Quantitative Analysis of Hotspots and Trends in User-generated Content

Yong XU*, Ya-li WU, Qian WANG, Xin-rui ZHANG and Fan-hua LI
School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu 233000, China
*Corresponding author

Keywords: User Generated Content, CiteSpace, Visualization analysis, Sentiment analyzing.

Abstract. User generated content (UGC) is a kind of Internet product mode leaded by user under Web 2.0, which is more active in recent years. In the paper, the text mining and visual analysis tool CiteSpace was used to analyze literatures with key words of "user generated content" published in Web of Science database from 2008 to 2019. The themes of literatures, institutes of authors, key words, and co-citation were analyzed to get focuses and trends of UGC. In the paper, the research hotspots of UGC were included to provide suggestion for further work: basic information, motivation, application and analysis of UGC, and legal issues.

Introduction

User generated content (UGC) is the information produced and published by Internet users on Internet platforms, also known as User Created Content, Consumer Generated Media, Consumer Generated Content, User Contributed Content. At the beginning, UGC was appeared in the form of letters from readers, calls from audiences and so on, which was known by a few of Internet users. With the progress of information technology, Internet has gradually advocated people-oriented, and the dominant position of users has become increasingly prominent.

As a mode of creation and organization for network information resources, UGC can improve user’s participation. So it has been accepted and applied to many fields, such as blog, Wiki, forum, business review, tourism review, unexpected events, marketing management.

Data Collection

The data for bibliometric analysis is collected on the Web of Science Core Collection. With the subject keyword of User generated content, 3224 articles were downloaded as analysis data (retrieval time: June 11, 2019). Authors, abstracts, keywords, citations and other information of each article are downloaded to achieve bibliometrics. The annual number of articles is displayed in Figure 1.

Data Processing and Analysis

Co-Authorship and Co-Institute Analysis
We use Minimum Spanning Tree algorithm to implement co-authorship and co-institute analysis. The data extraction object is set to Top 20. Then, a comprehensive analysis map of publishing
institutions and authors in the field of UGC is obtained, which is shown in Figure 2. Nodes represent authors or institutions respectively, and the size of nodes represents the amount of text sent.

![Figure 3. The Top 10 most productive institutes.](image1)

![Figure 4. Distribution of keywords.](image2)

Fig. 3 lists the institutions of Top 10, in which 52 papers were sent by CAS, which occupies the highest position.

**Co-Occurring Keywords Analysis**

In order to find the research hotspots of UGC, "Keyword" network node is set up in Cite Space 5.3, and the data extraction object is Top 20. The keyword co-occurrence map is drawn as shown in Figure 4. According to the occurrence frequency of keywords, the top 20 keywords are listed in descending order (Table 1).

<table>
<thead>
<tr>
<th>Freq.</th>
<th>Keyword</th>
<th>Freq.</th>
<th>Keyword</th>
</tr>
</thead>
<tbody>
<tr>
<td>411</td>
<td>social media</td>
<td>121</td>
<td>twitter</td>
</tr>
<tr>
<td>328</td>
<td>user generated content</td>
<td>121</td>
<td>social network</td>
</tr>
<tr>
<td>241</td>
<td>word of mouth</td>
<td>109</td>
<td>network</td>
</tr>
<tr>
<td>239</td>
<td>internet</td>
<td>108</td>
<td>communication</td>
</tr>
<tr>
<td>207</td>
<td>information</td>
<td>107</td>
<td>web</td>
</tr>
<tr>
<td>196</td>
<td>model</td>
<td>97</td>
<td>online</td>
</tr>
<tr>
<td>176</td>
<td>system</td>
<td>95</td>
<td>Facebook</td>
</tr>
<tr>
<td>157</td>
<td>impact</td>
<td>89</td>
<td>performance</td>
</tr>
<tr>
<td>155</td>
<td>media</td>
<td>88</td>
<td>algorithm</td>
</tr>
<tr>
<td>127</td>
<td>behavior</td>
<td>84</td>
<td>Community</td>
</tr>
</tbody>
</table>

**Document Co-Citation Analysis**

Co-citation analysis is an important method to reveal the basic topics of research field. Citation is a common phenomenon in the scientific research achievements of scientists. If two papers appear together in reference of the third article, they are co-citation. Documents with high co-citation frequency are usually valuable documents in this field. More than 110,000 citations were cited in this project. Figure 5 reflects the co-citation phenomenon in UGC field.

![Figure 5. Document co-citation analysis.](image3)
Through the analysis of the existing literatures by CiteSpace 5.3, the maps of co-authorship and co-institute analysis, co-occurring keywords analysis and document co-citation Analysis are obtained. The hotspots of UGC can be concluded into four points: Basic Concepts, Motivation, Application and Legal Issues of UGC.

Research Hotspots and Trends

Research Hotspots of UGC

**Basic Concepts.** In the 2005 Internet Industry Report, Mary Meeker introduced concept of UGC firstly which attracted extensive attention from scholars. Now, the definition of UGC in the report of OECD published in 2007 accepted by most of researchers. It figures that UGC has three key characteristics: (1) Information should be published on the Internet publicly; (2) Content may have some creative efforts; (3) UGC must be published by non-professional or authoritative persons [4].

UGC is regarded as a new form of media in modern society. Research on UGC actually is utilization of “Media”, the purpose is to connect different objects in the network environment [5]. Various forms of interactive information between user-user and user-media constitute strong relationship and weak relationship in UGC platform. Taking social media as an example, strong relationship refers to the interaction between users, such as attention, comments and so on. It often has high commercial value, such as the Bloggers with a large number of fans. However, the business value brought by weak relationship is generally low, such as browsing among users, content retrieval in Web pages, etc.

UGC analysis is also defined as a collection of methods and tools to extract value from user data [6]. Based on the value of content, UGC is located at the bottom of the pyramid. From bottom to top are Electronic Word-of-Mouth or Online Word-of-Mouth, Online Reviews, Online Recommendations [7]. The value of UGC is embodied in two aspects: User and Content. Through the research of users, we can mine the information of different users' preferences, characteristics, points of interest, and so on, so as to carry out user portrait or tagging processing. By analyzing the content, we can get the rules of topic distribution and topic dissemination. Users' preferences are reflected in the UGC created by them and their rating behavior to others. The value of UGC is reflected in its content and the ratings given by other users. ZHANG et al. proposed a personalized topic regression model (PTR), which combines hierarchical topic modeling and matrix decomposition, to provide an interpretable topic representation for users and UGC [8].

**Motivation.** (1) The division of motivation

In the early stage of UGC researching, many scholars have partitioned its generating motivation in detail. The basis of partitioning is mostly based on Flow Theory, Exchange Theory, Technology Acceptance Model (TAM), Maslow's hierarchy of needs.

Deci designed UGC motivation as a gradual spectrum, which was divided into external motivation and internal motivation from left to right [9]. UGC's creative motivation could be divided into social, psychological and technical perspectives. Consumers wanted to exchange opinions with other customers before making purchase decisions [10]. So the motivation of online reviews could be divided into two categories: helping others to buy products and ensuring that no one will buy products in the future [11].

At present, the classification of UGC generation motivation has been relatively perfect, but they are all based on experience, or through questioning and investigation. However, there is no research directly through the analysis of UGC to find out what kind of motivation it belongs to, which is also an interesting issue.

(2) Incentive methods

According users’ motivation, stakeholders can take some incentive measures to encourage users who would produce more high-quality UGC to share their opinion, complain and so on. Therefore, some scholars have done research on UGC incentive policy [12, 13].

The common incentive methods adopted to encourage users to produce UGC include single material incentive, virtual currency, give users a sense of mission, let users feel progress, let users
play their creativity and give feedback, create a sense of ownership, let users socialize in it, set up rare content, etc. Crowston found that altruistic consumers were more likely to create positive user generated content [14]. Therefore, marketers should study highly altruistic groups and implement incentives.

**Application.** (1) Pragmatic value of UGC

Comments analysis can help to solve the problem of cyberbullying and hate speeches [15], the risk of network crowdsourcing can be reduced too [16]. When disaster or accident occurs, supervisor could pay more attention to victim's microblog, especially track victim's negative reaction, and then find out the potential post-disaster events [17]. Question Answering in Knowledge Base is a research hotspot related to UGC. It can be regarded as a matching problem between questions and answers. This matching problem can be divided into two stages: candidate retrieval stage and answer selection stage [18].

In terms of tourism, UGC includes the views freely expressed by tourists. Therefore, it is considered as a source of information for National Tourism Organization (NTO), other decision makers, destination marketing organization (DMO) and potential tourists [19]. Shanshan collected on-line reviews of Macao from 2005 to 2013 in TripAdvisor, which included some tourists' cultural tourism experiences in Macao [20]. Based on such reviews, they identified the characteristics of tourists, divided the cultural tourism market, and forecasted tourists' preferences.

News communication is also one of applications of UGC. In many cases, users couldn’t describe the facts faithfully and comprehensively when producing UGC, which would leave a lot of space for imagination for network users. It is easy to make people prejudice or misunderstand the facts, mislead public opinion and confuse social audiovisual. UGC is produced by ordinary users. It has strong personal emotional tendency and is easy to be followed by other emotional users.

In e-commerce environment, information asymmetry is mainly reflected in product quality and communication between buyer and seller. Based on the convenience of digital payment, the lifestyle of consumers is also changed imperceptibly. Users communicate directly in the network platform, or publish UGC to provide purchase advice for others, which greatly reduces the information imbalance caused by space, time and other factors, and reduces the risk of users shopping.

(2) Quality Analysis of UGC

The quality of UGC can also be understood as the reliability of UGC. There are a lot of junk information or useless information in UGC, which seriously hinders the realization and utilization of UGC value. Therefore, it is interesting to mine the opinions of UGC, that is, to process, summarize, analyze and summarize the brief, discrete and decentralized UGC, so as to provide useful information for users. In addition, it can also judge the opinions or emotional tendencies that users want to express on specific matters, so as to provide important basis for other stakeholders.

On the view of big data, user's ability to process information couldn’t keep up with the development of the times because of mass data. Therefore, it is necessary to have a more intelligent way to help users find useful information without lost on spam information.

At present, spam review detection is mostly based on statistical data, so the problem of cold start appears. When the number of comments is too small, how to quickly identify whether it is false information? Wang X et al. believes that UGC published by network users is interrelated with their behavior information. The background information, motivation and interactive behavior style of commentators have a great influence on UGC and behavior information of commentators. They proposed that both of UGC and user behavior information would be embedded in the cold-start task, so that the neural network model could be used to detect junk comments in the cold-start problem without supervision [21]. To identify malicious Posts generated during news production, Dewan et al. deployed Facebook Inspector, which is different from the existing malicious content detection technology [22]. It does not depend on message similarity features, which were mainly used for detection activities in the past.

(3) Emotional Analysis of UGC

The classification of emotional information can be divided into subjective and objective emotion, which depends on whether the UGC contains emotional elements. If it does, then the strength of the
emotion should be studied and the emotional polarity of UGC would be calculated too [23, 24].

Now, many fields have been involved in emotional analysis research. It is of great practical
significance to judge users' emotional inclination by UGC, and more and more algorithms related to
emotional analysis are designed.

The emotional analysis of UGC can be summarized as extraction of emotional information,
classification of emotional information, retrieval and induction of emotional information. For a long
time, extraction of emotional information has been confronted with the problems of diversity,
ambiguity and structure in natural language processing (NLP), as well as the lack of a universal
emotional dictionary. Marc Egger and André Lang classified text-based UGC information
extraction into collection, analysis and visualization, which included such steps: data collection and
cleaning, document-level information extraction, sentence, phrase and word-level information
extraction, selection challenges and so on [25]. As the earliest English dictionary, emotional words
in General Inquirer (GI) Dictionary come from Harvard Dictionary and Rasville Dictionary. GI
dictionary contains 1914 commendatory words and 2293 derogatory words. Each emotional word is
labeled according to its polarity, intensity and part of speech. Senti WordNet Dictionary is an
extension of WordNet Dictionary. In addition to the emotional words and their corresponding
emotional extremes, there are also some additional contents such as synonyms and antonyms of
emotional words. To get more accurate emotional words and their corresponding extremes, Dragoni
et al. aggregates emotional extremes by integrating Senticnet, The General Inquirer Vocabulary and
MPQA [26]. Opinion Lexicon is an English Emotional Dictionary published by Bing Liu, which
has 6800 words of praise and derogation. Compared with English emotional dictionaries, Chinese
emotional dictionaries is slow-growing. Although Chinese natural language processing is a difficult
problem, there are some classic Chinese emotional dictionary which can be used in many fields, for
example, HowNet dictionary, DUTIR and NTUSD. DUTIR Emotional Vocabulary Ontology
Library is provided by Dalian University of Technology, it classifies Chinese Emotional Words into
7 categories and 21 sub-categories, and annotates emotional words with their lexical and emotional
intensity values. National Taiwan University Sentiment Dictionary (NTUSD) is published by
Taiwan University mainly expands the derogatory meanings. Dilip Raghupathi et al. proposed a
more accurate overall emotional rating algorithm to provide a comprehensive evaluation of product
reviews. Beginning with a text analysis, an impact language dictionary is used to evaluate the leaves
of a word tree, and a series of basic heuristics are used to calculate the backward overall emotional
rating for review. However, such algorithms pay more attention to the overall emotional tendency,
but ignore the role of individual UGC [27].

**Legal Issues.** Due to the difficulty of network management, and the randomness, anonymity and
massive data of UGC, users often ignore copyright of UGC which leads to infringe on the rights and
interests of others.

UGC may bring risks of disseminating pornography and violence, infringing other people's
portrait rights, privacy rights and intellectual property rights. Therefore, rules and regulations are
needed in the management of UGC and UGC platform. However, existing laws and regulations may
not be enough to deal with the widespread "dangers" of UGC platforms. George explored various
legal cases and the legal challenges posed by these "dangers" [28]. There are still some technical
difficulties in the supervision of UGC and UGC platform. For example, the real-name system can
prevent network fraud and false comments, but some people need "tree hole" to hide their identity.
Tree hole UGC is also used to cure depression. The identity information of patients is kept secret,
which is conducive to experts' psychological counseling.

Besides the restriction of laws and regulations, it is more important to cultivate the awareness of
copyright protection of all users and improve the ability of network autonomy [29]. In terms of
technology, technologies for junk identification and harmful information filtering should be
improved.
Conclusion and Future Works

The development of mobile Internet technology has enriched the dissemination and utilization of UGC, which prompted users to continue to play their initiative. According to the existing discussion on UGC, follow-up research can be carried out in the following aspects:

- **Design of Incremental Mining Algorithm**
  Data stream is dynamic data reaching in real time. Data stream mining is one of the hotspots in the field of data mining at present. In order to satisfy the timeliness and efficiency of information, incremental UGC classification and other algorithms can be considered.

- **Improving the Emotional Lexicon**
  The method based on dictionary and lexicon matching is a common method in Chinese word segmentation, but there is no universal emotion lexicon that can be applied to UGC emotional analysis. In the field of NLP, how to extract Chinese semantics accurately and correctly has always been a problem.

- **Dynamic Evolution of UGC Emotion**
  Because of the real-time fluidity of data, UGC emotions will also change with time. How to deal with the dynamic evolution of UGC emotional analysis is also attracting. In the future research, we will try to construct the emotional change curve for a single user and for a topic.

Acknowledgement

This research was financially supported by Anhui Provincial Natural Science Foundation under No. 1808085MF194; the Natural Science Foundation of Anhui Education Department of China under Grant No. KJ2019A0663.

References


