数据挖掘实验——社区问答网站专家发现

王晶 2120151040；李凯霞 2120151003；张林 2120151062；韩学博 2120150990

简介

找到社区中某个领域的专家的方式：

- 可以使用 Stack Overflow 中的口碑评分
  无法指定某个领域，比如 java。
  无法手动调节。
- 计算回答数量：
  无法度量答案的重要性
- 创建社交网络并计算用户的中心度
  PageRank，HITS

构建一个图：

- 节点：每个用户作为一个节点
- 边：一个问题的拥有者，指向一个被接收答案的拥有者。

方法概述：

1. 从输入中提取相关领域
2. 选择问题中包含某个领域的关键字，比如 Java
3. 通过找到被接受答案的拥有者创建图
4. 分析图

问题描述

1 实验数据集

来自于 Stack Overflow，下载地址是：

https://archive.org/download/stackexchange/stackoverflow.com-Posts.7z

更新时间：10-Mar-2016 02:36

数据大小：8.3G

文件名称：Posts.xml

2 问题分析

找到某个领域的专家的方式有很多种。使用 Stack Overflow 中的口碑评分，但无法指定某个领域(如 java)，无法手动调节；计算回答数量,无法度量答案的重要性。

我们采用创建社交网络，并计算用户的中心度的方法来发现某个领域的专家，并利用 PageRank 和 HITS 排序算法来评价用户的重要性。创建社交网络时，每个用户作为一个节点,以一个问题的拥有者指向一个被接收答案的拥有者为一条有向边。

3 数据分析

3.1 问题的 XML 格式及提取的属性

<row Id=”4” PostTypeId=”1” OwnerUserId=”8” AcceptedAnswerId=”7” Tags=”<c#><winforms><forms><opacity>” …/>

Field	Value
Id	4
PostTypeId	1
OwnerUserId	8
AcceptedAnswerId	7
Tags	C#, winforms, forms, opacity

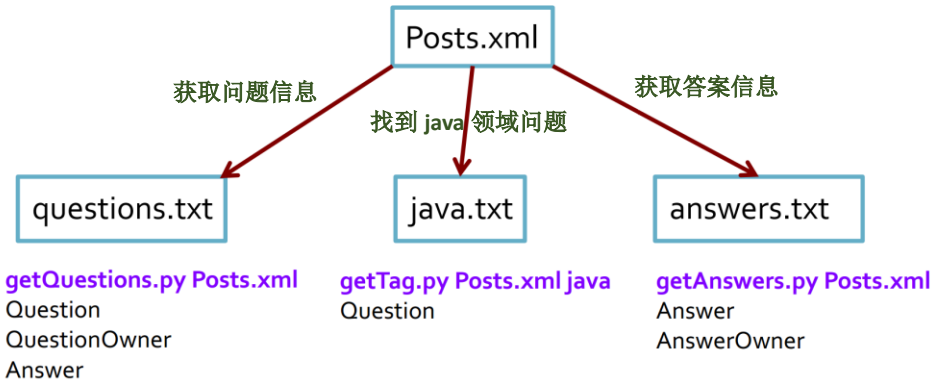
3.2 回答的 XML 格式及提取的属性

<row Id=”12” PostTypeId=”2” OwnerUserId=”1” …/>

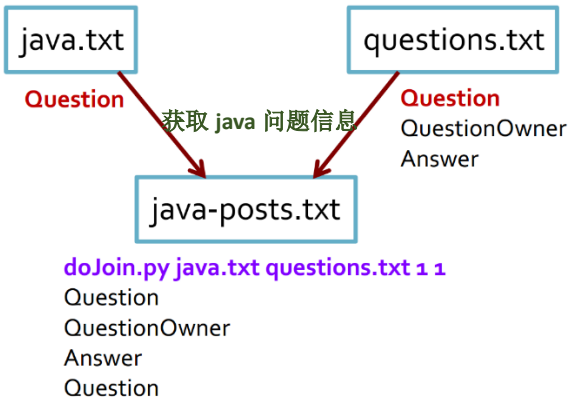
Field	Value
Id	12
PostTypeId	2
OwnerUserId	1

4 实现步骤

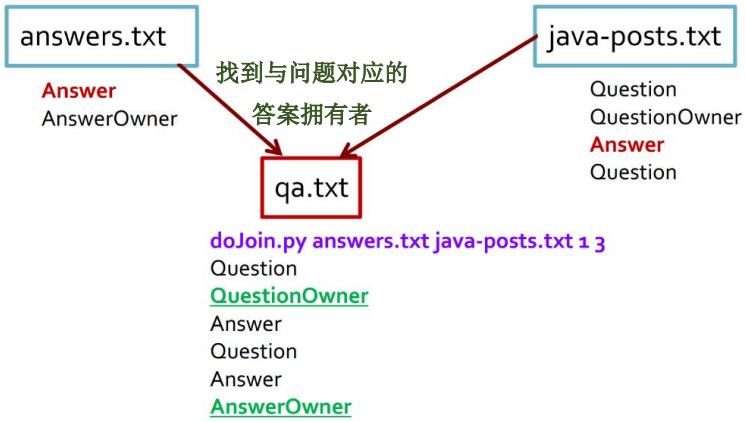
(1)处理输入文件，提取相关文件。获取问题和回答的列表，找到相关领域(如 java)的问题，共生成三个文件，question.txt，answers.txt，java.txt。



(2)选择只和相关领域(如 java)相关的问题：通过 java.txt 和 question.txt 合并出 java-posts.txt。



(3)通过找到被接收答案的拥有者，并创建图：通过 answers.txt 和 java-posts.txt 合并出 qa.txt，其中所存放的节点对就是图的有向边。



(4)分析图。提取网络相关信息，并分别利用 PageRank 和 HITS 算法找到 TOP-k java 专家。

① 网络中有向边和节点信息；
② 网络信息数据摘要；
③ 入读和出度的分布；
④ Top-k 出度用户；
⑤ Top-k 入度用户；
⑥ Pagerank 的 TOP-k 专家；
⑦ HITS 的 TOP-k 专家；
⑧ HITS 的 TOP-k 学习者。

技术方案

4 算法概述

4.1 Pagerank

PageRank 是 Google 专有的算法，用于衡量特定网页相对于搜索引擎索引中的其他网页而言的重要程度。它由 Larry Page 和 Sergey Brin 在 20 世纪 90 年代后期发明。一个页面的“得票数”由所有链向它的页面的重要性来决定，到一个页面的超链接相当于对该页投一票。一个页面的 PageRank 是由所有链向它的页面（“链入页面”）的重要性经过递归算法得到的。

假设一个由 4 个页面组成的小团体：A，B，C 和 D。其中 PR（*）为*的 PageRank 值，L(*)为*的出链数：

$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)}$$

由于存在一些出链为 0，也就是那些不链接任何其他网页的网，也称为孤立网页，使得很多网页能被访问到。因此增加阻尼系数（damping factor）d，一般 d=0.85。其意义是，在任意时刻，用户到达某页面后并继续向后浏览的概率。1-d= 0.15 就是用户停止点击，随机跳到新 URL 的概率。因此，没有页面的 PageRank 会是 0。所以，Google 通过数学系统给了每个页面一个最小值。

$$PR(A) = \left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)}\right)d + \frac{1-d}{N}$$

使用幂法计算，PageRank 公式可以转换为求解 $\lim_{n\to\infty} A^n$ 的值。其中，X 为设置初始每个网页的 PageRank 值。一般为 1。P 为概率转移矩阵。e^t 为 n 维的全 1 行。矩阵 $A =$

$d \times P + (1-d) \times ee^t /N$ 。最终计算收敛得到 R 为各页面的 PageRank 权值。

算法计算步骤如下：

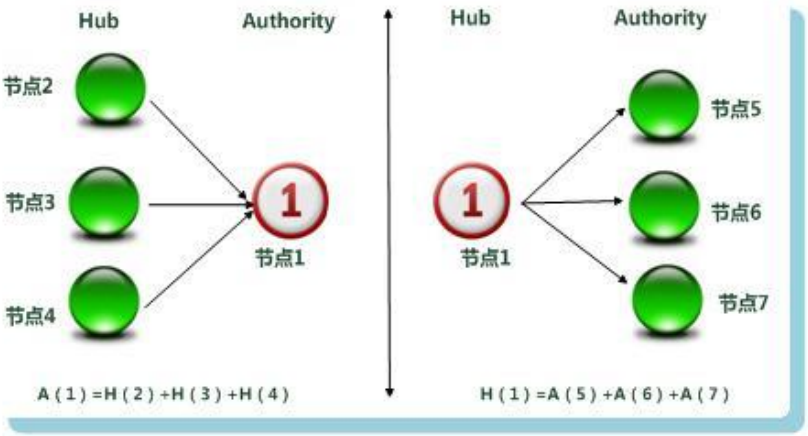
```
R = AX
while(1){
   if(|X-R| < ε ){
      return R;}
   else{
      X = R;
      R = AX;}
}
```

最后，根据 PageRank 权值得分由高到低排序，取权值最高的若干结点为 TOP-k 专家。

4.2 HITS

HITS 算法是由康奈尔大学的 Jon Kleinberg 博士于 1997 年首先提出的,为 IBM 公司阿尔马登研究中心的名为“CLEVER”的研究项目中的一部分。

按照 HITS 算法，用户输入关键词后，算法对返回的匹配页面计算两种值，一种是枢纽值（Hub Scores），另一种是权威值(Authority Scores),这两种值是互相依存、互相影响的。所谓枢纽值，指的是页面上所有导出链接指向页面的权威值之和。权威值是指所有导入链接所在的页面中枢纽之和。



算法计算步骤如下：

a ，h 初始化为 1，$a_0 = 1, h_0 = 1$

t=1

do

for each v in V

do $a_i\ (v) = \sum_{(w,v)\in E} h_{i-1}(w)$

$h_i\ (v) = \sum_{(v,w)\in E} a_{i-1}(w)$

$a_i\ = a_i\ / \|a_i\ \|$

$h_i\ = h_i\ / \|h_i\ \|$

$t = t + 1$

While $\|a_i\ - a_{i-1}\| + \|h_i\ - h_{i-1}\| < \varepsilon$

Return $(a_t\ ,\ h_t\ )$

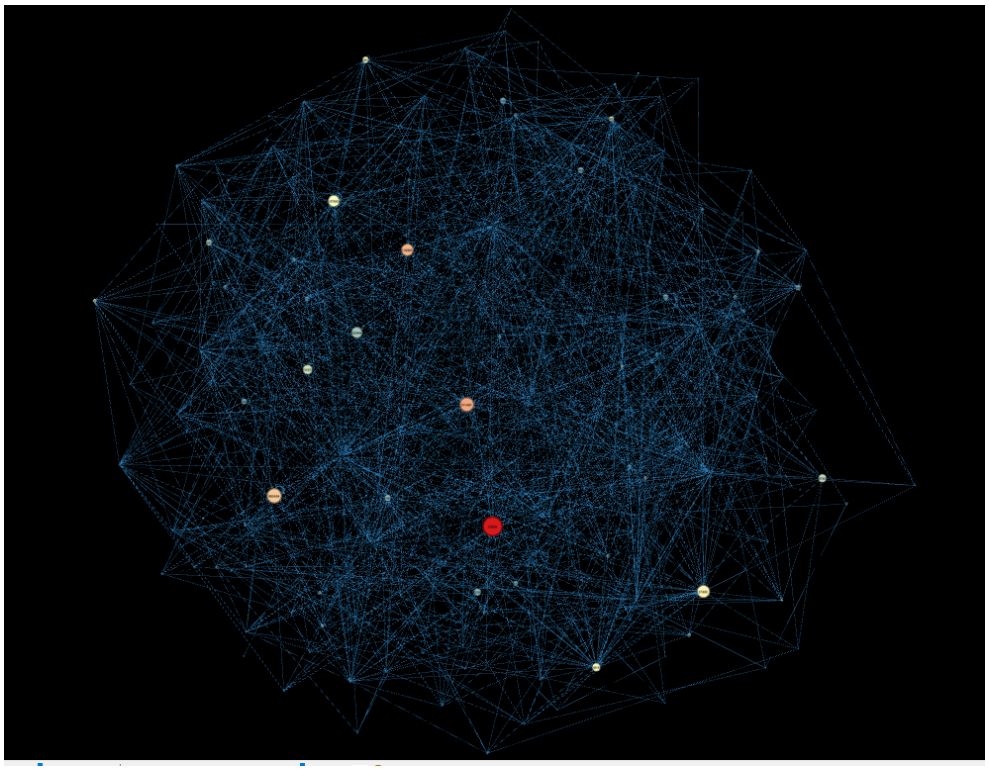
最后，根据 Authority 权值得分由高到低排序，取取权值最高的若干用户为 TOP-k 专家。根据 Hub 权值得分由高到低排序，取取权值最高的若干用户为 TOP-k 学习者。

实现和实验结果

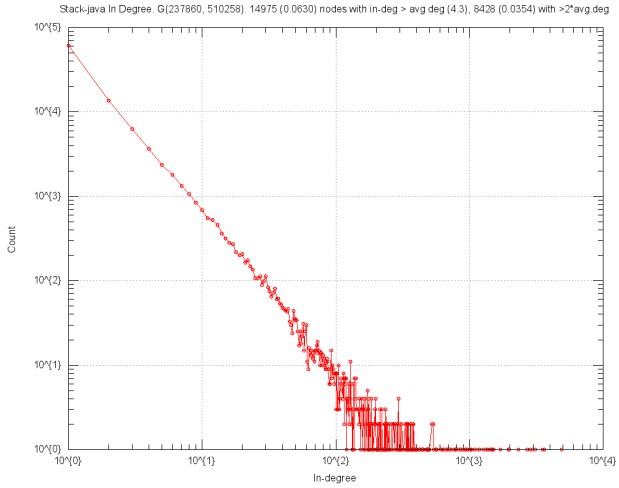
5 实验结果

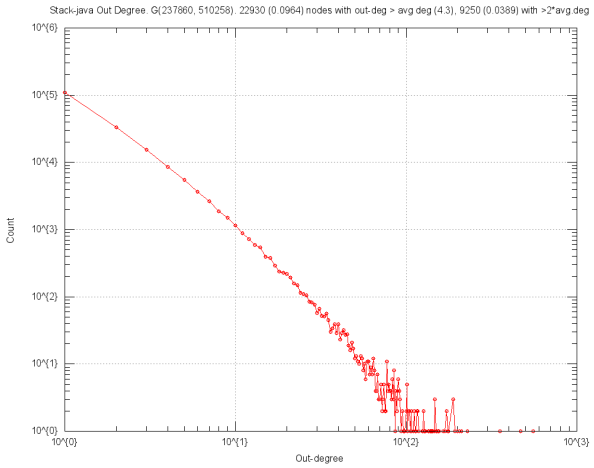
在实验中，我们分别对 java，C#，php 领域的问答建立了社交网络，进行了数据分析和挖掘，得到了各自网络的数据摘要、相关信息，以及 TOP 10 专家和学习者。下面以 java 领域为例说明，其余结果见提交材料。

创建的社交网络：



网络的入度和出度分布：





网络信息数据摘要：

```
QA Stats: Directed
  Nodes:                     237860
  Edges:                     510258
  Zero Deg Nodes:            0
  Zero InDeg Nodes:          137783
  Zero OutDeg Nodes:         47327
  NonZero In-Out Deg Nodes:  52750
  Unique directed edges:     510258
  Closed triangles:          51739
  Open triangles:            90688463
  Frac. of closed triads:    0.000570
  Connected component size:  0.880325
  Strong conn. comp. size:   0.023648
  Approx. full diameter:     11
  90% effective diameter:    5.617292
```

```
graph nodes 237860, edges 510258

# of connected component sizes 8
size 1, number of components 9768
size 2, number of components 7798
size 3, number of components 778
size 4, number of components 128
size 5, number of components 36
size 6, number of components 8
size 7, number of components 4
size 209394, number of components 1

max wcc nodes 209394, edges 489796
```

对比各算法得到的 Top 10 java 专家：

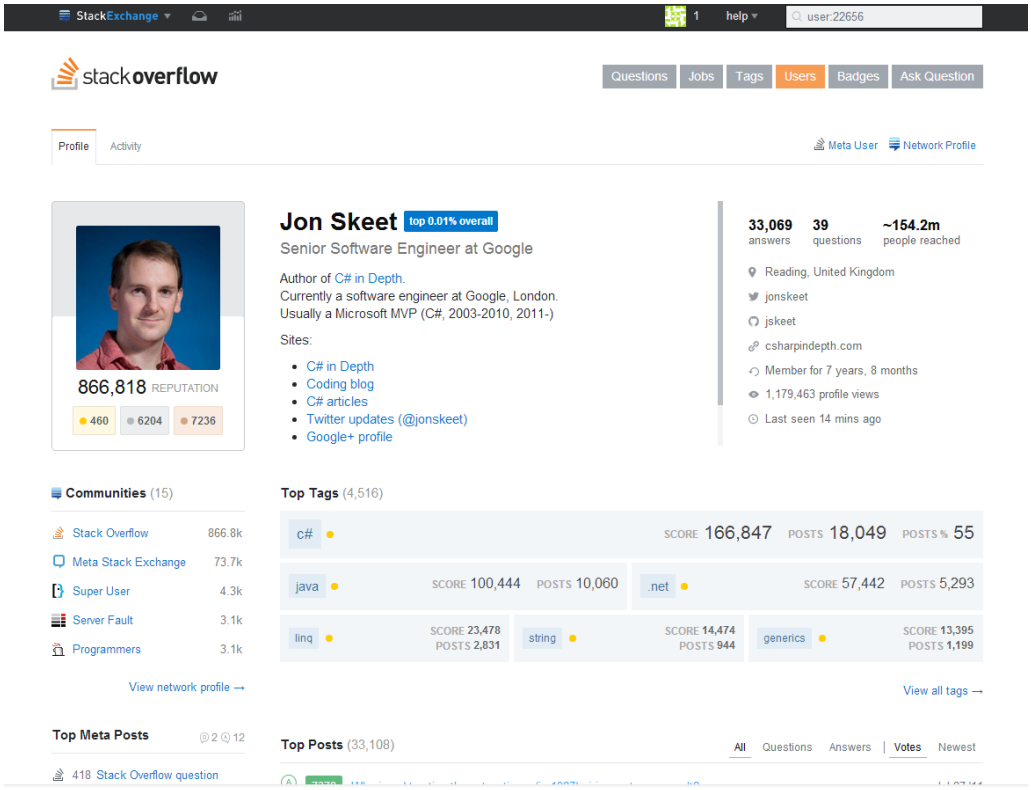
入度	PageRank	HITS（Authority）
22656	22656	22656
992484	4125191	571407
571407	571407	57695
57695	139985	139985
139985	992484	157882
157882	57695	992484
522444	4856258	203907
131872	157882	522444
438154	438154	131872
207421	5406012	438154

对比各算法得到的 Top 10 java 学习者：

出度	HITS（Hub）
1194415	892029
892029	1194415
785349	470184
470184	648138
454049	359862
853836	802050
2674303	384706
1833945	225899
359862	454049
44330	431769

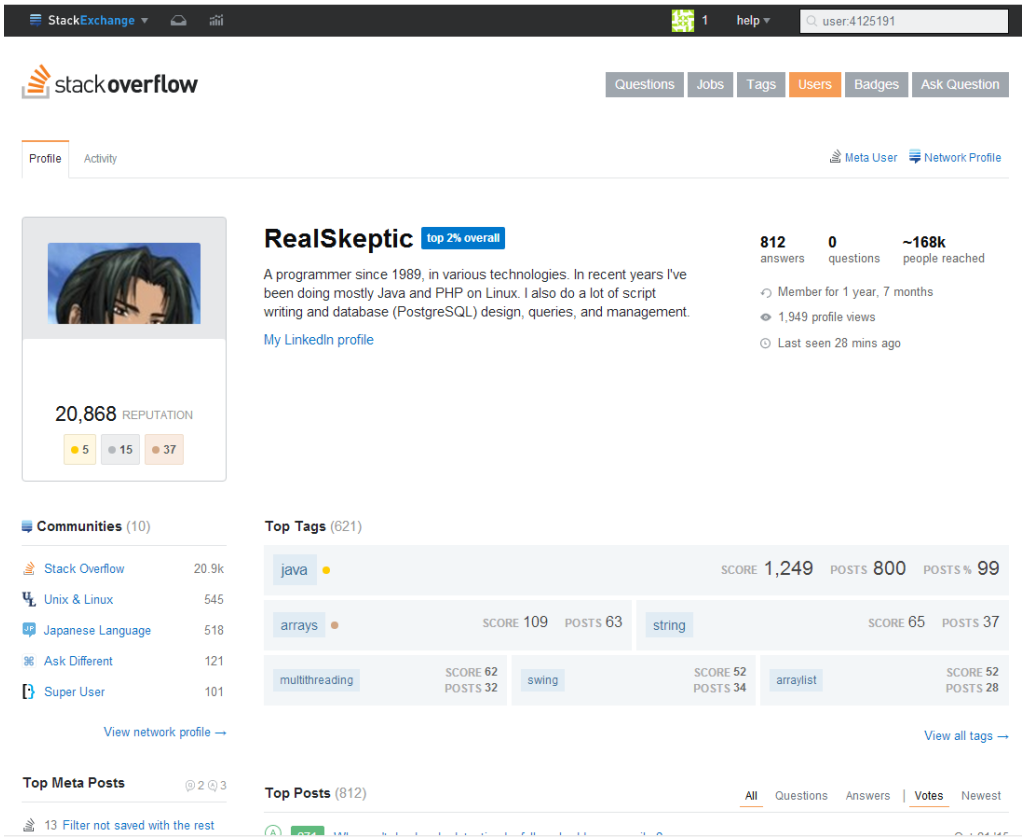
专家用户的网页：

TOP-1 专家 http://stackoverflow.com/users/22656

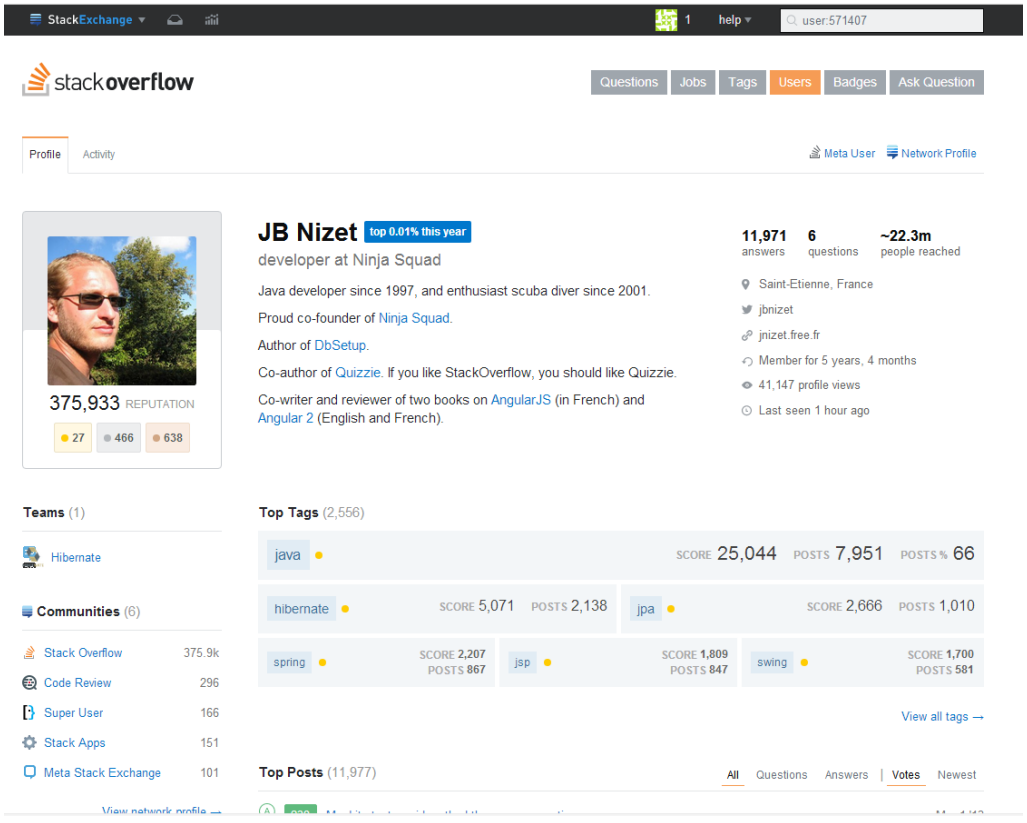


TOP-2 PageRank 专家 http://stackoverflow.com/users/4125191

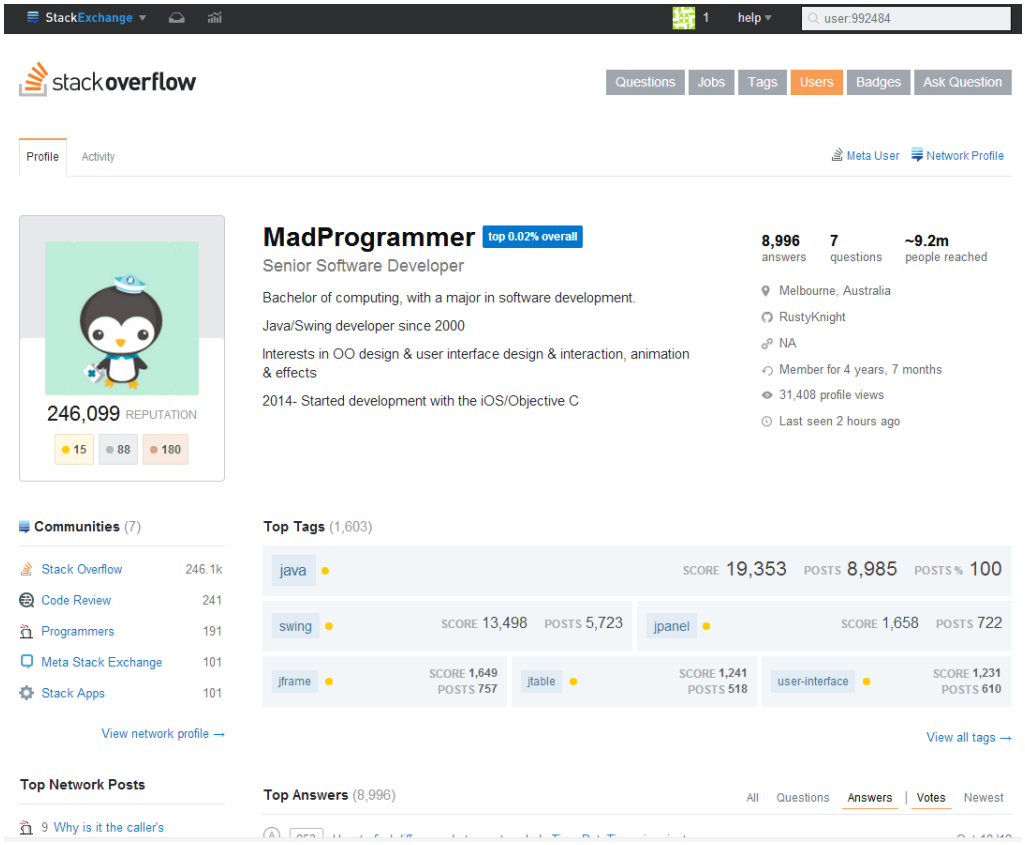
对 Top-2 专家得分较低但排名很高的解释，a. 该专家注册时间比较晚，所以得分较少，b. 由于该专家回答了其他专家提出的问题，产生一条由其他专家指向他的边，使得他的 PR 值增大，排名固然提升。



TOP-2 HITS 专家 http://stackoverflow.com/users/571407



TOP-2 入度用户 http://stackoverflow.com/users/992484


StackExchange
1
help
user:992484
stackoverflow
Questions
Jobs
Tags
Users
Badges
Ask Question
Profile
Activity
Meta User
Network Profile
246,099 REPUTATION
15
88
180
MadProgrammer
top 0.02% overall
Senior Software Developer
Bachelor of computing, with a major in software development.
Java/Swing developer since 2000
Interests in OO design & user interface design & interaction, animation & effects
2014- Started development with the iOS/Objective C
8,996 answers
7 questions
~9.2m people reached
Melbourne, Australia
RustyKnight
NA
Member for 4 years, 7 months
31,408 profile views
Last seen 2 hours ago
Communities (7)
Stack Overflow 246.1k
Code Review 241
Programmers 191
Meta Stack Exchange 101
Stack Apps 101
View network profile →
Top Tags (1,603)
java SCORE 19,353 POSTS 8,985 POSTS % 100
swing SCORE 13,498 POSTS 5,723
jpanel SCORE 1,658 POSTS 722
jframe SCORE 1,649 POSTS 757
jtable SCORE 1,241 POSTS 518
user-interface SCORE 1,231 POSTS 610
View all tags →
Top Network Posts
9 Why is it the caller's
Top Answers (8,996)
All
Questions
Answers
Votes
Newest