risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Iberian Journal of Information Systems and Technologies

Recebido/Submission: 29/11/2019
Aceitação/Acceptance: 31/01/2020

Minería de datos para detectar y prevenir la corrupción en los contratos: revisión sistemática de mapeo

Yeferson Torres-Berru[1,2], Vivian Félix López Batista[1], Pablo Torres-Carrión[3]

ymtorresb@usal.es, vivian@usal.es, pvtorres@utpl.edu.ec

[1] University of Salamanca, Plaza de la Merced, s/n,37008 Salamanca, Spain.

[2] Instituto Superior Tecnológico Loja, Av. Granada y Turunuma, , 1101608, Loja, Ecuador.

[3] Universidad Técnica Particular de Loja, San Cayetano Alto S/N, 1101608, Loja, Ecuador.

Pages: 13–26

Resumen: La corrupción según afirma la ONU está presente en sus diferentes formas y tipologías afectando directamente en la celebración de contratos tanto públicos como privados; en este contexto se realiza un mapping sistemático de investigaciones científicas (2015-2019) sobre corrupción en contratos en sus diversos formatos, aplicando minería de datos y sus técnicas. Se plantean seis preguntas de investigación a responder desde el análisis de 147 artículos obtenidos de las bases de datos WoS y Scopus. Las investigaciones se centran principalmente en la detección fraude, fraude financiero y corrupción, siendo las formas de corrupción más estudiadas el fraude (72.72%), y el sobreprecio (8,84%); las investigaciones se han realizado en Estados Unidos (16,32%), China (10,88%), Reino Unido (8,94%) y en LatinoAmerica Brasil (3,4%), con minimas contribuciones de Colombia y Paraguay.

Palabras-clave: Data mining, corrupción, mapping review.

Data Mining to detect and prevent corruption in contracts: Systematic Mapping Review

Abstract: Corruption according to the UN is present in various forms and types, affecting the realization of public and private contracts; in this context, a systematic mapping of scientific publications (2015-2019) is development, focused on contract corruption in its several forms, applying data mining and related techniques. Six research questions are presented to answer the analysis of 147 articles obtained from WoS and Scopus databases. The detection of fraud, financial fraud and corruption predominate in the researchs, exceling as forms of corruption, fraud (72.72%) and the overprice (8.84%). Researchs have been conducted in the United States (16.32%), China (10.88%) and the United Kingdom (8.94%); in Latin America emerge Brazil (3.4%) with minimum contributions from Colombia and Paraguay.

Keywords: Data mining, corruption, mapping review.

1. Introduction

Corruption is present in every country in the world, according to the (ONU , 2018, for the Transparency International Organization (2017) corruption is a "bribe, as an offer or receipt of any gift, loan, fee, reward or other advantage to or from any person as an incentive to do something that is dishonest, illegal or an abuse of trust, in the exercise of business activity". The word «corruption», derived from the Latin verb "corrumpĕre", with several negative meanings from the moral point of view: alter and disrupt, spoil, deprave, damage, rot, bribe, pervert, dodge (Martinez Fernandez 2015). Corruption is a situation in which a conflict of interest is used to satisfy a self-interest that is specified in obtaining a profit in breach of an existing legal framework (Cerillo Martinez 2015). This social behavior is present in different forms and aspects affecting different social contexts in family, institutional, private, governmental areas with negative repercussions.

In the public sector, corruption infers an abuse of an employee's power to obtain "profits," benefiting private entities, by taking advantage of a specific situation to break the law and benefit the other participant in the act (if any) and himself; "gains" for the corrupt include not only money but also material and intangible goods, which involve status and power (OECD 2005). Different forms and levels of corruption have been identified, such as bribery, embezzlement, fraud, extortion, abuse of trust, collusion and favoritism (Chan and Owusu 2017; Moran 2001; Vargas-Hernández 2009). The main mechanisms of corruption according to (Cassagne and Rivero Ysern 2007; Castro Cuenca 2017) are non-existence of contract, improper direct contracting, improper contracting, fractionation, contract modifications. As is visible, the study of corruption encompasses a large field of social and human sciences, with theoretical and scientific contributions.

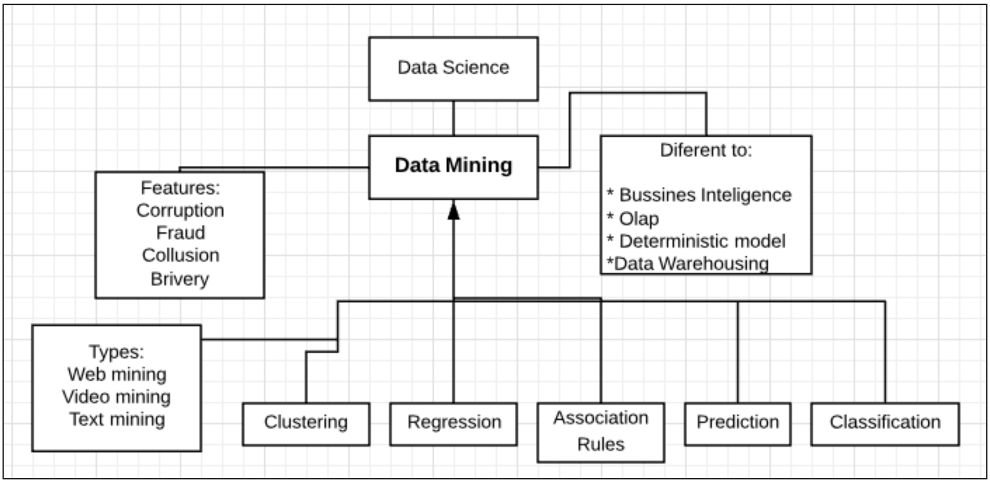

Figure 1 – Mentefacto Conceptual

As noted in (DiRienzo et al. 2007) technology has become a key ally in the fight against corruption, which shows that, in countries that use ICTS within the public administration, they get better levels in the corruption perception index. In this sense, the use of information technologies to combat corruption has shown good results in state agencies in several countries (Volosin 2015), excelling studies conducted by Goedhuys, Mohnen, and Taha (2016) demonstrating that reducing unnecessary interventions in procurement processes using technology improves transparency, because they usually cause abuse of power by public employees; (Shrivastava and Bhattacherjee 2015) consider using tools and resources to monitor the development of public employee functions at a low cost; (Lee and Lio 2016) explain how technology helps to promote transparency through the provision of information to the general public; and, (DiRienzo et al. 2007), (OECD 2007) explain how technology helps to eliminate bureaucratic procedures that generate higher costs and times. It is clear that the contribution of technology to improve anti-corruption management and control procedures, allows an improvement in the efficient administration, in search of transparency in public management processes.

Figure 1 shows the theoretical classification, detailing data science through data mining, its techniques (Han, et al 2011), and types (Hand 2013), and areas of knowledge different from data mining, but which are part of other branches of computer science and data science. In the "mentefacto conceptual", the relationships between characteristics of corruption and data mining are also graphically represented, among the primary forms of corruption are fraud, collusion, bribery, overprice, favoritism.

As related works, in the Scopus and Wos databases, two systematic reviews of scientific literature were found that relate the fight against corruption through the use of technology. (Indrajani et al. 2016) focus on finding fraud algorithms in online transactions, finding 25 scientific articles for study; and, (Ngai et al. 2011) reviews different data mining techniques for fraud detection in different fields of the financial field from 41 scientific articles. That is why this study is proposed as original, by encompassing corruption and data mining from a broader perspective (fraud, collusion, bribery, overprice, and favoritism) than the corruption form known as fraud.

Data science plays a vital role in detecting corruption and is commonly used to find hidden information in large amounts of data (Alvarez-Jareño et al. 2019). The corruption case known as Panama papers (Woodie 2016) revealed to the public opinion fiscal and financial fraud; in Brazil, the Observatory of Public Expenditures (Controladoria-Geral da União 2015) reviewed more than 120,000 public contracts and uncovered more than 7,500 cases involving $ 104 million in financial operations of doubtful legality. These examples illustrate the importance of data science in the fight against corruption.

Based on the forewent, this paper analyzes the scientific information published from 2015 to 2019 using the methodology of (Torres-Carrion et al. 2018) which is used for the systematic literature reviews, combining it with the methodology (García-González and Ramírez-Montoya 2019) whose components are detailed in the next section. The results are organized for presentation following the order of the research questions, with graphic explanations and referential analysis. Finally, the relevant conclusions regarding this systematic mapping are shared. The discussion section is omitted, taking into account that the objective of mapping is to describe the state of the art in terms of general areas of research on data mining and corruption.

2. Methodology

For the systematic search, the methodology of (Torres-Carrion et al. 2018) is followed, which divides the process into three phases: planning, conducting the review, and reporting the review. The third phase has been carried out following the methodology of (García-González and Ramíre\z-Montoya 2019) applied in its mapping report.

2.1. Research questions

Through the research questions, the investigation objective is established, as well as the variables to measure and answer the questions in Table 1 which pretend to inquire the number of published articles in the topic in mind, the geographical distribution, the context of the form of corruption studied and the primary line of research studied.

Question	Type of response sought
RQ1: How many studies are in the WOS and Scopus databases from 2015 to 2019?	• Number of articles in Scopus • Number of articles in WOS • Number of duplicated articles • Number of open access articles • Type of document (Review, etc)
RQ2: Who are the authors of the most cited articles?	• Most cited authors • Most cited articles
RQ3: What is the geographical distribution of the authors?	• Countries where the authors are from
RQ4: What are the journals with more publications on this line of research?	• Journals • Q1, Q2, Q3 or Q4 • Indice JCR
RQ5: In what contexts are these studies developed?	Fraud, Bribery, Collusion, Overpricing, Favoritism, Embezzlement
RQ6: What are the main topics addressed in this line of research?	

Table 1 – Research questions

2.2. Inclusion, exclusion and quality criteria

The inclusion and exclusion criteria that allowed discriminating articles that do not correspond to the areas of knowledge referred to in the study, the years of research, and the selected databases. Relevant articles that answer the raised research questions are considered. Details are presented in Table 2.

Criteria	Inclusion	Exclusion
Theoretical field	Corruption and data mining available.	Web and mobile phone applications.
Databases	Web of Science (WOS) o Scopus	Google Scholar and other index files in WoS o Scopus
Type	Article, review, editorial, conference proceedings.	Speech documents, book chapters, ESCI
Año	2015-2019	Before 2015 or later tan the article publication date.
Area of research	Computer Science, Social Science, Decision-making Science.	

Tabla 2 – Inclusion, exclusion and quality criteria

As a quality criterion, a thorough review of the articles resulting from the search and after applying the inclusion and exclusion criteria is established. Articles that do not explicitly refer to data mining applications for the prevention and detection of corruption will be separated. This phase requires the reading of each of the articles obtained and is carried out by expert researchers.

2.3. Semantic Search Structure

According to the methodology, the input for the semantic structure corresponds to the thesaurus and synonymy of the concepts obtained in the “mentefacto conceptual”. The base scripts have been established, supported by the logical conjunction and disjunction operators, as well as the sequence and word relationship operators (*W/n* y *NEAR/n*). The script is organized in three levels: in the first level, data mining is approached with its characteristics and techniques; in the second level, corruption and its typology are reviewed; and in the third level the contracts and their synonymy are analyzed, as shown in Table 3.

L1	(mining W/4 (data OR video OR text OR web)) OR classificat* OR cluster* OR regression OR (association W/2 rules) OR detection OR prediction OR (sequential W/2 patterns) OR (learning W/4 (machine OR deep OR reinforced))
L2	(corruption OR bribery OR collusion OR (embezzlement OR misappropriation) OR fraud OR (abuse W/0 of W/0 discretion) OR favoritism OR nepotism)
L3	(contract OR purchase OR investment OR procurement OR acquisition OR acquirement OR tendering)

Table 3 – Semantic Search Structure

The level structure is the input to generate the final search scripts, adaptable to the databases proposed in the methodology: Web of Science and Scopus, detailed in Table 4.

Script WOS	Script Scopus
TS=(((mining near/4 (data OR video OR text OR web)) OR classificat* OR cluster* OR regression OR (association near/2 rules) OR detection OR prediction OR (sequential near/2 patterns) OR (learning near/4 (machine OR deep OR reinforced))) AND ((corruption OR bribery OR collusion OR (embezzlement OR misappropriation) OR fraud OR (abuse near/0 of near/0 discretion) OR favoritism OR nepotism)) AND (contract OR purchase OR investment OR procurement OR acquisition OR acquirement OR tendering)) Refinado por: AÑOS DE PUBLICACIÓN: (2019 OR 2018 OR 2017 OR 2016 OR 2015) AND DOMINIOS DE INVESTIGACIÓN: (SCIENCE TECHNOLOGY) AND TIPOS DE DOCUMENTOS: (ARTICLE OR REVIEW OR EDITORIAL) AND ÁREAS DE INVESTIGACIÓN: (COMPUTER SCIENCE)	TITLE-ABS-KEY (((mining W/4 (data OR video OR text OR web)) OR classificat* OR cluster* OR regression OR (association W/2 rules) OR detection OR prediction OR (sequential W/2 patterns) OR (learning W/4 (machine OR deep OR reinforced))) AND (corruption OR bribery OR collusion OR (embezzlement OR misappropriation) OR fraud OR (abuse W/0 of W/0 discretion) OR favoritism OR nepotism) AND (contract OR purchase OR investment OR procurement OR acquisition OR acquirement OR tendering)) AND (LIMIT-TO (SUBJAREA , "COMP") OR LIMIT-TO (SUBJAREA , "DECI") OR LIMIT-TO (SUBJAREA , "SOCI")) AND (LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018) OR LIMIT-TO (PUBYEAR , 2017) OR LIMIT-TO (PUBYEAR , 2016) OR LIMIT-TO (PUBYEAR , 2015))

Tabla 4 – Search Script

3. Results

After the systematic search, 219 articles are obtained from Scopus and 250 from Web Of science, applying the inclusion, exclusion and quality criteria, 153 search articles are obtained in both databases; the number of repeated articles is 6, obtaining a total of 147 articles considered for mapping. The list of articles, as well as the search scripts, can be reviewed in the following link: http://bit.ly/DMCorruption.

3.1. RQ1: How many studies are in the WOS and Scopus databases from 2015 to 2019?

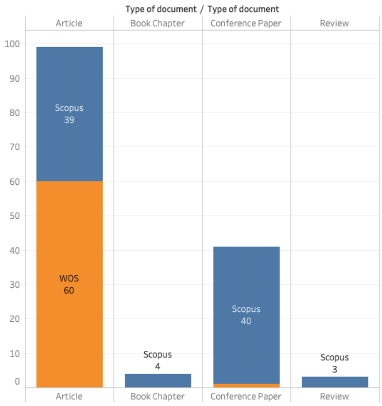

Figure 2 – Research Question 1

Of the 147 resulting articles, 61 correspond to Web of Science (WOS) equivalent to 41.49% and 86 to Scopus equivalent to 58.5%. Regarding the type of document, it is evident that the number of articles in the two databases is equivalent to 67.35%, the number of conference paper is 41, equivalent to 27.89%, of which 40 are from the Scopus database (Figure 2); the representation of book chapters is 2.72% and 2.04% revisions. Respecting the open access articles represent 16.32%.

3.2. RQ2: What are the papers with more cites?

The most cited study (Moro, Cortez, and Rita 2015) (c = 62) refers to an analysis of business intelligence literature for bank fraud using text mining, which extend to the field of study of the banking sector, to studies (7, 20, 30, 72, 84, 107) with appointments greater than 30. Types of corruption: Fraud (77.49%), overpricing (7.05%), bribery (5.05%) and favoritims (4.66%) generate greater citations in the articles. In Table 5, the articles are sorted by the number of appointments followed by the identification number.

Number of cites	Papers
>30	7, 20, 30, 72, 84, 107
21 - 30	52
16 - 20	5, 27,58, 97, 21
11 - 15	15,24,55,128,135, 144, 88,109, 147
6 - 10	2,3,43,76,112,144, 29,12, 16,63, 49,77, 83,102,123
5	39,46,56,96,131
4	35,48,69,103
3	8,37,53,59,71,98,143110,115,118,139,142
2	25,32,36,44,51,57,65,78,85,89,119,121,124, 125,132,146
1	1,6,11,14,26,33,42,45,64,67,68,70,73,80,87,95,99,106,108,111, 130,134,137
0	4,9,10,12,13,17,18,19, 22,23,28,31,34,38,40,41,47,50,54,60,61 62,66,74,75,79,81,82,86,90,91,92,93,94,100,101,104,105,113,116 117,122,126,127,129,133,136,138,140,141,145

Table 5 – Most cited articles

Analyzing the countries, in Turkey, Hungary, Egypt and South Africa, all the articles cited refer to fraud, and in the vast majority of countries they predominate widely Figure 3. Kazahistan is an exception because all quotes are from articles referring to Laundering, and countries such as Iran, Taiwan, France and Neatherlands excels Overpricing