
Ni Made Satvika Iswari*, Nunik Afriliana, Suryasari
Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Tangerang, Indonesia

Abstract: Scientific studies on user-generated content extraction began in 2007. User-generated content (UGC), which is all forms of content created by someone, is widely available on social media and can influence customer desire to shop. This study aims to systematically map research trends in the field of UGC extraction over the last 15 years using metadata taken from the Scopus database. Thus, novelties and opportunities will be found that will serve as a resource for researchers conducting research and determining the research theme. Bibliometric review analysis was carried out in this study by analyzing literature from year 2007 until 2022. The search using keywords related to UGC extraction resulted in 382 papers related to the specified keywords. The main findings of this study are 1) Research in the field of UGC extraction has emerged and has grown since 2007, 2) Research in this field has been conducted by researchers from various countries, mostly from China, followed by the United States, India, Italy, Germany, Spain, etc., 3) Several keywords were discussed in this field, which include UGC, sentiment analysis, opinion mining, social media, and information extraction. This bibliometric analysis has provided information on research opportunities/directions related to UGC extraction in the future. The originality of this study is that a bibliometric analysis was performed for the research trends in UGC with a focus on technical extraction. This topic is interesting to raise because mining and extracting knowledge from UGC is quite an expensive and labor-intensive undertaking.

Keywords: user-generated content, bibliometric analysis, research trend, country, co-occurrence.

用户生成的内容提取：研究文献的文献计量分析(2007-2022)

摘要：用户生成内容提取的科学研究始于2007年。用户生成的内容(教资会)是由某人创建的各种形式的内容，在社交媒体上广泛可用，可以影响客户的购物欲望。本研究旨在使用从斯科普斯数据库中获取的元数据系统地绘制过去15年教资会提取领域的研究趋势。因此，将发现新奇事物和机会，它们将作为研究人员进行研究和确定研究主题的资源。本研究通过分析2007年至2022年的文献进行了文献计量学评论分析。使用与教资会提取相关的关键词进行搜索，得到与指定关键词相关的382篇论文。本研究的主要发现是1) 自2007年以来，教资会提取领域的研究已经出现并不断发展，2) 各国的研究人员已经开展了该领域的研究，主要来自中国，其次是美国，印度+，3) 该领域讨论了几个关键词，包括教资会、情感分析、意见挖掘、社交媒体和信息抽取。该文献计量分析提供了有关未来与教资会提取相关的研究机会/方向的信息。本研究的独创性在于对教资会的研究趋势进行了文献计量分析，重点是技术提取。这个话题很有趣，因为从教资会中挖掘和提取知识是一项相当昂贵且劳动密集型的工作。

关键词：用户生成内容，文献计量分析，研究趋势，国家，共现。

Received: July 8, 2022 / Revised: September 4, 2022 / Accepted: October 2, 2022 / Published: November 30, 2022
Fund Project: The Ministry of Education, Culture, Research and Technology of the Republic of Indonesia (Grant 430/LL3/AK.04/2022); the Universitas Multimedia Nusantara
About the authors: Ni Made Satvika Iswari, Nunik Afriliana, Suryasari, Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Tangerang, Indonesia
Corresponding author Ni Made Satvika Iswari, satvikа@umn.ac.id
1. Introduction

The need for User-generated Content (UGC) is currently growing because it provides a valuable source of data for many parties to extract information that can be used as a competitive advantage. Media outlets often highlight user-generated brand content, namely in the form of UGC. Today’s business requires a large amount of information from customers to measure and evaluate competitive progress to continue to grow. Data is now a vital asset for any business to make healthy progress. Not only information is needed, but the results of the analysis play a significant role, especially with Big Data technology.

Customer information is typically derived from survey data, and numerous researchers have created various models for identifying customer information. For instance, a hierarchical approach is used to collect customer preferences from survey data, which includes verbatim constructs, superordinate constructs, and forced constructs. Then the subscribers were analyzed using the ART2 neural network. The neural network's findings can be used to examine customer segments and conduct market research. Customers, however, might be interested in different aspects of the product. An idea for a structural topic model based on permutations was made to group customers with comparable interests [1]. This model can be used to display the frequency and order of appearance of products with various features.

Online customer data is more extensive than the data from traditional customer surveys. This data enables organizations to understand customer information more granularly. One way to figure out how people feel about a product is to look up information about it on Google [2]. The search information is used to obtain the level of customer interest in certain products and to understand the product features that are considered important. Such information can also help predict customer preferences. Another scenario is that a customer might visit several different online stores to see price comparisons for the same product.

Users can share their shopping experiences and opinions on a product or service on social media and other online platforms [3]. Comments and ratings on these websites are referred to as electronic word of mouth (eWOM), and they have a significant impact on whether a customer will purchase a product [4]. Before making a purchasing decision, consumers frequently refer to User-Generated Content (UGC), such as reviews and comments on social media and networking sites like Yelp, Twitter, and Expedia [5]. Because there is so much unstructured UGC on social media, big data analytics techniques like sentiment analysis, geographic visualization, and frequency analysis have been extensively used to analyze UGC across many different research fields.

Unfortunately, mining and extracting knowledge from UGC is quite an expensive and labor-intensive undertaking. With the development of existing data today, UGC is available in various formats and sizes. To extract and analyze it, machine learning techniques and computationally strong machines are needed to handle the volume and variation of this data [6]. In the tourism domain, neural network-based sentiment analysis is used to support the tourism business information architecture [7]. In the restaurant domain, big data analysis and content analysis are used to explore UGC related to the dining experience for customers who have food allergies [8]. The supervised machine learning method is used in cloud-based big data analytics to help small and medium-sized enterprises (SMEs) improve their designs based on customer insights [9].

A Scopus database search was conducted and the results are presented in this work to investigate the current state-of-the-art extraction of user-generated content and to provide guidance on emerging trends in related studies. The goal was to evaluate the sources of publications, articles, journals, authors, nations and organizations, research areas, and the most referenced UGC extraction themes. This report provides essential information on emerging developments in UGC extraction research. It can also indicate hotspots that could be useful as research areas. This paper’s methodology is as follows: 1.) Part 2 presents the approach used to gather sources from the Scopus database and construct a bibliometric network. 2.) In part 3, the results and discussion of the Scopus data are presented. 3.) Section 4 discusses current state-of-the-art conclusions and significant research possibilities for UGC extraction based on keyword analysis.

2. Materials and Methods

This research was conducted in several stages, as shown in Figure 1. In general, this research was conducted in three phases, namely, Data Collection, Data Analysis and Visualization, and Data Interpretation. In the first phase, a search was conducted on Scopus documents using the keywords “User-Generated Content” AND (“Web Scraping” OR “Extraction”). The keyword "Web Scraping" is more general and widely used than "Extraction", so a logical OR is used for both keywords. Based on the search results, obtained 382 papers related to the specified keywords. After that, screening is carried out on the types of documents to be processed. Documents that are included in the next process are conference papers,
articles, books, book chapters, and reviews. Meanwhile, the document that is not included in the next process is the conference review.

In the second phase, data analysis and visualization processes are carried out. Papers that have been found in the previous stage are then processed with the Open Refine tools [10]. The features provided by the Open Refine tool help to group keywords that might be alternative representations of the same thing. For example, the two strings “Text Mining” and “text mining” are very likely to refer to the same concept and just have capitalization differences. From these features, 76 clusters were found, each cluster consisting of a group of keywords that might be an alternative representation. The keywords in each cluster are then merged so that one keyword is determined to represent each cluster.

After the keywords are purified, the next step is data visualization using the VOSviewer tool [11]. VOSviewer is software used for the construction and visualization of bibliometric networks. This software allows extracting information from publications, such as author, publisher, organization, country, and keywords. Some visualization charts are also created using Power BI tools [12].

In the third phase, the data interpretation process is carried out on the visualization results that have been produced. Bibliometric analysis is used to identify scientific trends and systematize research, ensuring the quality of information and the production of the resulting results. It can describe patterns of publication in a particular field using quantitative and qualitative indicators [13]. The analysis was carried out on several aspects, namely, the growth trend of scientific publications in the field of user-generated content extraction, the distribution of publications by country, and the main keywords that are often discussed. Based on the results of the analysis, a discussion was conducted and several perspectives on future research ideas were described that could be a reference for researchers in related fields.

3. Results and Discussion

According to the Scopus database, 382 documents have been published on user-generated content extraction from 2007 to August 2022. According to Figure 2, the number of publications related to this topic tends to fluctuate but has an increasing trend every year. The publication in 2022 noted a slight decline due to the collection of information carried out in August of the same year. Articles and conference papers have shown continued growth since 2007. In 2020 and 2021, conference papers experienced a slight decline. This is because, during the COVID-19 pandemic, several scientific seminars were postponed and in the transition period to virtual/hybrid seminars to reduce the prevention of the spread of COVID-19. Figure 3 shows a map of the distribution of co-authorship by country. Based on the map, it is known that China accounts for most publications related to user-generated content extraction, followed by the United States, India, Italy, Germany, and Spain.
media (31 occurrences), and information extraction (24 occurrences). The five keywords are close to each other, which shows that research with these keywords is related and more likely to be cited in the same situation.

The field of study of user-generated content extraction is mostly associated with several machine learning techniques for data processing and big data. Unstructured big data is an important by-product of the digital age. Online reviews, blogs, tweets, and Facebook posts are just a few examples of UGC being loaded with consumer behavior insights and bringing potential opportunities for academic researchers. Researchers apply analytical methods to big data sets to gain a deeper understanding of the customer’s shopping experience and brand reputation [14].

UGC contains valuable information, such as opinions, ratings, recommendations, experiences, or customer needs. Such information is available and freely accessible through the web and/or social media data, blog posts, tweets, product reviews, or survey responses in the form of images. UGC is usually distributed in a decentralized manner (information is stored on several web platforms) and consists of different data formats (text, images, videos), and the data quality is inconsistent (e.g., spam, spelling mistakes, unusual spelling, or multiple languages). UGC is usually only available in unstructured or semi-structured forms so it cannot be processed using standard data mining techniques. Figure 7 shows some examples of User Generated Content in the form of Online Reviews on e-commerce sites Tokopedia [15] and Shopee [16].

Based on Overlay Visualization as shown in Figure 6, research trends can be seen by the year of publication. Darker colors represent an earlier publication year, while lighter colors represent a more recent publication year. Based on the visualization, it is clear that future research in the field of user-generated content extraction could be conducted using deep learning methods.
Deep learning is an important tool for Big Data analysis because it uses massive amounts of data to automatically identify patterns and extract features from complex unsupervised data without the assistance of humans [17]. The data is currently expanding. To handle the volume and variety of this data, which comes in various formats and sizes, machine learning techniques and computationally powerful machines are needed [6].

To measure and assess competitive progress and maintain growth, today’s businesses need much information from their customers. Data is now a key resource for any business to advance successfully. In addition to information, analysis results are crucial, especially with the advent of Big Data technology. Businesses and organizations use big data to make the best decisions possible. Therefore, the analysis must be correct. If the analysis used to make decisions is flawed, it will lead to poor decisions that will ultimately harm the future success of the company. This is what enables Deep Learning algorithms, which are used in big data analysis and can use feature learning to make decisions similar to those made by humans. To interpret unstructured real-world data in the real world, big data analysis can improve intelligent computing [18].

Big data analytics and social media diversity may interact favorably to improve market performance [19]. Big data analytics offers various chances to create fresh perspectives on business management and decision-making. User reviews of services and online transactions are expanding on the Internet. Customers can post opinions, suggestions, or ratings about commercial services. Online reviews in the form of user-generated content (UGC) give business managers advantages by allowing them to receive customer feedback and improve the specific product or service characteristics to boost the value of their company and support marketing initiatives [20]. Customer experience and satisfaction are growing in popularity for SME businesses because they influence customer loyalty and repeat purchases, boost word-of-mouth marketing, and enhance operational efficiency.

This study provides an overview of the main themes related to User-generated Content extraction studied recently. The number of publications related to these topics tends to fluctuate but has an increasing trend every year. This shows that this theme is increasingly in demand by researchers. The People’s Republic of China currently stands out as the country with the most publications on UGC extraction, followed by the United States, India, Italy, Germany, and Spain.

Based on keyword analysis, several keywords are widely discussed in this field, which include user-generated content, sentiment analysis, opinion mining, social media, and information extraction. The field of study of user-generated content extraction is mostly associated with several machine learning techniques for data processing and big data. Based on research trends based on the year of publication, it is known that the use of deep learning methods provides future research potential for this field. Deep learning is an important tool for Big Data analysis because it can be used to mine massive amounts of data to automatically identify patterns and extract features from complex unsupervised data without the involvement of humans. Big data analytics and social media diversity can interact favorably in a number of ways, one of which is for market performance. Big data analytics can offer various chances to create fresh perspectives on business management and decision-making. The future directions and opportunities for user-generated content extraction research have been revealed by this bibliometric analysis.

4. Conclusion and Future Perspectives

Based on the bibliometric analysis carried out in this study, 382 papers related to the specified keywords have been analyzed. The findings of this study provided information on research opportunities/directions related to UGC extraction in the future. The analysis was carried out for the research trends in UGC with a focus on technical extraction. This topic is interesting to raise because mining and extracting knowledge from UGC is quite an expensive and labor-intensive undertaking. The limitation of this study is that the analysis was carried out using the Scopus database, that the authors overlooked certain peer-reviewed studies that were not listed in Scopus.

Acknowledgment

This research was made possible through a grant from the Ministry of Education, Culture, Research and Technology of the Republic of Indonesia (430/LL3/AK.04/2022), and it received support from the Universitas Multimedia Nusantara.

References


参考文献:


