一种高效的社交网络朋友推荐方案

程宏兵[1] 王 珂[2] 李 兵[2] 钱漫匀[1]
(浙江工业大学计算机科学与技术学院 杭州 310023)[1]
(国家电网浙江省电力有限公司金华供电公司 浙江 金华 321000)[2]

摘 要 当今社会,人们越来越多地通过社交网络来发言、聊天、交友。在互动过程中,除了用户主动关注感兴趣的人之外,社交网络也会为其推荐朋友。然而,所推荐的朋友大部分只是社交网络的推广,不一定符合用户的兴趣。针对社交网络推荐朋友的随机性和不可靠等问题,研究并提出了一种基于用户兴趣标签匹配的高效朋友推荐方案。首先,通过Word2Vec来训练语料库中的关键词,得到每个关键词的向量,产生一个词向量空间。其次,利用余弦相似度技术计算关键词之间的相似度并通过实验进行比较。实验中,综合选取合适的相似度值作为两个词向量是否相似的判断阈值。最后,将选取的相似度阈值应用到所提出的朋友兴趣匹配推荐算法中,并进行性能测试和各方案的仿真比较。结果表明,所提出的方案可靠且准确。

关键词 社交网络,朋友推荐,Word2Vec,相似度计算
中图法分类号 TP393.0 **文献标识码** A

Efficient Friend Recommendation Scheme for Social Networks

CHENG Hong-bing[1] WANG Ke[2] LI Bing[2] QIAN Man-yun[1]
(College of Computer Science & Technology, Zhejiang University of Technology, Hangzhou 310023, China)[1]
(Jinhua Branch of Zhejiang Power Company, National Grid, Jinhua, Zhejiang 321000, China)[2]

Abstract With the rapid development of modern network technology, human society has entered the era of information. An increasing number of people prefer to talk and make friends with others through social networks. Besides the people or events which users initiatively focus on, social network will also recommend alternative users. However, most of the alternative users are the promotion of social networks. In this paper, for the accuracy and reliability of social networks recommendation, a new scheme based on tag matching was proposed. First, each word in the corpus is trained by Word2Vec, and then a word vectors space can be obtained and the similarity among words can be obtained by using the cosine similarity. Secondly, through the similarity comparison experiments, this paper chose an appropriate similarity value as the threshold to judge whether two words are similar. Finally, the similarity threshold was applied to the matching algorithm. The simulation experiments show that the recommend users are relatively reliable and accurate.

Keywords Social networks, Friend recommendation, Word2Vec, Similarity degree computing

1 引言

近年来,社交网络在社会生活中变得越来越普遍,人们越来越多地通过社交网络聊天、互动和交友。精准预测并推荐新的朋友关系是社交网络演化、发展以及普及的重要途径之一。为社交网络中的目标用户推荐志同道合且安全可靠的朋友是社交网络价值体现的一种重要途径。因此,研究并设计一种兼顾准确性以及可靠性的高效、安全的朋友推荐方案是社交网络客户和运营商普遍关注的问题,也是近年来的研究热点。

安全的朋友推荐是为目标用户推荐与其最为相似的安全且可靠的用户作为朋友候选者,然后该目标用户从候选者中选择自己感兴趣的用户成为朋友。基于此,本文主要讨论目前社交网络的一个重要的基础问题——安全可靠的朋友推荐,即如何找到与目标用户最相似的朋友候选者并确保推荐过程中的隐私保护及其安全。

本文的研究思路如下:将朋友推荐问题归结为用户兴趣标签的安全匹配问题,然后通过分析用户的兴趣标签找到与目标用户志同道合的朋友。研究中,兴趣标签的匹配问题就是求用户之间兴趣标签的交集,即求两个集合的交集。例如:Amy的标签集为 $A=\{a_1,a_2,a_3,\cdots,a_n\}$,他想要关注与他兴趣相投的朋友,因此他向服务器提交了自己的标签集。服务器包含了由标签集合形成的标签库以及各自对应的用户ID,其中的用户都是通过被信任第三方认证的安全用户。

在标签库中:

Bob: $B=\{b_1,b_2,b_3,\cdots,b_m\}$

Candy: $C=\{c_1,c_2,c_3,\cdots,c_p\}$

David: $D=\{d_1,d_2,d_3,\cdots,d_q\}$

⋮

本文受国家自然科学基金项目(61402413)资助。

程宏兵(1979-),男,博士后,主要研究方向为网络与信息安全,E-mail:chenghb@zjut.edu.cn(通信作者);**王 珂**(1974-),女,硕士,主要研究方向为电网数据综合应用分析;**李 兵**(1972-),男,硕士,主要研究方向为电网数据分析技术;**钱漫匀**(1993-),女,硕士生,主要研究方向为社交网络与云安全。

服务器收到 Amy 的请求后，在标签库中开始匹配，即求 $A\cap B$，$A\cap C$，$A\cap D$，…，在这个过程中，不记录交集的具体内容，只记录用户 ID 以及与该用户交集的大小。按照要求获取需求用户之后，服务器将得到的结果发送给 Amy，Amy 可以从这些推荐用户中选取一些或全部用户进行关注，从而实现安全交友的目的。

2 相关工作

基于互联网衍生的社交网络已经深入到社会生活的诸多方面，人们越来越普遍地通过社交网络进行发言、聊天和交友，它为目标用户提供了一个扩充人脉的平台。目前，很多社交网络平台都基于一些现有的朋友推荐理论和技术进行推荐，考虑到利益、技术和认识层面的因素，现有的朋友推荐算法在效率、可靠性或隐私安全层面存在一些不足。

已有的成果中，Yin 等[1]首先提出了属性-社交网络的概念，对社交连接和属性间的关系进行了分析，并利用属性特征对用户间能否形成社交关系进行预测。在此基础上，文献[2-3]对属性-社交网络中的用户属性推测作了进一步分析，提高了链路预测的精确度。Wang 等[4]针对 1-邻居节点攻击，定义了一个关键隐私属性——概率的不可区分性，并基于此提出了启发式随机算法来生成匿名社交网络，以保护隐私。Shishodia 等[5]针对社交网络结构化匿名以及隐私属性保护，提出了一种贪婪算法，实现了 k-匿名和 l-多样性；另外，还提出了一个用来减少 $d>1$ 匿名代价的部分匿名概念。基于协同过滤，Khalid 等[6]提出了一种新的基于云的推荐框架——OmniSuggest，它结合蚁群算法、协同过滤和评价中心产生最佳推荐，通过用户兴趣的相似性来为个人或群体做更精确的推荐。Zhou 等[7]提出了一种 top-N 推荐方案，它是基于隐式用户偏好的协同推荐算法；另外，为了在贝叶斯网络中做出更好的近似推理，还开发了 EP 算法（Expectation Propagation (EP) message-passing algorithm），它适应于分布式实施，用户只跟与其直接连接的朋友交换信息，从而不向公众发布信息，进一步保护了用户的隐私。

李永凯等[8]研究了基于两种不同匹配类型的推荐方案：1）基于请求用户与被请求用户之间可用服务、可用应用等的匹配；2）基于请求用户与被请求用户之间兴趣、病症等的相似程度的匹配，针对用户不同的隐私要求，设计了 3 个不依赖同态加密的高效隐私内积计算协议，满足了不同匹配类型的需求。文献[9]基于网络结构局部特性的思想，将索引结构和推荐算法相结合，提出了一种新的朋友推荐排序索引树——RMPH-Tree，即通过将用户间的多属性相交值转换为二进制位码表示并组织成 PH_Tree 结构，利用树遍历过程来确定朋友推荐的最优顺序集。文献[10]提出了一个不采用任何可信任第三方的实际组匹配方案，介绍了一种模糊矩阵算法来生成用户的权限，减少了通信和计算的开销，然后建立一个公共集来存储所有组成员的属性和权限。文献[11]提出了一种新的可扩展性和隐私保护的朋友匹配协议——SPFM（Scalable and Privacy-preserving Friend Matching protocol），它在不用对外泄露个人资料的前提下，提供一个可扩展的朋友匹配和推荐方案，能防止 honest-but-curious mobile cloud 获取原始数据，同时支持多用户同时匹配。Guo 等[12]同时使用用户发布的信息和朋友关系来完成朋友推荐，该方法使用监督模型从用户发布的消息中提取代表用户兴趣的内容特征，同时从用户的朋友关系中抽取代表用户当前朋友链接的图特征，将上述 2 类特征作为监督模型的输入来预测 2 个用户之间是否会产生朋友关系。Huang 等提出了一种更加精确的自动推荐方案[13]，该方案分为两个阶段：第一阶段利用用户发布的信息以及用户的朋友关系选择一些“可能的朋友”；第二阶段利用图像特征和用户之间的关系建立了一个主题模型，进一步细化了推荐结果。上述已有研究成果在某些方面具有一定的优势，但在效率、安全和可靠性方面需要进一步加强。

3 本文模型

针对社交网站交友中兴趣标签的匹配问题，本文提出了一种基于 Word2Vec 的兴趣标签匹配技术，以获取安全、高效的朋友推荐方案。首先，我们获取了完整的维基百科文本，对其进行分析，再利用 Word2Vec 训练关键词，得到较为完整的关键词向量库；然后，获取新浪微博用户 ID 和标签作为测试集，并利用余弦相似度计算标签之间的相似性；最后，基于设计的相似度算法匹配用户兴趣标签，为目标用户选取合适的推荐用户。

按照标签的势将新浪微博用户进行分类，表 1 列出分类的样例。每个用户有自己的 ID 和标签集，标签集中有标签，每个标签都对应自己的词向量，我们通过向量标签来进行标签匹配。

表 1 分类样例

势	用户 ID	用户标签
1	1881455361	[摄影师]
2	1766722263	[创意，科技]
3	2387860933	[媒体，英语，小说]
4	1739268883	[旅行，摄影，摇滚，自由]
5	1726076897	[科学，哲学，人文，建筑，艺术]
6	1649717113	[动漫，街舞，科技，足球，武侠，金融]
7	1737472277	[理财，篮球，支付，游泳，双鱼座，金融，电子商务]

3.1 Word2Vec 简介

Word2Vec 是 Google 于 2013 年开源的一款将词表示为实数值向量的高效工具，采用了连续词袋模型（Continuous Bag-Of-Words，CBOW）和 Skip-Gram 两种模型。它实际上分为了两个部分，第一部分建立模型，第二部分通过模型获取嵌入词向量。它利用深度学习的思想，通过训练，可以把对文本内容的处理简化为 K 维向量空间中的向量运算，而向量空间上的相似度可以用来表示文本语义上的相似度。

Word2Vec 使用的是 Hinton[14]于 1986 年提出的 Distributed representation 的词向量表示方式。它的基本思想是通过训练将每个词映射成 K（K 一般为模型中的超参数）维实数向量，通过词之间的距离（即余弦相似度）来判断它们之间的语义相似度。

超参数：在机器学习中，超参数的值是在学习过程开始之前设置的；而其他参数的值是通过训练得到的。它定义了关于模型的更高层次的概念，如复杂性或学习能力，可以通过设置不同的值、训练不同的模型和选择更好的测试值来决定。本文将 K 值设置为 80，即词向量为 80 维。

余弦相似度：

$$\alpha=\cos\theta=\frac{\sum_{1}^{n}(x_i\times y_i)}{\sqrt{\sum_{1}^{n}x_i^2}\times\sqrt{\sum_{1}^{n}y_i^2}}$$

余弦值的范围在[－1，1]之间，值越趋近于 1，代表两个向量的方向越趋近于 0，即它们的方向更加一致，相应的相似度也越高。

本文采用一个三层(即输入层-隐层-输出层)的神经网络模型——Skip-Gram，如图 1 所示。这个三层神经网络本身是对语言模型进行建模，先基于训练数据构建一个神经网络，这个模型训练好后并不会被用来处理新的任务。训练模型的真正目的是获得模型基于训练数据学得的隐层权重，即这个模型通过训练数据所学得的参数。

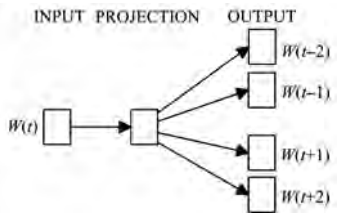

图 1　Skip-Gram 模型

3.2　集合的交集和势

集合的交集：设 A，B 是两个集合，将由所有属于集合 A 且属于集合 B 的元素所组成的集合叫做集合 A 与集合 B 的交集(intersection)。

假设：$A=\{a,b,c,d,e\}$，$B=\{c,d,e,f\}$，$C=A\cap B=\{c,d,e\}$，即 $A\cap B=\{x|x\in A\wedge x\in B\}$。

集合的势：用来度量集合规模大小的属性，即集合中元素的个数。如上文中 A 的势是 5，B 的势是 4，C 的势是 3。

标签匹配即求两个标签集的交集，简而言之就是求两个用户之间标签相似的个数。我们将用户的标签集看成数学中的集合，利用求集合交集的概念来获取两个用户之间标签集的交集和交集的势。

本文基于这种思想，将用户标签集看作一组词的集合，首先判断两个标签集中的兴趣标签是否相似，然后判断两个标签集交集的势是否满足相似要求。基于 Word2Vec，每个兴趣标签都能映射到 K 维向量空间，再利用余弦相似度求得兴趣标签之间的相似性。

3.3　相似性计算

为了得到较好的匹配结果，本文需要选取合适的相似度阈值。选取相似度阈值方案的具体步骤描述如下：

步骤 1　获取与每个词向量最相似的前 20 个词向量以及它们的相似度值 α_i。

首先将每个词向量与向量空间中的其他词进行比较，得到所有的相似度；然后对其进行排列，得到前 20 个词向量和对应的相似度值 α_i。

步骤 2　得到以 α_i 为比较阈值的词向量集合的交集大小。

1)基于词向量空间，自动生成 n 个词向量为一组的集合库，假设每一个集合代表一个用户；

2)以 α_i 为比较阈值，计算每两个用户之间集合的交集大小 S；

3)统计交集大小的数据，得到 $S=k$ 时的个数 NUM_s。

整个选取过程如算法 1 所示。

算法 1　相似度阈值的选取

输入：词向量空间

输出：NUM_s

```
1. WordsNearest(Word,20); //获取与词向量 Word 最相似的前 20 个词
2. similarity(Word,smWord);//得到与前 20 个相似词的相似度值 αi
3. AutomaticGenerator();//自动生成词向量组合而成的集合库
4.   For(i=1;i<=20;i++) //选择 αi 为阈值，得到交集 S 的大小
5.     S=0;
6.     if(similarity(tag1,tag2)>=αi)
7.       S=S+1;
8.   END For
9. StatisticInter(); //统计 S=k 时，NUMs 的值
```

如算法 1 所示，WordsNearest(*Word*，20)中，*Word* 表示核心词，20 表示获取与它最相似的 20 个词；similarity(*Word*，*smWord*)中，*Word* 表示核心词，*smWord* 表示与 *Word* 相似的词；similarity(*tag*1，*tag*2)中，*tag*1 表示匹配词，*tag*2 表示被匹配词，用以比较两个用户之间的相似度标签集。

表 2 以“音乐”为例，列出了与关键词“音乐”最相似的前 5 个词及其对应的相似度值(相似度值保留 6 位有效数字)。

表 2　与“音乐”最相似的前 5 个词及其对应的相似度值 α_i

音乐	摇滚	歌曲	乐队	录制	伴奏
相似度	0.907948	0.900203	0.896260	0.893945	0.886732

表 3 是以两个标签为例的 $Alpha_i$-NUM_s 表。我们从词向量库中选取了 500 个标签组合，当 $Alpha_i$ 取不同值时，分别计算 $Alpha_5$，$Alpha_{10}$，$Alpha_{15}$，$Alpha_{20}$ 的集合交集。从表中可以发现，随着 $Alpha_i$ 值的变化，NUM_s 值也在变化，具体为：从 $Alpha_5$ 到 $Alpha_{10}$ 时，$NUM_s=0$，即没有一个匹配成功，其值在减小；$NUM_s=1$，即有 1 个匹配成功，其值在增加；$NUM_s=2$，即 2 个都匹配成功，其值也在增加。从 $Alpha_{10}$ 到 $Alpha_{15}$，$NUM_s=0$，即没有一个匹配成功，其值在增加；$NUM_s=1$，即有 1 个匹配成功，其值减小；$NUM_s=2$，即 2 个都匹配成功，其值也在减小。从 $Alpha_{15}$ 到 $Alpha_{20}$，$NUM_s=0$，即没有一个匹配成功，其值在减小；$NUM_s=1$，即有 1 个匹配成功，此时的值在增加；$NUM_s=2$，即 2 个都匹配成功，其值也在增加，但幅度相比前面两个区间段的增减来说不大。由此可见，在 $Alpha_{10}$ 附近时，成功率匹配值最高。

表 3　两个标签的 $Alpha_i$-NUM_s 表

NUM_s \ $Alpha_i$	$i=5$	$i=10$	$i=15$	$i=20$
$NUM_s=0$	4428	4401	4326	4244
$NUM_s=1$	504	530	586	653
$NUM_s=2$	18	19	38	53

3.4　基于 Word2vec 的匹配算法

标签匹配主要是为了求两个标签的交集，即判断两个用户之间标签相似的个数。本文将选取用户的标签作为用户的特征向量，即最后是求特征向量之间的相似度。3.3 节中已经获取了标签的相似度值，本节将利用 3.3 节中的相似度值来进行标签匿的名匹配。所提出的匿名匹配方案的算法步骤具体如下：

步骤 1　得到符合要求的用户。

计算匹配者输入的标签个数 n。如果 n 为奇数，那么只有当相似标签的个数大于或等于$[n/2]+1$ 时，才认为匹配成功；如果 n 为偶数，那么只有当相似标签的个数大于$[n/2]$时，才认为匹配成功。

之后，过滤掉原始标签个数小于$[n/2]+1$ 的用户，从而得到新的被匹配用户集合。

步骤 2　得到匹配成功的前 20 个用户的 ID 和交集大小 S。

1)将匹配者的标签集的第一个标签作为原始标签,并与被匹配用户的每一个标签进行比较。用 3.2 节中的 α_i 作为阈值,如果与其中某个标签的相似度值大于 α_i,就执行 $S+1$,停止比较,并跳到第 2)步。

2)判断匹配者的标签是否匹配完,如果没有,则继续执行步骤 1),否则停止执行。

整个匹配过程如算法 2 所示。

算法 2 标签匿名匹配

输入:匹配发起者的标签集:A={a_1,a_2,a_3,…,a_n}

输出:匹配成功的前 20 个用户的 ID 和交集大小 S

```
1. SizeOf(A);//计算 A 标签的个数 n
2. Filter();//过滤掉标签个数小于[n/2]+1 或者[n/2]的用户
3. For(i=0;i<n;i++)
4.   S=0;
5.   SizeOf();//计算被匹配标签的个数 m
6.    For(j=0;j<m;j++)
7.        if(similarity(tag1,tag2)>=αi)
8.            S=S+1;
9.    END FOR
10. END FOR
11. Print(ID,S).
```

在算法 2 中,函数 Filter()在匹配之前就过滤掉了不符合要求的被匹配用户,因此可以在很大程度上提高匹配的效率。两个 For 循环就是标签匹配的过程,即求集合交集的过程。函数 similarity($tag1$,$tag2$)用于求得标签 $tag1$ 和 $tag2$ 的相似度,$tag1$ 为主动标签,即匹配发起者的标签;$tag2$ 为被动标签,即符合过滤要求的被匹配者标签。

4 仿真实验与结果分析

4.1 相似度阈值的选取

实验中从 600000 用户中选取了 3 个测试集,每个测试集有 500 个用户,包括 100 个 2 标签用户,100 个 3 标签用户,100 个 4 标签用户,100 个 5 标签用户以及 100 个 6 标签用户。同时,还获取了一个维基百科文本,对文本进行分词后,使用了现在流行的深度学习成果 Word2Vet 对语料库进行训练,最后得到一个对应的分布式词向量空间,大约包含 2000000 个词条。新浪用户的标签都能在词向量空间找到对应的向量表示。

对 3 个测试集的用户进行匹配实验,得到了匹配成功的 $Alpha_i$-NUM_s 对照图,结果如图 2－图 4 所示。图 2 是在测试集 D_1 上的实验结果,可以发现,随着 $Alpha_i$ 值的减小,匹配成功的 NUM_s 值增加,但当 $Alpha_i$ 达到某一阈值时,NUM_s 的值随着 $Alpha_i$ 的减小而减小。整体变化遵循这个规律,但当 i 值为 12 时,NUM_s 的值发生了突变。同样地,图 3 和图 4 分别展示了在测试集 D_2 和 D_3 上的实验结果,描述了匹配成功的 NUM_s 值随着 $Alpha_i$ 值变化的情况。

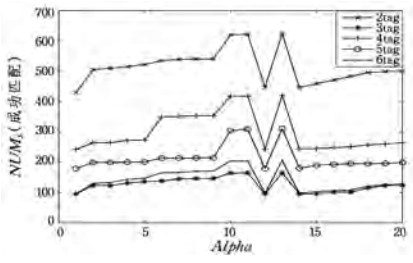

图 2 基于 D_1 测试集的 $Alpha_i$-NUM_s 结果

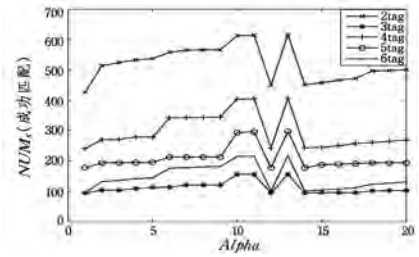

图 3 基于 D_2 测试集的 $Alpha_i$-NUM_s 结果

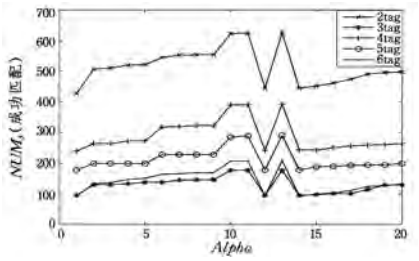

图 4 基于 D_3 测试集的 $Alpha_i$-NUM_s 结果

通过分析图 2 可以发现,在测试集 D_1 上,当 i 值为 11 和 13 时,匹配成功的 NUM_s 值最大。同样地,发现在测试集 D_2 和 D_3 上,即 i 值为 11 和 13 时,NUM_s 的值最大。为了给网络用户提供更多不同的选择,选取 $Alpha_{11}$ 和 $Alpha_{13}$ 作为相似度阈值进行了相关实验。同时,我们发现当 i 值为 12 时,NUM_s 的值骤然减小,呈不规则变化。为了实验的正确性,我们也需要验证当 i 值为 12 时的这种特殊情况。

4.2 不同方案的性能对比

将 4.1 节中测试集 2 的标签组混合在一起作为新的实验数据集。本实验将对比本文方案与目前常用的 LDA 主题分布方案在准确率和召回率上的优劣,同时对比在不同相似度阈值下两个方案的准确率和召回率的优劣。

LDA[15](Latent Dirichlet Allocation)是一种文档主题生成模型,也被称为一个三层贝叶斯概率模型,包含词、主题和文档三层结构。所谓生成模型,就是说,我们认为一篇文章的每个词都是通过“以一定概率选择了某个主题,并从这个主题中以一定概率选择某个词语”这样一个过程得到的。其中每个词语出现的概率为:p(词语|文档)$=\sum_{\text{主题}} p$(词语|主题)$\times p$(主题|文档),这个概率公式可以用矩阵表示为图 5 所示的形式。其中,“文档-词语”矩阵表示每个文档中每个单词的词频,即出现的概率;“主题-词语”矩阵表示每个主题中每个单词出现的概率;“文档-主题”矩阵表示每个文档中每个主题出现的概率。给定一系列文档,通过对文档进行分词来计算各个文档中每个单词的词频,从而可以得到左边的“文档-词语”矩阵。主题模型便是通过左边这个矩阵进行训练,进而学习出右边两个矩阵。本文的 LDA 模型中训练主题是 200 个,迭代 1000 次。

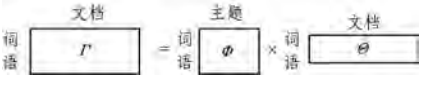

图 5 主题模型的矩阵表示

本文方案根据 Word2Vec 匹配算法,分别以每个用户的标签组为主动标签,设置 $Alpha_{11}$,$Alpha_{12}$ 和 $Alpha_{13}$ 为相似度阈值;在测试集 2 中进行 3 次匹配实验,得到最终的 3 个标签匹配的结果;最后再对匹配结果进行分类处理。本文以哈尔滨

(下转第 452 页)

Conference on Computational Intelligence and Software Engineering. IEEE,2010:1-4.
[8] FOGUE M,GARRIDO P,MARTINEZ F J,et al. Using Data Mining and Vehicular Networks to Estimate the Severity of Traffic Accidents[M]// Management Intelligent Systmes. Springer Berlin Heidelberg,2012:37-46.
[9] 李瑶.数据挖掘技术在交通事故分析中的应用[J].电子设计工程,2009(2):77-78,81.
[10] ZHANG C,WANG S. Application of Data Mining in Urban Traffic Accidents Governance Based on Association Rules[J]. Lecture Notes in Computer Science,2012,4(19):169-176.
[11] 董立岩,刘光远,苑森淼,等.数据挖掘技术在交通事故分析中的应用[J].吉林大学学报(理学版),2006,44(6):951-955.
[12] 王冬秀,李辉.关联规则在道路交通事故中的应用研究[J].福建电脑,2010,26(7):7-8.
[13] 魏玉晓,李宗平,李宵寅.基于加权关联规则的交通事故分析[J].交通信息与安全,2009,27(1):94-97.
[14] 肖冬荣,杨磊.基于遗传算法的关联规则数据挖掘[J].通信技术,2010,43(1):205-207.
[15] AGRAWAL R,IMIELIŃSKI T,SWAMI A. Mining Assocation Rules between Sets of Items in Large Databases[J]. Acm Sigmod Record,1993,22(2):207-216.
[16] AGRAWAL R,SRIKANT R. Fast algorithms for mining association rules(3rd ed.)[M]// Readings in database systems. Morgan Kaufmann Publishers Inc. 1998:2299-308.
[17] SRIKANT R,AGRAWAL R. Mining quantitative association rules in large relational tables[C]// ACM SIGMOD International Conference on Management of Data. ACM,1996:1-12.
[18] 哈工大停用词表[EB/OL]. https://wenku.baidu.com/view/b8b30382e53a580216fcfeb7.html.
[19] 四川大学机器智能实验室停用词库[EB/OL]. https://wenku.baidu.com/view/37f18269561252d380eb6e1e.html.
[20] 百度中文停用词表[EB/OL]. https://wenku.baidu.com/view/5059a59c2e3f5727a4e96245.html.

(上接第 436 页)

工业大学同义词林为标准,测试的实验结果如表 4 所列。

表 4 不同方案的性能对比

方案	准确率	召回率
LDA 主题分布	0.339494	0.767173
本文方案(i=11)	0.864290	0.994675
本文方案(i=12)	0.840233	0.993996
本文方案(i=13)	0.833646	0.993743

从表 4 中可以看出,本文所提方案的准确率和召回率都高于 LDA 主题分布。实验证明,当 i 取值为 11 时,准确率和召回率最高;此外,当 i 取值为 12 时,并不影响实验的准确率和召回率,只是召回总数 NUM_s 减少而已。

结束语 本文介绍了一种新的朋友推荐方案,该方案使用用户的兴趣标签,并基于 Word2Vec 设计了标签匹配算法。具体为:将用户标签看成一个集合,分析用户的兴趣标签;根据相似度计算方法,求不同用户标签之间的相似度,并根据相似度阈值求两个标签集的交集。此外,本文还详细介绍并讨论了相似度阈值的选取过程,并通过标签匹配实验比较了选取合适相似度阈值的方法。最后,与其他模型进行了对比实验,仿真结果证明了所提方案可以得到准确且可靠的朋友推荐效果。

参 考 文 献

[1] YIN Z,GUPTA M,WENINGER T,et al. LINKREC:a unified framework for link recommendation with user attributes and graph structure[C]// International Conference on World Wide Web(WWW 2010). Raleigh, North Carolina, USA, DBLP, 2010:1211-1212.
[2] YANG S H,LONG B,SMOLA A,et al. Like like alike:joint friendship and interest propagation in social networks[C]// International Conference on World Wide Web. 2011:537-546.
[3] GONG N Z,TALWALKAR A,MACKEY L,et al. Jointly Predicting Links and Inferring Attributes using a Social-Attribute Network(SAN)[C]// The 6th SNA-KDD Workshop(SNA-KDD'12). Beijing,China,2012.
[4] WANG G,LIU Q,LI F,et al. Outsourcing privacy-preserving social networks to a cloud[C]// IEEE INFOCOM. IEEE,2013:2886-2894.
[5] SHISHODIA M S,JAIN S,TRIPATHY B K. GASNA:greedy algorithm for social network anonymization[C]// IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE,2013:1161-1166.
[6] KHALID O,KHAN M U S,KHAN S U,et al. OmniSuggest:A Ubiquitous Cloud-Based Context-Aware Recommendation System for Mobile Social Networks[J]. IEEE Transactions on Services Computing,2014,7(3):401-414.
[7] ZOU J,FEKRI F. On top-N recommendation using implicit user preference propagation over social networks[C]// IEEE International Conference on Communications. IEEE,2014:3919-3924.
[8] 李永凯,刘树波,杨召唤,等.机会网络中用户属性隐私安全的高效协作者资料匹配协议[J].通信学报,2015,36(12):163-171.
[9] 梁俊杰,孙阳征.基于 PH-Tree 多属性索引树的朋友推荐算法[J].计算机科学,2015,42(4):156-159.
[10] LI F,WANG H,NIU B,et al. A practical group matching scheme for privacy-aware users in mobile social networks[C]// Wireless Communications and NETWORKING Conference. IEEE,2016.
[11] LI M,NA R,QIAN Q,et al. SPFM:Scalable and Privacy-Preserving Friend Matching in Mobile Cloud[J]. IEEE Internet of Things Journal,2017,4(2):583-591.
[12] GUO L,ZHANG C,FANG Y. A Trust-Based Privacy-Preserving Friend Recommendation Scheme for Online Social Networks[J]. IEEE Transactions on Dependable & Secure Computing,2015,12(4):413-427.
[13] HUANG S,ZHANG J,DAN S,et al. Two-Stage Friend Recommendation Based on Network Alignment and Series Expansion of Probabilistic Topic Model[J]. IEEE Transactions on Multimedia,2017,19(6):1314-1326.
[14] HINTON G. Learning distributed representations of concepts[C]// Eighth Annual Conference of the Cognitive Science Society. 1986:1-12.
[15] BLEI D M,NG A Y,JORDAN M I. Latent dirichlet allocation[J]. J Machine Learning Research Archive,2003,3:993-1022.