Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 46 (2025), No. 5, pp. 1679–1689
DOI : 10.47974/JIOS-1938

An automated approach for temporal specificity classification of user-generated content

Yanni Yang [§]
School of Literature and Media
China Three Gorges University
Yichang 443000
Hubei
China

and

School of Information Management
Central China Normal University
Wuhan 430079
Hubei
China

Jiawei Zhu [†]
School of Literature and Media
China Three Gorges University
Yichang 443000
Hubei
China

Qian Wu *
School of Information Management
Central China Normal University
Wuhan 430079
Hubei
China

[§] *E-mail:* `yangyanni@ctgu.edu.cn`
[†] *E-mail:* `zhujiawei0522@163.com`
* *E-mail:* `wuqian@mails.ccnu.edu.cn (Corresponding Author)`

Abstract

The temporal specificity of user-generated content (UGC) affects the information recommendation accuracy in online question answering (Q&A) community. Exploring the method for the temporal specificity classification of UGC is beneficial to optimize the content organization of online Q&A communities. In assessments of temporal specificity of UGC, it is relatively one-sided to judge only by the post time of the content. The relationship between the content's topic category and temporal specificity is often ignored. To solve these problems, an automated approach for assessing the temporal specificity of user-generated content is proposed. The temporal specificity of UGC in the online Q&A community was divided into three types: high temporal specificity, medium temporal specificity, and low temporal specificity. The temporal specificity of UGC is automatically classified based on the Word2Vec-XGBoost algorithm, and question text for different topics from the Zhihu Q&A community is collected for experimental verification. The results show that the classification's accuracy, recall, and F1 score with weighted temporal specificity are increased by 2.63%, 2.67%, and 2.69%, respectively, compared to those in the base case. The overall accuracy, recall, and F1 score reached 90.30%, 90.24%, and 90.12%, respectively, and the temporal specificity classification performance was good.

Subject Classification: *Primary 68T50, Secondary 03B65.*

Keywords: *Online question answering (Q&A) community, User-generated content (UGC), Temporal specificity, Text classification, Machine learning.*

1. Introduction

There is a large amount of complex user-generated content (UGC) in the online question answering (Q&A) community, and it is generally difficult for users to accurately retrieve content that meets the temporal specificity needs of the user. If you search for "today's gold price" in an online Q&A community on September 6, 2021, and sort the question discussion based on time, the top three answers are from August 31, September 2 and September 3, 2021. The results of this sorting approach are returned based on the amount of attention received or the degree of keyword matches, rather than in a temporal series. The temporal specificity of user-generated content (UGC) affects the information recommendation accuracy in online Q&A communities and is an important factor affecting the quality of UGC. If the online Q&A community can focus on the temporal specificity of UGC, it can improve the efficiency of user information retrieval and enhance the user's perception of the value of UGC.

The temporal specificity of UGC is generally assessed with a time tag (such as the post time). However, it is relatively rough to evaluate the temporal specificity only by the time tag, which easily causes misjudgment of the temporal specificity of UGC. In this regard, research on the temporal specificity of literature can provide a new assessment perspective, and the relationship between temporal specificity and the subject of literature is emphasized [1]. The temporal specificity of content is not only related to time but also affected by the topic feature of content. Therefore, to improve the accuracy of assessments of the temporal specificity of UGC, it is necessary to consider the topic feature of UGC.

Additionally, the amount of UGC in an online Q&A community can be enormous. For example, Zhihu is a Chinese online Q&A community that was founded on January 26, 2011, with a product form similar to that of an American website of the same type, Quora. The Zhihu Q&A community has accumulated more than 44 million questions and 240 million answers from users [2]. It is difficult and inefficient to manually label the temporal specificity of UGC. Therefore, the automatic classification of the temporal specificity of UGC in online Q&A communities is particularly important. To improve the efficiency of the classification, an automatic evaluation and classification approach for the temporal specificity of UGC is proposed in this paper based on the topic categories of UGC. The main contributions of our work include the following points. Besides, the factors that affect the temporal specificity of UGC are studied. Different from other studies, it focuses on the topic features of UGC, emphasizing the difference in the temporal specificity of UGC of various topics. Moreover, considering topic features, it constructed the temporal specificity vocabulary of UGC in an online Q&A community.

Based on the above analysis, this study formulated corresponding strategies of information organization for online Q&A communities. It is not only conducive to the online Q&A community to automatically label the temporal specificity of UGC, but also, based on this, optimizes the content recommendation of the online Q&A community and enhances the user's perception of the value of the recommended content.

2. Methods

The temporal specificity of UGC in an online Q&A community refers to the attribute characteristics of users' perception of the value of UGC within a specific time. The temporal specificity of UGC is closely related to the topic feature of content, and the decay period of different topics can significantly differ. Therefore, automatic temporal specificity classification is conducted based on the content topic types and temporal specificity keywords. In this work, we first sort the temporal specificity of UGC in an online Q&A community, which is the focus of this research. Second, the temporal specificity of UGC is judged based on the Word2Vec-XGBoost algorithm to realize the automatic temporal specificity classification of UGC. Finally, data from Zhihu Q&A community were collected for experiment.

2.1 *Types of Temporal specificity of UGC*

The value of UGC in the online Q&A community gradually decays over time. In the research on the temporal specificity of news information and geographic information, the rule of the third is commonly used to classify the types of information temporal specificity [3, 4]; the advantage of this approach is that it can avoid the increased difficulty of differentiation and inaccurate results caused by too fine division of temporal specificity. Users in an online Q&A community pay less attention to information that expires within a few hours or even minutes. Therefore, when classifying the temporal specificity of UGC in this paper, days and months are used as time intervals, and the temporal specificity is classified by the rule of the third according to the strength of temporal specificity of UGC. The types of temporal specificity of UGC in the online Q&A community include high temporal specificity, medium temporal specificity, and low temporal specificity.

High temporal specificity of UGC refers to the UGC whose value is greatly affected over time and is prone to rapid decay in a short period of time. This type of content is generally valid over a period of a few days and generally involves current affairs or real-time discussions. For example, news, stock market quotes, event analysis (short-term focus) and other information fall within this category.

Medium temporal specificity of UGC refers to the UGC whose value slowly decays over time but is relatively stable for a certain period. The period of validity of this type is generally a few days to a few months, and the rate of value decay is slower than that for UGC with high temporal specificity. The UGC with medium temporal specificity includes the contents about experience sharing and product or service recommendations or reviews. The content value will be attenuated due to related product iterations, service changes, etc. For example, the purchase experience of certain products may no longer be useful when the product is removed from the shelves; another example, the sharing of personal travel experiences, may be useful for a period of time, but these experiences may no longer be valuable after a period of time.

Low temporal specificity of UGC refers to UGC whose value is relatively stable and decays slowly over a relatively long period of time. The content value is generally valid for more than a few months or even a few years. UGC with low temporal specificity has the characteristics of high knowledge density and strong universality. For example, knowledge popularization and literary comments fall within this category.

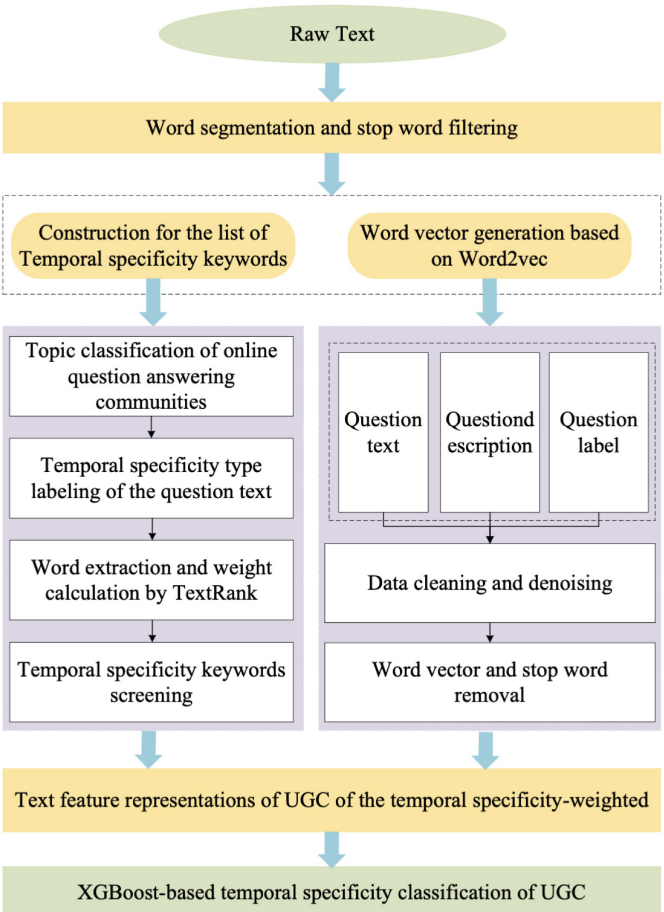

Figure 1

Temporal specificity classification approach based on Word2Vec-XGBoost

2.2 *Temporal specificity Classification Approach*

Based on supervised machine learning, the proposed approach of classifying the temporal specificity of UGC in an online Q&A community is shown in Figure 1. Based on the preprocessing of the text data, a list of temporal specificity keywords of UGC is constructed. The Word2Vec algorithm is applied to word vector representations of the temporal specificity keyword-weighted, and the XGBoost algorithm is used for the temporal specificity classification of UGC. The text data of UGC in an online Q&A community, including the content of finance, science, sports, campus, and other topics, are randomly scraped by Python. Half of the data are used for the construction of the list of temporal specificity keywords, and the other half is used for machine learning and verification according to the 80/20 principle.

2.3 *Data Validation*

By the end of 2020, the cumulative number of users of the Zhihu platform reached 370 million, and the number of monthly active users was 75.7 million. As an online Q&A community with the largest user scale in China, Zhihu provides a good environment for users to share knowledge and express their views. To verify the effectiveness of the temporal specificity classification method proposed in this paper, the UGC data in the Zhihu Q&A community was collected as the experimental data. The details of data are reported in the Supplementary Material. To ensure the reliability of the data, we randomly obtained question data for twelve topics, including finance, science, sports, and academics. There was an average of more than 1,600 pieces of data for each topic, totaling 20,000; the text data included question titles, detailed descriptions of questions, and question labels. After filtering and preprocessing the data with short titles and no problem descriptions and removing low-quality and redundant data, 17909 pieces of data were obtained, of which 9003 were used for the construction of the list of temporal specificity keywords of UGC and 7085 were used for machine learning training. The remaining 1821 pieces of data were used for testing.

3. Results

3.1 *Results of Temporal specificity Labeling*

According to the proposed temporal specificity types, the data were manually labeled. The basis for labeling is the validity period of UGC in the online Q&A community, that is, how long the content is still valuable after it is published. The value of the UGC, within a few days, a few days to a few months, and more than a few months or even a few years, respectively, labeled as high, medium and low temporal specificity of UGC.

After normalizing the data, the results showed that 31.63%, 9.11% and 59.26% of the data were classified in the high, medium and low temporal specificity of UGC, respectively. In the topics of law and sports, the UGC of high temporal specificity is relatively large; in the topics such as comics and cartoons, science, and film and television, the UGC of low temporal specificity accounts for a relatively large proportion; in topics such as digital products or examination and certification, the UGC of medium temporal specificity occupied a large proportion.

The main reasons for the above results are twofold. one is the attributes of the community. The UGC of low temporal specificity is more prevalent in online Q&A communities, which differs from real-time content sharing communities (e.g., Twitter or Facebook). The other is topics' properties. UGC for topics such as law and sports mainly involves news and the latest information. Moreover, the topics of science and fashion are mostly associated with knowledge content with a long time decay period, and the topics of digital products or examination and certification mainly involve experience sharing and evaluation-based recommendations.

3.2 *Results of Construction for the list of temporal specificity feature keywords*

According to the topic category in the Zhihu community, typical temporal specificity keywords were extracted for different topics with TextRank. Keywords with no temporal specificity were manually eliminated, and a word table of temporal specificity was constructed.

Temporal specificity keywords have strong topic features, and the keywords for each type of temporal specificity of UGC exhibit significant differences. UGC with high temporal specificity is mostly related to topical issues or real-time discussion. The keyword extraction results show that keywords with marked temporal features (e.g., "the afternoon of February 16"), implicit temporal features (e.g., "Teacher's Day") and the related temporal features (e.g., "recently") are typical keywords of high temporal specificity. In addition, the keywords that characterize the high temporal specificity of different topics have significant thematic features, such as "case" and "facts" for legal topics, "trends" and "opening" for financial topics, and “press conferences” and “initial release” for digital products topics.

The UGC of medium temporal specificity is mostly experience sharing and recommendation evaluation, including subjective experiences and feelings regarding products or services. The results of keyword extraction show that the keywords representing medium temporal specificity of different topics are relatively similar. For example, “experience”, “recommendation”, “feeling” and “evaluation” appear frequently in the UGC of medium temporal specificity for each topic. However, there are few keywords of medium temporal specificity that characterize thematic features in different topics.

The UGC of low temporal specificity involves teaching, the interpretation of phenomena, knowledge popularization, etc. Among them, similar keywords that characterize low temporal specificity often appear in various topics, such as “skills”, “techniques”, and “experiences”. Additionally, the keywords that characterize the low temporal specificity of different topics have significant thematic features, including “Trivia” and “core technology” for the science topic, “dress collocation” and “styling” for the fashion topic, and “operation” and “novice” for the game topic.

3.3 *Results of Temporal specificity Classification*

(1) Classification results based on Skip-Gram and CBOW

Analyze the difference between the Skip-Gram model and the CBOW model in the Word2Vec algorithm on the impact of temporal specificity classification results, and compare the three indicators of precision, recall and F1 score. The results are shown in Table 1 below. The results indicate that the classification based on Skip-Gram is more accurate. Notably, the

Table 1

Text temporal specificity classification results based on Skip-Gram and CBOW

Type	Skip-Gram			CBOW		
	Accuracy	Recall	F1	Accuracy	Recall	F1
High	0.8910	0.8780	0.8845	0.8020	0.8750	0.8369
Medium	0.8860	0.7368	0.8045	0.7727	0.5483	0.6415
Low	0.8595	0.9327	0.8946	0.9032	0.9105	0.9068
Overall	0.8767	0.8757	0.8743	0.8497	0.8512	0.8474

Skip-Gram model uses the available vocabulary to predict the surrounding words, which is more in line with the characteristics of the question text. However, the principle of the CBOW model is to predict the current word based on the surrounding vocabulary, and the trained word vector will be highly accurate only when the vocabulary is relatively concentrated. The accuracy of the CBOW model decreases because the question text in this research is generally independent and nonconcentrated.

The experimental results indicate that the accuracy, recall, and F1 score obtained for the Skip-Gram model are 87.67%, 87.57%, and 87.43%, respectively. Among them, the classification effect for the content text of low temporal specificity is the best, followed by that for the content text of high temporal specificity. The recall rate for the content text of medium temporal specificity is significantly lower than the rates for other content text. Because the sample size of content text of medium temporal specificity is smallest, the number of training samples for machine learning may be insufficient. Moreover, the content text of medium temporal specificity generally contains expressions similar to those in other types of text, which may result in possible misjudgments during machine learning.

(2) Classification results with temporal specificity weighting

Based on the comparison of the classification results of Skip-Gram and CBOW, the Skip-Gram model was used to generate word vectors. Additionally, the constructed temporal specificity vocabulary was combined to weight the word vectors, and the experimental results obtained with weighted and unweighted temporal specificity were compared, as shown in Table 2.

Table 2

Comparative experimental results for weighted and unweighted temporal specificity

Type	Unweighted			Weighted		
	Accuracy	Recall	F1	Accuracy	Recall	F1
High	0.8910	0.8780	0.8845	0.8906	0.9193	0.9047
Medium	0.8860	0.7368	0.8045	0.9183	0.7826	0.8450
Low	0.8595	0.9327	0.8946	0.9055	0.9504	0.9274
Overall	0.8767	0.8757	0.8743	0.9030	0.9024	0.9012

Experiments indicated that the overall classification effect was significantly improved when temporal specificity were weighted. Compared with temporal specificity classification without considering weighted temporal specificity, the accuracy, recall and F1 score of classification with the weighting of temporal specificity increased by 2.63%, 2.67%, and 2.69%, respectively. Notably, the overall accuracy, recall, and F1 score reached 90.30%, 90.24%, and 90.12%, respectively. Specifically, the recall and F1 score for the content text of high temporal specificity increased by 4.13% and 2.02%, respectively. The accuracy, recall, and F1 score for the content text of medium temporal specificity all significantly improved by 3.23%, 4.56%, and 4.05%, respectively. Moreover, the accuracy for the content text of low temporal specificity increased by 4.6%, and the recall and F1 score increased by 1.77% and 3.28%, respectively. These results verify the effectiveness of the proposed method of classifying the temporal specificity of UGC in an online Q&A community. It is important to consider temporal specificity for the automatic realization of temporal specificity classification of UGC.

4. Conclusion

In assessments of temporal specificity of UGC, it is relatively one-sided to judge only by the post time of the content. The relationship between the content's topic category and temporal specificity of the content is often ignored. The proposed temporal specificity classification approach of UGC in an online Q&A community can effectively solve these problems. The temporal specificity of UGC is divided into high temporal specificity, medium temporal specificity, and low temporal specificity. Additionally, the topic features of UGC are considered to construct the temporal specificity vocabulary, and the Word2Vec-XGBoost algorithm is used to realize the temporal specificity classification of UGC. The question text, as a typical UGC in the Zhihu Q&A community, was collected as the experimental data for method verification. The results highlight the effectiveness of the temporal specificity classification method for UGC developed in this study. In addition, the results suggest that it is necessary to consider topic category and temporal specificity in classification.

The temporal specificity classification of UGC in online Q&A communities is beneficial for the full identification and utilization of time-sensitive information. On the one hand, this approach avoids outdated content of high temporal specificity being recommended, and on the other hand, it prevents content of low temporal specificity from being overlooked due to its comparatively early publication time, thus ensuring reasonable recommendations for different temporal specificity content, optimizing the content recommendation pool and improving the recommendation quality and user experience in online Q&A communities. Moreover, the automation of content temporal specificity classification reduces the cost of manual labeling, improves the efficiency of content labeling, and optimizes the content organization of online Q&A communities.

The proposed temporal specificity classification method for UGC in an online Q&A community is universal and can be applied to assess content temporal specificity classification in related fields; however, the technical solutions may need to be further expanded. To ensure the accuracy of the experiment, a supervised learning method was adopted, and to reduce the difficulty of manual labeling, only 20,000 pieces of data were collected, which may affect the accuracy of the results. To further expand the research, follow-up research will be performed from the following perspectives. The temporal specificity of UGC includes not only diverse categories but also refined evaluations that can be performed with more indicators. For example, the temporal specificity of UGC is a dynamically changing variable, and content temporal specificity can be further evaluated from different dimensions. Furthermore it is necessary to further improve the accuracy of the automatic recognition of the content temporal specificity of medium temporal specificity. Finally, the application field can be expanded, and the model can be applied for temporal specificity evaluations of scientific and technological literature, books and other academic resources to further verify the effectiveness of the method.

References

[1] X. Tong, C. Cheng, R. Wang, L. Ding, Y. Zhang, G. Lai, and B. Chen, "An efficient integer coding index algorithm for multi-scale time information management," *Data & Knowledge Engineering*, vol. 119, pp. 123–138 (2019).

[2] China Economic Network, "2021 Xinzhi youth conference opens, zhihu will continue to increase support for creators," *China Economic Network*, Jan. 14 (2021). [Online]. Available: http://www.ce.cn/xwzx/gnsz/gdxw/202101/14/t20210114_36222648.shtml.

[3] S. U. R. Khan, M. A. Islam, M. Aleem, M. A. Iqbal, and U. Ahmed, "Section-based focus time estimation of news articles," *IEEE Access*, vol. 6, pp. 75452–75460 (2018).

[4] L. Sun, C. Zhang, C. Guo, and B. Ge, "Method of organizing spatial-temporal geographic metadata," *J. Computer Applications*, vol. 38, pp. 152–156 (2018).

Received August, 2024