A Study on the Correlation Between Weibo User Tags and Post Content: Postprint

Authors: Zhu Ling, Xue Chunxiang, Zhang Chengzhi, Fu Zhu

Date: 2017-10-11T00:00:00+00:00

Abstract

【Objective】 This study explores the potential relationship between Weibo user tags and the topics of their posted Weibo content, providing a reference for topic discovery and automatic user tag recommendation services on Weibo-like application platforms. 【Method】 A crawler program was utilized to collect user information and Weibo posts from Sina Weibo in the "natural language processing" domain, performing word segmentation on the collected Weibo content and semantic expansion on user tags, and employing the edit distance algorithm to match the tag set with users' Weibo content. 【Results】 Through sampling analysis of the matching results, it was found that on the Sina Weibo platform, academic domain Weibo user tags and user-posted content exhibit a certain degree of correlation. 【Limitations】 The research only focuses on the academic domain and Sina Weibo; the research domain and application platform need to be further expanded. 【Conclusion】 Weibo tag recommendation systems can utilize user Weibo content as an important data source for tag recommendation, providing users with more targeted personalized tags; simultaneously, when performing topic extraction and analysis on Weibo content, user tags can be leveraged to optimize analysis results.

Full Text

Research on the Correlation Between Weibo User Tags and Blog Content

Zhu Ling, Xue Chunxiang, Zhang Chengzhi, Fu Zhu
School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, China

Abstract

[**Objective**] This study explores the latent relationship between Weibo user tags and the topics of their posted content, providing a reference for topic discovery and automatic tag recommendation services on microblogging platforms. [**Methods**] We used crawler programs to collect user information and posts from the "natural language processing" community on Sina Weibo. After segmenting the collected post content and semantically expanding user tags, we matched the tag sets with users' microblog content using an edit distance algorithm. [**Results**] Sampling analysis of the matching results reveals that on the Sina Weibo platform, academic domain user tags exhibit a certain degree of correlation with the content of users' posts. [**Limitations**] This research focuses exclusively on the academic domain and Sina Weibo; both the research field and application platform require further expansion. [**Conclusions**] Microblog tag recommendation systems can utilize user post content as an important data source for tag recommendations, providing users with more targeted personalized tags. Conversely, when conducting topic extraction and analysis of microblog content, user tags can be leveraged to optimize analysis results.

Keywords: Subject analysis of posts; User tags; Correlation measure; Subject indexing; User modeling
Classification Number: G203

Introduction

Social tagging systems have become one of the most popular online services on the Internet, offering significant value for community discovery, information recommendation, and information integration. Weibo user tags are markers that users assign to themselves based on their professional fields or personal interests. While reflecting personalized characteristics, these tags also provide rich information sources for friend recommendations and user community segmentation. Since user tags are somewhat associated with the content users post, examining the degree of correlation between Weibo user tags and content holds important implications for automatic tag recommendation, friend recommendation, post tag recommendation, post topic retrieval, and information recommendation services.

However, few studies have quantified the correlation between user tags and microblog content. To address this gap, this paper takes academic users on Sina Weibo as an example, collecting their tags and microblog content to conduct statistical analysis on the correlation between user tags and content. This research further enriches the content of information organization studies and provides references for automatic user tag recommendation services on microblogging platforms to improve service quality.

Early research on tag-document correlation primarily focused on formal information resources such as web pages, books, and journal articles, measuring similarity by comparing tags with document subject terms or keywords. In 2006,

Al-Khalifa et al. [1] used Yahoo' s keyword extraction tool to extract web page keywords and performed pairwise matching among machine-extracted keyword sets, crowdsourced tag sets, and professional indexers' annotations. Results showed that professional indexers' annotations had higher overlap with crowdsourced tags than with machine-extracted keywords. In 2008, Rolla [2] compared subject headings with social tags and found that crowdsourced annotations provided more comprehensive and detailed descriptions of books, improving bibliographic retrieval performance, while subject headings could only serve as basic information indexing. Thomas et al. [3] reached the same conclusion in 2009. In 2010, Lu et al. [4] compared user tags on LibraryThing with Library of Congress Subject Headings assigned in libraries, finding that user tags could improve the accessibility of library resources. Also in 2010, Pan et al. [5] analyzed differences between tags and keywords on Del.icio.us, discovering significant variation in similarity between tags and keywords across entertainment and academic domains. However, their study scope was limited to these two broad domains with small sample sizes of only about a dozen objects per category. In 2011, Kipp [6] collected academic journal articles from three sources—user tags, author keywords, and subject terms—and found discrepancies in keyword-tag matching through descriptive statistics. In 2012, Lee et al. [7] compared MeSH subject terms for 231,388 papers in the Medline database with tags assigned by users on CiteULike, concluding that social tagging cannot replace traditional controlled indexing.

With the emergence of microblogs, scholars began investigating the correlation between microblog content and tags. Huang and Zhang [8] compared user tags with machine-generated tags, finding that user tags were somewhat associated with their microblog content. Zhang et al. [9] examined the topic expression capability of user tags on Tencent Weibo, revealing that among users with tags and high influence, approximately one-third of user tags were related to keywords in their posts. Xing et al. [10] assumed that both tags and microblogs could represent topics users follow, finding that users with more similar tags also had more similar post content.

In summary, scholars have conducted preliminary research on text topics and tags, noting that in specific domains, text topics and tags are both related and differentiated. Current research primarily focuses on academic and entertainment domains. Moreover, quantitative studies on the correlation between Weibo user tags and content analyze users' posts as an aggregate set, lacking research on individual posts as the unit of analysis.

Methodology

3.1 Basic Approach

This study uses crawler programs to collect user information and posts from the "natural language processing" community on Sina Weibo. Through correlation analysis between user tags and posts, we explore the thematic correlation

between Weibo user tags and their published content, thereby providing possibilities for tag-based microblog topic identification. The research framework is illustrated in [Figure 1: see original paper].

(1) **Microblog User Data Collection.** We selected users in the "natural language processing" field on Sina Weibo as our research subjects. Using "natural language processing" and "Chinese information processing" as keywords, we retrieved 835 users' profiles and 735,359 posts through the Sina Weibo API. User information included user ID, nickname, gender, and user tags.

(2) **User Tag Expansion.** To better represent content related to tag domains and improve semantic matching efficiency, we used the thesaurus API developed by Dr. Liang Bin from the Information Retrieval Group at Tsinghua University' s State Key Laboratory of Intelligent Technology and Systems to semantically expand each user tag, obtaining an expanded tag semantic set.

(3) **Microblog Data Processing.** We segmented posts using the ICTCLAS system. During segmentation, we imported user tags and tag expansion words into the segmentation system as a custom dictionary to improve segmentation performance.

(4) **User Tag and Post Content Matching.** We used the user tag set as a matching dictionary to match tags with users' microblog content.

3.2 Data Processing

(1) Microblog Data Preprocessing. During data preparation, we filtered out users with zero tags, ultimately obtaining 760 users with 703,635 posts. Consecutive forwarded or commented posts were treated as a single complete post.

We retrieved user-defined tags through the Weibo API. From the initial 835 users, after removing those with zero tags, we obtained 760 users with 4,689 tags, averaging approximately 6 tags per user. Using Liang Bin' s thesaurus API, we semantically expanded each user tag. shows partial expansion results for "natural language processing," where "artificial intelligence" is an expansion term and "0.184195" indicates the degree of word correlation. We processed the collected tags and expansion words to form a tag set for segmentation and post matching.

(2) Microblog Post Processing. Based on the above data, we segmented user posts, which is fundamental to text processing. Numerous Chinese segmentation algorithms and tools exist; this study used the ICTCLAS system, which outputs words with part-of-speech tags (e.g., noun/n, verb/v, adjective/a). User tags are primarily nouns. For example, one user' s self-assigned tags included: sentiment analysis, natural language processing, data mining, text classification,

pattern recognition, table tennis, PhD, Yangzhou, Nanjing, Beijing. The ICTCLAS system allows user dictionary import. This study imported the tag set obtained above as a custom segmentation dictionary. The system dictionary format arranges words line-by-line with part-of-speech tags, separated by a tab (e.g., "PhD n"). During processing, we tagged all tag words as "tag." compares processing effects before and after adding the tag dictionary.

As shown, phrases appearing in the tag set are segmented as complete units after dictionary addition. Without the tag dictionary, "natural language processing" was segmented into "natural language" and "processing" ; after adding the tag set, "natural language processing" was segmented as a single phrase.

3.3 Tag-Post Correlation Matching

Text semantic matching offers numerous methods for computing semantic similarity between words, such as corpus-based [11], dictionary-based [12], network or ontology-based [13-14], and edit distance-based [15] approaches. To simplify processing, this study used a thesaurus API for tag semantic expansion and introduced expansion words into the segmentation dictionary to improve segmentation precision. For subsequent word similarity matching, we selected the straightforward edit distance algorithm to match user tag sets with microblog content.

Edit Distance [15], also known as Levenshtein distance, refers to the minimum number of editing operations required to transform one string into another. Sim(Tag(u), Word(v)) denotes the similarity between tag u and microblog word v. Let Distance(Tag(u), Word(v)) be the edit distance between Tag(u) and Word(v), and length(x) represent the length of x. The similarity calculation formula is:

Sim(Tag(u),Word(v)) = 1 - Distance(Tag(u), Word(v)) / Max(length(Tag(u)), length(Word(v)))

Since edit distance reflects absolute differences between strings and is affected by word length, we only calculated similarity for tags longer than 3 characters during data processing, performing literal matching for tags of 3 characters or less. Through repeated experimental verification, we found that similarity thresholds greater than 0.5 yielded the most ideal matching results.

We illustrate tag-post matching results using user ID 1065269410, as shown in . In the matching results, "pattern recognition" is a user-assigned tag, while "computer vision" is an expansion term for "pattern recognition." "Expert," "academician," "UK," and "Chinese Academy of Sciences" are all expansion terms for the tag "PhD." The matching result "computer vision pattern recognition expert academician UK Chinese Academy of Sciences" demonstrates that "pattern recognition" is a self-assigned tag for this user type, while "computer vision expert academician UK Chinese Academy of Sciences" represents expansion terms for such tags.

We processed the segmented posts by aggregating the 703,635 post texts and 23,487 user tags by user ID to obtain each user' s post collection and tag collection, then performed post matching. Using the expanded user tag set as a matching dictionary, we matched the tag set with users' microblog content to achieve semantic matching between tag words and posts.

Results and Analysis

4.1 Analysis of Tag-Adding Behavior Among Professional Domain Users

Statistical analysis of users' personal tag-adding behavior yielded the results shown in [Figure 2: see original paper]. Among 835 users, 760 added at least one tag, while only 75 users assigned no tags. This differs significantly from Xing et al.' s [10] findings based on ordinary users, where 59.4% of users assigned no tags. Additionally, among users who added tags, 572 had more than five tags, indicating that professional domain users are more willing to add tags and assign as many as possible to gain peer attention.

Analysis of tag content in the "natural language processing" domain (excluding the search keywords "natural language processing"and "Chinese information processing," which appeared 622 and 35 times respectively) revealed the top 20 most frequently used tags, shown in . These high-frequency terms are mostly professional terminology related to the "natural language processing" field, whereas popular generic descriptors such as "travel," "food," "post-80s," "movies," and "music" appear only among the top five tags. These generic tags emerge because they are most easily recommended by the system without manual input and have universal applicability for users. Compared with Xing et al.' s [10] study of ordinary users, professional domain users prefer using more specialized terminology to describe their fields rather than mindlessly selecting popular recommended tags.

4.2 Correlation Measurement Between Professional Domain User Tags and Posts

Given the unique characteristics of microblog text, existing topic models cannot effectively analyze microblog content. Therefore, this study employed direct semantic matching between tags and content rather than keyword extraction. Statistical analysis of matching results is presented in [Figure 3: see original paper], where the matching rate is calculated as: (Number of posts associated with user tags) / (Total number of posts published by user).

As shown in [Figure 3: see original paper], user post-tag matching rates primarily concentrate below 70%, with only a few users achieving rates above 70%. Among 760 users, 341 achieved matching rates of 40% or higher, demonstrating a certain degree of correlation between user tags and post content.

To eliminate the influence of tag and post quantities, we conducted word fre-

quency statistics on post content from high matching rate intervals (>0.7) and low matching rate intervals (<0.1). shows the top 20 word frequencies. Judging professionalism by relevance to the "natural language processing" field, the high matching rate interval exhibits higher professional relevance than the low interval. Additionally, word frequencies in the high matching rate interval are substantially higher than in the low interval. Observing the top 100, 500, and 1000 words sorted by frequency reveals this consistent pattern. Word frequency partially reflects similarity between post contents, indicating that users in the high matching rate interval post more similar content than those in the low interval.

presents the number of professional terms among the top 20, 100, 500, and 1000 words. The high matching rate interval shows a gradually decreasing trend in professional vocabulary distribution, meaning more professional terms appear among higher-frequency words. In contrast, the low matching rate interval remains relatively balanced, indicating more dispersed word distribution. Observation of these users' posts reveals that high matching rate users primarily share research information, industry insights, or news, while low matching rate users post more about daily life and emotional expression with diverse content. This explains why both the proportion and frequency of professional vocabulary are higher in the high matching rate interval.

(1) Cause Analysis. The number of user tags and posts directly affects matching results. Analysis revealed that among 51 users with zero matching rate, 27 had only one tag, and the remainder posted fewer than 40 posts. After removing users with fewer than 5 tags and fewer than 100 posts, we examined the distribution of follower counts across different matching intervals, as shown in [Figure 4: see original paper]. Although no gradual increasing trend appears overall, users with matching rates above 0.6 have significantly higher average follower counts than those below 0.6. Follower count represents user influence and reflects user activity level, suggesting that more influential users exhibit stronger tag-post correlations.

Analyzing tags from users with matching rates above 0.6 and below 0.2, we categorized tag words into professional and non-professional vocabulary. Investigation showed professional tag vocabulary accounted for 81% in the >0.6 group versus 65% in the <0.2 group. Additionally, the <0.2 group' s non-professional vocabulary contained numerous internet slang terms, while only one such term appeared in the >0.6 group. For example, when describing programmers, terms like "IT elite female," "IT female laborer," "coding farmer buddy," and "software siege engineer" emerged. compares tag vocabulary across different matching intervals.

The comparison reveals that except for one instance of "naturally cute," all tags from the >0.6 group are traditional Chinese vocabulary. In contrast, among 168 non-professional terms in the <0.2 group, 70 are internet slang or descriptive phrases. This indicates that the proportion of professional vocabulary, combined with characteristics of non-professional vocabulary, can reflect users'

professionalism level. More professional users demonstrate stronger correlations between their posts and tags.

(2) Analysis of Different Matching Tag Sets. We matched both original tags and expanded tag sets with posts, with results shown in [Figure 5: see original paper]. Compared with the expanded tag set, original tags show lower matching rates, primarily concentrated below 10%. This demonstrates that original tags have some correlation with posts, but the number of correlated posts is limited, primarily because users have at most 10 tags, providing very limited expressive power. This validates the rationale for semantic tag expansion.

These results indicate that user original tags correlate with posts to some degree, but expanded tags show substantially higher correlation. This study achieved semantic rather than literal matching between tags and posts through edit distance algorithm and expanded tag sets, yielding more reasonable results than direct matching.

Conclusion

This study collected user tags and microblog content from academic users on Sina Weibo, analyzing tag-adding behavior patterns and tag content in the natural language processing domain. Results show that professional domain users are more willing to add numerous tags and prefer specialized terminology over popular recommended tags. Statistical analysis of tag-post correlation demonstrates that academic user tags exhibit certain correlations with post content on Sina Weibo. Besides post and tag quantities, user influence and professionalism also affect tag-post correlation.

Based on these findings, we recommend that general Weibo users value their self-defined tags as academic users do, avoiding arbitrary assignment of popular tags or neglecting tag assignment altogether. Microblog tag recommendation systems should consider user post content as an important data source for providing more targeted personalized tags. Conversely, user tags can optimize topic discovery and analysis results when processing microblog content.

References

[1] Al-Khalifa H S, Davis H C. Folksonomies Versus Automatic Keyword Extraction: An Empirical Study. *IADIS International Journal on Computer Science and Information Systems*, 2006, 1(2): 132-143.

[2] Rolla P J. User Tags Versus Subject Headings. *Library Resources & Technical Services*, 2011, 53(3): 174-184.

[3] Thomas M, Caudle D M, Schmitz C M. To Tag or not to Tag? *Library Hi Tech*, 2009, 27(3): 411-434.

[4] Lu C, Park J R, Hu X. User Tags Versus Expert-assigned Subject Terms: A Comparison of LibraryThing Tags and Library of Congress Subject Headings.

Journal of Information Science, 2010, 36(6): 763-779.

[5] Pan Chan, Feng Lifei, Ding Wanying, et al. Tag and Keyword-Based Analysis of Users' Behavior. *Journal of Intelligence*, 2010, 29(3): 139-142.

[6] Kipp M E I. Tagging of Biomedical Articles on CiteULike: A Comparison of User, Author and Professional Indexing. *Knowledge Organization*, 2011, 38(3): 245-261.

[7] Lee D H, Schleyer T. Social Tagging is no Substitute for Controlled Indexing: A Comparison of Medical Subject Headings and CiteULike Tags Assigned to 231,388 Papers. *Journal of the American Society for Information Science and Technology*, 2012, 63(9): 1747-1757.

[8] Huang Hongxia, Zhang Chengzhi. Investigation and Analysis of Chinese Microblog User Tags—Using Sina Weibo as Example. *New Technology of Library and Information Service*, 2012(10): 49-54.

[9] Zhang Chengzhi, He Lulin, Ding Peihong. Difference of Subject Expression Function of User Tags in Different Domains—Using Chinese Microblogging as Example. *Information Studies: Theory & Application*, 2013, 36(4): 68-71.

[10] Xing Qianli, Liu Lie, Liu Yiqun, et al. Study on User Tags in Weibo. *Journal of Software*, 2015, 26(7): 1626-1637.

[11] Baeza-Yates R, Ribeiro-Neto B. *Modern Information Retrieval*. New York: ACM Press, 1999.

[12] Kozima H, Furugori T. Similarity Between Words Computed by Spreading Activation on an English Dictionary. In: *Proceedings of the 6th Conference on European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 1993: 232-239.

[13] Jiang Min, Xiao Shibin, Wang Hongwei, et al. An Improved Word Similarity Computing Method Based on HowNet. *Journal of Chinese Information Processing*, 2008, 22(5): 84-89.

[14] Budanitsky A, Hirst G. Semantic Distance in WordNet: An Experimental, Application-oriented Evaluation of Five Measures. In: *Proceedings of the Workshop on WordNet and Other Lexical Resources, the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh. 2001.

[15] Levenshtein V I. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady*, 1966, 10(8): 707-710.

Author Contributions Statement

Xue Chunxiang: Conceived research ideas, designed research methodology, revised final manuscript.
Zhang Chengzhi: Data collection, research discussion.

Zhu Ling: Data analysis and processing, drafted manuscript.
Fu Zhu: Research discussion, manuscript revision suggestions.

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

Supporting data is self-archived by the authors. Contact: xuechunxiang@njust.edu.cn.

[1] Zhu Ling, Xue Chunxiang, Zhang Chengzhi, Fu Zhu. tagNum.xlsx. User tag quantity.
[2] Zhu Ling, Xue Chunxiang, Zhang Chengzhi, Fu Zhu. tagFreq.xls. User tag usage frequency.
[3] Zhu Ling, Xue Chunxiang, Zhang Chengzhi, Fu Zhu. wordFreq.xls. Word frequency distribution across different matching intervals.
[4] Zhu Ling, Xue Chunxiang, Zhang Chengzhi, Fu Zhu. funsNum.xlsx. User follower distribution across different matching intervals.
[5] Zhu Ling, Xue Chunxiang, Zhang Chengzhi, Fu Zhu. tagComp.xlsx. Tag vocabulary across different matching intervals.
[6] Zhu Ling, Xue Chunxiang, Zhang Chengzhi, Fu Zhu. matchResult.xls. Matching results of original and expanded tag sets with posts.

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