International Journal of Information Technology & Decision Making
Vol. 19, No. 1 (2020) 97–126

DOI: 10.1142/S0219622019300076

World Scientific
www.worldscientific.com

A Research Review and Taxonomy Development for Decision Support and Business Analytics Using Semantic Text Mining

Andrea Ko[*,‡] and Saira Gillani[†,§]

[*]*Department of Information Systems*
Corvinus University of Budapest
H-1093 Budapest, Fővám tér 13-15
Budapest, Hungary

[†]*Department of Computer Science*
Bahria University, Karachi Campus
Karachi, Pakistan
[‡]*andrea.ko@uni-corvinus.hu*
[§]*sairagilani@yahoo.com*

Published 28 January 2020

By 2018, business analytics (BA), believed by global CIOs to be of strategic importance, had for years been their top priority. It is also a focus of academic research, as shown by a large number of papers, books, and research reports. On the other hand, the BA domain suffers from several incorrect, imprecise, and incomplete notions. New areas and concepts emerge quickly; making it difficult to ascertain their structure. BA-related taxonomies play a crucial role in analyzing, classifying, and understanding related objects. However, according to the literature on taxonomy development in information systems (IS), in most cases the process is *ad hoc*. BA taxonomies and frameworks are available in the literature; however, some are excessively general frameworks with a high-level conceptual focus, while others are application or domain-specific. Our paper aims to present a novel semi-automatic method for taxonomy development and maintenance in the field of BA using content analysis and text mining. The contribution of our research is threefold: (1) the taxonomy development method, (2) the draft taxonomy for BA, and (3) identifying the latest research areas and trends in BA.

Keywords: Taxonomy; taxonomy development; business analytics; text mining; semantic technology; ontology.

1. Introduction

Taxonomies and ontologies play a key role in decision support and business analytics (BA), as effective decision-making requires knowledge of the underlying data

[§]Corresponding author.

structure and semantics. Clarification of terminology in these fields is important to business, as well as to data management and quality. The importance of BA has recently grown in tandem with the rising interest in big data, big data analytics and machine learning. BA remains the top priority for 2018, according to Gartner CIO Survey, and is considered strategic.[1] BA has kept its top position for 5–6 years; new BA-related competencies and jobs have appeared in organizations.[13–15] The BA field is increasingly complex, as new areas and concepts emerge. To make reliable business decisions, decision-makers must be familiar with BA and its key terms and they need to use them in a uniform way. To have a solid knowledge repository and to maintain reliable data quality, it is necessary to use consistent definitions and the same structure of concepts. This signifies the major importance of decision support and BA taxonomies; however, their development in information systems (IS) follows an *ad hoc* process in most cases.[4] Compared to several other domains, e.g., medical research, no mature BA taxonomies and ontologies are available in the literature, despite an increasing need for them in business and academia and despite the popularity of the field itself among researchers and practitioners. Our paper aims to present a novel semi-automatic method of taxonomy development and maintenance in BA using content analysis and text mining. We identify the latest research areas of the field as well.

Numerous papers on taxonomies in various areas, e.g., production, healthcare, etc., confirm the importance of taxonomies. We found 298 taxonomy-related papers in the Decision Support Systems and Electronic Commerce journal in December 2017, when we started to prepare the first version of BA taxonomy. However, we found only a few papers published by this journal that aimed to develop an ontology for BA, decision support and related fields in prior years.[2,5] Thirty-seven taxonomy-related papers were published in the International Journal of Information Technology & Decision Making journal in March 2019 and two of them are partially related to BA taxonomy. Peng *et al.*[81] provided a comprehensive framework for data mining and knowledge discovery (DMKD) using grounded theory. Zhang and Segall[82] present an overview and evaluation of web mining research and techniques available.

The present research discusses a semi-automatic method for taxonomy development and maintenance in BA while identifying the latest research topics and trends in this domain. The novelty of our research is the following: (1) the semi-automatic characteristics of our solution is better than fully automatic approach, because we have an opportunity to involve expert opinion, which can increase the accuracy and reliability of our result; (2) the combination of different technologies; visual analytics, clustering and our text mining solution called Promine provide unique features, which is different from existing methodologies; (3) we apply a new similarity measure in text mining. This paper is structured as follows: first, taxonomy development challenges are outlined, and related works are discussed. We detail the special features of the BA domain and present BA taxonomy development initiatives. Next, we detail our research into a semi-automatic method of taxonomy development and maintenance, followed by the discussion of the results. Finally, the conclusion is summarized.

2. Related Work — Literature Review

In this section, we discuss taxonomy development methods in Sec. 2.1 and provide an overview of BA taxonomies in Sec. 2.2.

2.1. *Taxonomy development*

One of various aspects of taxonomies' importance is a reduction in complexity of a domain due to a structure being imposed on the domain's objects. The classification of objects helps researchers and practitioners understand and analyze complex domains. Taxonomies play a key role in common understanding of a domain, in knowledge codification, structuring and also support knowledge base construction. The reduction of complexity and the identification of similarities and differences among objects are major advantages provided by taxonomies.[6,7] Taxonomy development has been researched extensively. Nickerson *et al.*[4] performed a comprehensive survey of literature on taxonomy development in IS, concluding that the process was *ad hoc* in most cases. The researchers analyzed 65 papers to identify methods used for taxonomy development. Each paper was classified by its principal domain: IS, computer science (CS), and non-IS business (Bus), and by the development approach: inductive, deductive, and intuitive. The inductive method includes observing and subsequently analyzing empirical cases to determine taxonomy's features. The analysis can apply statistical techniques, such as cluster analysis, or less rigorous techniques. The deductive approach derives a taxonomy through a logical theory- or conceptualization-based process. In the intuitive approach, the researcher uses his or her understanding of objects and this method is therefore essentially *ad hoc*. According to the authors' research, most papers (25) belong to the intuitive category, 11 applied statistical analysis and were hence classified under the inductive approach, 13 relied on informal analysis and were hence deemed to follow the inductive approach, while another 13 belonged to the deductive category. The authors provide a formal definition of a taxonomy and determined the characteristics of a useful taxonomy. According to their research, a proper taxonomy is concise, inclusive, comprehensive, extensible, and explanatory.

Science mapping analysis as a powerful bibliometric technique describes how disciplines, fields, specialties, and individual documents or authors are related to one another as a spatial representation. López-Herrera *et al.*[78] applied this method for the investigation of research themes of the first 10 years (2002–2011) of International Journal of Information Technology & Decision Making. Their bibliometric map was based on co-word analysis and provided an interesting insight into the main topics being discussed in the journal in these years. They highlighted the most productive themes (according to published papers) and the most impacting ones (according to received citations). Cobo *et al.*[79,80] applied science mapping to analyze fuzzy set theory field and the research on intelligent transportation systems. Their science mapping analysis tool, SciMAT[81] provides a workflow with three key modules: cleaning and preprocessing the raw bibliographical data, application of bibliometric

measures to study the impact of each studied element, and configuration of the analysis. Kang *et al.*[90] applies two Bayesian models DWET and HDWET to explore the latent semantic dimensions as the context in natural language. These two models outperform all baseline methods. Their proposed method works with contextual information to get the latent semantic dimensions. Their method predicts emotion for both word and document text. However, Bayesian models cannot converge if the semantic dimensions increase to any significant degree. Lv *et al.*[91] use social network analysis methods to analyze recent advances in transportation research. They summarize the main topics in traffic related research using social media data, and analyze the current collaboration patterns from the perspectives of researchers, institutions, and countries, which does not exhaustively discuss the representative methods adopted in detection processes.

Meijer *et al.*[8] propose a framework for automatic taxonomy construction called "Automatic Taxonomy Construction from Text" (ATCT). This framework has four stages: first, terms are extracted from the domain corpus; the second stage involves filtering for domain-relevant terms. Subsequently, a word disambiguation technique is applied to generate concepts. In the last stage, relations between concepts are determined by applying a submission technique.

Ontology development from text, and ontology learning (OL) types are discussed by Al-Arfaj and Al-Salman.[9] Ontology is a fundamental part of the semantic web. Gruber[10] provides the common definition of ontology in ICT: "An ontology is a formal, explicit specification of a shared conceptualization." The terms "taxonomy" and "ontology" are occasionally used synonymously; however, researchers make a point of distinguishing between them, providing clear definitions of each.[11] An ontology often includes a subclass-based taxonomic hierarchy; however, extra properties are added to the latter, compared to a taxonomy. OL is a process of creating an ontology automatically or semi-automatically. OL approaches can be classified using different dimensions: (1) by the type of knowledge resources used for OL, i.e., the data format: structured, semi-structured, or unstructured; (2) by the level of automation, with certain approaches being semi-automated and requiring user intervention, while others potentially being fully automated, with the system managing the entirety of the construction process; (3) by the learning target, i.e., concepts and relations, or definitions of concepts and axioms; and last, (4) by the purpose of the OL process, either to create an ontology from scratch or to enrich an existing ontology.[9] The authors also discuss and present a comparison of several well-known OL tools. The evaluation dimensions are the elements learned, the primary techniques used, the learning sources, the extent of user intervention, and the approach to evaluating results. Delir Haghighi *et al.*[5] developed a domain ontology (DO4MG) for mass gatherings. To construct an ontology, the authors first identified the scope and the objective of DO4MG, and subsequently prepared a corpus of the domain for knowledge acquisition. The corpus used for ontology construction consists of 27 scientific papers on emergency management during mass gatherings, several major journal and conference papers on emergency and crisis management,

and a public report manual. In this phase, the researchers, helped by domain experts, extract the domain concepts; however, whether the concept extraction process was manual or automated remains unclear. In the next stage, the authors use Protégé 4.0 (https://protege.stanford.edu/) to implement the DO4MG ontology. They tested the resulting ontology on a case-based reasoning decision support system for emergency medical management during mass gatherings. Basole *et al.*[12] provide a multidisciplinary classification and analysis of scholarly development of the literature on enterprise-level IT innovation adoption by examining papers from over the past three decades (1977–2008). The authors discovered new research trends and patterns across disciplines. To create a classification of relevant literature, the researchers used the previously identified subject areas within supporting disciplines, created five broad research streams, and subsequently collected all journals related to such five streams. To perform text analysis, the authors used Northern Light's MI Analyst engine to perform analysis and classification.[12] This engine is used to identify the key relationships and extract meaning from the corpus.

We analyze taxonomy development methods in the following dimensions: the main techniques used, learning sources, the way of user intervention, and the related domain. This comparison by dimensions provide the background for research gap identification. Table 1 summarizes the taxonomy development methods, their distinctive factors and the research gaps by dimensions. Typical learning sources used for taxonomy development are research papers from the literature,[12,24,25,78–80] while some authors use Wordnet and available ontologies as well.[71,72] In a case, when researchers manually select papers from different journals, the number of selected papers varying depending on the source and potentially affecting results. Taxonomy domains are diversified, from the medical domain to transportation, but there is no taxonomy focusing on BA.

User interventions are various, from the purely manual process[20,24,25] to the automatic methods. In a primarily manual process, a domain expert's intuition and knowledge of the domain will affect results. Experts' intervention is common in the selection of papers and in the evaluation of the results phases. Knowledge acquisition is usually not automatic, but is instead facilitated by a domain expert.[5] Manual evaluation by experts is a usual approach, as in Text-To-Onto[71] and OntoLearn.[72] Delir Haghighi *et al.*[5] evaluated DO4MG by using a case-based reasoning decision support system for emergency medical management during mass gatherings. Semantic approaches appear in evaluation in few investigated cases. Semantic precision, semantic recall, and the taxonomic F-measure are applied in the evaluation of ATCT.[8]

The main techniques applied in taxonomy development are diverse. Statistical and machine learning methods are the most popular ones, especially clustering, while the semantic techniques are rarer.[5,8] Science mapping analysis as a powerful bibliometric technique is also used for exploring the conceptual structure of a particular research field.[77–80] Semantic technologies are seldom combined with the previously mentioned statistical and machine learning methods.

Table 1. The major factors distinguishing the taxonomy development methods.

Taxonomy development method	Main techniques used	Learning sources	User intervention	Focus (domain)
Alter[25]	Human expertise	Literature review	Whole process	Decision support system
Basole *et al.*[12]	Statistical & text analysis methods	Research papers	Papers selection and classification	IT innovation and adoption
Cobo *et al.*[79]	CoPalRed, science mapping	Research papers	Paper selection and evaluation	Fuzzy sets theory
Cobo *et al.*[80] and Cobo *et al.* (2012)	SciMAT; science mapping	Research papers and documents	Paper selection and evaluation	Intelligent transportation systems
Delir Haghighi *et al.*[5]	Protégé 4.0	Domain corpus	Knowledge acquisition & Evaluation	Medical emergency management
Kang *et al.*[90]	Bayesian inference method	Blog papers	Whole process	Social media
López-Herrera *et al.*[78]	CoPalRed; science mapping	Research papers	Paper selection and evaluation	The first decade (2002–2011) of the International Journal of Information Technology & Decision Making
Lv *et al.*[91]	Social network analysis method	Multiple electronic citation databases, full text databases, and search engines	Paper selection and evaluation	Transportation
Meijer *et al.*[8]	Statistical & semantic approaches; ATCT	Text corpus	Evaluation	Economics and management; health and medicine

Table 1. (*Continued*)

Taxonomy development method	Main techniques used	Learning sources	User intervention	Focus (domain)
Nickerson's Taxonomy Development Method (2013)	Empirical approach or a conceptual approach	Literature review	Whole process	Information systems
Ontogen (2007)	Statistical analysis and clustering	Free text	Evaluation	General
OntoLearn (2005)	Linguistic analysis, machine learning and statistics	Free text and WordNet	Evaluation	General
Text-To-Onto (2000)	Statistical approach, pruning techniques and association rules	Free text, dictionaries and ontologies	Evaluation	General
Trieu[20]	Human expertise	Literature review	Paper selection and evaluation	Business intelligence
White[24]	Human expertise	Literature review/vendor reports	Paper/report selection and evaluation	Business intelligence
Our approach	Statistical and semantic approaches; clustering	Literature, domain corpus, WordNet	Evaluation	Business analytics
Research gap	Utilization of semantic technologies are limited. Semi-automatic solutions are rare.	Typical learning sources are limited for research articles of certain domains from the literature.	User intervention, manual processes are common, especially in evaluation phases, which increase subjectivity.	There is no taxonomy targeting BA domain

2.2. *BA and the related taxonomies*

BA, business intelligence, and data science have become popular due to the emergence of big data. WOS search for the term "BA" according to topic and title resulted in 373 journal papers published between 2009 and 2018. Figure 1 demonstrates that the publication frequency of academic journal papers related to BA has increased continuously, especially since 2012.

The business and IT communities started to use the term "business intelligence" from the 1990s, while "BA" was introduced to represent the key analytical component in business intelligence in the late 2000s.[16] Chen *et al.*[17] distinguish three phases in the history of business intelligence and BA (BI&A), denote BI&A 1.0, 2.0, and 3.0 according to Gartner BI reports on platforms' core capabilities and the hype cycle. As of 2018, BA, believed by global CIOs to be of strategic importance,[65] has for years been their top priority. It is also a focus of academic research, as shown by a large number of papers, books, and research reports.[2] BA-related skills are valuable in the labor market, with data analytics having become a mandatory core competency for professionals of all types beginning in 2017.[26] However, the BA field suffers from several incorrect, imprecise, and incomplete notions.[2] Terms, such as "business intelligence," "BA," "data analytics," "big data," "data mining," and "data warehousing" are often used interchangeably in the literature.[20] Big data has been characterized in the literature as having one or more of five dimensions: volume, velocity, variety, veracity, and value.[83,84] Chen *et al.*[17] uses business intelligence and analytics (BI&A) as a unified term. Turban *et al.*[18] define business intelligence as the set of techniques and tools for the transformation of raw data into meaningful and useful information for business analysis/decision support purposes. The terms "BA" and "business intelligence" are occasionally used synonymously. Larson and Chang[19] consider the emerging trends in business intelligence and explore the evolution of agile principles and practices with business intelligence, as well as the challenges of

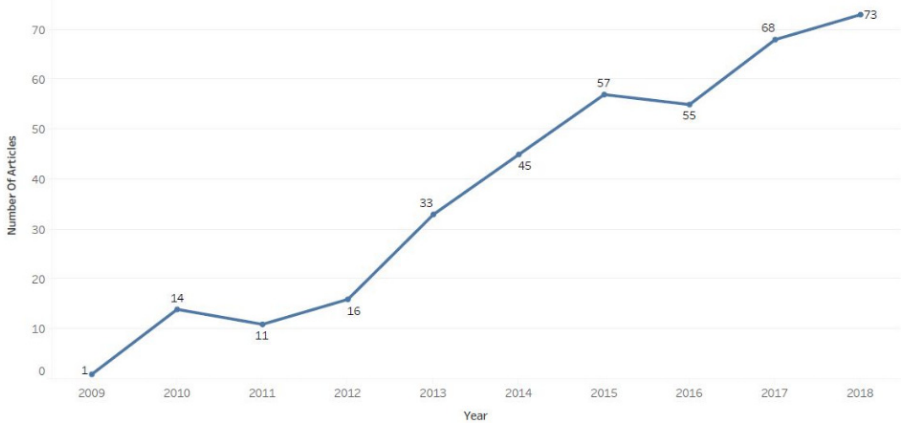

Fig. 1. BA academic journal papers from 2009 to 2018.

business intelligence and future directions. They propose an agile BI framework to compare the traditional BI and fast analytics lifecycles. The objective was to investigate the alignment between agile principles and BI delivery, fast analytics, and data science through the proposed framework. In respect to fast analytics, the researchers emphasize the need for a structure, as the majority of data was unstructured. They discuss the current challenges and future directions for adopting business intelligence platforms, applications, and services in all types of organizations.[19]

Trieu[20] explore business intelligence-related literature; however, the author focuses on a specific field, namely, the value creation aspects of business intelligence systems. The aim was to identify the parts of the BI business value process that have been studied and remain the most underexplored.[20] The author emphasizes, as we do as well, that the BI-related literature is fragmented and lacks an overarching framework for integrating findings and systematically guiding research. He provides a framework of BI value creation. The research framework combines the models of Soh and Markus,[21] Melville *et al.*,[22] Schryen[23] for IS business value. According to this framework, the author identifies five themes in the structure of research gaps and proposes research approaches, namely, context/environmental factors, the BI conversion process, the BI use process, the BI competitive process, and the latency effects. The limitations of their work include that the review process was manual, with two experts coding the literature and only the first 20 papers having been coded by both. White[24] notes the importance of BI-related taxonomies. The researcher distinguishes BI applications and platforms. The former include BI types, stores, rules, granularity, and latency subdomains, described by concepts. The latter consist of data integration suites, BI development suites, planning, and prediction tools, BI application packages, and the subdomains of BI delivery and collaboration suites that are described.[24] BA is a relatively new term and there is no common or established academic definition for it. Davenport and Harris explain BA as "the use of data, information technology, statistical analysis, quantitative methods, and mathematical or computer-based models to help managers gain an improved insight into their operations, and make better, fact-based decisions" (p. 7).[87] Holsapple defines BA as the "evidence-based problem recognition and solving that take place within the context of business situations."[2] In the Sharda *et al.* approach BA are defined as "the application of models directly to business data. BA involve using DSS tools, especially models, in assisting decision-makers (p. 393)."[18] "BA is a systematic thinking process that applies qualitative, quantitative, and statistical computational tools and methods to analyze data, gain insights, inform, and support decision-making." (p. 13) according to the Power *et al.*[85] Nerkar states that "BA provides the insight and understanding to support informed decisions and confident actions, and provides the feedback that is needed to create a learning organization" (p. 3).[88] Gartner IT glossary defines BA as being "comprised of solutions to build analysis models and simulations to create scenarios, understand realities, and predict future states."[89] According to these definitions, BA is an emerging discipline, covering

numerous activities and tasks. Several definitions emphasize that the goal of BA is to provide insights from data and support the decision-making process. The BA field is developing and changing rapidly, with the latest trends including ease-of-use and agility.[65] Advanced self-service BA platforms targeting end-users, e.g., Tableau, PowerBI, and Microstrategy solutions, appeared 12–13 years ago. Their appearance is a paradigm shift in this area, because people with no IT background, e.g., call center workers and sales managers, were able to perform analytical tasks. Currently, we are in the next phase of BA evolution. The common data preparation tasks, such as data imports and data quality checking, are no longer delegated to specialists, while the latest BA solutions interfaces are available in natural languages and with them users are able to select readymade models from BA frameworks.

New areas and concepts emerge quickly in the BA field, making it difficult to ascertain their structure. BA-related taxonomies play a crucial role in analyzing, structuring, and understanding related objects. BA taxonomies and frameworks are available in the literature; however, some are excessively general frameworks with a high-level conceptual focus, while others are application domain-specific.[66] Additionally, maintenance is a key issue in this rapidly changing domain.

Alter[25] developed a taxonomy for decision support systems that has been widely adopted; however, dating back to 1977, it cannot include the latest research fields. Holsapple *et al.*[2] presented a holistic framework/taxonomy for BA that summarized dimensions that were suitable for examining BA' possibilities. Such dimensions are domain, orientation, and technique. Domains are the subject fields that analytics are being applied to. Orientation refers to a direction of thought. The proposed framework includes six distinct classes of analytics: movement, collection of practices and technologies, transformational process, capability set, and activity type set as well as a decisional paradigm.[2] White[24] distinguishes BI applications and platforms in a BI-related taxonomy. The drawback of such taxonomy is that such subdomains are disjunctive sets. Trieu[20] investigates the value creation dimension of business intelligence systems. The researcher note the fragmentation of the literature and the absence of an integrated framework. Delir Haghighi *et al.*[5] construct a Domain Ontology for Mass Gatherings (DO4MG) and present an application of the DO4MG to an implementation of a case-based reasoning decision support system for medical emergency management during mass gatherings. Capgemini introduces a widely used threefold taxonomy for BA in 2010.[2,27,28] The company distinguishes descriptive, predictive, and prescriptive analytics within BA. Such taxonomy is common in BA literature.[3,17,18,86] We apply this structure in our paper and use it in the initial version of the BA taxonomy.

3. Semi-Automatic Method for Taxonomy Development and Maintenance in BA — Research Overview

The primary objective of our paper is to provide a semi-automatic method for taxonomy development and maintenance for the BA domain. The domain itself is

important and has been a focus of research in the past, due to the increasing interest in big data, data analytics, and data science.[14] Our research approach to taxonomy development beyond the taxonomy construction and maintenance additionally provides a method for identification of the latest research topics and trends in BA; hence, it can also be used for the literature review. We discuss the detailed research method and the corresponding system in the following Sec. 3.1. This includes text analytics that apply automated methods to extract and discover knowledge in unstructured data sources. Compared to the traditional literature analyses that are often time- and resource-intensive, text analytics offer automated or semi-automated solutions. Text analytics have not been used extensively to analyze literature, except in certain domains, e.g., biology, biomedicine, and bioinformatics, where researchers have mined various data repositories (e.g., MedLine) to identify gene functionality, molecular interactions, and disease progression.[29–31]

3.1. *Research steps and methods*

An outline of the research process is shown in Fig. 2, while Fig. 3 details the primary research steps. Our research involves six major phases (Fig. 3). Corpus preparation was performed first. We analyze this corpus with visual analytics and text mining solutions to identify the major areas, keywords and their relations, which helped the development of the first version of the BA taxonomy. Applying a visual analytics solution to the corpus, we were able to highlight the primary BA areas, and the associated metadata and characteristics. To obtain a clearer picture, we perform text mining to further investigate the corpus. This resulted in clusters and cluster descriptions (in terms of keywords) that we combined with the manually created

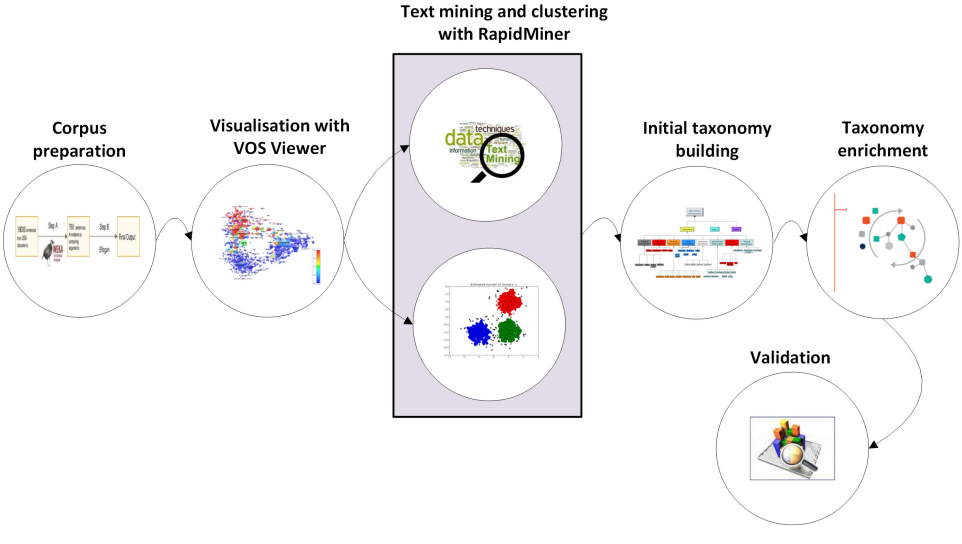

Fig. 2. Big picture of the research.

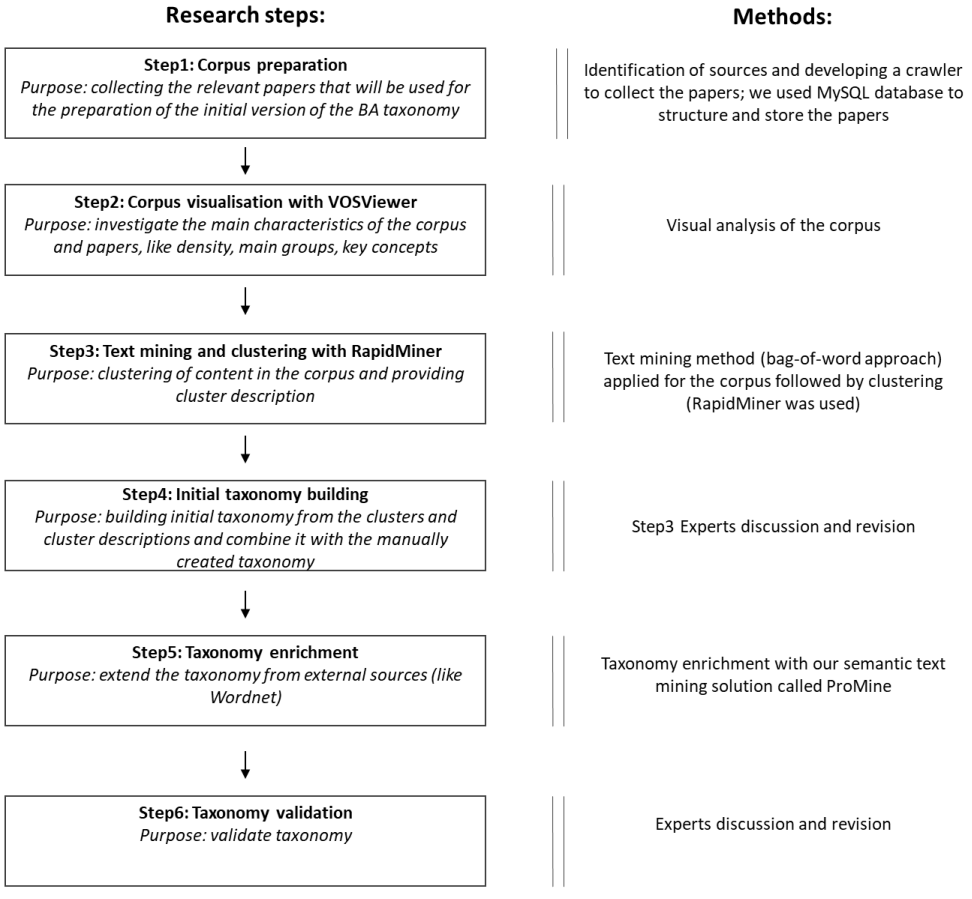

Fig. 3. Research steps.

taxonomy. The outcome of this phase is the initial BA taxonomy that highlights the latest research directions and trends.

Next, we apply a semi-automatic method, using our semantic text mining solution, to enrich the initial taxonomy (step 5 in Fig. 3). In this phase, we use our semi-automatic text mining solution, called ProMine.[32] Finally, we validate the resulting taxonomy (step 6 in Fig. 3).

Section 3.1.1 will detail the research steps.

3.1.1. *Corpus preparation*

Corpus building in general includes selecting the relevant research domain, publication source and period. Our domain of interest is BA; we select the Decision Support Systems and Electronic Commerce journal as the source for the corpus. The choice of the corpus to comprise papers published by this journal was made because it is a leading journal in the field (as shown by being ranked in SCImago Q1 category

from the beginning of 2000) that often publishes BA research, and, furthermore, one with a solid expert base and, among journals in the decision-making field, the best impact factor in 2017 and the highest number of Scopus citations. We apply a semantic text mining approach as a novel method of content analysis, relying on 42 volumes of the journal published between December 2010 (Volume 50) and November 2016 (Volume 91) that comprised 590 papers. The corpus included abstracts and keywords from the previous 5 years; we used the selected papers to build a MySQL database that supported a structured description of papers' metadata. In text mining, it is common to use a corpus based on abstracts and keywords, as such metadata help identify the primary message of each paper. Several researchers suggest this approach, e.g., O'Mara-Eves *et al.*[33] propose a (semi)-automated text mining approach to screening in a systematic review. Thomas *et al.*[34] propose this approach in automatic term recognition (ATR), while Guan *et al.*[35] follows a similar approach in the field of production research.

3.1.2. *Visual analysis of the corpus*

The second step of research involves a visual analysis of the corpus. Such an inspection helped us explore the primary segments of topics and facilitated further content analysis, e.g., clustering. It helped us analyze and explore the primary areas of BA research. It allowed us to create bibliometric maps and facilitated easy-to-interpret displays of large text-based maps.[36] We apply VOSviewer[37] to carry out visual analysis. VOSviewer presents a network map and a density map based on the preprocessing of the corpus content. A density visualization provides us with information on the possible primary groups/classes of terms. Such information is useful in the next phase (clustering) for identifying the possible number of clusters, and for taxonomy development. Network visualization (Fig. 4) and density visualization provided six primary segments. The parameters we used were full-counting (all occurrences of a term were counted in a corpus), with the minimum number of occurrences of a term set at 10 (resulting in 13,705 terms in our study); the relevance score is created, and the most relevant 60% of terms are selected (in our case, 251 terms remained). The segment to which it belongs (Fig. 4) determines an item's color. An item's label size and circle size depend on its weight.[37] Finally, selected terms are verified manually, resulting in the deletion of irrelevant terms.

3.1.3. *Clustering of corpus content with text mining*

A visual preliminary analysis of applying VOSviewer to the corpus resulted in six major segments. We use this information to build a text mining process in order to cluster the corpus content and obtain a corpus description. After clustering the corpus, the "bag-of-words representation" provided the clusters' descriptions. In the "bag-of-words representation" approach, a text document is represented by the set of words it contains. Such an approach is common in text mining research.[38,39] In general, document preprocessing includes such varied natural language processing

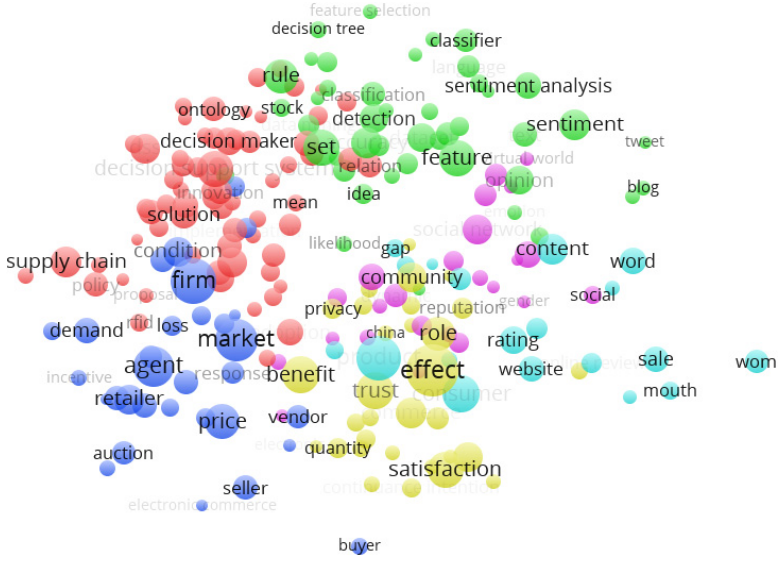

Fig. 4. Network visualization.

(NLP) and text mining techniques, as tokenization, removal of stop words, stemming or lemmatization, part-of-speech (POS) tagging, and frequency count.[40,41] The text mining process is carried out using RapidMiner, a popular data science platform with advanced text mining functionality.[42,43] RapidMiner provides a GUI-based platform for all data scientists. The core open-source code is available to expert data scientists who prefer to develop programs on their own. Gartner named it a leader in the 2017 Gartner Magic Quadrant for data science platforms,[44] while the Forrester Wave report also recognized its leading position in predictive analytics and machine learning in 2017.[45] The steps in our text mining process were as follows: read database (i.e., read the corpus content in the form of abstracts and keywords from the MySQL database), nominal-to-text (i.e., prepare text from the database output), process documents (document preprocessing), and clustering (Fig. 5).

Content preprocessing is carried out during the "process document" step. This step supported clustering and made it possible to determine the distance between content elements (in our case, the abstract, and keywords). We apply k-means clustering during the text mining process, where k was six, according to the visual analysis of the corpus in step 2. The k-means clustering algorithm is a popular unsupervised learning algorithm, using a set of k representatives, around which the clusters are built. One challenge arising during its application involves determining in advance the value of k, the number of clusters. To this end, we use visualization, as described in the previous section. The step "process document" included tokenization, filtering of stop words, stemming and filtering tokens by length. Tokenization (i.e., sentence segmentation) involves segmenting unstructured text into tokens or words, used as the processing units during the subsequent steps.[46] The process of

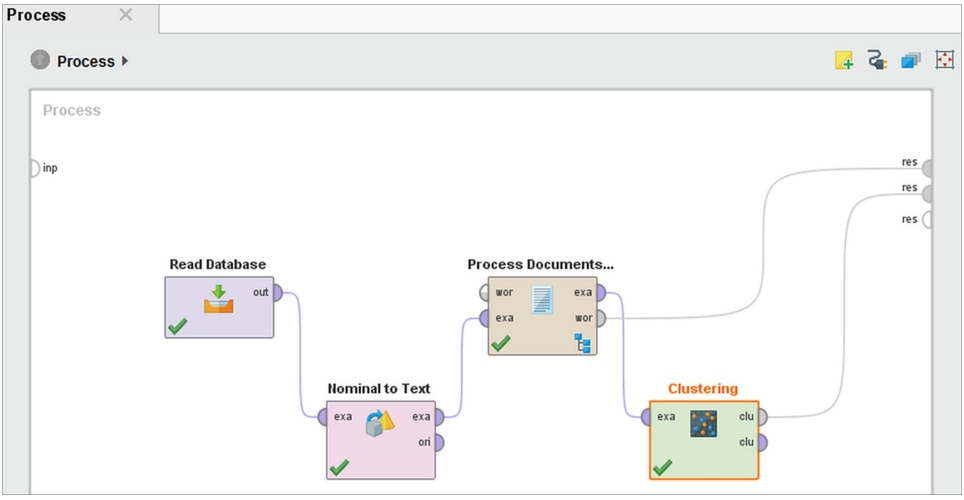

Fig. 5. Text mining process steps in RapidMiner.

filtering stop words removes words that have no semantic content relative to a specific domain. Stemming methods transform the words into their standardized forms, e.g., by removing the suffix "ing" from verbs, or other affixes. A stem is a natural group of words with equivalent (or very similar) meanings. Several algorithms, e.g., the snowball algorithm we applied,[47] are available for stemming, with most designed for texts in English.

3.1.4. *Initial taxonomy*

The initial taxonomy is a combination of a manually created taxonomy and the result of the text mining process. Figure 6 presents the manually created taxonomy for BA, prepared by us and based on a review of literature and discussions with experts. Experts from academia, software companies working with BA, and consultancy companies were consulted. Our research emphasizes taxonomy development and maintenance; hence, we did not attempt to prepare a comprehensive taxonomy. The primary fields of BA are descriptive analytics, predictive analytics, and prescriptive analytics, as mentioned in Sec. 2; hence, these are the subcategories of our BA taxonomy. Holsapple *et al.*[2] emphasizes the descriptive/explanatory, predictive, and prescriptive BA categories, while Refs. 17 and 18 apply the same grouping of BA solutions.[17,18] Descriptive analytics refers to knowing what happened and what is happening in an organization. In this situation, the business problem is well-defined. It involves business reporting (OLAP reports are typical), dashboards, scorecards, business performance management, data warehousing technology, data marts, ETL (extract, transform, load), data quality solutions, and visual analytics. Recently, visual analytics have gained prominence due to rapid development and the proliferation of self-service business intelligence solutions, e.g., Tableau and PowerBI.

Fig. 6. A manually prepared taxonomy.

Predictive analytics aims to determine what will happen in the future, and why. It refers to a group of methods that use historical data to predict a specific target variable in the future. This category of analysis involves applying statistical techniques, machine learning, data, web, and text mining. Well-known predictive methods include regression and neural networks.[18,19,48] Opinion mining and sentiment analysis have become popular recently with the proliferation of Web 2.0 initiatives. In Web 2.0, the focus is on user-generated content, with social media and social networking (e.g., Facebook) being the key areas. The resulting large amounts of content are very valuable to decision makers, as they can identify the customer's needs and opinions without the use of comparatively bothersome traditional techniques, e.g., surveys and questionnaires. Sentiment analysis and opinion mining are sometimes used interchangeably.[18,49–51] Liu and Zhang[49] define them as "the computational study of people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics, and their attributes."

Prescriptive analytics is defined by questions "what business should I pursue, and why." It is an emerging field that has drawn more attention with the advent of big data science. This analysis attempts to examine various courses of action in order to determine the best possible business decision. Real-world decision-making problems are usually complex, semi-structured or unstructured, where decision-makers have simultaneously consider of all pertinent factors that are related to the problem. Multiple-criteria decision making (MCDM) is the field that is devoted to the development and implementation of decision support tools and methodologies to confront complex decision problems involving multiple criteria, goals, or objectives of conflicting nature.[92,93] The goal here is to provide a decision or a recommendation for a specific action. Typical technologies of this category are expert systems, simulations and decision support systems.

The text mining and clustering process described in step 3 (Sec. 3.1.3) enriches the manually prepared taxonomy (Fig. 6) with several new areas according to clusters, describing the latest research fields and the most popular research domains. We put such fields in the domain category (Fig. 7), with six clusters providing the subcategories. Domain labels and their occurrences in the Decision Support Systems and Electronic Commerce journal (in January 2018) are summarized in Table 2. We name the clusters according to a set of descriptive representative words. Such representative words were prepared during the "bag-of-words representation" approach applied in text mining. The set of words was examined by experts to describe and name each cluster. The first cluster was labelled "social media analytics." Decisive words describing this category were as follows: sentiment, opinion, review, lexicon, emotion, text, social media, twitter, blog, feedback, tweet, positive, attitude, and vote, as shown in Fig. 7. This topic was decisive during the previous 5 years of BA history; hence, the result is in agreement with the literature and the latest software development initiatives in BA. Although sentiment analysis and opinion mining became a popular research topic beginning in approximately 2008, there are early related works.[52–55] From the software development side, the majority of the data, web and text mining packages added opinion mining and sentiment analysis components during the preceding years. The second cluster is labelled "health analytics." This result is somewhat surprising; however, it can be explained by a large volume of data created by this sector and the increase in applications of IoT devices in healthcare. The representative words are as follows: healthcare service, healthcare industry, healthcare information, hospital, clinic, examination, protocol, person and human. The biomedical community was very active in BA fields, especially in data and text mining. This field is particularly interesting not only for academia but also for the general public. Recent biomedical advances help understand disease mechanisms, and support the development of mechanisms for, and approaches to the prevention and cure of diseases.[56] The third cluster is labelled "security analytics." This is unsurprising, as security is a high-profile application in BA[17] and is one of the top trends in business intelligence/BA.[57] The representative words are fraud, detect, financial fraud, and phishing, and from the BA side, data and text mining, ontology

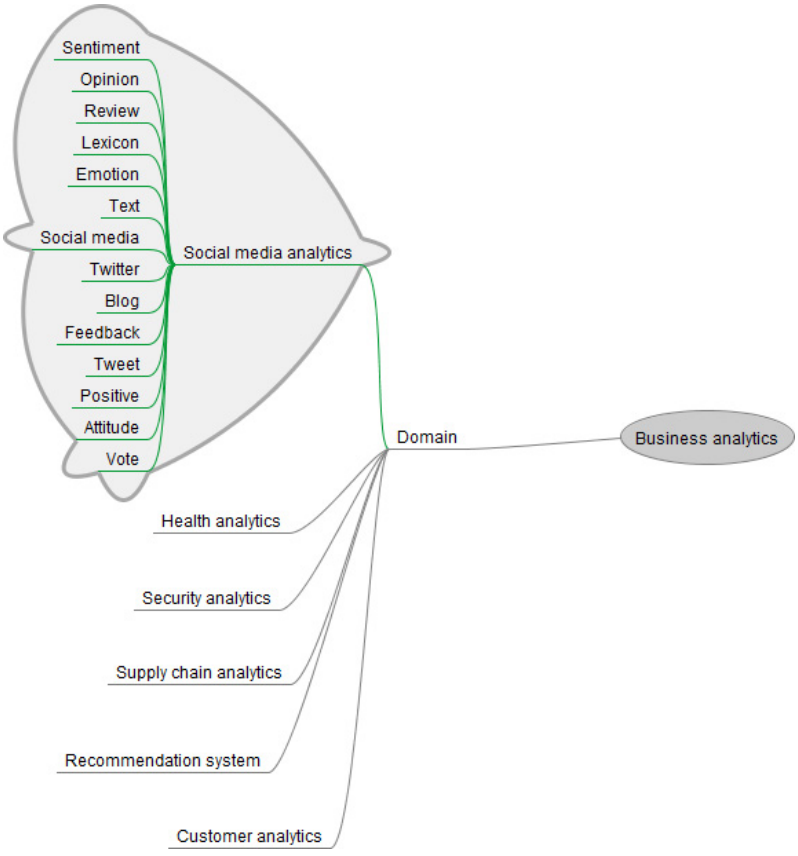

Fig. 7. Emerging research areas.

Table 2. Domain labels and their occurrences in the Decision Support Systems and Electronic Commerce journal (in January 2018).

Domain labels	Key terms in labels	The total number of papers mentioning the key term
Customer Analytics	Customer relationship management	1,195
Security Analytics	Security	1,110
Recommendation Systems	Recommendation	744
SCM Data Science	Supply chain management	731
Health Analytics	Health	666
Social Media Analytics	Social media	797
	Opinion mining	534

and semantic technology. Security issues are a major concern for most organizations due to a rapid growth in the number and types of risks and threats with increasing use of devices (mobile phones, smart devices, cars, etc.), more vulnerable than in the past.[58] Companies have to protect their intellectual assets and infrastructure,

requiring the prevention and detection of attacks. This "battle" occurs in cyberspace, where BA has a huge potential for the protection of information assets. Organizations gather large amounts of security-related data, e.g., log files, that have to be analyzed to detect security breaches. Due to the data-intensive nature of the field, BA technologies are widely applied to security in various areas. The key fields include rule mining and clustering, criminal network analysis, spatial-temporal analysis, analysis of log files, sentiment analysis, incident and cyberattacks analysis.[17] Using BA is mandatory for security organizations and agencies that collect large amounts of data on cybersecurity threats from several sources. Security analytics are challenging due to multiple sources and the unstructured nature of data they provide. Additionally, data are produced at a high rate (e.g., in the case of a log file). IT audit, a related field, is also becoming data-intensive; one example is continuous assurance. The latter is a combination of continuous monitoring by management, combined with continuous auditing of data streams and effectiveness of internal controls by an external auditor.[59] The fourth cluster is labelled "supply chain analytics" or supply chain management (SCM) data science. The representative words are supply, market, auction, negotiation, supply chain, retail, demand, product, firm, supplier, risk, risk management, vendor, decision, and simulation. SCM data science is a relatively new term, defined by Waller and Fawcett[60] as "the application of quantitative and qualitative methods from a variety of disciplines in combination with SCM theory to solve relevant SCM problems and predict outcomes, taking into account data quality and availability issues." Possible applications of BA in SCM and the related research questions have recently been discussed by several papers.[60–63] Trkman *et al.*[62] analyze the impact of BA use in various areas of the SC on the performance of the chain. They prove that BA use in critical process areas could affect the SC performance. Findings were confirmed in a large sample of companies in different industries and countries. The fifth cluster is labelled "recommendation systems." The representative words are recommend, recommendation system, filter, profile, collaboration, behavior, feature selection, and trust.

Recommendations, relied upon by decision-makers play crucial roles in situations requiring decisions. Recommendations help customize services and facilitate online shopping. In certain cases, e.g., that of Amazon.com,[18] recommendation systems are one of the major success factors.

The sixth cluster is labelled "BA in customer relationship management" or customer analytics in Holsapple's taxonomy.[2] The representative words are satisfaction, user satisfaction, trust, perceive, engagement, value, customization, market, cost, consumer, company, relationship, advertisement, and product. BA, e.g., data mining, has become an integral part of retailers' decision-making process and CRM activities. The goal of CRM is to build one-on-one relationships with customers. Companies have large customer datasets built through interactions with customers that can be combined with demographic and socioeconomic attributes, creating a valuable opportunity to improve customer relations and to be more competitive. Customer analytics can help companies improve understanding of customers'

behavior, tailor direct marketing offers to customer preferences and design promotions. Data mining applications are common in this field; the typical ones are customer profiling, churn analysis, association identification (cross-selling, up-selling) and identification of the most profitable customers.

3.1.5. *Taxonomy enrichment with ProMine*

The fourth step of our research resulted in the initial taxonomy, which we enrich in step five. To this end, we use ProMine,[32] our OL text mining tool. ProMine performs two basic tasks, one being information extraction, i.e., extracting new concepts from the domain document corpus using several other semantic sources (WorldNet and Wiktionary), and the other being a categorization of such extracted concepts into an already defined seed ontology with the help of a domain expert. In our experiments, we select a sample of keywords related to social media analytics. Included among these keyword lists was input into ProMine. To enrich the vocabulary of required knowledge elements, the selected keywords are linked to external lexical and semantic resources, e.g., WordNet[64] and Wiktionary. We extract similar words (synonyms) from the resulting lexicons. For instance, in the first step, we took the word "sentiment" and entered it into ProMine. Using WordNet and Wiktionary, the programme provides a list of synonyms and related words (opinion, sentiment, persuasion, view, thought, feeling, and sense) from the keyword "sentiment." To make the resulting new word list domain-specific, a domain corpus is required that can include domain glossaries, domain-related journal papers, or any type of domain-related documents. The concept enrichment and filtration modules of ProMine automatically filter out ambiguous words unrelated to the domain, and extract a set of keywords in the form of compound words by using the domain corpus. This reflects our belief that concepts, expressed in compound or multi-word terms can be more informative than single words. We obtained a large list of domain-specific concepts. At the end of the first task, concept ranking and selection, ProMine applies a statistical measure based on the information gain to identify the concepts most relevant to the starting keyword. Using the keyword "sentiment," we obtained 287 concepts. At the end of our experiments we obtained the respective lists of concepts (knowledge elements) for all keywords, with such words at that point ready to be used for ontology enrichment.

The second task of ProMine is semantic concept categorization for ontology enrichment. After extracting new concepts, ProMine puts them into an already developed seed ontology, in our case, into the initial taxonomy. To elucidate the conceptual relationships between such words and an existing initial ontology, ProMine uses a novel semantic concept categorization method to enrich an existing ontology. This method classifies new domain-specific concepts according to the existing taxonomy of the initial ontology. To categorize concepts, this method uses the knowledge of existing concept categories (taxonomy of classes) of the ontology together with the help of external knowledge resources, such as Wiktionary.[32] We detail the validation of our taxonomy in discussion below (Sec. 4).

4. Discussion

This paper provides a novel semi-automatic method of taxonomy development and maintenance in the field of BA using content analysis and text mining. First, we discuss the distinctive features of our approach in the context of the research gaps identified in Sec. 2; then we compare the results provided with our solutions with the results of the investigated solutions from the literature (Fig. 4). We identify four research gaps in the literature review section, related to the typical learning sources used for taxonomy development, degree of user intervention, the targeted domain and the main techniques applied.

We found that the typical learning sources used for taxonomy development are research papers from the literature.[12,24,25,78–80] A manual selection of papers from various journals showed that the number of selected papers vary depending on the source and these differences potentially affecting results. Our approach overcomes these potential biases through automatic extraction of papers. Additionally, we utilize the combination of various sources (literature, Wordnet, and domain corpus) in taxonomy development, which provides more enriched sources. User intervention, manual processes increase the subjectivity of the taxonomy development process. There are purely manual processes[20,24] and automatic methods.[8] In a manual process, a domain expert's intuition and knowledge of the domain will affect results. Our proposed taxonomy development method is a semi-automatic one; it combines different statistical, machine learning methods, and semantic technologies. We involve experts (manual process) where human judgment is needed, namely in the evaluation part, and apply automatic approaches when it is relevant and possible. We extract key concepts by using an unsupervised learning approach, performing clustering as in OntoGen,[76] where labels are assigned to clusters and regarded as concepts, while the terms in the cluster are considered its instances. In our work, domain concepts are extracted automatically without the involvement of a human expert. We enrich the existing ontology, while other approaches use the Protégé software program to actually create an ontology.[5] The domains related research gap reveals that the BA domain is not targeted, we found only "business intelligence" as a similar domain to ours.[20,24] The main techniques used for taxonomy development are diverse, the most popular ones are statistical and machine learning methods. Science mapping analysis is also widely used.[77–80] The main difference of our approach and the previously mentioned science mapping-based approach is that ProMine enriches the initial taxonomy, which we obtain in a semi-automatic way. The application of semantic technologies are also limited in the investigated solutions. Similarly, to Meijer's[8] approach, we also use semantic precision, semantic recall and the taxonomic F-measure in evaluation. However, our method is quite different from ATCT. The difference stems from our method of clustering before concept extraction for taxonomy enrichment. Table 3 summarizes the evaluation of taxonomy development methods.

According to Table 3 manual evaluation, based on experts' evaluation is typical. Our approach provides higher accuracy, than the other approaches where it is

Table 3. Evaluation of the taxonomy development methods.

Taxonomy development methods	Evaluation and results
Alter[25]	Manual
Basole *et al.*[12]	Accuracy is 65.3%
Cobo *et al.*[79]	Manual
Cobo *et al.*[80] and Cobo *et al.* (2012)	Manual
Delir Haghighi *et al.*[5]	N/A
Kang *et al.*[90]	Precision is 46.68%; recall is 83.24% of the document emotion prediction & precision is 84.00%; recall is 36.06% of the word emotion prediction
López-Herrera *et al.*[78]	Manual
Lv *et al.*[91]	Manual
Meijer *et al.*[8]	Precision is 84%; recall is 36.06% of the word emotion prediction
Nickerson's Taxonomy Development Method (2013)	Manual
Ontogen (2007)	N/A
OntoLearn (2005)	Recall ranges from 46% to 96%; precision ranges from 65% to 97%.
Text-To-Onto (2000)	Accuracy is 76% for taxonomic relationships.
Trieu[20]	Manual
White[24]	Manual
Our approach	Average precision is 92%; recall is 81%

counted. Our recall is also higher than the majority of other known recall values. We identify the latest research fields of BA domain (Fig. 7). Holsapple *et al.*[2] mentioned supply chain analytics, crisis analytics (corresponding to security analytics in our result) and customer analytics domains in their BA taxonomy. These dimensions overlap with our results; additionally, our BA taxonomy development method provides an approach to maintenance as well. The authors' first sample (descriptive analytics, predictive analytics, and prescriptive analytics) is the same as our subdomains, yet the authors do not detail such categories. Their third dimension, technique, refers to the way of performing an analytics task. The authors mention multiple perspectives in this context, while following the approach that distinguishes specific mechanisms used for analytics, which is similar to our taxonomy. To summarize, the authors' approach is conceptually higher-level, resulting in a meta-level synthesis of the BA field. A BI framework targeting business value that could be mapped to a taxonomy was provided.[20]

4.1. *Taxonomy validation*

There are various methods for taxonomy evaluation; additionally, we can also apply ontology evaluation techniques. Brank *et al.* distinguish four categories of ontology assessment that are also relevant to our case.[5,67] These are the "gold standard" evaluation,[8,68] the data-driven ontology evaluation,[69] the application-based evaluation,[5] and evaluation by manual inspection. Another categorization of assessment methods, proposed by Yu *et al.*,[70] suggests three primary methods: the "gold standard" evaluation, a task-based evaluation and criteria-based evaluation.

We use two methods to evaluate our taxonomy. One is the "gold standard" evaluation, a common approach to assessment of automatically built taxonomies.[8,68] In this approach, a constructed taxonomy is compared to the benchmark taxonomy. BA is a relatively new field; hence, we did not find a dedicated comprehensive "benchmark" taxonomy to apply in validation. To overcome this problem, we examine the possibility of applying more general taxonomies, such as the ACM Computing Classification System,[75] a keyword classification scheme for IS research proposed by Barki *et al.*[73] and a unified classification system for research in computing disciplines described by Vessey *et al.*[74] Both taxonomies by Barki and Vessey are higher-level than ours, while the ACM taxonomy is highly complex. Using a restricted and relevant part of the ACM taxonomy, we were able to compute precision and recall as suggested by Meijer *et al.*[8] and Staab abs Dellschaft[68]:

$$P = \frac{|\mathrm{Cc} \cap \mathrm{Cr}|}{|\mathrm{Cc}|}$$

$$R = \frac{|\mathrm{Cc} \cap \mathrm{Cr}|}{|\mathrm{Cr}|}$$

where Cc represents the concepts of the constructed taxonomy, while Cr represents those of the reference taxonomy. In our case, P is 0.53, while R is 0.6, which are higher than Meijer's results. To evaluate ProMine's categorization process, we calculated precision and recall for new concepts generated and categorized by the program (Table 4). The true-positive rate is also known as recall measures the proportion of actual positives, which are correctly identified

$$\mathrm{TP}_{\mathrm{rate}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}.$$

Recall is a measure to determine how many truly relevant results are returned. If we look at Table 4, the above rows show low recall and this gradually increases, and this high recall relates to a low false negative rate that proves that ProMine's categorization process is returning a majority of all the positive results (high recall).

Table 4. Precision and recall values of new concepts generated by ProMine.

No. of new concepts	TP	FP	FN	Precision	Recall
288	179	5	104	0.97	0.63
176	145	3	28	0.98	0.84
101	33	39	29	0.46	0.53
235	205	0	30	1	0.87
216	166	13	37	0.93	0.82
310	287	1	22	0.99	0.93
45	38	0	7	1	0.84
58	55	1	2	0.98	0.96
36	32	0	4	1	0.89

However, this will happen gradually by increasing the number of cycles because each time every category will be enriched with new concepts.

TP are those concepts that fit into our ontology, and our system has placed them in their corresponding categories. FP are those concepts that are wrongly categorized. FN are those concepts, which are incorrectly categorized by our system ProMine, while these concepts should be in a miscellaneous category (a new category that system generates itself in each cycle). TN describes a situation in concept categorization when ProMine correctly categories a negative test case into the miscellaneous category. TN does not exist in this case, because we had no miscellaneous category. The result is acceptable, with only a single concept (attitude) showing a low value. Inadequate precision and recall values were caused by "attitude" appearing related to positions/jobs as a part of competence criteria that are not relevant to our case.

Our approach is a semi-automatic one, it combines manual and automated steps, focusing on the BA domain and applying various techniques, including semantic, and statistical approaches. It is suitable for the identification of the most important domains of the latest BA research: customer analytics, security analytics, recommendation systems, SCM data science, health analytics, and social media analytics, which highlight several important research directions in this rapidly changing domain. The advantage of our method is that it is less time- and resource-intensive than the manual methods; and it provides better recall, compared with the investigated solutions. Initial taxonomy as a precondition is needed for the adoption.

5. Conclusion

The main objective of our paper is to present a semi-automatic method of taxonomy development and maintenance in the field of BA, using content analysis and text mining. The contribution of our research is threefold: (1) the taxonomy development method, (2) the draft taxonomy for BA, and (3) identification of the latest research areas and trends in BA. There is a growing interest in BA, and this field is becoming crucial to research, business and industry, as big data, data science and machine learning gain importance. However, as the literature review demonstrated, the BA field suffers from several incorrect, imprecise and incomplete definitions[2] that underscore the importance of taxonomies. We perform a review of literature on taxonomy development methods in Sec. 2, summarize the strengths and weaknesses of existing approaches in the discussion section (Table 1) and identify the research gaps. This review shows that most taxonomy development processes involve manual steps, in particular, for evaluation, with only a few methods targeting the BA domain.

Our approach has three limitations. One relates to the source and content of the corpus, whereas using a more diverse source could result in a different output. However, the corpus was applied in our approach to prepare the initial taxonomy that was enriched afterwards. As we mentioned in the discussion of corpus

preparation, we select the Decision Support Systems and Electronic Commerce journal as the source because it is a leading journal in the field. Another limitation arises from not preparing an exhaustive and complete taxonomy, since our focus was on demonstrating the taxonomy development process. The third limitation concerns the manual evaluation of clusters' labelling and initial taxonomy construction that involved experts. This limitation could be addressed by a comparison of the taxonomy with the relevant literature, as we did, and repeating the expert evaluation cycles.

Our approach, compared to the existing ones, combines manual and automated steps, focusing on the BA domain and applying various techniques, including semantic and statistical approaches. Additionally, we identify the most important domains of the latest BA research: customer analytics, security analytics, recommendation systems, SCM data science, health analytics, and social media analytics, which highlight several important research directions in this rapidly changing domain. This could constitute beneficial information for researchers who deal with BA and intelligent systems and it could also be used in education and training. BA is used in almost every company, but the majority of these BA environments have no semantic layer, which could be provided with our approach. This taxonomy could be integrated into the existing BA environment (e.g., BA reports) or into an intelligent system to provide contextual information for users. There should be two key consideration when we want to apply our method for any other domain: one is the initial taxonomy and the other is the domain expert (only for validation process).

The initial steps of our proposed taxonomy development method need human resource/experts in the domain, so we should also try to make it automatic. Research in this direction may contribute towards fully automatic taxonomy development. Future work should be concerned with the amplification of the taxonomy by adding additional levels of detail and applying nonfunctional computing models in order to leverage quantifiability of nonfunctional requirements. Another direction for future research is to develop taxonomies in other domains by using this framework. We identify six emerging research fields of BA: social media analytics, health analytics, security analytics, supply chain analytics, customer analytics, and recommendation systems. Future research should focus on the further exploitation of these fields.

References

1. Gartner (2018), 2018 CIO Agenda Report, https://www.gartner.com/imagesrv/cio-trends/pdf/cio_agenda_2018.pdf, 2018.
2. C. Holsapple, A. Lee-Post and R. Pakath, A unified foundation for business analytics, *Decision Support Systems* **64** (2014) 130–141, doi: https://doi.org/10.1016/j.dss.2014.05.013.
3. H. J. Watson, Business analytics insight: Hype or here to stay? *Business Intelligence Journal* **16**(1) (2011) 4–8.
4. R. C. Nickerson, U. Varshney and J. Muntermann, A method for taxonomy development and its application in information systems, *European Journal of Information Systems* **22**(3) (2013) 336–359, doi: https://doi.org/10.1057/ejis.2012.26.

5. P. Delir Haghighi, F. Burstein, A. Zaslavsky and P. Arbon, Development and evaluation of ontology for intelligent decision support in medical emergency management for mass gatherings, *Decision Support Systems* **54**(2) (2013) 1192–1204, doi: https://doi.org/10.1016/j.dss.2012.11.013.
6. K. D. Bailey, *Typologies and Taxonomies: An Introduction to Classification Techniques*, Vol. 102 (Sage, 1994).
7. R. Green, Typologies and taxonomies: An introduction to classification techniques. *Journal of the American Society for Information Science* **47**(4) (1996) 328–329, doi: https://doi.org/10.1002/(SICI)1097-4571(199604)47:4<328::AID-ASI10>3.0.CO;2-Y
8. K. Meijer, F. Frasincar and F. Hogenboom, A semantic approach for extracting domain taxonomies from text, *Decision Support Systems* **62** (2014) 78–93, https://doi.org/10.1016/j.dss.2014.03.006
9. A. Al-Arfaj and A. Al-Salman, Ontology construction from text: Challenges and trends, *International Journal of Artificial Intelligence and Expert Systems* **6**(2) (2015) 15–26.
10. T. R. Gruber, Toward principles for the design of ontologies used for knowledge sharing, *International Journal of Human — Computer Studies* **43**(5–6) (1995) 907–928, doi: https://doi.org/10.1006/ijhc.1995.1081.
11. R. Van Rees, Clarity in the usage of the terms ontology, taxonomy and classification, *CIB Report* **284**(432) (2003) 1–8.
12. R. C. Basole, C. D. Seuss and W. B. Rouse, IT innovation adoption by enterprises: Knowledge discovery through text analytics, *Decision Support Systems* **54**(2) (2013) 1044–1054, doi: https://doi.org/10.1016/j.dss.2012.10.029.
13. T. H. Davenport, *Big Data @ Work: Dispelling the Myths, Uncovering the Opportunities* (Harvard Business Review Press, 2014).
14. C. O'Neil and R. Schutt, Introduction: What is data science, in *Doing Data Science: Straight Talk from the Frontline*, 1st ed. (O'Reilly Media, Inc., USA, 2013).
15. S. Mohanty, M. Jagadeesh and H. Srivatsa, *Big Data Imperatives: Enterprise Big Data Warehouse, BI Implementations and Analytics* (2013), doi: https://doi.org/10.1007/978-1-4302-4873-6.
16. T. H. Davenport, L. Adams, Z. A. Ahmad, N. Karia, E. E. Anschutz, B. Becker and C. Young, Competing on analytics, *Harvard Business Review* **84**(1) (2006) 98–107, 134, doi: https://doi.org/10.1177/2158244011433338.
17. H. Chen, R. H. L. Chiang and V. C. Storey, Business Intelligence and Analytics: From big data to big impact, *Mis Quarterly* **36**(4) (2012) 1165–1188, doi: https://doi.org/10.1145/2463676.2463712.
18. E. Turban, R. Sharda and D. Delen, *Decision Support and Business Intelligence Systems*, 9th ed. (Pearson, New Jersey, 2011).
19. D. Larson and V. Chang, A review and future direction of agile, business intelligence, analytics and data science, *International Journal of Information Management* **36**(5) (2016) 700–710, doi: https://doi.org/10.1016/j.ijinfomgt.2016.04.013.
20. V.-H. Trieu, Getting value from business intelligence systems: A review and research agenda, *Decision Support Systems* **93** (2017) 111–124, doi: https://doi.org/10.1016/j.dss.2016.09.019.
21. C. Soh and M. L. Markus, How IT creates business value: a process theory synthesis, in *ICIS 1995 Proc.* (Association for Information Systems (AIS), 1995), Paper 4. Retrieved from http://aisel.aisnet.org/icis1995/4.
22. N. Melville, K. Kraemer and V. Gurbaxani, Review: information technology and organizational performance: An integrative model of IT business value, *MIS Quarterly* **28**(2) (2004) 283–322, doi: https://doi.org/10.2307/25148636.

23. G. Schryen, Revisiting IS business value research: What we already know, what we still need to know and how we can get there, *European Journal of Information Systems* **22**(2) (2013) 139–169, doi: https://doi.org/10.1057/ejis.2012.45.
24. C. White, A taxonomy for BI, *DM Review*, (2004) 70–71.
25. S. Alter, A Taxonomy of decision support systems, *Sloan Management Review* **19**(1) (1977) 39–56.
26. S. Adolph, M. Tisch and J. Metternich, Challenges and approaches to competency development for future production, *Educational Alternatives* **12** (2014) 1001–1010.
27. M. Liberatore and W. Luo, Informs and the analytics movement: The view of the membership, *Interfaces* **41**(6) (2011) 578–589, doi: https://doi.org/org/10.1287/inte.1110.0599.
28. D. Delen and H. Demirkan, Data, information and analytics as services, *Decision Support Systems* **55**(1) (2013) 359–363, doi: https://doi.org/10.1016/j.dss.2012.05.044.
29. L. J. Jensen, J. Saric and P. Bork, Literature mining for the biologist: From information retrieval to biological discovery, *Nature Reviews Genetics* **7**(2) (2006) 119–129, doi: https://doi.org/10.1038/nrg1768.
30. G. Nenadić, H. Mima, I. Spasić, S. Ananiadou and J. I. Tsujii, Terminology-driven literature mining and knowledge acquisition in biomedicine, *International Journal of Medical Informatics* **67**(1–3) (2002) 33–48, doi: https://doi.org/10.1016/S1386-5056(02)00055-2.
31. B. De Bruijn and J. Martin, Getting to the (c)ore of knowledge: Mining biomedical literature, *International Journal of Medical Informatics* **67**(1–3) (2002) 7–18, doi: https://doi.org/10.1016/S1386-5056(02)00050-3.
32. S. Gillani and A. Ko, Incremental ontology population and enrichment through semantic-based text mining, *International Journal on Semantic Web and Information Systems* **11**(3) (2015) 44–66, doi: https://doi.org/10.4018/IJSWIS.2015070103.
33. A. O'Mara-Eves, J. Thomas, J. McNaught, M. Miwa, S. Ananiadou, Using text mining for study identification in systematic reviews: A systematic review of current approaches, *Systematic Reviews* **4**(1) (2015) 5, doi: https://doi.org/10.1186/2046-4053-4-5.
34. J. Thomas, J. McNaught and S. Ananiadou, Applications of text mining within systematic reviews, *Research Synthesis Methods* **2**(1) (2011) 1–14, doi: https://doi.org/10.1002/jrsm.27.
35. J. Guan, A. S. Manikas and L. H. Boyd, The international journal of production research at 55: A content-driven review and analysis, *International Journal of Production Research* **57** (2017) 1–13, doi: 10.1080/00207543.2017.1296979.
36. L. Waltman, N. J. van Eck and E. C. M. Noyons, A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics* **4**(4) (2010) 629–635 doi: https://doi.org/10.1016/j.joi.2010.07.002.
37. N. J. van Eck and L. Waltman, Text mining and visualization using VOSviewer, *ISSI Newsletter* **7**(3) (2011) 50–54, doi: https://doi.org/10.1371/journal.pone.0054847.
38. Y. Li, S. Chung and J. Holt, Text document clustering based on frequent word sequences, *Data & Knowledge Engineering* (2005) 293–294, doi: https://doi.org/10.1016/j.datak.2007.08.001.
39. H. Ahonen-Myka, Finding all maximal frequent sequences in text, in *Proc. ICML Workshop on Machine Learning in Text Data Analysis* (Slovenian Language Technologies Society, 1999), pp. 11–17.
40. P. Monali and K. Sandip, A concise survey on text data mining, *International Journal of Advanced Research in Computer Science and Electronics Engineering* **3**(9) (2014) 8040–8043.

41. R. Feldman and I. Dagan, Knowledge discovery in textual databases (KDT), *International Conference on Knowledge Discovery and Data Mining (KDD)*, (1995) 112–117, doi: https://doi.org/10.1.1.47.7462.
42. Rapid-i, The RapidMiner GUI Manual. *October*, (2009) 1–14.
43. M. Hofmann and R. Klinkenberg, RapidMiner: Data Mining Use Cases and Business Analytics Applications, *Zhurnal Eksperimental'noi I Teoreticheskoi Fiziki*, (2013), doi: https://doi.org/78-1-4822-0550-3.
44. R. L. Sallam, J. Tapadinhas, J. Parenteau, D. Yuen and B. Hostmann, *Magic Quadrant for Business Intelligence and Analytics Platforms. Gartner RAS Core Research Notes* (Gartner, Stamford, CT, 2014).
45. M. Gualtieri, The forrester wave^{TM}: Predictive analytics and machine learning solutions, Q1 2017, *Forrester Research* (2017).
46. V. Korde and C. N. Mahender, Text classification and classifiers: A survey, *International Journal of Artificial Intelligence & Applications* **3**(2) (2012) 85–99, doi: https://doi.org/10.5121/ijaia.2012.3208.
47. M. F. Porter, Snowball: A language for stemming algorithms (2001), Available at: http://snowball.tartarus.org/texts/introduction.html.
48. G. Phillips-Wren, L. S. Iyer, U. Kulkarni and T. Ariyachandra, Business analytics in the context of big data: A roadmap for research, *Communications of the Association for Information Systems* **37** (2015) 448–472.
49. B. Liu and L. Zhang, A survey of opinion mining and sentiment analysis, in *Mining Text Data* (Springer US, 2012), pp. 415–463.
50. B. Liu, Sentiment analysis and opinion mining, *Synthesis Lectures on Human Language Technologies* **5**(1) (2012) 1–167, doi: https://doi.org/10.2200/S00416ED1V01Y201204HLT016.
51. J. Jin, Y. Liu, P. Ji and H. Liu, Understanding big consumer opinion data for market-driven product design, *International Journal of Production Research* **54**(10) (2016) 3019–3041, doi: https://doi.org/10.1080/00207543.2016.1154208.
52. B. Pang and L. Lee, Opinion mining and sentiment analysis, *Foundations and Trends® in Information Retrieval*, **1**(2) (2006) 91–231, doi: https://doi.org/10.1561/1500 000 001.
53. E. Breck, Y. Choi and C. Cardie, Identifying expressions of opinion in context, *IJCAI International Joint Conf. Artificial Intelligence* (2007), pp. 2683–2688, doi: https://doi.org/10.1016/j.jad.2005.02.015.
54. J. Yi, T. Nasukawa, R. Bunescu and W. Niblack, Sentiment analyser: Extraction sentiments about a given topic using natural language processing techniques, in *IEEE Intl. Conf. Data Mining (ICDM)* (IEEE, 2003), pp. 427–434.
55. H. Yu and V. Hatzivassiloglou, Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences, in *Proc. 2003 Conf. Empirical Methods in Natural Language Processing* (Association for Computational LinguisticsN, USA, 2003), pp. 129–136, doi: https://doi.org/10.3115/1119355.1119372.
56. C. C. Aggarwal and C. X. Zhai, *Mining Text Data*, Vol. 9781461432 (Springer, Boston, MA, 2013), doi: https://doi.org/10.1007/978-1-4614-3223-4.
57. M. Lebied, *Top 10 Analytics and Business Intelligence Trends for 2018*, https://www.datapine.com/blog/business-intelligence-trends (2018).
58. McAfee Lab, *McAfee Labs 2016 Threats Predictions McAfee Labs. McAfee Labs.* Retrieved from www.mcafee.com/us/mcafee-labs.aspx%0Ahttp://www.mcafee.com/us/resources/reports/rp-threats-predictions-2016.pdf (2016).
59. J. Kocken and J. Hulstijn, in *Providing Continuous Assurance*, VMBO Workshop Series (Luxembourg Institute of Science and Technology, Luxembourg, 2017), pp. 1–16.

60. M. A. Waller and S. E. Fawcett, Data science, predictive analytics and big data: A revolution that will transform supply chain design and management, *Journal of Business Logistics* **34**(2) (2013) 77–84, doi: https://doi.org/10.1111/jbl.12010.
61. G. C. Souza, Supply chain analytics, *Business Horizons* **57**(5) (2014) 595–605, doi: https://doi.org/10.1016/j.bushor.2014.06.004.
62. P. Trkman, K. McCormack, M. P. V. De Oliveira and M. B. Ladeira, The impact of business analytics on supply chain performance. *Decision Support Systems* **49**(3) (2010) 318–327, doi: https://doi.org/10.1016/j.dss.2010.03.007.
63. B. Chae and D. L. Olson, Business analytics for supply chain: A dynamic-capabilities framework. *International Journal of Information Technology & Decision Making* **12**(1) (2013) 9–26, doi: https://doi.org/10.1142/S0219622013500016.
64. G. A. Miller, WordNet: A lexical database for English, *Communications of the ACM* **38**(11), 39–41.
65. Gartner, *Magic Quadrant for Business Intelligence and Analytics Platforms*, https://www.gartner.com/home 2017 (2017).
66. D. R. Moscato and E. D. Moscato, A taxonomy of a decision support system for professional sports, *Issues in Information Systems* **5**(2) (2004) 633–639.
67. J. Brank, M. Grobelnik and D. Mladenić, A survey of ontology evaluation techniques, in *Proc. Conf. Data Mining and Data Warehouses* (Citeseer Ljubljana, Slovenia, 2005), pp. 166–170, doi: https://doi.org/10.1.1.101.4788.
68. K. Dellschaft and S. Staab, On how to perform a gold standard based evaluation of ontology learning, *Learning* **4273**(8) (2006) 228–241, doi: https://doi.org/10.1007/11926078_17.
69. C. Brewster, H. Alani, S. Dasmahapatra and Y. Wilks, Data driven ontology evaluation, in *Fourth Int. Conf. Language Resources and Evaluation (LREC'04)* (European Language Resources Association (ELRA), 2004), pp. 641–644, doi: https://doi.org/10.1.1.99.6070.
70. J. Yu, J. A. Thom and A. Tam, Ontology evaluation using wikipedia categories for browsing, in *Proc. Sixteenth ACM Conf. Information and Knowledge Management — CIKM '07* (ACM, 2007), doi: https://doi.org/10.1145/1321440.1321474, p. 223.
71. A. Maedche and S. Staab, The TEXT-TO-ONTO ontology learning environment, *Proc. Software Demonstration at ICCS-2000 — Eight Int. Conf. Conceptual Structures*, Retrieved from http://www.aifb.uni-karlsruhe.de/WBS (Academic Press, 2000), pp. 890–930.
72. P. Velardi, R. Navigli, A. Cucchiarelli and F. Neri, Evaluation of OntoLearn, a methodology for automatic learning of domain ontologies, *Ontology Learning from Text: Methods, Evaluation and Applications*, Vol. 123 (IOS Press, 2005), pp. 92–106.
73. H. Barki, S. Rivard and J. Talbot, A keyword classification scheme for IS research literature: An update, *Mis Quarterly* **17** (1993) 209–226, doi: https://doi.org/10.2307/249802.
74. I. Vessey, V. Ramesh and R. L. Glass, A unified classification system for research in the computing disciplines, *Information and Software Technology* **47**(4) (2005) 245–255, doi: https://doi.org/10.1016/j.infsof.2004.08.006.
75. B. Rous, Major update to ACM's computing classification system, *Communications of the ACM* **55**(11) (2012) 12–12, doi: https://doi.org/10.1145/2366316.2366320.
76. B. Fortuna, M. Grobelnik and D. Mladenić, OntoGen: semi-automatic ontology editor, *Human Interface and the Management of Information, Interacting in Information Environments* **4558** (2007) 309–318, doi: https://doi.org/10.1007/978-3-540-73354-6.
77. H. Small, Visualizing science by citation mapping, *Journal of the American Society for Information Science* **50** (1999) 799–813.

78. A. G. López-Herrera, E. Herrera-Viedma, M. J. Cobo, M. A. Martínez, G. Kou and Y. Shi, A conceptual snapshot of the first decade (2002–2011) of the international journal of information technology & decision making, *International Journal of Information Technology & Decision Making* **11**(2) (2012) 247–270.
79. M. J. Cobo, A. G. López-Herrera, E. Herrera-Viedma and F. Herrera, An approach for detecting, quantifying and visualizing the evolution of a research field: A practical application to the fuzzy sets theory field, *Journal of Informetrics* **5**(1) (2011) 146–166.
80. M. J. Cobo, F. Chiclana, A. Collop, J. de Ona and E. Herrera-Viedma, A bibliometric analysis of the intelligent transportation systems research based on science mapping, *IEEE Transactions on Intelligent Transportation Systems* **15**(2) (2014) 901–908.
81. Y. Peng, G. Kou, Y. Shi and Z. Chen, A descriptive framework for the field of data mining and knowledge discovery, *International Journal of Information Technology & Decision Making* **7**(4) (2008) 639–682.
82. Q. Zhang and R. S. Segall, Web mining: A survey of current research, techniques and software. *International Journal of Information Technology & Decision Making* **7**(4) (2008) 683–720.
83. IBM, *The Four V's of Big Data*. Retrieved from http://www.ibmbigdatahub.com/infographic/four-vsbig-data (2014).
84. P. Goes, Editor's comments: Big data and IS research. *MIS Quarterly* **38**(3) (2014) iii–viii.
85. D. J. Power, C. Heavin, J. McDermott and M. Daly, Defining business analytics: An empirical approach, *Journal of Business Analytics* **1**(1) (2018) 40–53.
86. R. Sharda, D. Delen, E. Turban, J. E. Aronson, T. Liang and D. King, *Business Intelligence: A Managerial Perspective on Analytics*, 3rd edn. (Prentice Hall, New York, 2014).
87. T. H. Davenport and J. G. Harris, *Competing on Analytics: The New Science of Winning* (Harvard Business Press, 2007).
88. A. D. Nerkar, Business Analytics (BA): Core of Business Intelligence (BI). *International Journal of Advanced Engineering, Management and Science* **2**(12) (2016) 2176–2178.
89. Gartner, *IT Glossary*, Retrieved from https://www.gartner.com/it-glossary/business-analytics (2019).
90. X. Kang, F. Ren and Y. Wu, Exploring latent semantic information for textual emotion recognition in blog articles, *IEEE/CAA Journal of Automatica Sinica* **5**(1) (2017) 204–216.
91. Y. Lv, Y. Chen, X. Zhang, Y. Duan and N. L. Li, Social media based transportation research: The state of the work and the networking, *IEEE/CAA Journal of Automatica Sinica* **4** (1) (2017) 19–26.
92. G. Kou, D. Ergu, C. Lin and Y. Chen, Pairwise comparison matrix in multiple criteria decision making, *Technological and Economic Development of Economy* **22**(5) (2016) 738–765, doi: https://doi.org/10.3846/20294913.2016.1210694.
93. G. Kou, Y. Peng and G. Wang, Evaluation of clustering algorithms for financial risk analysis using MCDM methods, *Information Sciences* **275** (2014) 1–12, doi: http://dx.doi.org/10.1016/j.ins.2014.02.137.