What Are They Talking about? Relation Extraction from News to Identify Research Directions in Emerging Technologies

Nicolas Prat
ESSEC Business School
prat@essec.edu

Abstract

Emerging technologies are characterized by their uncertainty, rapid evolution, and major impact. We focus on identifying research directions in these technologies. Identifying these research directions requires keeping in pace with the development of the technologies, and reaching out to individuals and society, beyond organizations. This is made possible by big data analytics. This paper uses design science research to propose and apply a methodology that performs text mining on news crawled from the Internet to identify research directions for an emerging technology. The methodology uses relation extraction on the news documents to extract relations between terms of the emerging technology and information systems constructs. These relations are then analyzed to suggest research avenues. We apply the methodology to blockchain, a major emerging technology, and derive insights from this application.

1. Introduction

Emerging technologies, such as blockchain, have a major impact on individuals, organizations and society. Apart from impact, noteworthy characteristics of these technologies are their radical novelty, their relatively fast growth, their uncertainty, unseen social and ethical concerns, and a lack of investigation and research [1, 2].

We focus on identifying research directions in emerging technologies. Typically, research directions are identified by spotting gaps in the literature, sometimes coupled with a bottom-up approach in an organizational context (e.g., action research). However, to identify research directions in emerging technologies, this approach is insufficient, due to the uncertainty and fast growth of these technologies, and their major impact not only on organizations, but also on individuals and society. Uncertainty and fast growth imply that academic publications are often lagging behind changes brought about or issues raised by the technologies [3]; it is often difficult for information systems (IS) researchers to "*keep up the pace in theory development in order to stay relevant to business organizations and practitioners*" [4, p.4]. Major impact implies that researchers investigating emerging technologies should identify research directions by reaching out to individuals and society, beyond organizations. In other words, they should listen to the crowd.

This paper applies big data analytics, more specifically text mining, to analyze news about an emerging technology (e.g., blockchain), and, from this analysis, identify research directions for the technology. The variety of available news sources, the velocity of big data, and the automation provided by text mining, help keep in pace with the concerns of individuals, organizations and society regarding emerging technologies. Text mining on news is a way of listening to the crowd to identify research directions. More specifically, we use relation extraction to identify research directions on an emerging technology from news. We establish an ontology of the terms characterizing the emerging technology (e.g., "blockchain", "token"), and a list of IS constructs (e.g., "use", "benefits"). With relation extraction, we study the influences of the terms on the constructs, as identified by relation extraction (e.g., the relation "*influences (token, use)*" may be extracted from a sentence in a news document). These relations suggest research directions: what constructs should research investigate more closely, in the context or in relationship with what components of the emerging technology (for example, what contributes to token use, or how do tokens influence blockchain use?). This work thus focuses on identifying research directions in behavioral positivist research. It draws on previous work by Li et al. [5], [6]. A major difference is that these authors apply relation extraction to published academic research.

This work is in line with recent calls to apply big data analytics to IS research, taking advantage of their complementarity [7-9]. More specifically, "*theory can help make sense of big data in that theory can inform the selection of constructs […]*" [7, p.vii]. In our case, relation extraction uses a predefined vocabulary (the

URI: https://hdl.handle.net/10125/71282
978-0-9981331-4-0

ontology of terms of the emerging technology and the list of IS constructs).

To answer our research question (*How can we use relation extraction to identify research directions on a specific emerging technology from news?*), we use design science research [10], contributing two main artifacts: (1) a methodology that identifies research directions in emerging technologies through relation extraction from news and (2) an application of the methodology to blockchain. The application demonstrates [11] and evaluates [12] the methodology.

Next, we review relation extraction, which is central to our work. We present our research approach, the methodology, its application to blockchain, discuss the contribution, and conclude the paper.

2. Literature review of relation extraction

Relation extraction is a form of information extraction, a branch of text mining that extracts structured information from unstructured documents [13]. Information extraction (IE) may extract entities, relations or events. Named entity recognition is a specific case of entity extraction. Entity extraction is a prerequisite for relation and event extraction.

To illustrate relation extraction, consider the text "*John Smith moved to California in 2018. He is the CFO of Apple*". In an IE system, the entity types Person and Location may be defined, as well as the relation types lives_in (Person, Location) and works_for (Person, Company). The system would recognize that John Smith is an entity, instance of the entity type Person, California is a location; it would recognize the relations lives_in (John Smith, California) and works_for (John Smith, Apple). The same individual (or relation) may be mentioned several times in a text. In the present example, "he" refers to John Smith (anaphora).

Early approaches to relation extraction have used rules or patterns. Modern approaches are based on machine learning, formulating relation extraction as a classification problem [14]: given two entity mentions in a sentence, are the entities related through a specific relation? Formulating the task as a classification problem makes supervised learning, e.g. support vector machines (SVM), applicable to relation extraction.

Relation extraction relies heavily on natural language processing (NLP), more particularly part-of-speech (POS) tagging, lemmatization, or syntactic parsing. POS tagging assigns tags to words (e.g., "proper noun in singular form"). Lemmatization replaces inflectional forms of words by their root form [13]. Syntactic parsing determines the syntactic structure of a sentence, e.g., by representing it as a tree.

Machine-learning based approaches to relation extraction may be feature-based, kernel, or a combination. Feature-based methods require feature engineering, while kernel methods (e.g., tree kernels) compute similarities for classification without needing feature engineering [15, 16]. Whatever the method, training classifiers is often a lengthy process of annotation. To alleviate this issue, several approaches have been proposed [14], such as active learning and bootstrapping. In active learning, the classifier reduces the number of examples to annotate by focusing on those for which the uncertainty is greatest. In bootstrapping, the classifier starts from a small set of examples and iteratively finds new relations [15].

To evaluate the performance of relation extraction, classical metrics apply. Precision penalizes false positives. Recall penalizes false negatives. Fβ score combines these two metrics (β is used to weigh precision and recall differently). In relation extraction, precision is typically emphasized [17]; in the IEPY tool [18], it is the metric that is optimized by default.

Using text mining to identify research directions in information systems is not new. For example, topic modeling may be used to cluster previous research [19]. However, to the best of our knowledge, only TheoryOn [5, 6] and the publications that preceded it [16, 17] apply relation extraction to automatically analyze and make sense of previous research in IS. A major advantage of relation extraction is the granular view that it provides. In TheoryOn, relation extraction is applied to automatically extract relationships between constructs from academic articles (focusing on articles that use positivist behavioral research). TheoryOn explores the section of academic articles where research hypotheses are formulated. This section is automatically identified, using a rule-based approach, and the constructs and their relationships are extracted from it. Based on the extracted construct relationships, nomological networks are constructed. A nomological network represents constructs as nodes, while edges represent relationships between constructs in a hypothesis [17].

Even if the present work draws on TheoryOn [5, 6], a major difference is that it applies relation extraction to identify research directions in a specific category of IS/IT (emerging technologies), and identifies these research directions from news (versus published academic research). This difference has fundamental implications for our methodology.

3. Research approach

Design science research (DSR) strives to produce useful and novel artifacts [10]. It is an iterative process, comprised of three closely related cycles [20]: the relevance cycle, the rigor cycle, and the design cycle. The relevance cycle ensures that DSR artifacts contribute to solving problems or take advantage of

opportunities in their environment. The rigor cycle ensures that DSR is grounded on knowledge from the knowledge base and contributes to the enrichment of this knowledge base. The design cycle is central to DSR, iteratively building and evaluating artifacts to address the problem identified in the environment.

In this paper, for convenience of presentation, we present our two main artifacts (the methodology and its application) linearly. However, our research process, based on DSR, comprised seven major build-evaluate iterations, which we will summarize in the discussion. As regards relevance, the problem the we address is the difficulty to identify research directions in emerging technologies, and the opportunity that we take advantage of is the advent of big data and analytics (more specifically text mining). Concerning rigor, our main source of knowledge is the literature of relation extraction. The artifacts that we contribute to the knowledge base are the methodology and its application to blockchain, with insights gained from this application as regards research directions for blockchain.

4. A methodology to identify research directions in emerging technologies through relation extraction from news

The methodology (Figure 1) follows the phases of a research process, i.e. research question, data collection, data analysis, and result interpretation [9]. It also draws on the big data analytics cycle [21] and considers the specificities of relation extraction [17].

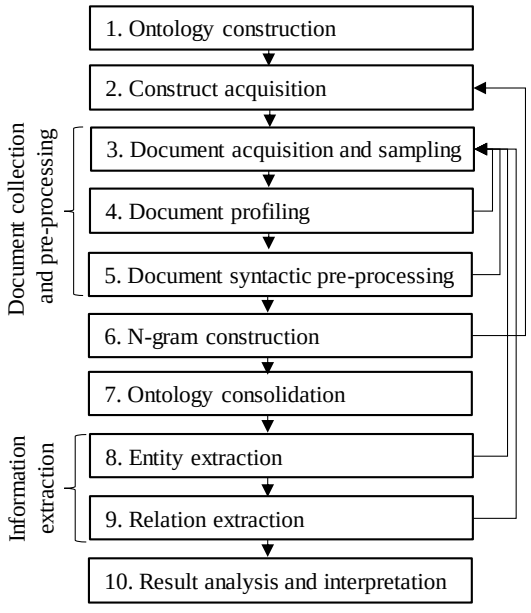

Figure 1. Overview of the methodology

The methodology does not have an explicit step of problem definition ("research question" in Müller et al. [9], or "business problem and opportunity identification" in the big data analytics cycle [21]). This is because the research question is the same for every instantiation of the methodology: *How can we use relation extraction to identify research directions on a specific emerging technology from news?* More specifically: *What constructs appear as worth investigating for positivist behavioral research, in relationship with what elements of the emerging technology*? The elements of the emerging technology are represented as an ontology of terms. The first step of the methodology, ontology construction, builds this ontology, based on academic articles that survey the emerging technology and present its key terms. Ontologies are relevant for information extraction [5, 6]. In this methodology, the ontology represents the terms of the emerging technology, their organization in a generalization / specialization hierarchy, their instances, and equivalences between terms or their instances. The second step, construct acquisition, establishes the list of IS constructs. This list is independent of the technology, although it may be completed with new constructs that appear relevant for the emerging technology but are missing from the list (as explained below).

The next steps of the methodology pertain to document collection and pre-processing, starting with document acquisition and sampling. Big data has brought about a new landscape in data acquisition, with the possibility of automating data extraction by crawling and scraping Web sites and by using application programming interfaces (APIs) [7]. In the methodology, the documents are news collected from the Web. The advantage with big data is that the collected news may span a wide variety of geographical, societal, and organizational borders [9]. This is especially important because, as mentioned above, researchers in emerging technologies should identify research directions by reaching out to individuals and society, beyond organizations. Document sampling should pay special attention to sampling bias, a major risk for Internet-mediated research [13]. In the methodology, we divide the sample into three subsamples: one for training the relation extraction algorithm, one for testing it, and one for executing it after training and testing. Document profiling follows document acquisition and sampling. In the big data analytics cycle, data preparation, comprising data profiling and data transformation [21], is crucial. With textual data, there is no real step of data transformation (documents pre-processing, mentioned below, replaces this step). However, data profiling remains crucial, because the trustworthiness of big data and their sources is often questionable [8, 9]. Document profiling may consider various characteristics, such as the sources of the documents, their freshness, their uniqueness (document deduplication), their style…

Based on the results of profiling, iteration on document acquisition and / or sampling may be required. Syntactic pre-processing follows document profiling. This step should be documented in detail [9]. Syntactic pre-processing often uses the Stanford parser[1]. Beyond its importance for n-gram construction and information extraction, syntactic parsing contributes to ensuring the quality of the collected documents. Depending on the results of parsing, iteration on document acquisition and / or sampling may be required.

The next step is n-gram construction. This step enriches and refines the ontology from step 1: while this ontology collects the terms characterizing the emerging technology from academic articles, n-grams are constructed from a sample of news documents. These n-grams may suggest new terms, refinement of terms… Thus, the news complete the academic articles, as also suggested in the methodology of taxonomy development for complex emerging technologies [3]. The sample of documents may be the one defined previously, or another sample, e.g. a larger sample. For n-gram generation, the value of *n* needs to be determined. In this paper, we consider that most of the terms consist of three words (trigrams) or less. N-grams are generated after syntactic pre-processing, and may thus use syntactic information (e.g., POS tags and / or parse trees). Apart from their use in refining the ontology from step 1, the n-grams may also serve to complete the list of constructs from step 2: although this list is independent of the emerging technology, it may be completed by frequent n-grams that appear as potential constructs and are missing from the list. Ontology consolidation (step 7) follows n-gram construction. It refines and completes the ontology from step 1, by considering the frequent n-grams that appear as important but were not accounted for in the ontology.

Information extraction comes next. We use relation extraction from news to determine *what constructs appear as worth investigating for positivist behavioral research, in relationship with what elements of the emerging technology*. We define one relation type: influences (Term of the emerging technology, Construct). Relation extraction identifies relations of this type from the news documents. This requires the preliminary step of entity extraction, i.e. the extraction of the terms of the technology and the constructs. The list of terms and constructs is known (as a result of steps 7 and 2), so we just need to look for their occurrences in the news, considering the possible inflections of words and capitalization (e.g., "token", "tokens", "Token"). Note that capitalization may in exceptional cases be used to distinguish terms (e.g., *B*itcoin refers to the eponymous blockchain, while *b*itcoin is the cryptocurrency). When the list of entities is known, gazettes are a simple solution for entity extraction, providing a list of entities to look for in the text. Relation extraction may be based on rules or machine learning. Several methods may be combined. Finally, the performance of relation extraction should be tested (precision, recall, Fβ). As mentioned above, data preparation is crucial in data analytics. Similarly, in the methodology, data profiling, sampling and pre-processing require a major effort. On the other hand, an existing algorithm for relation extraction may be used or adapted, instead of developing one from scratch.

Result analysis and interpretation is the final step of the methodology. In this step, we execute the relation extraction algorithm trained and tested in the previous step, and analyze and interpret the resulting relations. The analysis may be performed by term of the emerging technology, by construct, and by relation: what are the most frequent mentions of terms, the most frequent mentions of constructs, and the most frequent mentions of relations? The analysis of the most frequent relations helps answer the question: *What constructs appear as worth investigating for positivist behavioral research, in relationship with what elements of the emerging technology*? The analysis of the most frequent terms of the technology and the most frequent mentions of constructs may also suggest research directions. Note that the number of mentions of a term or construct is the number of times it is mentioned in a relation extracted by the algorithm. Counting the number of mentions of a term or construct in this manner provides focus. When counting the mentions of terms or relations, we should consider equivalences between terms (synonyms). (The methodology currently does not consider synonymous relations between constructs). The generalization / specialization relationships in the ontology may also be used. For example, if $term_i$ is a term in the ontology of the emerging technology, we may consider that a mention of the relation influences ($term_j$, $concept_k$), where $term_j$ is a specialization of $term_i$, also counts as a mention of the relationship influences ($term_i$, $concept_k$). In result interpretation, data-driven discoveries should be compared with the literature [9]. This means that the insights gained by analyzing the relations extracted from the news should be contrasted with the academic articles on the emerging technology: how do the research directions suggested by relation extraction from the news differ from those suggested by academic articles?

5. Application to blockchain

We apply to blockchain the methodology presented in the previous section.

[1] https://nlp.stanford.edu/software/lex-parser.shtml

5.1. Ontology construction

To build the ontology of blockchain terms, we use the following academic articles: [22-30].

A blockchain is a decentralized, immutable ledger for storing and exchanging financial assets, and more generally value. Traditional intermediaries are replaced by the nodes of the blockchain, which verify and certify transactions, using a consensus algorithm. Common consensus algorithms are proof of work, proof of stake, and practical byzantine fault tolerance.

The Bitcoin blockchain, with its cryptocurrency (bitcoin), was the first blockchain system. It uses proof of work. In consensus algorithms, cryptography plays a crucial role. Private keys are distinguished from public keys. Nodes that compete in proof of work are called miners. Ethereum, Hyperledger and Ripple are other examples of blockchain systems.

A blockchain is a chain of blocks of transactions. The first block is the genesis block. Forks may appear in a blockchain, leading to different versions.

Smart contracts execute automatically if certain conditions are met. Ultimately, they may lead to decentralized autonomous organizations.

Tokens, like cryptocurrencies, are cryptoassets. They have many possible applications, e.g., voting [26].

A blockchain may be permissionless or permissioned (i.e., regulated). It may be public (open to all), or private (restricted access).

We represent the ontology of blockchain terms with Protégé[2]. Terms may be classes or instances. For example, "proof of work" is an instance of the class "consensus algorithm". Classes are organized hierarchically. Equivalence between classes (synonymy) may also be represented.

5.2. Construct acquisition

To establish the list of IS constructs, we use the Inter-Nomological Network (INN)[3] [31], which integrates variables and items explored in the behavioral sciences. A variable is "*a measured entity of interest*". An item is "*a question or statement that is used to measure a variable*". Our methodology focuses on variables that may appear as dependent variables in IS positivist behavioral research. We call these variables IS constructs. With Excel, we build the list of IS constructs, proceeding as follows: from INN, we cut and paste all the (variable, journal (year)) tuples (82184 tuples); we deduplicate the tuples by removing the year from the journal; we keep only the variables explored in IS journals, obtaining 7861 variables (examples of such variables are "age" and "use of attribute-based decision support system"); among these variables, we keep only the ones with three words of less (this is because the name of variables is sometimes very detailed); from this list, we keep only the variables that may be dependent variables (e.g., we remove "age"), and remove variables that are too specific. The final list contains 105 constructs, such as adoption, anonymity, or use.

5.3. Document collection and pre-processing

For relation extraction, we use IEPY [18], a Python-based tool for information extraction focusing on relation extraction. This choice influences document collection and preprocessing. We need to check that IEPY will be able to process the documents.

5.3.1. Document acquisition and sampling. We acquire news documents from Webhose[4], a provider of unstructured data crawled from the Web. These data may be crawled from news, forums, or blogs. We are interested in news on blockchain ("blockchain" in the title), in English. We run a query specifying these constraints and the period (January 2017 to June 2019). As a result, Webhose provides 174391 documents.

We then iteratively build a sample composed of three subsamples. The objective is to get a final sample of about 5000 documents, with 20% of the documents used for training and testing the model (and inside these 20%, 80% for training and 20% for testing). An initial exploration of the data reveals that each document contains about five candidate instances of the relation type influences (Blockchain term, Construct).

For sampling, we use various criteria, including the length of the documents (minimum length to keep only informative documents), their source (avoid job ads for example), their style (familiarity), their "uniqueness" (the same news may appear in several documents), and the ability of syntactic parsing and IEPY to handle these documents. We try to keep the sources of documents as varied as possible. For documents that satisfy our constraints, we perform a random selection. Ultimately, we obtain 4987 documents from 1550 sources (examples of sources include www.bitcoinisle.com, www.forbes.com, and business.wapakdailynews.com). We keep 771 documents for training the relation extraction algorithm, 195 documents for testing, and the remaining 4021 documents for executing the trained and tested algorithm.

5.3.2. Document profiling. Document profiling is performed manually to avoid redundancies, check style (avoid familiarity) and relevance (e.g., documents investigating the rates of cryptocurrencies are not very

[2] http://protege.stanford.edu

[3] https://inn.theorizeit.org/

[4] www.webhose.io

relevant for this study). Even though we filter the documents as much as possible before manual profiling, we inspect 11000 documents, anticipating that we will keep about half of them. Ideally, artificial intelligence could assist in profiling, but to the best of our knowledge, application of artificial intelligence to data profiling is mostly restricted to structured data.

4.3.3. Document syntactic pre-processing. Syntactic pre-processing consists in applying the Stanford parser, used in IEPY.

5.4. N-gram construction

To construct the list of n-grams, we use a sample of 5920 news documents (period: May 2019) among the documents acquired from Webhose. Our algorithm for n-gram generation takes some syntactic information (from the previous step) into account. In particular, for unigrams, we consider nouns only.

5.5. Ontology consolidation

We use the list of n-grams generated previously to refine the ontology of blockchain terms. To this end, we examine the top unigrams, bigrams and trigrams. Among these n-grams, we select those that are relevant in the domain of blockchain, and complete and modify the blockchain ontology with the terms that were omitted in step 1. We thus complete the ontology with the following terms: mining, wallet, blockchain technology, public chain, digital identity, security (and utility) token, decentralized application, and distributed ledger technology. Figure 2 shows the ontology.

5.6. Information extraction

As mentioned above, we use IEPY [18] for information extraction.

5.6.1. Entity extraction. Entity extraction extracts the mentions of the terms of blockchain and of the IS constructs. To this end, we construct a gazette with the list of terms that result from ontology consolidation and construct acquisition. In the gazette, we consider the possible inflections of words and capitalization. Even though IEPY detects anaphora, this functionality would require assessing the accuracy of anaphora detection. We leave this for future work.

Figure 2. Consolidated ontology of blockchain

5.6.2. Relation extraction. For relation extraction, IEPY uses C-Support Vector Classification[5], a form of feature-based relation extraction based on SVM. Examples of features used in the classification are entity distance, verb count in between, bag of words in between, and bag of POS in between. "In between" means what is between a mention of a blockchain term and an IS construct in a sentence. Relation extraction boils down to a binary classification problem: for any

[5] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

blockchain term and IS construct that co-occur in a sentence, the blockchain term either influences the IS construct (the relation is present) or does not influence it (the relation is not present).

As mentioned above, we use 771 documents to train the classifier. This corresponds to a total of 4893 candidate relations (instances of the "influences (Blockchain term, Construct)" relation that possibly exist between the mention of a blockchain term and an IS construct in a sentence). We annotate each candidate relation manually, specifying if the relation exists or not.

Figure 3 shows a screenshot of IEPY with a document (from http://www.livetradingnews.com/uae-moving-government-to-blockchain-82135.html). For example, in the first sentence, the relation "influences" *is* labeled as *present* between "blockchain technology" and "adoption". In the second sentence, that mentions the use of courses, events, workshops and reports, the relation "influences" *is not present* between "blockchain" and "use" (what is used is the courses, events, workshops, and reports). In the third sentence, the relation "influences" *is present* between "Blockchain" and "control"…

IEPY uses active learning, but we annotate all 4893 candidate relations to be able to test the classifier with different numbers of annotated documents.

We use the subsample of 195 documents (with 1334 candidate relations) to test the classifier. With 2633 labeled candidate relations to train the classifier (out of a total of 4893 candidate relations), we get a precision of 0.96, a recall of 0.04, and a Fβ score of 0.48. (We choose a value of 0.2 for β, reflecting the fact that precision is much more important than recall). With 3701 labeled candidate relations, Fβ decreases (0.32). Testing the algorithm iteratively with several different numbers of labeled candidate relations (including the total number of candidate relations) reveals that the best results are obtained when the algorithm is trained with 2633 candidate relations. Performance does not necessarily increase with the number of labeled candidate relations, which may be explained by the fact that IEPY uses active learning and is very cautious in not generating false negatives (new information provided by annotation often results in lowering recall, without significantly improving precision). The very high precision score means that we are unlikely to suggest erroneous relations between blockchain terms and IS constructs. Low recall implies that we should be cautious in interpreting the results due to the limited sample size of extracted relations.

5.7. Result analysis and interpretation

5.7.1. Analysis of extracted relations. Having trained and tested the relation extraction algorithm, we can apply it to a large sample, our sample of 4021 news documents. This results in 362 relations. The relatively low number of relations is not a surprise, knowing that IEPY gives absolute priority to precision over recall.

The analysis by blockchain term reveals that the terms most mentioned in relations are blockchain (226 mentions), blockchain technology (96 mentions), transaction (11 mentions), smart contract, public blockchain, token, blockchain system, digital currency, and distributed ledger (3 mentions each). The exponential decline of the number of mentions, and the fact that blockchain and blockchain technology are the most mentioned terms, are not surprising. In our ontology, "blockchain" refers to a specific blockchain, as opposed to "blockchain technology". This is not always the case in news documents, where the term "blockchain" sometimes refers to the technology in general. Although any interpretation should take into account the sample size (362 relations), we notice that technical terms (e.g., relating to consensus algorithms and cryptography) are not among the top terms. We also notice that Bitcoin (or bitcoin) is absent from the top terms, suggesting that the crowd is considering diverse blockchain applications, while work in academia so far has often focused on Bitcoin and financial applications.

Figure 3. Labeling candidate relations with IEPY

The constructs that appear most frequently in extracted relations are (in this order): use, transparency, adoption, benefit, cost, trust, security, and efficiency. The fact that "use" is the top construct is not surprising, considering the importance of this construct in IS. The frequent mentions of "adoption" show that in practice, the adoption of blockchain raises many issues (the results would probably have been different with other emerging technologies like virtual reality or 3D printing for example). The ranking of constructs confirms that transparency, trust and security are important characteristics of blockchain. However, in the extracted relations, transparency is more frequently mentioned than trust. Traditionally, the IS discipline has devoted a lot of attention to trust. Since blockchain is sometimes referred to as "the trust machine", it is no wonder that trust in the context of blockchain is a major area of interest in academia. However, this study shows that in the context of blockchain, transparency may be a more important construct to explore than trust (naturally, transparency is likely to contribute to trust). Finally, the fact that governance [32] and traceability do not appear among the top constructs may come as a surprise. The reason is that they are absent from our list of constructs derived from INN. As mentioned above, a possibility would be to extend the list of constructs derived from INN, using the n-grams obtained for the emerging technology. Indeed, traceability and governance are among the top unigrams found at step 6.

Table 1 shows the number of extracted relations for the most mentioned blockchain terms and constructs (e.g., there are seven mentions of influences (transaction, cost)). This table synthetizes information by considering not only equivalences between terms (synonyms), but also generalization / specialization relationships in the blockchain ontology of Figure 2. This ontology considers blockchain as a specific case of distributed ledger, as is normally the case. Had we considered "blockchain" and "distributed ledger" as synonyms, the results of Table 1 would have been the same. The fact that "distributed ledger" and blockchain technology" appear as the top terms is not surprising. For these two terms, there is some consistency in the top relations (use, transparency, and then adoption are in the top three). This means that research should insist on the determinants of use, transparency and adoption (without overemphasizing trust, as mentioned previously). In particular, research should investigate what features of blockchain influence use and adoption [27]. For instance, it may explore the relation between the use of blockchain and the use of digital assets (Table 1 reveals three mentions of the relationship between use and digital asset).

Regarding transactions, the results of this study suggest that they are an important element of blockchain to zoom on, in particular as relates to cost, trust and transparency. For example, what are the determinants of the transparency of transactions in blockchain, or what is the influence of the transparency of transactions on other important constructs (like trust for example)?

As mentioned above, this analysis reveals little interest of the crowd in technical terms related to blockchain. For example, the crowd seems more interested in issues related to transactions, digital assets (including tokens) and smart contracts than in the effects of hard forks (a research direction suggested in [27]).

5.7.2. Complementary analysis with Word2vec. To benchmark the results obtained with relation extraction, we use another NLP technique to find frequent terms and term associations. The terms are the blockchain terms and the IS constructs. The technique that we choose is Word2vec[6] in its Python implementation[7]. Word2vec uses two-layer neural networks. This technique produces word embedding, i.e., represents words as vectors in a high-dimensional space. Like other techniques that find word associations, it is not as precise as relation extraction: just because a blockchain term appears frequently with an IS construct does not necessarily mean that the blockchain term influences the construct. However, this technique is powerful, does not require preliminary training of the algorithm, and may be used to confirm the results from relation extraction. For blockchain terms, Word2vec considers words separately (unigrams), but the first three unigrams are the same as the ones obtained in relation extraction, in the same order (blockchain, transaction, token). Regarding IS constructs, 6 of the 8 most frequent constructs from relation extraction also appear among the top 8 in Word2vec; the other two appear among the top 12 in Word2vec. Word2vec also confirms the importance of the constructs of use and adoption in relation with blockchain: predicting the words occurring most frequently in the context of the word "blockchain", we get the result: [('powered', 0.041), ('adoption', 0.020), ('enabled', 0.017), ('based', 0.015), ('technology', 0.010), ('using', 0.010), ('utilizing', 0.005), ('applications', 0.005), ('technologies', 0.005), ('application', 0.004)]. In the context of transactions, frequent words like "fees" or "speed" confirm the findings of relation extraction (transaction influences cost).

[6] https://code.google.com/archive/p/word2vec/

[7] https://radimrehurek.com/gensim/models/word2vec.html

Table 1. Relations (main blockchain terms and constructs)

	distributed ledger	blockchain technology	transaction	digital asset	blockchain system	smart contract
adoption	21	12		1		
benefit	21	7			1	
cost	13	1	7		1	
transparency	37	17	2		1	
trust	15	3	2	1		
use	45	28		3	1	1

6. Discussion and conclusion

Emerging technologies are characterized by their rapid evolution, their uncertainty, and their major impact. To identify research directions in emerging technologies, big data analytics appears relevant. In this work, we use relation extraction to identify research directions on a specific emerging technology from news. To this end, we use DSR to propose a methodology, which we apply to blockchain. From this application, we derive insights regarding possible research directions for blockchain.

In this paper, we have presented our two main artifacts – the methodology and its application, i.e. an instantiation [33] – linearly. However, the research process comprised seven major build-evaluate iterations, as summarized below (for "evaluate" iterations, we mention the evaluated criterion, according to the hierarchy of criteria of Prat et al. [12]). *Iteration 1* built the ontology of blockchain (a secondary artifact resulting from this research) and the list of IS constructs. *Iteration 2* evaluated the completeness of the ontology, leading to the consolidated ontology. *Iteration 3* trained, tested, and applied the IEPY classifier to extract relations from news on blockchain (this iteration was composed of multiple sub-iterations, including evaluation of accuracy with Fβ score). *Iteration 4* analyzed the extracted relations from iteration 3 to identify avenues for blockchain research, thus enabling us to evaluate the utility of our approach. In *iteration 5*, we abstracted from the blockchain example to build our methodology, i.e. the methodology to identify research directions in emerging technologies through relation extraction from news. In *iteration 6*, we got peer feedback on the methodology, including its completeness. This feedback lead to the introduction of Word2vec. *Iteration 7* completed the instantiation, applying Word2vec to the blockchain example. This enabled us to check the consistency between the results from relation extraction and the results from Word2vec.

This study focuses on identifying research directions for positivist behavioral research, but the insights from applying the methodology may be useful for other research traditions, e.g. DSR or qualitative research. Relation extraction from news complements other approaches, e.g. literature reviews on academic articles, to identify research directions.

Research increasingly relies on the crowd. Some researchers even consider the participation of the crowd to identify research questions [34]. We do not go this far, but consider that listening to the crowd can be a major source of inspiration to identify research directions in emerging technologies. News reflect the issues of interest to the crowd. To avoid biases related to news sources, a wide variety of sources should be considered, as in this research. Considering other types of sources, like blogs or forums, may be an avenue for further research. Syntactic parsing of these sources is more challenging. Another avenue for research is the more extensive use of word embedding, complementary to relation extraction, e.g. to identify synonymy relations between IS constructs, or IS constructs that often appear in the same context (suggesting possible relationships between these constructs, e.g. Word2vec revealed that the words "efficiency", "trust", "safety", "accountability", "traceability" and "security" often appeared in the context of the word "transparency"). Future work will also fine-tune the IEPY classification algorithm for relation extraction, and test and combine other algorithms to improve recall. We will also apply the approach to other emerging technologies.

7. References

[1] D. Rotolo, D. Hicks, and B. R. Martin, "What is an emerging technology?," *Research Policy*, vol. 44, no. 10, pp. 1827-1843, 2015.

[2] M. Halaweh, "Emerging technology: What is it?," *Journal of Technology Management & Innovation*, vol. 8, no. 3, pp. 108-115, 2013.

[3] O. Sangupamba Mwilu, N. Prat, and I. Comyn-Wattiau, "Taxonomy development for complex emerging technologies – The case of business intelligence and analytics on the cloud," *19th Pacific Asia Conference on Information Systems (PACIS 2015)*, Singapore, 2015.

[4] M. Chiarini Tremblay *et al.*, "Panel—Will artificial intelligence automate theory building? Are there lessons for academia from practice?," *39th International Conference on Information Systems (ICIS 2018)*, San Francisco, CA, 2018.

[5] J. Li, K. R. Larsen, and A. Abbasi, "TheoryOn: Designing a construct-based search engine to reduce information overload for behavioral science research," *11th International Conference on Design Science Research in Information Systems and Technology (DESRIST 2016)*, St. John's, NL, Canada, 2016.
[6] J. Li, K. Larsen, and A. Abbasi, "TheoryOn: A design framework and system for unlocking behavioral knowledge through ontology learning," *MIS Quarterly*, Forthcoming, 2020.
[7] A. Rai, "Editor's comments—Synergies between big data and theory," *MIS Quarterly*, vol. 40, no. 2, pp. iii-xi, 2016.
[8] R. Agarwal and V. Dhar, "Editorial—Big data, data science, and analytics: The opportunity and challenge for IS research," *Information Systems Research*, vol. 25, no. 3, pp. 443-448, 2014.
[9] O. Müller, I. Junglas, J. vom Brocke, and S. Debortoli, "Utilizing big data analytics for information systems research: challenges, promises and guidelines," *European Journal of Information Systems*, vol. 25, no. 4, pp. 289-302, 2016.
[10] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quarterly*, vol. 28, no. 1, pp. 75-105, 2004.
[11] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45-77, 2007.
[12] N. Prat, I. Comyn-Wattiau, and J. Akoka, "A taxonomy of evaluation methods for information systems artifacts," *Journal of Management Information Systems*, vol. 32, no. 3, pp. 229-267, 2015.
[13] G. Ignatow and R. Mihalcea, *Text Mining: A Guidebook for the Social Sciences*. Los Angeles, CA: SAGE Publications, 2016.
[14] J. Piskorski and R. Yangarber, "Information extraction: Past, present and future," in *Multi-source, Multilingual Information Extraction and Summarization*, T. Poibeau, H. Saggion, J. Piskorski, and R. Yangarber Eds. Berlin, Heidelberg: Springer-Verlag, 2013, pp. 23-49.
[15] J. Jiang, "Information extraction from text," in *Mining Text Data*, C. C. Aggarwal and C. Zhai Eds. Boston, MA: Springer US, 2012, pp. 11-41.
[16] J. Li and K. R. Larsen, "Tracking behavioral construct use through citations: A relation extraction approach," *34th International Conference on Information Systems (ICIS 2013)*, Milano, Italy, 2013.
[17] J. Li and K. R. Larsen, "Establishing nomological networks for behavioral science: A natural language processing based approach," *32nd International Conference on Information Systems (ICIS 2011)*, Shanghai, China, 2011.
[18] R. Carrascosa, J. Mansilla, G. García Berrotarán, F. M. Luque, and D. Moisset, "IEPY documentation." https://readthedocs.org/projects/iepy/downloads/pdf/latest/, 2017.
[19] M. J. Mortenson and R. Vidgen, "A computational literature review of the technology acceptance model," *International Journal of Information Management*, vol. 36, no. 6, Part B, pp. 1248-1259, 2016.
[20] A. R. Hevner, "A three cycle view of design science research," *Scandinavian Journal of Information Systems*, vol. 19, no. 2, pp. 87-92, 2007.
[21] N. Prat, "Augmented analytics," *Business & Information Systems Engineering*, vol. 61, no. 3, pp. 375-380, 2019.
[22] J. de Kruijff and H. Weigand, "Understanding the blockchain using enterprise ontology," *29th International Conference on Advanced Information Systems Engineering (CAiSE 2017)*, Essen, Germany, 2017.
[23] F. Glaser, "Pervasive decentralisation of digital infrastructures: a framework for blockchain enabled system and use case analysis," *50th Hawaii International Conference on System Sciences (HICSS 2017)*, Waikoloa Village, HI, 2017.
[24] M. Iansiti and K. R. Lakhani, "The truth about blockchain," *Harvard Business Review*, vol. 95, no. 1, pp. 118-127, 2017.
[25] O. Labazova, T. Dehling, and A. Sunyaev, "From hype to reality: A taxonomy of blockchain applications," *52nd Hawaii International Conference on System Sciences (HICSS 2019)*, Grand Wailea, Maui, HI, 2019.
[26] L. Oliveira, L. Zavolokina, I. Bauer, and G. Schwabe, "To token or not to token: Tools for understanding blockchain tokens," *39th International Conference on Information Systems (ICIS 2018)*, San Francisco, CA, 2018.
[27] M. Risius and K. Spohrer, "A blockchain research framework," *Business Information Systems Engineering*, vol. 59, no. 6, pp. 385-409, 2017.
[28] M. Rossi, C. Mueller-Bloch, J. B. Thatcher, and R. Beck, "Blockchain research in information systems: Current trends and an inclusive future research agenda," *Journal of the Association for Information Systems*, vol. 20, no. 9, pp. 1388-1403, 2019.
[29] G. Salviotti, L. M. De Rossi, and N. Abbatemarco, "A structured framework to assess the business application landscape of blockchain technologies," *51st Hawaii International Conference on System Sciences (HICSS 2018)*, Waikoloa VIllage, HI, 2018.
[30] Z. Zheng, S. Xie, H. Dai, X. Chen, and H. Wang, "An overview of blockchain technology: Architecture, consensus, and future trends," *IEEE 6th International Congress on Big Data (Big Data 2017)*, Boston, MA, 25-30 June 2017 2017.
[31] K. R. Larsen and C. H. Bong, "A tool for addressing construct identity in literature reviews and meta-analyses," *MIS Quarterly*, vol. 40, no. 3, pp. 529-551, 2016.
[32] R. Beck, C. Müller-Bloch, and J. L. King, "Governance in the blockchain economy: a framework and research agenda," *Journal of the Association for Information Systems*, vol. 19, no. 10, pp. 1020-1034, 2018.
[33] S. T. March and G. F. Smith, "Design and natural science research on information technology," *Decision Support Systems*, vol. 15, no. 4, pp. 251-266, 1995.
[34] S. Beck, T. Brasseur, M. Poetz, and H. Sauermann, "What's the problem? How crowdsourcing contributes to identifying scientific research questions," *Academy of Management Annual Meeting Proceedings*, vol. 2019, no. 1, pp. 644-649, 2019.