Metadata Connector: Exploiting Hashtag and Tag for Cross-OSN Event Search

Yuqi Gao, Jitao Sang, Chengpeng Fu, Zhengjia Wang, Tongwei Ren, and Changsheng Xu, *Fellow, IEEE*

***Abstract*—Social media has revolutionized the way people understand and keep track of real-world events. Various related multimedia information in different modalities such as texts, images and videos is updated on social media and reflects the events. These quantities of information distributes on different Online Social Networks (OSNs), which provides rich, wide coverage, comprehensive information about the trending events. Faced with such large amounts of data, searching has become a handy tool for event understanding and tracking on social media. However, existing single-OSN search mainly involves with single modality on single platform. Moreover, most OSNs usually focus on biased perspective of events, which significantly limits the coverage and diversity of single-OSN based event search. In this paper, we introduce a novel cross-OSN framework to help integrate these cross-OSN information regarding the same event and provide an immersive experience for information retrieval. Since social media information is widely distributed in different OSNs where semantic gap exists among these heterogeneous spaces, we propose to utilize hashtag and tag, which are user-generated metadata for organizing and labeling in many OSNs, as bridges to connect between different OSNs. In our four-stage solution framework, various methods are adopted for hashtag and tag filtering, search results representation, clustering and demonstration. Given an event query, in the first stage we generate related items with corresponding tags and hashtags from OSNs and filter the hashtags and tags we need. Then, topical representation is generated for hashtag and tag. The third stage leverages the derived representation for cross-OSN hashtag and tag clustering. Finally, demonstration for each query is produced and the results are organized hierarchically. Experiments on a dataset containing hundreds of search queries and related items demonstrate the effectiveness of our cross-OSN event search framework.**

***Index Terms*—Social media search, cross-OSN application, social multimedia.**

This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0100604, in part by the National Natural Science Foundation of China under Grants 61832002, 61632004, 61672518, and 61602115, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20191248, and in part by the Science, Technology and Innovation Commission of Shenzhen Municipality under Grant JCYJ20180307151516166. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Pradeep K. Atrey. *(Corresponding author: Jitao Sang.)*

Yuqi Gao and Tongwei Ren are with the State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China (e-mail: gaoyq@smail.nju.edu.cn; rentw@nju.edu.cn).

Jitao Sang, Chengpeng Fu, and Zhengjia Wang are with the School of Computer and Information Technology and the Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China (e-mail: jtsang@bjtu.edu.cn; 18120358@bjtu.edu.cn; 17281195@bjtu.edu.cn).

Changsheng Xu is with the National Laboratory of Pattern Recognition Institute of Automation Chinese Academy of Sciences, Beijing 100190, China (e-mail: csxu@nlpr.ia.ac.cn).

I. Introduction

With the rapid development of social media, a great amount of information discussing the real-world events is uploaded and shared in real time. Various Online Social Networks (OSNs) play a major role in real-time information acquisition and extensive sharing. These OSNs contain abundant multimedia information such as texts, images and videos which provide opportunity for us to integrate these multimedia items to provide a unified searching experience. In addition to the real-time feature, information on social media is also characterized by its multi-source distribution: when real-world events happen, the discussions around certain perspectives of the events distribute and propagate between different OSNs [25]. For example, considering the event 2018 NBA all-star, audience follow real-time progress and support their home teams on Twitter, watch and discuss highlights on YouTube, and share their favorite players' posters on Instagram and Flickr. These cross-OSN information enables comprehensive event understanding and description in different modalities and from different perspectives.

Faced with massive information on social media, online search has become a window for people to understand the world. In spite of the cross-OSN distribution feature, current social media search functions focus on information access to one single OSN. Taking the search in Twitter for example, although ranking options like time and popularity are supported, the following problems prevent from a better experience: (1) Information richness. Prevailing OSN usually focuses on single modality of multimedia, e.g., text on Twitter, image on Flickr, video on YouTube. Although Twitter also supports multimedia attachments like image and video, we observed in our data analysis that the embedded images and videos in Twitter are inferior to those from other platforms in terms of both quality and endorsement. (2) Information coverage. Different perspectives are described between OSNs, together contributing to full-scale event understanding and tracking. While Twitter features in abundant data and effective propagation, Flickr features in visual information demonstration and user group interaction, YouTube provides living stream and social discussion. Therefore, an immersive cross-OSN search framework is what we needed: given an search query, relevant information from different OSNs is filtered, aggregated, organized, and demonstrated as search results.

Fig. 1. The search results from different OSNs with tags and hashtags by issuing "Nba All Star 2018" to different OSNs.

TABLE I
EXAMPLE OF QUERIES WITH RETRIEVED HASHTAGS AND TAGS.

Pro Bowl 2018	NFL, Pro Bowl 2018, Bowl 2018 live, #NFLProBowl, #Steelers...
Nba All Star 2018	#NBAAllStar, #TeamCurry, nbaallstar, all-star, team lebron vs. team curry, stephen curry ...
Oscar nominations 2018	#Oscars2018, oscar nominations 2018, Oscars, Academy Awards, #oscarnominations2018...
Australian Open 2018	australian open, tennis, mes's final, #TennisNews, #MensTennis, #AusOpen...
Grammys 2018	Grammy Winner, GRAMMY2018, #Grammys, #musiclover, #Billboard...

A direct solution is to directly aggregate and arrange the returned items related to the search query from different OSNs. However, there are several problems with this solution: (1) Relevance. It is difficult for common users to choose an appropriate search query to describe the event accurately and succinctly. Searching only with an inaccurate query will make the search results biased and noisy. (2) Organization. Different OSNs support different search options and avoid a consistent ranking solution.

Twitter sorts search results by "the latest" and popularity; Flickr displays listings by date, interest and relevance; and Youtube by date, relevance, rating and viewed times.

Moreover, the different modality focuses aggravate the difficulty in aggregation and organization. Fig. 1 shows search results of "Nba All Star 2018," where different modalities of information from Twitter, Flickr and YouTube are returned with ranking options of popularity, interest and #. view. In addition to the inconsistent ranking options and different modalities, different information of a certain topic marked with the same color cannot be organized across the platforms directly. Further, we provide some other examples in Table I to illustrate the connection between events and retrieved tags&hashtags.

This paper proposes to exploit the hashtags&tags as bridge to address the above challenges and solve the cross-OSN search problem. Hashtag and tag are typical social media metadata widely used on different OSNs. These metadata are useful to help ensure the relevance of noisy results and the organization of multi-modal information: (1) Since hashtags and tags are user-generated annotations, the relevance of annotated items to events is partially guaranteed. Moreover, the hashtag and tag themselves serve as suitable search queries and in this way more related items will be retrieved by further querying. (2) Hashtag and tag are originally adopted for information management and indexing, making them natural tools for cross-OSN and multi-modal information organization [24].

Fig. 1 also shows hashtags and tags where subtopics are marked with different colors. Three quick observations about tags and hashtags are derived: (1) Regarding the same event query, multiple different hashtags or tags are adopted on each OSN and vary between OSNs. The number of tags on Flickr and YouTube is higher than that of hashtags on Twitter. (2) Different hashtags and tags describe different aspects, i.e., subtopics of the event. The straightforward solution based on raw ranking from respective OSNs will mix these subtopics. (3) There is a considerable overlap between the hashtags and tags from different OSNs, inspiring us to exploit the inner semantic structure within these metadata for cross-OSN aggregation and organization.

Moreover, with the example in Fig. 1, different social networks provide hashtags and tags in diverse fields. Hashtag in Twitter is a representative one-word item which can be made up of multiple words. Flickr can provide tags which aim to index like "lol" for humor images, or record like "LeBron James wins third career NBA All-Star Game MVP, tying Michael Jordan". YouTube can provide tags for fields like "2k18 myleague" discussing basketball video game "NBA 2k18" which usually been discussed and shared by long-time game-play videos. Also, different social websites have different fields of users. Gamers of basketball video games simulate the all-star game and share their videos on YouTube. Sports fans follow the real-time progress of all-star game on Twitter. Photographers record the wonderful moments on Flickr.

Considering the above observations and to exploit hashtag&tag better for cross-OSN information aggregation and organization, the fundamental question is how to discover the underlying subtopics regarding certain events, and to organize the filtered multiple hashtags&tags as well as the annotated items under the subtopics. As shown in Fig. 2, the hashtag and tag-centric immersive search framework consists of four stages. The first stage aims to filter the hashtags and tags to generate a better subset for aggregation and organization. The second stage generates topical representation over a unified vocabulary set for hashtags&tags on each OSN. In third stage, filtered hashtags&tags are clustered into subtopics considering both the semantic correlation between subtopics and the hashtag and tag co-occurrence constrain. Finally, the generated hashtag and tag clusters are ranked according to the relevance to the query for search result demonstration. The main contributions of this work are summarized as following:

- We formalize the cross-OSN immersive search problem. Information in multiple modalities and from different OSNs is around the same event.
- We propose a four-stage framework to exploit the hashtag and tag as bridge for cross-OSN information aggregation and organization.

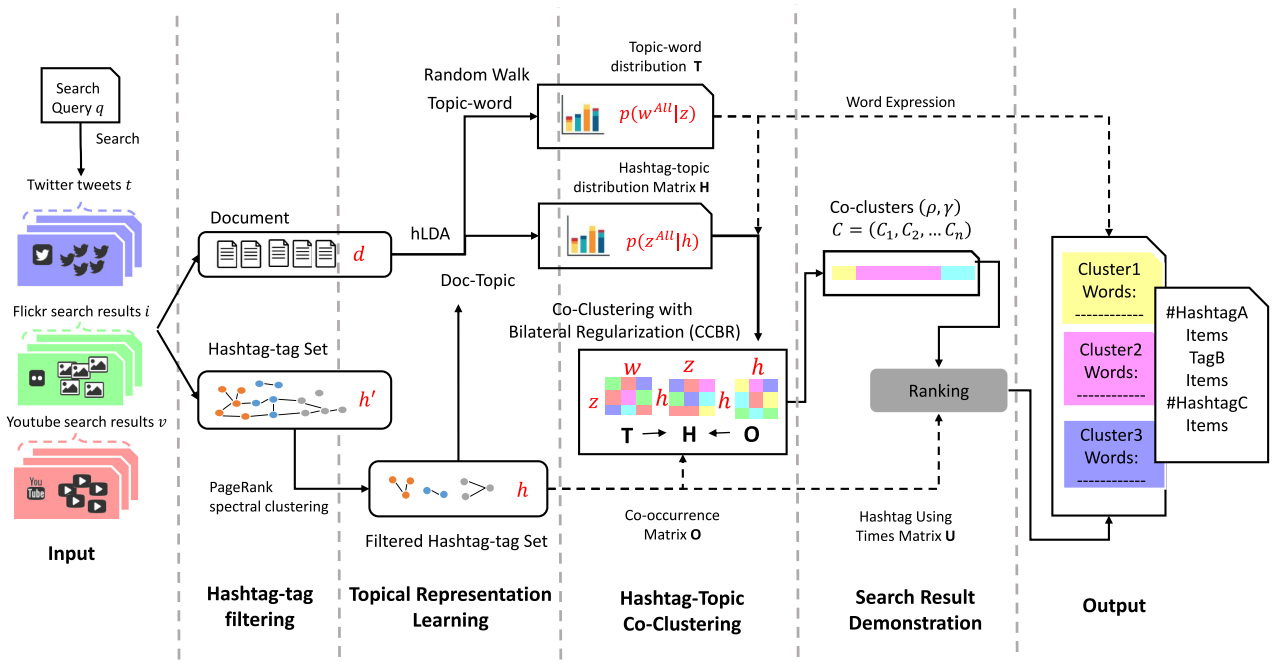


Fig. 2. The solution framework.

- We implement an online demo for search result demonstration. Quantitative and qualitative evaluation on real-world dataset demonstrates the effectiveness of the proposed solution.

A conference version of this work is published in [36]. The major difference of this version lies in the integration of tag for better information aggregation and organization. To further address the motivation, importance as well as new solution in integrating tag, we extend the paper in the following aspects: (1) Related work in Section II. A separate review of studies is introduced in utilizing tags to facilitate information retrieval. (2) Data analysis in Section III. A new dataset is prepared, on which we conduct additional data analysis to demonstrate the necessity of integrating tag. (3) Methodology in Section IV. Tag-hashtag graph is constructed and new modules of graph filtering are added to derive an integrated solution considering tags. (4) Evaluation in Section V. Evaluation on the extended modules is described in Section V-A and further quantitative/qualitative comparison is shown in Section V-C. The online demo is also updated correspondingly.

II. Related Work

A. Searching Application

Searching are widely attractive for researchers and many studies have focused on the comparison and multimedia application of them. In [5], coverage rates of websites returned by search engines were analyzed with search engines from different domains and countries. In [6], a search engine that aims to multimedia information retrieval such as video and audio was introduced.

In addition to web search, social multimedia search has been analyzed and presented in several studies. In [9], Flickr and Wikipedia were adopted to improve the matching of indexing vocabulary and query vocabulary. In [8], the author provided a hierarchical visualization method which enables an understanding of the query topic from different perspectives and a corresponding dataset on YouTube was developed. However, aggregating and organizing search results from different OSNs has not been widely researched and discussed.

B. Cross-OSN Analysis and Application

Social multimedia researchers also pay attention on cross-OSN applications. One momentous research area is user-centric application, and most of the user-centric applications focus on the same user's information from different social networks and information in different modalities for user modeling. In [7], the authors presented a cross-OSN user modeling solution utilizing user's social data and behavior data and the solution was evaluated by the task of personalized video recommendation. [3] proposed a cross-OSN recommendation framework to drive the network traffic from videos on YouTube and find suitable Twitter followees to promote the videos.

The other important research line is content-centric application, and it intends to bridge the topic or event across OSNs. [4] introduced SocialTransfer, a novel transfer learning framework, to connect different social media with two representative use cases mutually. [1] focused on events of interest discovery, and proposed a framework to unify cross-domain media streams. These pilot studies and applications enlighten us to focus on the issue of cross-OSN searching.

C. Hashtag Usage Pattern and Application

Hashtag was originally adopted by Twitter and many researchers have utilized hashtag for applications and datasets.

In [38], a large-scale dataset on Twitter is developed and hashtags are collected and marked. [35] combined heterogeneous features of users and images, and employed these features for hashtag recommendation.

D. Tag Usage Pattern and Application

Tag is widely used in many social tagging systems and its user-generated features has attracted a lot of attention from researchers. [15] proposed a ranking method for tag and exploited it for tag-based image search, tag recommendation, and group recommendation on Flickr. Apart from that, tag usage motivation has also been analyzed in [28] from two perspectives: function and sociality, it is argued that the goal of tagging could be various such as information organization and social communication.

Tag also plays an important role in the research of social media application and cross-domain application. [34] introduced a social game "Herd it" for tagging music. In [32], the authors proposed TagCDCF, which exploited tag that is common to different domains to build a cross-domain recommender system. The authors in [33] also utilized tag in recommender system to solve cold-start problem. A recent work [39] proposed a personalized recommendation approach of social image by employing deep features and tag trees to model user interest. [40] focused on the problem of zero-shot image tagging with deep multiple instance learning.

The above studies show the effectiveness of hashtag and tag in social media information organization and indexing, constitute the basis and inspire us to implement a hashtag&tag-centric framework in this work. Different to the above applications which mostly adopt hashtag or tag as additional information or auxiliary context, we intend to construct a novel cluster-hashtag&tag-item hierarchical structure to organize and integrate the search results with them.

III. Data Analysis

Data analysis in this section is processed to answer three questions: (1) Why is it necessary to study cross-OSN search? (2) Why could hashtag and tag be appropriate metadata connectors to solve the cross-OSN search problem? (3) What are the advantages of integrating tag into the framework?

For the first question, we compare search results from different single OSNs to demonstrate the advantage of integrating the single-OSN search results. For the second and third questions, we examine how users employ tag and hashtag across different OSNs to discuss the availability and also the challenges in using hashtag and tag, and the superiority of integrating tags. The challenges discussed will be resolved later in the solution section.

A. Single-OSN Search Comparison

We compare search results from different OSNs in terms of information richness and user interaction.

1) Information Richeness: First, we compare information richness of returned results. For images, we compare the resolution of them and for videos, we compare the duration of them between the OSNs. 205 search queries are chosen on Google Trends,[1] involving subjects from entertainment, sports, economy to public events. We get search results by APIs[2] from the platforms with the queries respectively, and 18,891 tweets are acquired from Twitter, 112,262 image items from Flickr and 96,881 video items from YouTube. Average resolution and duration of images and videos are illustrated in Fig. 3(a) (b) for each query separately. Specifically, images and videos on Twitter refer to the embedded multimedia of tweet and when comparing, only tweets with embedded images or videos are involved. There is no doubt that the Flickr and YouTube items contain significant richer information than those from Twitter regarding to image resolution and video duration. Besides, it is shown is the Fig. 3(b) there are even no video embedded in some search results of Twitter. We then further make a data analysis on what percentage of results includes video or image information on Twitter. The result shows that about 10% search results include images and 0.3% search results include videos. About 84% queries' search results do not have videos in them. This statistical result supports the superiority of the other two platforms over Twitter on information richness. While the superiority of twitter is that it provides effective dissemination of text information, and as supplement and extension, Flickr and YouTube could supply images and videos with high quality.

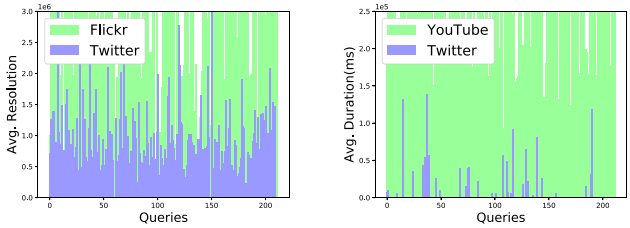


Fig. 3. Information richness comparison.

2) User Interaction: We compare the differences in user interactions attracted by search results on three OSNs. Two typical interactions, comment and endorsement, are examined. *Retweets* on Twitter are counted as comment; *like*/*dislike* on YouTube and *favorites* on Twitter and Flickr are calculated as endorsement. The mean number of comments and endorsements is shown in Fig. 4(a) (b) in log-scale among the OSNs. It is shown that the two figures reach similar observations that more user interactions occur on YouTube than Flickr and Twitter which enables us to utilize multi-platform information to promote better cross-OSN searching experience.

Through the above comparisons, we point out the limitations of single-OSN search. In conclusion, to ensure a better search experience and exploit more advanced features such as social interaction, it is necessary to integrate the single-OSN search results.

[1][Online]. Available: https://trends.google.com

[2]Flickr: [Online]. Available: https://www.flickr.com/services/api/
Twitter: [Online]. Available: https://dev.twitter.com/overview/api
YouTube: [Online]. Available: https://www.youtube.com/yt/dev/api-resources.html

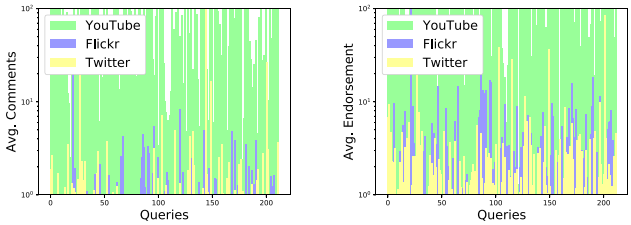


Fig. 4. User interaction comparison.

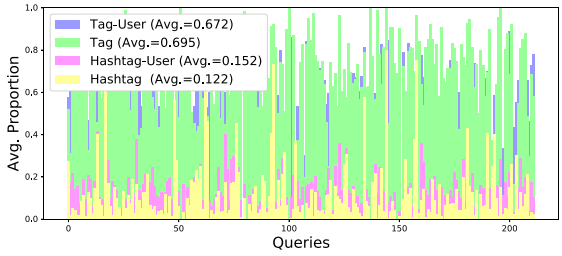


Fig. 5. Tag and Hashtag usage comparison on Flickr.

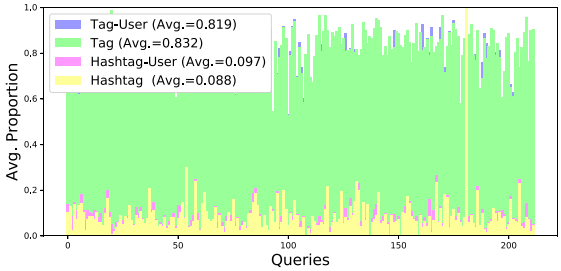


Fig. 6. Tag and Hashtag usage comparison on YouTube.

B. *Cross-OSN Hashtag and Tag Usage Analysis*

The usage analysis subsection points out the availability and challenge of exploiting hashtag and tag to integrate cross-OSN search results and demonstrates the benefits of integrating tags. Corresponding data analysis below consists of four stages that elaborate and compare popularity, diversity, topic coverage and semantic expression of hashtag and tag respectively. Inspired by the data analysis, Section IV will introduce how to solve the challenges and how to take full advantage of the features of hashtag and tag to improve the performance of cross-OSN search.

1) Popularity: In Fig. 5, we count and compare the percentage of search results as well as users using hashtag and them using tag within the search results returned by Flickr per query.

The average percentage of hashtag is 15.2% for user and 12.2% for search result on Flickr, on the other hand, the average percentage of tag is 67.2% for user and 69.5% for search result on average. When it comes to YouTube shown in Fig. 6, the percentage of hashtag shows similar characteristics and the average percentage is 9.7% for user and 8.8% for search result and the percentage of tag is higher with average search result and user percentage above 81%. Combining the results from Fig. 5 and Fig. 6, it is obvious that tag shows higher popularity than hashtag on both Flickr and YouTube, which makes tag more

TABLE II
THE NUMBER OF UNIQUE HASHTAG AND TAG USED PER QUERY.

Twitter	Flickr-tag	Flickr-hashtag	YouTube-tag	YouTube-hashtag
19.32	1002.87	62.49	2732.06	84.93

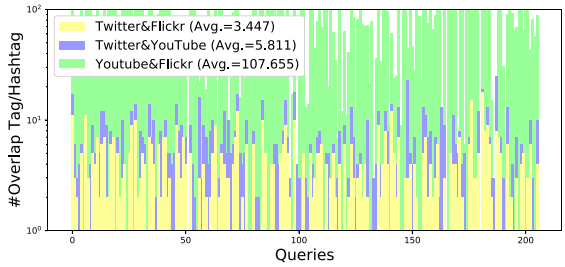


Fig. 7. The number of overlapping tags and hashtags between different OSNs.

suitable and informative for cross-OSN integration considering popularity. Regarding to hashtag on Twitter, we also perform the analysis that shows hashtag is popular on Twitter, with average search result and user percentage above 18%. Considering that tag is not supported on Twitter and the high popularity of hashtag on Twitter, it is reasonably suitable for cross-OSN integration.

2) Diversity: Apart from popularity, hashtag and tag also differ in usage diversity. In discussions surrounding certain topics, users may create multiple hashtags and tags which vary across OSNs. We count the unique hashtags and tags used by each query on the three OSNs and summarize the results in Table II. It is obvious that compared with hashtag, tag is far more popular on these platforms, which makes it more appropriate for future processing and a rich information source.

Furthermore, we calculate the overlapping hashtags and tags between every two platforms and tags are considered on Flickr and YouTube and hashtags on Twitter. Statistical result of overlapping hashtags and tags is shown in Fig. 7. Considering the numbers of tags and hashtags shown in Table II, these overlapping hashtags and tags achieve a high proportion in every two platforms which indicates their importance to bridge the platforms. These overlapping hashtags and tags facilitate connection between different OSNs and help us generate the representative tags and hashtags for integration and organization. Also, overlapping hashtags and tags show that there are similar and different discussion fields among platforms to some extent.

In addition, we examine the sequence of hashtags and tags between OSNs. Spearman's footrule [20] [21] is an extensively used method to calculate the distance between permutations. Normalized spearman's footrule which we used is calculated as:

$$NFr(\mu_1, \mu_2) = 1 - \frac{Fr^{|S|}(\mu_1, \mu_2)}{\max \quad Fr^{|S|}} \tag{1}$$

where μ_1, μ_2 are two permutations, $|S|$ is the number of overlapping items between two permutations, $max \quad Fr^{|S|}$ calculated as $1/2|S|^2$ when $|S|$ is even and $1/2(|S|+1)(|S|-1)$ when $|S|$ is odd, $Fr^{|S|}(\mu_1, \mu_2)$ is the standard Spearman's footrule

TABLE III
NFr SCORE TO EXAMINE HASHTAG AND TAG USAGE DIFFERENCE BETWEEN OSNS.

Twitter&Flickr	Twitter&Youtube	Flickr&Youtube
0.3107	0.3588	0.4104

which is:

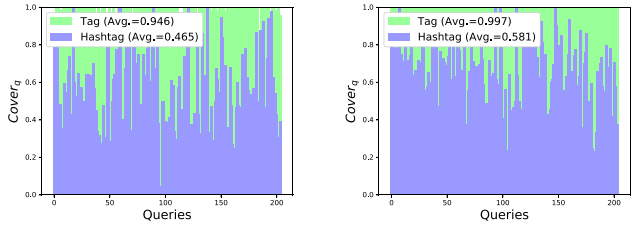
$$Fr^{|S|}(\mu_1, \mu_2) = \sum_{i=1}^{|S|} |\mu_1(i) - \mu_2(i)| \quad (2)$$

where $\mu_1(i)$ is the rank of i^{th} item in permutations μ_1. The NFr score varies from 0 to 1 and higher NFr score indicates the two permutations are more similar. To calculate the NFr score, we sorted the hashtags returned from Twitter and the tags from Flickr and YouTube in descending order based on the number of search results annotated by them. Table III expresses the average NFr score of tag and hashtag lists between OSNs over the queries. Combined with the former analysis of overlapping hashtags and tags, two observations are derived: (1) Hashtag and tag lists among OSNs are generally different, which indicates cross-OSN hashtags and tags provide abundant information for aggregating. (2) Shared hashtags and tags appear high in ranking showing their high representativeness.

Furthermore, integrated with the previous data analysis on popularity, we make the conclusion that both hashtag and tag are diffusely used and appropriate as bridge to integrate and organize cross-OSN search results, but integration and organization still face challenges owing to the usage diversity and quantity difference.

3) Topic Coverage: Moreover, topic coverage needs to be considered when comparing the comprehensiveness of hashtags and tags retrieved by the query. The more subtopic of search results covered by hashtags or tags, the more topical comprehensiveness and suitable for organization they are. To explore the semantic meaning and compare the topic coverage of hashtags and tags, we employ the hierarchical topic modeling method, hLDA [11]. hLDA models the topic in a hierarchical topic tree with depth M(M is set to 2 in our analysis), while classic topic model LDA has a horizontal structure. Through one path from the root of the tree, each document could be produced by the corresponding path. Topic modeling is conducted on the original search result collection $\mathcal{D}_{o_q}^{F}$ for Flickr and $\mathcal{D}_{o_q}^{Y}$ for YouTube respectively, with the textual content of each item $\mathbf{d_o}_q^{F,Y}$ as document.[3] After topic modeling, take Flickr as an example, each document $\mathbf{d_o}^F$ is affixed with a 2-dimension topic distribution $[p(z_o{}^{F,root}|\mathbf{d_o}^F), p(z_{o_k}^{F,leaf}|\mathbf{d_o}^F)]$. $z_o{}^{F,root}$ represents the root topic and $z_{o_k}^{F,leaf}$ represents the k^{th} leaf topic. The root topic is presumed to be fully covered by the documents, so we only compare the topic coverage of tag and hashtag on leaf topic distribution. For each topic distribution $p(z_o|\mathbf{d_o})$, we compare the documents with hashtag and those with tag in order to discover whether the corresponding topic appears in the leaf topics.

[3]Textual content is collected from title & description of YouTube video and Flickr image.



(a) (b)

Fig. 8. Topic coverage comparison.

TABLE IV
#. TAGS AND HASHTAGS CONTAINS SEMANTIC INFORMATION OF SEARCH QUERY.

metadata	Tag	Hashtag
Flickr	27.08	5.04
Youtube	381.24	6.10

Coverage score of hashtag for a certain query q is then calculated as:

$$Cover_q = \frac{\sum_{k=1}^{K} \mathbb{I}(\bigvee_{\mathbf{d_o}^h}(z_{\mathbf{d_o}} \in z_k))}{K} \quad (3)$$

where K is the number of leaf topics, $\mathbf{d_o}^h$ is the document containing hashtag, $z_{\mathbf{d_o}}$ is the topic distribution of document $\mathbf{d_o}$. When calculating coverage score of tag, h is replaced with tag in Eqn. (3). The coverage score ranges from 0 to 1 and the higher the coverage score, the more topics are covered by hashtag set or tag set. Fig. 8 illustrates the comparison of hashtag and tag on Flickr and YouTube respectively. It is obvious that tags from Flickr and YouTube cover more subtopics, which indicates that tag can provide a more comprehensive understanding of the event on Flickr and YouTube.

4) Semantic Expression: In addition to the structural feature of hashtag and tag, semantic feature of hashtag and tag is also important. As shown in Fig. 1, hashtags and tags reflect literal information about their semantic meaning. For example, "#TeamLeBron" refers to the NBA All star team of the player LeBron and tag "stephen curry" refers to the player Stephen Curry. However, some hashtags might be created by accident, for example, "#1" might be created to indicate No.1 or episode 1 and tags won't face this situation. Analyzing and comparing the semantic information of tag and hashtag would help us understand the difference in usage pattern between hashtag and tag on social media and make use of it. Direct comparison of semantic information is difficult since hashtag usually concatenated by several words. To solve the problem, we utilize [37] to segment hashtags into analyzable words. Considering tag, it supports the use of multiple words and that brings semantic variety. We conduct a simple analysis to calculate how many tags and hashtags contain semantic information of the query q and the results are shown in Table IV. Word set of hashtag is generated by the segmentation method mentioned above and word set of tag is generated by directly dividing with space. Whether the hashtag or tag contains semantic information of the query can be checked by the intersection of words from query and word set of the hashtag or

tag (The words in different cases are equally calculated because they have the same literal meaning). Tags and hashtags filtered by this standard show strong connection with the search query on semantic level. Table IV shows that there are many tags and hashtags considering the literal meaning relationship of themselves and the query and there are more tags than hashtags in this respect.

The main observations of this subsection and the inspirations for later section are summarized as follows: (1) Popularity: Tag is much more popular than hashtag on Flickr and YouTube, so the former can provide more comprehensive information of social events. (2) Diversity: The usage of tag and hashtag is diverse on different platforms but there are also a large number of frequently used overlapping hashtags and tags among platforms. Diversity of tag and hashtag provides challenges and advantages. Inspired by the above analysis, overlapping hashtags and tags can bridge the quantity gap of them between platforms and detailed solution will be introduced later. (3) Topic coverage: Documents with tag cover more subtopics of the event, indicating that tag is more comprehensive and widely distributed when considering subtopics. (4) Semantic expression: Both tag and hashtag contain rich semantic information. In most cases tags show better characteristic in simple semantic relationship such as literal meaning. Based on the observations, we intend to utilize both tags and hashtags as bridges to integrate and organize cross-OSN information.

IV. Solution

A. Hashtag & Tag Filtering

As shown in data analysis, the number and usage of hashtag and tag between OSNs are not the same, directly integrating tags and hashtags from different OSNs would cause imbalance and noise. Therefore, to make the best use of tags, we need a method to filter the generated tags and hashtags and maintain the tag's advantage in terms of topic coverage at the same time. Two issues are addressed: (1) Information loss. Regarding the same query, multiple subtopics are discussed among OSNs. If we only select commonly used hashtags and tags or overlapping ones, information of some subtopics might be incomplete or inaccurate. In this case, we employ spectral clustering [31] to divide tags and hashtags and maintain the diversity. (2) Imbalance. The number of hashtags on Twitter is less than the number of tags on Flickr and YouTube significantly. We utilize PageRank [29][30] to generate more representative tags and hashtags. As illustrated in Fig. 9, we elaborate the solution to the above issues in three stages as follows:

1) Graph Conduct: To discover the relation between tags and hashtags, a graph connecting tags and hashtags from different OSNs is built. Given a graph $G = (V, E)$ with vertex set $V = \{v_1, v_2, \ldots, v_{N_{h^{all}}}\}$, a hashtag or tag is represented with the vertex and $N_{h^{all}}$ is the number of all hashtags and tags. With the assumption that hashtags or tags co-occurring in the same item indicates they have a high probability of belonging to the same subtopic, we build a matrix $\mathbf{O}_{N_{h^{all}} \times N_{h^{all}}}$ with element O_{ij} denoting the times that the i^{th} and j^{th} hashtag or tag co-occur

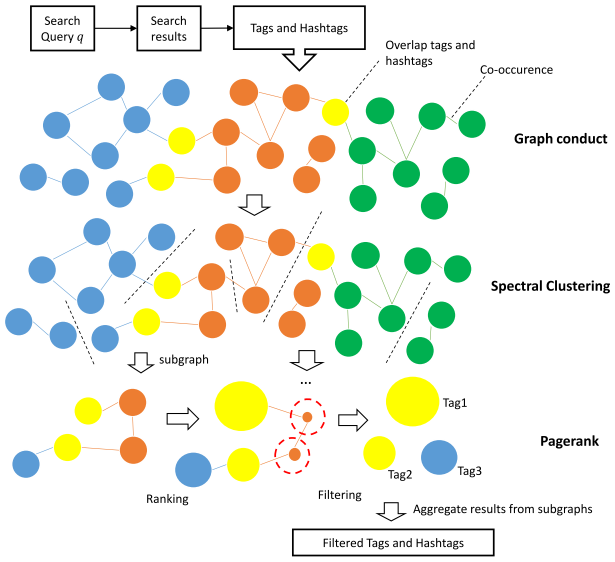


Fig. 9. Illustration of hashtag and tag filtering.

in the same item. To deal with the cross-OSN co-occurring, the overlapping tags and hashtags are applied as the same vertex in the graph G. The normalized matrix $\mathbf{O}$ is adopted as the vertex similarity matrix of the graph G.

2) Spectral Clustering: Based on the similarity matrix O, spectral clustering is adapted as follows. Let D be a diagonal matrix and D_{ii} calculated as the sum of O's i-th row, matrix L is constructed as $L = D^{-1/2}OD^{-1/2}$. Then we find L_{row} largest eigenvectors, and establish matrix X with these eigenvectors as columns. After that, X is normalized to X^{norm} and each row of X^{norm} is assigned as a sample for clustering. The tag or hashtag v_i belongs to cluster j if and only if X_i^{norm} belongs to cluster j, and to generate clusters with X^{norm}, we utilize k-means algorithm. After clustering, for each query, L_{row} clusters containing hashtags and tags are generated. Throughout the process, we obtain a preliminary classification that maintains the diversity of subtopics. A deeper exploration considering semantic structures will be discussed in the following stage.

3) PageRank: With spectral clustering, we obtain L_{row} clusters which divide graph G into subgraphs $\{G_1, G_2, \ldots, G_{L_{row}}\}$. The number of tags and hashtags are still imbalanced both inside and between subgraphs, to solve this problem, we employ PageRank. In PageRank considering the edge weights, rank $r(v_i)$ of vertex v_i is calculated as:

$$r(v_i) = \frac{(1-\lambda)}{N_{h^{all}}} + \lambda \sum_{v_j \in {v_i}^{in}} \frac{w_{ij} r(v_j)}{|{v_j}^{out}|} \tag{4}$$

where ${v_i}^{in}$ is a set of vertices link to v_i, $|{v_j}^{out}|$ is the number of out links of vertex v_j, w_{ij} is the normalized weight of vertex v_i and v_j, and $N_{h^{all}}$ is the number of vertices.

The above process will generate a rank score for each hashtag or tag and hashtags and tags are selected with a threshold within each subgraph. In this way, we obtain N_h tags and hashtags while maintaining diversity and avoiding imbalance at the same time.

B. Topical Representation Learning

To fully express the hashtags as well as tags and integrate more related information, for each filtered hashtag or tag, we further gather more items annotated by the corresponding hashtag or tag by searching with it on all OSNs(referred as *extended search results*).[4]

The second stage generates hashtag and tag topical representation and two issues are confronted: (1) Most search results are associated with a certain query and share a general topic. To avoid topic distribution mingled with each other, hLDA [11] as mentioned in Section III is utilized to discover the semantic structure. (2) Topic modeling is executed on the OSNs respectively and the generated topical distribution of cross-OSN hashtags and tags cannot be directly integrated with distinct vocabularies. To bridge the vocabularies, random walk is adopted to construct a unified vocabulary set. Detailed solution is expounded as follows.

1) Hierarchial Topic Modeling on Respective OSNs: We further conduct hLDA on extended result $\mathcal{D}_q^T, \mathcal{D}_q^Y, \mathcal{D}_q^F$ for each query to promote the mining of semantics. hLDA is performed over each OSN collection, with the textual information of $\mathbf{d}_q^{T,Y,F}$ as document (Tweet is adopted on Twitter). After topic modeling, take Twitter as an example, the topical distribution of the i^{th} hashtag h_i^T on the leaf topic space is combined over all its corresponding tweets:

$$p(z_k^{T,leaf}|h_i^T) = \frac{\sum_{\mathbf{d}^T \in \mathcal{D}^T_{h_i^T}} p(z_k^{T,leaf}|\mathbf{d}^T)}{\sum_{k=1}^{K^T} \sum_{\mathbf{d}^T \in \mathcal{D}^T_{h_i^T}} p(z_k^{T,leaf}|\mathbf{d}^T)} \quad (5)$$

where K^T is the number of leaf topics from Twitter, $\mathcal{D}^T_{h_i^T}$ represents the tweets denoted by h_i^T. As mentioned above, the root topic is expected to be fully representative, only leaf topics are involved for hashtag or tag. In consequence, three topic spaces $\{\mathbf{z}^{T,leaf}, \mathbf{z}^{Y,leaf}, \mathbf{z}^{F,leaf}\}$ are produced on vocabulary sets $\mathcal{W}^{T,Y,F}$ separately. Besides, each hashtag's and tag's topic distribution are generated on homologous space.

2) Random Walk-Based Cross-OSN Vocabulary Integration: To process the cross-OSN analysis, an integral vocabulary set $\mathcal{W}^{all}=\mathcal{W}^T \bigcup \mathcal{W}^Y \bigcup \mathcal{W}^F$ is what we need. The similarity π_{ij} of word w_i and w_j can be measured with WordNet [27], which could be used to examine the semantic relevance and connect the isolated vocabulary sets. We then build the word graph $\mathbf{G}$ with $w \in \mathcal{W}^{all}$ as node and similarity π as edge. Random walk [14]–[16] is an effective method and we utilize it to disseminate the similarities of words on the word graph. Then we construct transition matrix $R_{|\mathcal{W}^{all}|\times|\mathcal{W}^{all}|}$, in which the transition probability from word w_i to w_j is computed as $R_{ij} = \pi_{ij} / \sum_{w_k \in \mathcal{W}^{all}} \pi_{ik}$. The relevance score of node i at iteration l is represented as $s_l(i)$, the vector $\mathbf{s}_l = [\ldots, s_l(i), \ldots]^T$ consists of all these scores. The random walk is then formulated as:

$$\mathbf{s}_{l+1} = \alpha \sum_i \mathbf{s}_l R + (1-\alpha)\mathbf{t} \quad (6)$$

where $\mathbf{t}$ is the initial relevance scores substituted with former topic-word distribution, and α is a weight parameter that ranges from 0 to 1.

This process will strengthen the nodes with cognate neighbors and weaken the separate ones. It is proved to converge to a fixed point $\mathbf{s} = (1-\alpha)(\mathbf{l} - \alpha R)^{-1}\mathbf{t}$ [15]. After random walk process, a cross-OSN topic space $\mathbf{z}^{all}$ over the unified vocabulary $\mathcal{W}^{all}$ is generated.

C. Hashtag&Tag-Topic Co-Clustering

After topic modeling, we use the generated topic distribution to cluster the filtered hashtags and tags. The remaining problems are: (1) Each hashtag or tag only has topic distribution on the corresponding OSN. (2) There are internal relations of topics within and cross OSNs. To solve the problems, we propose a hashtag&tag-topic co-clustering solution, which considers both semantic connection of topics and hashtag&tag co-occurrence. The following part of subsection first introduces the standard Bregman co-clustering and expounds our hashtag&tag-topic co-clustering solution next.

1) Bregman Co-Clustering: Bregman co-clustering [12] is an effective approach in multi-dimension clustering. It intends to discover the optimal row and column mapping (ρ^*, γ^*) of an existing matrix $\mathbf{H}$ defined on two sets $\mathcal{H}$ and $\mathcal{T}$. Element of matrix $\mathbf{H}$ takes values following ν, and $\nu = \{\nu_{ij}; i = 1, \ldots, |\mathcal{H}|, j = 1, \ldots, |\mathcal{T}|\}$ represents the joint probability measure of (H, T) defined on $\mathcal{H}$ and $\mathcal{T}$ severally (That means $H_{ij} \sim \nu_{ij}$).

Make matrix $\hat{\mathbf{H}}$ an approximation of $\mathbf{H}$ that determines only upon (ρ, γ) and summary statistics. Then the quality co-clustering (ρ^*, γ^*) can be measured through minimizing the expected Bregman divergence on ν between $\hat{\mathbf{H}}$ and $\mathbf{H}$:$(\rho^*,\gamma^*)=\arg\min_{\rho,\gamma} E[d_\phi(\mathbf{H},\hat{\mathbf{H}})]=\arg\min_{\rho,\gamma} \sum_i \sum_j \nu_{ij} d_\phi(H_{ij},\hat{H}_{ij})$, where ϕ is a convex function and $d_\phi(z_1, z_2)$ represents *Bregman divergence* which is defined as:$d_\phi(z_1, z_2) = \phi(z_1) - \phi(z_2) - < z_1 - z_2, \bigtriangledown\phi(z_2) >$, $\bigtriangledown\phi$ represents the gradient of ϕ.

2) Hashtag & Tag-Topic Co-Clustering With Bilateral Regularization: Bregman co-clustering can be processed iteratively and three subproblems are solved during each iteration. First, with mapping (ρ_i, γ_i) at i^{th} step, the approximation matrix $\hat{\mathbf{H}}$ is updated by resolving a *Minimum Bregman Information* problem [12]. Randomly shuffle the rows or columns, we get a permuted matrix $\tilde{\mathbf{H}}$ from $\hat{\mathbf{H}}$ The second and third subproblem utilize the permuted matrix $\tilde{\mathbf{H}}$ to choose the optimal column and row mappings by optimizing the functions:

$$\gamma_{i+1}(t) = \arg \min_{1,\ldots,L_{col}} E_{H|t}[d_\phi(\mathbf{H}, \tilde{\mathbf{H}})] \quad (7)$$

$$\rho_{i+1}(h) = \arg \min_{1,\ldots,L_{row}} E_{T|h}[d_\phi(\mathbf{H}, \tilde{\mathbf{H}})] \quad (8)$$

where L_{col}, L_{row} represent the number of column and row clusters, $E_{H|t}, E_{T|h}$ are the expectations under marginal distribution of ν by setting $T = t$ and $H = h$.

Considering our problem, filtered hashtag&tag collections and topic collections are $\mathcal{H}$ and $\mathcal{T}$ respectively. With hashtag&tag-topic distribution $p(\mathbf{z}^{all}|\mathbf{h}^{T,Y,F})$, we build matrix $\mathbf{H}_{N_h \times N_t}$, in which N_h, N_t represent the number of filtered hashtags&tags and topics. We introduce a method named

[4]We further obtain 179,008 tweets from Twitter, 316,196 images from Flickr, 369,101 videos from YouTube

Hashtag&Tag-Topic Co-Clustering with Bilateral Regularization (HTCCB) to cluster the filtered hashtags&tags. With the information of semantic connection of topics and hashtag&tag co-occurrence, we update the second and third subproblems in Bregman co-clustering separately.

Topic clustering is considered in the second subproblem. The optimal topic clustering γ is presumed to not only consider the topic-hashtag&tag distribution but also the topic-word information. With matrix $\mathbf{T}_{N_t \times |\mathcal{W}^{all}|}$ consists of topic-word distribution $p(\mathcal{W}^{all}|\mathbf{z}^{all})$, row clustering is processed on $\mathbf{T}$ with column clustering on $\mathbf{H}$ at the same time. The news optimal function is:

$$\gamma_{i+1}(t) = \arg \min_{1,\ldots,L_{col}} E_{H|t}[d_\phi(\mathbf{H}, \tilde{\mathbf{H}})] + E_{W|t}[d_\phi(\mathbf{T}, \tilde{\mathbf{T}})] \tag{9}$$

where the right part of the equation represents the row clustering on $\mathbf{T}$, $E_{H|t}$ denotes expectation according to marginal distribution by setting $T = t$, and the production of $\tilde{\mathbf{T}}$ is similar with the processing of $\tilde{\mathbf{H}}$. Because the row clustering of $\mathbf{T}$ is what we concerned, the right part is practically the one-sided Bregman clustering problem [17].

Similar to the assumption in Section III, we presume that hashtag or tag co-occurring in an item have high possibility of being classified into the same subtopic and with the assumption we address the hashtag&tag clustering subproblem. With element O_{ij} representing the times of co-occurrence of i^{th} and j^{th} hashtags/tags, we construct matrix $\mathbf{O}_{N_h \times N_h}$ to cluster the filtered hashtags&tags. Similar to the processing of Eqn. (9), the optimal hashtag&tag clustering ρ is coherent with the clustering on $\mathbf{O}$. Then the modified optimal function is:

$$\rho_{i+1}(h) = \arg \min_{1,\ldots,L_{row}} E_{T|h}[d_\phi(\mathbf{H}, \tilde{\mathbf{H}})] + E[d_\phi(\mathbf{O}, \tilde{\mathbf{O}})] \tag{10}$$

By changing Eqn. (7) (8) with Eqn. (9) (10), the hashtag&tag-topic co-clustering with bilateral regularization is processed iteratively on three subproblems.

D. Search Result Demonstration

After the co-clustering, L_{row} hashtag&tag clusters $\{C_1, C_2, \ldots, C_{L_{row}}\}$ are generated for each query respectively and each cluster is composed of N_{C_l} hashtags&tags. To measure the significance of hashtag/tag belonging to C_l, we generate the cluster-hashtag&tag weight $p(h|C_l)$ of each hashtag/tag. The demonstration section contains two parts: search result organization and description. Search results are organized under a cluster-hashtag&tag-item hierarchy structure (illustrated with Fig. 18). Inner items of hashtag/tag are demonstrated chronologically. Inner hashtags&tags of cluster are sorted with cluster-hashtag&tag weight $p(h|C_l)$ in descending order. When ranking clusters, we utilize importance score of them to sort and the importance score is based on two rules: (1) The number of times that hashtags&tags emerge in the search results. (2) The semantic similarity between clusters.

We then present how to calculate the semantic similarity between clusters. The cluster-topic distribution $p(z^t; z^t \in \mathbf{z}^{all}|C_l)$ as:$p(z^t|C_l) = \sum_{h \in C_l} p(h|C_l) \cdot p(z^t|h)$ can be calculated with hashtag&tag-topic distribution $p(z^t|h)$ and cluster-hashtag&tag distribution $p(h|C_l)$. Then we can get the semantic similarity κ_{ij} of cluster C_i and C_j:

$$\kappa_{ij} = exp\left(-\frac{\sum_{z^t \in \mathbf{z}^{all}} (p(z^t|C_i) - p(z^t|C_j))^2}{2\sigma^2}\right) \tag{11}$$

where σ represents the average of pairwise Euclidean distance between clusters. By minimizing the cost function, we then get the importance score of clusters $\boldsymbol{\eta} = [\eta_1, \eta_2, \ldots, \eta_{L_{row}}]$:

$$Q(\boldsymbol{\eta}) = \sum_{i=1,j=1}^{L_{row}} \kappa_{ij} \left(\frac{1}{\sqrt{D_{ii}}}\eta_i - \frac{1}{\sqrt{D_{jj}}}\eta_j\right)^2 + \psi \sum_{i=1}^{L_{row}} (\eta_i - U_i)^2 \tag{12}$$

where $D_{ii} = \sum_{j=1}^{L_{row}} \kappa_{ij}$, ψ denotes the weight parameter, and U_i denotes the frequency that hashtags/tags within C_i used in the search results. The cost function can be optimized by iteratively updating the importance score:[5] $\boldsymbol{\eta}^{(t+1)} = \frac{1}{1+\psi}(\boldsymbol{\eta}^{(t)} S + \psi U)$ where $S = D^{-1/2} W D^{-1/2}$, $D = Diag(D_{11}, D_{22}, \ldots, D_{L_{row}L_{row}})$, and $U = [U_1, U_2, \ldots, U_{L_{row}}]$. $\boldsymbol{\eta}^*$ can be adopted for ranking the hashtag&tag clusters after convergence. As introduced above, the filtering process have already filtered out the hashtags&tags in low quality, so that there is no need to remove the clusters in row rank since they still reflect the query from certain point of view.

In terms of search result description, the hashtag&tag clusters are assumed to be related to subtopics, our solution is to create semantic description for each cluster. In detail, the cluster-word semantic weight can be generated with cluster-topic distribution and topic-word distribution:

$$p(w|C_l) = \sum_{z^t \in \mathbf{z}^{all}} p(z^t|C_l) \cdot p(w|z^t) \tag{13}$$

Consequently, each cluster (subtopic) can be described with 5 10 words with the highest $p(w|C_l)$.

V. Experiments

Based on the same dataset described in data analysis section, we present the experimental results of the four stages in the following separately.

A. Hashtag & Tag Filtering

As introduced in the solution section, given a query q, Spectral Clustering is conducted on the graph G established with h^{all} hashtags and tags assigned to the search results of q.

Based on the number of hashtags shown in Fig. 10, the number of cluster is set as $L_{row} = 8$ in most cases (L_{row} ranges in [5, 9] in some cases based on the number of hashtags and tags from query q).

Next, L_{row} clusters which divide graph G into subgraphs $\{G_1, G_2, \ldots, G_{L_{row}}\}$ are generated through clustering.

PageRank is than adopted within each subgraph and $\lambda = 0.85$ which is an empirical setting of most PageRank applications.

[5]Detailed processing is introduced in [19].

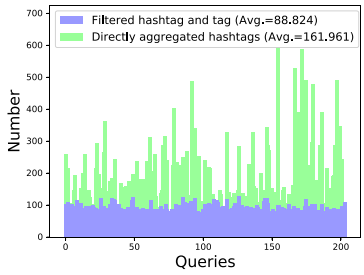


Fig. 10. Quantity Comparison on hashtag and tag before and after filtering.

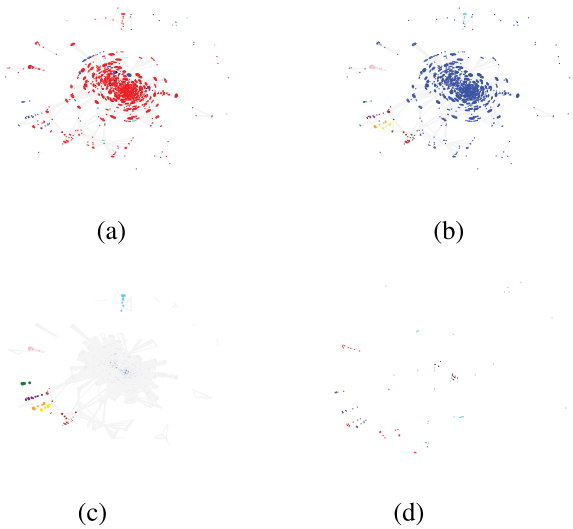


Fig. 11. An example of filtering process. Please use high resolution for a better view.

We make the assumption that the higher rank score a hashtag or tag get by PageRank, the closer relation to the query it has. In this way, top tags and hashtags are selected within a subgraph and for each subgraph, at most 8 tags from a certain OSN are selected to guarantee there is not too much noisy information. After the PageRank process, we obtain $N_h = 88.82$ hashtags and tags for each query on average and there is a significant decline in quantity that can be observed. In addition, we make a further analysis to find the source platform of these filtered hashtags and tags. From the data analysis, the proportions of the three platforms are 10.5% for Twitter, 56.20% for Flickr and 63.3% for YouTube. Since there are overlapped items, sum of the proportions is greater than 1.

We provide an example of filtering process in Fig. 11 of the query 'Nba All Star 2018'. Fig. 11(1) shows the constructed graph before filtering where the imbalance can be seen intuitively and tags and hashtags from different platforms are marked with different colors where cyan for Twitter, blue for Flickr, red for YouTube and black for overlapped items. Fig. 11(2) shows the result of Spectral Clustering where different clusters are marked with different colors. Fig. 11(3) shows the result of PageRank inner subgraphs of previous step, the size of nodes represents the score generated by PageRank. For example, the cluster with most tags and hashtags (i.e. the blue cluster in the figure) are ranked as: 'nba,' 'basketball,' 'NBA,' 'lebron james,' 'sports'... etc. Filtered tags and hashtags with are show in Fig. 11(4) where the meaning of color is the same to Fig. 11(1). It is shown that after filtering, the imbalance are solved and the diversity is still kept.

TABLE V
NFr SCORE TO EXAMINE HASHTAG AND TAG USAGE DIFFERENCE BETWEEN OSNS.

Twitter&Flickr	Twitter&Youtube	Flickr&Youtube
0.3230	0.3744	0.4207

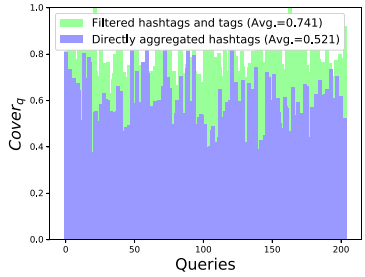


Fig. 12. Topic coverage comparison.

In addition to the example, we further make some quantitative experiments to show the result of stage 1. Following the data analysis in Section III-B from different perspectives, we validate the method from corresponding different point of view, which includes diversity, topic coverage and semantics.

1) Diversity: Table V illustrates the NFr score of filtered tags and hashtags between OSNs. It is shown that the NFr score is similar before and after clustering and that means there is not much loss in terms of usage pattern and diversity.

2) Topic Coverage: To validate the effectiveness of our method, first we make a comparison between the filtered hashtags&tags and the hashtags directly aggregated from three OSNs. Similar to the comparison of topic coverage described in Section III-B, coverage score of query q is then calculated as: $Cover_q^{all} = 1/3(Cover_q^{Twitter} + Cover_q^{Flickr} + Cover_q^{YouTube})$.

Fig. 12 shows that tags and hashtags after filtering achieve a higher topic coverage score than directly integrating hashtags with less quantity. It indicates that our method maintains the diversity with clustering and reduces the quantity of tags and hashtags as much as possible at the same time with PageRank. Now we have shown that filtered hashtags and tags have advantage in topic coverage.

3) Semantics: Semantic features of hashtag and tag themselves also need to be compared. First we follow the experiment in Section III-B and compare the semantic information of search query. It is shown that 12.4% filtered hashtags and tags contains the semantic information related to the search query on average and 7.7% for directly aggregated hashtags. Our method keeps the semantic information from this point of view. We further make a comparison based on the semantic information in topical structures. Following the segmentation method and topic modeling method in Section III-B, we make a comparison on semantic information coverage. For each word distribution $p(w_o|z_o)$ of

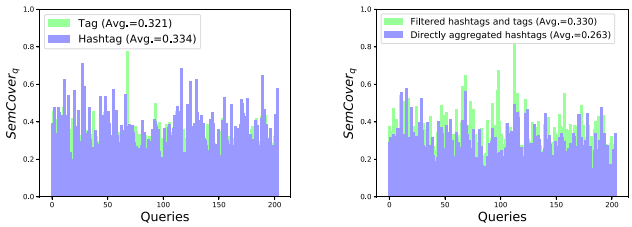


Fig. 13. Semantic coverage comparison.

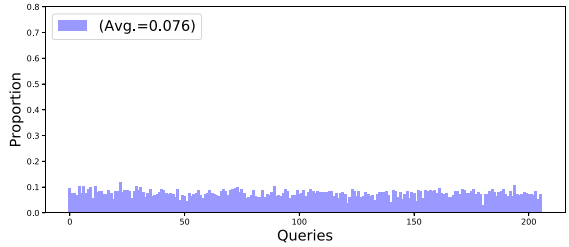


Fig. 14. Vocabulary overlap proportion.

topics from a certain query q, we examine the segmentations of hashtags and tags whether themselves or lower case versions appear in w_o. Take hashtag as an example, the semantic coverage score can be calculated as:

$$SemCover_q = \frac{\sum_{n=1}^{N_h} \mathbb{I}(\bigvee_{h^q}(w_{h^q} \in w_o))}{N_h} \quad (14)$$

where N is the number of hashtags, h^q is the hashtags retrieved by q, w_{h^q} is the word set segmented from h^q, w_o is the word distribution of topics. The semantic coverage score measures the percentage of hashtag or tag captures semantic information in the topics directly and comparison on semantic coverage score is shown in Fig. 13. Before filtering, tags and hashtags are compared on Flickr and YouTube which is illustrated with Fig. 13 (1) and $SemCover_q = 1/2(SemCover_q^{Flickr} + SemCover_q^{YouTube})$. It is shown that the semantic score is similar in tags and hashtags. Filtered hashtags&tags and hashtags directly integrated are compared on three platforms illustrated with Fig. 13 (2) and $SemCover_q = 1/3(SemCover_q^{Twitter} + SemCover_q^{Flickr} + SemCover_q^{YouTube})$ here. It is shown that the semantic score gets higher after filtering, indicating that our filtering method maintains the semantic information coverage of hashtags and tags.

B. Results of Representation Learning

As elaborated in the solution section, given a query q, hLDA is executed on collections $\mathcal{D}_q^T, \mathcal{D}_q^Y, \mathcal{D}_q^F$ over the unified vocabulary set $\mathcal{W}^{T,Y,F}$ respectively. Parameters are set with $\alpha = 10$ $\gamma = 1$ and $\eta = 0.1$ empirically [11]. After hLDA, random walk is processed with $\alpha = 0.5$ [14] to construct the unified set $\mathcal{W}^{all}$. Fig. 14 illustrates the percentage of overlapping vocabulary set $\mathcal{W}^{overlap}=\mathcal{W}^T \bigcap \mathcal{W}^Y \bigcap \mathcal{W}^F$. As the figure shows, only about 7.6% vocabulary is shared cross OSNs, which emphasizes the importance and necessity of vocabulary integration by random walk. Table VI shows some of the learned leaf topics on different OSNs

TABLE VI
VISUALIZATION OF TOPICS FROM DIFFERENT OSNS OF QUERY "NBA ALL STAR 2018".

Platform	Topic
Twitter	all-star,popovich,coach,oklahoma ...
	harden,live, james,night,...
Flickr	pictures,sports,media,basketball,gamers...
	stephen,cavaliers,highlights,jordan...
Youtube	durant,kevin,players,highlights,leonard,...
	thunder,talent, george,paul, playoffs,...

for the query "Nba All Star 2018". Each topic is described by the top probable words, where the words in the original vocabulary space is marked with black and the words extended by random walk from Twitter are highlighted with cyan, Flickr with blue and YouTube with red. Two observations are obtained therefrom: (1) The learned topics covers a wide range of themes and there are shared themes and words on different OSNs. (2) Random walk bridges the vocabulary spaces and provide cross-OSN words for better expression of topics.

C. Results of Hashtag & Tag Clustering

As show in Section V-A, filtered hashtags and tags maintain the diversity and topic coverage. However, the topic coverage score only captures the information based on the search results directly generated by the search query and the clustering only considers the co-occurrence information. Furthermore, to validate the effectiveness of the co-clustering, we elaborate the experimental setting and results of co-clustering in subsections later.

1) Experimental Setting: After clustering, hashtag&tag clusters $\{C_1, C_2, \ldots, C_{L_{row}}\}$ are generated with hashtag&tag-topic distribution $\mathbf{H}$. To evaluate the quality of clustering, we utilize Normalized Mutual Information (NMI) [26] which is an effective metric. With label assignment C_1 on h hashtags&tags, the entropy can be calculated as $H(C_1) = \sum_{i=1}^{|C_1|} P(i)\log(P(i))$, where $P(i) = |C_{1i}|/h$ is the probability of selecting a hashtag/tag randomly from C_1 and the hashtag/tag belonging a cluster C_{1i}. Then the NMI between two cluster label assignments C_1 and C_2 can be calculated as:

$$NMI(C_1, C_2) = \frac{\sum_{i=1}^{|C_1|}\sum_{j=1}^{|C_2|} P(i,j)\log\left(\frac{P(i,j)}{P(i)P(j)}\right)}{\sqrt{H(C_1)H(C_2)}} \quad (15)$$

where $H(C_2) = \sum_{j=1}^{|C_2|} P(j)\log(P(j))$, $P(j) = |C_{2j}|/h$ and $P(i,j) = |C_{1i} \cap C_{2j}|/h$.

To measure the clusters with NMI, the clusters generated by 100 randomly chosen queries are labeled by 5 volunteers manually. With regard to the labeling strategy, volunteers were under the guideline that they need to divide the hashtags&tags into clusters where the number of clusters are determined by the number of hashtags&tags as introduced in subsection of hashtag&tag filtering. After that, the co-clustering is processed with the labeled truth as the number of hashtag clusters L_{row}. The number of topic cluster L_{col} varies between 5 and 30 with the step of 5 and the NMI of them is reported in Fig. 15 on 20 randomly selected queries. As the figure shows, the NMI keep

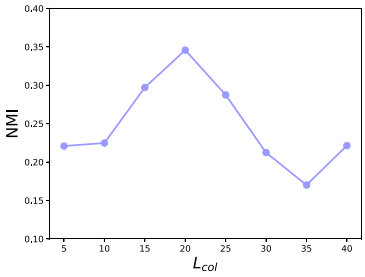


Fig. 15. Comparison on different settings of L_{col}.

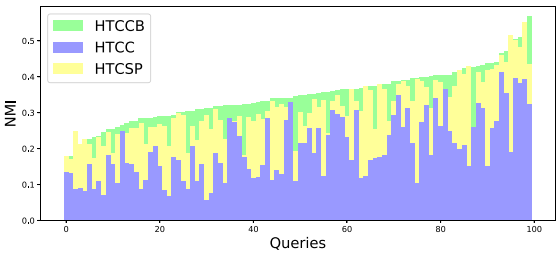
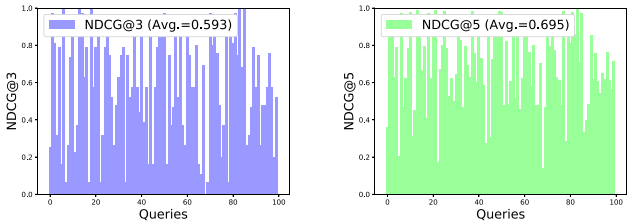


Fig. 16. Experiment performance comparison with NMI of different methods.



(a) (b)

Fig. 17. NDCG for different queries.

increasing till $L_{col} = 20$ and decreases afterwards. That indicates dividing underlying number of leaf topics to 20 clusters achieves the better result, thus we set $L_{col} = 20$ in our experiments. For other parameters of co-clustering, we follow the empirical settings from [12] and select basis $\mathcal{C}_2$ and Squared Euclidean distance as d_ϕ.

2) Experimental Results and Analysis: Fig. 16 illustrates the performance of different models. Original Bregman co-clustering processed on filtered hashtags&tags is denoted as $HTCC$, the hashtag&tag clustering result of Spectral Clustering is denoted as $HTCSP$, which is introduced in subsection of hashtag&tag filtering where only filtered hashtags and tags and co-occurrence information are considered, and the proposed model with bilateral regularization is denoted as $HTCCB$ The experimental results are shown in ascending order of the NMI of $HTCCB$. It is shown that the bars $HTCCB$ appears atop other curves of bars for most of the queries, demonstrates the advantages of bilateral regularization. An interesting observation is that $HTCSP$ achieves acceptable performance and it indicates that co-occurrence and usage information are both important for hashtag&tag integration.

D. Search Result Demonstration

1) Experimental Setting: We focus on the evaluation of hashtag&tag cluster ranking at this subsection. After hashtag&tag-topic clustering, the search result demonstration is processed on the hashtag&tag clusters with the parameter $\psi = 0.5$. We provide a simple example to illustrate the important score which considering both frequency and semantic similarity. The first cluster which comes from query 'Nintendo Labo' includes hashtags and tags such as 'NintendoLabo,' 'amazon,' 'NintendoSwitch' and whose score is 0.329. The second cluster includes 'games,' 'Mario,' 'Labo ASCA,' 'mosaic,' 'news' and so on and whose score is 0.180. The second cluster's total usage of tags and hashtags is greater than the first cluster (Specific numbers are 45 and 16), since both frequency and semantic similarity between clusters are considered, the first cluster achieves the highest importance score.

To evaluate the goodness of ranking quantitative and general, we employ Normalized Discounted Cumulative Gain (NDCG) and the NDCG metric is calculated as:

$$NDCG@k = \frac{1}{Z}\sum_{j=1}^{k}\frac{2^{r(j)}-1}{\log(1+j)} \tag{16}$$

where $r(\cdot)$ represents the relevance score between the query and the corresponding cluster which is calculated by Eqn. (12).

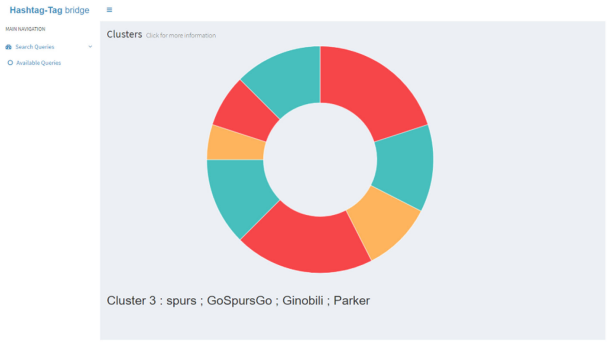
2) Experimental Results and Analysis: To make use of NDCG, 5 volunteers voted for the top-5 appropriate clusters for each of the examined 100 queries. The ground-truth is the mean score of the votes from volunteers. NDCG@3 and NDCG@5 for the queries are illustrated in Fig. 17(1)(2). Observation can be derived that when considering top-5 clusters, the rank method achieves a high average NDCG which is 69.5%. Note that ground-truth of most queries is 8 clusters, this high NDCG@5 demonstrates the effectiveness of the solution. When only rank-1 cluster is considered, the rank method still achieves a satisfied performance with average NDCG@1=34%.[6]

As mentioned above, overlapping tags and hashtags play an important role in our framework. We calculate the percentage of overlapping items in top-3 ground-truth clusters and it achieves 30.7% on average over the queries. It shows the importance of overlapping tags and hashtags.

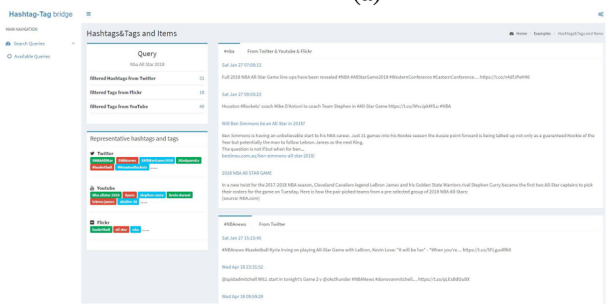
We develop a demo that shows the demonstrated search results on the website.[7] After querying, the related search results with corresponding hashtags&tags from Twitter, Flickr and YouTube are automatically collected and processed. Search results are organized and demonstrated under the cluster-hashtag&tag-item structure hierarchy, as illustrated in Fig. 18. Fig. 18(a) shows the clusters layer, the proportion of each cluster in pie chart is determined by the rank score and a cluster is described with the words extracted according to Eqn. (13). Clicking certain hashtag&tag cluster from the pie chart, the assigned hashtags&tags with related items within the cluster are displayed as in Fig. 18(b). Items are listed chronologically in a certain hashtag or tag and in this

[6]Since NDCG@1 is either 0 or 1, detailed NDCG@1 results are not illustrated for each query in Fig. 17.

[7][Online]. Available: https://hashtagasbridge.github.io/HashtagTag/



(a)



(b)

Fig. 18. Illustration of the cross-OSN event search demo interface.

way, search results of a certain query and demonstrated in the cluster-hashtag&tag-item structure.

VI. Conclusion

This study has pointed out the cross-OSN immersive search problem. A preliminary hashtag and tag-centric solution is introduced. Hashtags&tags are collected based on the corresponding OSNs and exploited to organize the search results from different OSNs to help understand social events in a coarse-to-fine scheme. This work, however, is more of an attempt than a solid real-world application. Based on this, suggestions for future research are as follows: (1) Consider the time distribution of the collected hashtags and tags, to visualize, track and further forecast the evolution of events among OSNs; (2) Explore the social interaction potential of hashtag and tag, e.g., analyzing users with hashtags or groups using tags and creating event-oriented user channels to enrich the immersive search experience. (3) Utilize more contextual information, e.g., geographic information as well as hyperlink, and integrate them into a unified framework to provide a better performance.

References

[1] W. Y. Lee, W. H. Hsu, and S. Satoh, "Learning from cross-domain media streams for event-of-interest discovery," *IEEE Trans. Multimedia*, vol. 20, no. 1, pp. 142–154, Jan. 2018.

[2] M. Yan, J. Sang, and C. Xu, "Unified youtube video recommendation via cross-network collaboration," in *Proc. 5th ACM Int. Conf. Multimedia Retrieval*, 2015, pp. 19–26.

[3] M. Yan, J. Sang, C. Xu, and M. S. Hossain, "Youtube video promotion by cross-network association: @Britney to advertise gangnam style," *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1248–1261, Aug. 2015.

[4] S. D. Roy, T. Mei, W. Zeng, and S. Li, "Socialtransfer: Cross-domain transfer learning from social streams for media applications," *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 649–658.

[5] L. Vaughan and Y. Zhang, "Equal representation by search engines? A comparison of websites across countries and domains," *J. Comput.-Mediated Commun.*, vol. 12, no. 3, pp. 888–909, 2007.

[6] J. M. Van Thong *et al.*, "Speechbot: An experimental speech-based search engine for multimedia content on the web," *IEEE Trans. Multimedia*, vol. 4, no. 1, pp. 88–96, Mar. 2002.

[7] J. Sang, Z. Deng, D. Lu, and C. Xu, "Cross-OSN user modeling by homogeneous behavior quantification and local social regularization," *IEEE Trans. Multimedia*, vol. 17, no. 12, pp. 2259–2270, Dec. 2015.

[8] Y. Jiang, J. Wang, Q. Wang, W. Liu, and C. Ngo, "Hierarchical visualization of video search results for topic-based browsing," *IEEE Trans. Multimedia*, vol. 18, no. 11, pp. 2161–2170, Nov. 2016.

[9] A. Popescu and G. Grefenstette, "Social media driven image retrieval [C]," in *Proc. 1st ACM Int. Conf. Multimedia Retrieval*, 2011, pp. 1–8.

[10] W. X. Zhao, J. Jiang, J. Weng, J. He, and E. P. Lim, "Comparing twitter and traditional media using topic models," in *Proc. Eur. Conf. Inf. Retrieval*, 2011, pp. 338–349.

[11] T. L. Griffiths *et al.*, "Hierarchical topic models and the nested chinese restaurant process," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 17–24.

[12] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D. S. Modha, "A generalized maximum entropy approach to bregman co-clustering and matrix approximation," *J. Mach. Learn. Res.*, vol. 8, pp. 1919–1986, 2007.

[13] B. K. Bao, C. Xu, W. Min, and M. S. Hossain, "Cross-platform emerging topic detection and elaboration from multimedia streams," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 11, no. 4, 2015, Art. no. 54.

[14] Z. Deng, J. Sang, and C. Xu, "Personalized celebrity video search based on cross-space mining," in *Proc. Pacific-Rim Conf. Multimedia*, 2012, pp. 455–463.

[15] D. Liu *et al.*, "Tag ranking," in *Proc. 18th Int. Conf. World Wide Web*, 2009, pp. 351–360.

[16] W. H. Hsu, L. S. Kennedy, and S. F. Chang, "Video search reranking through random walk over document-level context graph," in *Proc. 15th ACM Int. Conf. Multimedia*, 2007, pp. 971–980.

[17] A. Banerjee *et al.*, "Clustering with Bregman divergences," *J. Mach. Learn. Res.*, vol. 6, pp. 1705–1749, 2005.

[18] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 321–328.

[19] D. Lu, X. Liu, and X. Qian, "Tag-based image search by social re-ranking," *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1628–1639, Aug. 2016.

[20] P. Diaconis and R. L. Graham, "Spearman's footrule as a measure of disarray," *J. Roy. Statist. Soc. Ser. B*, vol. 39, no. 2, pp. 262–268, 1977.

[21] J. Bar-Ilan, M. Mat-Hassan, and M. Levene, "Methods for comparing rankings of search engine results," *Comput. Netw.*, vol. 50, no. 10, pp. 1448–1463, 2006.

[22] A. K. McCallum, "Mallet: A machine learning for language toolkit[J]," 2002. [Online]. Available: http://mallet.cs.umass.edu.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[24] L. Yang *et al.*, "We know what you# tag: Does the dual role affect hashtag adoption," in *Proc. 21st Int. Conf. World Wide Web*, 2012, pp. 261–270.

[25] J. Sang, C. Xu, and R. Jain, "Social multimedia ming: From special to general," in *Proc. IEEE Int. Symp. Multimedia*, 2016, pp. 481–485.

[26] A. Strehl and J. Ghosh, "Cluster ensembles—A knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, 2002.

[27] G. A. Miller, "WordNet: A lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[28] M. Ames and M. Naaman, "Why we tag: Motivations for annotation in mobile and online media," in *Proc. SIGCHI Conf. Human Factors Comput. Syst. ACM*, 2007, pp. 971–980.

[29] L. Page *et al.*, "The PageRank citation ranking: Bringing order to the web," Stanford InfoLab, Stanford, CA, USA, 1999.

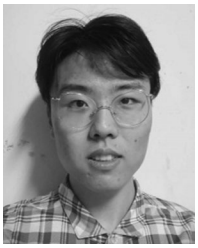
[30] A. N. Langville and C. D. Meyer, "A survey of eigenvector methods for web information retrieval," *SIAM Rev.*, vol. 47, no. 1, pp. 135–161, 2005.

[31] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 849–856.

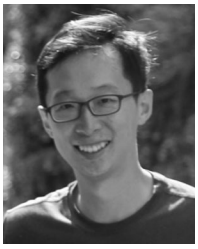
[32] Y. Shi, M. Larson, and A. Hanjalic, "Tags as bridges between domains: Improving recommendation with tag-induced cross-domain collaborative filtering," in *Proc. Int. Conf. User Model., Adaptation, Personalization*, 2011, pp. 305–316.

[33] M. Enrich, M. Braunhofer, and F. Ricci, "Cold-start management with cross-domain collaborative filtering and tags," in *Proc. Int. Conf. Electron. Commerce Web Technologies*, 2013, pp. 101–112.

[34] L. Barrington, D. O'Malley, D. Turnbull, and G. Lanckrie, "User-centered design of a social game to tag music," in *Proc. ACM SIGKDD Workshop Human Comput.*, 2009, pp. 7–10.
[35] E. Denton, J. Weston, M. Paluri, L. Bourdev, and R. Fergus, "User conditional hashtag prediction for images," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 1731–1740.
[36] Y. Gao, J. Sang, T. Ren, and C. Xu, "Hashtag-centric immersive search on social media," in *Proc. ACM Multimedia Conf.*, 2017, pp. 1924–1932.
[37] C. Baziotis, N. Pelekis, and C. Doulkeridis, "Datastories at semeval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis," in *Proc. 11th Int. Workshop Semantic Eval.*, 2017, pp. 747–754.
[38] Y. Hu, L. Zheng, Y. Yang, and Y. Huang, "Twitter100k: A real-world dataset for weakly supervised cross-media retrieval," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 927–938, Apr. 2018.
[39] J. Zhang, Y. Yang, L. Zhuo, Q. Tian, and X. Liang, "Personalized recommendation of social images by constructing a user interest tree with deep features and tag trees," *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2762–2775, Nov. 2019.
[40] S. Rahman, S. Khan, and N. Barnes, "Deep0tag: Deep multiple instance learning for zero-shot image tagging," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 242–255, Jan. 2020.

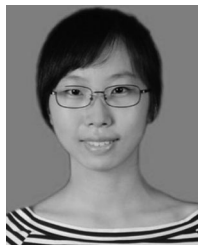


Yuqi Gao received the bachelor's degree in software engineering from Nanjing University, Nanjing, China, in 2017, where he is currently working toward the graduate degree. His research interests include social multimedia computing and web data mining.



Jitao Sang received the B.E. degree from the Southeast University, Nanjing, China and the Ph.D. degree (Hons.) from CASIA, Beijing, China with the Special Prize of President Scholarship. He is a Professor with Beijing Jiaotong University, Beijing, China. He has co-authored more than 80 peer-referenced papers in multimedia-related journals and conferences. His research interests include social multimedia computing, web data mining, and machine learning interpretation.

Chengpeng Fu received the bachelor's degree from Liaoning University, Shenyang, China, in 2018. He is currently working toward the graduate degree with Beijing Jiaotong University, Beijing, China. His research interests include knowledge graph and social multimedia computing.



Zhengjia Wang is currently an undergraduate student with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China. Her research interests include information retrieval and social event mining.

Tongwei Ren received the bachelor's, master's, and doctoral degrees from Nanjing University, Nanjing, China, in 2004, 2006, and 2010, respectively. He joined the Software Institute of Nanjing University as an Assistant Professor in 2010, and at present he is an Associate Professor with Nanjing University. He visited the Hong Kong Polytechnic University in 2008 and the National University of Singapore from 2016 to 2017. He has authored more than 50 papers in international journals/conferences, such as TIP, TOMM, TNNLS, MM, ICCV, and AAAI, and won the best paper honorable mention of ICIMCS 2014, the best paper runner-up of PCM 2015, the champions of ECCV 2018 PIC challenge and MM 2019 VRU challenge, and the second places of ICME 2019 SVU challenge and MM 2019 CBVRP challenge. His research interest mainly includes visual multimedia computing and its application.

Changsheng Xu (Fellow, IEEE) is a Distinguished Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. He holds 40 granted/pending patents and authored/coauthored more than 300 refereed research papers in these areas. His research interests include multimedia content analysis/indexing/retrieval, pattern recognition, and computer vision. He was an Associate Editor, a Guest Editor, the General Chair, the Program Chair, the Area/Track Chair, a Special Session Organizer, the Session Chair, and a TPC Member for more than 20 prestigious multimedia journals of the IEEE and ACM, conferences, and workshops, including the IEEE TRANSACTIONS ON MULTIMEDIA, *ACM Transactions on Multimedia Computing*, and Communications and Applications and ACM Multimedia conferences. He is an IAPR Fellow and ACM Distinguished Scientist.