Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 55 (2015) 261 – 267

Procedia Computer Science

Information Technology and Quantitative Management (ITQM 2015)

Text Mining Business Intelligence: a small sample of what words can say

Célia Satiko Ishikiriyama[a,]*, Diego Miro[b], Carlos Francisco Simões Gomes[a]

[a,]Universidade Federal Fluminense, Engineering School, Rua Passo da Pátria, 156, Niterói 24.210-240, Brazil
[b]Escola Nacional de Ciências Estatísticas, Rua André Cavalcanti, 106, Bairro de Fátima , Rio de janeiro, 20.231-050, Brazil

Abstract

Business Intelligence (BI) has been an object of study for many researchers around the world. From collecting, treating and storing data, to systems solutions, database administration, and analysis techniques applied to various fields such as retail, call centers, financial institutions, and telecommunication. This paper aims to present a small sample of what is possible to achieve by analyzing text data from academic papers by using the software R-project. The methodology consisted of analyzing a sample of the top 35 most relevant papers regarding Business Intelligence, obtained through an academic search engine and offered the results of this text mining study.

Peer-review under responsibility of the Organizing Committee of ITQM 2015

Keywords: Business Intelligence; Data Analysis; Text Mining; R-project

1. Introduction

Much has been said about the importance of the application of information management and bringing knowledge to business in order to make better decisions on a day-by-day basis. Business Intelligence (BI) is here to stay. Intelligent companies achieve competitive intelligence. Agility and responsiveness to changes give businesses the ability to compete in a changing global economy and knowing the business environment has become key to keep businesses profitable and competitive [1]. Information systems can make or significantly facilitate the emergence of innovations [2].

* Célia Satiko Ishikiriyama. Tel.: +55-21-98814-9483.
E-mail address: csatiko@gmail.com.

Peer-review under responsibility of the Organizing Committee of ITQM 2015
doi:10.1016/j.procs.2015.07.044

BI solutions have been made a priority by organizations who implemented these solutions [3]. BI not only supports the decision making process but also allows businesses to have a better insight into their operations by applying data analysis techniques to their information [1]. The use of BI also makes it possible for an organization to include intelligent behavior in its base functions [4].

BI provides businesses the necessary support for decision making by using a collection of techniques and tools [5]. It is possible to identify three main groups of activities to achieve intelligence in their business [4]:

- Access, integrate and store data from different sources;
- Analyze data and transform it into information;
- Present information.

To implement BI sounds easy, but it can be tricky and challenging, depending on the complexity of the business, the number of different operational systems in use and the quality of available data. Because of the amount of ever-increasing information, major difficulties and challenges are being posed towards the decision-making process [5]. Integrating business intelligence and software agents can provide solutions to these problems [4].

Within this scenario, Information Technology (IT) plays a paramount role in the success of BI in many ways. The offer of BI solutions in the market can illustrate the complexity of dealing with so much data from so many different origins. These solutions are considered the main tools to analyze and monitor performance in organizations [6].

Using a BI system and related applications can provide decision support, competitive intelligence, operational intelligence, early warning systems, sense making and support problem solving [7,12]. Implementing a BI solution must take into consideration previous planning. Having a good BI architecture plan is fundamental for the success of BI implementation [8]. Also, the use of IT has helped organizations to explore customer relationships as never before [9].

After having some explanations about BI roles and its importance and association with IT, we can now briefly explain the technique used for our study. We chose to use text mining because we wanted to identify the main concepts surrounding BI within the 35 most relevant papers. Text mining refers to extracting information from unstructured textual data and a “bag-of-words” text representation based on a vector space [10], aiming to discover new knowledge [11].

The methodology applied to this study was divided into two main steps. The first step was to search over literature papers that discussed the object of our study: BI. The second step refers to the use of a text mining technique, using the abstracts and keywords of the selection of papers, in order to provide an understanding of the main concepts surrounding BI nowadays. To achieve this understanding, all data needed was collected from the papers and analyzed using the software R-project.

2. Researching Business Intelligence

The first step of our study consisted of collecting the papers we needed. Our goal was to select the top 35 most relevant papers BI. First, we researched “Business Intelligence or Business intelligence or business intelligence” in a Brazilian academic search engine, resulting in 100,115 items.

The first filter we selected was by type, “papers only”, resulting in 73,055. As we wanted to understand the most recent discussed issues, we filtered by year, aiming for the last ten years of publications, from 2004 to 2014 (47,736 results). Then, we chose only peer reviewed (23,068), and finally, by topics, selecting only topics related to decision making and data analysis, resulting in 9,050 papers. Since they are presented ordered by relevance, we selected top 35 to give a closer look.

After this, we decided to trace the profile of selected papers by year of publishing, country, institution or university and journal. No big trends were obtained when analyzing frequency of journals or institution or

university, because the results were sprayed across the papers. When we see the series by year in Figure 1, we can see a growing interest throughout time until 2012, going down in 2013.

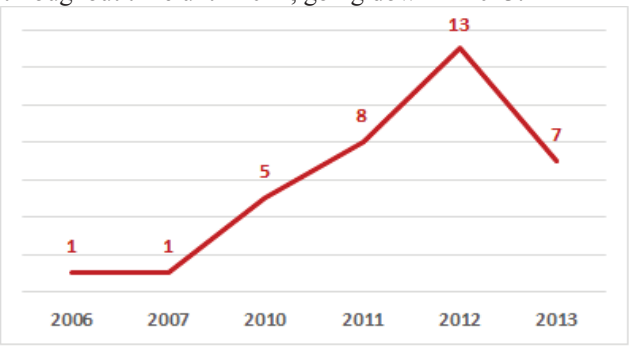

Fig. 1. Selected papers by year of publishing

Regarding the countries of universities or institutions involved in the research of the selected paper, The United States was the country that participated in the most papers, 8 out of 35, followed by Romania which participated in 7 out of 35 papers. The United States presented papers in partnership with other countries: Hong Kong (2) and Belgium (1), while Romania was only involved in papers carrying its name alone.

Grouping these countries by continent, we can see in Figure 2, Europe is the continent with the most participation in papers, with 15 papers, corresponding to 37%, followed by Asia with 29% (12) and finally North America with 10 papers, being 8 from United States and 2 from Canada.

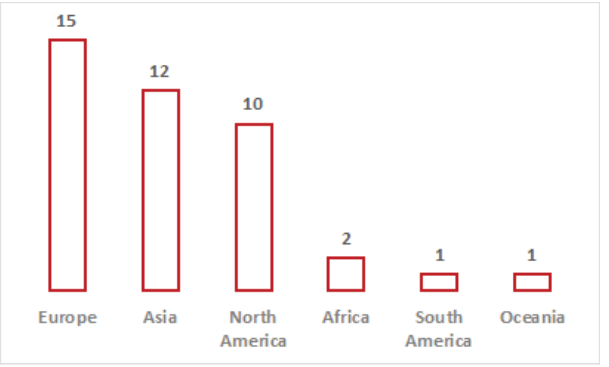

Fig. 2.Number of countries involved in the production of papers selection grouped by continent

3. Text Mining Business Intelligence

Before we got started, we had to organize the abstracts of the selected papers in a database format. First of all, we had to treat our abstract database before getting started with text mining by correcting some misspelled words and by grouping different words with the same meaning, such as “system” and “systems” or “approach” and “approaches”. It can also be part of text mining process to correct spelling mistakes and to replace acronyms and abbreviations, in order to eliminate noisy data [11].

We also excluded all the words classified as articles, prepositions, pronouns, conjunctions and determiners. After this, we ended up with the final number of 1,393 words. Frequency distribution of words can be seen in Table 1. We can see that 84.6% of the words appear in the texts up to three times with just a few words appearing a considerable number of times. Words that were mentioned 10 times or more total 46 words (3.3%).

Table 1. Frequency of words in the sample

Frequency	Quantity of Words	% Total	Frequency	Quantity of Words	% Total
1	848	60.9%	14	5	0.4%
2	231	16.6%	16	4	0.3%
3	99	7.1%	17	2	0.1%
4	90	6.5%	18	1	0.1%
5	32	2.3%	19	2	0.1%
6	16	1.1%	21	2	0.1%
7	15	1.1%	24	1	0.1%
8	7	0.5%	30	1	0.1%
9	9	0.6%	53	1	0.1%
10	8	0.6%	54	1	0.1%
11	7	0.5%	63	1	0.1%
12	3	0.2%	100	1	0.1%
13	5	0.4%	119	1	0.1%

One of the approaches to Text Mining is the integration and querying of text data after it was properly stored [11]. Using R-project, version 3.1.2 and packages tm, for text mining, and wordcloud, for word cloud, we started by analyzing the text of selected abstracts. Firstly, we generated word clouds in three different levels. In the first level, we used all the words found in the abstracts. Then, in the second level, we took out the most frequent words, to generate a second cloud and the same procedure was adopted to generate the third cloud. The results of the first level word cloud can be seen in Figure 3 and in Table 1 it is possible to verify the most frequent words per level.

Fig. 3. (a) First level word cloud; (b) Second level word cloud; (c) Third level word cloud

From this initial word cloud, apart from obvious words in our first cloud, we can observe the most frequent concepts surrounding BI (system, data, information, management and support) in Table 2. Putting together those words we come to the classic definition of what BI is: systems that provide data and information to management and support decisions.

Table 2. Frequency of words for each level of word clouds

Most frequent words	First level	Second level	Third level
Business	**119**	x	x
Intelligence	**100**	x	x
Data	63	**63**	x
Information	54	**54**	x
System	53	**53**	x
Management	30	30	**30**
Support	24	24	**24**

Having our definition constructed, we decided to dig deeper into the keywords, this time aiming at a more detailed output of what BI is connected to in recent years. In Table 3, it is possible to see the frequency of words. From 178 keywords, the concept with the highest frequency (34) is obviously BI, which has proved to be an extremely diverse world; almost 90% appeared just once among all the keywords. In Table 4, the most frequent words can be seen.

Table 3. Frequency of keywords in the sample

Frequency	Quantity of Words	% Total
1	105	87.5%
2	9	7.5%
3	2	1.7%
4	2	1.7%
7	1	0.8%
34	1	0.8%

Table 4. Most frequent words in the sample

Frequency	Words
7	Data Warehouse
4	Performance Management, OLAP (On-line Analytical Processing)
3	Data Mining, Ontology
2	Artefacts, Customer Relationship Management, Dashboard, Decision Support System, Design Science, Evaluation, Social Network, Software as a Service, Success

Our next step was to classify the keywords into six different word groups:

- Data: data related concepts, for example, data integration;
- Data Analysis: techniques applied for data analysis, like data mining;
- Process: keywords related to processes, for example, knowledge management;
- Subject: words related to the cases of papers, having no direct relation to BI ;
- Technology: systems or technology related words, for example, cloud computing;
- Concept: words that couldn´t be classified in the other five groups above like intelligence or success.

The results after grouping keywords suggest a great importance of technology when it comes down to BI. By results in Table 5, disregarding subject, since it is not our purpose to identify fields of application, Data Analysis techniques are also in the highlights.

Table 5. Frequency of keywords in the sample by words group

Group	Quantity of Words	% Total
Technology	38	31.9%
Subject	26	21.8%
Data Analysis	18	15.1%
Concept	17	14.3%
Process	15	12.6%
Data	5	4.2%

4. Conclusions and Highlights

The purpose of this study was to identify the main concepts surrounding BI. With a small but relevant sample of papers, using accessible software, we could show how analyzing the words can bring information and knowledge about a certain subject.

Main inferences:

- BI seems to be losing academic interest in the past couple of years. Our guess is that new related areas are gaining importance among academics, like Big Data phenomenon;
- A good definition for BI, given to us from analyzed is ‘systems that provide information to management and support decisions’;
- Concepts surrounding BI are diverse, showing how all embracing it can be;
- Technology is highly associated with BI, even more than Data Analysis related issues or Processes.

References

[1] Thompson, WJJ, Van der Walt JS. Business intelligence in the cloud., *SA Journal of Information Management* 2010; 12:5 pages.
[2] Sanner, TA, Manda TD, Nielsen P. Grafting: Balancing Control and Cultivation in Information Infrastructure Innovation. *Journal of the Association for Information Systems* 2014; 15:220-243.
[3] Isik O, Jones MC, Sidorova. A Business Intelligence (BI) success and the role of BI capabilities. *Intell. Sys. Acc. Fin. Mgmt.* 2011; 18:161–176.
[4] Bologa A, Bologa R. Business Intelligence using Software Agents. *Database Systems Journal* 2011; 2:31-42.

[5] Mikroyannidis A, Theodoulidis B. Ontology management and evolution for business intelligence. *International Journal of Information Management* 2010; 30:559–566.
[6] Rusaneanu A. Comparative Analysis of the Main Business Intelligence Solutions. *Informatica Economica* 2013; 17:148-156.
[7] Skyrius R, Kazakevičienė G, Bujauskas V. From Management Information Systems to Business Intelligence: The Development of Management Information Needs. *International Journal of Artificial Intelligence and Interactive Multimedia* 2013; 2:31-37.
[8] Ong I, Siew P, Wong S. A Five-Layered Business Intelligence Architecture. *IBIMA Publishing* 2011; 2011:11 pages.
[9] Phan, D. D.; Vogel, D. R.. A model of customer relationship management and business intelligence systems for catalogue and online retailers. *Information & Management* 2010; 47:69-77.
[10] Radovanović M., Ivanović M.. Text Mining: Approaches and Applications. *Novi Sad J. Math* 2008; 38:227-234.
[11] Stavrianou A., Andritsos P, Nicoloyannis N. Overview and Semantic Issues of Text Mining. *SIGMOD Record* 2007; 36:23-34.
[12] Gomes CFS, Ribeiro PCC. *Gestão da Cadeia de Suprimentos Integrada à Tecnologia da Informação*. 2nd ed. São Paulo: Cengage Learning; 2014.

Appendix A. R-project syntax for word clouds

```
library(tm)
library(wordcloud)
color <- c(rgb(0.750,0.750,0.750)
        ,rgb(0.902,0.722,0.718)
        ,rgb(0.753,0.200,0.200)
        ,rgb(0.470,0.000,0.000))
data.base <- read.csv('file.csv')
dat <- as.character(data.base[,1])
words <- VectorSource(dat)
words <- Corpus(words)
words <- tm_map(words,tolower)
words <- tm_map(words,removePunctuation)
words <- tm_map(words,removeNumbers)
words <- tm_map(words,stripWhitespace)
words <- tm_map(words, PlainTextDocument)
my_stopwords <- c(stopwords('english'),'also','will','one','many')
words <- tm_map(words,removeWords,my_stopwords)
words <- tm_map(words, PlainTextDocument)
tdm <- TermDocumentMatrix(words)
tdm <- as.matrix(tdm)
w <- rownames(tdm)
f <- rowSums(tdm)
words.joint <- matrix(c('system','systems', 'use','used','technology','technologies','analysis','analyses','service',
'services','develop','developed','benchmark','benchmarks','benchmark','benchmarking','approach','approaches','ap
plication','applications', 'challenges','challenge','company','company's','BI','(BI)'), ncol=2,byrow=TRUE)
for(i in 1:dim(words.joint)[1]){
if(sum(which(w==words.joint[i,2]))>0){
f[which(w==words.joint[i,1])] <- f[which(w==words.joint[i,1])] +
                        f[which(w==words.joint[i,2])]
f <- f[-(which(w==words.joint[i,2]))]
w <- w[-(which(w==words.joint[i,2]))]
}
}
wordcloud(w,f,c(5,1),6,colors=color)
```