Proceedings of the Tenth International AAAI Conference on
Web and Social Media (ICWSM 2016)

Cross Social Media Recommendation

Xiaozhong Liu
School of Informatics and Computing
Indiana University Bloomington
Bloomington, IN, USA, 47405
liu237@indiana.edu

Tian Xia
School of Information Resource
Management, Renmin University of China
Beijing, China, 100872
xiat@ruc.edu.cn

Yingying Yu
College of Transportation Management
Dalian Maritime University
Dalian, China, 116026
uee870927@126.com

Chun Guo
School of Informatics and Computing
Indiana University Bloomington
Bloomington, IN, USA, 47405
chunguo@indiana.edu

Yizhou Sun
College of Computer and Information Science
Northeastern University
Boston, MA, USA, 02115
yzsun@ccs.neu.edu

Abstract

The proliferation of rich social media data revolutionizes the way people perceive and understand the world. Unfortunately, so far, there does not exist a single social media system that efficiently *globalizes* users around the world. Two well-known social media systems, Twitter and Facebook, are strictly blocked in mainland China for political reasons. Similarly, the second-largest microblogging system in the world, Sina Weibo, features a default system language of Chinese, which rules out many users from other countries. As a result, if we are interested in modeling the knowledge of the world, all research findings based on a single social media system (within a bubble) can be biased, and the social networks or knowledge networks generated from a single system or social community cannot fully represent people from around the world. In this study, we generate a pseudo-social heterogeneous network - Pseudo Global Social Media Network (PGSMN), which bridges the topics of Twitter and Weibo. On this network, all Weibo and Twitter nodes are interconnected via an interim knowledge layer, and user or topic nodes from Twitter can randomly walk to the nodes on Weibo (via different kinds of paths), and vice versa, which enables cross-network information recommendation and knowledge globalization.

1 Introduction

Social media is bringing about significant changes in how people perceive and make sense of their world (Pak and Paroubek 2010), and the algorithms and methods for social media mining are well-documented. Millions of individuals communicate with each other through a variety of social media platforms. Unfortunately, there is no social media system that efficiently globalizes users from around the world. For example, two popular social media systems, Twitter and Facebook, are strictly blocked in mainland China due to political concerns (Wikipedia 2014), which means 21.97%[1] of Internet users are excluded from these systems and social networks. Similarly, the world's second-largest microblogging system, Sina Weibo, boasted more than 176 million active users by 2015, but because the system's default language is Chinese, almost all its users come from China or speak Chinese. Because Twitter and Weibo serve different communities, users in different networks may be interested in different kinds of topics.

The degree to which isolated communities (a.k.a. "bubbles") are shaped and separated by the lack of direct communication on social media is largely unknown and remains an under-studied topic in the scientific literature. Such understanding must, however, form the foundation for the development of information technology that can effectively address this issue by connecting otherwise separated entities. In this study, we focus on constructing **"Pseudo Global Social Media Network (PGSMN)"** to interconnect the users and topics across two distinct social media 'islands'. To the best of our knowledge, PGSMN will be the first pseudo-social media network to interconnect users and topics from different language, culture, and network bubbles. PGSMN can be important for three reasons. First, for advertisers or law enforcement agencies, PGSMN can help to find out the topics and users following or related to them across Twitter and Weibo and to explore some potential application. Second, by leveraging PGSMN, scholars can investigate and compare the similarity/difference of Chinese and American societies when consuming the same or similar topics. Last but not least, PGSMN enables cross-social media information recommendation, e.g., recommend Twitter topics or users to Weibo users, and vice versa. While not all users are interested in the novel topics from another social media, cross social network information recommendation alleviates the problem of information isolation, and some particular group of users may potentially benefit from it, e.g., some Chinese Weibo users can be interested in the 'outside' dynamic topics that are censored by government.

There are two main challenges for PGSMN construction. First, obviously, Twitter and Weibo users and topics belong to two distinct social networks, and no physical linkage exists among them. Very few users own both Twitter and Weibo accounts (Chinese users cannot register Twit-

[1] Statistics of China Internet Users (2014): http://www.internetlivestats.com/internet-users/china/

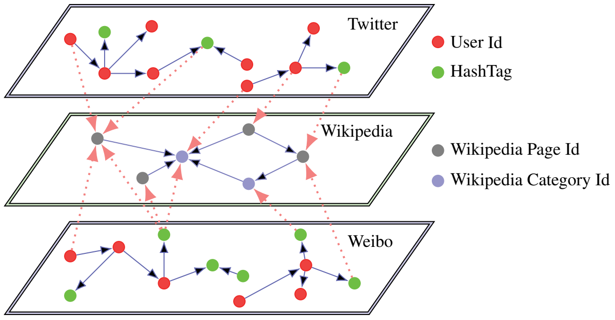

Figure 1: 3-layer Pseudo Global Social Media Network

ter account), and we cannot map the information/metadata across these networks[2]. For this reason, most prior works in "user identification across multiple social networks" cannot be applied to solve this problem and interconnect those networks. Second, Weibo and Twitter, in most cases, are written in two different languages. Meanwhile, because of the message length limitation, users tend to use very colloquial languages and acronyms to express complex semantics, which challenges traditional machine translation algorithms.

Unlike most related studies, we won't use machine translation for PGSMN construction. Instead, PGSMN is a three layer heterogeneous graph, which includes a Twitter layer, a Weibo layer, and a Wikipedia bridge layer (as Figure 1 shows). The Wikipedia multi-lingual corpus is used as a "Rosetta Stone" in this study to connect culturally disjoint language and network bubbles. All the Weibo and Twitter users and topics are interconnected by Wikipedia articles or category nodes and semantic paths on the Wikipedia category tree. A number of different meta-paths will be employed for user and topic recommendation across Weibo and Twitter via learning-to-rank. The rich linkages between Wikipedia pages and categories provide great potential to interconnect users and topics from different communities.

The contribution of this paper is threefold. First, we propose to solve the innovative information-recommendation problems across different social networks and globalize different local communities, i.e., Twitter and Weibo. Different social networks target exclusive communities, and they do not necessarily share the same language or culture background. This work can also be generalized to other social networks. Second, we propose using our three-layer PGSMN to interconnect users and topics from Twitter and Weibo by using Wikipedia and its associated linkages, i.e., incoming/outgoing links among Wikipedia articles, article-categorical links, and Wikipedia category hierarchy links. Two types of links, as Figure 1 shows, are employed to interconnect the nodes between Weibo and Twitter layers, text match the edge between the Twitter/Weibo user/hashtag node and the Wikipedia article node(s), and ESPM (Explicit Semantic Path Mining) edge between the Twitter/Weibo user/hashtag node and the Wikipedia category path(s). Then, by using the paths between and within layers, we can calculate the supervised random walk probability from one node (in Twitter) to another (in Weibo), or vice versa. We also share the constructed PGSMN data to motivate other studies for this newly proposed topic. Last but not least, we provide a case study, recommend Twitter hastag to Weibo user, to verify the usefulness of PGSMN. By employing meta-path and learning to rank, we investigate the optimized recommendation model to recommend Twitter hashtags to Weibo users. Note that we use cross social media hastag recommendation as the case study because of evaluation reason (ground truth is partially available), but, more practically, other recommendation task, e.g., user and advertisement recommendation, can be more significant.

2 Literature Review

[Social Media and User Community]: The proliferation of rich social media data revolutionizes the way people understand the world, and the algorithms and methods for social media mining are well documented (Baucom et al. 2013; Kouloumpis, Wilson, and Moore 2011). The way people both perceive and exploit social media, especially microblogging systems, has been observed and well-documented. For instance, sentiment analysis (Kouloumpis, Wilson, and Moore 2011), social network analysis (Kwak et al. 2010; Ediger et al. 2010), and event prediction (Achrekar et al. 2011) studies are well-documented. Existing studies found that microblogging systems can be used to accurately characterize different social events at a low cost.

Unfortunately, so far, there hasn't been a social media system which efficiently globalizes users around the world. Our study builds on existing knowledge and techniques, and contributes a body of novel problems and algorithms that aims to achieve global knowledge integration.

[Twitter and Weibo Comparison]: While using Twitter to characterize real-world events is well documented, Weibo is becoming an important means to understand the Chinese community. For instance, Zhao et al., (Zhao et al. 2011) employed Weibo data to investigate event discussion by using term-message-user networks and compared it with those found on Twitter. They used random-walk algorithms to study the temporal event information diffusion, and the event is pre-defined by domain expert. Similarly, Guan et al., (Guan et al. 2014) studied 21 hot events of Weibo by utilizing 32 prestigious users.

Unfortunately, due to both language and political barriers, most users from each community can only access one system exclusively. Although most previous studies treat Twitter and Weibo as comparable social media outlets except for language, some other researchers (Li et al. 2012) found that Weibo may have some unique features. Some scholars have only recently became aware of the importance of comparing Weibo and Twitter (Shuai et al. 2014).

[Meta-Path on Heterogeneous Graph]: The concept of meta-path was first proposed in (Sun et al. 2011b), which can systematically capture the semantic relation between objects in a heterogeneous information network scenario. A *meta-path* $\mathcal{P}$ is a path defined on the graph of network schema $T_G = (\mathcal{A}, \mathcal{R})$, and is denoted in the form of $\dot{A}_1 \xrightarrow{R_1}$

[2] Chinese users with foreign IP addresses can register Twitter, but very few Chinese users have a foreign IP.

$\dot{A}_2 \xrightarrow{R_2} \ldots \xrightarrow{R_l} \dot{A}_{l+1}$, which defines a composite relation $R = R_1 \circ R_2 \circ \ldots \circ R_l$ between types $\dot{A}_1$ and $\dot{A}_{l+1}$, where $\circ$ denotes the composition operator on relations. Different meta-path-based mining tasks are studied, including similarity search (Sun et al. 2011b), relationship prediction (Sun et al. 2011a), user-guided clustering (Sun et al. 2012), and recommendation (Liu et al. 2014). It turns out that meta-path serves as a very critical feature extraction tool for most of the mining tasks in a heterogeneous information network. From a random walk viewpoint, one node may random walk to another node based on different meta-paths. In this paper, we propose a novel meta-path-based approach by using a Wikipedia category tree. On the new meta-path, a node on the path could be a path on the Wikipedia category tree. The random walk function, then, is re-defined.

[Cross-Domain User Identification and Recommendation]: Cross-social media user identification has become popular in the recent years. It is common that users register in multiple social media or commercial sites. Identifying the same user in multi-social network environments can enhance the user profiling accuracy (Zafarani and Liu 2013), collaborative filtering performance(Li and Lin 2014), and friend recommendation precision (Zafarani and Liu 2014). Vosecky et al., (2009), for example, employed real-life data, i.e., user name, gender, and addresses from two popular social networks, and a machine learning algorithm to identify the same users across Facebook and StudiVZ. They found the user profile features are important for this task. Similarly, Zhang et al.,(2014) used the profile features (e.g., user description) to identify the same users across Twitter and LinkedIn. Zafarani (2013) identified the same users across 32 popular U.S. sites, including Flickr, Reddit, and YouTube, by using more sophisticated features extracted from username strings. Authors studied user behavior when selecting username, typing pattern, and modifying previous usernames as features to train a user identification function. More recently, Li and Lin (2014) proposed a new user identification method across various domains from collaborative filtering viewpoint (for item rating prediction).

Another kind of related research is cross-domain information recommendation, e.g., recommending movies using book rating data. Fernández-Tobías et al.(2011), for example, suggested the use of semantic-based framework to integrate knowledge of different domains for cross-domain item recommendations. Shi et al. (2011) employed cross-domain tag information to perform cross-domain recommendation. Sahebi and Brusilovsky (2013) used social links and cross-domain community detection to perform cross-domain recommendations.

All of these studies share the same premise that a user can register in different social media or commercial sites, and the selected experimental systems share the same language. Unfortunately, we cannot use these methods to integrate Weibo and Twitter data, because Weibo and Twitter communities are exclusive to each other.

3 Research Methods

In this section, we discuss the methodology in detail, which includes: constructing PGSMN, generating heterogeneous edges to interconnect Weibo and Twitter users and hashtags, and designing a random walk method by using PGSMN.

PGSMN Construction

To achieve the goals of this project, we need to interconnect Weibo and Twitter data. We utilize Wikipedia as the bilingual global knowledge-base to link Weibo and Twitter on a heterogeneous graph because of the following reasons:

1) Wikipedia provides concept definitions in Chinese and English. For instance, in Wikipedia's 2014 March dumps, we find 397,689 important articles (at least three incoming links) defined in both English and Chinese, which cover essential universal knowledge. The English article can be projected into the Twitter topic space, and the Chinese counterpart for the same concept (the same wiki article node) can be used to bridge the Weibo topics.

2) All concepts in Wikipedia are interlinked via Wikipedia hierarchical categories and incoming/outgoing links among Wikipedia articles. For instance, the articles *"NBA"* and *"LeBron James"*are connected via the path *"[wiki article: NBA] $\overset{b}{\rightarrow}$ [wiki category: Basketball] $\overset{b}{\leftarrow}$ [wiki article: LeBron James]"* and path *"[wiki article: LeBron James] $\overset{l}{\rightarrow}$ [wiki article: NBA]"* ($\overset{b}{\rightarrow}$represents *"belong to"* relation, and $\overset{l}{\rightarrow}$represents *"link to"* relation). In other words, all articles in Wikipedia are inter-connected through heterogeneous links and cross-language equivalents. Then, all Twitter and Weibo hashtags/users are also interlinked via a Wikipedia bridge.

3) Both Weibo and Twitter are written in colloquial language, and Wikipedia provides colloquial-language-like Redirected Links for both Chinese and English articles. For instance, the concept *"Patient Protection and Affordable Care Act"* can be redirected from *"Obamacare"*, which can be helpful for social media text mining.

Based on above reasons, we propose Pseudo Global Social Media Network (PGSMN), which is a 3-layer heterogeneous graph including the following information: **Weibo Layer**, which has Weibo user (U_{Weibo}) and Weibo HashTag (HT_{Weibo}) nodes; **Wikipedia bridge layer**, which has Wikipedia article (A) and category (C) nodes. The articles are connected via page hyperlinks, and the category nodes are organized on a category tree. Each node may belong to one or multiple categories. All the article and category nodes are defined in both Chinese and English Wikipedia space, and each article node has at least three incoming links; **Twitter Layer**, which has Twitter user ($U_{Twitter}$) and Twitter HashTag ($HT_{Twitter}$) nodes. Based on this information, we constructed a novel heterogeneous graph (Figure 1), and the relations are defined in the following table:

Each edge on the graph is associated with a weight which denotes the transitioning probability from one node to another. $U_X \overset{E}{\rightarrow} \mathbb{P}_C$ and $HT_X \overset{E}{\rightarrow} \mathbb{P}_C$ are different compared with the other edges, and they utilize a different data structure. For these edges, user or hashtag node links to a path, $\mathbb{P}_C$, on the Wikipedia category tree (part of the Wikipedia

Table 1: Relations in the constructed heterogeneous graph

Edge	Description
Edges within Twitter/Weibo Layer	
$U_{X_i} \xrightarrow{t} U_{X_j}$	User-to-User relation (reply/mention/comment)
$U_{X_i} \xrightarrow{u} HT_{X_j}$	User use HashTag (user message has HashTag)
$HT_{X_i} \xrightarrow{c} HT_{X_j}$	HashTag-HashTag (HashTag co-occur in a message)
Edges between Twitter/Weibo and Wikipedia Layer	
$U_X \xrightarrow{r} A$	User's message content relevant to a Wikipedia article
$HT_X \xrightarrow{r} A$	HashTag's message content relevant to a Wikipedia article
$U_X \xrightarrow{E} \mathbb{P}_C$	User's message content relevant to a path $\mathbb{P}_C$ on the Wikipedia category tree (part of the Wikipedia layer)
$HT_X \xrightarrow{E} \mathbb{P}_C$	HashTag's message content relevant to a path $\mathbb{P}_C$ on the Wikipedia category tree (part of the Wikipedia layer)
Edges within Wikipedia Layer	
$A_i \xrightarrow{l} A_j$	Wikipedia article incoming/outgoing link
$A \xrightarrow{b} C$	Wikipedia article belongs to Wikipedia category
$C_i \xrightarrow{h} C_j$	Wikipedia category has Wikipedia category
For U_X and HT_X, X could be either $Weibo$ or $Twitter$.	

layer). For example, for a Twitter hashtag *"#iPadAir2"*, based on ESPM algorithm (will be introduced in the next section), the node is linking to two exemplar Wikipedia category paths (each path has a weight):

1. $C_{Technolgy} \xrightarrow{h} ...C_{Mobile\_operating\_systems} \xrightarrow{h} C_{IOS\_(Apple)}$ (weight = 0.45), and 2. $C_{Technolgy} \xrightarrow{h} ...C_{Personal\_computers} \xrightarrow{h} C_{Tablet\_computers}$ (weight = 0.55).

Similarly, the Twitter or Weibo user node, U_X, is also linked to multiple category paths. The weight of the path is the probability that the path can represent the semantics (content) of the target user or hashtag node, $P(\mathbb{P}_C|U_X)$ or $P(\mathbb{P}_C|HT_X)$. The links within each layer are generated from social media and Wikipedia data, and the links across different layers are generated using graph mining algorithms (will be introduced in the next section). So, the quality of the inter-layer links can be lower, and we will need to use the supervised model to optimize different possible paths for cross-media information recommendation.

Wikipedia category hierarchy for different languages may be different. In this study, we use the English category hierarchy to keep the process simple and comparable. When pre-processing, we construct a homogeneous Wikipedia category graph $G_C =< V, E >$, where each vertex, $C_i \in V$, is a Wikipedia category, and the edge links two categories, $C_i \xrightarrow{h} C_j \in E$. Then a tree-like category graph is constructed by Algorithm 1.

By using this method, we construct a tree-like structure as part of the Wikipedia layer on PGSMN. G_C has one root node R, and one category node may have multiple parents. R is the root category defined on Wikipedia category hierarchy, which links to 26 1st level categories, i.e., *Culture, Education, Environment, Politics*, and *Science*. Inside the for loop (line 7 - 11), all the children nodes of the target node (C_i) that have appeared in current V are removed from $Children(C_i)$ set ($\{C_j|(C_j \notin V) \wedge (C_i \in Parents(C_j)\}$). Therefore, all potential loops in Wikipedia categories are eliminated in G_C. The generated G_C is then used in ESPM algorithm and category path-based random walk introduced in the following sections.

Algorithm 1: G_C generation algorithm

```
1  set root vertex $R$ = 'Main topic classifications'
2  $V = \{R\}$
3  $Q = \{R\}$ //temporary vertex queue
4  $V_{temp} = \emptyset$ //temporary vertex set
5  $E_{temp} = \emptyset$ //temporary edge set
6  while not empty $Q$ do
7      for each $C_i \in Q$ do
8          $Children(C_i) = \{C_j|(C_j \notin V) \wedge (C_i \in Parents(C_j)\}$
9          $E_{temp} = E_{temp} \cup \{(C_i \xrightarrow{h} C_j)|C_j \in Children(C_i)\}$
10         $V_{temp} = V_{temp} \cup Children(C_i)$
11     end
12     $V = V \cup V_{temp}; E = E \cup E_{temp}; Q = V_{temp}$
13     $V_{temp} = \emptyset; E_{temp} = \emptyset$
14 end
15 return $G_C =< V, E >$
```

Interconnect Weibo and Twitter Nodes

In this paper, as Table 1 shows, we employ two kinds of links to interconnect Weibo and Twitter users/hashtags with the Wikipedia bridge layer. For either method, the user/hashtag is represented by a message text index, which merges all the messages sent by the target user, T_{U_X}, or includes the target hashtag, T_{HT_X}. The first method is a text exact match, $P(A|U_X)$ or $P(A|HT_X)$, which denotes the probability of a Wikipedia article given a user or hashtag (node) in Twitter or Weibo. For this approach, each Wikipedia concept (article), A, is represented by a phrase collection $PC(A) = A_{p1}, A_{p2}...A_{pk}$, where each A_{pi} is the Wikipedia article or its redirect page name. For instance, for article "Barack_Obama", $PC(A)$ is {*"barack_obma", "barry_obama", "obamma" "o'bama", "barak_obama", "barack_obamaca", "barack_obamaca"*}. The redirected article names can be important for mining the semantics of the target user or hashtag in a colloquial language context. For instance, the abbreviation (of a long article title) or common spelling mistakes can be important to extract semantics of colloquial Twitter or Weibo messages. The transition probability from user node U_X to a Wikipedia article A_i can be calculated by:

$$P(A_i|U_X) = \frac{\sum_{j=1}^{|PC(A_i)|} freq(A_{i_{pj}}, T_{U_X})}{\sum_{t=1}^{|A|} \sum_{k=1}^{|PC(A_t)|} freq(A_{i_{pk}}, T_{U_X})}$$

The probability score equals the sum of the frequency of all the strings in the $PC(A_i)$ in the message text T_{U_X} (exact match), $\sum_{j=1}^{|PC(A_i)|} freq(A_{i_{pj}}, T_{U_X})$, divided by the total frequency count of all the Wikipedia articles, $|A|$. Similarly, we can calculate the hashtag to article transitioning probability $P(A_i|HT_X)$ of the edge $HT_X \xrightarrow{r} A_i$.

Unfortunately, string match cannot solve semantic match problem entirely. For instance, a user may be interested in a specific concept but not necessarily use the target Wikipedia title (or redirected page titles) in the message content. A well-known solution for this problem is Explicit Semantic

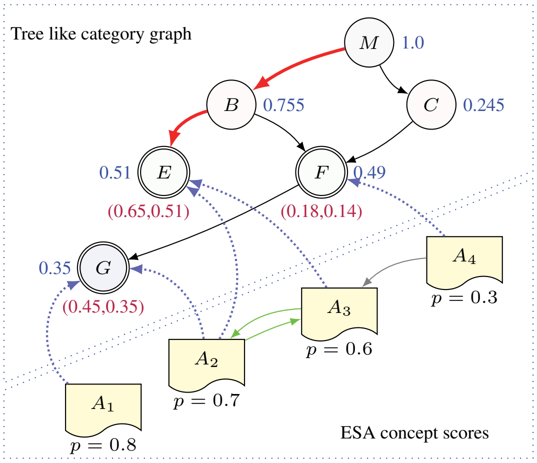

Figure 2: ESPM Example

Analysis (ESA) (Gabrilovich and Markovitch 2007), which calculates the related articles, $P(A_i|t)$, given a text, t, from a latent semantic perspective. In ESA, a word is represented as a column vector in the TFIDF matrix of the Wikipedia, and the input text is represented as the centroid of the vectors representing its words. Experiments (Anderka and Stein 2009; Scholl et al. 2010) have shown that ESA, when combining word and concept features, can enhance text categorization performance on standard corpus over the bag-of-words approach (Gabrilovich and Markovitch 2007). Although a number of studies successfully employed ESA to enhance the text mining performance, the accuracy, or to say the quality, of ESA vector is still problematic (Jiang, Chen, and Liu 2014). For instance, given the text *“Iraq’s Top Shiite Cleric Calls for New Government...”*, the top ranked ESA concepts include *“John Flower”*, *“Hammadi Ahmad”*, and *“Promised Day Brigades”*, can hardly represent the accurate semantics of the given text, even though they may be statistically useful.

In order to cope with this problem, in this study, we propose a new method **Explicit Semantic Path Mining (ESPM)** by leveraging the rich linkage and categorical relationships of Wikipedia. For each given text, i.e., T_{U_X}, instead of generating the semantic article distribution, ESPM identifies optimized semantic path(s), i.e., $U_X \xrightarrow{E} \mathbb{P}_C$, on the Wikipedia category tree. As an example, by using ESPM for the same input above, we get the following semantic path: $Politics \xrightarrow{h} Politics\_by\_country \xrightarrow{h} Politics\_of\_Iraq \xrightarrow{h} Iraqi\_nationalism$(each node is a wiki category), which makes more sense compared with the ESA concept vector.

On the back end of ESPM, Wikipedia provides high-quality, user-oriented hierarchical category definition. The top levels of categories, in most cases, are defined by professional editors. For instance, the first-level includes 26 general categories, such as *Culture*, *Education*, *Environment*, and *Politics*, while Wikipedia page authors and contributors define most of the bottom level categories, such as *American military personnel killed in the War of 1812*, which provides potential to interconnect Weibo and Twitter nodes.

Figure 2 visualizes the ESPM generation progress. Given a text, the related Wikipedia articles (non-zero $P(A_i|t)$ calculated by ESA) vote for the significant semantic path(s) on the Wikipedia category tree graph. Note that one article can belong to multiple Wikipedia categories. In this step, we follow two premises. First, all of the related articles are more likely to be connected via incoming and outgoing page hyperlinks, and the links can be important to help us filter noisy articles. As the following formula shows, the importance of a Wikipedia category given text, $w(C_i|t)$, is calculated by all the articles belonging to C_i, $A_j \in C_i$, and all the linked articles who share the target category, $A_j \xrightarrow{l} A_k \cap A_k \in C_i$. $|A_j \in C_i|$ is the total number of Wikipedia pages in the target category.

$$w(C_i|t)$$
$$= \frac{\sum_{A_j \in C_i} (\lambda \cdot P(A_j|t) + (1-\lambda) \cdot \frac{\Sigma_{A_j \xrightarrow{l} A_k \cap A_k \in C_i} P(A_k|t)}{|A_j \xrightarrow{l} A_k \cap A_k \in C_i|})}{|A_j \in C_i|}$$

In other words, based on this formula, if a number of highly-ranked articles are interconnected in the ESA vector and all belong to a specific category, $\frac{\Sigma_{A_j \xrightarrow{l} A_k \cap A_k \in C_i} P(A_k|t)}{|A_j \xrightarrow{l} A_k \cap A_k \in C_i|}$, this category, C_i, can be important. As Figure 2 shows, because A_2 and A_3 are well-connected, the probability that category E and its related paths are selected is higher than other categories on the tree. λ controls the importance of article content and the importance of the links between articles. In this paper, $\lambda = 0.6$. Once we calculate the seed category node importance for the given text, we need to normalize them.

$$w^{norm}(C_i|t) = \frac{w(C_i|t)}{\Sigma_{C_j \in seeds} w(C_j|t)}$$

If node C does not belong to seed categories, $w^{norm}(c|t) = 0$. Now we can calculate the category probability, $P(C_i|t)$, on the tree-like category graph. First, we normalize the category probabilities to ensure the root node probability will always equal 1.0, which means any text must belong to “something” defined by Wikipedia categories. Second, we transfer every node’s probability to their parents iteratively (bottom up); all the nodes’ probabilities will be transferred to the root node through all possible paths.

$$P(C_i|t) = \Sigma_{C_i \xrightarrow{h} C_{child_k}} \frac{P(C_{child_k}|t)}{|C_j \xrightarrow{h} C_{child_k}, \forall j|} + w^{norm}(C_i|t)$$

Through the bottom-up method, all the possible nodes would be assigned values. Then we use the top-down method to find all possible paths from root node to seed nodes. We define the path weight as the sum of all the category nodes on the path:

$$P(path_k|t) = \frac{\Sigma_{C_i \in path_k} P(C_i|t)}{|C_i \in path_k|}$$

Take Figure 2 for example. Suppose (A_1, A_2, A_3, A_4) with score $(0.8, 0.7, 0.6, 0.3)$ is the ESA analysis result for given text t, and then we have $w(E|t) = 0.65$, $w(F|t) = 0.18$, $w(G|t) = 0.45$. After normalization, we get $w^{norm}(E|t) = 0.51, w^{norm}(F|t) = 0.14, w^{norm}(G|t) = 0.35$ respectively. For leaf node E and G, their final probability is 0.51 and 0.35, for node F, its probability equals child category G’s probability plus its own normalized importance, i.e., $0.35 +$

$0.14 = 0.49$, and $P(E|t) = (0.51 + 0.49 \div 2) = 0.755$. After all propagation, we get the probability 1.0 on root node M. There are three paths from root M to leaf node E and G: $path_1 = [M] \rightarrow [B] \rightarrow [E]$, $path_2 = [M] \rightarrow [B] \rightarrow [F] \rightarrow [G]$and $path_3 = [M] \rightarrow [C] \rightarrow [F] \rightarrow [G]$. Since $P(path_1|t) = (1.0 + 0.755 + 0.51) \div 3 = 0.755$, $P(path_2|t) = 0.649$, $P(path_3|t) = 0.521$, $path_1$ is better than $path_2$ and $path_3$.

As another premise, for ESPM, we need to find a number of independent paths on the Wikipedia category tree. For instance, if we find $C_1 \overset{h}{\rightarrow} C_2 \overset{h}{\rightarrow} C_3$ we don't want to find another path $C_1 \overset{h}{\rightarrow} C_2 \overset{h}{\rightarrow} C_3 \overset{h}{\rightarrow} C_4$, as these two paths are highly dependent and provide very similar information. To characterize this assumption, we use greedy algorithm to identify the top k independent important paths on the tree.

First, we calculate all the relevant paths' weight with the aforementioned method. Then we generate a graph where each path is conceptualized as a node on the graph (with the node weight = $P(path_k|t)$), and if any two paths are dependent, there will be an edge connecting these two nodes. For path dependence measure (Chakrabarti and Mehta 2010), we utilize the similarity between two paths. For example, if $Sim(p_i, p_j) > \delta$ ($\delta = 0.7$), path p_i and p_j are dependent. On this graph, we first pick the node with the largest weight then remove all of its connected nodes. We will repeat this process until all the nodes on the graph are removed and picked. Note that after this step, we will get a list of ranked paths, which are independent to others. This greedy algorithm has proven useful in prior feature selection studies (Chakrabarti and Mehta 2010).

Random Walk between Twitter and Weibo

In this section, we design the random walk functions to enable cross-media information recommendation. From a random walk viewpoint, on the heterogeneous PGSMN, there are different alternatives (functions) to random walk from one node (in Twitter) to another (in Weibo), and vice versa, i.e., $RW(N_X \rightsquigarrow N_Y)$, which denotes the random walk probability from a node in X to a node in Y (X and Y could be Weibo or Twitter). For example, if we want to recommend Twitter hashtags to Weibo users (random walk from target Weibo user to Twitter hashtags), the meta-path based random walk functions can be defined in Table 2.

Each random walk function in this exemplar table is a meta-path, $\mathcal{P}$, on the heterogeneous graph. The ranking score of a candidate node, N_Y, given a meta-path, $\mathcal{P}$, is the cumulated random walk probabilities (tours $\in$ meta-path) starting from N_X to N_Y, $RW_{\mathcal{P}}(N_X \rightsquigarrow N_Y) = \sum_{t_{\mathcal{P}} \in \mathcal{P}} RW(t_{\mathcal{P}})$. The $\mathcal{P}$ is defined by $N_X \xrightarrow{R_1} \dot{A}_1 \xrightarrow{R_2} \dot{A}_2 \ldots \xrightarrow{R_l} N_Y$(length = l), and $t_{\mathcal{P}}$ is a tour from N_X to N_Y following the specified meta-path $\mathcal{P}$. $RW(t_{\mathcal{P}})$ is the random walk probability of the tour $t_{\mathcal{P}}$. This meta-path based scoring function has been defined in (Liu et al. 2014).

However, previous meta-path based random walk methods cannot solve the ESPM related problems, like $N_X \overset{E}{\rightarrow} \mathbb{P}_C \rightsquigarrow \mathbb{P}_C \overset{E}{\leftarrow} N_Y$. We need to define the random walk probability from one path collection (m paths related to node N_X based on $P(\mathbb{P}_C|T_{N_X})$), $\{\mathbb{P}_{C_X}\}$, to another (n paths related to node N_Y), $\{\mathbb{P}_{C_Y}\}$:

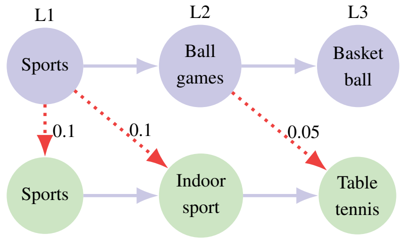

Figure 3: Example of Random Walk between Paths.

$$RW(\{\mathbb{P}_{C_X}\} \rightsquigarrow \{\mathbb{P}_{C_Y}\}) = \frac{\sum_{i=1}^{m} \underset{1 \leqslant j \leqslant n}{max} (RW(\mathbb{P}_i \rightsquigarrow \mathbb{P}'_j))}{m}$$

where, the random walk probability equals the average random walk probability from one path $\mathbb{P}_i$ in $\{\mathbb{P}_{C_X}\}$ to its closest path $\mathbb{P}'_j$ in $\{\mathbb{P}_{C_Y}\}$. Then, we need to define the random walk function between two paths on the Wikipedia tree. If $\mathbb{P}_i = C_{\mathbb{P}_i 1} \overset{h}{\rightarrow} .C_{\mathbb{P}_i 2} \overset{h}{\rightarrow} ....C_{\mathbb{P}_i k}$ and $\mathbb{P}'_j = C_{\mathbb{P}'_j 1} \overset{h}{\rightarrow} .C_{\mathbb{P}'_j 2} \overset{h}{\rightarrow} ....C_{\mathbb{P}'_j l}$, then $RW(\mathbb{P}_i \rightsquigarrow \mathbb{P}'_j) = RW(\mathbb{P}_{i(1))} \rightsquigarrow \mathbb{P}'_{j(1))})$, and $RW(\mathbb{P}_{i(t))} \rightsquigarrow \mathbb{P}'_{j(t))}) = \alpha \cdot RW(C_{i(t)} \rightsquigarrow \mathbb{P}'_{j(t)}) + (1 - \alpha) \cdot P(C_{i(t+1)}|C_{i(t)}) \cdot RW(\mathbb{P}_{i(t+1))} \rightsquigarrow \mathbb{P}'_{j(t+1))})$, where $RW(C_{i(t)} \rightsquigarrow \mathbb{P}'_{i(t)})$ is the random walk probability from t^{th} category node $C_{i(t)}$ (on path $\mathbb{P}_i$) to sub path $\mathbb{P}'_{j(t)}$ (on path $\mathbb{P}'_j$) from $(t)^{th}$ position. $P(C_{i(t+1)}|C_{i(t)})$ is the transition probability from node $C_{i(t)}$ to $C_{i(t+1)}$ on path $\mathbb{P}_i$. $RW(C_{i(t)} \rightsquigarrow \mathbb{P}'_{j(t)}) = \sum_{C_{jx} \in \mathbb{P}'_{j(t)}} RW(C_{i(t)} \rightsquigarrow C_{jx})$.

This is a recursive random walk definition. For each node on path $\mathbb{P}_i$, there is α chance to walk from the current node on $\mathbb{P}_i$ to $\mathbb{P}'_j$, $RW(C_{i(t)} \rightsquigarrow \mathbb{P}'_{j(t+1)})$, and there is $(1 - \alpha)$ chance to walk to the next node on $\mathbb{P}_i$, and it may potentially walk to $\mathbb{P}'_j$ at the next step, $P(C_{i(t+1)}|C_{i(t)}) \cdot RW(\mathbb{P}_{i(t+1))} \rightsquigarrow \mathbb{P}'_{j(t+1))})$. Based on this definition, *level 1 category node* on the Wikipedia category tree is most important, which have α chance to walk to another path. For other nodes, like t^{th} node ($t > 1$), the importance is $\alpha \cdot (1 - \alpha)^{(t-1)}$. This definition is based on the truth that the higher (general) level categories inference can be more accurate than lower (detailed) level ones. For example, it's easier to estimate a hashtag is about *"Technology"* (level 1) than to label it as *"Tablet_computers"* (level 5).

Figure 3 visualizes an example. Suppose we have two paths: If $\mathbb{P}_i = Sports \rightarrow Ball\_games \rightarrow Basketball$ and $\mathbb{P}'_j = Sports \rightarrow Indoorsports \rightarrow Table\_tennis$. $RW(\mathbb{P}_i \rightsquigarrow \mathbb{P}'_j) = \alpha \cdot (0.1 + 0.1) + (1 - \alpha) \cdot 0.1 \cdot RW(\mathbb{P}_{i(2)} \rightsquigarrow \mathbb{P}'_{j(2)})$, and $RW(\mathbb{P}_{i(2)} \rightsquigarrow \mathbb{P}'_{j(2)}) = \alpha \cdot 0.05 + (1 - \alpha) \cdot 0.05 \cdot RW(\mathbb{P}_{i(3)} \rightsquigarrow \mathbb{P}'_{j(3)})$. For $RW(\mathbb{P}_{i(3)} \rightsquigarrow \mathbb{P}'_{j(3)}) = 0$ (because there is no link from node $basketball$ to path $\mathbb{P}'_j$, and node $basketball$ is the last node on path $\mathbb{P}_i$). When $\alpha = 0.6$ (we use this value for this study), then random walk probability is $RW(\mathbb{P}_i \rightsquigarrow \mathbb{P}'_j) = 0.12012$. In another case, if all the nodes $\in \mathbb{P}_i$ are not linked to $\mathbb{P}'_j$, $RW(\mathbb{P}_i \rightsquigarrow \mathbb{P}'_j) = 0$.

Case Study: Recommend Twitter Hashtag to Weibo

In this section, we use a case study to verify and evaluate the usefulness of PGSMN: **recommend Twitter Topic (hashtag) to Weibo Users** (easier for evaluation). Based on the random walk functions defined in Table 2, the recommendation problem can be conceptualized as a ranking problem. Given a Weibo user, U^{*}_{Weibo}, we rank all the Twitter hashtags, $HT^{?}_{Twitter}$, based on the random walk probability on PGSMN, $RW(U^{*}_{Weibo} \rightsquigarrow HT^{?}_{Twitter})$. Note that, PGSMN supports other recommendation tasks, such as recommending Weibo user to Twitter user (as friend), but auto-evaluation can be hard.

As the last section shows, there are a number of methods we can use to calculate this random walk probability. So, in this study, we use the learning to rank method to combine different ranking features while avoiding manual parameter tuning. As this study is not focusing on learning to rank, we used a relative simple algorithm, Coordinate Ascent (Metzler and Croft 2007), which iteratively optimizes a multivariate objective ranking function, for meta-path PRF feature integration and algorithm evaluation.

In order to generate the training data, we need to find a number of Twitter hashtags that are of interest to some Weibo users. Two methods are used to generate the ground truth data: (1) as the most straightforward approach, we find a number of Weibo hashtags sharing the same strings as the Twitter hashtags in the experiment dataset. However, as most Weibo users are Chinese, we can only find a small number of hashtags for this approach. So (2) we use Google translation API to translate all the Weibo hashtags into English, and we match the translated Weibo hashtags with Twitter hashtags. Note that, only a small proportion Twitter hashtags can be successfully translated into Chinese. The selected hashtags will be removed from the PGSMN, which guarantees the algorithms can find "new Twitter information" for Weibo user.

4 Experiment

[Data]: For this experiment, we employ two datasets sampled from Twitter and Weibo corpora. Both corpora were sampled between 2012/09/17 and 2012/09/23 (1 week data). Based on 3,296,945 Weibo messages, 20,128,826 Twitter messages, and the Wikipedia March 2014 Dumps (Chinese and English page dumps and Chinese language links dump), we generate the PGSMN following the method introduced in the method section. Based on our previous description, we build the 3-layer heterogeneous graph. The number of each type of node or edge is listed in Table 3.

All the hashtags used for fewer than 10 messages and all the users who composed fewer than five messages are removed from the datasets. The PGSMN data can be downloaded from the project website[3], and other researchers can use this data to reproduce this experiment or test and verify other usage of PGSMN.

The pre-processing step contains two parts: (1) Construct Chinese and English datasets and generate cross-language ESA models (prior for ESPM calculation); (2) Construct Wikipedia tree-like category graph and ESPM model.

In order to build Chinese and English ESA models and maintain the concept mapping relationships between these two language bubbles, we traverse the Chinese Wikipedia page dump first. Any Wikipedia article written in both English and Chinese and has at least three incoming links will be incorporated into the experiment PGSMN. Following this process, we get 400,275 Chinese pages in total, corresponding to a total of 397,689 English pages. Because different Chinese pages may feature the same related English page, these two values are not the same, so we merge this type of Chinese pages as one page, and finally form the dataset which consisted of 397,689 articles featuring both Chinese and English content. When building ESA models, all Chinese text is tokenized by word segmentation. Based on the algorithm introduced in method section, we construct a Wikipedia category tree with 871,978 nodes and 1,229,833 edges (Wiki-layer). We then use this tree to calculate the ESPM based random walk probability between Weibo user and Twitter hashtags.

[Result]: In this experiment, we locate 459 hashtags used by both Twitter and Weibo users (based on the methods in Section 3). For the string match method, we find 129 hashtags, and, by using Google translation API, we find 330 matched hashtags from 2,931 Weibo hashtags (because most Weibo hashtags and messages are written in colloquial Chinese). We also find 401 Weibo users who used at least three hashtags from this test collection. However, we noticed that the majority of these 401 users used fewer than 6 out of the 459 selected hashtags. The resulting training dataset has 401 users, 459 hashtags and 20,248 user-hashtag training/testing pairs. To validate the graph mining performance, all the 459 testing hashtags (in Weibo part) were removed from the PGSMN, which guarantees the algorithms can find "new Twitter information" for Weibo user.

For baseline algorithm, we used **Machine Translation (MT)**. The ranking score for a Twitter hashtag, $HT_{twitter_i}$, given a Weibo user, U_{weibo_j}, can be estimated by: $\sum_{m_k \in weibo_j} S_{BM25}(\{m_{HT_{twitter_i}}\}, MT(m_k))$, where S_{BM25} is the BM25 similarity function, and $MT(m_k)$ is the translated Chinese message (via Google Machine Translation API) from the target user. For this method, we use all of the Weibo user's translated messages as queries to rank all the hashtags from Twitter corpus. Each Twitter hashtag is represented by all the Twitter messages, including the target hashtag, $\{m_{HT_{twitter_i}}\}$.

As Section 3 mentioned, we use precision (**P**), mean average precision (**MAP**), and normalized discounted cumulative gain (**NDCG**) to evaluate the recommendation ranking performance. For each Weibo user in the test collection, we recommend the Twitter hashtags via meta-path based ranking function introduced in Table 2. To make the result clear, we also compare the mean reciprocal rank (**MRR**), which is the average of the multiplicative inverse of the rank of the first correct Twitter hashtag for the Weibo users.

The recommendation performance, by using different kinds of feature combinations, is reported in Table 4. We utilize Coordinate Ascent (Metzler and Croft 2007) as learning

[3] http://scholarwiki.indiana.edu/data

Table 2: Meta-path in the constructed heterogeneous graph

ID	Meta-path & Ranking Hypothesis	ID	Meta-path & Ranking Hypothesis
F1	$U^{*}_{Weibo} \xrightarrow{r} A \xleftarrow{r} HT^{?}_{Twitter}$	**F2**	$U^{*}_{Weibo} \xrightarrow{t} U_{Weibo} \xrightarrow{r} A \xleftarrow{r} U_{Twitter} \xrightarrow{u} HT^{?}_{Twitter}$
	The candidate Twitter HashTag is important, if it is relevant to the same Wikipedia page as the query Weibo User.		The candidate Twitter HashTag is important, if it is mentioned by the Twitter User who is relevant to the same Wikipedia page as the query Weibo User's related users.
F3	$U^{*}_{Weibo} \xrightarrow{u} HT_{Weibo} \xrightarrow{r} A \xrightarrow{l} A \xleftarrow{r} HT_{Twitter} \xrightarrow{c} HT^{?}_{Twitter}$	**F4**	$U^{*}_{Weibo} \xrightarrow{u} HT_{Weibo} \xrightarrow{r} A \xrightarrow{l} A \xleftarrow{r} U_{Twitter} \xrightarrow{u} HT^{?}_{Twitter}$
	The candidate Twitter HashTag is important, if its co-occur HashTag is relevant to some Wikipedia page which links to the one relevant to the Weibo HashTag which is mentioned by the query Weibo User.		The candidate Twitter HashTag is important, if it is mentioned by the Twitter User relevant to some Wikipedia page which links to the one relevant to the Weibo HashTag mentioned by the query Weibo User.
F5	$U^{*}_{Weibo} \xrightarrow{t} U_{Weibo} \xrightarrow{r} A \xrightarrow{l} A \xleftarrow{r} U_{Twitter} \xrightarrow{u} HT^{?}_{Twitter}$	**F6**	$U^{*}_{Weibo} \xrightarrow{t} U_{Weibo} \xrightarrow{r} A \xrightarrow{l} A \xleftarrow{r} HT_{Twitter} \xrightarrow{c} HT^{?}_{Twitter}$
	The candidate Twitter HashTag is important, if it is mentioned by the Twitter User who is relevant to some Wikipedia page which links to the one relevant to the query Weibo User's related users.		The candidate Twitter HashTag is important, if its co-occur HashTag is relevant to some Wikipedia page which links to the one relevant to the query Weibo User's related users.
F7	$U^{*}_{Weibo} \xrightarrow{u} HT_{Weibo} \xrightarrow{r} A \xrightarrow{b} C \xleftarrow{b} A \xleftarrow{r} HT^{?}_{Twitter}$	**F8**	$U^{*}_{Weibo} \xrightarrow{r} A \xrightarrow{b} C \xrightarrow{h} C \xleftarrow{b} A \xleftarrow{r} HT^{?}_{Twitter}$
	The candidate Twitter HashTag is important, if its relevant Wikipedia page's category is the same as the page which is relevant to the query Weibo User's mentioned HashTag.		The candidate Twitter HashTag is important, if its relevant Wikipedia page's category has category which has page that is relevant to the query Weibo User.
F9	$U^{*}_{Weibo} \xrightarrow{u} HT_{Weibo} \xrightarrow{r} A \xrightarrow{b} C \xleftarrow{b} A \xleftarrow{r} HT_{Twitter} \xrightarrow{c} HT^{?}_{Twitter}$	**F10**	$U^{*}_{Weibo} \xrightarrow{E} \mathbb{P}_C \rightsquigarrow \mathbb{P}_C \xleftarrow{E} HT^{?}_{Twitter}$
	The candidate Twitter HashTag is important, if its co-occur HashTag is relevant to some Wikipedia page which belongs to the same category as the one relevant to the Weibo HashTag which is mentioned by the query Weibo User.		The candidate Twitter HashTag is important, if its relevant category path is related to the query Weibo User's relevant category path.
F11	$U^{*}_{Weibo} \xrightarrow{u} HT_{Weibo} \xrightarrow{E} \mathbb{P}_C \rightsquigarrow \mathbb{P}_C \xleftarrow{E} HT_{Twitter} \xrightarrow{c} HT^{?}_{Twitter}$	**F12**	$U^{*}_{Weibo} \xrightarrow{t} U_{Weibo} \xrightarrow{E} \mathbb{P}_C \rightsquigarrow \mathbb{P}_C \xleftarrow{E} U_{Twitter} \xrightarrow{u} HT^{?}_{Twitter}$
	The candidate Twitter HashTag is important, if its co-occur HashTag's relevant category path is related to the query Weibo User's mentioned hashtag's relevant category path		The candidate Twitter HashTag is important, if it is mentioned by some Twitter User whose relevant category path is related to the query Weibo User's related users' relevant category path
F13	$U^{*}_{Weibo} \xrightarrow{t} U_{Weibo} \xrightarrow{u} HT_{Weibo} \xrightarrow{E} \mathbb{P}_C \rightsquigarrow \mathbb{P}_C \xleftarrow{E} HT^{?}_{Twitter}$		
	The candidate Twitter HashTag is important, if its relevant category path is related to the query Weibo User's related users' mentioned hashtag's relevant category path		

Table 3: Statistics on the 3-layer heterogeneous graph

Graph 1: Weibo Layer (generated from 3,296,945 messages)	
Node/Edge	Number
U_{Weibo}	85,572
HT_{Weibo}	2,931
$U_{Weibo} \xrightarrow{t} U_{Weibo}$	296,438
$U_{Weibo} \xrightarrow{u} HT_{Weibo}$	41,896
$HT_{Weibo} \xrightarrow{c} HT_{Weibo}$	3,456
Graph 2: Wikipedia Layer	
Node/Edge	Number
A	397,689
C	871,978
$A \xrightarrow{l} A$	6,044,535
$A \xrightarrow{b} C$	1,044,002
$C \xrightarrow{h} C$	1,229,833
Graph 3: Twitter Layer (generated from 20,128,826 messages)	
Node/Edge	Number
$U_{Twitter}$	797,869
$HT_{Twitter}$	6,995
$U_{Twitter} \xrightarrow{t} U_{Twitter}$	1,396,296
$U_{Twitter} \xrightarrow{u} HT_{Twitter}$	384,474
$HT_{Twitter} \xrightarrow{c} HT_{Twitter}$	172,089

to rank method for feature integration and algorithm evaluation. NDCG30 is used as the training metric, and 5-fold cross validation is used to evaluate the recommendation performance. All of the ranking features (in Table 2) are classified into three categories based on the random walk functions within the Wikipedia layer: **A**: Meta-paths based on Wikipedia articles (A) and the hyperlinks, $A_i \xrightarrow{l} A_j$, among them. We use features **[F1-F6]** in Table 2 for this group. **A-C**: Meta-paths based on Wikipedia articles (A), categories (C), the links between article and category, $A \xrightarrow{b} C$, and the links among categories, $C_i \xrightarrow{h} C_j$. Features **[F7-F9]** are used for this group. **ESPM**: Meta-paths based on Wikipedia category path and ESPM based random walk features, $\mathbb{P}_{C_i} \rightsquigarrow \mathbb{P}_{C_j}$. Feature **[F10-F13]** are used for this group with ESPM algorithm. **ALL** combines A, A-C, and ESPM.

Experiment results support our hypothesis that machine translation (**MT**) is not a good option for cross-Twitter and Weibo recommendation. As most messages in Weibo are written in colloquial Chinese, machine translation and content matching performance is not ideal. On the other hand, all the PGSMN recommendation methods can significantly ($p < 0.01$) enhance the recommendation performance. Compared with different categories of PGSMN features, results show that, for most metrics, Wikipedia article and category **A-C** and **A + A-C** outperforms other kinds of features for overall ranking (NDCG and MAP). **ESPM** based random walk (between Wikipedia category paths) performs well for top ranked result. Evaluation result also shows that when we use all the features **A + A-C + ESPM**, recommendation performance reaches its peak ($p < 0.01$), which means different kinds of random walk methods between

Table 4: Experiment Result

Metrics	**MT**	**A**	**A-C**	**A + A-C**	**ESPM**	**ALL**
MRR	0.0146	0.1816	0.4780	0.4653	0.3428	**0.4887***
P5	0.0187	0.0777	0.1222	0.1138	**0.1668**	0.1549
P10	0.0174	0.0722	0.0785	0.0769	**0.1164**	0.1150
P15	0.0160	0.0648	0.0707	0.0665	0.0962	**0.0982**
P20	0.0155	0.0595	0.0682	0.0675	0.0841	**0.0897***
MAP10	0.0031	0.0054	0.0117	0.0110	0.0125	**0.0144***
MAP30	0.0038	0.0079	0.0146	0.0138	0.0152	**0.0178***
MAP50	0.0040	0.0095	0.0168	0.0160	0.0167	**0.0198***
MAP100	0.0043	0.0127	0.0197	0.0194	0.0197	**0.0231***
MAP	0.0056	0.0255	0.0308	0.0333	0.0299	**0.0370***
NDCG10	0.0123	0.0419	0.0781	0.0748	0.0780	**0.0961***
NDCG30	0.0137	0.0428	0.0662	0.0646	0.0605	**0.0763***
NDCG50	0.0151	0.0497	0.0732	0.0734	0.0645	**0.0820***
NDCG100	0.0191	0.0684	0.0866	0.0904	0.0810	**0.0985***
NDCG	0.0518	0.1664	0.1657	0.1923	0.1557	**0.2017***
* Significant p <0.01						

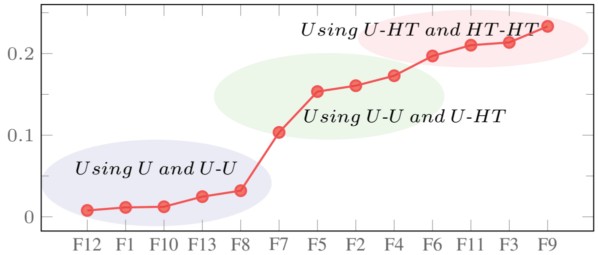

Figure 4: NDCG Comparison for Ranking Features

different layers can all contribute to the recommendation model. Overall, for most evaluation metrics, **ESPM** > **A-C** > **A**, which proves our hypothesis that Wikipedia category and category tree are helpful to enhance the recommendation performance.

When we use ALL features, MRR reaches 0.4887 while P5 is 0.1549, which means that the top-ranked Twitter hashtags are highly likely to interest targeted Weibo users. MAP score is relatively low because we can only find 459 hashtags for evaluation in the test collection, and these hashtags can only cover partial interests of the target Weibo user.

To test the feature performance, we also compare each feature individually from an NDCG perspective (see Figure 4). The ranking features are ordered in an ascending order. Based on the result, we find recommendation performance is also closely related to the sub-paths within the Weibo and Twitter layers. It is clear that user and hashtag relations $U_X \xrightarrow{u} HT_X$, and hashtag co-occur relations $HT_X \xrightarrow{c} HT_X$, are more important than other kinds of relations. For instance, we find user-to-user reply relation, $U_X \xrightarrow{t} U_X$, is less useful. If we don't use any relationship within Weibo and Twitter layer (i.e., F1, $U^*_{Weibo} \xrightarrow{r} A \xleftarrow{r} HT^?_{Twitter}$), the ranking performance is not strong.

This result proves our initial assumptions that the rich linkage within and across different layers and ESPM-based random walk on the Wikipedia category tree can be important for cross-social media information recommendation.

5 Analysis and Conclusion

In this study, we propose a new method to interconnect Twitter and Weibo datasets while enabling cross-media information recommendation. Unlike earlier studies, Twitter and Weibo belong in exclusive bubbles, i.e., language bubbles, culture bubbles, and network bubbles. As we cannot identify the same user across these two social networks, we employed a novel method to construct a heterogeneous graph, PGSMN, to interconnect Weibo and Twitter users and hashtags. Wikipedia categories and articles are employed as the "Rosetta Stone" to bridge Twitter and Weibo networks. Our research will help the underrepresented groups fully participate in the global cultural and political conversation that is now increasingly taking place online and through social media. Our results may mitigate the digital divide that results from social, and linguistic disparities.

Our evaluation result shows that PGSMN can efficiently recommend Twitter hashtags to Weibo users, which significantly ($p < 0.0001$) outperforms machine translation. Meanwhile, all the proposed meta-path based ranking features are potentially useful. We also find that ESPM and Wikipedia category based random walk features can significantly enhance the recommendation performance ($p < 0.0001$). Simply using Wikipedia articles (and links) may not be accurate enough for cross Twitter and Weibo recommendation. We assume this may be because of the impact of the culture difference. For instance, if a Weibo user is interested in (or linked to) a local basketball star, we can hardly find the corresponding or related hashtags in Twitter if we simply use Wikipedia articles. However, if Wikipedia categories are used, we can find a path related to the 'basketball category' or ESPM based random walk between similar Wikipedia category paths like $C_{Sport} \xrightarrow{h} C_{Basketball} \xrightarrow{h} C_{Chinese\_Basketball}$ and $C_{Sport} \xrightarrow{h} C_{Basketball} \xrightarrow{h} C_{NBA}$. All of these Wikipedia category related meta-paths provide good potential to random walk cross different media.

In the future, we would like to test more sophisticated random walk methods and learning to rank methods to further enhance the recommendation performance. Meanwhile, we will test other applications of PGSMN. For instance, by using this graph, we could calculate the relatedness between a Weibo user and a Twitter user. Then, we could develop innovative community detection algorithm on PGSMN to group "similar" Weibo and Twitter users into the same cross-domain community.

References

Achrekar, H.; Gandhe, A.; Lazarus, R.; Yu, S.-H.; and Liu, B. 2011. Predicting flu trends using twitter data. In *IEEE Conference on Computer Communications Workshops*, 702–707. IEEE.

Anderka, M., and Stein, B. 2009. The esa retrieval model revisited. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 670–671. ACM.

Baucom, E.; Sanjari, A.; Liu, X.; and Chen, M. 2013. Mirroring the real world in social media: twitter, geolocation, and sentiment analysis. In *Proceedings of the 2013 International Workshop on Mining Unstructured Big Data Using Natural Language Processing*, 61–68. ACM.

Chakrabarti, D., and Mehta, R. 2010. The paths more taken: matching dom trees to search logs for accurate webpage clustering. In *Proceedings of the 19th international conference on World wide web*, 211–220. ACM.

Ediger, D.; Jiang, K.; Riedy, J.; Bader, D. A.; Corley, C.; Farber, R.; and Reynolds, W. N. 2010. Massive social network analysis: Mining twitter for social good. In *39th International Conference on Parallel Processing (ICPP)*, 583–593. IEEE.

Fernández-Tobías, I.; Cantador, I.; Kaminskas, M.; and Ricci, F. 2011. A generic semantic-based framework for cross-domain recommendation. In *Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, 25–32. ACM.

Gabrilovich, E., and Markovitch, S. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, 1606–1611.

Guan, W.; Gao, H.; Yang, M.; Li, Y.; Ma, H.; Qian, W.; Cao, Z.; and Yang, X. 2014. Analyzing user behavior of the microblogging website sina weibo during hot social events. *Physica A: Statistical Mechanics and its Applications* 395:340–351.

Jiang, Z.; Chen, M.; and Liu, X. 2014. Semantic annotation with rescoredesa: Rescoring concept features generated from explicit semantic analysis. In *Proceedings of the 7th International Workshop on Exploiting Semantic Annotations in Information Retrieval*, 25–27. ACM.

Kouloumpis, E.; Wilson, T.; and Moore, J. 2011. Twitter sentiment analysis: The good the bad and the omg! *ICWSM* 11:538–541.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, 591–600. ACM.

Li, C.-Y., and Lin, S.-D. 2014. Matching users and items across domains to improve the recommendation quality. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 801–810. ACM.

Li, D.; Zhang, J.; Sun, G. G.-z.; Tang, J.; Ding, Y.; and Luo, Z. 2012. What is the nature of chinese microblogging: Unveiling the unique features of tencent weibo. *arXiv preprint arXiv:1211.2197*.

Liu, X.; Yu, Y.; Guo, C.; Sun, Y.; and Gao, L. 2014. Full-text based context-rich heterogeneous network mining approach for citation recommendation. In *23rd ACM International Conference on Information and Knowledge Management*.

Metzler, D., and Croft, W. B. 2007. Linear feature-based models for information retrieval. *Information Retrieval* 10(3):257–274.

Pak, A., and Paroubek, P. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *The International Conference on Language Resources and Evaluation*.

Sahebi, S., and Brusilovsky, P. 2013. Cross-domain collaborative recommendation in a cold-start context: The impact of user profile size on the quality of recommendation. In *User Modeling, Adaptation, and Personalization*. Springer. 289–295.

Scholl, P.; Böhnstedt, D.; García, R. D.; Rensing, C.; and Steinmetz, R. 2010. Extended explicit semantic analysis for calculating semantic relatedness of web resources. In *Sustaining TEL: From Innovation to Learning and Practice*. Springer. 324–339.

Shi, Y.; Larson, M.; and Hanjalic, A. 2011. Tags as bridges between domains: Improving recommendation with tag-induced cross-domain collaborative filtering. In *User Modeling, Adaption and Personalization*. Springer. 305–316.

Shuai, X.; Liu, X.; Xia, T.; Wu, Y.; and Guo, C. 2014. Comparing the pulses of categorical hot events in twitter and weibo. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14, 126–135. New York, NY, USA: ACM.

Sun, Y.; Barber, R.; Gupta, M.; Aggarwal, C.; and Han, J. 2011a. Co-author relationship prediction in heterogeneous bibliographic networks. In *Proc. 2011 Int. Conf. Advances in Social Network Analysis and Mining (ASONAM'11)*.

Sun, Y.; Han, J.; Yan, X.; Yu, P. S.; and Wu, T. 2011b. PathSim: Meta path-based top-k similarity search in heterogeneous information networks. In *Proc. 2011 Int. Conf. Very Large Data Bases*.

Sun, Y.; Norick, B.; Han, J.; Yan, X.; Yu, P. S.; and Yu, X. 2012. Integrating meta-path selection with user guided object clustering in heterogeneous information networks. In *Proc. of 2012 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*.

Vosecky, J.; Hong, D.; and Shen, V. Y. 2009. User identification across multiple social networks. In *1st International Conference on Networked Digital Technologies (NDT'09)*, 360–365. IEEE.

Wikipedia. 2014. List of websites blocked in china. [Online; accessed 2014-02-07].

Zafarani, R., and Liu, H. 2013. Connecting users across social media sites: a behavioral-modeling approach. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 41–49. ACM.

Zafarani, R., and Liu, H. 2014. Finding friends on a new site using minimum information. In *International Conference on Data Mining*. SIAM.

Zhang, H.; Kan, M.-Y.; Liu, Y.; and Ma, S. 2014. Online social network profile linkage. In *The Tenth Asia Information Retrieval Societies Conference*.

Zhao, B.; Zhang, Z.; Gu, Y.; Gong, X.; Qian, W.; and Zhou, A. 2011. Discovering collective viewpoints on micro-blogging events based on community and temporal aspects. In *Advanced Data Mining and Applications*. Springer. 270–284.