Research trends in CISTI's unveiled through text mining

Sérgio Moro
Instituto Universitário de Lisboa
(ISCTE-IUL)
ISTAR-IUL
Lisboa, Portugal
scmoro@gmail.com

Bráulio Alturas
Instituto Universitário de Lisboa
(ISCTE-IUL)
ISTAR-IUL
Lisboa, Portugal
braulio.alturas@iscte.pt

Joaquim Esmerado
Instituto Universitário de Lisboa
(ISCTE-IUL)
DCTI
Lisboa, Portugal
joaquim.esmerado@iscte.pt

Carlos J. Costa
Instituto Universitário de Lisboa
(ISCTE-IUL)
ISTAR-IUL
Lisboa, Portugal
carlos.costa@iscte.pt

Abstract — **CISTI (Iberian Conference on Information Systems and Technologies) is a technical and scientific annual event dating back to 2006, whose purpose is to present and discuss knowledge, new perspectives, experiences and innovations in the Information Systems and Technologies domain (IST). The dynamics associated with IST together with the growing interest of the global community in CISTI have resulted in many topics and articles published in each conference, justifying a comprehensive analysis of the literature published. In this study, we present such analysis encompassing the last four editions, between 2013 and 2016, and a total of 677 articles. To accomplish such challenge more efficiently, we adopted text mining.**

We assessed through topic modeling how the unveiled research trends are aligned with the main conference themes. Data-driven empirical research has proven it is still a hot subject for researchers. Likewise, education and learning are also playing a significant role in CISTI's contributions. Notwithstanding, Internet and social media are highly relevant topics for the conference, although not figuring as major themes. On the other side, health is receiving less attention. Thus, this study can lead to recommendations for future CISTI's themes, in addition to providing an overview of current research trends.

Keywords – CISTI; text mining; literature analysis; research trends; topic modeling.

I. INTRODUCTION

CISTI (Iberian Conference on Information Systems and Technologies) has existed since 2006 and is held alternately in Portugal (in odd years) and in Spain (in even years). While it started as a small conference that brought together researchers from Portugal and Spain, it quickly gained a larger dimension, and today it is recognized internationally.

The conference is a technical and scientific event, whose purpose is to present and discuss knowledge, new perspectives, experiences and innovations in the Information Systems and Technologies field. In recent editions the articles accepted and published in CISTI have been made available in the IEEE Xplore Digital Library and sent for indexing in ISI, Scopus, EI, INSPEC and Google Scholar.

The main themes proposed for the conference are:

A) OMIS - Organizational Models and Information Systems
B) KMDSS - Knowledge Management and Decision Support Systems
C) SSAAT - Software Systems, Architectures, Applications and Tools
D) CNMPS - Computer Networks, Mobility and Pervasive Systems
E) HCC - Human Centered Computing
F) HIS - Health Informatics
G) ITE - Information Technologies in Education
H) AEC – Architecture and Engineering of Construction

In recent years, a number of CISTI workshops have also emerged which address other issues related to information systems and technology (IST). Nevertheless, the main contributions remain in the form of full articles, which are required to contain valuable and innovative material related to complete studies, whereas the remaining types of contributions (posters, company and short articles) are not required to provide completed research. Thus, our analysis will focus on the full articles published.

It is believed that one of the key success factors of CISTI is the fact that it allows submissions in three languages: English, Spanish, and Portuguese. Such a blend can lead to some misunderstanding, although it emphasizes the relevance of studies published in non-English languages, which are accountable for a large number of relevant localized research studies [1]. Furthermore, there are still language barriers difficult to overthrown for some researchers, who have a chance to publish their work in conferences such as CISTI, opened to submissions in other languages [2].

Considering the profusion of articles presented and published, this study sought out to unveil the main research trends of interest of researchers publishing in CISTI, and how such trends have shifted during the past four years. The vast literature included (677 articles) justified the adoption of an automated literature analysis using text mining. Furthermore, as three languages are involved, the results from independent automated analysis of articles published in each language may be compared to understand the influence of the language in the contributions to CISTI.

II. OVERVIEW

Fig. 1 shows the number of articles published over the four years distributed by the three languages. The values were obtained directly from CISTI's proceedings. While all languages made significant contributions, there seems to be a relationship between the country where the event occurs and the number of papers in the corresponding native language, with the years of 2014 and 2016 being more profitable in Spanish articles, while 2013 and 2015 show more articles in Portuguese (Tab. I). Tab. II shows the distribution of authors' nationality in the four years analyzed.

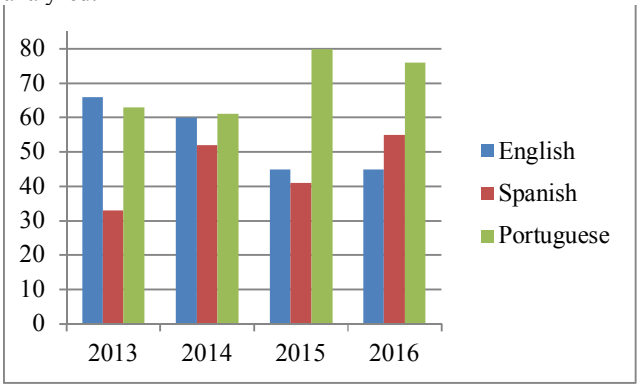

Figure 1. Number of articles published in the three languages.

TABLE I. COUNTRIES AND CITIES WHERE THE EVENTS TOOK PLACE

Year	Country	City
2013	Portugal	Lisbon
2014	Spain	Barcelona
2015	Portugal	Águeda
2016	Spain	Las Palmas

TABLE II. FULL ARTICLE AUTHORS BY COUNTRY OF ORIGIN

Country of origin	2013	2014	2015	2016
Portugal	256	194	278	207
Brazil	97	128	174	156
Spain	78	145	86	51
Colombia	23	54	50	46
Mexico	26	27	24	14
Ecuador		8	38	22
Chile		1	14	22
Argentina	3	3	8	11
Peru			4	11
USA	7	1		7
Romania	5	8		
Turkey	2	8		3
Germany	3	3	2	
Malaysia	2		6	
Croatia		6	1	
China			4	2
Czech Republic				5
South Africa		3		2
India				4
Italy		1	3	
Others	3	22	5	9

Portuguese articles are the most representative; however, the Spanish has become the second most represented language in 2016, showing that the conference is gaining traction within the Spanish speaking researchers. Also worth of note is the fact that the English language, although it declined slightly since 2013, still remains with a significant number of articles. Such figures prove the perceived value of this pluri-language conference and strengthen its core strategy.

Tab. II confirms previous statement regarding the influence of the location of the conference to increase local participants, especially on the Portuguese language. However, the year of 2016 shows a shift in this tendency, as there were fewer authors from Spain. These numbers may also be a result of the insular localization of Gran Canaria. In fact, continental Spanish may be more tempted to afford lower travel expenses to Portugal than to the Canarias. The recent years of 2015 and 2016 also show a dominance of Latin American researchers, with a wider number of authors' nationality. It should be noted that both Spanish and Portuguese are among the ten most spoken languages due to imperialism and colonialism starting in the XV century. As a result, most of the Spanish and Portuguese native speakers are not from Spain and Portugal, but essentially from the Latin America [3].

In the next sections, we dive deeply into the contents of the articles published to uncover the focus and trends in the last four CISTI editions.

III. METHODOLOGY

Text mining comprehends a set of techniques and methods for harnessing large volumes of textual documents, aiming at extracting useful knowledge from qualitative unstructured data [4]. Considering the number of articles involved, we adopted text mining in order to benefit from an automated method for analyzing the textual contents of the literature published in the four editions of the CISTI.

While topic modeling has been used for different purposes (e.g., analyzing hotel online reviews [5]), it may also be adopted for facilitating literature analysis. Hence, as a baseline for the proposed approach, two previous studies were considered. In the first, Moro et al. proposed using topic modeling for categorizing the literature through the latent Dirichlet allocation algorithm in logical topics, using a contextualized dictionary for their problem defined with the help of domain experts [6]. In the second study, the same authors went further on their approach by considering a dictionary based on articles' keywords, thus addressing the issue regarding the subjectivity of using human expertize [7]. For the present study, the approach for defining the dictionary differs: first, a simple text mining procedure is used for counting the frequency of words occurring in the indexed sections of each of the articles (title, abstract, and keywords). This is executed three times, one per language. Then, all words occurring more than ten times are considered for evaluation by a panel of two experts belonging to CISTI's organizing and scientific committees. Since one of the goals is to assess the difference in the themes of research for the three languages, the results from their analysis is a mixed dictionary

with all the three languages. Such approach addresses the issue posed by the usage of English native terms in both Spanish and Portuguese writing (e.g., "data mining"). Fig. 2 illustrates the adopted method. Tab. 3 shows the lexicon defined through this method (only English words are shown).

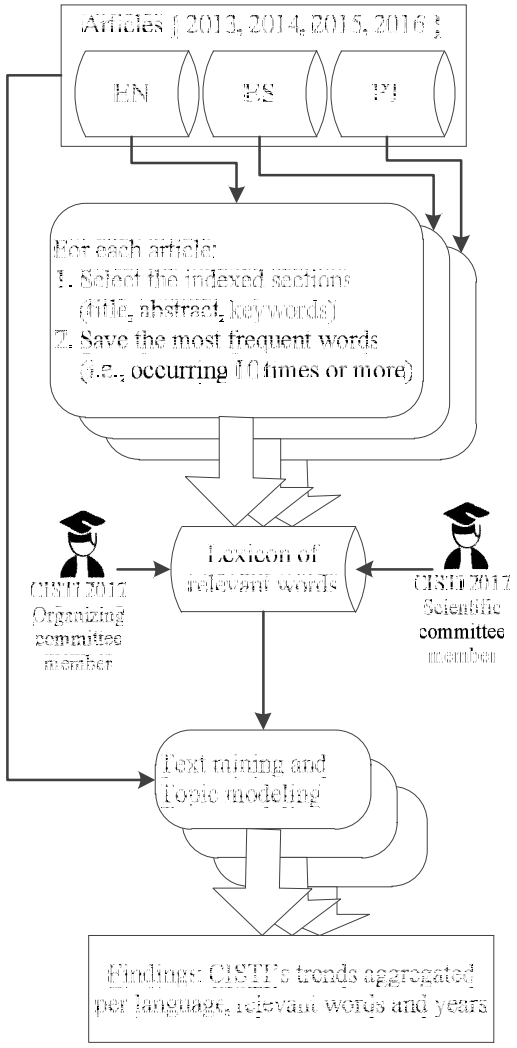

Figure 2 Methodological approach

After obtaining the dictionary of relevant words, full text articles are now parsed through text mining for computing the topics using the latent Dirichlet allocation algorithm and thus unveiling the trends of research in articles published in the three languages in CISTI's last four editions.

The R statistical tool was adopted for conducting all the experiments, as it offers an open source language with a vast number of packages for data analysis. Specifically, both the "tm" and "topicmodels" packages were used for gathering word frequency and extracting the topics.

TABLE III. LEXICON OF RELEVANT WORDS

application	human	performance
collaboration	information	planning
communication	interaction	quality
data	internet	replication
decision	knowledge	requirements
device	learning	risk
ERP	maturity	social
evaluation	mobile	usability
framework	model	video
game	network	virtual
health		

IV. RESULTS AND DISCUSSION

While topic modeling computes relations between every word, document and topic, in a three-dimensional model, given the large volume of data, for the present analysis, only the topic to which an article is mostly related is considered. The number of topics was computed in a similar procedure to Moro et al. [6], and was tuned to ten topics. Also, only the three words that best match a topic, thus serving the purpose of characterizing it, are analyzed (a β distribution value closer to zero represents a stronger relationship with the topic). The results for the three independent language-based analyses are shown on Tab. 6 (English), Tab. 7 (Spanish), and Tab. 8 (Portuguese).

The three tables show one line per topic, and are sorted in a descending order according to the number of articles that best match the topic (second column). In the first column, the topics are numbered for facilitating the discussion on the findings. The third to the eighth columns display the words that best characterize the topics, sorted by its proximity (i.e., column third shows the best matching word) and the corresponding β for each word, to understand the degree of the relationship. Finally, the last four columns show the articles from the second column distributed through the four editions, for unfolding the timely evolution of each thematic comprehended within each topic.

In Tab. 4, the most representative topic is highly related with "data", showing also some relationship with "health" and "information". One example of a study under such topic is the article by Nery et al. [8], addressing the problem of predicting Leptospirosis through data mining. In fact, health is one of the themes proposed by the conference, which may emphasize the contributions on health informatics. This topic reveals a steady number of articles over the four years. "Data" is also present in the third topic for Spanish articles, although related to the more generic word of "model" [9]. In fact, for Portuguese, the first topic, by far the most representative, is also characterized by "data". Such findings are a confirmation of the relevance of empirical data-driven research in the recent years, especially to support managerial decisions [10].

TABLE IV. TOPICS DISCOVERED FOR THE ENGLISH LANGUAGE

#	Nr.	word	β	word	β	word	β	2013	2014	2015	2016
1	36	data	0.40	health	2.24	information	2.26	11	8	8	9
2	30	application	1.70	human	1.72	usability	1.79	12	6	7	5
3	29	learning	0.23	knowledge	2.80	interaction	3.55	8	4	10	7
4	27	model	0.21	performance	3.35	decision	3.39	5	9	6	7
5	24	information	0.91	knowledge	1.68	human	1.95	10	5	5	4
6	23	game	1.07	performance	1.54	quality	1.67	8	8	3	4
7	20	internet	0.93	social	1.71	communication	1.81	2	10	2	6
8	12	risk	0.58	ERP	2.22	framework	2.28	6	3	1	2
9	8	network	0.24	social	2.54	human	3.62	1	3	3	1
10	7	virtual	0.87	video	1.46	interaction	2.65	3	4	0	0

TABLE V. TOPICS DISCOVERED FOR THE SPANISH LANGUAGE

#	Nr.	word	β	word	β	word	β	2013	2014	2015	2016
1	45	model	0.48	knowledge	1.07	human	4.5	7	12	9	17
2	33	learning	0.30	game	1.70	knowledge	3.4	8	13	3	9
3	28	data	0.10	model	2.57	human	5.25	4	8	6	10
4	16	quality	0.89	requirements	1.16	usability	2.02	4	2	7	3
5	14	device	0.78	virtual	1.46	human	2.09	4	6	2	2
6	13	human	0.35	social	1.95	knowledge	3.17	1	4	4	4
7	12	risk	0.14	model	2.84	requirements	3.86	1	2	5	4
8	8	internet	0.18	human	2.48	network	3.51	1	3	1	3
9	8	network	0.10	social	3.24	data	3.88	3	2	3	0
10	4	health	0.53	quality	1.86	human	2.24	0	0	1	3

TABLE VI. TOPICS DISCOVERED FOR THE PORTUGUESE LANGUAGE

#	Nr.	word	β	word	β	word	β	2013	2014	2015	2016
1	71	data	0.16	performance	2.24	quality	4.89	13	13	21	24
2	56	learning	0.08	knowledge	4.09	virtual	4.29	15	15	13	13
3	34	model	0.07	data	4.17	planning	4.39	7	10	9	8
4	31	quality	1.21	requirements	1.26	maturity	1.92	7	5	9	10
5	25	human	0.86	device	1.27	risk	2.38	8	4	7	6
6	19	network	0.56	social	1.52	human	1.73	3	5	6	5
7	14	game	0.44	virtual	1.71	human	2.91	3	3	5	3
8	12	usability	0.32	human	1.79	performance	3.03	2	4	2	4
9	11	internet	0.06	data	4.12	Framework	4.16	2	2	4	3
10	7	knowledge	0.21	data	2.35	Human	3.65	3	0	4	0

Other theme that seems to occur often in CISTI is gaming, as it is represented in the sixth topic in English, second topic in Spanish, and seventh topic in Portuguese, although English articles appear to be more related to performance issues, while Spanish are associated to "learning" and Portuguese to "virtual". One common trend in the conferences is "learning", as it is highly represented (second topic for Spanish and Portuguese, and third for English). Inevitably it is associated with knowledge, although gaming seems also to play a role in Spanish articles [11].

One issue that seems to be of great concern for CISTI researchers writing in Spanish or Portuguese is "quality"; on the opposite, the subject is scarcely mentioned in the English literature (appears in the sixth topic as the third most related word). "Internet" and "social" are common denominators for several topics. Nevertheless, research numbers are steady since

2013 (e.g., topic seven in English or topic six in Portuguese), which confirms the level of maturity that social media related topics have reached, remaining a top theme of research [12].

"Usability" appears in Portuguese articles as a highly relevant theme, although it has not received the same attention in the remaining two languages, as it emerges only as the third most relevant word in two topics.

V. CONCLUSIONS

CISTI is an international conference of growing interest and is spanning throughout the globe, with a special focus on the Latin American, given it allows submissions in Spanish, Portuguese, and English, providing a fast forward publishing service. In this study, we propose analyzing the last four editions of CISTI, between 2013 and 2016, using text mining and topic modeling.

While the location of the conference tends to increase the number of articles published in the host country native language (Portuguese or Spanish), all the three languages have been represented in the four editions, with a reasonably high number of contributions.

The topics unveiled through topic modeling show research trends generally aligned with the main themes proposed for the conference, although some have received further attention than others. As an example, data-driven research is in high demand, given the figures presented in the topics, with all three languages showing a large number of related articles. On the other hand, health, which is one of the main proposed themes, has paled in comparison in terms of articles published. Education and learning appears as the second dominant theme, well represented in all the languages. Research regarding quality issues has been more active in both Spanish and Portuguese, when compared to the English articles.

A research trend unveiled through the topics that is not explicitly recognized as a major conference theme is Internet and social media. Nevertheless, the number of relevant contributions has remained in quite significant figures in the last four editions. Other relevant subjects uncovered from the literature include gaming and usability. Finally, it should be stated that while some recommendations may rise from this study, a literature analysis is a task that should be carried out frequently for assessing shifts in research and maintain authors up-to-date on the most recent hot themes [13].

REFERENCES

[1] D. Hicks, P. Wouters, L. Waltman, S. De Rijcke & I. Rafols "The Leiden Manifesto for research metrics," *Nature*, vol. 520, n.º 7548, pp. 429-431, 2015.

[2] Å. Gornitzka, The internationalisation of research and higher education. In Borderless knowledge (pp. 1-11), Springer Netherlands, 2008.

[3] N. Ostler, *Empires of the word: A language history of the world* (p. 615). New York: HarperCollins, 2005.

[4] I. H. Witten, E. Frank, M. A. Hall & C. J. Pal, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

[5] A. C. Calheiros, S. Moro & P. Rita, "Sentiment Classification of Consumer Generated Online Reviews Using Topic Modeling," *Journal of Hospitality Marketing & Management*, *in-press*, 2017.

[6] S. Moro, P. Cortez & P. Rita, "Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation," *Expert Systems with Applications*, vol. 42, n.º 3, pp. 1314-1324, 2015.

[7] S. Moro, P. Cortez & P. Rita, "An Automated Literature Analysis on Data Mining Applications to Credit Risk Assessment," In *Artificial Intelligence in Financial Markets* (pp. 161-177). Palgrave Macmillan UK, 2016.

[8] N. Nery, D. B. Claro & J. C. Lindow, "Classification model analysis for the prediction of leptospirosis cases," *Information Systems and Technologies* (CISTI), 2016 11th Iberian Conference on. IEEE, 2016.

[9] I. Bonet, P. A. Pena, C. Lochmuller & A. Patino, "Fuzzy credibility for mixing different data sources in evaluating operational risk: Modelling operational risk," In *Information Systems and Technologies* (CISTI), 2014 9th Iberian Conference on (pp. 1-6). IEEE, 2014.

[10] S. Moro, P. Cortez & P. Rita, "A data-driven approach to predict the success of bank telemarketing," *Decision Support Systems*, vol. 62, pp. 22-31, 2014.

[11] J. I. P. Osma, J. A. G. Suarez, C. E. M. Marin & J. I. R. Molano, "Metric LMS: Educational evaluation platforms," In *Information Systems and Technologies* (CISTI), 2016 11th Iberian Conference on (pp. 1-6). IEEE, 2016.

[12] S. Moro, P. Rita & B. Vala, "Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach," *Journal of Business Research*, vol. 69, n.º 9, pp. 3341-3351, 2016.

[13] Y. Levy & T. J. Ellis, "A systems approach to conduct an effective literature review in support of information systems research," *Informing Science: International Journal of an Emerging Transdiscipline*, vol. 9, pp. 181-212, 2006.