Research on Influencing Factors of Tag Quality in Social Tagging Systems: A Postprint Based on Random Forest Algorithm

Authors: Zhang Yunzhong, Qin Yiyuan

Date: 2023-07-26T00:00:00+00:00

Abstract

[目的/意义] In social tagging systems, tag quality often affects users' experiences in classifying, querying, browsing, and retrieving web resources, and identifying the key factors influencing tag quality helps further optimize the core resource organization functions of social tagging systems. [方法/过程] Taking tags in social tagging systems as the research object, this study reconstructs a tag quality influence factor model from dimensions including tagging subject, tagging object, tagging environment, tagging motivation, tagging method, and tagging product, attempts to explore the key factors influencing social tag quality, collects data through questionnaire surveys, and establishes a decision tree model for tag quality influence factors by combining supervised learning random forest algorithms. [结果/结论] Results show that the tagging subject is the primary key dimension affecting tag quality, with the subject's knowledge structure and cognitive level, tagging frequency, and perceived usefulness having prominent impacts on tag quality; the tagging method is the secondary key dimension affecting tag quality, with tag recommendations and standardized tag prompts being important factors influencing tag quality.

Full Text

Preamble

Research on Influencing Factors of Tag Quality in Social Tagging Systems: Based on Random Forest Algorithm

Zhang Yunzhong, Qin Yiyuan
Department of Library, Information and Archives, Shanghai University, Shanghai 200444

Abstract

[**Purpose/Significance**] In social tagging systems, tag quality directly affects users' experience in classifying, querying, browsing, and acquiring online resources. Identifying the key factors influencing tag quality is essential for optimizing the core resource organization functions of social tagging systems. [**Method/Process**] This study examines tags in social tagging systems and reconstructs a tag quality influence factor model from six dimensions: tagging subject, tagging object, tagging environment, tagging motivation, tagging method, and tagging product. We explore the key factors affecting social tag quality through a questionnaire survey and construct a decision tree model using supervised learning with the random forest algorithm. [**Result/Conclusion**] Results demonstrate that the tagging subject is the primary dimension affecting tag quality, with the subject's knowledge structure, cognitive level, tagging frequency, and perceived usefulness having prominent effects. Tagging method is the secondary key dimension, where tag recommendation and standard tag prompts emerge as important factors influencing tag quality.

Keywords: social tagging system, tag quality, machine learning, random forest

1. Introduction

Social Tagging Systems (STS), emerging as novel network information resource organization and management systems in the Web 2.0 era, allow users to freely describe and annotate online resources. The resulting tags prove highly effective for organizing and indexing web resources. STS represent platforms where users, based on self-cognition, freely attach tags to achieve description, classification, and navigation of network resources. However, the quality of user-generated tags varies significantly.

Tag quality measures how precisely tags can describe resources to facilitate user classification, querying, browsing, acquisition, and utilization. It is closely related to factors including users, network resources, tagging environment, tagging motivation, tagging methods, and tagging products. With the dramatic increase in tag quantity across current social tagging systems, quality remains inconsistent, severely impacting user experience in resource classification, querying, browsing, and sharing. While numerous scholars have approached this issue through tag evaluation and recommendation to suggest high-quality tags, research on reducing low-quality tags and understanding their causes remains scarce. Since tag quality varies, identifying its influencing factors and determining their relative importance is crucial for fundamental quality improvement.

This study addresses the weight determination problem in tag quality assessment by combining weight coefficients with existing statistical indicator systems. By calculating each factor's influence weight on tag quality, we can not only improve tag quality assessment models but also enable social tagging systems to implement effective measures for enhancing tag quality and promoting network information resource organization. This holds significant reference value for

perfecting user-centered social tagging system functions in the Web 2.0 environment.

2. Literature Review

Currently, most scholars do not treat tag quality influencing factors as an independent research topic but rather as one component of tag quality assessment studies. Two perspectives emerge: multi-dimensional comprehensive influencing factors and single-dimensional key influencing factors.

(1) Multi-dimensional Comprehensive Influencing Factors. This view holds that tag quality results from combined effects across multiple dimensions including tagging subject, environment, motivation, and method. The Zhang Chengzhi team exemplifies this approach, arguing that subjects' excessive freedom and subjectivity constitute primary causes of low tag quality. Tagging environmental factors such as incomplete system functional mechanisms also affect quality, while tagging method factors like standard tag prompts and spelling suggestions can reduce ambiguity, synonymy, and misspelling. Different tagging motivations further lead to quality variations.

(2) Single-dimensional Key Influencing Factors. This perspective emphasizes particular critical factors, most commonly "tagging method." Research on WorldCat, Flickr, Bibsonomy, and Douban indicates that quality can be improved through standard tag prompts, spelling suggestions, error detection mechanisms, and tagging guidelines. Similar studies on Flickr add that managing popular tags enhances quality. N. Sogol et al. propose that tag recommendation improves quality, while C. Hall suggests using controlled vocabularies for tag recommendation. M. Guy also emphasizes the importance of input prompts, spell-checking, and tag recommendation.

Some scholars consider "tagging environment" as key, particularly system interface design. F. Floeck et al. demonstrate that social tagging system interface design affects tag quality, while S. Sen et al. argue that interface improvements can enhance quality. Others emphasize "tagging subject" as critical, with Zhu Qinghua suggesting that user scale, structure, and tagging frequency affect retrieval quality, and Luo Lin's empirical research showing that information quality, perceived usefulness, and perceived ease-of-use positively influence tagging intention.

While scholars have reached consensus on potential influencing factors, current research focuses on identifying relevant factors rather than systematically exploring their general weight relationships with tag quality—precisely the gap this study addresses.

3. Research Methods

This study aims to determine the influence weights of multiple factors on tag quality, essentially a weight determination problem. Common weighting meth-

ods include principal component analysis, multiple regression analysis, analytic hierarchy process (AHP), and Support Vector Machine (SVM)-based machine learning.

Principal component analysis suits dimensionality reduction in multi-variable problems but requires transforming numerous factors into principal components, obscuring individual factor weights. Multiple regression analysis measures factor significance through significance levels but assumes strict conditions requiring all explanatory variables, otherwise risking spurious regression. AHP suits multi-scheme optimization but becomes problematic with many indicators as judgment matrix order increases, making assignment difficult and reducing precision. While theoretically sound, SVM primarily suits binary classification with poorer performance on multi-class problems.

Considering these limitations, this study employs the random forest machine learning algorithm to establish a predictive model between influence factor features and tag quality, then classifies and predicts using the classifier. Random forest offers several advantages: (1) Unlike principal component analysis, it directly obtains factor weights through objective training without dimensionality reduction; (2) Unlike regression, it doesn't require exhaustive explanatory variables and resists overfitting through sample and feature randomness, avoiding spurious regression; (3) Unlike AHP, it handles large samples, resists noise, and yields more reliable, precise influence factors; (4) Unlike SVM, the resulting decision tree model is interpretable, establishing relationships between factors and tag quality through intuitive if-then rules.

4. Tag Quality Influence Factor Model

Based on existing research and extensive literature review, interviews with researchers and social tagging system users, we propose a tag quality influence factor model measuring impacts from six dimensions: tagging subject, tagging object, tagging environment, tagging motivation, tagging method, and tagging product (see Figure 1 [Figure 1: see original paper]).

(1) Tagging Subject refers to taggers themselves. Factors including disciplinary background, knowledge structure, cognitive level, tagging frequency, interest preferences, and emotional state during tagging inevitably affect tag quality.

(2) Tagging Object refers to resources being tagged. The quantity, quality, and type of online resources awaiting tagging also influence tag quality.

(3) Tagging Environment primarily concerns the social tagging system platforms where tagging occurs. Functional completeness, performance superiority, and platform stability all impact tag quality.

(4) Tagging Motivation refers to users' driving forces for tagging, generally covering seven types: revealing resource themes/classification/attributes (what it's about), describing resource carriers/types (what it is), identifying

owners, refining existing tags, describing resource characteristics, self-reference, and task/personal resource management.

(5) Tagging Method includes free tagging and intervention-based tagging. Five main intervention mechanisms exist: tag recommendation, standard tag prompts, spelling suggestions, tagging guidelines, and error detection.

(6) Tagging Product refers to the tags themselves. Tag quality directly relates to their form; high proportions of ambiguous, synonymous, or misspelled tags directly indicate low quality.

5. Empirical Study on Social Tag Quality Influencing Factors

5.1 Questionnaire Design

Following general questionnaire design principles and our model hypotheses, we structured the questionnaire into three sections (see Table 1):

(1) Basic Information Section. While not directly part of our influence factor model, this information might affect prediction results. It includes six questions (Q1-Q6) covering gender, age, education level, professional background, resource ownership, and system usage duration.

(2) Influence Factor Feature Section. This comprises 29 questions (Q7-Q35) on tag quality influencing factors across six dimensions, using a five-point Likert scale to measure user acceptance levels from "strongly disagree" to "strongly agree."

(3) Target Feature Section. This includes one question (Q36) representing five tag quality levels, where users indicate their acceptance of the statement: "In social tagging systems, 'tag quality' is crucial for measuring whether tags can accurately describe resources to facilitate user classification, querying, browsing, acquisition, and utilization."

5.2 Data Collection and Validation

To ensure scientific rigor and accuracy, we targeted both users with information organization backgrounds and ordinary users through two channels: (1) Email distribution of electronic questionnaires to teachers (76) and students (113) in information organization research at universities and institutions; (2) Online distribution via shared links and QR codes to users of Douban (books: 18, movies: 27, music: 31), Flickr (17), blogs (44), Diigo (19), Pinterest (28), and Haowangjiao (19). Over two months, 523 questionnaires were distributed, yielding 429 responses with 392 valid questionnaires (82% recovery rate).

The sample comprised 136 males and 256 females, predominantly aged 21-30 (277 people), the main user demographic. Education levels concentrated on bachelor's (137) and master's degrees (172), with fewer 专科/high school or below

(36) and doctoral degrees (47). The ratio of ordinary to professional users was 203:189 (approximately 1:1). Resource publishers to non-publishers ratio was 164:228. 86% had used social tagging systems like Douban, indicating broad familiarity.

We imported the 392 questionnaires into SPSS 21.0 for reliability and validity testing. Cronbach's Alpha coefficient was 0.931 (>0.8), indicating excellent scale reliability. Structural validity was tested using KMO and Bartlett's test, yielding KMO = 0.912 (>0.7) and Bartlett's sphericity test ^{2} = 5661.566 (df = 435, sig = 0.000), confirming statistically significant and reasonable scale structure.

5.3 Data Preprocessing

Survey data existed as text requiring quantification before machine learning model training and testing.

5.3.1 Quantification of Ordinal Features

For Likert-scale questions Q2 and Q7-Q36, the five options exhibit clear ordinal relationships reflecting user acceptance levels. These ordinal features were assigned integer values (1-5 for Q7-Q36; 1-4 for Q2). Since machine learning is sensitive to scale differences, Z-score standardization unified the feature spaces using the transformation:

$$X^* = \frac{X - \mu}{\sigma}$$

where X is the unnormalized feature, μ is the mean, σ is the standard deviation, and X^* is the normalized feature (0-1 real numbers preserving ordinal information).

5.3.2 Quantification of Nominal Features

For nominal questions Q1 (gender), Q3 (education), Q4 (background), Q5 (resource ownership), and Q6 (STS usage duration) without inherent ordering, we applied one-hot encoding. For example, Q1 was encoded as male = 10, female = 01.

5.4 Model Training and Evaluation

We employed random forest to build a decision tree model, training on all 392 questionnaires and evaluating via leave-one-out cross-validation, achieving 0.1475 error rate and 85.25% prediction accuracy. Each sample included 35 influence factor features (Q1-Q35) and one target variable (Q36: tag quality), making this a five-class supervised learning problem.

5.4.1 Information Gain and Influence Factors

Random forest measures factor influence weights through information gain,

which quantifies how much a feature contributes to classification. Higher information gain indicates greater importance. Information gain represents the reduction in dataset uncertainty after introducing feature a, calculated as:

$$g(D|a) = H(D) - H(D|a)$$

where $H(D)$ is the entropy of sample D:

$$H(D) = -\sum_{k} p_k \log_2 p_k$$

and $H(D|a)$ is the conditional entropy:

$$H(D|a) = -\sum_{n} \frac{|D_i|}{|D|} H(D_i)$$

5.4.2 Threshold Determination

After ranking features by information gain, decision tree classification rules require threshold selection for decision nodes. For discrete features, we exhaustively test all possible values as thresholds, selecting those maximizing information gain. For instance, Q8 (the most influential feature) has an optimal threshold of 4.53.

5.4.3 Model Training and Assessment

With feature ranking and thresholds determined, we constructed random forest decision trees. Using leave-one-out validation across 100 random trials (each using one sample as test set, 391 as training set), we achieved 92% accuracy on test data, demonstrating strong predictive performance.

Figure 2 [Figure 2: see original paper] illustrates one sample's prediction process. Rectangles represent decision nodes, circles chance nodes, and triangles leaf nodes. The model first evaluates Q8 (threshold 4.53), then Q27 (threshold 3.82), sequentially applying the most information-gaining features until reaching a leaf node, which predicts the tag quality level (Q36 = 3 in this example).

5.4.4 Influence Factor Analysis

During training, random forest calculated each factor's information gain (influence factor). Figure 3 [Figure 3: see original paper] shows feature-level influence factors (by question number), while Figure 4 [Figure 4: see original paper] displays dimension-level influence factors via PCA.

The six dimensions rank by influence weight: tagging subject, tagging method, tagging motivation, tagging product, tagging object, and tagging environment.

Tagging subject is the most influential dimension. Within it, subject knowledge structure/cognitive level (Q8), tagging frequency (Q9), and perceived usefulness (Q13) rank 1st, 4th, and 5th among individual factors, all significantly impacting quality.

Tagging method ranks second. Tag recommendation (Q27) and standard tag prompts (Q28) rank 2nd and 3rd among individual factors, also substantially affecting quality.

Tagging motivation ranks third, slightly below tagging method. Revealing resource themes/classification/attributes (Q20) shows the strongest motivational impact, ranking 6th among individual factors.

While the remaining three dimensions (tagging product, object, environment) have relatively lower overall influence, specific factors within them significantly affect quality: misspelling (Q34) in tagging product, resource type (Q14) in tagging object, and system functional completeness (Q17) in tagging environment.

6. Conclusion and Outlook

This study investigated tag quality influencing factors in social tagging systems, employed questionnaire surveys to capture user perceptions, and used random forest machine learning to construct a decision tree model yielding factor influence weights. Key findings: tagging subject is the primary dimension, with knowledge structure, cognitive level, tagging frequency, and perceived usefulness exerting prominent effects; tagging method is secondary, with tag recommendation and standard tag prompts as important factors.

Compared to existing research, this study offers four contributions: (1) More universal focus on commonalities across different social tagging systems, yielding more generalizable conclusions; (2) More comprehensive factor coverage across six dimensions, providing a more robust theoretical framework; (3) More scientific weight calculation using objective random forest training, reducing subjective judgment; (4) More targeted improvement strategies based on key influencing factors.

The resulting predictive model is highly interpretable, establishing explicit expressions between influence factors and tag quality, and offering a novel multi-attribute weighting method. These weights clarify causes of low-quality tags and provide references for improvement. Current tag quality assessment relies on statistical indicators (resource views, recommendations, comments) with unknown weights; our influence factors can serve as reference weights to enhance assessment systems—this represents a future research direction.

References

[1] Tai Yangfang, Chen Xinguo. User Tagging Behavior and Differences in Social Tagging Systems[J]. Library, 2017(10): 42-49, 61.

[2] Xiong Huixiang, Yang Xueping. Research on Personalized Information Recommendation in Social Tagging Systems[J]. Journal of the China Society for Scientific and Technical Information, 2016, 35(5): 549-560.

[3] Li Xuhui, Li Yuanyuan, Ma Feicheng. Main Issues in Social Tagging Research in China's Library and Information Science Field[J]. Library and Information Service, 2018, 62(16): 120-131.

[4] Zhang Chengzhi, Li Lei. Research on Automatic Quality Assessment of Social Tags[J]. New Technology of Library and Information Service, 2015(10): 2-12.

[5] Zhang Chengzhi, Zhao Hua, Li Lei, et al. Comparative Study on Quality Differences Between Chinese and English Image Tags: A Case Study of Flickr[J]. Information Studies: Theory & Application, 2018, 41(4): 123-127.

[6] Huang Ruhua, Ren Qixiang. Investigation and Analysis of Popular Tags in WorldCat[J]. Library and Information, 2012(5): 7-10.

[7] Wu Fangzhi. Quality Control Strategies for User Tags on Flickr[J]. Research on Library Science, 2012(11): 26-28.

[8] Sogol N, Arash B, Chen D. An Improved Collaborative Recommendation System by Integration of Social Tagging Data[EB/OL]. [2018-08-01]. https://doi.org/10.1007/978-3-319-14379-8_7.

[9] Hall C E, Zarro M A. What Do You Call It?: A Comparison of Library-created and User-created Tags[C]//Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries. New York: ACM, 2011: 53-56.

[10] Guy M, Tonkin E. Folksonomies: Tidying Up Tags?[EB/OL]. [2018-08-02]. http://www.dlib.org/dlib/january06/guy/01guy.html#1.

[11] Floeck F, Putzke J, Steinfels S, et al. Imitation and Quality of Tags in Social Bookmarking Systems[C]//Advances in Intelligent and Soft Computing. Berlin: Springer-Verlag, 2010: 75-91.

[12] Sen S, Harper F M, Lapitz A, et al. The Quest for Quality Tags[EB/OL]. [2018-08-01]. http://www.doc88.com/p-3724514050617.html.

[13] Wu Kewen, Zhu Qinghua, Zhao Yuxiang, et al. Simulation Study on Tag Retrieval Quality in Social Tagging Systems[J]. Journal of the China Society for Scientific and Technical Information, 2011, 30(1): 29-36.

[14] Luo Lin, Yang Yang. Research on Influencing Factors of User Tagging Behavior in Social Tagging Systems[J]. Document, Information & Knowledge, 2018(3): 85-94.

[15] Gao Bing, Sun Lin, Xie Biao, et al. Establishment and Application of Weighted Probabilistic Principal Component Analysis Model[J]. Chinese Journal of Health Statistics, 2018, 35(6): 802-805.

[16] Xu Shaocheng, Li Dongxi. Weighted Feature Selection Algorithm Based on Random Forest[J]. Statistics & Decision, 2018, 34(18): 25-28.

[17] Scott A G, Bernardo A H. Usage Patterns of Collaborative Tagging Systems[J]. Journal of Information Science, 2006, 32(2): 198-208.

[18] Liu Jingyan, Zhang Ke, Wang Guihua. Comparative Study on Data Standardization Methods in Comprehensive Evaluation[J]. Digital Technology & Application, 2018, 36(6): 84-85.

Author Contributions

Zhang Yunzhong: Conceptualized the research, proposed the overall framework, and revised the manuscript.
Qin Yiyuan: Conducted data processing and analysis, and drafted the manuscript.

Research on Influencing Factors of Tag Quality in Social Tagging System: Based on Random Forest Algorithm

Zhang Yunzhong, Qin Yiyuan
Department of Library, Information and Archives, Shanghai University, Shanghai 200444

Abstract: [Purpose/significance] Tag quality is often related to users' experience of classification, query, browsing, acquisition of online resources in social tagging systems. Identifying key influencing factors of tag quality can optimize the core resource organization functions of STS. [Method/process] Based on tags, we provided the influencing factors model of tag quality from six perspectives covering tagging subject, tagging object, tagging environment, tagging motivation, tagging methods and tagging products. The study attempted to explore key influencing factors of tag quality by questionnaire, and established the decision tree model of influencing factors of tag quality based on Random Forest. [Result/conclusion] Tagging subject is the primary key dimension affecting tag quality. The impact of the subject's knowledge structure and cognitive level, the subject's tagging frequency, and the subject's perceived usefulness are prominent. Tagging methods are the secondary one, and tag recommendation and standard tag tips are main influencing factors.

Keywords: social tagging system, tag quality, machine learning, random forest

Note: Figure translations are in progress. See original paper for figures.