基于内容与社会过滤的好友推荐算法研究

高永兵,杨红磊,刘春祥,胡文江
(内蒙古科技大学 信息工程学院,内蒙古 包头 014010)

摘 要:基于内容算法与社会过滤算法都是迄今为止在社交网络中较为成功的好友推荐算法。结合两者的优点,根据用户已有的好友来给用户推荐新的好友,并与用户的兴趣爱好、地理位置等个人信息相结合的方式来处理好友推荐问题。通过实验验证以及准确率和召回率的评测显示,改进的算法比传统的好友推荐算法在推荐性能上有较为明显的提高。

关键词:社会过滤;好友推荐;内容相似性;基于内容算法
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1674-7720(2013)14-0075-04

Friends recommendation algorithm based on the content and social filtering

Gao Yongbing, Yang Honglei, Liu Chunxiang, Hu Wenjiang
(School of Information Engineering, Inner Mongolia University of Science and Technology, Baotou 014010, China)

Abstract: Content-based algorithm and social filtering algorithm are all successful algorithms in social network friend recommendation. This paper combines the advantages of the two algorithms, basing make new friends by means of connecting with users' old friends, and combining personal information such as users' interests, geographical location etc. to solve the problem of friends recommendation. Through experiments and the precision rate and recall rate evaluation, it showed that the new algorithm is more improved than the traditional friend recommendation algorithm in recommendation functions.

Key words: social filtering; recommendation of friends; content similarity; content-based algorithm

伴随着 Web2.0 的到来,各种各样的社交网站不断涌现。如国外的 Facebook、twitter、Flickr,国内的新浪微博、人人网等。在这些社交网站上,用户能够添加自己在日常生活中已经认识的好友,也可以在网络上结交新的朋友[1]。

经研究,Savage[2]发现在 Facebook 这类网站上用户主要是和自己在生活中认识的人进行交流;Dimicco[3]则发现在 twitter 等网站上用户则更倾向于和自己不认识的人结交好友;Ehrlich[4]介绍了依据用户的聊天信息来帮助用户寻找专家的方法,但是不能够给用户推荐个性化好友。

在社交网络中,用户想要添加的好友不仅仅是自己在日常生活中的好友,同时那些虽然用户不认识但极为感兴趣的其他用户也是理想的好友。无论是为用户推荐现实中已经认识的朋友还是推荐和用户有共同兴趣的好友,目前的好友推荐算法都不能很好地解决问题。本文通过分析现有好友推荐算法的不足之处,综合考虑用户日常生活中的好友以及用户的兴趣爱好等个人信息,有效地解决了好友推荐中遇到的冷启动(为新注册用户推荐好友)、用户个人信息过少而无法推荐好友等问题。

1 社会过滤算法

社会过滤算法(social filtering)建立在这样一个前提下:如果甲的朋友是乙的朋友,那么甲有可能是乙的朋友。已经有很多社会网络分析方法采用了类似的方法找到了邻居和合适的途径。这种推荐方法不仅仅是通过考虑用户的兴趣爱好,还通过分析隐含在用户每一个好友身上的信息,来给用户准确地推荐好友[5]。这种算法主要用在社交网络中的"你可能认识的人"这一板块。

在介绍算法之前,给出以下定义:在社交网络中,如果用户 b 是用户 a 的好友,那么定义为 $F(a,b)$。算法描述如下:

假设存在用户 a 和用户 u,用户 u 的推荐好友候选集定义为 $RC(u)$,用户 a 是用户 u 的好友,同时用户 c 又是用户 a 的好友,则用户 c 为用户 u 的推荐候选集中的一个用户。用公式表达为:

$$RC(u)=\{c|F(u,a)\wedge F(a,c)\wedge\neg F(u,c)\}$$

共同好友集定义为 $MF(u,c)$，公式表达为：

$$MF(u,c)=\{a|F(u,a)\wedge F(a,c)\}$$

通过共同好友的关系，在用户 u 和用户 c 之间添加了联系。然后通过计算共同好友集 $MF(u,c)$ 得出用户 c 的可推荐百分比。通过候选集中用户可推荐百分比的高低排序给目标用户 u 推荐得分最高的用户 Top-N。

该算法较为适合与现实社会有较大联系的社交网站，其较大的不足之处为目标用户必须有一定的好友数量的积累。对于一个新注册或者好友数量较少的用户来说，不能够使用该算法来给用户推荐好友。

2 基于内容推荐算法

基于内容推荐算法基于以下思想：如果两个人有相似的话题，他们也许会更愿意去认识对方。换句话说，这个算法是努力地寻找与目标用户有相似爱好的用户。这与信息挖掘领域的发现文档之间相似内容的方法极为相似。

首先使用文本内容创建一个词向量代表每一个用户。从用户的个人设置项和状态信息（发布的文章信息、对个人的描述等）中提取关键词[6]，也可以提取用户工作所在地等信息。所有保留的词通过一个词向量 $V_u=(v_u(w_1),\cdots,v_u(w_m))$ 来描述用户 u，m 代表所有单词的数量，每一个 $v_u(w_i)$ 代表用户 u 的兴趣词，w_i 代表这个词在用户所有的兴趣中的权重。单词 $v_u(w_i)$ 的权重通过 TF-IDF 算法来计算：

$$TF_u(w_i)=\frac{u(w_i)}{W}$$

其中 $u(w_i)$ 代表用户 u 使用过的保留词，W 代表用户 u 使用过的所有单词。

$$IDF_u(w_i)=\log\frac{E}{U}$$

其中 E 代表所有的用户，U 代表在所有用户中使用过单词 $v_u(w_i)$ 的用户数。

$$v_u(w_i)=TF_u(w_i)\times IDF_u(w_i)$$

通过余弦相似度来计算用户 a 和用户 b 的两个向量 V_a 和 V_b 的相似度。可以直观地认为如果用户 a 和用户 b 在日常使用中分享了相同的关键词，而其他用户很少分享这些关键词，则他们有很大的相似性。作为一个被推荐的用户 c，在所有分享的关键词中，只显示前 10 个数量积最高的关键词。直观地认为它们是用户 u 和用户 c 分享的最具有代表性的关键词。

基于内容和链接的算法主要是通过使用社交网络中的社交链接信息加强基于内容匹配算法的准确度[7]。算法通过将那些社交网络中的弱约束和隐式用户显示出来，目标用户更乐意于接受这种算法。此算法与基于内容的算法中计算相似度的方法有很大的相似之处。然而，与向用户推荐前几位相似度最高的用户方法不同的是，如果用户 u 和用户 c 之间存在有效的链接，给用户 u 和用户 c 之间的相似度加 50%的权重，即如果用户 u 和用户 c 之间存在联系，那么在推荐时它的推荐顺序将会排在基于内容相似度之前。

一个有效的链接的定义为：将若干个用户排成一队，第一个用户作为目标用户，最后一个用户作为被推荐用户，每两个用户 a 和 b 之间都必须至少满足以下 3 个条件之一：

(1)a 主动联系过 b；

(2)a 对 b 有过评论；

(3)b 主动联系过 a。

该定义确保了两个用户之间存有社会链接并且最低限度地认为他们或者他们的好友之间是熟人或者有一定的互动关系。例如用户 a 给用户 c 评论过，而用户 b 和用户 c 又是好友关系，则认为用户 a 和用户 b 之间存在一个有效链接。

在推荐时使用有效链接，同时还考虑相同关键词的内容匹配技术，也可以把链接作为一种扩展，包括考虑用户 u 和候选集中用户 c 之间的所有链接。在推荐的用户中，至少 77.8%都需要考虑有效链接信息。

3 个性化好友推荐算法

为了解决社会过滤算法遇到的冷启动问题以及基于内容相似性算法的准确率较低问题，根据对现有算法的总结，本文提出了改进的个性化好友推荐算法。经过实验验证，本算法能够有效地解决这些问题。

根据用户的个人特征信息，计算出与目标用户词特征向量最为相似的用户集，即要产生一个与用户 u 的特征信息相似性从大到小排列的推荐集。对于目标用户 u，通过他的个人特征信息及特定相似度函数，计算出与他的特征信息最相近的 N 个用户作为目标用户 u 的最近邻居集，即为目标用户 u 的 Top-N 推荐集。

(1)收集用户信息

在社交网站中，用户会描述自己的兴趣以及个人信息。例如在人人网中，用户注册时会选择自己所在学校、专业、班级、地理位置等，这些就代表了用户的个人特征；在微博中，用户会选择自己感兴趣的方向、擅长的领域等标签，这些也同样代表了用户的个人特征。推荐算法给用户推荐好友时，应该充分利用用户的这些个人特征信息。

(2)建立用户的词特征向量(UserVector)

建立一个词特征向量 $V_u=(w_1,\cdots,w_i,\cdots,w_m)$ 来描述用户 u，其中 m 代表用户的单词数量，w_i 代表用户的个人信息（兴趣爱好、地理位置等）。此处按照每个网站中的特定顺序来给用户的词特征向量中的每一个词排序。

(3)计算用户特征向量之间的相似度

通过余弦相似度来计算用户 u 和用户 a 的两个向量 $\overline{V}_u$ 和 $\overline{V}_a$ 的相似度。

$$Sim(u,a)=\cos(\overline{V}_u,\overline{V}_a)=\frac{\overline{V}_u\cdot\overline{V}_a}{|\overline{V}_u|\cdot|\overline{V}_a|}$$

通过相似度的计算，得到与目标用户 u 特征信息最为相似的 Top-N 推荐集。

(4)生成推荐好友候选集

取出 N 个最靠前的用户作为目标用户的推荐好友候选集，即产生一个与目标用户 u 的个人信息相似度从高到低排列的推荐好友候选集 r。

(5)检测目标用户好友数

根据目标用户 u 的好友数量来确定是否继续使用社会过滤推荐算法。如果目标用户 u 没有好友，则直接将推荐出来的推荐好友候选集推荐给用户 u；如果目标用户 u 有好友，则继续使用社会过滤推荐算法给用户 u 推荐好友。

(6)计算目标用户和推荐好友候选集中每个用户的共同好友数

目标用户 u 的推荐好友候选集定义为 $RC(u)$，用户 a 是用户 u 的好友，用户 c 又是用户 a 的好友，同时用户 c 不是用户 u 的好友。则用户 c 为用户 u 的推荐候选集中的一个用户。用公式表达为：

$$RC(u)=\{c|F(u,a)\wedge F(a,c)\wedge\neg F(u,c)\}$$

共同好友集定义为 $MF(u,c)$，公式表达为：

$$MF(u,c)=\{a|F(u,a)\wedge F(a,c)\}$$

通过共同好友的关系，在用户 u 和用户 c 之间添加了联系。然后通过计算共同好友集 $MF(u,c)$ 得出用户 c 的可推荐百分比。

设 $\boldsymbol{P}$ 是 n 个用户之间的好友关系矩阵。在这个矩阵里，如果用户 i 和用户 j 是好友关系，则 P_{ij} 为 1，否则为 0。

$$\boldsymbol{P}=\begin{bmatrix} p_{11} & \cdots & p_{1j} & \cdots & p_{1n} \\ \vdots & & \vdots & & \vdots \\ p_{i1} & \cdots & p_{ij} & \cdots & p_{in} \\ \vdots & & \vdots & & \vdots \\ p_{n1} & \cdots & p_{nj} & \cdots & p_{nn} \end{bmatrix} \quad \boldsymbol{A}=\begin{bmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1n} \\ \vdots & & \vdots & & \vdots \\ a_{i1} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & & \vdots & & \vdots \\ a_{n1} & \cdots & a_{nj} & \cdots & a_{nn} \end{bmatrix}$$

设 $\boldsymbol{A}$ 为从矩阵 $\boldsymbol{P}$ 计算得出的关联矩阵，含 n 个用户彼此关联规则的置信度。$\boldsymbol{A}$ 是 $n\times n$ 的一个矩阵，n 为用户的数量，$a_{i,j}$ 是 $i\Rightarrow j$ 关联规则的置信度，$a_{i,j}$ 表示同时是用户 i、j 的好友的用户在所有用户 N 中的比例。

目标用户 u 的偏好向量 $\boldsymbol{u}$ 为一个 $1\times n$ 的矩阵，u_{ij} 表示目标用户 u 和用户 j 的共同好友关系，它是 $\boldsymbol{P}$ 矩阵的横向量。为目标用户推荐的矢量 s 可以从计算关联矩阵 $\boldsymbol{A}$ 和目标用户的偏好向量 $\boldsymbol{u}$ 的乘积得出，计算公式为：

$$\boldsymbol{s}=\boldsymbol{u}\times\boldsymbol{A}$$

(7)根据共同好友数对推荐好友候选集重新排序

根据目标用户与推荐集中用户的共同好友个数，产生一个与目标用户的共同好友数从多到少的好友推荐候选集 s。

(8)选定合适的权重

选定权重的值，新算法的计算公式如下：

$$NF=\alpha\times r+(1-\alpha)s$$

其中 NF 表示新算法，α 表示权重。

(9)将重新排序后的 Top-N 好友推荐给目标用户。

4 实验结果及分析

4.1 测试数据集

实验采用的数据是从人人网收集的好友信息数据集。本次实验共收集了将近 5 万个用户信息，为提高实验算法的准确性，此处过滤掉好友数量少于 20 的用户，最终得到 7 630 个用户，包含 268 943 个好友关系，每个用户约有 20~50 个好友关系。本实验采用交叉验证[8]，将数据集 80%的训练集和 20%的测试集对不同的算法进行分析。同时为验证实验的准确性，实验也将每一个用户的好友随机分为 80%的使用集和 20%的验证集，并对实验数据进行多次运算取平均值。

验证算法所用的硬件平台为 Intel® Core™ 2 Duo CPUE7400，主频为 2.8 GHz，2 GB 内存，320 GB 硬盘。操作系统为 Windows Professional sp3，所有算法用 Visual C++语言实现。

4.2 评价标准

测试结果的评价指标采用 Top-N 推荐中使用的准确率(Precision)、召回率(Recall)和 F 度量(F-measure)。准确率定义为：

$$\text{Precision}=\frac{\text{hit}}{N}=\frac{[\text{Test}\cap\text{Top-}N]}{N}$$

召回率定义为：

$$\text{Recall}=\frac{\text{hit}}{|\text{Test}|}=\frac{|\text{Test}\cap\text{Top-}N|}{|\text{Test}|}$$

其中，hit 为命中的数量，Test 为验证集，N 为向用户推荐的好友数量。

将这两个度量值融合成一个度量值，就是 F 度量(F-measure)：

$$\text{F-measure}=\frac{2\times\text{Recall}\times\text{Precision}}{\text{Recall}+\text{Precision}}$$

此处首先根据实验结果，取得个性化算法最优推荐时的 α 值，并对个性化算法(NF)、社会过滤算法(SF)、基于内容推荐算法(CB)3 种算法进行评价。在本实验中，对推荐出的 Top-N 的个数 N=2，4，6，8，10 这 5 种情况分别进行评价。

4.3 实验结果

图 1 显示了个性化好友推荐算法在 α 取不同值时的 F-measure 值。结果显示，当 α 取 0.4 时 F-measure 值最大，此时个性化推荐算法(NF)最优。

图 2 显示了 3 种不同推荐算法 F-measure 的比较结果。表 1 显示了不同情况下，各算法详细数据记录，数据显示当推荐用户不断增加时，各个指数性能也随之增加，在 4~8 个推荐用户时达到最大。这说明一次给用户推荐的好友数不宜太多，6 个左右最佳，同时也显示出本文的好友推荐算法比单一算法效率更高。

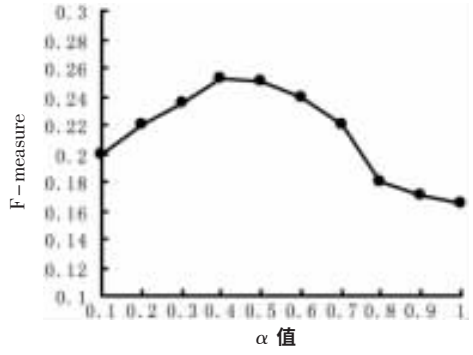

图 1 NF 算法 α 值与 F-measure 的关系

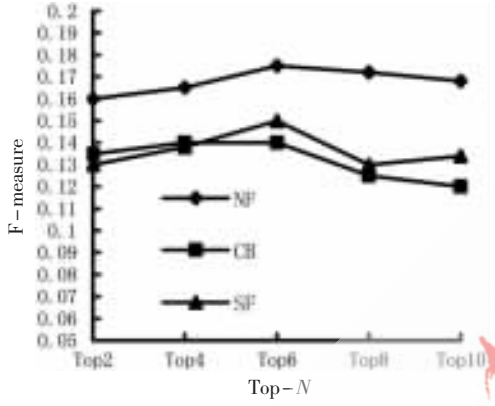

图 2 算法性能对比

由实验结果分析可知，本文提出的结合社会过滤算法和内容推荐算法的个性化好友推荐算法，能够有效地处理社交网络中好友推荐时遇见的冷启动、标签冗余等问题，同时推荐的准确性也有了进一步的提高。

在以后的研究中应更加重视用户在使用社交网络中的动态信息，多考虑用户的兴趣变化，根据用户的兴趣变化实时地给用户推荐好友。

参考文献

[1] GOU L, YOU F, GUO J, et al. Sfviz: interest-based friends exploration and recommendation in social networks[C]. In Proceedings of the 2011 Visual Information Communication-International Symposium, ACM, 2011.

[2] SAVAGE S, BARANSKI M, CHAVEZ N E, et al. I'm feeling loco: a location based context aware recommendation system[C]. In Advances in Location-Based Services: 8th International Symposium on Location-Based Services, Vienna, 2011.

[3] DIMICCO J, MILLEN D, GEYER W, et al. Motivations for social networking at work[C]. ACM CSCW, 2008.

[4] EHRLICH K, LIN C, MILLEN D, et al. Recommending topic for self-descriptions in online user profiles[C]. ACM RecSys, 2008.

[5] GROH G, EHMIG C. Recommendations in taste related domains: collaborative filtering vs. social filtering[C]. Proc. ACM Group'07: 127-136.

[6] LINDEN G, SMITH B, YORK J. Amazon.com recommendations: Item-to-Item collaborative filtering[J]. IEEE Internet Computing, 2003, 7(1): 76-80.

[7] HALPIN H, ROBU V, SHEPHERD H. The complex dynamics of collaborative tagging[C]. In Proc. of WWW'07: 211-220.

[8] ALJANDAL W, BAHIRWANI V, CARAGEA D, et al. Ontology-aware classification and association rule mining for interest and link prediction in social networks[C]. In SSS'09: AAAI Spring Symposia 2006 on Social Semantic Web, 2009.

(收稿日期：2013-03-16)

作者简介：

高永兵，男，1974 年生，副教授，硕士，主要研究方向：Web 信息检索。

表 1 算法性能对比

		SF	CB	NF
Top2	准确率	0.098	0.112	0.124
	召回率	0.193	0.169 8	0.225
	F-measure	0.13	0.135	0.16
Top4	准确率	0.102	0.108	0.117
	召回率	0.213	0.199 8	0.208 9
	F-measure	0.138	0.14	0.165
Top6	准确率	0.11	0.12	0.141
	召回率	0.235 7	0.168	0.230 6
	F-measure	0.15	0.14	0.175
Top8	准确率	0.115	0.13	0.138
	召回率	0.149 5	0.120 37	0.228
	F-measure	0.13	0.125	0.172
Top10	准确率	0.126	0.136	0.142
	召回率	0.143	0.107	0.205 6
	F-measure	0.134	0.12	0.168