基于 SimRank 的用户相似度挖掘技术

1 技术背景

社交网络在近几年获得了空前发展，常见的有 Facebook、Weibo、WeChat、Instagram 等，而用户相似度计算作为社交网络应用中不可或缺的基础研究，是一个具有挑战的研究课题。在对社交网络的数据进行挖掘与分析中，我们常常需要知道用户之间差异的大小，进而评价用户的相似性和类别。常见的应用有：用户相关性分析，连接预测，分类，聚类，推荐系统等。相似度计算已经成为其中的一个重要环节，采用什么样的方法进行相似度计算直接关系到后续应用效果。**如何结合基于内容和基于结构两种相似度度量方式来进行有效的用户相似度计算，从而提高数据分析的质量，是社交网络应用中不得不解决的一个基础问题，但目前还没有一个系统完善的解决方案。**

2 技术方案与创新

利用社交网络数据，以 SimRank 相似度模型为基础，以快速有效计算用户相似度为主要目标，围绕融合图结构与用户的相似度模型、加快模型计算速度、多社区结构下的计算框架 3 个方向，利用统计学、人工智能、机器学习等学科的相关理论和方法，以及相关工作基础和研究成果，将研究内容分为 ConSim 模型的定义与特性、基于图拓扑结构的采样技术、分块计算框架 3 三项子内容，并在此基础上实现科研突破。

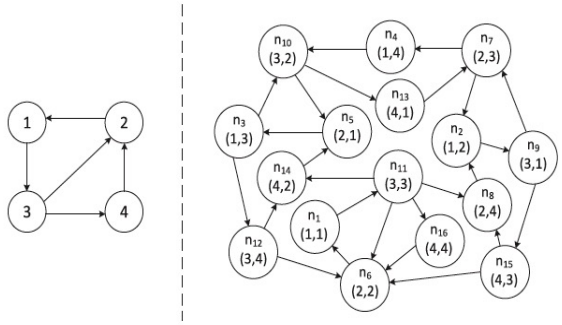

Fig. 1. A simple graph G and its corresponding G^2.

Algorithm 3. Preconditioned Bi-Conjugate Gradient Stabilized Method

$\vec{r}_0 = \vec{q}$
$\vec{r} = \vec{r}_0$
$\rho_0 = \alpha = \omega_0 = 1$
$\vec{v}_0 = \vec{p}_0 = \vec{0}$
for $k = 1, 2, 3, \ldots$ **do**
  $\rho_k = (\vec{r}, \vec{r}_{k-1})$
  $\beta = (\rho_k / \rho_{k-1})(\alpha / \omega_{k-1})$
  $\vec{p}_k = \vec{r}_{k-1} + \beta(\vec{p}_{k-1} - \omega_{k-1}\vec{v}_{k-1})$
  $\vec{y} = \mathbb{M}^{-1}\vec{p}_k$ based on Equation (18)
  $\vec{v}_k = \mathbb{L}\vec{y}$ based on Equation (19)
  $\alpha = \rho_k / (\vec{r}, \vec{v}_k)$
  $\vec{s} = \vec{r}_{k-1} - \alpha\vec{v}_k$
  $\vec{z} = \mathbb{M}^{-1}\vec{s}$ based on Equation (18)
  $\vec{t} = \mathbb{L}\vec{z}$ based on Equation (19)
  $\omega_i = (\vec{t}, \vec{s}) / (\vec{t}, \vec{t})$
  $vec(S_{k+1}) = vec(S_k) + \alpha\vec{y} + \omega_k\vec{z}$
  $\vec{r}_k = \vec{s} - \omega_k\vec{t}$
end for
return $vec(S_{k+1})$

为提高计算速度，在高维度 G2 图上使用 CG 和 Bi-CG 方法搜索最优收敛方向，加快迭代速度，同时可以使用图模型的拓扑结构加快采样精度，最终可以使用并行计算实现分块计算框架。

3 技术创新点

（1）一套简单有效的基于图模型与用户标签的相似度模型

（2）新颖通用的基于图模型的采样技术。

（3）基于多社区结的相似度计算技术

4 应用案例

将上述方法用于社交网络用户相似度计算中，结果显示 CG 和 Bi-CG 方法的计算速度优于其他方法。

TABLE 3
Running Time on Different Datasets

Datasets	NS(BI)PCG	(BI)PCG	OIP	SOR
Blogs	0.2s	34s	9.8s	170s
Hamster	4s	5*s*	7.8s	17.2s
Movielens	7s	7.5s	13s	20s
Facebook	27s	33s	50s	131.1s
Reactome	34.1s	101.2s	95.6s	357s
CaHepph	120s	160s	255s	542s
Google	250s	350s	422s	588s
AstroPh	533s	550s	870s	1,406s
DBLP	31,125s	33,425s	55,520s	80,232s

计算精度与迭代速度如下。

TABLE 6
Time Efficiency and Effectiveness of Our Method and SSJ

Dataset	NS(BI)PCG			(BI)PCG			SSJ		
	Time	AvgDiff	NDCG	Time	AvgDiff	NDCG	Time	AvgDiff	NDCG
Blogs	0.2 s	$2.74E-07$	0.999	34 s	$2.74E-07$	0.999	27.902 s	0.00270	0.94971
Hamster	0.4 s	0.0024547	0.8970	0.5 s	0.0024547	0.8970	3.169 s	0.00311	0.82367
Movielens	1.8 s	0.0024343	0.9988	1.9 s	0.0024343	0.9988	22.742 s	0.01287	0.91432
Facebook	3 s	0.0016255	0.9669	3.6 s	0.0016255	0.9669	32.555 s	0.00263	0.88515
Reactome	3.1 s	0.0010905	0.8962	9.2 s	0.0010905	0.8962	334.781s	0.00172	0.86190
CaHepph	12 s	0.0003374	0.9140	16 s	0.0003374	0.9140	45.786	0.00063	0.83759
Google	25 s	0.0003192	0.9273	35 s	0.0003192	0.9273	22,832.024	0.000853	0.766785
AstroPh	41 s	0.0001179	0.9246	43 s	0.0001179	0.9246	79.317 s	0.000357	0.856212

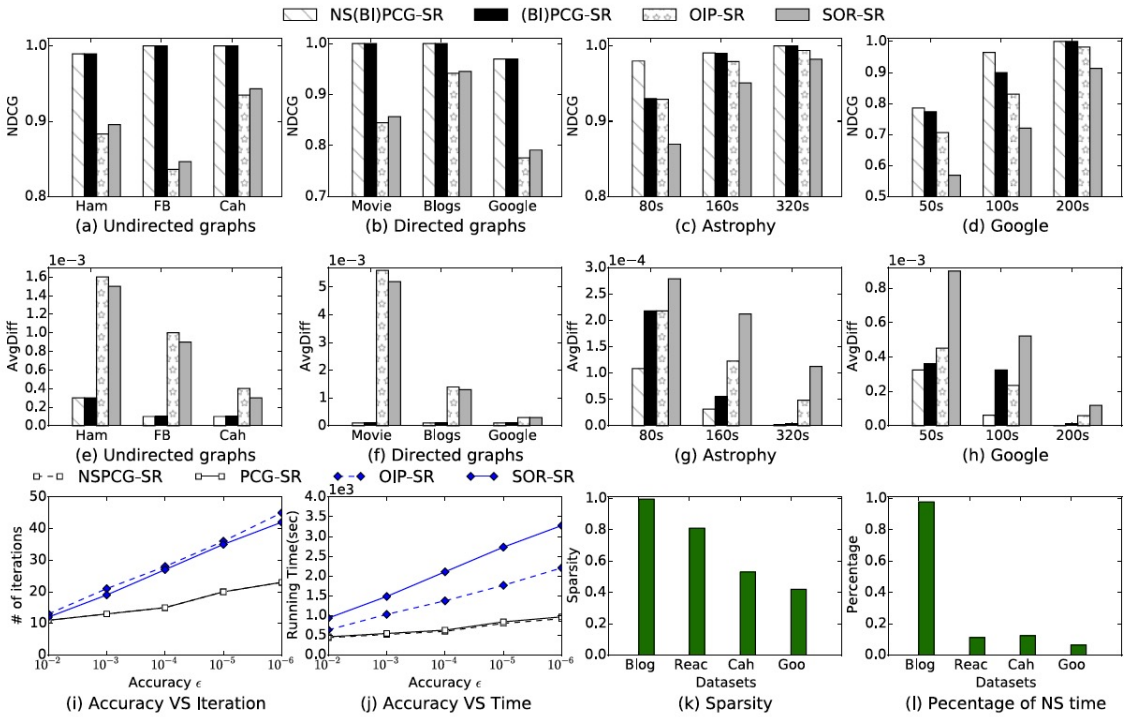

Fig. 2. Performance evaluations on real and synthetic datasets.

5 对接联系

联系人：卢娟 （信息工程学院博士）

邮　箱：lujuan@bipt.edu.cn