Trends of Intangibles and Intellectual Capital: State of Art and Research

Florinda Matos[1], Valter Vairinhos[2,3] and Ana Josefa Matos[2]
[1] DINÂMIA'CET-IUL - ISCTE-IUL, Lisbon, Portugal
[2] ICLab - ICAA - Intellectual Capital Association, Santarém, Portugal
[3]CINAV - Naval Research Centre - Escola Naval, Almada, Portugal
florinda.matos@iscte-iul.pt
valter.vairinhos@icaa.pt
anajosefa.matos@icaa.pt

Abstract: Conference proceedings about intellectual capital and knowledge management are important sources of current ideas about intellectual capital, intangibles, knowledge management, authors, institutions, trends and how these are related. Since those meetings are periodic concentrations of the main sources – the papers' authors - of innovative ideas about those subjects, it is believed that an adequate analysis and synthesis of those documents can be useful to identify emerging concepts, topics, trends, directions and relations involving those concepts, their creators and the places where they were presented. The purpose of this paper is to provide, using as a data source the texts of conference proceedings, a comprehensive knowledge about the state of art of the research on intangibles and intellectual capital over the last decade and to identify the trends on those issues for future research. This study consists of a review of abstracts, titles, authors' names, emails and institutions, keywords and main texts of all the papers in the Proceedings of the European Conference on Intellectual Capital, presented between 2009 and 2017. The study also involves the identification and characterization of patterns, such as the main topics subjacent to such texts, including associations of concepts, and the trends of such associations, involving concepts of intellectual capital and intangibles, throughout that period and conference locations. The innovative methodology used in this study is text mining, based on the classic bag-of-words model and in more recent natural language processing approaches, incorporated in R or Python packages. This work also highlights some needs not covered by the present packages and presents directions for future researches and software development. The paper can be classified as a pilot study to support the construction of new computational and knowledge management methodologies in this area.

Keywords: Intellectual Capital; Intangibles; Text Mining; Knowledge Management

1. Introduction

ECIC's Proceedings corresponding to the years from 2009 to 2017 are the data source for this paper, which has, as the main objective, the eventual discovery of regularities, patterns or trends in the evolution of the main concepts, language and methods related with the topic of Intellectual Capital, seen in a European perspective. The work is descriptive and aims to obtain answers for the following questions: What are the main topics of interest in this series of conferences? Is there any trend in the definition of IC? What has been the evolution of important issues such as the measurement of IC? Are there any new questions and/or concepts that emerged during these 9 years of research? What are the concepts more frequently related with research in this domain?

One recurring question has to do with the relation between the concepts of Intellectual Capital (IC) and Knowledge Management (KM), so frequently employed as synonymous. What does the ECIC literature indicates about the relations between those concepts?

This kind of questions have traditionally been answered using qualitative methodologies of text analysis. In this work, quantitative statistical text analysis (data mining) through methodologies incorporated in *R packages* are exclusively used.

The paper structure is as follows: Part 1 gives a brief literature review of the evolution of the concept of IC and text mining based on R packages.

Part 2 describes the data used. The statistical methodology and software used - both *R packages* and other software - are identified in Part 3.

Part 4 presents the synthesis and main results of the data analysis. In section 4.1, a synthesis of countries participation in each ECIC edition is presented. In sections 4.2 to 4.4, the analysis of the content of the 565 papers integrating the ECIC´s proceedings is explained. The paper closes with a synthesis of the findings, conclusions, limitations and research agenda.

2. Text Mining with R Language

The use of *R language* to perform statistical analysis of texts (text mining and natural language processing) is having a fast development, especially in the last twelve years, as documented by the edition and reedition of large and high-quality R-packages. This evolution – both in methods, algorithms and software - has accelerated exponentially in the last three years (2016 to 2018). In relation to the statistical methodology, the interest in aspects of statistical inference connected with *topic modeling* is evident and it is implemented in *R packages* such as the *latent Dirichelet allocation (LDA)* by Chang (2015), which implements models connected with *Dirichelet* distribution, "*topicmodels"* by Grün and Hornik (2011) or "*structural topic models (STM)"* by Roberts, Stewart and Tingley (2018) (see also Blei & Lafferty [2007]).

Natural Language Processing (NLP) is another active area that is capturing great attention, namely in connection with the development of software related with deep learning (Chollet and Allaire, 2018).

All this is having a huge impact in the practical use of text mining in disciplines such as literature, sociology and politics, with real consequences in the evolution of society (for applications of text mining in scientific literature, see Lin and Wilbur [2007] and Boyack *et al.* [2011]).

3. Data

The data used in this work is formed by the 565 papers presented in the successive editions of ECIC from 2009 to 2017.

From a general point of view, this source of literature about IC can be considered slightly biased, being formed only by papers written for a specific source. Nevertheless, the authors believe that it can be considered an important, reliable and representative sample of the global literature that has been published about this topic. The papers were organized in a *corpus* with 565 texts with a total volume of about *20 Mb* and it was assumed that all the papers had the following structure: Title, Authors, Institutions (Authors Affiliations), Emails, Abstract, Keywords, Text and References.

It was necessary to program a tailor-made software to identify, parse and extract from those papers the aforementioned data, allowing the creation of a data-base (in MS Access) containing tables that followed the structure presented above. This set up allowed a much greater flexibility in the treatment of those documents than the usual *corpus* organization would allow, considering the current literature and the *R packages* available.

4. Methodology and Software

Since the entire data set can be seen as a time series formed by the texts presented in ECIC conferences throughout the years 2009, 2010, …, 2017, the main methodological concern in this exploratory paper was to describe the evolution of some text features relevant to obtain answers for the formulated questions.

Each text analyzed was subjected to the usual tasks of tokenization (extraction of words), elimination of stop words in English and lemmatization, reducing considerably the number of words to analyze.

The computations were based mainly in the *R-Package "quanteda"* - Quantitative Analysis of Textual Data (Benoit *et al.*, 2018), complemented with other *R packages* such as *"tm"* - Text Mining (Feinerer, Hornik and Artifex Software Inc., 2018) and the *R package "ggplot2"* by Wickham *et al.* (2018) (see also Welbers, Atteveldt and Benoit [2017]).

For the purpose of data management and analysis, the structure assumed for each paper is the one presented section 2 of this paper. The text corresponding to each part was loaded in the corresponding field of a database table, in such a way that each table entry (row) corresponds to one paper and the whole set of texts for the 565 papers forms a specific *corpus*.

In terms of meaning, each part "sees" the paper content from a specific point of view. Titles are texts used by authors to attract attention and convey to potential readers a brief but suggestive idea of the paper content.

Authors descriptions are formed by names and identify the team members that created the content and their role in the process. Emails identify the authors' emails, including the organizations to which the authors were affiliated during the paper production process. The Abstract is a brief but reliable description of the paper content, allowing a potential reader to decide to completely read or skip the paper. Texts are, obviously, the true content of the papers. The References identify documents that, in some way, influenced the paper construction, defining its context in a specific domain of knowledge.

Since the text of each part describes the paper content from a specific point of view and purpose, it is expected the existence of some degree of overlap among the sets of words used – but not a perfect coincidence in the word distributions.

Being the main objective of this paper to "sense" the evolution of ECIC's language throughout the period of 2009 to 2017, the method used in the analysis consisted in the definition of a set of eight "features", meaningful for all parts, and the observation of their values for each of the papers in each of the respective *corpus*, after the papers aggregation by year, as described before. The "features" employed were: IC (Intellectual Capital), Human, Innovation, Research, Economy, Society and Intangibles. These "features" were defined using a dictionary compatible with the packages "*quanteda*" and/or *"tm*". For example, the feature "IC" was defined to occur or be observed in a text when the words of the set {"capital", "intellectual", "knowledge", "creation", "intelligence", "asset"} occur in the text. *Innovation (2012) = 250* means that the feature/concept "Innovation" occurred 250 times for that year.

Using the function text-frequencies of the package "*quanteda*" (see Benoit et al. [2018] for the *corpus* of papers' texts), Table 1 was produced. This same function was used to count word occurrences and discover the top occurring words.

Table 1: Framework for the tables used in the analysis, containing for each year the number of occurrences of the "feature" for the papers considered that year

Year	IC	Human	Innovation	Research	Economy	Society	Intangible
2009	115	172	245	105	258	53	52
2010	55	196	186	68	188	83	33
2011	84	133	225	81	161	57	49
2012	55	146	250	57	217	64	37
2013	44	135	127	68	137	59	22
2014	37	173	121	75	120	70	17
2015	33	133	102	56	168	48	20
2016	26	100	035	43	75	56	14
2017	79	237	161	63	195	118	64

Similar tables were produced for *corpora* corresponding to Titles, Authors, Institutions, Emails, Abstract, Keywords and References. Those tables were analyzed with biplots using the software *BiplotsPmd* by Vairinhos (2003) (see section 4.3).

For *corpora* corresponding to authors, institutions and emails, what matters are the possible relations between the authors, institutions and countries involved. For this kind of result, "*quanteda" "plot_network"* was employed.

5. Data Analysis Results

5.1 Countries Involved in ECIC from 2009 to 2017

Using the information in the authors' affiliations, it is possible to assign to each paper a country and use that information to generate Figure 1, corresponding to the whole set of ECIC conferences (from 2009 to 2017).

Figure 1 shows a synthesis of the countries' participation in the successive editions of ECIC.

Because of space and graphical limitations, not all countries are displayed in figure1. The countries with a number of papers above 20 are: Romania (62), Spain (53), Portugal (45), UK (32), Poland (27), USA (26), Russia

(24), Italy (24) and Germany (21). From this result, it can be observed that the greatest number of participations in ECIC editions come from countries such as Romania, Spain and Portugal, where the subject of IC seems to have captured great attention that is translated in the number of presented works.

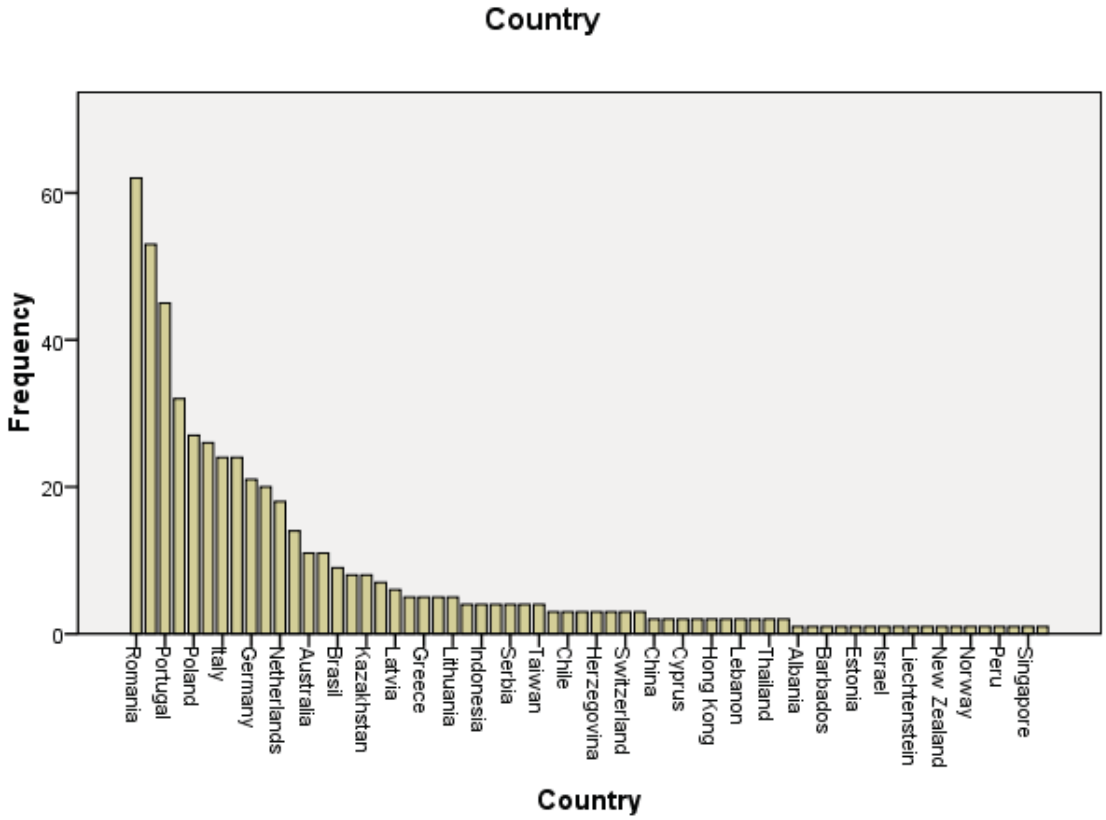

Figure 1: Countries' Participation

5.2 Top Frequent Words

Table 2 below presents both for the complete set of ECIC editions (2009 to 2017) and for each year´s edition, the list of the 20 more frequent terms present in the *corpus* of the papers' Titles, Abstracts and Texts.

Examining that table, it is possible to notice that there exists a considerable stability in the top 20 words throughout the years, suggesting, perhaps, that the emergence of new concepts has slowed down.

The concept of innovation is frequently present in these top 20 ranks confirming its association to the set {IC, Capital, Intelligence, Knowledge}. The concept of knowledge seems to be associated and frequently used as synonymous of IC.

Expressing graphically these perceptions, Figure 2, obtained using the "*quanteda"* function "*textplot _wordcloud"*, shows a word cloud of occurrence of words: the greater the frequency the bigger the label; proximities of words mean frequency of co-occurrence. In this figure, the numbers associated with the colored sectors identify the successive ECIC years. This figure, built with the results of *corpus* text analysis, attributes a central position to the concept of knowledge and its associations to capital, intellectual and management. The order associated to the sequence "Intellectual → Capital → Knowledge" points to the years 2015 and 2016, suggesting a trend that consists in using the term Knowledge (and Knowledge Management) as a synonymous of Intellectual Capital.

Table 2: Top 20 More Frequents Words in the Papers' Texts

	Texts									
	09 to 17	***Year 2009***	***Year 2010***	***Year 2011***	***Year 2012***	***Year 2013***	***Year 2014***	***Year 2015***	***Year 2016***	***Year 2017***
1	knowledge	knowledge	knowledge	knowledge	capital	capital	capital	knowledge	knowledge	Capital
2	capital	capital	capital	capital	knowledge	knowledge	knowledge	capital	capital	knowledge
3	management	value	management	management	intellectual	innovation	companies	research	management	research
4	companies	management	process	companies	companies	companies	intellectual	management	companies	companies
5	intellectual	ic	development	development	management	research	management	companies	development	intellectual
6	value	companies	value	intellectual	innovation	ic	value	organism	process	innovation
7	development	development	companies	organism	value	intellectual	process	development	ic	Studies
8	research	use	intellectual	process	development	management	development	value	model	Ic
9	process	intellectual	organism	ic	ic	value	research	studies	research	development
10	ic	research	information	value	organization	process	performance	ic	firm	management
11	innovation	process	innovation	research	research	development	innovation	intellectual	value	Value
12	organism	measure	model	innovation	process	model	measure	relation	intellectual	Relation
13	relation	asset	research	model	relation	studies	human	firm	studies	Result
14	model	organism	importance	result	employment	relation	information	human	performance	universal
15	information	relation	relation	social	human	firm	relation	social	resource	information
16	result	information	human	employment	firm	organism	importance	model	work	Level
17	human	social	metaphoric	human	result	result	ic	process	employment	performance
18	studies	intangible	organization	measure	performance	indication	organism	factor	data	organism
19	performance	difference	learning	performance	model	performance	result	employment	business	Activity
20	Import	business	difference	relation	importance	human	market	result	innovation	Human

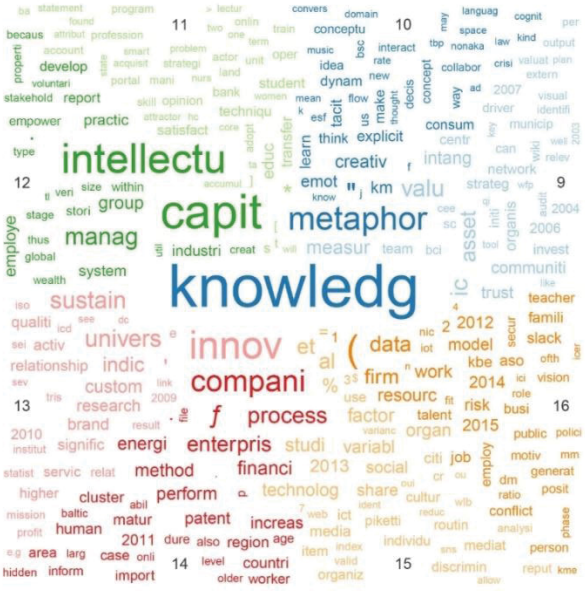

Figure 2: Word Cloud obtained with *corpus* text analysis using the "quanteda" function

5.3 Detecting Associations using Biplots

As explained in section 3 of this paper, it was decided to define 8 "features" – the same to the whole set of *corpora* – to observe the yearly aggregated sets of papers corresponding to the papers' parts. The observation of those features on the groups of aggregated texts for the same year generated Table 1 that shows, for each pair (year, feature), the frequency of observation of that feature for that year. For example, the frequency $f_{4,5}$= *217* corresponds to the frequency of occurrence of the "feature" "Economy" for the paper texts aggregated for the year 2012 and its value is 217.

Figure 3 shows a biplot corresponding to this table.

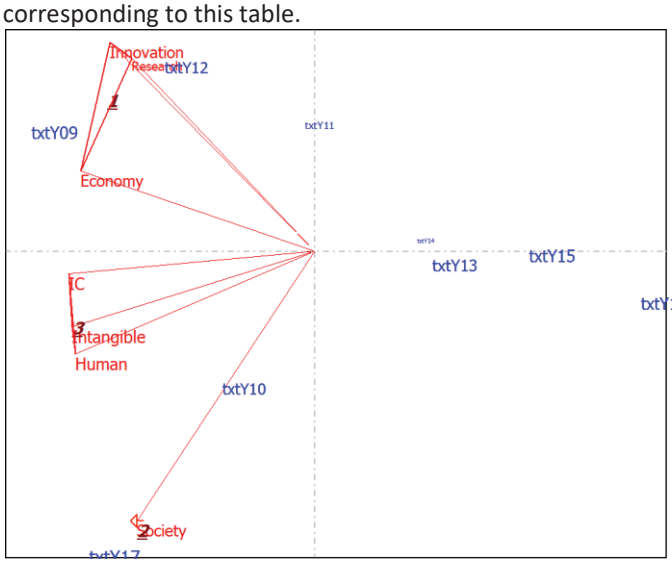

Figure 3: Biplot corresponding to Table 1, relating the features/concepts of IC with the year of occurrence. The prefix "txt" of the years' labels means that the biplot was built with the papers' texts

On this biplot, rows (corresponding to the proceedings of each ECIC) are represented by points labelled with the identification of the year prefixed with "txt". The red arrows represent the "features" (the columns of Table 1). Through an automatic classification of the coordinates of these points in the biplot, the "features" originate the clusters T_1= *{Economy, Innovation, Research}*, T_2= *{Society}*, T_3= *{IC, Intangible, Human}*. These associations could be interpreted as meaning that subjacent to the papers' texts presented in ECIC, there are 3 topics: T_1= *Economy*, T_2= *Intangibles* and T_3= *Society*. Examining the years (proceedings texts) near these groups it can be observed that the years that most contribute to explain *T1* are: *{2009, 2010, 2011, 2012}*. The set of years (meaning the associated proceedings texts) *{2013, 2014, 2015, 2016}* occupy a region in the graph corresponding to the frequencies of occurrence of the features in the set *T3*. Finally, the year 2017 is strongly associated with the feature *T2 = {Society}*.

5.4 Networks of Authors and Affiliations

Given the distinct nature of the texts corresponding to Authors, Emails and Affiliations (Institutions), a different methodology was applied. For those *corpora*, formed by identifiers of objects (persons, institutions, emails) and not by texts in the syntactic sense, the relevant aspects to analyze are the objects' networks (authors, emails, institutions), that eventually emerge from those texts.

For this kind of analysis, the *"quanteda*" function "*textplot_network"* seems especially adequate (due to space limitations, it is not possible to present here the graphics originated from the application of this function).

6. Conclusion

This work can be classified as an instrumental pilot study for more profound studies that contributes for the state of art of the concept of Intellectual Capital in the world. The main finding was that there is a progressive overlap between the concepts of IC and KM, even thought, in our view, from a logical point of view, there is a clear distinction between the two concepts. Therefore, IC concept expresses more accurately the management concepts involved.

If in the early years - 2009, 2010, 2011, 2012 -, the focus was on innovation as a growth factor of the economy, in the following years - 2013, 2014, 2015, 2016 -, the focus was on the management of intangibles. The year 2017 seems to represent a turning point, probably influenced by the Sustainable Development Challenges of the United Nations, since the focus is on society, i.e. the concept of IC appears in a more society focused perspective.

Another aspect that the research showed was the fact that the research in Intellectual Capital predominates in Portugal and Romania, countries that in the context of the European Union are associated with more difficult economies and that were very affected by the European crisis of the last decade.

One relevant result of this research, that should be highlighted, is the frequent association between the concept of knowledge management and intellectual capital. This association is not new, appearing constantly in the literature as a result of the lack of unanimity within the scientific community on the definition of intellectual capital.

Finally, this research allowed also to set up a methodology to study the evolution of texts and sets of texts throughout time.

7. Limitations

The main limitation of this exploratory research was the definition of "features" to compare texts. As explained in section 3, there is, inherently, in this approach, a subjective component that can be reduced or eliminated in future work, using text modeling and statistical inference.

Another limitation is the sample used. The sample is composed of ECIC' papers, very concentrated in countries of Europe and therefore may have left out other relevant publications realized by authors from countries that did not publish in ECIC.

8. Research Agenda

Considering the results of this pilot study, the research agenda will be a refinement of the methodology used in this study, making it easier to understand the results obtained.

Further work on the analysis of the main scientific publications on the topic of intellectual capital, published in the same period, can confirm the conclusions reached.

References

Benoit, K. et al. (2018) 'quanteda: An R package for the quantitative analysis of textual data', Journal of Open Source Software, 3(30), p. 774. doi: 10.21105/joss.00774.

Blei, D. M. and Lafferty, J. D. (2007) 'A correlated topic model of Science', The Annals of Applied Statistics. The Institute of Mathematical Statistics, 1(1), pp. 17–35. doi: 10.1214/07-AOAS114.

Boyack, K. W. et al. (2011) 'Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches', PLOS ONE. Public Library of Science, 6(3), pp. 1–11. doi: 10.1371/journal.pone.0018029.

Chang, J. (2015) 'Package "lda": Collapsed Gibbs Sampling Methods for Topic Models'. CRAN Repository. Available at: https://cran.r-project.org/package=lda.

Chollet, F. and Allaire, J. J. (2018) Deep Learning with R. New York: Manning Publications.

Feinerer, I., Hornik, K. and Artifex Software Inc. (2018) 'Package "tm": Text Mining Package'. CRAN Repository. Available at: https://cran.r-project.org/package=tm.

Grün, B. and Hornik, K. (2011) 'topicmodels: An R Package for Fitting Topic Models', Journal of Statistical Software, 40(13), pp. 1–30. doi: 10.18637/jss.v040.i13.

Lin, J. and Wilbur, W. J. (2007) 'PubMed related articles: a probabilistic topic-based model for content similarity', BMC Bioinformatics, 8(1), p. 423. doi: 10.1186/1471-2105-8-423.

Roberts, M., Stewart, B. and Tingley, D. (2018) 'stm: R Package for Structural Topic Models', Journal of Statistical Software. doi: 10.18637/jss.v000.i00.

Vairinhos, V. M. (2003) Desarrollo de un Sistema de Minería de Datos Basado en los Métodos de Biplot. Universidad de Salamanca, España.

Welbers, K., Atteveldt, W. Van and Benoit, K. (2017) 'Text Analysis in R', Communication Methods and Measures. Routledge, 11(4), pp. 245–265. doi: 10.1080/19312458.2017.1387238.

Wickham, H. et al. (2018) 'Package "ggplot2": Create Elegant Data Visualisations Using the Grammar of Graphics'. CRAN Repository. Available at: https://cran.r-project.org/package=ggplot2.