIT Sloan Management and the Harvard Business Review; and
- since the review of Ngai et al. (2009) presented a very large list of references, journals cited only once or twice in this review were discarded, except for Information & Management, which was selected since it was also used in the review of Huang and Hsu (2005).

The final result is a list of fourteen journals from eight different publishers (Table 2) that set the sources in this study for searching relevant articles.

Table 2
Journals selected and search results.

Journal	Publisher (search engine)	[1]☆	[2]☆	[3]☆	Hits
Expert systems with applications	Elsevier (SciVerse Science Direct)		X		126
Decision support systems		X	X	X	25
European journal of operational research			X		48
Information & management		X	X		2
IEEE Trans. knowledge and data engineering	IEEE (IEEE Xplore)		X		2
IEEE intelligent systems			X		2
Information systems research	INFORMS (Informs online)	X		X	0
Journal on computing				X	1
Journal of the association for IS	Association for IS (AIS elect. library)	X		X	1
Communications of the association of IS				X	1
Data mining and knowledge discovery	Springer (Springer link)		X		4
Communications of the ACM	ACM (ACM digital library)	X			1
Journal of management IS	M. E. Sharpe (jmis-web.org)	X		X	3
MIS Quarterly	MIS research center (misq.org)	X		X	3
					Total: 219

A “X” indicates the journal was used as a source in the corresponding column reference.
☆ [1] = Huang and Hsu (2005); [2] = Ngai et al. (2009); [3] = Chen et al. (2012).

3.2. Article search

The searches were performed through each of the publishers online search engine. Most of the search engines are optimized, allowing complex search queries through the use of specific fields and Boolean operators "AND" and "OR". It should be noted that a few of the search engines did not provide the flexibility needed (e. g. Boolean AND/OR operators, search field specification), thus the search was partitioned in searches within the main search.

The query used is the same for every journal, and consists in a Boolean expression containing two OR connected expressions, one for banking terms and another for BI, and both are connected through an AND, meaning that any article should include at least one banking term and another BI term:

> (banking OR bank OR credit) AND ("business intelligence" OR "data mining" OR "decision support system" OR "knowledge discovery" OR "business analytics" OR forecasting OR "modern optimization" OR modeling OR "machine learning" OR "artificial intelligence" OR prediction OR predictive).

The composition of such query is always subjective. To reduce such subjectivity, the authors and two banking domain experts conducted several broader searches with single keywords such as "banking" and "business intelligence", reaching to a consensus consisting in the query presented above. Some remarks should be mentioned. First, credit is a subject on its own, although closely related to banking, so it is considered in the search. For BI terms, the choice is on high-level concepts, discarding specific methods and techniques such as data warehouses, neural networks and decision tables.

All searches were performed in 2014, with the corresponding journal 2013 volumes already published, and included only the article title, abstract and keywords, since those are the most visible article areas where, if a certain concept is relevant, should be mentioned. It should also be noted that some online databases search engines only allow searching in these types of contents, rendering unfeasible a full-text search.

The first search results included a total of 240 articles. A manual analysis, consisting in reading each title, abstract and keywords, detected several articles where the terms occurred with a different meaning, such as "blood bank" or "credit" mentioned in a non-financial context. This manual pruning led to a pool of 219 articles. Table 2 shows each journal contribution in terms of search hits (where each hit denotes an article).

3.3. Text mining for literature review

Since 219 articles is quite a large number for a manual analysis, in this study TM was used to facilitate in producing organized information to analyze the literature. Considering the goal is set specifically on applications of BI to banking, in order to keep the scope within a manageable list of terms, it makes sense to define a dictionary that encompasses both BI and banking more common terms and concepts, rather than let the TM algorithms to search, group and count words indiscriminately. Hence, two dictionaries were defined, one for banking and another for BI, each of them containing a list of terms composed of one or more words (n-grams).

Stemming is a technique often applied in TM, in order to reduce similar words to a unique term (e.g., "banking" and "banks" are transformed in "bank"). Rather than just performing usual stemming, an extended list of related terms was created that includes other concepts in the same domain. For example, "loyalty" and "lifetime value" are the opposite of "defection" and "churning", but all of them concern with the problem of customer "retention", thus all of them were grouped under this reduced term.

Both the definition of dictionaries and the grouping of terms under a unique reduced term are subjective. To reduce this subjectivity, the three authors of this paper analyzed all decisions. It should be mentioned that, while all three authors are experienced in information systems and BI, one of them is a full-time information systems manager in a retail bank since 2001, having coordinated projects in distinct areas such as marketing and risk. Additionally, two experienced banking professionals in different areas were consulted (one of them has 3 years as a technical Contact Center support, and 10 years as a technician in Marketing, while the other has 6 years in the Commercial Area, plus one year in the Risk Department).

To further extend the validation of the dictionaries, considering these will guide the entire TM approach, and also the relatively small number of articles, each of the articles was analyzed in terms of title, abstract and keywords for prospecting the adequacy of the terms in each dictionary for the articles. For a large number of articles, an alternative would be to pick up a reasonable randomly selected number of articles for validating the dictionary.

The resulting dictionaries and grouping of terms defined for banking and BI are shown in Tables 3 and 4, respectively.[1] Some considerations should be made about the dictionaries. First, both the terms "banking" and "business intelligence" were not included, since are the two broader terms that characterize every article found. An also relevant term that was not included is risk and its variations, since it is a research subject by its own and it is implicit to other specific banking domains such as credit scoring, fraud and bankruptcy detection and churning (considering the risk of losing customers).

For the literature analysis of the articles collected, the full-text is considered. Since this analysis encompasses two distinct areas, BI and banking, it is likely that some of the terms from the dictionaries may not be present in the title + abstract + keyword, for they are not the main focus of the research (e.g., certain BI techniques applied). Also the full-text analysis allows a better evaluation of term frequencies, since a term expressed numerous times through an article is probably more relevant than another that is only mentioned in the abstract. The exception is the references section, which was pruned from all articles. By proceeding this way, it is assured that no term from the dictionary will match any from publication titles cited in the article. If some term in the dictionary is relevant for some study, then it is likely mentioned through the article text.

The TM procedure adopted included several steps over the corpus of documents, for striping extra whitespaces, converting all words in lowercase, reducing the terms of the dictionary to a common term, and finally defining the document term matrix, which is a bi-dimensional representation used as an input for the LDA (the dimensions are the articles and the terms, and each cell contains the frequency which $term_x$ appears in $article_y$).

There are a wide variety of tools and software that can be used to perform TM. For this review, the **R** statistical tool was chosen (<www.r-project.org>), since it is open source and provides a high flexibility through the installation of packages. In particular, the **tm** package chosen was adopted, since it offers a large number of functions for managing text documents and provides an abstraction of the process of document manipulation (Meyer, Hornik, & Feinerer, 2008).

For demonstration purposes, part of the R code is exposed (Code 1). This code was used first to create the corpus of documents based on a path containing all documents (line 1), perform cleaning by removing extra spaces (line 2) and converting all words to lowercase (line 3). Then the list of equivalent terms for reducing them to a common unique term (Tables 3 and 4) are loaded into

[1] Also available online at: <https://fenix.iscte.pt/homepage/smcmo@iscte.pt/BIinBankingReview>.

Table 3
Dictionary for the banking domain.

Reduced term	Similar terms or from the same domain*
Bankruptcy	systemic risk, crisis, contagion, financial distress, solvency
Branches	bank branch, banking center, financial center
Central bank	central banks
Credit	loan
crm	customer relationship management
Deposit	savings, bank account, bank accounts, deposits
e-bank	e-banking, electronic banking, electronic bank, homebanking, homebank, home banking, home bank, internet banking, internet bank, online banking, online bank, netbanking, net banking, netbank, net bank, mobile banking, m-banking, m-bank, sms banking, sms bank, mobile bank, technology acceptance, tam
Fraud	fraud detection, fraud evaluation, fraud detect, fraud prevention, fraud risk, money laundering
Interest rate	interest rates, annual percentage rate, annual percentage rates, bank rate, bank rates, borrowing rate, borrowing rates, lending rate, lending rates, prime rate, prime rates, rate of interest, rates of interest
Investment	investments
Retention	defection, churning, churn, loyalty, lifetime value
Segmentation	client segment, profiling, client profile, client profiles, customer profile, customer profiles
Stocks	stock price, stock exchange, stock market, commodity, commodities
Targeting	direct marketing, database marketing, telemarketing, cross-selling

* All terms are in lower case and separated by commas.

Table 4
Dictionary for the BI domain.

Reduced term	Similar terms or from the same domain*
Adaptive	adaptive
Analytic	analytics, data science, data sciences
Artificial intelligence	machine learning, intelligent agent, intelligent agents
Association rule	association rules
Big data	terabytes, massive data
cbr	case-based reasoning
Classification	classifier, classifiers
Cluster	clusters, clustering, clusterings
Data mining	data miner, datamining
Data warehouse	datawarehouse, datawarehouses, data warehouses
Decision support system	decision support systems, expert system, expert systems
Decision table	decision tables
Decision tree	decision trees, random forest, random forests
Genetic algorithm	genetic algorithms, genetic programming
Knowledge discovery	knowledge discovering
Modeling	modeling, data model
Naive bayes	naivebayes, bayesian
Neural network	neural networks, artificial network, artificial networks, multilayer perceptron, multilayer perceptrons
Optimization	optimize
Predict	prediction, predictive, predicting, forecasting, forecast
Regression	time series, time series
Self-organizing map	self-organizing feature map, self organizing map, sofm, kohonen map, kohonen network
Set theory	rough set, rough sets, fuzzy set, fuzzy sets, sets theory
Support vector machine	support vector machines

* All terms are in lower case and separated by commas.

a lookup table (line 5) and the reduced terms (first element of the R lookup table list) are checked against the dictionaries previously loaded through the intersect function, constituting the reduced terms dictionary (line 6). Next follows a computationally expensive mapping to perform a stem function which uses the terms in the lookup table to reduce them to a common term (line 7). Finally line 10 defines a function to allow tokens up to three words (the maximum words for the terms in the considered dictionaries) and line 11 builds the document term matrix (Delen & Crossland, 2008; Meyer et al., 2008).

3.4. Classification of topics

To obtain a structure that groups articles in order to allow a deeper analysis, the R package **topicmodels** is a logical choice, since it takes advantage of the data structures produced by the **tm** package in order to provide basic infrastructure for fitting topic models (Hornik & Grün, 2011).

Within the **topicmodels** package, the latent Dirichlet allocation (LDA) algorithm (Blei et al., 2003) is implemented and can be applied by receiving just two parameters, the document term matrix created for the TM and the desired number of topics. The result is a complex structure from which can be obtained the topics and terms that define it, characterized through a beta (β)

R Code 1: Creating the corpus, cleaning it and build the document term matrix

```
1  articles <- Corpus(DirSource(pathOut), readerControl = list(
       language = "en"))
2  articles <- tm_map(articles, stripWhitespace) # remove spaces
3  articles <- tm_map(articles, tolower) # lower case
4
5  termDomains <- stemFromFileLoad("equivalent.txt")
6  reducedDictionary <- as.vector(intersect(unique(termDomains
       [[1]]),dictionary))
7  articles <- tm_map(articles, function(x) stemFromFile(doc=x,
       equivTerms=termDomains))
8
9  # create the document term matrix
10 phraseTokenizer <- function(x) RWeka::NGramTokenizer(x,
       Weka_control(min = 1, max = 3)) # terms up to 3 words
11 dtm <- DocumentTermMatrix(articles, control = list(tokenize =
        phraseTokenizer, dictionary = reducedDictionary))
```

For a general characterization of the literature, the frequency of each term was obtained for the combined dictionary including both Tables 3 and 4. Also a word cloud was designed to allow a visual interpretation of the obtained results.

distribution computed for each term for a given topic. Also for each article, it can be obtained the likelihood of matching it to each of the topics. In this study, only the most probable topic according to LDA for a given article was considered. Also, the three most significant terms for characterizing each topic according to the β distribution will be analyzed.

As stated previously, one necessary parameter for LDA is the number of topics. Following the approach of Delen and Crossland (2008), this value was set to half of the terms considered. To simplify the analysis conducted, the topics will be presented in tables referring the number of articles in each topic published through the considered period of the last twelve years.

4. Results and analysis

The results are presented in two Sections: in the first, the results are analyzed based on term frequencies for the whole 219 articles collected. The respective results are shown using a table and a word cloud, which uses a larger font size for the most frequent terms. After the global analysis, the topics generated with LDA are displayed and analyzed. In the second Section, a representative article for each topic is selected and scrutinized with the goal to understand if the trend suggested by the topic characterization is aligned with such article.

4.1. Text mining and latent Dirichlet allocation topics

The global results are presented in Table 5, with a total of 38 terms. The respective word cloud is shown in Fig. 1. Overall, the BI terms are much more evenly distributed: credit is the top term, followed by four BI terms, and next comes two banking related terms, fraud and bankruptcy. This is an expected result, since banking defines the problems being addressed, to which many different BI solutions can be applied, including more specific algorithms and tools or more general approaches, such as modeling and knowledge discovery. Only three of the fourteen bank terms are among the eleven most cited. This global analysis allows taking a glimpse on what seems to be an interesting hypothesis to test: most of the BI research efforts are directed toward a few (and probably more relevant) of the banking domains. The word cloud on Fig. 1 seems to help support this claim, since it makes more visible that credit is the dominant term, followed by several BI terms, and only then comes the next two banking problems: fraud and bankruptcy. The second level analysis, using a LDA parameterized to 19 topics and presented in Table 6, is more interesting for this study, as it allows to relate BI terms to banking problems, thus identifying research trends and eventually gaps for further research. Each topic is presented in horizontal lines, with the column labeled "topics" presenting the most relevant terms and β distribution values (converted to positives, since they are used only for comparison purposes) in respect to a given topic (defined by the row). The number of articles column presents the number of articles that were included in the topic and that were published through the analyzed twelve year period.

The results of Table 6 show an increasing although not steady interest in BI applied to banking. For each topic, there is always a dominant term, with a β value that matches it to closer to a certain banking problem or to a type of BI technique, tool or context. Given that the three most relevant terms are shown for each topic, most of them have at least one of the top 3 terms belonging to banking and another to BI, which enables to analyze each topic as a BI application to banking. Still, there are four topics that focus specifically on BI (topics 3, 9, 14 and 16), with the three dominant terms matching all BI terms, and one equivalent topic for banking (topic 5).

Table 5
Most relevant term frequencies for the BI applied to banking.

#	Term	Frequency
1.	Credit	7299
2.	Predict	4053
3.	Regression	2022
4.	Decision support system	1765
5.	Neural network	1735
6.	Fraud	1358
7.	Bankruptcy	1152
8.	Decision tree	997
9.	Data mining	874
10.	Cluster	839
11.	Modeling	793
12.	e-bank	621
13.	Investment	545
14.	Retention	536
15.	Interest rate	493
16.	Artificial intelligence	444
17.	Deposit	389
18.	Optimization	382
19.	Support vector machine	379
20.	Classification	336
21.	Set theory	335
22.	Data warehouse	267
23.	Naive bayes	261
24.	cbr	260
25.	Genetic algorithm	235
26.	Adaptive	204
27.	Association rule	168
28.	Branches	158
29.	Segmentation	157
30.	Stocks	139
31.	Targeting	118
32.	crm	108
33.	Central bank	67
34.	Knowledge discovery	64
35.	Decision table	47
36.	Analytic	42
37.	Self-organizing map	33
38.	Big data	3

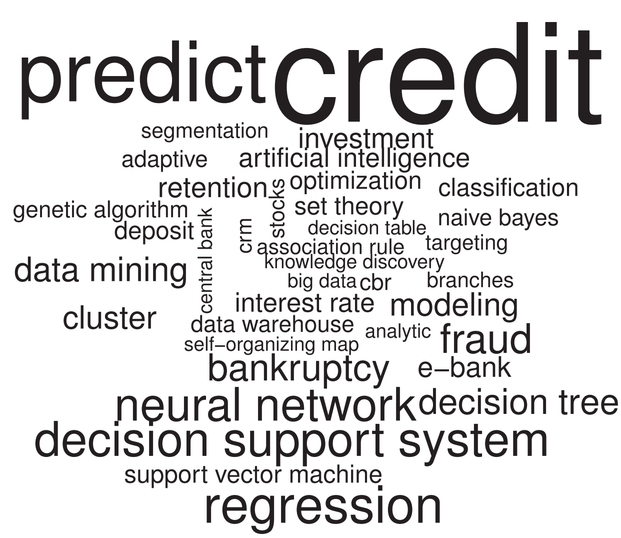

Fig. 1. Word cloud for BI applied to banking.

The topic best identified with credit gets 70 matching articles, although second and third terms for this topic, predict and segmentation, have a significantly higher β value (greater than 3.3), meaning that its relation is not so tight. This puts emphasizes on numerous applications of BI to benefit credit business and risk evaluation. In fact, credit gets into the top 3 of six more topics

Table 6
Relevant topics for BI applied to banking.

Topic	#	1st Term		2nd Term		3rd Term		2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
		term	β	*term*	β	*term*	β												
1.	70	Credit	0.08	Predict	3.34	Segmentation	4.37	2	0	2	4	4	6	3	6	13	13	11	6
2.	25	Predict	0.15	Set theory	2.58	Stocks	3.65	0	0	1	3	1	2	1	8	1	2	5	1
3.	22	Neural network	0.85	Predict	1.18	Support vector machine	2.43	0	0	2	1	1	3	0	6	0	4	4	1
4.	12	Credit	0.80	Neural network	1.59	Adaptive	2.41	1	0	0	1	1	2	0	2	1	0	3	1
5.	12	Retention	0.89	Interest rate	1.07	Targeting	2.50	1	0	2	1	1	1	1	0	0	2	2	1
6.	9	Fraud	0.26	Classification	3.06	Regression	3.43	0	0	1	0	0	0	0	2	0	2	3	1
7.	8	Optimization	0.96	Deposit	1.19	Branches	1.67	0	0	0	1	0	0	0	2	4	0	1	0
8.	7	Decision tree	0.57	Classification	1.83	Credit	3.20	0	0	2	1	0	1	1	1	0	0	1	0
9.	7	Decision support system	0.19	Naive bayes	2.06	Adaptive	4.39	1	1	0	0	0	0	0	3	1	0	1	0
10.	7	Bankruptcy	0.25	Predict	2.18	Deposit	2.83	0	0	0	0	2	1	0	0	0	2	1	1
11.	7	Regression	0.09	Predict	3.10	Credit	4.14	0	1	0	0	1	1	0	1	0	1	1	1
12.	6	Cluster	0.13	Credit	3.08	Neural network	4.09	0	0	0	0	1	1	1	1	2	0	0	0
13.	5	e-bank	0.09	Decision support system	3.09	predict	4.26	0	0	0	0	1	0	0	1	2	0	0	1
14.	5	Artificial intelligence	0.69	Association rule	1.20	Decision table	2.75	0	1	1	0	0	0	0	1	0	0	0	2
15.	5	Modeling	0.36	Credit	1.40	optimization	3.96	0	0	0	0	0	0	0	0	0	2	2	1
16.	5	Data mining	0.28	Decision support system	2.56	Knowledge discovery	2.88	0	0	0	0	0	0	0	2	0	2	1	0
17.	4	Investment	0.03	Predict	4.35	Analytic	4.70	0	0	0	1	0	0	0	0	0	1	2	0
18.	2	cbr	0.35	Credit	2.00	Decision support system	2.68	0	0	0	0	0	0	0	1	1	0	0	0
19.	1	Data warehouse	0.18	Decision support system	2.71	Investment	3.56	0	0	1	0	0	0	0	0	0	0	0	0

β corresponds to the correlation between the topic and term; # is the number of articles in the topic.

while being also the top term for the fourth topic, confirming the diversity of this subject.

The year of 2008 seems to be an outlier, containing a smaller number of articles when compared to its surrounding years (only seven articles). Probably the global financial crisis, which culminated in 2008 with the failure of major financial institutions, also helped to boom research in the following year of 2009, with a total of 37 articles for the set analyzed. The second topic, with 25 articles in total, had eight publications just for 2009, the highest number for the topic in the twelve years studied. Furthermore, the topic includes stocks as the third more relevant term, while predict and set theory are the first and second, respectively.

Concerning the banking domain, fraud and bankruptcy prediction get a match of nine (topic 6) and seven articles (topic 10) respectively, although most of them are recent, which can be also a result of the financial crisis. Neural networks are the dominant specific learning technique adopted, topping the third topic with more articles (22). Topic 5, with 12 articles, has the three most relevant terms for banking only: retention, interest rate and targeting. This is an interesting topic, since it shows an evenly distributed publication number for the period considered, with most years having just one or two articles, with the exception of the years 2003, 2009 and 2010. Considering that the three terms have significantly close β values, one can hypothesize that by targeting customers with attractive interest rates in the products offered may also serve the purpose of retaining them, thus reducing churn.

DSS are a thematic rather old, but far from outdated. From the topics in Table 6, it is possible to confirm the wide reference to DSS by counting five occurrences of the term decision support systems in different topics, with an apparent even distribution in the years considered. On the other hand, data mining has only one reference in the top 3 terms for every topics, which is on topic 16, with just 5 articles. This an unexpected result, since the state of the art for prediction is the application of data mining techniques. Nevertheless, it should be noted that dominant data mining techniques include neural network and regression, which have several references spread through the 19 topics.

In respect for the four topics which are best identified by three terms all related to BI, and some other topics, one may hypothesize that it is probably an indication that the main focus is on BI applications, not evaluating in-depth benefits to banking.

Looking at the end of the table, data warehouse is surprisingly low in publications, although banks continue to invest on those systems as a way to unify data otherwise spread through an organization. Other recently proposed terms for BI, such as adaptive (Michalewicz, Schmidt, Michalewicz, & Chiriac, 2005) (mentioned in topics 4 and 9, but only as the third most relevant term in both cases), still get few publications and others are not even on any of the top 3 terms list (e.g., big data).

4.2. Analysis of representative articles per topic

In previous Section, LDA was applied to unveil topics which group articles, characterized by the terms identified on Table 6, suggesting the major trends of research concerning BI applications to banking. However, such automated approach has a significant limitation (Thomas, McNaught, & Ananiadou, 2011): document clustering is completely dependent on the technique used for creating the clusters, which is based on term identification; the problem consists in terms with different meanings based on the remaining text (e.g., risk may refer to credit default risk or to bankruptcy risk). In this Section, this issue is addressed by identifying the most representative articles for each topic. Then a full text manual analysis of each of the nineteen articles is performed in order to confirm or not the hypotheses suggested by the topics found. Table 7 identifies the articles chosen.

Considering the fact that the three most relevant terms were selected for characterizing the topics (Table 6), in order to select the most relevant article two metrics were considered, by the following order of relevance: the number of different terms mentioned in each article (from one to the whole three most relevant terms, displayed for each topic), and the total number of times each of the three terms occurred, regardless of the specific term. Such procedure is best explained through an example: topic 10

Table 7
Core article per topic.

Topic	Article	Different terms	Frequency
1.	Chi and Hsu (2012)	3	186
2.	Kumar and Ravi (2007)	3	189
3.	Huang et al. (2004)	3	221
4.	Malhotra and Malhotra (2002)	3	175
5.	Prinzie and Van den Poel (2006)	3	27
6.	Abbasi et al. (2012)	3	426
7.	Azadeh et al. (2012)	3	27
8.	Sinha and May (2004)	3	208
9.	Ben-David and Frank (2009)	2	89
10.	Hu et al. (2012)	3	167
11.	Zhao et al. (2011)	3	179
12.	Lim and Sohn (2007)	3	71
13.	Gu et al. (2009)	3	162
14.	Hsieh (2004)	1	52
15.	Liu et al. (2012)	3	67
16.	Chen et al. (2011)	3	64
17.	Soper et al. (2012)	3	190
18.	Park et al. (2009)	3	123
19.	Hwang et al. (2004)	3	255

grouped a total of seven articles; for each of those articles, the number of occurrences for each of the three most relevant terms was extracted from the document term matrix (hence resulting in a sub-matrix limited to the seven articles and the terms "bankruptcy", "predict" and "deposit" from topic 10); next, the sub-matrix was ordered decreasingly by the number of different terms and the number of times each of the three terms occurred, resulting in four articles which mentioned the three terms, and three referring only two terms; from those four articles, the one with higher frequency of terms was selected, which was the study of Hu et al. (2012).

Topic 1 is best represented by the work of Chi and Hsu (2012), which is a typical research for predicting credit risk of default, suiting perfectly in the two most relevant terms, "credit" and "predict" (Table 6), whereas "segmentation" (the third most relevant term) is also used in their work for defining homogeneous risk groups. This study confirms the hypothesis arose from the previous Section, which pointed this research trend of predicting credit behavior as the major application for BI to banking.

By looking at the title of the article best identified with topic 2, the work of Kumar and Ravi (2007), "Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review", one could argue why it did not match topic ten, which focus precisely on predicting bankruptcy. However, a deeper analysis of such article revealed it is a work more focused in applying set theory as well as other techniques for comparing their performance when addressing a prediction problem, which happens to be bankruptcy. In fact, the third term for topic ten is "deposit", while for topic two is "stocks", which is much more related to bankruptcy: it is mentioned several times through the text.

Topic 3 is more focused on the techniques applied rather than the banking problem itself, which fits perfectly with the chosen article (Huang et al., 2004). This study is focused on comparing machine learning techniques, using corporate credit rating for benchmarking their performance.

While seventeen of the nineteen topics were best matched by one article which referred the three most relevant terms (from Table 6), there remain two for which the best article only contained two of the three most relevant terms (topic 9) or just one (topic 14). In case of topic 9, the significantly higher β value for "adaptive", the third term, more than twice as the second term, may justify the result. However, topic 14 shows clearly a weakness of this approach: although it groups five articles, none is related with more than one term of the three most relevant (e.g., the work of Hsieh (2004) is dedicated to association rules, without even mentioning the remaining two terms). One may hypothesize that this is a direct consequence of the ill-posed problem of clustering: the data-driven nature of clustering makes it very difficult to correctly find clusters in the given data (Jain, 2010). LDA faces the same challenge of other clustering algorithms, implying that there will inevitably exist articles that cannot match to any of the existing topics, leading to issues such as the one in topic 14.

5. Conclusions

This literature analysis paper focused on the main banking problems and BI solutions used to solve them. Banking is a competitive industry where innovation thrives, due to the importance of this sector for the economy, thus making it an attractive field for researchers. Banking is also a domain that generates large amount of data and where BI applications can potentially benefit business, increasing the visibility and recognition of research achievements. This recent analysis encompassed the last twelve years (2002–2013), being a period that includes the effect of the global financial crisis and its impact on research on this sector. Thus, this study can potentially benefit researchers by allowing the identification of new research trends and possible gaps for future research.

For analyzing literature, a text mining approach using the latent Dirichlet allocation was performed. As a result, several topics grouping articles were found, being each of those topics characterized by the three most relevant terms. Such topics suggested several research trends. However, intrinsic limitations of clustering algorithms such as LDA have lead efforts toward validating the hypotheses for relations between the several terms and corresponding trends. To address this issue, the most relevant article for each topic was exhaustively analyzed, in order to confirm the hypothesized trends. Expert systems conducting automated text analyses may benefit from the suggested procedure with little additional human effort for analyzing the few more relevant documents.

The most relevant conclusion is that credit maintains its status as the dominant field of research in the banking industry. Other relevant banking subjects are fraud and bankruptcy, mainly for detection and prevention, thus mitigating risks taken by banks. Concerning BI, the main goal consists in prediction, rather than modeling and knowledge discovery, which emphasizes the importance of estimating what is going to happen on the future in order to better support decision-making. There are some studies that use banking problems to test and evaluate BI techniques and tools, but possibly not accounting for real business benefits for banks, since banking terms are lesser relevant for those articles. Regarding the evolution of publications per year, 2009 is a milestone year, triggering a boom in the research publications on the domains analyzed. Most likely, this effect was motivated by business pressures due to the global financial crisis. Still, through the time period studied, publications related with BI approaches applied to banking had a steady increase until 2012, indicating this is a domain application much studied. Nevertheless, research has diminished in 2013, although some lack of research in newer concepts such as big data may suggest there is still open room for research.

The results highlight some possibly interesting research gaps. DSS in banking is a subject far from exhaust. Emerging concepts such as adaptive BI and optimization can be applied to enhance DSS and improve banking efficiency in several areas. For example, targeting customers to sell deposits is an application domain where there seems to be a lack of research. Although some articles mention deposits and others targeting, none of these words top any of the topics computed. Still concerning customer domains, it is interesting to verify that CRM is not a top banking domain for BI applications. This comes as a surprise, since CRM is a subject where research has been quite active, although the results here presented show it is not the case for the banking industry.

With the intensifying global competition within the financial sector and namely involving the banking industry, CRM has become critical. Thus, future studies in this topic are paramount in order to understand clearly what is more successful according to the size and nature of the financial organizations being at stake. E-banking offers a wide spectrum of services to customers. Some involve non-transactional tasks such as viewing of account balances or recent transactions, downloading account and bank statements. Others demand real transactions like fund transfers, bill payments, loan applications and transactions, investments in stocks and bonds. Banks offering all these services online are becoming financial "supermarkets" and demand further research in this area. Mobile devices such as smartphones and tablets are at the forefront of electronic consumer products. Their penetration is increasing rapidly in a diverse range of markets across the globe. Mobile banking solutions are nowadays a major challenge to the banking sector in order to be able to adapt its approach to new customer demands and expectations. Hence, this is another important focus for further research.

Bankruptcy associated with systemic risk is also a recent interesting subject, with its visibility set to a high level thanks to the global financial crisis which is far from over. Furthermore, it is now known that prior to the crisis, systems failed to predict it, and prediction is precisely a top keyword for BI, thus applications for this case must also be enhanced in the next years in order to try to prevent future financial crisis.

Acknowledgments

The authors would like to thank the valuable effort of the two bank professionals who helped in defining both the search query and the dictionary used. Also, the authors wish to thank the anonymous reviewers for their helpful comments.

References

Abbasi, A., Albrecht, C., Vance, A., & Hansen, J. (2012). Metafraud: A meta-learning framework for detecting financial fraud. *MIS Quarterly*, 36.
Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting, Finance and Management, 18*, 59–88.
Amayri, O., & Bouguila, N. (2010). A study of spam filtering using support vector machines. *Artificial Intelligence Review, 34*, 73–108.
Andersen, J., Belmont, J., & Cho, C. T. (2006). Journal impact factor in the era of expanding literature. *Journal of Microbiology, Immunology, and Infection/Wei Mian yu Gan Ran za Zhi, 39*, 436–443.
Azadeh, A., Saberi, M., & Jiryaei, Z. (2012). An intelligent decision support system for forecasting and optimization of complex personnel attributes in a large bank. *Expert Systems with Applications, 39*, 12358–12370.
Ben-David, A., & Frank, E. (2009). Accuracy of machine learning models versus "hand crafted" expert systems–A credit scoring case study. *Expert Systems with Applications, 36*, 5264–5271.
Berger, A. N. (2003). The economic effects of technological progress: Evidence from the banking industry. *Journal of Money, Credit, and Banking, 35*, 141–176.
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993–1022.
Bragge, J., Korhonen, P., Wallenius, H., & Wallenius, J. (2012). Scholarly communities of research in multiple criteria decision making: A bibliometric research profiling study. *International Journal of Information Technology & Decision Making, 11*, 401–426.
Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36.
Chen, J., Wu, G., Shen, L., & Ji, Z. (2011). Differentiated security levels for personal identifiable information in identity management system. *Expert Systems with Applications, 38*, 14156–14162.
Chi, B.-W., & Hsu, C.-C. (2012). A hybrid approach to integrate genetic algorithm into dual scoring model in enhancing the performance of credit scoring model. *Expert Systems with Applications, 39*, 2650–2661.
Cronin, P., Ryan, F., & Coughlan, M. (2008). Undertaking a literature review: A step-by-step approach. *British Journal of Nursing, 17*, 38–43.
Dahlberg, T., Mallat, N., Ondrus, J., & Zmijewska, A. (2008). Past, present and future of mobile payments research: A literature review. *Electronic Commerce Research and Applications, 7*, 165–181.
Delen, D., & Crossland, M. D. (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems With Applications, 34*, 1707–1720.
Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the power of text mining. *Communications of the ACM, 49*, 76–82.
Fethi, M. D., & Pasiouras, F. (2010). Assessing bank efficiency and performance with operational research and artificial intelligence techniques: A survey. *European Journal of Operational Research, 204*, 189–198.
Gu, J.-C., Lee, S.-C., & Suh, Y.-H. (2009). Determinants of behavioral intention to mobile banking. *Expert Systems with Applications, 36*, 11605–11616.
Han, J., Kamber, M., & Pei, J. (2006). *Data mining: Concepts and techniques*. Morgan Kaufmann.
Han, J., Wang, C., & El-Kishky, A. (2014). Bringing structure to text: Mining phrases, entities, topics, and hierarchies. In *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1968). ACM.
Hornik, K., & Grün, B. (2011). Topicmodels: An r package for fitting topic models. *Journal of Statistical Software, 40*, 1–30.
Hsieh, N.-C. (2004). An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert Systems with Applications, 27*, 623–633.
Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: A market comparative study. *Decision Support Systems, 37*, 543–558.
Huang, H.-H., & Hsu, J. S.-C. (2005). An evaluation of publication productivity in information systems: 1999 to 2003. *Communications of the Association for Information Systems*, 15.
Hu, D., Zhao, J. L., Hua, Z., & Wong, M. (2012). Network-based modeling and analysis of systemic risk in banking systems. *MIS Quarterly*, 36.
Hwang, H.-G., Ku, C.-Y., Yen, D. C., & Cheng, C.-C. (2004). Critical factors influencing the adoption of data warehouse technology: A study of the banking industry in Taiwan. *Decision Support Systems, 37*, 1–21.
Jain, A. K. (2010). Data clustering: 50 Years beyond k-means. *Pattern Recognition Letters, 31*, 651–666.
Jesson, J. K., & Lacey, F. M. (2006). How to do (or not to do) a critical literature review. *Pharmacy Education, 6*, 139–148.
Jourdan, Z., Rainer, R. K., & Marshall, T. E. (2008). Business intelligence: An analysis of the literature. *Information Systems Management, 25*, 121–131.
Karakostas, B., Kardaras, D., & Papathanassiou, E. (2005). The state of crm adoption by the financial services in the UK: An empirical investigation. *Information & Management, 42*, 853–863.
Kumar, P. R., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques–A review. *European Journal of Operational Research, 180*, 1–28.
Levy, Y., & Ellis, T. J. (2006). A systems approach to conduct an effective literature review in support of information systems research. *Informing Science: International Journal of an Emerging Transdiscipline, 9*, 181–212.
Lim, M. K., & Sohn, S. Y. (2007). Cluster-based dynamic scoring model. *Expert Systems with Applications, 32*, 427–431.
Lin, H.-F. (2011). An empirical investigation of mobile banking adoption: The effect of innovation attributes and knowledge-based trust. *International Journal of Information Management, 31*, 252–260.
Liu, Y., Zhang, H., Li, C., & Jiao, R. J. (2012). Workflow simulation for operational decision support using event graph through process mining. *Decision Support Systems, 52*, 685–697.
Malhotra, R., & Malhotra, D. (2002). Differentiating between good credits and bad credits using neuro-fuzzy systems. *European Journal of Operational Research, 136*, 190–211.
Marqués, A., García, V., & Sánchez, J. (2012). A literature review on the application of evolutionary computing to credit scoring. *Journal of the Operational Research Society, 64*, 1384–1399.
Meyer, D., Hornik, K., & Feinerer, I. (2008). Text mining infrastructure in r. *Journal of Statistical Software, 25*, 1–54.
Michalewicz, Z., & Michalewicz, M. (2008). Machine intelligence, adaptive business intelligence and natural intelligence. *Computational intelligence magazine* (Vol. 3, pp. 54–63). IEEE [research frontier].
Michalewicz, Z., Schmidt, M., Michalewicz, M., & Chiriac, C. (2005). Case study: An intelligent decision support system. *Intelligent systems* (Vol. 20, pp. 44–49). IEEE.
Ngai, E., Hu, Y., Wong, Y., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems, 50*, 559–569.
Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems With Applications, 36*, 2592–2602.
Park, Y.-J., Choi, E., & Park, S.-H. (2009). Two-step filtering datamining method integrating case-based reasoning and rule induction. *Expert Systems With Applications, 36*, 861–871.
Prinzie, A., & Van den Poel, D. (2006). Investigating purchasing-sequence patterns for financial services using Markov, mtd and mtdg models. *European Journal of Operational Research, 170*, 710–734.
Shuaibu, B. M., Norwawi, N. M., Selamat, M. H., & Al-Alwani, A. (2013). Systematic review of web application security development model. *Artificial Intelligence Review*, 1–18.
Shu, W., & Strassmann, P. A. (2005). Does information technology provide banks with profit? *Information & Management, 42*, 781–787.

Sinha, A. P., & May, J. H. (2004). Evaluating and tuning predictive data mining models using receiver operating characteristic curves. *Journal of Management Information Systems, 21*, 249–280.
Sontag, D., & Roy, D. (2011). Complexity of inference in latent Dirichlet allocation. In *Advances in neural information processing systems* (pp. 1008–1016).
Soper, D. S., Demirkan, H., Goul, M., & St Louis, R. (2012). An empirical examination of the impact of ict investments on future levels of institutionalized democracy and foreign direct investment in emerging societies. *Journal of the Association for Information Systems, 13*, 116–149.
Soper, D. S., & Turel, O. (2012). An n-gram analysis of communications 2000–2010. *Communications of the ACM, 55*, 81–87.
Sunikka, A., & Bragge, J. (2012). Applying text-mining to personalization and customization research literature–who, what and where? *Expert Systems with Applications, 39*, 10049–10058.
Thomas, J., McNaught, J., & Ananiadou, S. (2011). Applications of text mining within systematic reviews. *Research Synthesis Methods, 2*, 1–14.
Turban, E., Sharda, R., & Delen, D. (2011). *Decision support and business intelligence systems*. USA: Prentice Hall Press.
Vatanasombut, B., Igbaria, M., Stylianou, A. C., & Rodgers, W. (2008). Information systems continuance intention of web-based applications customers: The case of online banking. *Information & Management, 45*, 419–428.
Wilson, J. O., Casu, B., Girardone, C., & Molyneux, P. (2010). Emerging themes in banking: Recent literature and directions for future research. *The British Accounting Review, 42*, 153–169.
Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann.
Zhao, H., Sinha, A. P., & Bansal, G. (2011). An extended tuning method for cost-sensitive regression and forecasting. *Decision Support Systems, 51*, 372–383.