交叉社交媒体的推荐

论文标题：Cross Social Media Recommendation

作者：Xiaozhong Liu，Tian Xia，Yingying Yu，Chun Guo，Yizhou Sun

会议：2016 The International AAAI Conference on Web and Social Media (ICWSM)

研究背景

社交媒体改变了人们认知和理解世界的方式，社交媒体挖掘的方法目前已经比较完备。尽管每天都有数以万计的人在社交平台上相互交流，不幸的是，没有一个社交媒体能够连接来自全球的用户。由于 Twitter 和 Facebook 无法在中国使用，21.97%的用户被排除在这两种社交媒体之外。类似地，至 2015 年，新浪微博在中国拥有多达 1 7600 0000 的用户，然而，由于新浪微博的默认语言是中文，它的用户群体几乎都是中国人或懂中文的人。不同社交媒体针对不同的人群，不同的人群感兴趣的话题也不同。因此，没有任何一个社交媒体可以很好地反映全球用户的信息。

本文的重点在于如何构建“虚拟全球社交媒体网络”，从而把两个不同社交媒体（island）中的用户和话题连接到一起。就我们所知，这是第一个把不同语言、文化、和社交网络（bubles）中的用户和话题连接到一起的虚拟社交网络。PGSMN 的作用体现在以下三个方面：

1. 广告商和法律执行代理人可以在 PGSMN 网络中，发现对某些话题关注的用户群，从而发掘一些可以利用的潜在关系；
2. 通过 PGSMN 网络，学者们可以调查中国社会和美国社会对待相同的问题会有哪些异同；
3. PGSMN 网络使交叉社会媒体中的推荐得以实现。

PGSMN 网络的构建主要有两方面的阻碍：

1. Twitter 和微博的用户和话题分别隶属于两个独立的社交网络，由于只有极少数用户同时拥有 Twitter 账号和微博账号，两个网络之间几乎没有任何已存在的联系；
2. Twitter 和微博分别使用两种不同的语言。不仅如此，由于字数限制，用户经常 使用俗语和缩写来表达复杂的语法，这对传统的机器翻译的方法是一个很大的挑战。

PGSMN 网络的构建

本文构建的 PGSMN 网络是一个三层异构网络，包括 Twitter 层、微博层，以及维基百科过渡层，如下图所示：

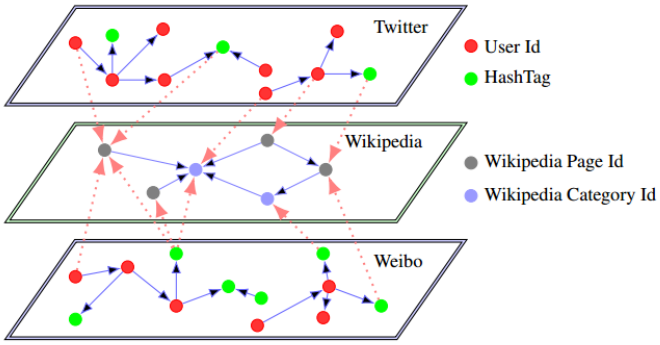


Figure 1: 3-layer Pseudo Global Social Media Network

之所以选择维基百科作为中间过渡层，原因有三：首先，维基百科为所有的词语提供了英文和中文的定义；其次，维基百科中的所有词语都通过类别归属或者相关链接连接在一起；最后，维基百科提供了俗语和普通词汇的转换机制。

如下图所示，PGSMN 网络中节点的关系有以下三类：

1. Twitter 层和微博层各自内部节点间的关系，包括用户之间的关系、用户和话题标签之间的关系，以及话题标签之间的关系；

2. Twitter 层和微博层与维基百科层之间的关系，包括用户与维基百科中文章的关系、话题标签与维基百科中文章的关系、用户与维基百科中路径的关系，以及话题标签与维基百科中路径的关系；

3. 维基百科层内部节点间的关系，包括文章之间的关系、文章与分类的关系，以及分类之间的关系。

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 $Twitter$.	

在构建 PGSMN 网络前，需要先构建一个同构树状的维基百科的类别层次结构图（Category tree），构建该图的算法如下：

Algorithm 1: G_C generation algorithm

set root vertex R = 'Main topic classifications'
$V = \{R\}$
$Q = \{R\}$ //temporary vertex queue
$V_{temp} = \emptyset$ //temporary vertex set
$E_{temp} = \emptyset$ //temporary edge set
while *not empty* Q **do**
  for *each* $C_i \in Q$ **do**
    $Children(C_i) = \{C_j | (C_j \notin V) \wedge (C_i \in Parents(C_j)\}$
    $E_{temp} = E_{temp} \cup \{(C_i \xrightarrow{h} C_j) | C_j \in Children(C_i)\}$
    $V_{temp} = V_{temp} \cup Children(C_i)$
  end
  $V = V \cup V_{temp}; E = E \cup E_{temp}; Q = V_{temp}$
  $V_{temp} = \emptyset; E_{temp} = \emptyset$
end
return $G_C =< V, E >$

网络各层之间链接的构建

为了将微博层与 Twitter 层中的节点与维基百科层联系在一起，本文先定义了用户（User）与话题标签（Hashtag）与维基百科中的词条节点（Article）之间相关性的计算方法（ESA）：

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

为了使实验结果更精确，本文还定义了用户和话题标签与维基百科中某条分类路径（Path）相关性的计算方法（ESPA）。

在 ESPA 方法中，首先定义了用户与话题标签与维基百科中的分类节点（Category）之间相关性的计算方法：

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

之后对每个分类节点的相关性进行归一化：

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

完成归一化后，以自底向上的顺序（Botton up）求出分类树种每个节点的相关性：

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

最后，以自顶向下的顺序（Top down）求出每条分类路径关于该用户（话题标签）的相关性：

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

路径集合之间的随机游走

经过 ESPA 方法计算后，为每个用户和话题标签筛选出相关性最高的若干条分类路径，将其放入一个分类路径的集合。本文首次定义 了两个路径集合之间的随机游走：

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

$$RW(\mathbb{P}_i \rightsquigarrow \mathbb{P}'_j) = RW(\mathbb{P}_{i(1)} \rightsquigarrow \mathbb{P}'_{j(1)})$$

$$RW(\mathbb{P}_{i(t)} \rightsquigarrow \mathbb{P}'_{j(t)}) = \alpha \cdot RW(C_{i(t)} \rightsquigarrow \mathbb{P}'_{j(t)}) + (1-\alpha) \cdot P(C_{i(t+1)}|C_{i(t)}) \cdot RW(\mathbb{P}_{i(t+1)} \rightsquigarrow \mathbb{P}'_{j(t+1)})$$

微博与 Twitter 间的交叉推荐

为了更加易于验证，本文选择的实验方式是将 Twitter 上的。

实验开始前，本文先使用 Google API 将所有的微博标签翻译成英文，之后将翻译后的微博标签与 Twitter 标签进行匹配，过滤掉微博和 Twitter 重复的话题，这样做的目的在于保证推荐给微博用户的话题是微博中没有的“新话题”。

实验部分

数据来源：本文使用的实验数据是微博和 Twitter 自 2012.9.17 至 2012.9.23 这一周时间的数据，以及维基百科 2014 年三月份的数据，数据共包含 3296945 条微博消息和 20128826 条 Twitter 消息。

实验构建了如下的 PGSMN 网络：

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

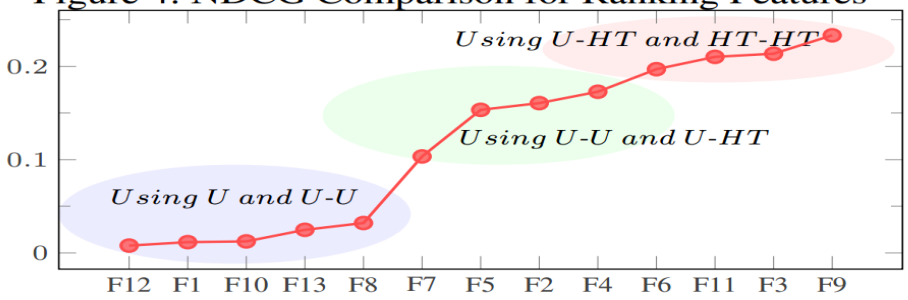
不同评价指标下的实验结果如下：

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 $p < 0.01$						

NDCG 指标下每个元路径特征的实验结果如下：

Figure 4: NDCG Comparison for Ranking Features



该结果说明，本文的实验结果可靠有效，远优于传统机器翻译推荐算法，且包含了用户与话题标签联系和标签之间相互联系的元路径的实验结果优于其他元路径，表明本文选用了微博与 Twitter 内部节点的多种关系，有效提高了实验效果。