USING TEXT MINING TECHNIQUES TO ANALYSE RESEARCH TRENDS AND FOCUS, CASE STUDY OF EDUCATIONAL RESEARCH

Emmanuel Imiere. August, 2019

ABSTRACT

The goal of this study is to identify major research trends and focus in the field of educational research. To this end, 22,000 titles consisting of PhD dissertations, journal papers and conference proceedings were retrieved from the ERIC (Education Resources Information Center - https://eric.ed.gov/?) database in the period from 2009 to 2018. A topic modeling tool from Google code was downloaded and used for topic analysis, the researcher performed topic modeling for a ten year span and for each year individually using LDA. A total of 1000 keywords extracted from these titles i.e a list of 10 topics comprising of ten words each for each year was generated , which I then labeled with descriptive titles. Both word frequencies and topic modeling were analyzed at the article title level these results were visualized using tableau so that they could be easily understood. A comparison of the extracted topics by LDA with subject headings in education shows that there several distinct sub research domains strongly tied with the field. They include study, student and learning in the domain of ''Student learning''. Professional, education, teacher in the domain of ''Teacher education''. Special, education and support in the domain of special education. Reading and language in the domain of ''reading comprehension and instruction'' Secondly, the results show that there is a markedly strong rise in prominence of topics about special education and an overall dominance of topics on student learning.

INTRODUCTION

Global research efforts have increased significantly in recent years (Oecd, 2008) As the number of documents increases, traditional search-engine techniques, in which a user's keyword search simply returns a potentially long list of documents containing the keywords, are becoming increasingly inefficient because the user has to spend valuable time determining which documents contain information that is relevant to their needs. In particular, the underlying topic i.e the idea underlying the article, which may be shared with similar articles – cannot always be detected using keyword searches (Srivastava, A., and M. Sahami. ,2009).These variety and the overwhelming amount of research literature published yearly can make it difficult to effectively access and use the output of these research publications. As a result, scientists are suddenly faced with millions of publications, overwhelming their capacity to effectively use these collections and to keep track of new research (Larsen and von Ins,2010).Education policy makers and academic library's do not make as much use of research evidence as they might, this is partly because it is fragmented, difficult to find and is sometimes written in inaccessible language. As a consequence, identifying the developments and trends within educational research can be challenging and time-consuming. Developments in social science research, such as systematic reviewing, make the need for more efficient searching even more critical.

Systematic reviews can take more than a year to complete, with up to half of that time being spent searching and screening hits. Most of the work on the analysis of scientific research deals with citations (R. Rubin. 2004). This includes the examination of the frequency, patterns and graphs of citations in articles and books. citations analysis augments traditional approaches to literature searching in the systematic review process . It serves primarily as a search strategy to retrieve documents relevant to a given topic using search systems and citation index database. However, citation analysis literature (Meho) shows that 90% of papers published in academic journals are never cited. Moreover 50% of papers are never read by anyone else but their authors, referees, and journal editors. As a consequence, the sample of a study employing citation analysis method would be small and would provide less accurate mean values, unidentified outliers that would have been identified in data of much larger samples and also a larger margin of error compared to smaller margin of error for much larger samples. This would make searching and assessing vast amount of research literature difficult and time consuming. This is problematic because education policy makers often need to know the state of research evidence over a much shorter time scale than current methods allow. Educational research is being increasingly challenged for not contributing effectively enough to the improvement of policy and practice worldwide. Critics call for more relevant, cumulative, accessible and cost-effective studies. Secondly, as with any subjective review, there is the problem of selection bias researchers are often guilty of selecting the research literature best fitting their pre-conceived notions. As a result of such difficulties, text-mining techniques are receiving increasingly more attention (Ananiadou et al. 2009).Text-mining techniques have been used to enhance search and discovery options across scientific disciplines. With text mining, education policy makers, researchers and academic library can leverage these vast data and information resources to good effect. Text mining can dramatically reduce the amount of work required by the user—instead of being presented with potentially tens of thousands of documents to sift through and comb for relevant knowledge, text mining offers the possibility of automatically extracting and presenting to the user precise facts retrieved from relevant documents. Furthermore, interesting associations may be found among disparate extracted facts, leading to the discovery of new or unsuspected knowledge. The ability of decision makers within the education sector and academic library to plan educational research projects make informed collection decisions and provide customized research support is dependent in large part on how well the research interests and focus of the university or college 's faculty are mapped out. Decision-making in such areas can be facilitated through timely,formal and precise expressions for the uncertainties involved. Today, the use of data collected through Computer-based technologies and techniques is supporting a second-round of transformation in all areas of learning and research with different achievements. Data mining is a powerful new technology with great potential to help Schools and Universities focus on the most important information in the data they have collected about the behavior of their students and potential learners (Sun , 2010). Text mining a subfield of data mining is a burgeoning new field that tries to extract meaningful information from natural language text . It deals with machine supported analysis of text. According to Wikipedia, "Text mining, also referred to as text data mining, roughly equivalent to text analytics, is the process of deriving high-quality information from text." It may be characterized as the process of analyzing text to extract information that is useful for a specific purpose. The study of text mining concerns the development of various mathematical, statistical, linguistic and

pattern-recognition techniques which allow automatic analysis of unstructured information as well as the extraction of high quality and relevant data. A text document contains characters which together form words, which can be further combined to generate phrases. These are all syntactic properties that together represent already defined categories, concepts, senses or meanings [7]. Text mining must recognize, extract and use the information. Instead of searching for words, we can search for semantic patterns, and this is therefore searching at a higher level. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. The Text mining process usually involves a series of activities to be performed in order to efficiently mine the information. These activities are

*Text Pre-processing which involves Tokenization, Stop word removal, Stemming

*Text Transformation which involves feature generation and selection and attribute selection

*Text mining techniques , there are several methods that can be used in the text mining process some of them are given as follows, clustering, classification, information retrieval, topic modeling, summarization, topic extraction.

*Evaluation and interpretation of results in terms of calculating precision and recall, accuracy etc.

Text mining uses the techniques from information retrieval, information extraction as well as natural language processing and connects them with the algorithms and methods of KDD, data mining, machine learning and statistics (Sathees and Karthika, 2014). This study aims to overcome the limitations of previous approaches by applying this bottom-up approach in which research topics automatically emerge from the statistical properties of the documents. In doing so, the topics
are automatically uncovered without prior human labeling, categorization, or predefined classification of publications, and they are thus not biased by researchers' top-down subjective choices. For this purpose, a probabilistic topic model algorithm called latent Dirichlet allocation (LDA) (Blei et al., 2003) which belongs to the field of unsupervised machine learning algorithms, was used to reveal research topics within the field of education that are/were published in PhD dissertations , conference papers and peer-reviewed journals. By utilizing unsupervised machine learning, this study aims to provide comprehensive information on the focus and topical trends within educational research for education policy makers and practitioners .By automatically retrieving knowledge from unstructured text, text-mining techniques can provide enhanced views of search results, which permit users to perform more focused searches than previously possible, and allow them to locate relevant information within the retrieved documents in a more timely and efficient manner. However, the current technology of natural language processing has not yet reached a point to enable a computer to precisely understand natural language text, but a wide range of statistical and heuristic approaches to mining and analysis of text data have been developed over the past few decades. They are usually very robust and can be applied to analyze and manage text data in any natural language and about any topic. The advantage of text mining is that it enables researchers to collect, maintain, interpret, curate and discover knowledge needed for research or education, in an efficient and systematic way.

METHODS AND MATERIALS

The methodology of this study consist of 3 phases; data collection and processing ,topic analysis , and visualization or presentation.

Data collection

22,000 titles consisting of PhD dessertations, journal papers and conference proceedings were retrieved from the ERIC (Education Resources Information Center - https://eric.ed.gov/?) database in the period from 2009 to 2018.

Topic analysis

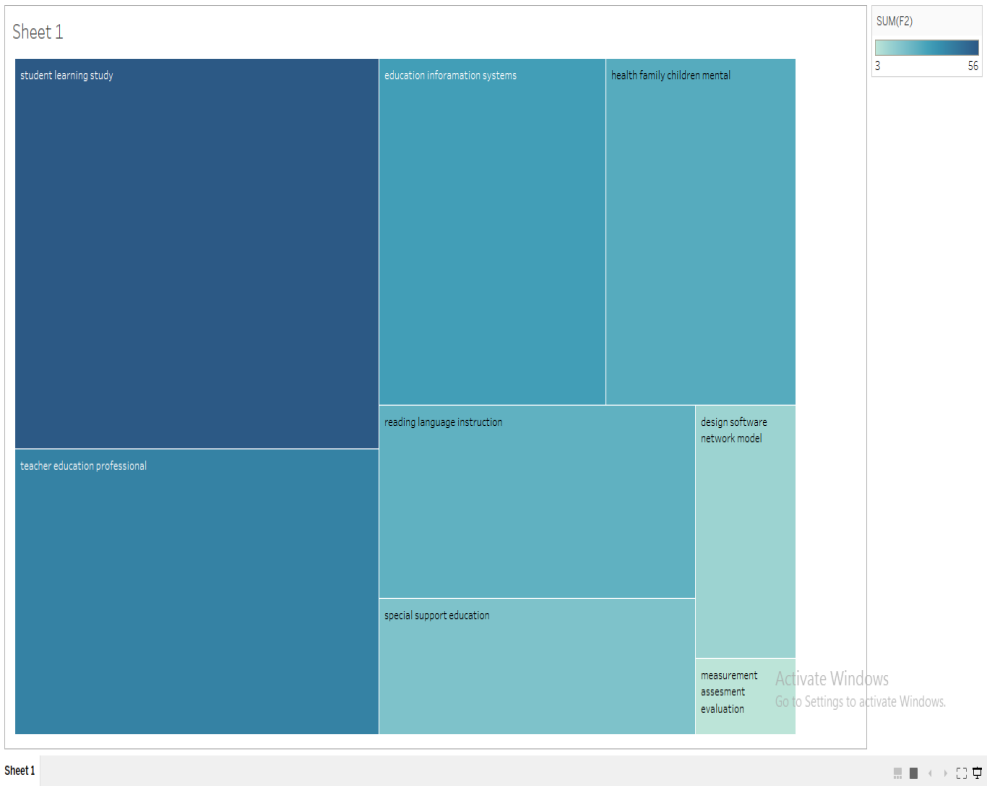
A topic modeling tool from Google code was downloaded and used for topic analysis. It is a graphical user interface tool using LDA topic modeling. It was chosen for the project due to its ease of use and the fact that it does not require coding by the user. The software gives users the ability to define the number of topics to be included and to import a unique stop words list. It's a Jupyter notebook environment that requires no setup to use and runs entirely in the cloud. The researcher examined the topics identified with latent Dirichlet allocation), and their analogues in educational research modeling.The LDA model is a generative probabilistic topic model that represents documents (i.e., educational research publications) as discrete distributions over K latent topics; each topic is subsequently represented as a discrete distribution over all the words (i.e., vocabulary) used. The words with high probability within the same topic are frequently co-occurring words, which can be seen as clusters or constellations of words that are often used to describe an underlying topic or theme (DiMaggio et al.,2013).A total of 1000 keywords were extracted from the titles i.e a list of 10 topics comprising of ten words each for each year was generated for further analysis.

Visualizations

The researcher used Tableau 10.0 to visualize text data to identify major research trends and focus. The results were visualized so that they could be easily understood. .Tableau enables the user to interactively select the criteria and explore the data on different levels. It is an effective way of portraying results for exploratory analysis.

RESULTS

A comparison of the extracted topics by LDA with subject headings in education shows that there several distinct sub research domains strongly tied with the field. They include study, student and learning in the domain of ''Student learning''.Professional, education , teacher in the domain of ''Teacher education''. Special , education and support in the domian of special education. Reading and language in the domain of ''reading comprehension and instruction'' Secondly, the results show that there is a markedly strong rise in prominence of topics about special education and an overall dominance of topics on student learning . Figure 1 shows the word frequency for all the publications.



Combining ten years of data, the topics generated were: student learning edducation school children study teacher language . Within each year, additional topics are identified. Table 1 lists the top ten topics for each year.

TABLE 1: Educational Research Publications Topics by Year

Year	Topics
2009	Education school study language learning student college information access technology
2010	Student teacher learning American information insight teens technology study adult
2011	Research professional paper teaching education school university young college program
2012	Learning children impact education language social analysis information paper technology
2013	School college teacher leadership reasoning education student early American information
2014	Teacher education learning school base college student information paper research
2015	Research learning children training teacher education school study student perception
2016	learning children development student teaching design education school reading analysis
2017	School study learning children teaching social education decision impact experiment
2018	Learning teacher student using data children study language early English research

DISCUSSION AND RECOMMENDATIONS

The academic study of education is made up primarily of the sociology, psychology, philosophy and history of education. The subject has enormous scope from the development of young children, through learning in Higher Education, to the workplace and a study of lifelong learning. The subject is significant with respect to many aspects of individual and social life. Educational research refers to the systematic collection and analysis of data related to the field of education. Research may involve a variety of methods (Lodico, Marguerite G.; Spaulding, Dean T.; Voegtle, Katherine H. 2010). Research may also involve various aspects of education including student learning, teaching methods, teacher training, and classroom dynamics(IAR: Glossary 2011). With the vast and increasing research publications, addressing pressing research questions, such as how to enhance child development and learning . Leveraging these to good effect would no doubt rely on computer and web based technology developments and also emerging new methods and techniques. In this new era of data-driven learning and teaching, researchers need to be equipped for the change with an advanced set of competencies that encompass areas needed for computationally intensive research (Buckingham Shum et al. 2013). For example, new data-management techniques are needed for big data, and new knowledge is needed for working with interdisciplinary teams with members who understand programming languages as well as the cognitive, behavioral, social and emotional perspectives on learning. A new horizon of professional knowledge is needed, including new heuristics, which incline a researcher or teaching-researcher toward computational modeling when tackling complex research problems (Gibson 2012).Interdisciplinary research will thus facilitate the integration of information, data, techniques, tools, perspectives, concepts, and theories from two or more disciplines or sources of specialized knowledge to advance fundamental understanding and to solve problems whose solutions are beyond the scope of the field of educational research. As research continue to take on a more transdiciplinary nature , we can say that future research will be so complex as to require insights from multiple disciplines. On the importance of

a transdisciplinary and multidimensional approach Jason M. Lodge, Sakinah S. J. Alhadad, Melinda J. Lewis and Dragan Gaševic discuss the need for systematic collaboration across different paradigms and disciplinary backgrounds in interpreting big data for enhancing learning. One must also acknowledge the importance of transdisciplinary conversations in moving forward research collaborations as various disciplines attempt to infer learning from big data using different methodologies. Educational policy makers and researchers need to explore these range of methodologies for synthesizing research knowledge, identify new substantive areas, new research models and modes, and insufficiently understood issues for policy-relevant research. Furthermore, an improved understanding of educational modeling approaches could help researchers to more easily synthesize historical and current research developments. This , coupled with an understanding the aspirations of educational researchers will go long way of helping educational policy makers and academic library plan educational research projects make informed collection decisions and provide customized research support. Scientific research results are of high significance as they not only affect policies and implications in scientific areas but also form an empirical basis for implications and serve as a guide for implementers (Karadag, 2009).

CONCLUSION

The results of the study showed how text mining techniques and topic modeling software, along with data visualization tools, can provide insight into the nature of the types of research being conducted in the field of education in an efficient and timely manner. The data used for the project was extracted directly from the ERIC database. It includes research journal papers, conference papers and PhD dissertations from the period 2009 to 2018 . This study demonstrates that research in the field of education shows a slight shift of scientific focus in topics and subtopics over the last 10 years. In conclusion, text mining approach revealed interesting insights into the topical trends of a large dataset of articles published in the education field. This approach enables researchers to identify research topics and shifts in research focus, and it provides a bigger picture that captures the main ideas prevailing scientific publications.

REFERENCES

Ananiadou & McNaught 2006)Ananiadou, S. & Mcnaught, J. (eds) 2006. Text mining for biology and biomedicine. Boston,MA/London, UK: Artech House.

Ananiadou S., Rea B., Okazaki N., Procter R., Thomas J. Supporting systematic reviews using text mining. Soc. Sci. Comput. Rev. 2009;27:509–523. doi: 10.1177/0894439309332293.

Blei, D. M., A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. J. Machine Learn. Res., 3: 993–1022 (2003).[Crossref], [Web of Science ®],, [Google Scholar]

Chalmers I. Trying to do more good than harm in policy and practice:the role of rigorous, transparent, up-to-date evaluations. Ann. Am. Acad. Polit. Sci. 2003;589:22– 40. doi: 10.1177/0002716203254762.

Church, K. W., and Hanks, P. 1990. Word association norms, mutual information, and lexicography. Computational Computational Linguistics 16(1):22.

DiMaggio, P., M. Nag, and D. Blei. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. Poetics, 41(6): 570–606 (2013). doi:10.1016/j.poetic.2013.08.004. [Crossref], [Web of Science ®],, [Google Scholar]

H. Sun, “Research on Student Learning Result System based on Data Mining,” Int. J. Comput. Sci. Netw. Secur., vol. 10, no. 4, pp. 203–205, 2010.

IAR: Glossary. (n.d.)". Instructional Assessment Resources. University of Texas at Austin. 21 September 2011.

Karadag, 2009. (PDF) Trends in Educational Research: A Content Analysis of the Studies Published in International Journal of Instruction. Available from: https://www.researchgate.net/publication/318115182_Trends_in_Educational_Research_A_Content_Analysis_of_the_Studies_Published_in_International_Journal_of_Instruction [accessed Jan 07 2019].

K. Sin and L. Muthu, “Application of big data in education data mining and learning analytics-A literature review,” Ictact J. Soft Comput. Spec. Issue Soft Comput. Model. Big Data, vol. 5, no. 4, pp. 1035–1049, 2015.

L. Meho. The Rise and Rise of citation analysis. Physics World, to appear

Larsen and von Ins, 2010 Larsen, P. O., and M. von Ins. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. Scientometrics, 84(3): 575–603 (2010). doi:10.1007/s11192-010-0202-z.[Crossref], [PubMed], [Web of Science ®],, [Google Scholar]

Lodico, Marguerite G.; Spaulding, Dean T.; Voegtle, Katherine H. (2010). Methods in Educational Research: From Theory to Practice. Wiley. ISBN 978-0-470-58869-7.

Oecd. Main science and technology indicators. Sci. Technol., 2008: 104 (2008).

R. Rubin. Foundations of Library and Information Science. 2nd ed. New York: Neal-Schuman, 2004

Sathees Kumar and Karthika R A survey on text mining process and techniques; International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 7, July 2014

Srivastava and Sahami, 2009 Srivastava, A., and M. Sahami. Text mining: Classification, clustering, and applications. Boca Raton, FL: CRC Press (2009).[Crossref],, [Google Scholar]