Multi-source Information Fusion Microblog Query Likelihood Model (Postprint)

Authors: Wu Shufang, Zhang Xiongtao, Zhu Jie

Date: 2023-04-01T16:16:00+00:00

Abstract

[Purpose/Significance] The query likelihood model suffers from the zero-probability problem. Integrating multi-source information to extend the model can not only resolve the zero-probability problem but also enable differentiated processing of global information and reduce noise. [Method/Process] Through LDA topic mining and historical microblog interest mining, topic-related information and interest-related information of initial microblogs are respectively obtained, and both are fused with global information to improve the language model estimation of initial microblogs, thereby deriving an extended microblog query likelihood model. A web crawler tool is employed to crawl data from Sina Weibo, and the effectiveness of the extended model is verified through empirical research. [Results/Conclusion] Experimental results demonstrate that, compared with existing query likelihood model extension methods, the new model exhibits superior retrieval performance.

Full Text

Preamble

The query likelihood model suffers from the zero-probability problem. Fusing multi-source information to extend this model not only resolves zero-probability issues but also enables differential processing of global information, thereby reducing noise. This study obtains initial microblog topic-related information and interest-related information through LDA topic mining and historical microblog interest mining, respectively. These two information sources are then integrated with global information to improve the estimation of the initial microblog's language model, yielding an extended microblog query likelihood model. Using web crawler tools to collect data from Sina Weibo, we verify the effectiveness of the extended model through empirical research. Experimental results demonstrate that the proposed model achieves superior retrieval performance compared to existing query likelihood model extensions.

With the further development of mobile internet, microblogging has become an important platform for people to produce, share, and consume information. To address information overload caused by massive microblogs, microblog retrieval has emerged as a crucial pathway for users to obtain effective information. Microblog retrieval differs from traditional text retrieval in several key aspects: its retrieval objects are characterized by fragmentation, internet slang, and heavy use of symbols; and its ranking principles must consider not only query-microblog similarity but also other information such as user interests and temporal factors. Consequently, directly applying traditional text retrieval models to microblog retrieval is inappropriate.

The query likelihood model represents the current mainstream approach for microblog retrieval. Its similarity calculation comprises document prior probability and document language model estimation (i.e., the probability distribution of terms in the document language model). The accuracy of document language model estimation directly affects retrieval performance. To address zero-probability problems that may arise due to data sparsity, scholars have conducted extensive research on extending the query likelihood model, primarily focusing on document language model estimation.

In traditional text retrieval, research on document language model estimation has evolved through two stages. First-stage methods estimate language models by introducing global information, such as the Jelinek-Mercer (JM) method and Dirichlet Prior (DIR) method. While these approaches effectively solve zero-probability problems in traditional language model estimation, they fail to differentially process global information, introducing substantial noise that compromises retrieval accuracy. Second-stage methods improve language model estimation by fusing global information with other relevant information. For example, X. Liu et al. utilized clustering information to revise global information, proposing a language model estimation method that integrates clustering and global information to extend the traditional query likelihood model. T. Tao et al. incorporated content neighbor information with global information to develop an improved query likelihood model. Empirical studies show that second-stage methods enable more accurate language model estimation than first-stage approaches that directly introduce global information, thereby effectively extending the query likelihood model.

The aforementioned research primarily targets traditional text retrieval. Considering the differences between microblogs and traditional texts, researchers have extended microblog query likelihood models by incorporating microblog-specific characteristics. The basic approach involves first identifying microblog-relevant information based on microblog features, then fusing this information with global information to improve the microblog query likelihood model. For instance, M. Efron et al. treated short microblogs as queries and combined them with relevance feedback methods to obtain relevant microblog information, which was then fused with global information to estimate the microblog language model, resulting in an improved microblog query likelihood model.

Li Rui et al. considered the timeliness and interactivity of microblogs, obtaining relevant microblogs based on users' historical microblogs and interaction information, then fused these with global microblogs to develop an improved microblog language model estimation method. M. Efron et al. used hashtags to obtain relevant microblogs and proposed a microblog query likelihood model fusing hashtag and global information. Hashtags are special tags in microblogs; microblogs with the same hashtag belong to the same topic, making hashtag information effective for obtaining relevant microblogs for the current microblog.

In summary, microblog language model estimation is the key component of microblog query likelihood models, and its accuracy directly impacts retrieval performance. Obtaining effective relevant microblog information is crucial for accurate estimation. Building upon existing research, this paper comprehensively considers four aspects: microblog self-information, global information, topic information, and author interest information. We propose a multi-source information fusion microblog query likelihood model that obtains relevant microblogs from multiple dimensions. In information retrieval, topic mining is an important means of acquiring text semantic information. Since its proposal in 2003, the LDA (Latent Dirichlet Allocation) topic model has been widely applied to topic mining. This paper employs the LDA model for topic mining to obtain microblog topic-related information. Additionally, user interest mining is a key technology for personalized retrieval. Effective interest mining can improve retrieval performance. Therefore, this paper also mines author interests based on historical microblogs to obtain interest-related information.

Research Design

Acquiring multi-source information is critical to the proposed extended model. Self-information refers to the microblog itself, and global information refers to all currently available information—both are easily obtainable. For topic-related information, this paper adopts an empirical research approach, conducting topic mining based on the LDA model. For interest-related information, we first mine author interests from historical microblogs, then calculate interest similarity to obtain the interest-related microblog set. The research framework is shown in Figure 1 [Figure 1: see original paper], with main contributions including:

1. **Acquisition of topic-related information for microblog d_i**: Based on LDA topic modeling, microblog texts are represented as probability distribution vectors across m topics. Microblog relevance is calculated based on topic distribution differences to obtain the topic-related microblog set T for microblog d_i.

2. **Acquisition of interest-related information for microblog d_i**: Author interests are mined from the historical microblogs of microblog d_i's author. To reflect the dynamic nature of interests, a dynamic weight calculation method for interest terms is proposed. Similarity between each microblog and author interests is calculated to obtain the author's interest

microblog set I.

3. **Multi-source information fusion**: Sets T, I, and the complete microblog collection are fused to smooth the initial microblog d_i, re-estimating term probability distribution to obtain the extended microblog query likelihood model.

Traditional Query Likelihood Model

In information retrieval, language models represent texts as probability distributions of words. The query likelihood model proposed by J.M. Ponte and W.B. Croft is a classic application of language models to information retrieval, calculated as:

$$P(d_i|q) = \log P(d_i) + \sum c(k,q) \log P(k|M_{d_i}) \tag{1}$$

where q represents the query, d_i represents the document, k represents a term, V represents the set of all terms, and M_{d_i} represents the document language model. P(d_i|q) denotes the probability of retrieving document d_i given query q. P(d_i) represents the prior probability of document d_i, typically measured using equal probability (assuming uniform prior probabilities for all documents to be retrieved). c(k,q) represents the frequency of term k in query q. p(k|M_{d_i}) is the probability distribution of term k in document language model M_{d_i}, i.e., the document language model estimation, calculated as:

$$P(k|M_{d_i}) = P_{ml}(k|M_{d_i}) = \frac{c(k,d_i)}{|d_i|} \tag{2}$$

In formula (2), P_{ml}(k|M_{d_i}) represents the maximum likelihood estimation. c(k,d_i) denotes the frequency of term k in document d_i, and |d_i| represents the number of terms in document d_i. For term k in the complete vocabulary, if document d_i does not include this term, the zero-probability problem occurs, making log P($\cdot$) calculation meaningless. This problem becomes more severe with shorter documents. In reality, although term k may not appear in document d_i, if its related terms appear in d_i, this probability value should not be zero. Therefore, the key to solving the zero-probability problem is finding relevant information to effectively smooth the initial document d_i and proposing improved probability estimation methods for p(k|M_{d_i}), which is also the focus of this paper.

Multi-source Information Fusion Microblog Query Likelihood Model

To overcome the limitations of traditional query likelihood models, accurately estimate term probability distributions in documents, and improve the comprehensive performance of microblog retrieval based on query likelihood models,

this paper extends existing research by using topic-related information (semantic relevance) and interest-related information (personalized information) to differentially process global information, smoothing initial microblogs and proposing an extended microblog query likelihood model.

4.1 Acquisition of Topic-Based Related Microblog Set

The LDA model is the current mainstream method for text topic mining. After training with the LDA model, texts can be mapped from term space to topic space, achieving semantic representation. This paper employs the LDA topic model for microblog text modeling: first using Python's Gensim toolkit to train m topics, then representing each microblog as a probability distribution across m topics to obtain the microblog-topic probability distribution matrix shown in Table 1 .

After this training, microblog d_i can be represented as a topic vector composed of probability distributions across different topics:

$$d_i = (P_{i1}, P_{i2}, \dots, P_{im}) \tag{3}$$

Based on this representation, we use JS distance to calculate the topic relevance between any two microblogs d_i and d_j:

$$JS(d_i, d_j) = KL\left(d_i, \frac{d_i + d_j}{2}\right) + KL\left(d_j, \frac{d_i + d_j}{2}\right) \tag{4}$$

In formula (4), KL($\cdot$) measures asymmetric distance between two quantities, calculated as shown in formula (5), where (d_i + d_j)/2 represents the mean distribution of microblogs d_i and d_j across m topics. Larger JS distance indicates greater distribution differences and lower relevance. Based on this value, we sort JS distances in ascending order and select the Top-N_1 microblogs (N_1 determined through experiments) to form the topic-related microblog set for the current microblog.

$$KL(d_i, d_j) = \sum P_{ir} \log \frac{P_{ir}}{P_{jr}} \tag{5}$$

In formula (5), P_{ir} represents the probability distribution of microblog d_i on topic Topic_r, and P_{jr} represents the probability distribution of microblog d_j on topic Topic_r.

4.2 Acquisition of Author Interest-Based Related Microblog Set

Historical microblogs can effectively reflect user interests. This paper mines author interests based on historical microblogs, calculates similarity between each microblog and author interests, and finally obtains the author's interest

microblogs through threshold judgment. All interest microblogs constitute the author interest-based related microblog set I. Assuming user u_i is the author of microblog d_i, with historical microblog set D and any term k_j in the historical microblog set, the initial weight calculation formula for k_j is:

$$w_{kj-original} = \frac{|\{r : k_j \in d_r\}|}{|D|} \times \frac{n_{ji}}{|d_i|} \tag{6}$$

In formula (6), w_{kj-original} represents the initial weight of term k_j, n_{ji} represents the frequency of term k_j in microblog d_i, |d_i| represents the number of terms in microblog d_i, |D| represents the number of microblogs in the historical microblog set, and |{r: k_j d_r}| represents the number of microblogs containing term k_j in the historical microblog set.

Considering that microblog users' interests gradually decay over time, this paper updates term weights based on the publication time of the microblogs containing the terms. Following the idea of exponential decay, the updated weight calculation formula is:

$$w_{kj-new} = w_{kj-original} \times e^{-\mu \Delta t} \tag{7}$$

where w_{kj-new} represents the updated weight, Δt represents the time distance between microblog publication time and the latest time in the historical microblog set, and is the exponential decay parameter, which is set to 0.02 based on J. Choi's experimental results.

Since term k_j may have different weights in different microblogs, this paper uses the average weight of k_j across the entire historical microblog set as the term's weight in the historical microblog set:

$$w_{kj-D} = \frac{\sum_{|D|} w_{kj-new}}{|D|} \tag{8}$$

Through the above calculations, we can obtain the weight of each term in the author's historical microblog set. This paper selects Top-N_2 terms based on term weights to represent author interests (N_2 determined in the experimental section). For example, author u_i's interests can be represented as:

$$u_i - interest = \{k_1, k_2, \ldots, k_{N_2}\} \tag{9}$$

After obtaining the user interest representation, formula (10) calculates the similarity between any microblog d_r in the microblog collection and author interest u_i-interest:

$$sim(d_r, u_i - interest) = \frac{\sum_{j=1}^{N_2} w_{kj-D} \times |d_r - k_j|}{|d_r|} \tag{10}$$

where sim(d_r, u_i-interest) represents the similarity between microblog d_r and author interest u_i-interest, N_2 represents the number of author interest representation terms, w_{kj-D} represents the weight of author interest term k_j calculated according to formula (8), |d_r-k_j| represents the frequency of author interest term k_j in microblog d_r, and |d_r| represents the number of terms in microblog d_r. Microblogs with similarity greater than threshold δ are selected as the author's interest microblogs to form author u_i's interest microblog set I.

4.3 Extended Microblog Query Likelihood Model

After the above processing, we obtain the topic-related microblog set T for microblog d_i and the interest microblog set I of d_i's author. By fusing term k's distribution in the original microblog d_i (M_{d_i}), in the topic-related microblog set T (M_T), in the interest-related microblog set I (M_I), and in global information (M_C), we derive the extended microblog query likelihood model shown in formulas (11) and (12):

$$P(d_i|q) = \log P(d_i) + \sum c(k,q) \log P(k|M_{d_i})_{improve} \tag{11}$$

$$P(k|M_{d_i})_{improve} = \beta_1 P_{ml}(k|M_{d_i}) + \beta_2 P_{ml}(k|M_T) + \beta_3 (P_{ml}(k|TI) + \beta_4 P_{ml}(k|M_C) \tag{12}$$

Examining formula (11) reveals that the proposed extended model primarily improves document language model estimation P(k|M_{d_i}), avoiding the shortcomings of traditional query likelihood models. P(k|M_{d_i}){improve} *represents the improved microblog language model estimation. M_T denotes the language model constructed from the topic-related microblog set T for microblog d_i. M_I denotes the language model constructed from the interest microblog set I of d_i's author.* P{ml}(k|M_{d_i}) represents the maximum likelihood estimate of language model M_{d_i}. P_{ml}(k|M_T) represents the maximum likelihood estimate of language model M_T. P_{ml}(k+M_I) represents the maximum likelihood estimate of language model M_I. P_{ml}(k|M_C) represents the maximum likelihood estimate of language model M_C. All these estimates are calculated using formula (2). *βi (i = 1, 2, 3, 4) are smoothing parameters with* Σ\{i=1\}^4 *βi = 1. The inclusion of global information (*P{ml}(k|M_C)) avoids zero-probability problems because term k may not belong to a particular microblog d_i but certainly originates from the global collection. The drawback of fusing global information is noise introduction. To address this, the paper differentially processes global information based on

topic-related information (P_{ml}(k|M_T)) and interest-related information (P_{ml}(k|M_I)), increasing probabilities of relevant terms. During feature selection, low-probability terms will be eliminated, effectively preventing noise introduction.

This paper employs the Analytic Hierarchy Process to determine smoothing parameters β_i (i = 1, 2, 3, 4) in formula (12). First, based on the 1-9 importance scale, pairwise comparisons of each component's importance yield the judgment matrix shown in Table 2 . Then, based on this matrix, the maximum eigenvalue is calculated as 4.1389, with eigenvector (0.54, 0.25, 0.15, 0.06), consistency index 0.0463, and consistency ratio 0.0514. Since the consistency ratio is less than 0.1, the judgment matrix passes the consistency test. This paper sets β_i values as 0.54, 0.25, 0.15, and 0.06 respectively.

Empirical Research

5.1 Experimental Data

Sina Weibo is currently the most authoritative microblog platform in China. This paper uses web crawler tools to collect 661,845 Sina Weibo posts and constructs a microblog retrieval system based on the query likelihood model. The crawled data includes three types of information: microblog text content, publication time, and author.

To avoid interference from invalid data, this paper follows TREC (Text Retrieval Conference) evaluation requirements combined with experimental needs to process the crawled Sina Weibo data: (1) remove invalid microblogs or those containing only emoticons; (2) remove microblogs shorter than 30 characters; (3) convert all traditional Chinese characters to simplified Chinese; (4) use Python's jieba package for word segmentation of each microblog and remove stopwords using the stopword list compiled by Harbin Institute of Technology.

After processing the crawled microblog corpus, this paper follows small-scale test collection construction methods in information retrieval to select 17,010 microblogs as the document collection for the microblog retrieval system. Five queries and their corresponding query sets were constructed, with relevance judged using the Pooling method. The queries and their relevant/irrelevant document counts are shown in Table 3 .

5.2 Evaluation Metrics

This paper employs Precision at k (P@k) and Mean Reciprocal Rank (MRR) to evaluate microblog retrieval performance. P@k is the official evaluation metric in TREC microblog retrieval tasks. MRR is a popular metric in recent years that considers positional factors on top of precision, effectively measuring the position information of relevant documents. The calculation formulas are:

$$P@k = \frac{\sum_{j=1}^{k} r_j}{k} \tag{13}$$

$$MRR = \frac{1}{|R|} \sum_{i=1}^{|R|} \frac{1}{rank_i} \tag{14}$$

In formula (13), k represents the top k retrieval results. If the j-th document in the retrieval results is relevant, then r_j = 1; otherwise, r_j = 0. This paper sets k = 30 because the first two pages of microblog web search contain 30 results. In formula (14), |R| is the total number of relevant documents, and rank_i is the position of the i-th relevant document in the returned results. Higher MRR values indicate that relevant documents appear earlier in the result list, reflecting better retrieval performance. Model explanations are shown in Table 4 .

5.3 Experiments and Analysis

The experiments consist of two parts: parameter setting and retrieval performance comparison using different query likelihood models for microblog retrieval. The abbreviations and explanations of microblog query likelihood models involved in the second part are shown in Table 4 .

5.3.1 Parameter Analysis The relevant parameters in this paper's experiments include: microblog topic count m, topic-related microblog count N_1, author interest representation term count N_2, and author interest threshold δ. These parameters are determined through repeated experiments.

Determination of microblog topic count m: This paper calculates the perplexity of microblog texts to determine the optimal number of topics. Lower perplexity indicates stronger model text generation capability and better performance. The calculation formula is:

$$perplexity = \exp\left(-\frac{\sum_{i=1}^{M} \log P(k_{d_i})}{\sum_{i=1}^{M} N_{d_i}}\right) \tag{15}$$

where M represents the number of microblogs in the complete collection D, d_i is any microblog in D, N_{d_i} represents the number of terms in microblog d_i, k_{d_i} represents terms in microblog d_i, and P(k_{d_i}) represents the probability of term k_{d_i} appearing in microblog d_i, calculated using formula (16):

$$P(k_{d_i}) = \sum_{z} P(z) \cdot P(k_{d_i}|z) \tag{16}$$

where z represents a topic involved in microblog d_i. Since microblog texts are typical short texts with limited topic coverage, based on the scale of experimental data crawled in this paper, we calculate model perplexity for topic counts of 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10. The results are shown in Figure 2 [Figure 2: see original paper]. The figure shows that when topic count m = 4, the LDA model achieves relatively low perplexity, so this paper sets the topic count to 4.

Determination of topic-related microblog count N_1: Using an appropriate amount of topic-related microblogs for smoothing can effectively improve microblog language model estimation accuracy. Too few microblogs produce insufficient smoothing effects, while too many may introduce noise. To obtain a reasonable N_1, this paper first conducts empirical research with data intervals of 10, 100, and 1000. The results show that when the interval is 10, topic information cannot be fully utilized; when the interval is 1000, substantial noise is introduced. Therefore, this paper sets the microblog quantity interval to 100. Experiments based on the LM-JM-Topic method estimate the microblog language model with N_1 taking values of 100, 200, 300, 400, 500, 600, 700, 800, 900, and 1000, calculating P@30 for 5 queries in the microblog retrieval system. The results are shown in Table 5 .

Table 5 shows that when N_1 = 300, the average P@30 for 5 queries reaches a relatively high value, indicating that selecting Top-300 topic-related microblogs can effectively estimate the microblog language model. The value 300 represents a roughly optimal setting; more precise values could be obtained by reducing the interval and repeating the experiment. Notably, for Query 2, P@30 shows a slight rebound when topic-related microblogs increase to 700, and for Query 3, a rebound occurs at 900. Analysis reveals that although these microblogs have lower topic relevance to the original microblog, they exhibit higher relevance in author interaction and publication time distribution, leading to higher comprehensive relevance and causing these minor rebounds. However, these rebounds do not exceed the maximum P@30 values, and from the average P@30 perspective, the minor rebounds do not affect the overall decreasing trend. Therefore, this paper preliminarily sets N_1 to 300.

Determination of author interest representation term count N_2: This paper uses formula (6) to calculate term weights in the author's historical microblog set, sorts them, and selects Top-N_2 terms to represent author interests. Too small N_2 cannot adequately represent author interests, while too large N_2 reduces representation distinctiveness. To select an appropriate number of terms for representing author interests, this paper uses the Average User Satisfaction (AUS) metric to determine N_2. The specific process involves: first randomly selecting 10 microblog authors from the dataset and using an expert panel to annotate interest terms for each author, selecting interest term sets that can represent each author's interests (size ≤ 40); then calculating AUS values for these 10 authors' Top-N_2 terms and selecting the N_2 corresponding to higher AUS values. The AUS calculation formula is:

$$AUS = \frac{1}{M}\sum_{m=1}^{M}\frac{num}{N_2} \tag{17}$$

In formula (17), M represents the number of randomly selected authors (M = 10 in this paper), num represents the number of interest terms annotated by experts for author u_m, and N_2 represents the number of terms selected from the interest term list. This paper sets N_2 to values of 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100, obtaining the relationship between different interest term counts and AUS shown in Figure 3 [Figure 3: see original paper]. The figure shows that when the author interest representation term count is 40, the AUS value is relatively high, so this paper preliminarily sets N_2 = 40.

Determination of author interest threshold δ: To obtain a reasonable author interest threshold, experiments based on the LM-JM-Interest query likelihood model (which primarily considers interest-related information) calculate P@30 for 5 queries in the microblog retrieval system with δ taking values of 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. The results are shown in Table 6 .

Table 6 shows that when $\delta = 0.8$, the average P@30 for 5 queries reaches a relatively high value. Notably, for Query 2, P@30 shows a slight rebound when the interest threshold decreases to 0.4, and for Query 5, a rebound occurs when the threshold decreases to 0.2. These minor rebounds do not exceed the maximum P@30 values, and from the average perspective, they do not affect the overall decreasing trend. Therefore, this paper preliminarily sets δ to 0.8.

5.3.2 Performance Comparison of LM-JM-Topic, LM-JM-Interest, and LM-JM-Topic-Interest This paper compares the performance of LM-JM-Topic, LM-JM-Interest, and LM-JM-Topic-Interest query likelihood models using P@30 and MRR metrics.

P@30 Comparison: Figure 4 [Figure 4: see original paper] shows the P@30 values when using LM-JM-Topic, LM-JM-Interest, and LM-JM-Topic-Interest as representation models in the microblog retrieval system. The results demonstrate that when using the LM-JM-Topic-Interest query likelihood model, all five queries achieve higher P@30 values than the other two methods. This indicates that the proposed LM-JM-Topic-Interest model provides more accurate estimation values compared to LM-JM-Topic and LM-JM-Interest models, thereby improving the precision of the microblog retrieval system.

MRR Comparison: Figure 5 [Figure 5: see original paper] shows the MRR values for the three models across five queries in the test collection. The results reveal that when using the LM-JM-Topic-Interest query likelihood model, all five queries achieve higher MRR values than the other two methods. This suggests that the proposed LM-JM-Topic-Interest model can rank relevant documents higher than LM-JM-Topic and LM-JM-Interest models.

In summary, when using the LM-JM-Topic-Interest query likelihood model for microblog retrieval, all five queries achieve better P@30 and MRR performance in the microblog retrieval system compared to the other two models. This occurs because the LM-JM-Topic-Interest method considers both topic-related microblog information and author interest information, while the other two methods introduce relatively one-sided relevant information, leading to insufficient accuracy in microblog language model estimation and affecting the comprehensive performance of microblog retrieval systems.

5.3.3 Comparison of LM, LM-JM, and LM-JM-Topic-Interest This experiment compares the proposed LM-JM-Topic-Interest query likelihood model with traditional query likelihood model LM and global information-based extended query likelihood model LM-JM.

P@30 Comparison: Figure 6 [Figure 6: see original paper] compares the P@30 values obtained using the three models for five queries in the test collection. The results show that when using the LM-JM-Topic-Interest model for microblog retrieval, all five queries achieve higher P@30 values than the other two methods, indicating that the proposed LM-JM-Topic-Interest method yields higher precision for the microblog retrieval system compared to LM and LM-JM methods.

MRR Comparison: Figure 7 [Figure 7: see original paper] shows the MRR values for the three models. The results demonstrate that when using the LM-JM-Topic-Interest model for microblog retrieval, all five queries achieve higher MRR values than the other two methods, indicating that the proposed LM-JM-Topic-Interest method can rank relevant documents higher than LM and LM-JM methods.

In conclusion, when using the LM-JM-Topic-Interest model for microblog retrieval, all five queries achieve better retrieval performance in the microblog retrieval system compared to the other two models. Therefore, the LM-JM-Topic-Interest model outperforms the other two models. This is because the traditional query likelihood model only considers microblog self-information, suffering from zero-probability problems due to data sparsity. Global information-based smoothing methods solve the zero-probability problem but introduce excessive supplementary information and noise. The proposed microblog query likelihood model uses topic-related information and author interest information to differentially process global information, effectively increasing probabilities of relevant terms and decreasing probabilities of noise terms, thereby improving microblog language model estimation accuracy and microblog retrieval performance.

Conclusion

Addressing limitations of existing query likelihood models, this paper comprehensively utilizes microblog self-information, topic information, author inter-

est information, and global information to propose a multi-source information fusion microblog query likelihood model. Compared with existing research, this study improves microblog retrieval performance to some extent, but limitations remain. Future research will focus on: (1) This paper works with offline microblog data, whereas real microblog data is updated in real-time as data streams. Future research will incorporate online learning ideas to improve the microblog query likelihood model. (2) This paper primarily combines four aspects of information to improve the query likelihood model, but other information (such as temporal information and author interaction information) could also benefit microblog language model estimation. Future research will deeply explore other effective information to further improve microblog language model estimation accuracy and microblog retrieval performance.

References

[1] Wu Shufang, Zhang Xiongtao, Zhu Jie. A microblog retrieval model fusing user interests and hybrid estimation [J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(4): 411-419.

[2] KATZ S. Estimation of probabilities from sparse data for the language model component of a speech recognizer [J]. IEEE transactions on acoustics speech & signal processing, 2003, 35(3): 400-401.

[3] GANGULY D, ROY D, MITRA M, et al. A word embedding based generalized language model for information retrieval [C]// Proceedings of the 38th international ACM SIGIR conference. Santiago: ACM, 2015: 795-798.

[4] LIU X, CROFT W B. Cluster-based retrieval using language models [C]// Proceedings of the 27th annual international ACM SIGIR conference. Sheffield: ACM, 2004: 186-193.

[5] TAO T, WANG X, MEI Q, et al. Language model information retrieval with document expansion [C]// Proceedings of the human language technology conference of the north American chapter of the ACL. New York: Association for Computational Linguistics, 2006: 407-414.

[6] EFRON M, ORGANISCIAK P, FENLON K. Improving retrieval of short texts through document expansion [C]// Proceedings of the 35th international ACM SIGIR conference. Portland: ACM, 2012.

[7] Wei Bingjie, Shi Liang, Wang Bin. A new microblog ranking method fusing clustering and temporal information [J]. Journal of Chinese Information Processing, 2015, 29(3): 177-183, 189.

[8] EFRON M. Hashtag retrieval in a microblogging environment [C]// Proceedings of the 33rd international ACM SIGIR conference. Geneva: ACM, 2010: 787-788.

[9] Zhang Xiaopeng, Lü Xueqiang, Li Zhuo, et al. A topic phrase extraction method combining LDA and lexical chains [J]. Journal of Chinese Computer

Systems, 2018, 39(11): 107-.

[10] BLEI D M, NG A Y, JORDAN M I, LAFFERTY J. Latent dirichlet allocation [J]. Journal of machine learning research, 2003(3): 993-1022.

[11] JIANG Y, XU Y, SHAO L. A personalized microblog search model considering user-author relationship [C]// Proceedings of international conference on data science in cyberspace. Changsha: IEEE, 2016: 508-513.

[12] PONTE J M, CROFT W B. A language modeling approach to information retrieval [C]// Proceedings of the 21st annual international ACM SIGIR conference. New York: ACM, 1998: 275-281.

[13] Liu Dexi, Fu Qi, Wei Yaxiong, et al. A social short text retrieval method based on multiple enhanced graphs and topic analysis [J]. Journal of Chinese Information Processing, 2018, 32(3): 110-.

[14] Tang Xiaobo, Fang Xiaoke. A query expansion method for microblogs [J]. Library and Information Service, 2014, 58(1): 130-135.

[15] Xiong Caiwei, Cao Yanan. Research on microblog user interest mining method based on posted content [J]. Application Research of Computers, 2018(6): 63-71.

[16] CHOI J, CROFT W B. Temporal models for microblogs [C]// Proceedings of the 21st ACM international conference on Information and Knowledge Management. Maui: ACM, 2012: 2491-2494.

[17] VAIDYA O S, KUMAR S. Analytic hierarchy process: An overview of applications [J]. European journal of operational research, 2006, 169(1): 1-29.

[18] LIN J, ROEGIEST A, TAN L, et al. Overview of the TREC 2016 real-time summarization track [C]// Proceedings of the 25th text retrieval conference. Maryland: TREC, 2016.

[19] Xu Jianmin, Wang Ping. Construction and analysis of small Chinese information retrieval test collections [J]. Information Science, 2009, 28(1): 13-16.

[20] CORMACK G V, PALMER C R, CLARKE L A. Efficient construction of large test collections [C]// Proceedings of the 21st annual international ACM SIGIR conference. Melbourne: ACM, 1998: 282-.

[21] WANG Y, HUANG H, FENG C. Query expansion based on feedback concept model for microblog retrieval [C]// Proceedings of the 26th international conference on World Wide Web. Perth: International world wide web conferences steering committee, 2017: 559-568.

[22] Guan Peng, Wang Yuefen. Research on determining optimal topic number for LDA topic model in scientific and technical intelligence analysis [J]. New Technology of Library and Information Service, 2016(9): 42-50.

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