https://doi.org/10.26552/com.C.2017.1.3-8

Andrej Gogora*

THE TEXT MINING OF ETHICS AND INFORMATION TECHNOLOGY

The aim of contribution is to provide computational analysis of the journal Ethics and Information Technology by means of application of digital text mining and statistical tools on the journal's data extracted from 17 volumes. The procedure consists of data set building, application of various digital tools and interpretation of outputs. The analysis is conducted in order to acquire the basic topic structure of articles, to identify most frequently used terms, collocations and their occurrences with respect to the year of publishing, and to expose basic statistical author's data. The purpose of contribution is to present statistical overview of the journal between the years 1999 and 2016, thus to illustrate, by means of computational method, its main trends and bibliometrics data.

Keywords: *Text mining, digital humanities, bibliometrics, ethics, information technology.*

Abbreviations: *EIT - Ethics and information technology; DH – digital humanities; IT/ICT – information and communications technology*

1. Introduction

Ethical aspects of new technologies have been in focus of scholars for a relatively long time – Norbert Wiener, professor of mathematics and engineering, during the 1940s created "cybernetics", a branch of applied science in which he foreshadowed most of topical issues such as computer and security, digital privacy, ethics of programmer, information networks, virtual communities. Later on in the 1970s Walter Maner coined a new field of applied ethics "computer ethics" which is concerned with ethical problems created and aggravated by computers. After that James Moor [1] offered more profound definition of computer ethics reflecting the fundamental nature and social impact of computer technology. In the 1990s Donald Gotterbarn redefined computer ethics as a professional ethics that deals with codes of conduct for computing professionals; and finally, Floridi [2] based his "information ethics" on metaphysical assumption that world is made of informational objects (humans, animals, artifacts, electronic objects, data...) regarded as potential ethical agents.

This is a brief overview of the field of computer and information ethics. The peculiarity of the article is that it doesn't provide an armchair inquiry about particular issue of this domain, but it analyzes the representative sample of journal *Ethics and information technology* (EIT) by means of computational methods. We are asking these questions: what kind of information can be computationally retrieved from EIT corpus? What kind of digital tools can be applied to analyze EIT in order to obtain relevant statistical outputs? There are three main aims:

a) to gather and prepare digital resources and to compile EIT corpus;
b) to extract various statistical data sets from corpus and to interpret the outputs;
c) to examine methodological possibilities of applied digital tools.

The purpose is to offer the overview of computer and information ethics (represented by EIT journal) through the use of quantitative computational methods. We are going to work with digital tools that fall under the approaches elaborated in statistics, computational linguistics, natural language processing, bibliometrics and digital humanities (DH).

2. Data set

Considering the topic we decided to choose the EIT journal as a representative sample for it is included in most prestige peer-reviewed journals published in English and focused on the dialogue between ethics and ICT. In addition, it is the oldest journal exclusively dealing with the given topic and abstracted in scientific indexes with high impact factor (2014: 1.021; 2015:

* **Andrej Gogora**
Department of General and Applied Ethics, Faculty of Arts, Constantine the Philosopher University, Nitra, Slovakia
E-mail: gogora@gmail.com

0.739). There are also other journals concerning with ethical consideration of ICT such as *Information, Communication and Society*; *International Review of Information Ethics*; *Journal of Information, Communication and Ethics in Society*; *Journal of Information Ethics*; *The Ethicomp Journal*. However, processing of those resources would require the amount of time-consuming work. We claim that for the purpose of this article EIT journal represents a well-balanced and sufficient data resource.

Firstly, we downloaded set of files (.html, .pdf) from EIT website [3]. In total, we collected 502 abstracts and open access articles from volume 1, issue 1 to volume 18, issue 2 (period 1999-2016) with bibliographic data (.bib, .ris). Then, we organized files according to the year of publication, number of volume/issue, and each file renamed after specific DOI number. Then we automatically converted all files (.html, .pdf) into text format (.txt) and manually corrected scattered footnotes, page numbers, paragraphs and special characters. Considering the original articles layout we identified a set of key elements – journal, year, volume, number, pages, DOI, title, author, affiliation, author's address, e-mail, abstract, keywords, text, section, subsection, paragraph, footnotes, references and acknowledgment. Each element contains particular information that can be computationally utilized. For our purpose we processed just year, DOI, title, author/s, affiliation, address, abstract, keywords and open access articles text (we stored away references and footnotes for further research). Finally, we obtained basic data set with 1,471,502 total words and 18,137 unique word forms.

Regarding the building of resources, we have to point out that this phase of digital research is commonly the most laborious and long-lasting one. On the other hand, it's highly probable that producing of widely accessible and applicable digital resources would stimulate the research activity – after the compiling of CEPHIT corpus [4] a number of corpus-based studies raised. Thus, it is crucial to support the creating of utilizable digital resources [5, p. 80-82] as well as making of hand-lists of resources [6].

3. Digital tools application

In the next section we present the result of computational processing – instead of a single tool we intentionally applied various types of accessible text mining tools within the limits of our data set. It follows that we don't postulate a specific philosophical hypothesis that would be computationally resolvable. Our aim is to demonstrate the methodological potential of digital text analysis for the needs of ethics and philosophy.

3.1. Corpus terms

In the first step, we chose a standalone version of Voyant Tools [7] a web-based text analysis environment, to retrieve the ordered list of most frequent words from tokenized EIT corpus. The first 10 terms in raw (total count) frequency are: information (14.916), privacy (8.474), moral (8.425), data (6.927), ethical (6.313), human (6.297), social (6.112), technology (5.712), ethics (5.237), computer (4.650). It perfectly matches with basic themes covering EIT journal as were declared in the first editorial [8, p. 3]. By the same tool and manual filtering we obtained a list of most frequent authors and ethical trends. Statistically, the result shows that besides contemporary scholars specialized in ethics and new technology: Floridi (1.194), Nissenbaum (309), Lessig (215), Allen (162), Friedman (139), van den Hoven (132), Coeckelber (128), Tavani (120) there is ongoing dialogue between traditional theories and authors such as Kant (1.029), Aristotle (660), utilitarianism (500), consequentialism (304), Heidegger (249), deontology ethics (236) or Kierkegaard (171). In addition to elementary statistics, this function is often the basis for another text mining analysis, for example, the tracking of changes of term frequency according to various criteria.

Moreover, considering the data representation we notify that Voyant as well as other end-user tools include powerful functions to visualize statistical-linguistics data. The Voyant functions such as "Trends" (Fig. 2), "Bubblelines", "ScatterPlot" fulfill the requirement of visualization to communicate the piece of knowledge and to help others to acquire new knowledge.

3.2. Vocabulary density

In the next step, we provide the vocabulary density statistics (in Voyant) that is defined as the number of lexical words (content words) divided by the total number of words. Simply, lexical density is a measure of how informative a text is. In this case, it's indicative that first 5 documents with highest vocabulary density are editorials and book reviews. It's explicable by short document length of these texts that brings about higher vocabulary density. In the same way, Fig. 1 shows distribution of vocabulary density across EIT by years – it illustrates that in the first 7 years (except of 2002) there is a noticeable increased incidence of higher vocabulary density then in subsequent years. It is also explainable by the direct proportion of high lexical density of given year to short document length – or more precisely, the longer text contains more re-used (non-unique) words that reduce lexical density. Johansson [9] provides an overview of solutions to problem of lexical density of texts of different lengths (VocD, Type-Token Ratio, Theoretical vocabulary analysis). Basic end-user text mining tools do not include these algorithms, however in the case of successful vocabulary evaluation, this function may

be contributive in comparing authors, genres, text types in terms of information value.

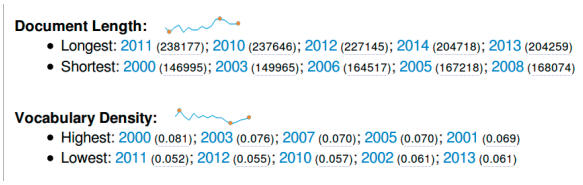

Fig. 1 Document length and vocabulary density

3.3. Collocations

Another useful feature is "Collocation Tool" that allows to search for collocations of a particular term. In this case, we took advantage of AntConc [10] software that offers more setting options for this function than Voyant. Firstly, we generated lemmatized word list (along with using basic stop-list) and then, according to top 10 terms and the interest to find out which types of ethical theories are most frequently used in EIT, we decided to look up the collocations of "ethics". We configured a query to find out all first left collocations (1L_"ethics") and to sort them out by frequency. First 10 1L collocations of "ethics" are: computer, information, applied, virtue, machine, business, discourse, global, environmental and professional. It is remarkable that bulk of the total amount consists of computer ethics (923) and information ethics (490). Statistically, it confirms the importance of these fields as we have indicated in the introduction (Fig. 2). Moreover, third right collocations of "ethics of" ("ethics of"_3R) aptly suggest the common areas covered by ethical reflection: ethics of technology, ethics of information, ethics of robotics, ethics of research, ethics of computer, ethics of warfare, ethics of care, ethics of surveillance, ethics of internet, ethics of games. Considering other frequent terms in EIT, 1L collocations of "information" are: personal, genetic, medical, digital, sensitive, health, global and private. 1R collocations of "moral" are: responsibility, agency, consideration, status, value, philosophy. Finally, 1R collocations of "social" are: media, network, justice, context, relation. These figures confirm the anticipated phrases most frequently occurring in general overviews of ethics and ICT [11, p. 25-48].

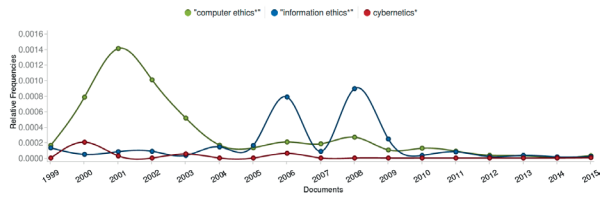

Fig. 2 Trends of collocations by years

Collocation demonstrates the sequence of words that co-occur with a frequency greater than a chance, thus it indicates the habitual juxtaposition of terms in the field of ethics and ICT. This contribution is useful in considering terminology or in more complex overview of a specific domain. Moreover, linking of various digital tools and data types, for example, searching for multiple overlapping collocations in combination with analysis of distribution of terms within the structured units of corpus, would produce valuable results (see project *Minerva – Data visualization to support the interpretation of Kant's work* [12]). The limits of computing are in principle restricted by scholar's requirement.

3.4. Cluster analysis

The cluster analysis is text mining practice in which the task is to group terms on the basis of their relatedness; it means the distance between terms in text (the smaller the distance between terms, the stronger are related to each other). It follows that terms in the same group are more similar to each other than to those in other groups – each cluster may be seen as a topic. In our case, we used VOSviewer software [13] that support creating network maps based on terms co-occurrences in documents. Our task is to build up a cluster map of the co-occurrence relations between key terms in EIT corpus. We reduced the minimum number of occurrences of a term (noun phrase) at least to 100 and it resulted in 365 terms from which we manually excluded irrelevant terms (general or non-informative). VOSviewer generated a 2-D map (Fig. 3) in which terms are located according to their co-occurrence frequencies (smaller distance = stronger relation).

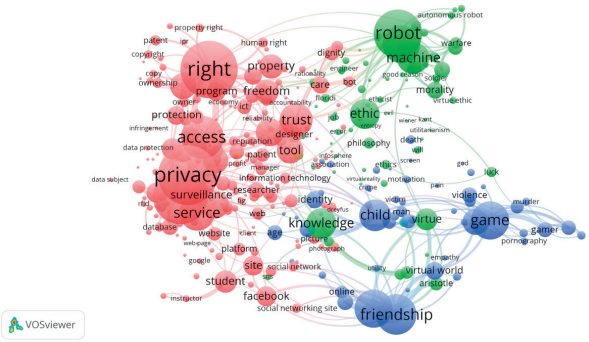

Fig. 3 Clusters network of EIT

We have to point out that the map has a number of limitations based on data availability and graphical simplification. It should be treated as an auxiliary tool to improve the knowledge of certain domain, not a perfectly valid representation. However, it is possible to suggest some findings. In brief, we can identify three main clusters in different colors with more or less central terms. First one (red) is concentrated around terms with highest

frequency of co-occurrence: privacy, right, data, access. Second one (green): ethics, robot, human, machine, knowledge; and third one (blue): friendship, game, virtual world, player, identity. In general, these clusters are the most frequent topics in EIT – dominance of "privacy/right" and "ethics/robot" clusters is not a surprise, both of them are prominent manifestations of ICT in ethics literature. Heersmink et al. [14, p. 241-249] conducted similar research by analyzing only titles and abstracts of 1.027 articles from EIT (2003-2009) and other journals. In comparison, our three clusters relatively fit their results, but in the third one there is a significant absence of the term "internet" in our map. It may be explained by the fact that the frequency graph of "internet" in corpus shows that it has the highest incidence between 2000-2003, but in the following years it is greatly reduced (on the other hand, terms "privacy" and "data" have very similar trend graph - Fig. 4). Finally, this kind of topical macro-view is for our purpose sufficient, however the further challenge is to conduct a more complex analysis of distributional semantics in natural language processing [15, p. 394-428].

3.5. Authorship's bibliometrics

Another kind of bibliometric method is statistical analysis of authorship and co-authorship. By means of VOSviewer and RIS files containing bibliographic data we received the list of authors (496) with the largest number of published articles in EIT: Floridi (14), Tavani (14), Coeckelber (8), De Laat (8), Spinello (8), Van den Hoven (8), Ess (7); and also the most frequent co-authorship: Tavani (7), Grodzinsky (5), Floridi (4), Johnson (4). On the basis of these data VOSviewer created the map of co-authorship (Fig. 5). This bibliometric result has no serious heuristic importance; nevertheless it provides the synoptic view of co-authors relationships useful in meta-philosophical or discourse analysis. Moreover, these network data may be later used for analysis of author's impact by comparing the number of his articles and occurrence of the most frequent topic clusters.

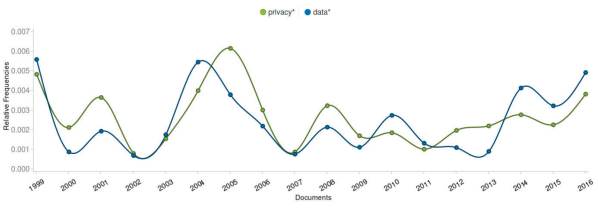

Fig. 4 Trends of "privacy" and "data" by years

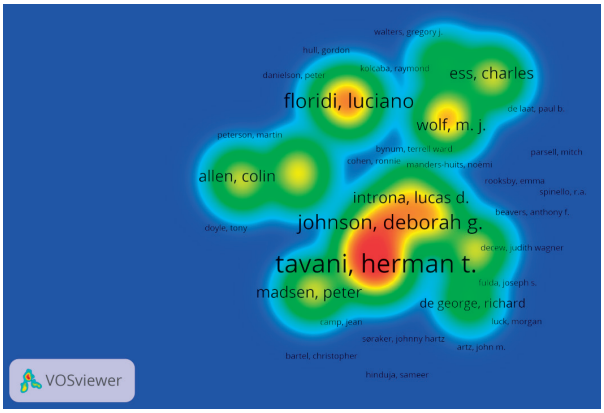

Fig. 5 Co-authorship density map of EIT

3.6. Geo-mapping

According to author's contact information in header of EIT articles, as well as given the data from previous analysis, we constructed a geographical map with layer compounds of affiliation localities. Firstly, we created a CSV spreadsheet consisting of author, affiliation, city, country, year and geocoding data (longitude, latitude). Then we calculated occurrences (including co-authors) and compiled the list of most frequent countries: USA (299), UK (106), Netherlands (101), Australia (52); cities: New York (39), Delft (56), Oxford (55), Enschede (26), Eindhoven (20), Lancaster (19); affiliations: University of Oxford (24), Technical University Delft (23), University of Twente (22), Rivier University (16), Charles Sturt University (13), Boston College (10), Lancaster University (9), Eindhoven University of Technology (9), Dartmouth College (9), CUNY (9), Carnegie Mellon University (8). Then, we uploaded CSV data set to online tool CARTO [16] and generated the interactive geographical map displaying localities, their frequency and timeline (Fig. 6). The results confirm that majority of EIT articles originate in the USA, UK and Netherlands – it is easily explainable by the country of origin of Springer publishing company (NY, USA); Springer Humanities Department based in the Netherlands; the composition of editorial board; and last but not least the English language and professional competency. However, the timeline map displays that there is an increase incidence of articles coming from other parts of the world in the last years (it fulfills the strive for international relevance in EIT editorials). In addition, there is no contribution from Central Europe region.

4. Conclusion

In conclusion, we demonstrated the way in which it is possible to apply end-user text mining software (Voyant, AntConc, VOSviewer, CARTO) to particular corpus data set. The purpose

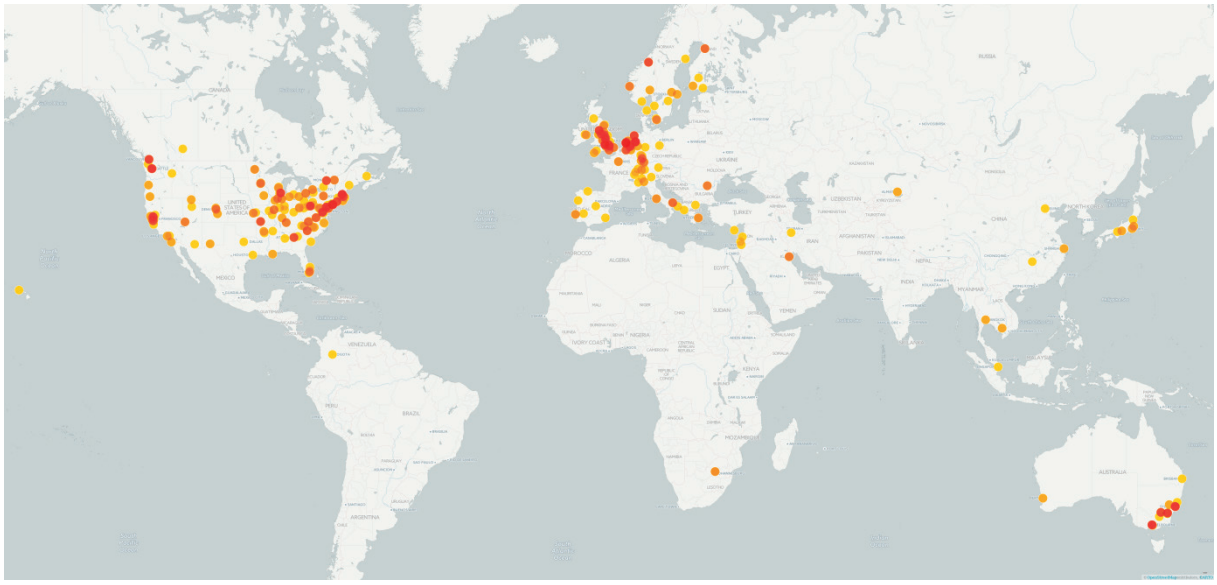

Fig. 6 Geographical map of affiliation occurrence

was to evaluate the applicability of these standardized digital tools, and to provide the basic statistical and bibliometric over-view of 16 volumes of EIT. In general, this task coincides with interdisciplinary challenge of DH to extent the benefits of digital research to practice of humanities, namely philosophical disciplines [17].

Firstly, we showed the outcome of basic term frequency analysis, interpreted it and suggested the potentiality of this data source for further research. Afterward, we tried to measure out the vocabulary density of articles, but we found out that available end-user tools are not able to accurately compare documents with different length – it needs to be processed by complex algorithms. Then we retrieved collocations of particular terms and by interpretation confirmed correctness of these results. Later on, we built up and mapped the clusters of co-occurring terms, and interpreted it according to general trends in computer and information ethics. In this case, we indicated the great opportunity to proceed with advanced types of computational semantic analysis. Lastly, we provided bibliometrics analysis of authorship and co-authorship data, and the map displaying the localities of authors´ affiliations (in respect to bibliometrics there remains the task to survey articles references).

Finally, we claim that beside the review of applicability of text mining tools we offered new pieces of knowledge concerning EIT journal. Some of them validate or contradict intuitive suppositions in a quantitative way, another one may inspire to create new interpretations or to suggest original hypotheses, but we have to draw attention to the fact that all of these digital tools are merely supportive instruments for philosophical-ethical inquiry.

Acknowledgement

The paper is part of grant project realized at the Institute of World Literature, Slovak Academy of Sciences, VEGA 2/0107/14, Hypermedia artefact in the postdigital age, 01/2014 – 12/2017.

References

[1] MOOR, J. H.: What is Computer Ethics? *Metaphilosophy*, vol. 16, No. 4, 1985, 266-275. ISSN 1467-9973.

[2] FLORIDI, L.: *Philosophy and Computing: An Introduction*. New York: Routledge, 1999. ISBN 0-415-18025-2.

[3] Available on the Internet at: http://www.springer.com/computer/swe/journal/10676

[4] MOSKOWICH, I., CAMINA, G., LAREO, I., CRESPO, B.: *The Conditioned and the Unconditioned: Late Modern English Texts on Philosophy.* Amsterdam: John Benjamins, 2016. ISBN 978 90 272 1229 0.

[5] KONVIT, M.: On Access to University Information Resources. *Communications - Scientific Letters of the University of Zilina*, vol. 14, No. 1, 2012, 80-82. ISSN 1335-4205.

[6] KRALIK, R., PAVLIKOVA, M.: The Reception of Kierkegaard's Work in Slovakia (in Slovakia). *Filozofia*, vol. 68, No. 1, 2013, 82-88. ISSN 0046-385X; KRALIK, R.: The Reception of Soren Kierkegaard in Czech language (in Czech). *Filosoficky casopis,* vol. 61, No. 3, 2013, 443-351. ISSN 0015-1831.

[7] Available on the Internet at: http://voyant-tools.org/
[8] VAN DEN HOVEN, J., INTRONA, L. D., JOHNSON, D. G., NISSENBAUM, H.: Editorial. *Ethics and Information Technology*, vol. 1, No. 1, 1999, p. 3. ISSN 1388-1957.
[9] JOHANSSON, V.: Lexical Diversity and Lexical Density in Speech and Writing: A Developmental Perspective. *Department of Linguistics and Phonetics: Working Papers in Linguistics*, vol. 53, 2008, pp. 61-79, *Lund University.*
[10] Available on the Internet at: http://www.laurenceanthony.net/software/antconc/
[11] BYNUM, T. W.: Milestones in the History of Information and Computer Ethics. *The Handbook of Information and Computer Ethics*. Kenneth E. Himma, Herman T. Tavani (eds.), New Jersey: John Wiley & Sons, Inc., 2008, 25-48. ISBN 978-0-471-79959-7.
[12] Available on the Internet at: http://www.densitydesign.org/2013/08/minerva-data-visualization-to-support-the-interpretation-of-kants-work/
[13] Available on the Internet at: http://www.vosviewer.com/
[14] HEERSMINK, R., VAN DEN HOVEN, J., VAN ECK, N. J., VAN DEN BERG, J.: Bibliometric Mapping of Computer and Information Ethics. *Ethics and Information Technology*, vol. 13, No. 3, 2011, 241-249. ISSN 1388-1957.
[15] FOX, CH.: Computational Semantics. *The Handbook of Computational Linguistics and Natural Language Processing*. Alexander Clark, Chris Fox & Shalom Lappin (eds.), Wiley – Blackwell, 2010, 394-428. ISBN 978-1-4051-5581-6.
[16] Available on the Internet at: http://www.carto.com/
[17] BRADLEY, P.: "Where Are the Philosophers?" Thoughts from THATCamp Pedagogy. *J. of Digital Humanities*, vol. 1, No. 1, 2011. ISSN 2165-6673.