Information Fusion 91 (2023) 566–581

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

A survey on cross-media search based on user intention understanding in social networks

Lei Shi [a], Jia Luo [b,*], Chuangying Zhu [c], Feifei Kou [d], Gang Cheng [e], Xia Liu [f]

[a] State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, 100024, China
[b] College of Economics and Management, Beijing University of Technology, Beijing, 100124, China
[c] Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin, 541004, China
[d] School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing, 100876, China
[e] School of Computer Science, North China Institute of Science and Technology, Beijing, 101601, China
[f] School of Physics and Electronic Information, Yantai University, Yantai, 264005, China

ARTICLE INFO

Keywords:
Social network
User intention understanding
Cross-media search
Deep learning

ABSTRACT

With the increasing popularity of online social networks, more and more people are posting information, updating their statuses, and searching for topics there. Massive cross-media big data has been gathered by online social networks, with high dynamics, context-sparsity, and cross-media semantic gaps. In addition, it can be challenging to understand users' search intentions in the setting of social networks. The above problems have brought severe challenges and obstacles to cross-media searches, which also attracted more and more attention on social networks. As a relatively new research topic and interest, the concept, methodology, and overall research idea of cross-media search based on user search intention understanding are not evident in the literature. The research also lacks a unified paradigm and relatively complete research ideas on social networks. To solve these problems, we reviewed more than 100 references based on our preliminary exploration and research experience in this field from the whole process involved. We also detailed methodology, datasets, evaluation indicators, experiment evaluation, and research trends and analyzed the challenges. These works will help beginners quickly establish research ideas and processes in this field, and enable them to focus on algorithm design without paying too much attention to datasets, evaluation metrics, and research frameworks. We believe this review will attract more researchers to focus on social network cross-media search based on user search intention understanding and benefit their work.

1. Introduction

Online social networks' rapid development and widespread use have greatly facilitated users' daily lives. Users can use their mobile phones, PCs, tablets, and other terminals to post a range of topics and information through social networking platforms anytime, anywhere, including national policies, news themes, common dialog, and mood sharing. Data from Sina Weibo's official website indicates that as of June 2022, there were 252 million daily active users in addition to 582 million monthly active users, a net gain of 22 million over the same period in 2021. With the rise in active users, the platform generates massive amounts of cross-media data daily. Users are eager to do cross-media searches using one of the modalities, such as text, to look for results in other relevant modalities, such as images, videos, audio, etc. Cross-media search successfully realizes the mutual search between extra modals in cross-media data by returning the search results of other modals relevant to a modal query phrase. The results can intuitively reflect and supplement the semantic information of other modals, which is helpful to describe the topic or event more deeply in recent years. Social network cross-media search is becoming the forefront and hotspot of current academic research and has received extensive attention from the academic circle. It is a crucial branch for the advancement of information search in the future. In addition, different users have their own needs and expectations for search results. How to understand and mine users' search intentions and return topics that users expect or are interested in has important practical significance. Therefore, the research on social network cross-media search based on user intention understanding has important academic significance and practical utility.

* Corresponding author.
E-mail address: jialuo@bjut.edu.cn (J. Luo).

https://doi.org/10.1016/j.inffus.2022.11.017
Received 12 August 2022; Received in revised form 13 November 2022; Accepted 15 November 2022
Available online 17 November 2022

Social network cross-media search technology involves machine learning, image processing, computer vision, natural language processing, and other fields. It is closely related to many different disciplines, such as mathematics, statistics, and computer science. The research on cross-media search will significantly promote the development and application of many machine learning theories. Typical theories include topic models, subspace learning, metric learning, deep learning, hash learning, and multi-view learning. The social network cross-media search method based on user search intent understanding is based on the application of cross-media search in the specific field of social networks. The social network structure is complex, and it contains various attributes and features such as users, following relationships, topics, multimodality, time information, and location information. Therefore, compared with general cross-media search, social network cross-media search is more complex. It involves both cross-media search and cross-media semantic learning and knowledge fusion, user search intent understanding and mining, bursty topic discovery, etc. Researching social networks cross-media search based on user intent understanding can provide theoretical support for the emergence and development of new search technologies. In addition, it effectively meets users' needs for diverse information retrieval methods and promotes establishing a more effective personalized search application.

Cross-media search refers to the search modal in which the input query data and the output queried data belong to different modes. Due to the gap in semantic understanding of data in other feature spaces and the diversity of data in various modals, how to reduce the semantic gap and retain the effective comparative features of data is the key issue of cross-media search. The main idea of cross-media search is to use heterogeneous data representing the same semantics to build the corresponding relationship between different modals, build a mathematical model and optimize the solution, and then compare the data similarity of each modal to search for other modal information of the same semantics. Due to the gap in semantic understanding of data in different feature spaces, the data of each modal has diversity and integrity. Researchers of cross-media search mainly use various technologies to model and compare similarities to reduce the gap and retain data integrity. The social network cross-media search problem definition is as follows:

To facilitate the introduction of our related content, we define cross-media search in two modalities (image and text). Given a social network cross-media dataset $\bar{D} = \{d_1, d_2, \cdots, d_n, \cdots, d_N\}$, where $d_n = (v_n, t_n, E_n)$ represents each data instance in the social network cross-media dataset, and v_n and t_n represent the images and text in that instance. $E_n = [E_{n1}, E_{n2}, \cdots, E_{nE}] \in R^E$ denotes the category to which each data instance belongs. N represents the number of instances in the social network cross-media dataset, and E represents the number of categories. Each element E_{ne} in the vector of class labels has a value of 1 or 0. Where 1 means that the instance belongs to the e-th category, and 0 means that the instance does not belong to the e-th category. Constructing the label matrix S indicates whether the two samples belong to the same category, where s_{ij}=0 means that the two samples do not belong to the same category, and s_{ij}=1 demonstrates that the two samples belong to at least the same category. The goal is to learn a cross-media similarity measure $sim(\cdot)$ for a given input sample i, which can be text or an image. It expects to return the sample of another modality similar to the input sample j, s_{ij}=1, for the social network cross-media search problem. The returned samples can be text or images.

According to the aforementioned research, general cross-media search techniques concentrate on the shared semantic representation of various modals and alleviate the semantic gap. In addition to general cross-media search problems, social network cross-media search also has its own unique problems: (1) Social networks contain time, location, topic, cross-modal data, and other attributes and features and are faced with the problem of cross-media semantic learning and knowledge fusion. (2) Social network is an Internet application with users as the core, so it faces the problem of understanding users' search intention in social networks. (3) Relevant hot events and discussions in social networks are often based on topics, and the problem of topic discovery in social networks needs to be solved. (4) Social network cross-media search is facing the problem of cross-media semantic gap. Given the above problems, this paper mainly focuses on the social network cross-media search based on user search intention understanding, including cross-media semantic learning and knowledge fusion, user search intention understanding and mining, bursty topic discovery, social network cross-media search, and other research directions. Cross-media semantic learning and knowledge fusion mainly solve the semantic gap faced by cross-media search on social networks. User search intention understanding and mining mainly solve the problem of user preference perception. By obtaining users' behavior, users' intentions can be mined, and then search results can be returned more accurately. Topic discovery and recognition can achieve accurate cross-media search on social networks. Cross-media search mainly solves the semantic gap and search results ranking problem. By integrating cross-media semantic learning and knowledge fusion, user search intention understanding and mining, and information obtained from sudden topic discovery, it can achieve cross-media search based on user search intention understanding on social networks. By referring to the research progress of cross-media search, combined with our research experience for cross-media search in social networks, we propose a research framework of cross-media search based on user search intention understanding in social networks. The research framework is shown in Fig. 1.

Wang et al. [1] gave a comprehensive survey on cross-media search in a typical cross-media search overview. Kaur et al. [2] summarize cross-media search methods from multiple perspectives. However, they focus on the cross-media search algorithm that does not consider the complex application scenario of social networks. In addition, they do not believe the whole process of implementing cross-media search on social networks. Moreover, its research focus is too scattered to cover all the essential issues of the application of social network cross-media search tasks.

The research on cross-media search review is different from the above surveys. Based on our years of experience in cross-media search research on social networks, we mainly focus on the broad application scenario of social network cross-media search, and take users as the center. From the perspective of all aspects involved in social network cross-media search, we review the four important processes involved in social network search step by step. In addition, we focus on the latest research progress, recent open data sets, and evaluation methods in recent years, point out the problems and challenges in this field at the emerging stage, and look forward to the future direction. The researchers have provided various ideas and technologies for different challenging problems cross-media search faces. In this paper, we mainly summarize the latest research results of depth cross-media search, which differ from previous related research. We take online social network cross-media search based on user search intention understanding as the research object. We summarize the relevant research progress from four aspects: semantic learning and knowledge fusion of online social network cross-media big data, user search intention understanding and mining in social networks, online social network bursty topic discovery, and online social network cross-media search. We also discuss cross-media datasets, evaluation methods, and future research trends for cross-media searches in social networks. The main contributions of this paper are as follows:

(1) We summarize the current relevant theories and methods from the whole process of cross-media search based on user search intention understanding in social networks. It can help promote the continuous progress of this research field and provide a complete and cutting-edge research perspective for relevant researchers.

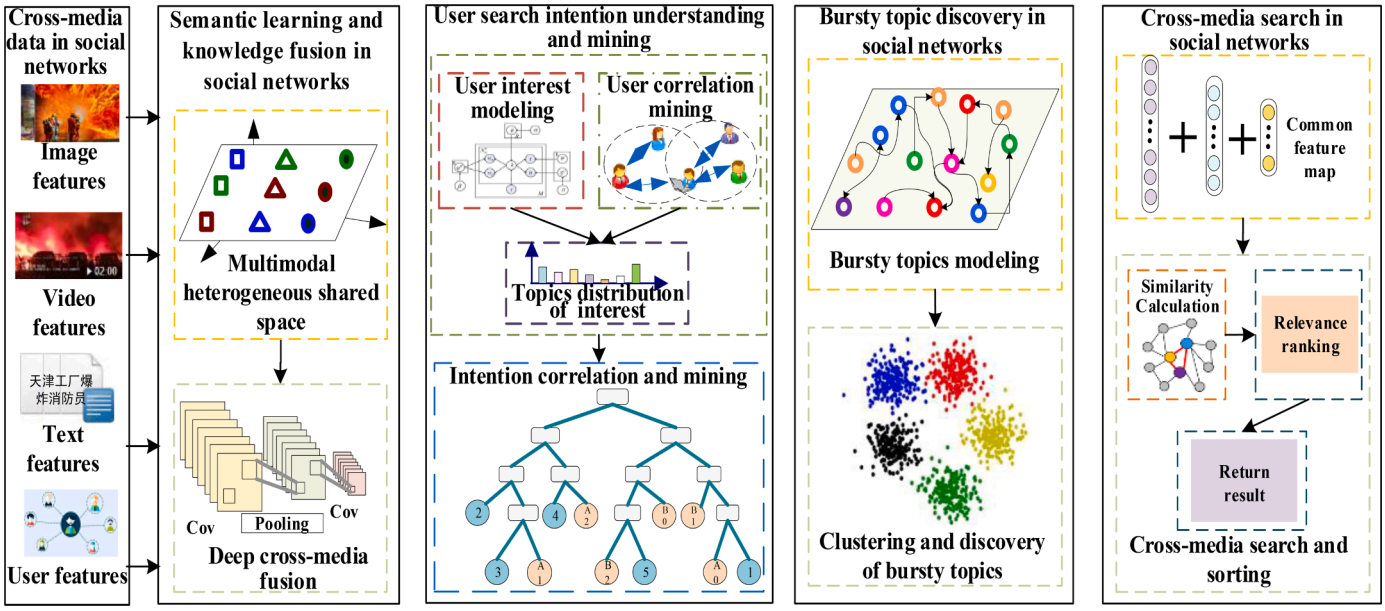

Fig. 1. The research framework of social network cross-media search based on user search intent understanding.

(2) We classify the full-cycle social network cross-media search methods through the layered progressive logic of the study and expound on the differences between different ways, which will help readers better understand the use of social network cross-media search. Various new technologies facilitate beginners to familiarize themselves with the field quickly.
(3) We analyze and summarize the data sets and evaluation indicators related to cross-media searches. We also selected typical cross-media search methods to conduct search experiments on public data sets. It can help researchers, especially students, to focus on algorithm design without spending more time investigating data sets, evaluation indicators, and benchmark methods.
(4) We summarize the challenges and opportunities in the social network cross-media search field and point out future development directions.

The rest of this paper is organized as follows: Section 2 summarizes the research status and progress of cross-media search based on user intent understanding in social networks. Section 3 summarizes the widely-used datasets for cross-media search. Section 4 presents evaluation indicators. Section 5 reports experiments. Section 6 presents a discussion and future trends, and Section 7 concludes this paper.

2. Research status and progress

For problems facing cross-media search on social networks, we discuss the theory and technology from four aspects: semantic learning and knowledge fusion, user search intention understanding and mining, social network bursty topic discovery, and social network cross-media search.

2.1. Semantic learning and knowledge fusion of cross-media big data in social networks

The cross-media data has various features, such as text, image, time, location, and users in social networks. By comprehensively using multiple features, the semantic learning of cross-media data can obtain a more accurate, consistent semantic representation and associate and fuse the learned knowledge in social networks. It can also provide the basis for subsequent intention understanding, topic discovery, and accurate search. Currently, the semantic learning methods for cross-media big data mainly include traditional and deep learning methods in social networks. The typical techniques of knowledge fusion are based on the knowledge graph. Table 1 summarizes the representative works of semantic learning and knowledge fusion for social networks.

2.1.1. Traditional method

Cross-media semantic learning mainly realizes the mapping of public semantic features by learning the semantic representation of different modals in social networks. In the traditional method of semantic learning, Yang et al. [3] used the shared self-interest-based classification model to interactively and directly train different modals to learn consistency embedding and obtain better cross-media consistency semantics. Kim et al. [4] used hypergraphs and attention to learn the public

Table 1
Representative works of semantic learning and knowledge fusion.

Categories	Typical research	Characteristics
Traditional method	SMIMM [3], HAN [4], PCME [5], SDMSA [6], SADTM [7], SMMTM [8], CITM [9], MFAF [10], MEPGM [11]	It can introduce multi-label information to learn the consistent semantic space of different modalities comprehensively. The time complexity is nonlinear with the number of samples, and the computational time complexity is too high in the face of massive data.
Based on deep learning	UniT [12], SkeletonNet [13], DR.-CNN [14], MM-GNN [15], CASC [16], S^3CA [17], AsymFusion[18], STGCM [19], HetEmotionNet [20], M3P [21], HFFCA [22], CSM-GAN[23]	It uses nonlinear mapping to learn the mapping relationship between different modalities and learns more fine-grained features through varying levels of deep network structures. The retrieval effect is better. The focus is still on the underlying global features, ignoring the primary, secondary, and central features.
Based on Knowledge Graph	CR^2[25], TransRHS [26], LCLDKG [27], UCKG [28], text2edges [29], RPJE [30], DihEdral [31]	It can integrate multiple attributes and modals, which is more suitable for data scenarios such as social networks. There are still problems of high noise and sparse semantics, and there is a lack of a systematic system

semantic representation of cross-media big data and effectively learned the semantic representation of multiple modals by constructing the public attention graph between subgraphs in the semantic space. Chun et al. [5] learned the probability representation of multimodal data in the embedded area to model the one-to-many association widely existing in cross-media data. They obtained a better cross-media semantic learning effect. Zhang et al. [6] used matrix decomposition and semantic auto-coding to learn potential semantic. They incorporated the label matrix into the loss function instead of the similarity matrix, effectively obtaining cross-media semantic information.

In addition, Shi et al. [7] constructed a dynamic self-aggregation topic model to enrich the semantic information by aggregating short texts into long documents. It realizes the dynamic semantic learning of short texts through incremental sampling and updating methods. Zhang et al. [8] proposed an unsupervised multimodal topic model to learn the cross-media semantic information in social networks. Various features and attributes accompany social network cross-media big data. By using various features and attributes for semantic learning of cross-media big data, we can obtain a more accurate semantic representation and support top-level research in social networks. Liang et al. [9] used users' preferences and social network context information to model users' dynamic preferences. Sahoo et al. [10] used the related social network user profiles, including communication information and user interaction information, to deeply study and analyze user behavior characteristics and apply them to false news detection. In addition, cross-media data contains rich contextual information, potentially including users' behavioral preferences in social networks. Extracting potential contextual information from social network cross-media big data to analyze users' behavior is a current research hotspot. Kou et al. [11] reduced the semantic space by constructing spatial-temporal regions combined with social information and alleviated the problem of social network context sparsity.

However, the primary purpose of traditional methods is to learn the discriminant shared subspace by maximizing the correlation, ignoring the local data structure in each modal, and the structure matching between modals. In addition, most traditional methods are linear mapping, which is difficult to effectively model the high-order correlation of different modals in social networks.

2.1.2. Based on deep learning

With the improvement of hardware computing power in recent years, semantic learning methods based on deep learning and graph neural network have gradually become a research highlight. Typically, Hu et al. [12] used the encoder of the transformer model to encode each input modal and used the shared decoder to predict the encoded input representation. It realized effective unified semantic learning and modeling cross-modal from different fields. Yang et al. [13] combined tensor decomposition and multi-view depth self-encoder to build a hybrid depth network for modeling low-level and high-level views. Zhang et al. [14] proposed encoding semantic context-aware representation based on multi-region CNN and realizing the expression of image information by learning the interactive characteristics of the background. Gao et al. [15] expressed the image as a graph composed of three subgraphs describing visual, semantic, and digital information. They used three aggregators to guide the transmission of messages for learning fine-grained semantic features. Xu et al. [16] proposed a semantic hybrid method of cross-modal semantic consistent attention based on image-text matching, which realizes the local alignment of cross-modal attention and the global semantic consistency of multi-label prediction. Yang et al. [17] learned the common semantic space between different modals by aligning the distribution between the activation layers in a specific modal network. Wang et al. [18] achieved an implicit fusion of modal by learning feature representation in the same network. They introduced asymmetric fusion operations to obtain a fine-grained semantic consistent representation of different fusion features. Song et al. [19] introduced a graph convolution network to learn the representation of spatial-temporal graphs for the coding of spatial-temporal interaction between different modals. Then they used Bert to dynamically encode the context of sentences for the joint spatial learning of cross-media data. Jia et al. [20] constructed a two-stream heterogeneous graph recurrent neural network to fuse spatial, spectral, and time-domain features in a unified framework for multimodal fusion. It realized the emotional recognition of multimodal physiological signals. Ni et al. [21] combined multilingual pre-training into a unified framework through multi-task pre-training and obtained consistent semantics in different modals. Zhang et al. [22] integrated the complementary information of cross-media segments through multiple deep interactions. They aggregated the obtained local features into an embedded vector, realizing the consistent association of cross-media semantics. Tan et al. [23] introduced a convolutional neural network to capture and highlight the features in the local visual text description and used cross-modal semantic matching to generate adversarial network model for improving cross-modal semantic consistency.

However, the existing semantic learning and knowledge fusion methods only target single-modal data or specific application scenarios and ignore the multi-attributes of social network cross-media information and related background and context information. It is challenging to understand comprehensively and deeply capture latent semantic information and user intent in social networks.

2.1.3. Knowledge fusion method based on knowledge graph

The knowledge graph is a branch of knowledge engineering. It takes the semantic network in knowledge engineering as the theoretical basis and combines the latest achievements of machine learning, natural language processing, and knowledge representation and reasoning [24]. Using the knowledge map, we can effectively fuse different attributes and features of social networks. Li et al. [25] used a best-first search algorithm to obtain entity-relationship subgraphs in the knowledge graph. They sorted the subgraphs based on entity-based documents and internal correlations, effectively enriching the relationship between entities. Zhang et al. [26] used the relative positions between vectors to model neutron relationships in knowledge graphs and obtained good results on link prediction and triple classification tasks. Wang et al. [27] adopted graph convolutional representation and contrastive learning methods for zero-shot learning. It effectively obtained the consistency of class representations from different knowledge graphs. Cao et al. [28] used unsupervised learning to construct a knowledge graph and associated source code entities with resource concepts based on word embedding and clustering techniques. Prokhorov et al. [29] proposed the mapping and fusion method of knowledge graph entities and realized the entity path prediction from the root node to the target node. Niu et al. [30] used the interpretability and accuracy of logical rules, knowledge graph embedding, and the supplementary semantic structure of paths to explicitly create entity relationship embedding. Xu et al. [31] proposed a dihedral symmetric group model to learn the embedding and association of knowledge graph and effectively capture the natural composition relationship of entities.

However, in the face of social networks' high noise and semantic sparseness, the fused content is sparse by knowledge graph. It lacks a systematic system, which makes it challenging to learn the fusion semantics and relevance of cross-media knowledge of social networks comprehensively and accurately. The comprehensive utilization of multimodal and multi-attribute information is needed to improve further cross-media semantic learning and knowledge fusion of social networks.

2.1.4. Discussion

The above methods carry out cross-modal semantic learning in the global feature space of different modal data without considering the salience attention mechanism to achieve more fine-grained feature semantic learning and knowledge fusion. In addition, the above method ignores the user's search intention and background preference in the

social network. Therefore, it is necessary to realize the fusion and association of multi-attribute and multi-modal information according to the semantic correlation and knowledge association between users and resources of social networks. In addition, cross-media big data contains multiple modal and contextual attributes in social networks. Extensive multi-modal and contextual features can further enhance cross-media relevance learning.

2.2. Understanding and mining of user intent of the social network

To realize an accurate cross-media search, it is necessary to establish a mechanism and algorithm for users' search intention understanding. Understand and mine users' search intention according to users' information, and finally return the search results that meet users' search intention. Therefore, using users and the posted content, understanding and mining users' search intention is of great significance to developing social network-related applications, such as social network precise search, topic clustering, and topic recommendation. The existing research on understanding and mining users' search intention mainly focuses on the methods based on the topic model, multi-attribute modeling, and comprehensive modeling of users' search intention in online social networks. Table 2 summarizes the representative works of understanding and mine users' search intention.

2.2.1. Based on topic model

The traditional topic model is used to mine and cluster the hidden topics of social networks. By understanding the topics of standard long documents and combining post-processing steps, the understanding and mining of users' search intentions are completed. However, the traditional topic model is designed to model the semantic information of standard news or long documents. Due to the sparse semantics and the lack of contextual word co-occurrence information, it cannot have a good effect on user search intention understanding and mining when applied to the social network context. To fully understand and mine users' search intention, many researchers improve the traditional topic model by solving the sparsity of social network context. Zuo et al. [32] used social network text's extracted word co-occurrence information to build a word network and directly modeled the word network using the topic model. The social network context sparsity problem was effectively solved by directly modeling word co-occurrence. Wang et al. [33] introduced hashtags as weak supervision information in the topic model, revealed the potential semantic relationship between words by learning the relationship between hashtags, and effectively solved the problem of social network context sparsity. Shi et al. [34] used the information of users and followers, combined with the user topic model, to mine users' search intentions and preferences in social networks. It effectively solves the problems of semantic sparseness and historical data. Yang et al. [35] proposed the Bayesian hierarchical topic model to solve the short text context sparsity problem. However, the above methods do not fully consider the user characteristics of online social networks, so they cannot accurately understand and mine users' search intentions.

Table 2
Representative works of understanding and mine users' search intention.

Categories	Typical research	Characteristics
Based on Topic Model	WNTM [32], HGTM [33], UATM [34], CTM [35]	It can effectively model social network text information with good interpretability and low complexity. The user characteristics of online social networks are not fully considered, so they cannot accurately understand and mine users' search intentions.
Based on user attributes modeling	DSM [36], ACN [37], AMEIR [38], UCIT [39], TGCN [40], MB-GMN [41], AOTM [42], SocialLGN [43], MPSR [44]	Integrating multiple attributes and modal information of social networks enables users' intentions and preferences to be effectively mined. Introducing multiple attributes will bring additional noise, semantic sparseness, and overfitting.
Based on historical data comprehensive modeling	CBSUM [45], HEE [46], CLM [47], CMTF [48], DTCAM [49], KPRN [50], KTUP [51], KHGT [52]	The accuracy of understanding and mining users' intentions is higher, and the results obtained are closer to users' ideas. It is heavily dependent on users' search history, click history, and other data, and these data are difficult to obtain.

2.2.2. Based on user attributes modeling

Users are the core attribute of social networks. Therefore, modeling combined with user information can effectively improve the performance of user search intention understanding and mining. Many researchers explore using users' social attributes to cluster to understand and mine users' search intentions. Qiu et al. [36] proposed a user clustering method based on the dynamic topic model, which realizes the clustering of user search intention by modeling the dynamic characteristics and social relationships of user topics in the social network text stream. Li et al. [37] adopted a transformer to obtain the interaction characteristics of users. They used the capsule network based on the attention to capturing a variety of interests from the user's behavior history. Zhao et al. [38] divided the recommendation model into three stages: behavior modeling, interactive exploration, and multi-layer perceptron aggregation, and introduced a new search space to realize a one-time random search and search result summary. Liang et al. [39] introduced the user's follower information to improve the performance of user search intention understanding and mining. Huang et al. [40] constructed a temporal graph convolution network to learn the node embedding of sequential semantics. They encoded the position information by interleaving cosine embedding and attention mechanism to capture the emotion-related information, realizing multimodal emotion recognition. However, the above method does not consider the association between words in social networks. It also ignores the impact of familiar words on users' understanding and mining of search intention.

Based on the multi-attribute and multimodal features of social networks, such as user relations and spatial-temporal information, some scholars use deep learning, graph neural network, and attention to mine users' intentions and preferences to realize personalized search and recommendations of users in different preferences and spatial-temporal situations. Xia et al. [41] incorporated the correlation of various types of user behavior into the collaborative filtering architecture to model the user's multiple behavior patterns. They realized the precise recommendation of learning interaction. Yang et al. [42] proposed an author co-occurrence topic model to model ordinary documents and their brief user comments and obtain individual user preferences based on author-level distribution. Liao et al. [43] used graph convolution networks to model users and projects and aggregated user representations based on graph fusion. Liu et al. [44] proposed a representation learning method based on a graph convolution network to construct hierarchical user preferences. They used the statistical information of historical user behavior to build a multi-angle graph network, and effectively mined users' intentions and preferences. However, this method can only realize the static modeling search, making it challenging to realize real-time in social networks.

2.2.3. Based on historical data comprehensive modeling

When using social network platforms, online social network users will generate many access logs, search history, and click records. This information can effectively understand and mine users' search intentions. Therefore, comprehensively adopting the user's search history,

access log, and click history to model the user's search intention has been a hot research area. Zhao et al. [45] mined the preferences and intentions based on behavioral factor decomposition in social media users. Shi et al. [46] combined user interest distribution and social network short text context to mine users' preferences and intentions and completed the dynamic understanding of users' intentions. Xun et al. [47] proposed a unified language model based on matrix decomposition to learn the intentions and preferences of social network users by considering complementary global and local context information simultaneously. Mehrotra et al. [48] use users' subject interests, search task behavior, and historical search behavior to learn users' preferences. Yin et al. [49] proposed a probability generation model to understand user intentions and preferences depending on the region. The sparsity of user sign-in records is solved by using the public's access behavior in the target region. However, the above method requires specific data and heavily depends on the relevant data generated by users, such as search history and click history, which is very difficult for researchers to obtain. Therefore, it is necessary to study a universal method that does not rely on user-related data to understand and mine users' search intentions effectively. In addition, in the current mainstream research, most ignore the dynamic diversity of users' intentions and preferences in social networks under different situations. Wang et al. [50] used knowledge graph to learn the semantics of social network entities and relationships to generate path representations and infer user preference information. Cao et al. [51] used knowledge graph to convey social relationships to understand user preferences at a finer granularity. Xia et al. [52] constructed a hierarchical graph neural network class based on knowledge enhancement to capture specific behavioral semantics, combined with multimodal graph attention mechanism and temporal coding strategy to complete the understanding based on user interaction behavior. However, the existing methods have not fully explored the connectivity of knowledge to infer user preferences, especially in the order dependence of paths and the path of overall semantics.

2.2.4. Discussion

Social network big data has multi-modal and multi-attribute characteristics. Users have different backgrounds and preferences, and users have similar social relations. Existing user intention perception methods do not consider user behavior, background information, and social relationship information, and it is difficult to obtain user status, intention, and preference information. Therefore, it is necessary to integrate the background and social relationship information of social network users, build a dynamic three-dimensional accurate portrait of users, and perceive the user's intentions and preferences in real-time, so as to make up for the shortcomings of traditional search and recommendation methods in user intention perception, knowledge association, and reasoning, and realize accurate search and intelligent recommendation oriented to the personalized needs of social network users.

2.3. Social network bursty topic discovery

Social networks attract many user groups as an essential platform for daily topic publishing. Ordinary users and many authoritative media release and disseminate the latest topics and news events through social networks, making social networking platforms generate a large amount of topic content every day. Some emergencies were first released through social network platforms, formed hot discussions, and sharing, and finally evolved into hot and cutting-edge topics on social networks. However, social networks generate a large number of sudden topics every day. Suppose users or relevant institutions want to understand the situation of a sudden topic quickly. In that case, they need effective methods to find the bursty topic and combine it with social network topic search to obtain relevant information quickly and automatically. The current mainstream research on bursty topic discovery can be summarized into two categories: bursty topic discovery methods based on the topic model and based on deep learning. Table 3 summarizes the representative works of social network bursty topic discovery.

Table 3
Representative works of social network bursty topic discovery.

Categories	Typical research	Characteristics
Based on Topic Model	STM [53], BBTM [54], TUS-LDA [55], ST-SRTM [56], ED-SWE [57], ST-ETM [58], DTE [59], SRTM [60], EVDE [61], Topicsketch [62]	It can integrate multiple attribute information to model the topic semantic information of social networks and carry out initial clustering of topics. The method has strong interpretability and low complexity. This method requires a complex post-processing process, and the findings may be mixed with general topics.
Based on deep learning	PP-GCN [63], Text-SimCLNN [64], KCN [65], DRMM [66], MVGAN [67], APC [68], TERAN [69], AEP [70], FinEvent [71]	Effectively learn the semantic association and representation of different modals of topics, obtain fine-grained semantic information, and the accuracy of topic discovery is high. Considering only the global semantics, it is difficult to model the multi-attribute information of social network topics.

2.3.1. Based on topic model

In the research of the topic model, use the topic model to model the semantic information of social network topics, and realize burst topic discovery through post-processing and clustering methods. Yan et al. [53] believe that the sudden nature of words can determine topics in social networks. Using the bursty nature of words as a prior, we can realize the automatic discovery of sudden topics. Shi et al. [54] adopted the topic model to model the burst of social network topics. They introduced sparse prior to focusing on the discovered burst topics, which achieved good performance in multiple topic discovery evaluation indicators. Xu et al. [55] introduced the emotional characteristics of the topic and the sudden calculation method of words based on the traditional topic model. They monitored the bursty topic through the bursty features. Zhu et al. [56] built a spatial-temporal topic model to model sudden topics in social networks and introduced sparse priors and RNN priors to optimize the bursty topics found jointly. Sun et al. [57] proposed an event detection model based on scoring and text embedding, which can detect emergencies from data streams. Dai et al. [58] introduced spatial-temporal regions and word pair information into the topic model, which can effectively alleviate semantic sparseness, and realized the automatic discovery of sudden topics by modeling the bursty nature of words in user information. Du et al. [59] introduced word embedding into the topic model to fully extract and integrate the text's topic distribution, semantic knowledge, and syntactic relationships and use classifiers to identify topics in the news automatically. Shi et al. [60] built a multi-feature topic model to model social network short text messages and introduced topic burst and RNN as a priori of the model, which can effectively obtain burst topics. Yang et al. [61] integrated a snapshot network topology index to quantify the structural characteristics of its sudden topics, and identified emergencies by mining network structure changes through the topic model. Then, they can judge whether there is an emergency by investigating the change degree of the network structure characteristics of adjacent snapshots. Xie et al. [62] found abnormal words by monitoring the temporal change characteristics of words in the data stream. They used the sudden nature of words as a prior to mixing them into the topic model for the real-time monitoring of sudden topics in social networks. However, the above methods require complex post-processing or multi-step operations such as clustering to discover emergent topics. It is challenging to realize the automatic discovery of bursty topics. In addition, the results of the above

methods are prone to serious overfitting. It is necessary to consider introducing more optimally related aggregation and clustering mechanisms to make the discovered topics more focused.

2.3.2. Based on deep learning

With the advantage of deep learning in the topic representation in social networks, many researchers have researched bursty topic monitoring and mining based on deep learning. It has become a research highlight in the current academic circles. Peng et al. [63] classified social network topics based on pairwise manifold graph convolution network with weighted meta-path instance similarity and text semantic representation. They realized dynamic detection and evolutionary discovery of topics using meta-path similarity search, meta-path historical information, and heterogeneous DBSCAN clustering methods. Tong et al. [64] constructed a text similarity comparative learning neural network to learn the similarity probability of text pairs from the perspective of semantics and structure. It effectively bridges the gap between text representation learning and similarity measurement in topic monitoring. Cao et al. [65] introduced incremental learning into event detection. They used a hierarchical distillation mechanism to learn previous knowledge from the original model, which effectively alleviated the semantic ambiguity and data imbalance problems in event detection. Tong et al. [66] used pre-trained Bert and RESNET to encode sentences and images to achieve mutual enhancement of modals, effectively improving the performance of social network public opinion monitoring. Cui et al. [67] described the emergencies by constructing a semantically compact heterogeneous graph of social networks. By learning the emergencies in a single view. In addition, a multi-view graph attention mechanism based on label guidance is designed to integrate the characteristics of emergencies in semantic and temporal perspectives to achieve efficient event detection. Yu et al. [68] monitored abnormal social network events by establishing the adversarial learning of the past and future. They realized the correlation between the current and future events in the training steps. Messina et al. [69] proposed a transformer encoder reasoning and calibration network, which performs fine-grained matching between the essential components of images and sentences, and realizes the accurate monitoring of topics. Yu et al. [70] established an anti-learning abnormal event prediction method for the potential feature space of joint learning in video and motion streams. They applied mutual information constraints to improve the recognition of learning representation, achieving good performance in detecting abnormal events. Peng et al. [71] proposed a deep reinforcement learning guided multi-graph neural network social network event monitoring method. They realize the dynamic monitoring of social network events by modeling social messages as learning social messages embedded by fusing rich meta semantics and strengthening weighted multi-graph neural network framework. However, these techniques focus more on the semantic properties of topics, associated data types, application scenarios, lack generalizability and portability rely disproportionately on particular scenarios and data, and largely ignore pertinent social network properties and background and context data. To overcome these issues, we must investigate the generalizable model to achieve the practical application of topic discovery and introduce the pertinent social network traits to direct the acquisition of more precise bursty topics.

2.3.3. Discussion

The above methods lack comprehensive utilization of social network users' behavior patterns and emotional characteristics. Therefore, it is challenging to realize the dynamic and accurate perception of bursty topics of cross-platform and cross-media big data and to track bursty topics with complex communication modes in real-time. To dynamically and accurately discover bursty topics, it is necessary to comprehensively utilize the behavior patterns of users and user emotion analysis methods in social networks. Combining the multimodal information, spatial-temporal information, and propagation path, we build an intelligent analysis and mining model of bursty topics in social networks and then achieve the intelligent and accurate discovery of bursty topics.

2.4. Social network cross-media search

Through cross-media search, different modal forms can be used as input to return the results of another modal, making the search results more vivid. For example, when the user enters a topic about the "Sichuan earthquake," if only the relevant text description about the "Sichuan earthquake" is returned, or the news from the news media, the user cannot visually understand the news extent of the disaster. The current research on cross-media search in social networks is mainly based on text and image. The key idea of cross-media search is to map different modal data to a public semantic space, measure the similarity of the two modals in the public semantic space, and get the best matching results back to the user.

The core problem of cross-media search in social networks is understanding users' search intention and spanning the semantic gap between two modals. The key to solving the semantic gap is to improve the quality of semantic representation between different modals and the quality of public subspace mapping. The cross-media search methods can be divided into three categories: cross-media search based on shallow semantics, cross-media search based on deep semantics, and cross-media search based on the cross-modal hash. Table 4 summarizes representative works of social network cross-media search.

2.4.1. Based on shallow semantics

The shallow semantic method mainly maps the features of different

Table 4
Representative works of social network cross-media search.

Categories	Typical research	Characteristics
Based on shallow semantics	GSS-SL [72], M3R [73], KDM [74], MVMLCCA [75], TNN—CCA [76], ml-CCA [77], SCSL [78]	The model structure is simple, and the operation efficiency is high. Unable to deal with nonlinear problems, the model has poor scalability, and it is challenging to learn high-order semantic information
Based on deep semantics	CCL [79], HRL [80], AGCN [81], RCBT [82] COTS [83], URL [84], GCR [85], ACMR [86], P-GNN—CON [87], DAGNN [88], AME-Net [89], CAGS [90], ITMeetsAL [91], CM-Gans [92], SC-ACMR [93], SSAL [94], DAML [95]	Use nonlinear mapping to learn the mapping relationship between different modalities and learn more fine-grained features through varying levels of deep network structures. Compared with methods based on shallow features, deep feature methods have better cross-media search performance. When applied to cross-media search, it is difficult to learn the association between two modalities comprehensively, and the central features of image modalities are ignored.
Based on cross-modal hashing	TSDH [96], MIAN [97], CMIMH [98], GCDH [99], DSAH [100], DDSJH [101], CPAH [102], ATFH—N [103]	It can learn a more accurate unified hash code of multi-modalities and easily expand to more modalities, which effectively improves the efficiency of cross-media search and is more suitable for cross-media search scenarios with massive data. By compressing the original data into binary hash coding for Hamming space search, in the process of compression and coding, the original features of the data are lost more, and the accuracy of cross-media search is sacrificed while improving efficiency.

shallow modalities to a common subspace through traditional subspace learning techniques or associates two modalities through other methods. In the shallow semantic method, Typical definitions are as follows. For two views after the mean value is removed $X = [x1, x2, x3, \cdots, xN] \in R^{dx \times N}$ and $Y = [y1, y2, y3, \cdots, yN] \in R^{dy \times N}$, By finding two projection directions $wx \in R^{dx \times 1}$ 和 $wy \in R^{dy \times 1}$, Maximize the correlation coefficient of $X^{\mathrm{T}}wx$ and $Y^{\mathrm{T}}wy$ to achieve cross-modal correlation. As shown in Eq. (1):

$$\underset{w_x, w_y}{max}\rho\left(X^{\top}w_x, Y^{\top}w_y\right) = \frac{w_x^{\top}S_{XY}w_y}{\sqrt{\left(w_x^{\top}S_{XX}w_x\right)\left(w_y^{\top}S_{YY}w_y\right)}} \tag{1}$$

where N represents the number of samples, d_x and d_y represent the characteristic dimensions of view X and view Y, respectively, and S represents the covariance matrix.

The main advantage of the above definition is that different semantic spaces can be mapped into public semantic spaces through simple projection, which is clear and relatively low in complexity. Shallow semantic methods mainly include the topic model and subspace learning method. Zhang et al. [72] leveraged use a label graph constraint to ensure that the internal geometric structure of different feature spaces is consistent with the internal structure of the label space. They also use tag space as a link to model the correlation between different modes. Wang et al. [73] proposed a supervised multimodal topic enhancement modeling based on the probabilistic topic model, which learns the potential representation of multimodal data by introducing topic interaction and label information to improve the quality of semantic learning. Xu et al. [74] proposed a multi-label cross-media search framework combining semantic association and subspace learning, which preserves the semantic consistency between modals by using the relevant information between multiple labels. Sanghavi et al. [75] proposed a multi-view and multi-tag canonical correlation analysis method. Through the high-level semantic information provided in the form of multi-tag annotation in each view, the semantic relationship between multiple views is established in the multi-tag annotation, and effective cross-media search performance is obtained. Zeng et al. [76] proposed an end-to-end supervised learning network structure based on clustering CCA for audio-video retrieval. Shu et al. [77] obtained consistent semantic information by learning the common semantic space of the two-modal data. Xu et al. [78] proposed a cross-modal subspace learning model based on kernel dependency maximization, which learns the subspace representation of each modal by maximizing kernel correlation. Although the above shallow cross-media search methods can consistently correlate data of different modals, simple probability generation or linear mapping of data of different modals cannot effectively learn the matching information within and between modals, and cannot learn higher-order semantic information.

2.4.2. Based on deep semantics

The core idea of the cross-media search method based on deep semantics is to improve the effect of feature learning by using the deep learning method to learn complex high-level features. Typical definitions of such problems are as follows.

The image dataset is expressed as $\{I_{c,i}\}$, where $x_{c,i}$. C is the number of category labels, n_c is the number of image samples in category c, and n is the total number of image mode samples. For the original image sample $I_{c,i}$, we want to learn its characteristics $\varphi_I(\theta_I, I_{c,i})$ in this modal and input this feature into the projection layer network $f_i(w_i)$ to obtain its characteristics $x_{c,i}$.

$$x_{c,i} = f_I\left(\omega_I, \varphi_I\left(\theta_I, I_{c,i}\right)\right) \tag{2}$$

Similarly, $y_{c,t}$ can be jointly characterized by text data $T_{c,t}$ learning, as shown in Eq. (3):

$$y_{c,t} = f_T\left(\omega_T, \varphi_T\left(\theta_T, T_{c,t}\right)\right) \tag{3}$$

$$\varphi_T\left(\theta_T, T_{c,t}\right) = ReLU\left(w_2 ReLU\left(w_1 T_{c,t} + b_1\right) + b_2\right) \tag{4}$$

where $c = 1, 2, 3, \ldots, C; \cdot t = 1, 2, 3, \ldots, m_c$, w_1 and b_1 are the network parameters of the first full connection layer in the text network, w_2 and b_2 are the network parameters of the second full connection layer in the text network. ReLU $(\bullet)$ is the activation function after the full connection layer

To alleviate the semantic gap, the common representations X and Y are transformed into the probability distribution of labels in each category as the common semantic representation. For a common representation z (which can be $x_{c,i}$ or $y_{c,t}$), its semantic feature representation can be obtained by the softmax function transformation.

Assume z belongs to category c, then the c element of l is 1, and the rest is 0. If the data of image and text modals are subject to semantic consistency, the loss of semantic consistency can be defined as Eq. (5):

$$L_{\{S\}} = L_{\{I\}} + L_{\{T\}} = \sum_{c=1}^{c}\sum_{i=1}^{n_c} L_{ce}\left(x_{c,i}\right) + \sum_{c=1}^{c}\sum_{t=1}^{m_c} L_{ce}\left(y_{c,t}\right) \tag{5}$$

where $l = \{0, 1\} \in R^C$ is the category of z, expressed as one-hot form.

The advantage of this kind of problem definition is that it can learn nonlinear fine-grained semantic representations of different modals through complex deep learning and conduct semantic mapping through public semantic learning, and then achieve accurate cross-media search. In the deep semantics, Peng et al. [79] proposed a multi-granularity fusion cross-modal correlation learning method based on a hierarchical network, which uses the multi-level correlation of joint optimization to keep the complementary context free from the influence of intra-modal and inter-modal correlation. Cao et al. [80] effectively mapped between modals by stacking bimodal auto-encoders to obtain a common representation of each modal. Dong et al. [81] reconstructed sample representations using graph convolutional networks and reconstructed node feature based on local graphs to map the features of the two modalities to a common space, obtaining hidden high-level semantic information. Peng et al. [82] captured cross-media correlations by bidirectional translation training between bilingual pairs of visual and textual descriptions, providing complementary cues for learning cross-media associations. Lu et al. [83] proposed a collaborative dual-stream visual language pre-training cross-media search method, which improves cross-media search performance by using token and task-level interaction to enhance cross-media interaction. Cheng et al. [84] built a unified framework to learn cross-media alignment in public space and combined it with the optimization paradigm of ranking learning to achieve effective cross-media search. The above methods based on deep learning have greatly improved the search performance compared with shallow learning methods, mainly due to deep methods' excellent feature learning ability, representation, and computing ability. However, deep semantic learning methods still need to use the feature representation of their respective modals. They focus on the underlying global features, ignoring the primary and secondary features and the focusing features of the image.

With the development of adversarial learning, many researchers apply adversarial learning to cross-media search based on deep semantic learning, which effectively improves the performance of cross-media search. The main idea of adversarial learning is to get high-quality common semantic representation by the adversarial and game of the generative model and discriminant model. Zhang et al. [85] constructed a semantic constraint learning framework based on graph neural network to learn the image-text semantic matching relationship, combined with cross-media adversarial learning to improve the identification ability further and realize reliable cross-media retrieval. Wang et al. [86] used generative adversarial networks to learn the feature representation of modals. In the generative process, the feature projection

was used to generate the modal invariant representation of the common subspace. In the discrimination process, different modals were distinguished according to the generated representation, effectively improving the cross-media search performance. Wu et al. [87] added tag information to adversarial training, and the model trained in one direction achieved better search performance in the opposite direction. Qian et al. [88] introduced a supervised cross-modal retrieval network composed of double generative adversarial networks and a multi-hop graph neural network. It can effectively obtain label information, construct the relationship between labels, and learn the corresponding classifier, to reduce the heterogeneity between different modals in the public space to a certain extent. Han et al. [89] used an adversarial learning strategy to use local relations and global features in heterogeneous modal information to learn a better joint embedding space further to narrow the semantic gap between different modal representations. Shi et al. [90] combined adversarial learning and complementary attention to learn the semantic consistency of cross-media information and obtained wonderful performance in cross-media search. Chen et al. [91] introduced calculating information entropy to evaluate the uncertainty of the modal classification. Eliminating modal heterogeneity is an essential task of cross-modal retrieval. To meet this challenge, Peng et al. [92] proposed a cross-modal Gan structure to simulate the joint distribution of different modal data. The correlation of heterogeneous data is effectively realized by exploring the correlation between modals and within modals simultaneously in the generation and discrimination model. Ou et al. [93] used adversarial learning to obtain the public semantic representation of cross-media data and introduced semantic consistency restrictions to improve the quality of semantic representation. Wang et al. [94] proposed a self-supervised adversarial learning method, using unlabeled rotating image data to train the model, constraining the learning process through self-supervised learning and obtaining better search performance. Xu et al. [95] used adversarial learning to distinguish the data representation within and between modals, effectively improving cross-media data's common semantic representation quality. However, the current cross-media search methods based on adversarial learning are still based on global features and do not consider the impact of focused and unfocused features. Moreover, the above approach ignores the importance of users' search intention in social networks.

2.4.3. Based on cross-modal hashing

The main idea of cross-modal hash is to map the data of different modals to a common binary hamming space, calculate the similarity between different modal, and realize a cross-media search. Typical definitions of issues related to cross-modal hashing are as follows.

First, convert the multi-dimensional feature vector $X \in R^{d \times n}$ into the corresponding k-bit hash code $z=\{z1, z2, \ldots, zk\}$, and obtain from the corresponding hash function using Eq. (6):

$$y = h(x) = \operatorname{sgn}\left(\sum_{t=1}^{T} \omega_t K(s_t, x) + b \right) \tag{6}$$

When $z \geq 0$, sgn (z)=1, on the contrary, sgn (z)=-1, $\{S_t\}$ represents the randomly selected classical sample, $\{\omega_t\}$ represents the weight value w is the projection vector, and b is the offset variable. In Hamming space, hamming distance d_{ij}^h is used to describe the distance between hash codes y_i and y_j, and Hamming distance is the corresponding different digits between the two hash codes $z \geq 0$:

$$d_{ij}^h = \sum_{m=1}^{M} \delta\left[y_{im} \neq y_{jm} \right] \tag{7}$$

The definition of this problem is that mapping the original code to the compact Hamming space can improve search efficiency, reduce time and space complexity, and is suitable for large-scale search scenarios.

One of the most significant characteristics of cross-modal hash is to encode different modal data into binary codes, which effectively improves the efficiency of cross-media search and is more suitable for cross-media search scenarios with massive data. Zhang et al. [96] proposed a hierarchical supervised discrete hash method. By mapping the potential modal representation to the public Hamming space, the recognition ability to learn binary codes is enhanced. Combined with the discrete hash optimization method, the quantization loss of cross-media hash is reduced. Zhang et al. [97] proposed a modal invariant asymmetric network architecture, which takes pairing, segmentation, and transformation semantics into a unified semantic preserving hash code, and realizes the similarity preservation within and between asymmetric modals under the framework of probabilistic modal alignment. Hoang et al. [98] proposed a maximum hash method of cross-modal information, which realizes the unsupervised learning of binary hash code by using the maximum mutual information method and efficient cross-modal retrieval. Bai et al. [99] proposed an approach based on graph convolution network (GCN) discrete hash, which uses GCN to represent each tag as a word embedding, and combines a discrete optimization strategy to learn discrete binary codes to enhance cross-modal feature representation. Yang et al. [100] aligned the similarity between different modal features and the similarity between hash codes through the semantic alignment loss function and used the hash code of one modal to reconstruct the characteristics of the other modal, realizing the unsupervised cross-modal accurate search. Zhang et al. [101] proposed a deep discriminant feature learning method for cross-modal semantic understanding. It uses mutual information to exchange features in semantic space and paired features, calculates the loss between semantic space and hash space, and realizes the collaborative utilization of corresponding information in cross-modal data. Xie et al. [102] used adversarial learning to obtain consistent representations of different modals, combined with cross-modal depth hashing to achieve an effective cross-modal search. Liu et al. [103] used adversarial training to alleviate the semantic gap and data imbalance, combined with the tag prediction network to guide feature learning and hash code learning, and achieved good results on multiple cross-media search tasks. However, by compressing the original data into binary hash coding for hamming space, the original characteristics of the data are lost more in the process of compression coding, which improves the efficiency and sacrifices the accuracy of cross-media searches.

2.4.4. Discussion

Most of the above methods focus on solving the semantic gap of modal space information by learning the features of different modes and mapping them to a unified semantic space. Additional similarity calculations are used to realize cross-media searches. It ignores the search intentions of users in social networks and the different attributes and background information accompanying social networks. More intelligent search results can be obtained by integrating user intention, multi-attribute, multi-feature information, context, and background information for semantic reasoning and intention perception. Therefore, it is necessary to research how to use different attributes and background information of social networks to align and sum and combine the understanding and mining of user search intention and topic information. It achieves a cross-media accurate search of social networks based on user search intention understanding.

3. Public datasets

As a current trend, cross-media search attracts the attention of many researchers, so more cross-media search datasets for different task applications and research have been opened up. It is worth noting that the research on cross-media search of social networks based on understanding users' search intention has developed for a short time. Therefore, there are few public datasets for cross-media search research on social networks, and the amount of data and modal types are insufficient. Next, we introduce several standard public datasets of cross-media

search. Table 5 summarizes the representative public dataset in cross-media search.

(1) IAPR TC-12 [104]: The dataset consists of 20,000 image-text pairs, including photos of different movements and movements and photos of people, animals, cities, landscapes, and many other aspects of contemporary life. Each image is paired with annotations in English, German and Spanish. However, since the relevant data are collected from non-field travel companies, there may be bias in the data-related information.
(2) NUS-WIDE [105]: The dataset was collected from Flickr by the national university of Singapore media research laboratory through a web crawler. It mainly includes images and their corresponding image tags, including 269,648 images. After data cleaning, there are 5108 independent labels, and each image includes about six labels on average. This dataset has apparent advantages in data volume but still contains only two modal types: image and text.
(3) Wikipedia [106]: The dataset was collected from Wikipedia and is one of the most used datasets in cross-media search research. It contains 2866 image-text data pairs comprising 10 semantic categories. Because the dataset was published earlier, it is not sufficient in semantic categories. In addition, it only includes two different modals: image and text.
(4) PASCAL VOC [107]: The dataset consists of 9963 images and 24, 640 annotation objects, which are divided into 20 different categories. The notes mention classes, bounding boxes, views, truncated entities, and complex entities. The main disadvantage of this data set is that it is not rich in categories. In addition, the data set was initially used for classification and detection tasks, not cross-media search.
(5) Wiki-CMR [108]: The dataset is mainly concentrated on geography, humanities, nature, culture, and history. A total of 74,961 documents contain images, paragraphs, hyperlinks, or category labels. Documents are divided into 11 different semantic categories. Images are represented by eight features, including dense filtering, gist, Phog, LBP, and other features, and texts are represented by TF – IDF.
(6) Flickr30k [109]: The dataset is created from the Flickr website, and the data is mainly on some actions of people or animals, which can be used for images and extended text. It contains 31, 783 daily images and 158,915 related subtitles. The dataset has a high degree of matching between pictures and texts. However, the data set lacks specific category information and content diversity.
(7) MS COCO [110]: The dataset comprises 328,000 images and 250, 000 pictures of daily scenes of marked instances, including 91 different categories, and each photo has five corresponding comments. Annotation is divided into three types: concept location in the marked image, all instances of the marked concept, and segmentation of each object instance. The initial design of this dataset was mainly used for target detection, target recognition, and target segmentation. Later, researchers used it for cross-media search tasks.
(8) PKU XMEDIA [111]: The dataset comprises 5000 texts, 5000 images, 500 videos, 1000 audio clips, and 500 3D models. The data sources include Wikipedia, Flickr, YouTube, etc., a total of 20 categories, such as bicycles, pianos, insects, etc., and each class has 600 media instances. This dataset is rich in modal information and comes from various sources. However, this dataset is still not rich in categories.
(9) PKU XMEDIANET [112]: The dataset is divided into 200 categories and five modal types. The file formats are TXT, JPG, avi, WAV, and obj, and the data volume is 40,000, 40,000, 10,000, 10,000, and 2000 in turn. Data sources include Wikipedia, Flickr, YouTube, etc. This dataset is currently the largest cross-media data set containing five media types worldwide, and its scale is 10 times that of xmedia. Compared with PKU XMEDIA, this dataset significantly improved the magnitude and category of the data set. However, the proportion of different modals in this data set varies greatly.
(10) Twitter-100k [113]: Twitter-100k contains 100,000 image text pair data randomly crawled from Twitter, including image description, emotion, and comment information, and there are no restrictions on image categories. The data set is used for weakly supervised cross-media search learning and tests the retrieval method's robustness under massive data. Its specific feature is that it has rich data, which can avoid overfitting in the training process, and the platform of the disadvantage source is too single.
(11) M5product [114]: The dataset contains 6 million multimodal samples, including five different modal types such as image, text, table, video, and audio, with coarse-grained and fine-grained annotations of 1 million merchants for electronic goods, 6 million category annotations, including more than 6000 categories, 5000 attributes, and 24million values, which is 500 times larger than the most significant publicly available dataset with similar modal numbers. Its advantage is that there is a large amount of data. Still, its disadvantage is that the data is related to e-commerce, more attention is paid to commodity and merchant information, and there is a lack of non-commodity support.
(12) Wukong-100 M [115]: The dataset is the first large-scale open--source cross-media search data set in the Chinese field, which contains about 100 million image text pairs from the Internet. The data is collected from around 200,000 essential keywords, and 1000 samples are reserved at most for each basic keyword. Its advantage is that the amount of information is large, which is suitable for training complex models, but its disadvantage is that the types of modals are not rich enough.

4. Evaluation indicators

Common evaluation indicators of cross-media search include

Table 5
Representative public datasets in cross-media search.

Name	Category	Images number	Image text scale	Year	Modal type
IAPR TC-12	various	20, 000	1:1	2006	Image/multilingual text
NUS-WIDE	81	270, 000	1:1	2009	Image/label
Wikipedia	10	2, 866	1:1	2010	Image/text
Pascal VOC	20	9, 963	1:1	2010	Image/label
WIKI-CMR	11	74,961	1:1	2013	Image/paragraph/hyperlink
Flickr30k	various	31, 783	1:5	2014	Image/sentence
MS COCO	91	328, 000	1:5	2014	Image/comment
PKU XMedia	20	5, 000	1:1	2015	Image/text/video/audio/3d
PKU XMediaNet	200	40, 000	1:1	2017	Image/text/video/audio/3d
Twitter-100K	various	100,000	1:1	2017	Image/broken/comment/emotion
M5Product	6000	6, 000, 000	–	2022	Image/text/table/video/audio
Wukong	various	100, 000, 000	1:1	2022	Image/text

average accuracy (MAP) [116], precision-recall curves [103], and precision scope curves [117] in social networks. Next, we will introduce them in detail.

(1) The mean average precision (MAP)

The mean average precision (MAP) is used to evaluate the search results, mainly to measure whether the searched modals are consistent with the query modal categories. It is often used to evaluate the performance of cross-modal retrieval algorithms. Ideally, the higher the ranking of the search results, the more significant the correlation between the samples and the query samples, the better. The MAP can better reflect the retrieval effect. The specific calculation is shown in Eq. (8) and Eq. (9):

$$MAP = \frac{1}{C}\sum_{c=1}^{C} AP(C) \tag{8}$$

$$AP = \frac{1}{T'}\sum_{t=1}^{T} prec(t)\delta(t) \tag{9}$$

Where C represents the number of queries in the query set, T and T' are the total number of results and the number related to the query respectively, $prec(t)$ is the accuracy of the t-th result, $\delta(t)$ is an indicative function, $\delta(t) = 0$ means that the returned result is irrelevant, $\delta(t) = 1$ means that the returned result is related.

(2) Precision-recall curves

Accuracy rate and recall rate affect each other and sometimes contradict each other. Ideally, it must be high in both. But generally, a high accuracy rate and low recall rate, low recall rate and high accuracy rate can be achieved. By using different thresholds, the accuracy rate and recall rate under a group of different thresholds can effectively reflect the effect of cross-media searches on social networks. In precision-recall curves, the abscissa is the recall rate, and the ordinate is the accuracy rate. The closer the P-R curve is to the upper right corner, the better the model effect is. The area of the P-R curve is average precision.

(3) Precision-scope curves

Similar to the P-R curve, the abscissa represents the scope, that is, the range of search samples, and the ordinate represents the accuracy.

5. Experiments

The search experiment is carried out by text search image, image search text between modals, and search average value. The public cross-media datasets Wikipedia and NUS-WIDE are used as experimental data. The mean average precision (MAP) evaluation metric is used to verify and evaluate the results of different cross-media search algorithms. For cross-modal hash, we set the length of the hash code to 64 bits on Wikipedia and NUS-WIDE datasets, the comparison of MAP values of different algorithms is shown in Table 6.

The performance of the deep learning based cross-media search algorithm and the cross-modal hash algorithm on Wikipedia and NUS-WIDE datasets is better than that of the shallow learning algorithm. This result shows that the cross-media search algorithm based on deep semantics and cross-modal hashing can more fully learn the feature representation and more consistent semantic information between different modals. The cross-media search algorithm based on deep semantics has obtained better mean average precision. It can effectively improve the common semantic learning ability and obtain the best cross-media search results.

In addition, as seen from Table 6, the MAP value of each method on

Table 6
Comparison of MAP of different algorithms on Wikipedia and NUS-WIDE datasets.

Datasets	Categories	Methods	Text2Image (MAP)	Image2Text (MAP)	Avg MAP
Wikipedia	Shallow semantics	GSS-SL [72]	0.33	0.39	0.36
		M3R [73]	0.37	0.23	0.30
		KDM [74]	0.48	0.46	0.47
		ml-CCA [75]	0.43	0.56	0.50
		SCSL [78]	0.42	0.41	0.42
	Deep semantics	CCL [79]	0.46	0.50	0.48
		HRL [80]	0.67	0.65	0.66
		AGCN [81]	0.73	0.62	0.68
		URL [84]	0.50	0.56	0.53
		GCR [85]	0.49	0.54	0.52
		ACMR [86]	0.49	0.62	0.56
		DAGNN [88]	0.53	0.66	0.60
		CAGS [90]	0.51	0.59	0.55
		CM-Gans [92]	0.47	0.52	0.50
		SSAL [94]	0.44	0.48	0.46
		DAML [95]	0.48	0.56	0.52
	Cross-modal hashing	TSDH [96]	0.77	0.39	0.58
		MIAN [97]	0.69	0.52	0.61
		CMIMH [98]	0.76	0.61	0.69
		GCDH [99]	0.68	0.62	0.65
		DSAH [100]	0.71	0.70	0.71
		DDSJH [101]	0.66	0.71	0.69
		CPAH [102]	0.72	0.61	0.67
		ATFH—N [103]	0.70	0.33	0.52
NUS-WIDE	Shallow semantics	GSS-SL [72]	0.40	0.54	0.47
		M3R [73]	0.24	0.30	0.27
		KDM [74]	0.28	0.35	0.32
		ml-CCA [75]	0.83	0.83	0.83
		SCSL [78]	0.38	0.41	0.40
	Deep semantics	CCL [79]	0.68	0.67	0.68
		HRL [80]	0.60	0.60	0.60
		AGCN [81]	0.63	0.61	0.62
		URL [84]	0.61	0.63	0.62
		GCR [85]	0.54	0.54	0.54
		ACMR [86]	0.54	0.54	0.54
		DAGNN [88]	0.76	0.75	0.76
		CAGS [90]	0.72	0.71	0.72
		CM-Gans [92]	0.68	0.69	0.69
		SSAL [94]	0.62	0.66	0.64
		DAML [95]	0.53	0.51	0.52
	Cross-modal hashing	TSDH [96]	0.79	0.67	0.73
		MIAN [97]	0.76	0.80	0.78
		CMIMH [98]	0.79	0.80	0.80
		GCDH [99]	0.75	0.74	0.75
		DSAH [100]	0.80	0.82	0.81

(continued on next page)

Table 6 (*continued*)

Datasets	Categories	Methods	Text2Image (MAP)	Image2Text (MAP)	Avg MAP
		DDSJH [101]	0.67	0.63	0.65
		CPAH [102]	0.67	0.63	0.65
		ATFH—N [103]	0.72	0.62	0.67

the Wikipedia dataset is significantly higher than that on the NUS-WIDE dataset, and there is a significant difference. The main reason is that a paragraph of text describes an image in the Wikipedia dataset, while an image in the NUS-WIDE dataset is described by multiple tag information. In contrast, the algorithm is easier to extract effective text features from the sentences in the NUS-WIDE dataset, and there are relatively few redundant interference words, thus improving the algorithm's accuracy.

Combining the characteristics of the method through the experimental results, we analyzed the advantages and disadvantages of existing methods. In addition, we provide a series of guidelines or insights on how to select the existing methods. The advantages and disadvantages of analysis and selection guidance for the existing methods are shown in Table 7.

6. Discussion and future trends

In recent years, from the traditional method to the deep learning, the performance of cross-media search continues to refresh historical achievements and obtain excellent performance. However, with the rapid development of different forms of social networks, in real life and applications, users are more eager to obtain cross-media search results that meet their intentions and preferences. The traditional cross-media search scenarios have been challenging to meet users' expectations. Therefore, cross-media search is a trend that needs in-depth research and attracts more and more researchers to the social network environment. In this section, we mainly put forward the following thoughts on some urgent research interests in the future.

(1) Massive cross-media information collection and processing in social networks. Due to the vast amount and dynamic changes and related data, it is an urgent problem to establish a large-scale cross-media streaming data processing mechanism for social networks. In addition, in view of the user-centered, multi-attribute, and multi-modal characteristics of social networks, most of the existing datasets are oriented to single image-text matching information, lacking effective data that contains multiple attributes, multiple modalities, and have a certain specificity. A large number of social networks are validated cross-media benchmark datasets. There is a big gap between the application level and the actual situation of social networks. It is challenging to effectively describe social networks' real environment and modal representation. Therefore, we must combine distributed technologies such as cloud computing to process and analyze social network cross-media data intelligently. Establish social network cross-media datasets with multiple semantic categories, rich modal types, multiple attributes, and other practical applications to provide favorable research conditions for cross-media search tasks and practical applications in social networks.

(2) The balance between accuracy and efficiency of cross-media search. In cross-media search algorithms proposed in recent years, most build complex models and computing frameworks based on deep learning and other methods, which has significantly improved the accuracy of cross-media search. However, such algorithms and models have high complexity and requirements for basic computing power. When applied to natural scenes and practical applications, they face low search efficiency and time-consuming problems, which challenges using relevant models. The cross-modal hash converting the consistent semantic features into hamming space based on binary coding has achieved excellent performance in cross-media search efficiency. Still, coding into binary hamming code inevitably causes a loss of search accuracy. Therefore, in future research on cross-media search for social networks, we need to focus on the balance between the accuracy and efficiency of cross-media search and use a lightweight hybrid model to improve the semantic correlation between different modal data, to achieve the accuracy and efficiency of cross-media search in social networks.

(3) Semantic modeling of cross-media consistency with multi-attribute and multimodality in social networks. In the current research, different modal data are mapped to the public semantic space through deep semantic uniform representation when learning social networks' cross-media consistent semantic representation. Then the direct cross-modal measurement is carried out in this space. However, the above method is too rough when modeling multimodal joint representation, and it is difficult to obtain the consistent semantics of different modals in fine-grained. Recently, a series of fine-grained semantic consistency modeling methods have been proposed to effectively mine the corresponding relationship between image and text segment levels and obtain better cross-modal correlation modeling results. However, the key to cross-media search is to establish the association between different modal, and mine the user-centered fine-grained semantics combined with social networks' social relations and attention relations. Therefore, it is necessary to establish consistent semantic information and association from the perspectives of semantic class association between modals, modal symbiosis association, multi-attribute, and multi-situational association. In addition, we can obtain more efficient, consistent semantics by introducing users' emotional information, background, and situational information, getting higher-level semantics in cross-media information, and using strong or weak supervised learning to focus on fine-grained features and carrying out consistent semantics according to attention mechanism and adversarial learning.

(4) Cross-network, cross-platform, cross-media big data association, and integration. In the current research on cross-media search, most of the research is aimed at a single data source or a single social network platform, such as Twitter, Sina Weibo, and flicker social networks. Cross networks, cross-platform, and cross-media big data are complementary, integrating network information from different social network platforms. We can get a complete development track of relevant information through forwarding, sharing, and other functions between platforms to comprehensively obtain relevant information. To fully compensate for the semantic sparsity and cross-media semantic gap of specific information on a single platform, we must think about how to carry out fine-grained association and fusion of the generated big data of cross-network and cross-platform.

(5) The accurate understanding of users' intentions in social networks. The massive, multi-source, heterogeneous, and polymorphic data is full of semantic ambiguities, which makes it difficult to penetrate and understand users' real intentions. It is necessary to use the semantic association between multiple attributes and multimodal, combined with the context, spatio-temporal characteristics, scene perception, action emotion, and other ways of user requests, to achieve an accurate understanding of users' intentions at the semantic level. The specific difficulties and challenges include:1) User intention modeling supporting spatio-temporal characteristics. For various time and space information of online social networks, it is necessary to study the user behavior pattern analysis and mining technology based on

Table 7
The advantages and disadvantages analysis and selection guidance for the existing methods.

Categories	Methods	Advantages	Disadvantages	Selection guidance
Shallow semantics	GSS-SL [72]	Using labels as link pairs to improve the correlation between different modals	Need to utilize the label information in the pair association	Suitable for small-scale data with a complete paired relationship
	M3R [73]	Bayesian modeling process is used, which is interpretable	There are strong assumptions about the distribution of cross-media topics, which do not conform to the actual application environment.	Applicable to situations with supervision and multi-attribute information
	KDM [74]	The subspace representation of modals is learned through kernel maximization, which can maintain kernel semantic similarity	There are many parameters involved, and the efficiency is not high	Applicable to small-scale data scenarios with low-efficiency requirements
	ml-CCA [75]	Using multi-tag information to learn the common semantic space can improve the correlation of modals	Large-scale training samples lead to high computational time complexity	It is suitable for multi-label data, especially for non-one-to-one cross-media search tasks
	SCSL [78]	The consistency between the modal similarity of each modal can be maintained	There are many parameters involved, the algorithm structure is complex, and the efficiency is not high	Multi-label cross-media search task applicable to small-scale situations
Deep semantics	CCL [79]	Fine-grained semantic representation can be learned by using the relationship between modals	A large number of parameters need to be set, and the learning time is long	Applicable to scenarios with supervision and low requirements for operation efficiency
	HRL [80]	Weight learning based on likelihood estimation can learn more complex features	Adaptability to long text search data sets needs to be improved	Suitable for the image to text search scenarios
	AGCN [81]	The complementary semantic information is used to enhance similar information with the same semantic samples	It needs to use paired modal information and requires high data	It applies to the retrieval of text image pairs in the case of large-scale data
	URL [84]	Using the interaction between images and text, semantic relationships can be effectively learned	Lack of structural comprehensiveness among different modalities, requiring a lot of repetitive training	Scenarios suitable for large-scale common modal data
	GCR [85]	Ability to learn the intrinsic structure of training data and enable reliable cross-media search	The process of building graphs is complex and requires large-scale data	Suitable for large-scale labeled data scenarios
	ACMR [86]	Using the adversarial learning mechanism, more effective features can be generated	Intra-modal classification loss barely works when using pre-trained models	It is suitable for application scenarios with low search efficiency requirements and multiple tags
	DAGNN [88]	Ability to effectively learn cross-media semantic information	Lack of data on local structure within modalities and semantic class structure associations between modalities	It is suitable for application scenarios where search efficiency is not high and there are tags, especially multi-tags
	CAGS [90]	Through adversarial learning and complementary attention mechanisms, better common semantic representations can be obtained	Involves multiple different loss functions, and the complexity is high	Applicable to specific data scenarios requiring data to have label information
	CM-Gans [92]	Can effectively eliminate the semantic gap between different modalities	The search accuracy is not high, and the tag information is not used	It is suitable for application scenarios with low requirements on algorithm efficiency and unlabeled data
	SSAL [94]	It has better data feature extraction ability in different modalities	The algorithm is complex and requires large-scale training data	It is suitable for scenarios with large-scale labeled data and high retrieval accuracy requirements.
	DAML [95]	Intra-modal and inter-modal losses are established to obtain consistent semantics globally	The corresponding semantic relationship between fine-grained regions in images and text words is ignored	Suitable for application scenarios that do not require high operating efficiency
Cross-modal hashing	TSDH [96]	Capable of effectively capturing all discriminative information present in the raw multimodal data	Separate feature learning and hash code learning fail to form a unified framework	It is suitable for scenarios with a large amount of data and occasions that require high search efficiency
	MIAN [97]	Heterogeneity between different modalities can be eliminated	A large amount of data is required for model training, and the model has poor scalability	Applicable to supervised cross-media search scenarios
	CMIMH [98]	The balance between reducing modal differences and modal loss of private information	The modal similarity is not well preserved in Hamming space	It applies to large-scale search problems where label information is mostly unavailable
	GCDH [99]	It can enrich semantic information with the extension of multi-tag information	Specific labeled data is required, and the algorithm complexity is high	It applies to large-scale tagged data scenarios that require high search efficiency
	DSAH [100]	The semantic information can be effectively modeled by using the internal relationship between the image and its description.	A large amount of data is required for model training, and the model has poor scalability	Deep Semantic Alignment Hashing for Unsupervised Cross-Media Search
	DDSJH [101]	High-quality semantic features can be obtained by combining latent semantic information between different modalities	It has high computational cost and quantization loss	Data suitable for large-scale complex modalities
	CPAH [102]	Scalable, no need to repeatedly train existing data	The annotation label needs to be operated manually, which does not conform to the reality	Applicable to application scenarios with the fast search of tag data or multi-tag data
	ATFH—N [103]	A triple loss function is constructed that is able to preserve the original semantic similarity between hash codes	There are many parameters and loss functions involved, and the algorithm structure is complex	Applicable to application scenarios with unbalanced data

spatio-temporal characteristics to achieve the modeling of users' real search intentions. 2) User intention understanding with multi-semantic information. Because user search intention is related to the context, we need to study user intention understanding methods based on the fusion of multi-semantic information and emotion-based user intention understanding methods to achieve semantic-level user intention understanding. 3) Interactive user search intention understanding. Given the

frequent interaction of online social network users, it is necessary to study interactive user intent understanding technology and achieve an accurate understanding of user-level intent through user interaction and feedback mechanisms. Aiming at the ambiguity, fuzziness, and time-varying of social network data, we can build a knowledge graph of virtual space and mine the hidden user intention to achieve a comprehensive and complete expression and understanding of social network data information. It also is necessary to further study the understanding of users' intentions according to the different semantics of images and texts when they appear at other times and places. It is necessary to break through the semantic barriers between shallow features and deep semantics of cross-media big data knowledge space based on advanced methods such as deep learning. Establish a cross-media and cross-space mining system that supports spatio-temporal and social characteristics. It is necessary to build a user intent understanding model based on the spatio-temporal features, social characteristics, user behavior characteristics, real space real-time data, and combined with a domain ontology knowledge base.

(6) Develop better metrics, benchmarks, and simulation environments. To develop the excellent intention understanding method, we need to consider the characteristics of the social network itself and combine multiple attributes and multimodal characteristics to deeply understand and mine the big data at the semantic level. It mines the user's intention pattern and features in social network activities from multiple perspectives, real-time acquisition of the user's intention characteristics, and highlighting the real-time expression of the user's intention. In addition, benchmarks and simulation, and experimental environments are crucial. We need to build typical benchmarks based on social network analysis, deep learning, federated learning, and integrated learning, to serve as an effective baseline. Combined with the open-source deep learning framework and the framework of social network analysis tools, we will build an experiment and simulation environment for understanding intentions in line with the actual situation.

7. Conclusions

In this paper, to encourage the promotion and growth of pertinent research, we evaluated more than 100 references and summarized the research progress of the whole cycle of social networks cross-media search based on user search intention comprehension. We mainly discuss it layer by layer from four aspects: semantic learning and knowledge fusion, understanding and mining of user intention social network, social network bursty topic discovery, and social network cross-media search. We also introduced standard social network cross-media search datasets, including IAPR TC-12, NUS-WIDE, Wikipedia, Pascal VOC, WIKI-CMR, Flickr30k, MS COCO, PKU XMedia, PKU XMediaNet, Twitter-100 K, M5Product and Wukong. We introduced the commonly used social network cross-media search evaluation indicators and conducted experimental analysis and comparison on typical methods based on the evaluation indicators. In addition, we also discuss the future development trend of cross-media search in social networks.

Although researchers have done a lot of work in cross-media search, the problems faced in cross-media search have not been well solved for social networks. To better deal with the issue of cross-media search and realize the implementation of the research, we need to introduce further algorithms that meet the attributes and characteristics of social networks and carry out research on algorithms with generalization ability in social networks. Therefore, we still have a lot of work to explore and implement in social network cross-media search. We hope our review will attract more researchers to pay attention to the latest progress in cross-media search and stimulate more meaningful research and applications on social networks.

Data availability declaration

The data used to support the findings of this study are available from the corresponding author upon request.

CRediT authorship contribution statement

Lei Shi: Conceptualization, Methodology, Data curation, Writing – original draft, Writing – review & editing. **Jia Luo:** Writing – review & editing. **Chuangying Zhu:** Writing – review & editing. **Feifei Kou:** Writing – review & editing. **Gang Cheng:** Writing – review & editing. **Xia Liu:** Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities (No. CUC220C011), National Natural Science Foundation of China (No. 62002027, No. 72104016), R&D Program of Beijing Municipal Education Commission (No. SM202110005011), Youth Fund Project of Guangxi Natural Science Foundation (No. 2021JJB170032), Central Guidance on Local Science and the Technology Development Fund of Hebei Province (No. 226Z5404G), Natural Science Foundation of Hebei Province of China (No. D2022508002), Guangxi Key Laboratory of Trusted Software.

References

[1] Wang K., Yin Q., Wang W., et al. A comprehensive survey on cross-modal retrieval[J]. arXiv preprint arXiv:1607.06215, 2016.
[2] P. Kaur, H.S. Pannu, A.K Malhi, Comparative analysis on cross-modal information retrieval: a review[J], Comput. Sci. Rev. 39 (2021), 100336.
[3] Y. Yang, C. Zhang, Y.C. Xu, et al., Rethinking Label-Wise Cross-Modal Retrieval from A Semantic Sharing Perspective[C], IJCAI, 2021.
[4] E.S. Kim, W.Y. Kang, K.W. On, et al., Hypergraph attention networks for multimodal learning[C]//, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14581–14590.
[5] S. Chun, S.J. Oh, R.S. De Rezende, et al., Probabilistic embeddings for cross-modal retrieval[C]//, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8415–8424.
[6] D. Zhang, X.J. Wu, Scalable Discrete Matrix Factorization and Semantic Autoencoder for Cross-Media Retrieval[J], IEEE Trans. Cybern. (2020).
[7] L. Shi, J. Du, M. Liang, et al., Dynamic topic modeling via self-aggregation for short text streams[J], Peer Peer Netw. Appl. 12 (5) (2019) 1403–1417.
[8] H. Zhang, C. Yi, B. Zhu, et al., Multimodal Topic Modeling by Exploring Characteristics of Short Text social media[J], IEEE Trans. Multimedia (2022).
[9] S. Liang, Collaborative, dynamic and diversified user profiling[C]//, Proc. Conf. AAAI Artif. Intell. 33 (2019) 4269–4276.
[10] S.R. Sahoo, B.B Gupta, Multiple features based approach for automatic fake news detection on social networks using deep learning[J], Soft. Comput. 100 (2021), 106983.
[11] F. Kou, J. Du, C. Yang, et al., A multi-feature probabilistic graphical model for social network semantic search[J], Neurocomputing 336 (2019) 67–78.
[12] R. Hu, A. Singh, Unit: multimodal multitask learning with a unified transformer [C]//, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1439–1449.
[13] S. Yang, L. Li, S. Wang, et al., SkeletonNet: a hybrid network with a skeleton-embedding process for multi-view image representation learning[J], IEEE Trans. Multimedia 21 (11) (2019) 2916–2929.
[14] M. Zhang, W. Li, Q. Du, Diverse region-based CNN for hyperspectral image classification[J], IEEE Trans. Image Process. 27 (6) (2018) 2623–2634.
[15] D. Gao, K. Li, R. Wang, et al., Multi-modal graph neural network for joint reasoning on vision and scene text[C]//, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12746–12756.
[16] X. Xu, T. Wang, Y. Yang, et al., Cross-modal attention with semantic consistence for image–text matching[J], IEEE Trans. Neural Netw. 31 (12) (2020) 5412–5425.
[17] Z. Yang, Z. Lin, P. Kang, et al., Learning shared semantic space with correlation alignment for cross-modal event retrieval[J], ACM Trans. Multim. Comput. Commun. Appl. (TOMM) 16 (1) (2020) 1–22.
[18] Y. Wang, F. Sun, M. Lu, et al., Learning deep multimodal feature representation with asymmetric multi-layer fusion[C]//, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 3902–3910.

[19] X. Song, J. Chen, Z. Wu, et al., Spatial-temporal graphs for cross-modal text2video retrieval[J], IEEE Trans. Multimedia (2021).
[20] Z. Jia, Y. Lin, J. Wang, et al., HetEmotionNet: two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition[C]//, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 1047–1056.
[21] M. Ni, H. Huang, L. Su, et al., M3p: learning universal representations via multitask multilingual multimodal pre-training[C]//, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3977–3986.
[22] G. Zhang, S. Wei, H. Pang, et al., Heterogeneous feature fusion and cross-modal alignment for composed image retrieval[C]//, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 5353–5362.
[23] H. Tan, X. Liu, B. Yin, et al., Cross-modal semantic matching generative adversarial networks for text-to-image synthesis[J], IEEE Trans. Multimedia (2021).
[24] S. Ji, S. Pan, E. Cambria, et al., A survey on knowledge graphs: representation, acquisition, and applications[J], IEEE Trans. Neural Netw. (2021).
[25] S. Li, Z. Huang, G. Cheng, et al., Enriching documents with compact, representative, relevant knowledge graphs[C]//, in: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, 2021, pp. 1748–1754.
[26] F. Zhang, X. Wang, Z. Li, et al., TransRHS: a representation learning method for knowledge graphs with relation hierarchical structure[C]//, in: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, 2021, pp. 2987–2993.
[27] J. Wang, B. Jiang, Zero-shot learning via contrastive learning on dual knowledge graphs[C]//, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 885–892.
[28] K. Cao, J. Fairbanks, Unsupervised construction of knowledge graphs from text and code[C]. //, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 15–22.
[29] V. Prokhorov, M.T. Pilehvar, N. Collier, Generating Knowledge Graph Paths from Textual Definitions Using Sequence-to-Sequence Models[C], NAACL, 2019, pp. 1–9.
[30] G. Niu, Y. Zhang, B. Li, et al., Rule-guided compositional representation learning on knowledge graphs[C]//, in: Thirty-fourth AAAI conference on artificial intelligence (AAAI 2020), 2020, pp. 1–9.
[31] C. Xu, R. Li, Relation embedding with dihedral group in knowledge graph[C], in: The 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), 2019, pp. 1–10.
[32] Y. Zuo, J. Zhao, K. Xu, Word network topic model: a simple but general solution for short and imbalanced texts[J], Knowl. Inf. Syst. 48 (2) (2016) 379–398.
[33] Y. Wang, J. Liu, Y. Huang, et al., Using hashtag graph-based topic model to connect semantically-related words without co-occurrence in microblogs[J], IEEE Trans. Knowl. Data Eng. 28 (7) (2016) 1919–1933.
[34] L. Shi, G. Song, G. Cheng, et al., A user-based aggregation topic model for understanding user's preference and intention in social network[J], Neurocomputing 413 (2020) 1–13.
[35] G. Yang, D. Wen, N.S. Chen, et al., A novel contextual topic model for multi-document summarization[J], Expert Syst. Appl. 42 (3) (2015) 1340–1352.
[36] Z. Qiu, H. Shen, User clustering in a dynamic social network topic model for short text streams[J], Inf. Sci. (Ny) 414 (2017) 102–116.
[37] D. Li, B. Hu, Q. Chen, et al., Attentive capsule network for click-through rate and conversion rate prediction in online advertising[J], Knowl. Based Syst. 211 (2021), 106522.
[38] P. Zhao, K. Xiao, Y. Zhang, et al., AMEIR: automatic behavior modeling, interaction exploration and MLP investigation in the recommender system[C]//, in: Thirtieth International Joint Conference on Artificial Intelligence (IJCAI), 2021.
[39] S. Liang, E. Yilmaz, E Kanoulas, Collaboratively tracking interests for user clustering in streams of short texts[J], IEEE Trans. Knowl. Data Eng. 31 (2) (2018) 257–272.
[40] J. Huang, Z. Lin, Z. Yang, et al., Temporal graph convolutional network for multimodal sentiment analysis[C]//, in: Proceedings of the 2021 International Conference on Multimodal Interaction, 2021, pp. 239–247.
[41] L. Xia, Y. Xu, C. Huang, et al., Graph meta network for multi-behavior recommendation[C]//, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 757–766.
[42] Y. Yang, F Wang, Author topic model for co-occurring normal documents and short texts to explore individual user preferences[J], Inf. Sci. (Ny) 570 (2021) 185–199.
[43] J. Liao, W. Zhou, F. Luo, et al., SocialLGN: light graph convolution network for social recommendation[J], Inf. Sci. (Ny) (2022).
[44] H. Liu, C. Zheng, D. Li, et al., Multi-perspective social recommendation method with graph representation learning[J], Neurocomputing 468 (2022) 469–481.
[45] Z. Zhao, Z. Cheng, L. Hong, et al., Improving user topic interest profiles by behavior factorization[C]//, in: Proceedings of the 24th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2015, pp. 1406–1416.
[46] L.L. Shi, L. Liu, Y. Wu, et al., Event detection and user interest discovering in social media data streams[J], IEEE Access 5 (2017) 20953–20964.
[47] G. Xun, Y. Li, J. Gao, et al., Collaboratively Improving topic discovery and word embeddings by coordinating global and local contexts[C]//, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2017, pp. 535–543.
[48] R. Mehrotra, YilmazE. Terms, Topics & tasks: enhanced user modelling for better personalization[C]//, in: Proceedings of the 2015 International Conference on The Theory of Information Retrieval, ACM, 2015, pp. 131–140.
[49] H. Yin, B. Cui, L. Chen, et al., Dynamic user modeling in social media systems[J], ACM Trans. Inf. Syst. 33 (3) (2015) 1–44.
[50] X. Wang, D. Wang, C. Xu, et al., Explainable reasoning over knowledge graphs for recommendation[C]//, Proc. Conf. AAAI Artif. Intell. 33 (2019) 5329–5336.
[51] Y. Cao, X. Wang, X. He, et al., Unifying knowledge graph learning and recommendation: towards a better understanding of user preferences[C]//, in: The world wide web conference, 2019, pp. 151–161.
[52] L. Xia, C. Huang, Y. Xu, et al., Knowledge-enhanced hierarchical graph transformer network for multi-behavior recommendation[C]//, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 4486–4493, 35(5).
[53] L. Shi, J. Du, F. Kou, A sparse topic model for bursty topic discovery in social networks[J], Int. Arab J. Inform. Technol. 17 (5) (2020) 816–824.
[54] X. Yan, J. Guo, Y. Lan, et al., A probabilistic model for bursty topic discovery in microblogs[C]//, in: Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.
[55] K. Xu, G. Qi, J. Huang, et al., Detecting bursts in sentiment-aware topics from social media[J], Knowl. Based Syst. 141 (2018) 44–54.
[56] X. Zhu, Y. Han, S. Li, et al., A spatial-temporal topic model with sparse prior and RNN prior for bursty topic discovering in social networks[J], J. Intell. Fuzzy Syst. 42 (4) (2022) 3909–3922.
[57] X. Sun, L. Liu, A. Ayorinde, et al., ED-SWE: event detection based on scoring and word embedding in online social networks for the internet of people[J], Digit. Commun. Netw 7 (4) (2021) 559–569.
[58] L. Dai, H. Wang, X. Liu, ST-ETM: a spatial-temporal emergency topic model for public opinion identifying in social networks[J], IEEE Access 8 (2020) 125659–125670.
[59] Q. Du, N. Li, W. Liu, et al., A topic recognition method of news text based on word embedding enhancement[J], Comput. Intell. Neurosci. (2022) 2022.
[60] L. Shi, J.P. Du, M.Y. Liang, et al., SRTM: a sparse RNN-topic model for discovering bursty topics in big data of social networks[J], Int. J. Comput., Inf., Syst. Sci., Eng. 35 (4) (2019).
[61] J. Yang, Y Wu, An approach of Bursty event detection in social networks based on topological features[J], Appl. Intell. 52 (6) (2022) 6503–6521.
[62] W. Xie, F. Zhu, J. Jiang, et al., Topicsketch: real-time bursty topic detection from Twitter[J], IEEE Trans. Knowl. Data Eng. 28 (8) (2016) 2216–2229.
[63] H. Peng, J. Li, Y. Song, et al., Streaming social event detection and evolution discovery in heterogeneous information networks[J], ACM Trans. Knowl. Discov. Data 15 (5) (2021) 1–33.
[64] C. Tong, H. Peng, X. Bai, et al., Learning discriminative text representation for streaming social event detection[J], IEEE Trans. Knowl. Data Eng. (2021).
[65] P. Cao, Y. Chen, J. Zhao, et al., Incremental event detection via knowledge consolidation networks[C]//, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 707–717.
[66] M. Tong, S. Wang, Y. Cao, et al., Image enhanced event detection in news articles [C]//, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 9040–9047, 34(05).
[67] W. Cui, J. Du, D. Wang, et al., MVGAN: multi-view graph attention network for social event detection[J], ACM Trans. Intell. Syst. Technol. (TIST) 12 (3) (2021) 1–24.
[68] J. Yu, J.G. Kim, J. Gwak, et al., Abnormal event detection using adversarial predictive coding for motion and appearance[J], Inf. Sci. (Ny) 586 (2022) 59–73.
[69] N. Messina, G. Amato, A. Esuli, et al., Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders[J], ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) 17 (4) (2021) 1–23.
[70] J. Yu, Y. Lee, K.C. Yow, et al., Abnormal event detection and localization via adversarial event prediction[J], IEEE Trans. Neural Netw. (2021).
[71] H. Peng, R. Zhang, S. Li, et al., Reinforced, Incremental and Cross-lingual Event Detection from Social Messages[J], IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.
[72] L. Zhang, B. Ma, G. Li, et al., Generalized semi-supervised and structured subspace learning for cross-modal retrieval[J], IEEE Trans. Multimedia 20 (1) (2017) 128–141.
[73] Y. Wang, F. Wu, J. Song, et al., Multi-modal mutual topic reinforce modeling for cross-media retrieval[C]//, in: Proceedings of the 22nd ACM International Conference on Multimedia, ACM, 2014, pp. 307–316.
[74] M. Xu, Z. Zhu, Y. Zhao, et al., Subspace learning by kernel dependence maximization for cross-modal retrieval[J], Neurocomputing 309 (2018) 94–105.
[75] R. Sanghavi, Y. Verma, Multi-view multi-label canonical correlation analysis for cross-modal matching and retrieval[C]//, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4701–4710.
[76] D. ZENG, Y. YU, K OYAMA, Deep triplet neural networks with cluster-cca for audio-visual cross-modal retrieval[J], ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) 16 (3) (2020) 1–23.
[77] X. SHU, G ZHAO, Scalable multi-label canonical correlation analysis for cross-modal retrieval[J], Pattern Recognit. 115 (2021), 107905.
[78] M. Xu, Z. Zhu, Y. Zhao, Towards learning a semantic-consistent subspace for cross-modal retrieval[J], Multimed. Tools Appl. 78 (1) (2019) 389–412.
[79] Y. Peng, J. Qi, X. Huang, et al., CCL: cross-modal correlation learning with multigrained fusion by hierarchical network[J], IEEE Trans. Multimedia 20 (2) (2017) 405–420.

[80] W. Cao, Q. Lin, Z. He, et al., Hybrid representation learning for cross-modal retrieval[J], Neurocomputing 345 (2019) 45–57.
[81] X.F. Dong, L. Liu, L. Zhu, et al., Adversarial graph convolutional network for cross-modal retrieval[J], IEEE Trans. Circuits Syst. Video Technol. 32 (3) (2022) 1634–1645.
[82] Y. Peng, J. Qi, Reinforced cross-media correlation learning by context-aware bidirectional translation[J], IEEE Trans. Circuits Syst. Video Technol. 30 (6) (2020) 1718–1731.
[83] H. Lu, N. Fei, Y. Huo, et al., COTS: collaborative two-stream vision-language pre-training model for cross-modal retrieval[C]//, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15692–15701.
[84] Q. Cheng, Z. Tan, K. Wen, et al., Semantic pre-alignment and ranking learning with unified framework for cross-modal retrieval[J], IEEE Trans. Circuits Syst. Video Technol. (2022).
[85] L. Zhang, L. Chen, C. Zhou, et al., Exploring graph-structured semantics for cross-modal retrieval[C]//, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 4277–4286.
[86] B. Wang, Y. Yang, X. Xu, et al., Adversarial cross-modal retrieval[C]//, in: Proceedings of The 25th ACM International Conference on Multimedia, ACM, 2017, pp. 154–162.
[87] S. Qian, D. Xue, Q. Fang, et al., Integrating multi-label contrastive learning with dual adversarial graph neural networks for cross-modal retrieval[J], IEEE Trans. Pattern Anal. Mach. Intell. (2022).
[88] S. Qian, D. Xue, H. Zhang, et al., Dual adversarial graph neural networks for multi-label cross-modal retrieval[C]//, in: Thirty-Fifth AAAI Conference on Artificial Intelligence, Held virtually, Feb 2-9, 2021, Palo Alto, AAAI, 2021, pp. 2440–2448.
[89] N. Han, J. Chen, H. Zhang, et al., Adversarial multi-grained embedding network for cross-modal text-video retrieval[J], ACM Trans. Multim. Comput. Commun. Appl. (TOMM) 18 (2) (2022) 1–23.
[90] L. Shi, J. Du, G. Cheng, et al., Cross-media search method based on complementary attention and generative adversarial network for social networks [J], Int. J. Intell. Syst. 37 (8) (2022) 4393–4416.
[91] W. CHEN, Y. LIU, E.M. BAKKER, et al., Integrating information theory and adversarial learning for cross-modal retrieval[J], Pattern Recognit. 117 (2021), 107983.
[92] Y. Peng, J. Qi, Cm-Gans: cross-modal generative adversarial networks for common representation learning[J], ACM Trans. Multim. Comput. Commun. Appl. (TOMM) 15 (1) (2019) 1–24.
[93] W. Ou, R. Xuan, J. Gou, et al., Semantic consistent adversarial cross-modal retrieval exploiting semantic similarity[J], Multimed. Tools Appl. (2019) 1–18.
[94] Y. WANG, S. HE, X. XU, et al., Self-supervised adversarial learning for cross-modal retrieval[C]//, in: Proceedings of the 2nd ACM International Conference on Multimedia in Asia, Virtual Event Singapore,Mar 7, 2021, ACM, New York, 2021, pp. 1–7.
[95] X. Xu, L. He, H. Lu, et al., Deep Adversarial Metric Learning for Cross-Modal Retrieval[J], World Wide Web, 2018, pp. 1–16.
[96] D. Zhang, X.J. Wu, T. Xu, et al., Two-stage supervised discrete hashing for cross-modal retrieval[J], IEEE Trans. Syst. Man. Cybern. Syst. (2022).
[97] Z. Zhang, H. Luo, L. Zhu, et al., Modality-invariant asymmetric networks for cross-modal hashing[J], IEEE Trans. Knowl. Data Eng. (2022).
[98] T. Hoang, T.T. Do, T.V. Nguyen, et al., Multimodal mutual information maximization: a novel approach for unsupervised deep cross-modal hashing[J], IEEE Trans. Neural Netw. (2022).
[99] C. Bai, C. Zeng, Q. Ma, et al., Graph convolutional network discrete hashing for cross-modal retrieval[J], IEEE Trans. Neural Netw. (2022).
[100] D. Yang, D. Wu, W. Zhang, et al., Deep semantic-alignment hashing for unsupervised cross-modal retrieval[C]//, in: Proceedings of the 2020 International Conference on Multimedia Retrieval, 2020, pp. 44–52.
[101] H. Zhang, F. Liu, B. Li, et al., Deep discriminative image feature learning for cross-modal semantics understanding[J], Knowl. Based Syst. 216 (2021), 106812.
[102] D. Xie, C. Deng, C. Li, et al., Multi-task consistency-preserving adversarial hashing for cross-modal retrieval[J], IEEE Trans. Image Process. 29 (2020) 3626–3637.
[103] X. Liu, Y. Cheung, Z. Hu, et al., Adversarial tri-fusion hashing network for imbalanced cross-modal retrieval[J], IEEE Trans. Emerg. Topics Comput. Intell. 5 (4) (2021) 607–619.
[104] M. Grubinger, P. Clough, H. Müller, et al., The iapr tc-12 benchmark: a new evaluation resource for visual information systems[C]//, in: International workshop ontoImage, 2006.
[105] T.S. Chua, J. Tang, R. Hong, et al., Nus-wide: a real-world web image database from national university of singapore[C]//, in: Proceedings of the ACM international conference on image and video retrieval, 2009, pp. 1–9.
[106] N. Rasiwasia, J. Costa Pereira, E. Coviello, et al., A new approach to cross-modal multimedia retrieval[C]//, in: Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 251–260.
[107] C. Rashtchian, P. Young, M. Hodosh, et al., Collecting image annotations using amazon's mechanical turk[C]//, in: Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk, 2010, pp. 139–147.
[108] W. Xiong, S. Wang, C. Zhang, et al., Wiki-cmr: a web cross modality dataset for studying and evaluation of cross modality retrieval models[C]//, in: 2013 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2013, pp. 1–6.
[109] P. Young, A. Lai, M. Hodosh, et al., From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions[J], Trans. Assoc. Comput. Linguist. 2 (2014) 67–78.
[110] T.Y. Lin, M. Maire, S. Belongie, et al., Microsoft coco: Common objects in context [C]// European conference On Computer Vision, Springer, Cham, 2014, pp. 740–755.
[111] Y. Peng, X. Zhai, Y. Zhao, et al., Semi-supervised cross-media feature learning with unified patch graph regularization[J], IEEE Trans. Circuits Syst. Video Technol. 26 (3) (2015) 583–596.
[112] Y. Peng, X. Huang, Y. Zhao, An overview of cross-media retrieval: concepts, methodologies, benchmarks, and challenges[J], IEEE Trans. Circuits Syst. Video Technol. 28 (9) (2017) 2372–2385.
[113] Y. Hu, L. Zheng, Y. Yang, et al., Twitter100k: a real-world dataset for weakly supervised cross-media retrieval[J], IEEE Trans. Multimedia 20 (4) (2017) 927–938.
[114] X. Dong, X. Zhan, Y. Wu, et al., M5Product: self-harmonized contrastive learning for e-commercial multi-modal pretraining[C]//, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 21252–21262.
[115] Gu J., Meng X., Lu G., et al. Wukong: 100 million large-scale Chinese cross-modal pre-training dataset and a foundation framework[J]. arXiv preprint arXiv: 2202.06767, 2022.
[116] S.A. Curiskis, B. Drake, T.R. Osborn, et al., An evaluation of document clustering and topic modelling in two online social networks: twitter and Reddit[J], Inf. Process. Manag. 57 (2) (2020), 102034.
[117] F. Kou, J. Du, W. Cui, et al., Common semantic representation method based on object attention and adversarial learning for cross-modal data in IoV[J], IEEE Trans. Veh. Technol. 68 (12) (2019) 11588–11598.