ANALYSIS METHOD OF RESEARCH PAPERS PUBLISHED FOR AUDIT DOMAIN, BASED ON TITLES AND KEYWORDS

GREAVU-ŞERBAN VALERICĂ
LECTOR UNIVERSITAR DOCTOR LA UNIVERSITATEA „ALEXANDRU IOAN CUZA" IAŞI, FACULTATEA DE ECONOMIE ŞI ADMINISTRAREA AFACERILOR, DEPARTAMENTUL CONTABILITATE, INFORMATICĂ ECONOMICĂ ŞI STATISTICĂ (COLECTIVUL DE INFORMATICĂ ECONOMICĂ),
e-mail: valy.greavu@feaa.uaic.ro

ABSTRACT

Representing a strong instrument of control and feedback used by top management executives, regulators institutions or independent bodies, the audit, its methods and techniques incite the interest of specialists, professionals, professors and researchers from all socio-economic activities.

The way domain experts write about audit itself is often reflected in the manner in which they choose the keywords for the title and for the article. This study is a detailed analysis of assignment to the specific thematic areas of articles published in "Financial Audit" journal, for all public appearances in electronic format from the period 2003-2015. The study is different from other similar researches by the methodology and the type of information extracted addressed. The main purpose is to identify the most used keywords in the title and content of articles published over time and insight traceability to future research directions.

The conclusions of the analysis from this article give a comprehensive picture of audit multidisciplinary, thus providing researchers, on several economic fields, an image about the content of the publication, quality information for readers, authors and future authors.

Keywords: *audit, Google Scholar, web scraping, data mining, DataMiner, R, keywords*
Jael Classification*: G00; M40*

Introduction

Audit domain in general, and financial audit in particular, is in a continuous transformation. National and international regulatory changes and improvement processes control and prevention tools provide specialists and researchers in the fields of economic and social activity an abundance of topics for discussion and investigation. When researching, any new direction or idea must reasoned and framed in an existing scientific context.

In Romania, the most prestigious publication dedicated to audit domain is journal "Financial Audit", that provides access to publications to a large number of specialists from various specializations, which entails a variety of thematic research areas: audit, accounting, computer science, finance, corporate governance.

Writing a scientific publication involves, in addition to a thorough knowledge of domain expertise, the process of organizing a scientific paper [1]. In the organization of work, titles and keywords play an important role. Journals, search engines and indexing tools classify articles based on both words used in the title and keywords specified in article [2]. Thus, the keywords are actually those that provide the correct classification of the article in a thematic area or more.

We particularly appreciate the bibliometric analysis conducted by the authors Chersan and Mironiuc in the paper *Incursion in the Audit and Accounting Research over One Decade. Intuitive Analysis on the Articles Published in "Audit Financiar" Journal*. [3]. Analyzing the same information content, they identified a "thematic structure of publications, their dynamics depending on the status and affiliation of the authors, the degree of collaboration in research / publishing journal articles and citations number".

Given the increasing interest of authors for national publications, and journals increased interest in the hierarchy of influence factors in the research world, we believe that a thorough analysis of what has been published so far is necessary from several points of view: (1) to identify trends for research topics, (2) coherence and consistency of titles compared with keywords and (3) to identify some important factors that ensure proper indexing on dissemination tools to ensure the visibility of journal and articles.

1. Literature review

Romanian literature offers a wide range of definitions concerning the concept of audit, regardless of scope: financial [4], chartered accountant [5], statutory [6], information technology [7], [8] or for business information systems [9] and for principles and values that underlie it: ethics [10], transparency [11], performance [12], governance [13] and responsibility [14], [15].

The research directions in auditing are currently divided between academic centers in Romania and professional companies working in the field of auditing. Some authors [3] identified, in their study that more than 50% of the articles are published by researchers of the Romanian universities, but at the same time, we see a steady decline in the practitioner's researchers. We believe that this decrease is not auspicious, but is justified due to the rigors of scientific research, practiced increasingly in universities, doctoral and research schools.

Specifically, a number of Romanian studies that directly address the future of research in the field of audit, are based on a theme of international regulations forecasted developments in the context of a paradigm shift to local management [16], also an intensification of attention to corporate governance and its mechanisms [17]. Another topical subject is the concern for harmonization and convergence of accounting standards, the creation of a common accounting language, in order to increase the comparability, the transparency and the relevance of financial reporting information [18].

Our view is consistent with the report of Association of Chartered Certified Accountants [19] regarding the digital evolution, how it will help reduce audit risk, and the importance of the internal control and centralized reporting.

The fundamental purpose of a research on socio-economic area of specialization is to ensure consistent development of the industry. A special role in achieving this goal is the dissemination and improvement of our articles, by indexing them into categories of interest using search engines or specific tools. Best practices in the publication of scientific papers indicate a complementarity between headlines and keywords used in the content [20]. Words in headlines are automatically indexed in databases, the main role being to guide the reader toward an area of specialization of each article. In this paper, we expose an assessment as to use keywords in titles and in article content complementarity for the journal "Financial Audit" for the period 2003-2015.

2. Substantiation of the working hypotheses

Conventionally, any scientific journal has its own analytical tools for information classification of articles for a certain period. In this way, trends and areas of science insufficiently treated in specialized articles can be determined, guiding authors in these research directions. An overview, relative to the size and scale of research at the property level, in terms of articles published by a particular journal, it is more difficult to accomplish by a reader or an external evaluator. Search engines offer very often enough results to conceive an idea or suitable starting points for a subsequent analysis. Mostly, these searches are based on titles of articles or authors' names, seldom being targeted by keywords. Specialized indexing tools use keywords from science journals for the classification on specific categories. The following assumptions establish our research work:

H1: Web scraping tools can provide a feasible sample data used for research.

H2: In articles of the journal "Financial Audit" keywords are homogeneous, providing a breakdown by subject areas.

H3: Keywords are used with a certain tendency ensuring predictability of future research areas.

Similar studies have been conducted directed to the field of medical journals [21] or social science [22].

3. Research methodology

In this section we describe the data collection procedures, structure and size and a short summary of the instruments used in data analysis. There has been applied quantitative and descriptive statistics to demonstrate the hypotheses. For data collection we used two types of web scraping tools: *DataMiner* plug-in from Chrome browser for data collection of published articles from journal "Financial Audit" indexed on scholar.google.com database, and package *rvest 0.2.0* [23] from R platform for data collection about published articles on the website revista.cafr.ro. Data analysis was accomplished using the *wordcount*, function from the package *tm* [24] of R platform.

3.1. Data structure and gathering tools

To study the number of citations and their evolution over time it has been executed a query on site scholar.google.com (last query on 10.06.2015) using the advanced search of all articles published in "Financial Audit" excluding mentions and patents. Google has a very ingenious method for protecting the content against automatic extraction tools of search results, few people being able to manage to automate these processes. For this

article it has been used the plug-in DataMiner (http://data-miner.herokuapp.com/) from Chrome browser with custom *xpath* queries.

The number of results was 386 articles. Only 380 of these articles had complete details that were be used in this article (Figure 1). The data has been exported in *CSV* format assuring the possibility to import it in Excel or R for analysis.

Table 1 - **Physical and logical structure of data collected from Google Scholar**

Entity	Logical structure	Physical structure	xpath query
Article	All data about article	All data about article	//DIV[@class='gs_ri']
Title	Article title	Article title	h3/a
Authors	Authors	Author + Name of Journal + Year + indexing tool	div[@class='gs_a']
Descriptions	Short description of article	Short description of article	div[@class='gs_rs']
Citation	Citations	Citations + various links	div[@class='gs_fl']

Raw data was imported in Excel and using *Text to Columns* operations, the physical structure being converted into a new logical structure containing: Title, Description, Citations, Authors, Year.

To collect details about keywords used in articles, it was used package *rvest* from R platform. There were queried 837 records between 2003 and 2015, from 905 articles published on site (Figure 1). Dynamic URL used in the query *"http://revista.cafr.ro/revista.php?id=", i,"&p=sumar"*, where *"i"* has generated using a "*for*" loop statement, between 1 and 143. Last issue published on site has the id 143, representing the month June of year 2015.

Table 2 – **Physical and logical structure of data collected from http://revista.carf.ro**

Entity	Logical structure	Physical structure	xpath query
Articles	All details about articles	All details about articles	//div[2]/p
Issue No	Issue No	Issue No	//div[1]/section/div[1]/a[1]
Year	Publishing Year	Publishing Year	//div[1]/section/div[1]/a[2]
Article	Details about an article	Title + Authors + Keywords	//div[2]/p/p

Raw data has been exported from R in CSV format and processed with text to columns operations in Excel, obtaining a new logical structure: Year, Issue, Title, Authors, and Keywords.

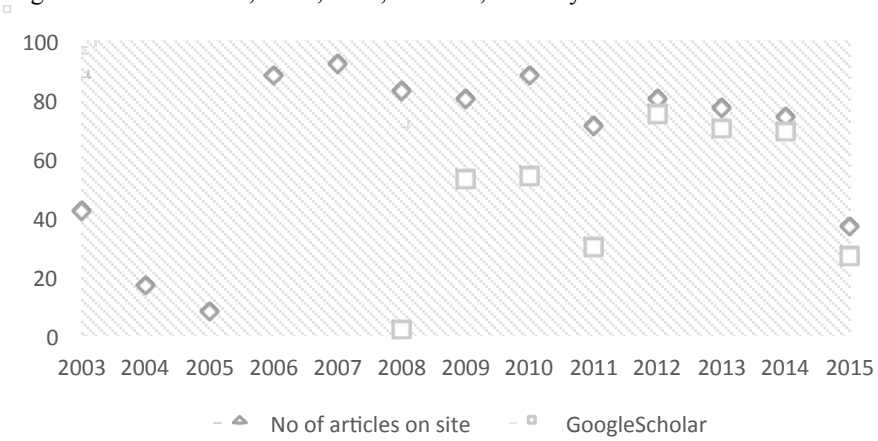

Figure 1 – **Graphical representation of data collected using different tools and methods**

Source: Own projection

The difference between the numbers of articles published on Google Scholar from those extracted directly from the site is the fact that a number of articles published in the magazine are without scientific overviews, others are written by the Big4 firms: KPMG, PwC, and etc. or other disclosures synthesis of national conferences of auditors. Google Scholar was launched in 2006 in his experimental forms, thus explaining that only find articles published after 2008.

4. Results and their interpretation

Using web scraping tools and data mining for extracting information from databases by disseminating articles provide data reproducibility study's lead researcher and at the same time the feasibility of easily updates. Alternatively, publishers can quickly identify periodically the status of dissemination, relevance and impact of published articles. In Figure 2, we present the evolution of the number of articles published in journal "Financial Audit", the trends in the number of articles cited and the actual amount of citations for articles on to each year. This number of citations is increasing year by year. The multiannual average of citations is 9.5 representing 16.24% from total amount of published articles. The number of citations of articles is an important indicator for listing and classification of scientific journals [25]. Therefore, the editorial policy of many journals recommends authors citing articles already published in that journal.

Figure 2 – **Evolution of number of articles and citations from Google Scholar database**

Source: Own projection

Web scraping tools have allowed the extraction of information about items from 2008 and 2015 but were not used this information in graphical representation because in 2008 there are only two articles on Google Scholar, and in 2015 there is no active citations.

We are convinced that the number of citations is directly related to the content of the articles, but to be able to quickly find an article for a specific research theme, each author, in the section of the literature review will attempt to make a selection of articles based on relevance titles, short abstracts and keywords within articles. Often, in addition author's name and data about publisher are the only public information indexed by certain special indexing sites. Our analysis helps readers of journal "Financial Audit" to identify most common key phrases or words used in the articles published so far, providing decision support for a specific research direction.

For counting keywords from articles published in the period 2003-2015, it was used the package *'tm' version 0.6-1* [24]. For text formatting raw data we used functions like *corpus()* for converting text, and vector analysis, and *tm_map()* to remove special characters from the text, punctuation, common prepositions or numbers.

Using *wordcount* functions we identify 1965 unique words or derivatives words used in articles classification. Analyzing the results based on keywords expressions (sample: corporate+governance, financial+audit) we found a prevalence of certain expressions presented in Table 3.

Table 3 – **Top 20 keyword expressions used in published articles in "Financial Audit" (2003-1015)**

Expression	No of appearances	Root words	Root words appearances as number	Total number of appearances in combinations
corporate governance	55	*governance*	15	84
audit	46	*audit*	238	568
financial audit	45	*financial*audit*	13	74
internal audit	39	*internal*audit*	27	79
ifrs	33	*IFRS*	15	53
independence	31	*independence*	11	43
financial reports	30	*financial*reports*	9	45
evaluation	29	*evaluation*	24	61
fair value	28	*fair*value*	3	34
performance	24	*performance*	20	55
responsibility	23	*responsibility*	6	32
financial reporting	21	*financial*reporting*	14	44
audit committee	20	*audit*committee*	8	28
internal control	20	*internal*control*	5	25
transparency	20	*transparency*	7	28
risk	18	*risk*	80	181
risks	18	*risks*	23	57
audit evidences	17	*audit*evidence*	2	18
ethics	15	*ethics*	11	32
fraud	15	*fraud*	12	32

The data in Table 3 are presented, in descending order on the number of expressions, giving the false impression that the most important topic of the articles is directed to the corporate governance research. To detail the major themes of debate we decomposed independent expressions in words or root words keywords present in the same article (columns 3, 4 and 5 of Table 3), thus offering readers the chance to identify the fact that the most used word is *audit* with its derivatives (568 appearances) and on the second place there is the word *risk* (181 appearances).

To highlight the importance of the word *audit*, we conducted a quantitative analysis of the derivatives of the word analyzing its presence in the keywords of the articles published (Table 4), identifying a few exceptions from the "root" words used in section keywords of published articles.

Table 4 – **Derivate of word audit used in section Keywords of published articles**.

Derivate	Appearance derivate in expressions	Cumulative appearance in different expressions
audit(ing)	7	9

audit(ed)	5	6
audit(ee)	1	1
audit(ee)	3	4
audit(able)	2	2
audit(or)	40	75
audit(ors)	10	16
audit(.)	15	18
audit(ors)	1	1
audit(.)	26	30
Total	110	162

Derivate *audit(ors)* appears in expression *auditors+independence* (article from 2013) and derivate *audit(ee)* appears in expression *familiarity+with+auditee+entity* (article from 2009). We have not proposed to achieve a lexical analysis of these expressions, but consider that these kind of exceptions expressions will fall from automatic indexation based on keywords practiced by certain scientific databases [26].

Given the limited space allocated for this article we will not present a table with the most common words used in independent keywords of articles, deciding to present only those that exceed the number of 100 iterations: audit (474) financial (316), risk (177), accounting (111) internal (101).

Using text mining tools and *wordcount* we found very interesting analysis of the words used in articles titles (Table 5) to determine if they are different from the keywords from the article content.

Table 5 – **The present words in the titles of articles published in the journal Financial Audit (2003-2015)**

Word	Number	Word	Number	Word	Number
looking	205	accounting	47	reporting	25
audit	156	financial	44	case	23
financial	100	analyze	42	performance	23
financial	96	aspects	36	companies	23
audit	93	international	33	auditors	22
study	85	context	31	framework	22
audit	76	ifrs	31	accounting	22
intern	75	evaluation	30	risk	22
for	73	governance	27	listed	21
from	70	Romanian	26	financial	21
on	61	auditing	25	accounting	20
of	60	auditor	25	reporting	20
Romania	59	corporate	25	standards	20
considerations	47	economics	25	entities	19

Based on the results presented in Table 5, it results that most of the articles or studies *concerning internal financial audit of Romania in the international context, offering analytical considerations aspects such as governance, performance, risks, reporting and standard alignment to IFRS.*

5. Research limits and future research

The limits of that method are linked to the use of *wordcount* method for "cleaning" the text by removing special characters, common words, and numbers. In this regard they were omitted from the analysis keywords like: ISA 545, ISA 240, ISA 570, IAS 24, and others.

More important for such an analysis would have been applying the techniques of map-reduce that requires automatic analysis of words and bringing them to the same root. Currently on open-source applications market, map-reduce-type applications are used only for analysis of texts in English. Romanian, as with other national languages, diacritics used in declination of certain words. A map-reduce processing type is more difficult in Romanian and in any other disclaimers of words there using diacritics.

Conclusions

In the context of the increasing volume of information to be analyzed by the professionals of different fields and the predominant role of computers in financial accounting records, we believe that the future direction of research that will be increasingly present is that of IT domains.

Articles that use bibliometric analysis as a research methodology present citations as the main quality indicator for themselves. We believe that publishing an article is the desire to open future opportunities for researchers' debate, analysis and research. Each researcher or professional has his own methods of analysis and interpretation of the literature, so we can conclude that *no matter how good is the instrument, the specialist is the one to give the supreme expertise.*

Acknowledgements

This work was co-financed from the European Social Fund through Sectorial Operational Program for Human Resources Development 2007-2013, project number POSDRU/159/1.5/S/134197 „Performance and excellence in doctoral and postdoctoral research in Romanian economics science domain".

References

[1] **A. Nadim**, "How to Write a Scientific Paper?," *ASJOG,* vol. 2, pp. 255-258, 3 2005.

[2] **V. Rodrigues**, "How to write an effective title and abstract and choose appropriate keywords," 4 11 2013. [Online]. Available: http://www.editage.com/insights/.

[3] **I.-C. Chersan and M. Mironiuc**, "Incursiune în cercetarea de audit şi contabilitate pe orizontul unui deceniu," *Audit Financiar,* no. 122, pp. 52-64, 2 2015.

[4] **M. Toma**, Iniţiere în auditul situaţiilor financiare ale unei entităţi, Bucuresti: Editura CECCAR, 2005.

[5] **I. Florea, I.C. Macovei, R. Florea and M. Berheci**, Introducere în expertiza contabilă şi în auditul financiar, Bucureşti: Editura Cecar, 2008.

[6] **M. Manolescu, A.G. Roman, C. Roman and M. Mocanu**, "Rolul auditului statutar în evaluarea şi consolidarea controlului intern al entităţilor auditate.," *Audit Financiar,* vol. 8, no. 2, pp. 16-20, 2 2010.

[7] **I. Ivan, G. Noșca and S. Capisizu**, Auditul sistemelor informatice, Bucureşti: 2005, 2005.

[8] **D. Homocianu and D. Airinei**, "Business Intelligence Facilities with Applications in Audit and Financial Reporting," *Audit Financiar,* pp. 17-29, 9 2014.

[9] **A. Munteanu**, Auditul sistemelor informaționale contabile. Cadrul General, Iaşi: Editura Polirom, 2001.

[10] **A. Ardelean**, "Study regarding the Clarification of Ethical Dilemmas in Financial Audit. Audit Financiar," *Audit Financiar,* no. 5, pp. 75-90, 5 2015.

[11] **M.T. Fülöp**, "Rolul guvernantei corporative eficiente în vederea întelegerii si aplicarii adecvate a principiului transparentei de catre entitatile românesti," *Audit Financiar,* pp. 48-53, 8 2012.

[12] **L. Feleagă, N. Feleagă and M. Dumitrașcu**, "Perceperea performantei organizationale în firmele de contabilitate si audit," *Audit Financiar,* pp. 36-40, 2013.

[13] **C.L. Dobroțeanu, A.S. Răileanu and L. Dobroțeanu**, "Trinomul audit extern - comitet de audit - audit intern, în contextul reglementarilor privind guvernanta corporativa," *Audit Financiar,* pp. 3-11, 4 2011.

[14] **S. Briciu, C.T. Mihăilescu and A.-M. Cordoș**, "Consideratii privind responsabilitatea si raspunderea auditorului independent în auditul statutar privind frauda," *Audit Financiar,* pp. 21-26, 12 2010.

[15] **C. Apostol**, "Corporate Social Responsability in Romania between Declaration and Implementation," in *Euro and The European Banking System: Evolutions and Challenges*, Iaşi, Editura Universitatii Al. I. Cuza, 2015,

pp. 584-590.

[16] **M. Dobre and A. Hodgson**, "Controlul intern si credibilitatea situatiilor financiare - noi directii de cercetare pe plan international," *Audit Financiar,* pp. 25-31, 11 2010.

[17] **C. Boța-Avram**, "Directii de cercetare în mediul stiintific românesc privind relevanta functiei de audit în contextul guvernantei corporative," *Audit Financiar,* pp. 10-22, 4 2011.

[18] **R. Bălășoiu**, "Consideratii privind cercetarea contabila normativa si normalizarea contabila – trecut, prezent si viitor," *Audit Financiar,* pp. 47-56, 5 2012.

[19] **ACCA**, "Darwinismul digital: evoluţie în contextul modificărilor tehnologice," *Adit Financiar,* pp. 16-20, 2 2014.

[20] **Y. Joshi**, "Why do journals ask for keywords?," 27 2 2014. [Online]. Available: http://www.editage.com/insights/why-do-journals-ask-for-keywords.

[21] **K. Jeung-Im, L. Eun-Hee, K. Hee Sun, O. Hyun-Ei, L. Eun-Joo, J. Eun-Mi and C. Suk-Hee**, "Analysis of Published Papers by Keywords and Research Methods in the Korean Journal of Women Health Nursing (2007-2009)," *Korean J Women Health Nurs,* vol. 16, no. 3, pp. 307-316, 9 2010.

[22] **J. Whittaker**, "Creativity and Conformity in Science: Titles, Keywords and Co-word Analysis," *Social Studies of Science,* vol. 19, no. 3, pp. 473-496, 8 1989.

[23] **H. Wickham**, "rvest: Easily Harvest (Scrape) Web Pages," 01 01 2015. [Online]. Available: http://cran.r-project.org/web/packages/rvest/.

[24] **I. Feinerer and K. Hornik**, "tm: Text Mining Package. R package version 0.6-1," 07 05 2015. [Online]. Available: http://cran.r-project.org/web/packages/tm/.

[25] **J.Y. Chan, K.C. Chan, J.-Y. Tong and F. Zhang**, "Using Google Scholar citations to rank accounting programs: a global perspective.," *Review of Quantitative Finance and Accounting,* pp. 1-27, 2014.

[26] **T. Berber-Sardinha**, "Using KeyWords in text analysis: Practical aspects.," *DIRECT Papers,* vol. 42, pp. 1-9, 1999.

[27] **I. Mihăilescu and C.T. Mihăilescu**, "Audit Financiar versus audit statutar. Clarificări necesare în practica profesională.," *Audit Financiar,* vol. 8, no. 2, pp. 3-9, 2 2010.

[28] **A.T. Ciuhureanu and N. Balteș**, "Etică sau creativitate în activitatea financiar-contabilă - opinii şi realităţi în organizaţiile româneşti," *Audit Financiar,* no. 6, pp. 23-31, 6 2009.