

基于社会网络的人名检索结果重名消解

郎 君 秦 兵 宋 巍 刘 龙 刘 挺 李 生

(哈尔滨工业大学计算机科学与技术学院信息检索研究室 哈尔滨 150001)

摘 要 人物重名现象十分普遍,搜索引擎的人名检索结果通常是多个同名人物相关网页的混合.该文依据同名的不同人物具有不同的社会网络的思想,利用检索结果中共现的人名发现并拓展检索人物相关的潜在社会网络,结合图的谱分割算法和模块度指标进行社会网络的自动聚类,在此基础上实现人名检索结果的重名消解.在人工标注的中文人名语料上进行实验,整体性能达到较好水平,图聚类算法能帮助连通社会网络的进一步划分,从而提高消解效果.

关键词 社会网络;重名消解;谱分割;模块度

中图法分类号 TP18 **DOI号**: 10.3724/SP.J.1016.2009.01365

Person Name Disambiguation of Searching Results Using Social Network

LANG Jun QIN Bing SONG Wei LIU Long LIU Ting LI Sheng

(Information Retrieval Laboratory, School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001)

Abstract The person names are so ambiguous that the results for searching a person name are usually a mixture of pages about the namesakes. This paper presents a novel approach leveraging the fact that each namesake has a unique social community. Firstly, the social network of the person name to search is found and extended by employing the co-occurrence of person names in snippets returned by a search engine, then automatically clustered into different social communities by the algorithm combining spectral partition and modularity evaluation. Finally, the search results are clustered into different groups where each contains pages referring to the same individual. On the corpus of Chinese person names, experimental results show that the whole performance achieves high level and graph clustering algorithm benefits improving disambiguation effect from further dividing the connecting social network.

Keywords social network; name disambiguation; spectral partition; modularity

1 引 言

利用搜索引擎检索人物信息是互联网用户的主要活动之一.然而现实世界中,多个人物共享一个人名是很普遍的现象.这导致搜索引擎对某一特定人

名的检索结果往往是共享这一人名的不同人物相关网页的混合.例如,百度检索“王刚”返回的前10个结果中就有“国家著名演员”、“中央政治局委员”、“西北工业大学副教授”、“山东黄金篮球队队员”、“建筑师”、“中国作家协会会员”等6位不同的人物.重名消解就是根据上下文或篇章信息来区分同一人

收稿日期:2007-09-17;最终修改稿收到日期:2009-04-10.本课题得到国家自然科学基金(60675034,60803093)和国家“八六三”高技术研究发展计划探索类专题项目(2008AA01Z144)资助.郎君,男,1981年生,博士研究生,研究方向为信息抽取、共指消解、机器学习. E-mail: bill_lang@ir.hit.edu.cn.秦兵,女,1968年生,博士,教授,主要研究领域为自然语言处理、信息检索.宋巍,男,1983年生,博士研究生,主要研究方向为自然语言处理、信息检索.刘龙,男,1985年生,硕士研究生,主要研究方向为自然语言处理.刘挺,男,1972年生,博士,教授,博士生导师,主要研究领域为自然语言处理、信息检索.李生,男,1943年生,博士,教授,博士生导师,主要研究领域为自然语言处理、机器翻译.

名表示的不同人物的过程。

虽然现在有些系统能对检索结果进行聚类处理,例如 iBoogie^①、SnakeT^②、Vivisimo^③、Apex 搜索^④、Bbmao^⑤等,但它们都把人名当成普通词汇进行处理,聚类结果的标签也是这个人名相关的一些词汇,没有对人名的重名结果进行区分。现在尚未完全公开的 Spock^⑥系统检索人名时能够找出重名的不同人物,但这个系统不是针对搜索引擎检索结果进行处理,而是通过不同途径抓取并索引了超过一亿人的相关资料,在检索时根据同名人物的个人信息来实现重名消解。同时,这个系统只能检索英文人名,不支持中文人名。

人名检索结果的重名消解,可以采用类似于自然语言处理中词义消歧的方法,利用人名的上下文信息来实现。常见的方法将人名检索结果对应的 Snippet 或者网页内容采用向量空间模型表示^[1-2],或者抽取上下文中的关键性短语^[3],然后采用计算向量相似度的方法来实现最终的检索结果聚类。更为深入的方法是在网页中抽取人物的相关信息,例如性别、民族、籍贯、出生年月、家庭关系、住址、职务等,然后在人物属性集上计算人物的相似度,从而实现人物同一性判别^[4]。

对于检索结果的重名消解,基于文本聚类的方法考虑了很多无用词汇,而且需要人工设定阈值或者类别数量;基于信息抽取的人物相关属性相似度的方法对于人物信息的抽取具有很大的依赖性,各种属性在抽取时的错误容易导致错误级联。针对这

些问题,本文提出基于社会网络的人名检索结果重名消解方法。主要依据是“物以类聚,人以群分”,即重名的不同人物所属的社会网络具有区分性^[5-6]。例如演艺圈的“王刚”和政界的“王刚”在社会网络上明显不同的。本文利用检索结果背后潜在的社会网络关系来解决检索结果重名问题。与文献[5-6]不同的是,本文针对中文人名进行处理,并采用了不同的社会网络构建方法;而且,本文结合谱分割和模块度的方法能自动确定最优的类别数量。

本文第2节介绍整体系统框架及每步采用的方法;第3节说明语料准备以及标注情况;第4节是评价方法;第5节展示实验设计、实验结果及分析等;最后是结论以及展望。

2 系统框架及方法

整个系统处理流程主要分为3步:(1)社会关系获取。在检索结果中抽取全部人名来构建同一人名的初始社会网络,并对其中各个子图进行拓展,使得有所关联的离散的社会网络连接起来。(2)社会网络聚类。在拓展后的社会网络图上采用图分割的算法来实现自顶向下的图聚类算法,结合模块度的评价指标自动得到局部社会圈子。(3)社会网络映射。将形成的各个局部社会圈子映射回到最早检索人名的 Snippet 上,从而实现检索结果的重名消解。系统框架如图1所示。

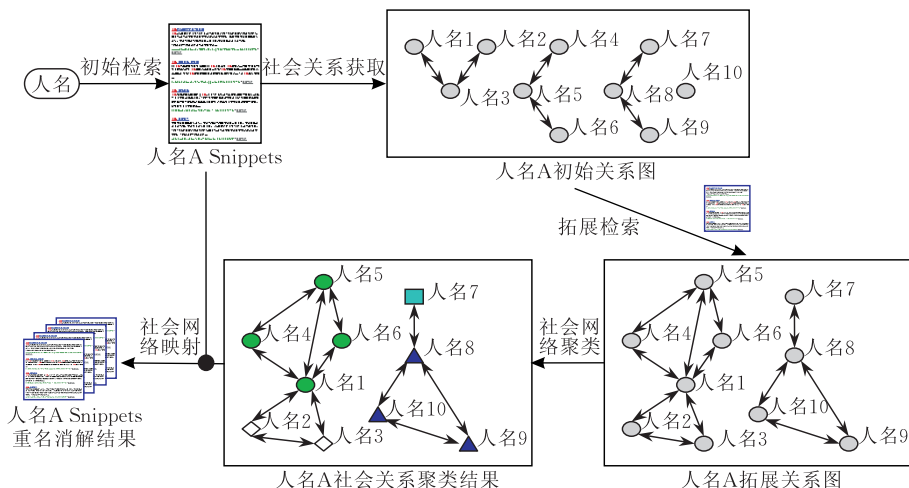


图1 重名消解系统框架

2.1 基于人名检索结果的社会网络构建

利用社会网络进行相关处理的前提是构建一个合理的社会网络。本文利用待检索的中文人名在搜索引擎上返回的 Snippet 进行社会网络构建。这里

① <http://www.iboogie.com/>
② <http://snaket.di.unipi.it/>
③ <http://earch.vivisimo.com/>
④ <http://apex.sjtu.edu.cn:50183/>
⑤ <http://www.bbmao.com/>
⑥ <http://www.spock.com/>

的 Snippet 包括检索结果的标题以及紧随的片断文本. 社会关系建立在至少两个人物的基础上, 所以本文定义有效 Snippet 为包含至少两个不同人名的 Snippet. 系统最后的聚类对象就是这些有效的 Snippet.

以检索人名 A 为例, 初始检索返回一组 Snippet, 抽取每个 Snippet 中的人名. 假设任何两个人名共同出现在某个 Snippet 中就认为两人具有社会关系, 共现的次数作为这种关系的度量. 从而可以对出现在所有 Snippet 中的人名构建关系矩阵 M , 矩阵元素 $M_{i,j}$ 表示人名 i 和人名 j 的共现次数. 由于是利用人名 A 的社会网络来对人名 A 检索得到的有效 Snippet 进行重名消解, 关系矩阵 M 中不包含人名 A .

限于检索一个人物获得的有效 Snippet 数量有限, 这样得到的关系矩阵往往会比较稀疏, 形成的社会网络图中有很多的孤立子图, 事实上有些子图之间在真实的网络环境中又是有关联的. 例如图 1 中的人名 A 初始关系图. 本文希望借助更多的网络信息, 对孤立子图进一步扩展, 来丰富初始的社会关系网络.

拓展方法是在初始关系图中找出所有连通子图, 然后依次在每个子图中选取最能够代表该子图的节点来进行拓展检索. 本文引入带权重 (Weighted degree) 来衡量扩展节点的重要程度. 带权重即为与该节点相连接的所有边的权值之和. 这是基于以下两种假设:

(1) 与节点相连的边越多, 说明该节点在这个网络中交际的范围越广, 影响力越大.

(2) 边上的权值越大, 说明该节点与相连节点共现的频率越大, 二者的关系越紧密.

利用带权重将以上两点结合起来.

本文采用两种不同的拓展方式: (1) 单点拓展: 选取子图带权重最大的一个节点; (2) 两点拓展: 选取子图中带权重最大的两个节点. 假设子图 X 中带权重最大的节点名为人名 B . 为了拓展出来的人物尽量都和初始检索的人名 A 有关, 每次拓展检索时 Query 都包含人名 A , 例如对子图 X 扩展时, 检索 Query 为“人名 B 人名 A ”. 拓展检索时, 选取除人名 A 和人名 B 外至少包含一个人名的 Snippet. 将拓展得到的所有 Snippet 直接加入到初始检索到的 Snippet 集合中, 采用构建关系矩阵 M 的方法重新构建新的包含更多人名的关系矩阵 M' . 显然, M' 比 M 包含更多的人名和社会关系, 使得 M 的社会

关系网络进一步丰富与完善.

对于初始社会网络的拓展有如下两种处理方法:

(1) 平均拓展. 矩阵 M' 中会引入很多初始检索中不包含的人名, 剔除这些新引入的人名得到的矩阵为 M'' . 在 M'' 中, 如果两个人物不认识 (对应关系数为 0), 但同时 M' 中有很多人同时认识他们, 则可以利用两个人物之间的中间人来求取两个人物的关系数. 平均拓展采用 M' 中两个人物的中间人的关系数平均值来进行更新. 例如, M'' 中, 对于任意两个人名 $a, b (a \neq b)$, 如果 $M''_{a,b} = 0$, 但是 M' 中存在人名 n_1, n_2, \dots, n_m , 同时满足 $M'_{a,n_i} \neq 0$ 且 $M'_{b,n_i} \neq 0 (i = 1, 2, \dots, m)$, 则更新 $M''_{a,b}$ 为

$$M''_{a,b} = \frac{\sum_{i=1}^m (M'_{a,n_i} + M'_{b,n_i})}{2m} \quad (1)$$

这样更新得到的新矩阵 M'' 将拓展 M 中人名之间的关系, 并且将原来没有直接相邻的节点之间的关系数进行更新, 可将初始图中不连接的若干子图连接起来.

(2) 最大拓展. 考虑现实世界中的两个人物, 如果有一位中间人与他们的关系都非常密切, 这两个人之间的关系就应该很密切; 如果此时还有一位和这两个人虽然认识但是关系很不密切的中间人, 也不应该使得这两个人之间的关系数减少. 事实上, 方法 1 中取平均的方法就可能存在这样的问题. 这里利用两个人物之间关系最为密切的中间人来进行关系数更新. 更新方法类似于方法 1, 只是更新公式变为 (2).

$$M''_{a,b} = \max_{i=1,2,\dots,m} \frac{M'_{a,n_i} + M'_{b,n_i}}{2} \quad (2)$$

2.2 社会网络上的图聚类方法

纷繁复杂的社会网络本质上是一种紧致的小世界网络, 著名的六度理论说明了人际关系的社会网络图的直径很小^[7]. 社会网络一般由若干个“社团” (community) 构成, 每个社团内部的节点之间的链接相对非常紧密, 但各个社团之间的链接相对来说却比较稀疏. 上面得到的关系矩阵, 实际上就是初始检索时全部重名人物的社会关系混在一起的网络结构. 因此需要采用相关方法来将这些不同的社团区分开, 这里选用社会网络中的谱分割算法和模块度评价指标来自动确定不同的社团.

2.2.1 谱分割算法

谱分割 (spectral partition) 算法是一种复杂网络中社团结构发现的算法, 主要基于 Laplace 图特征值来进行图的划分^[7-8].

$G=(V,E)$ 是一个无向无环图, 其中 $V=\{v_1, v_2, \dots, v_n\}$ 为点集, $E=\{e_1, e_2, \dots, e_m\}$ 为边集, 边权函数为 $w(e)$, $e \in E$. $A(G)=(a_{i,j})_{n \times n}$, 其中

$$a_{i,j} = \begin{cases} w(v_i, v_j), & (v_i, v_j) \in E, \\ 0, & (v_i, v_j) \notin E \end{cases} \quad (3)$$

$D(G) = \text{diag}(d_1, d_2, \dots, d_n)$, 其中 $d_i = \sum_{k=1, k \neq i}^n a_{i,k}$. G 的 Laplace 矩阵为

$$L(G) = D(G) - A(G) \quad (4)$$

Laplace 矩阵的特征值最小为 0. 如果图 G 是连通图, 那么 Laplace 矩阵的所有特征值均大于 0; 若图 G 不连通, 则 Laplace 矩阵的特征值中 0 的个数即为图 G 的连通分支数, 0 特征值对应的特征向量中的元素非 0 值对应的点构成 G 的一个连通子图.

Laplace 矩阵的次小特征值 λ_2 表示图 G 的代数连接度. 如果 G 连通, 可以将 λ_2 对应的特征向量 x_2 中元素值大于 0 的对应点和数值不大于 0 的对应点划分开; 如果图 G 不连通, 可以根据任一 0 特征值对应的特征向量中非 0 值的情况将全部点划分开. 这样图 G 就被分解成两个子图, 这种划分使得被切掉边的权值总和达到最小.

用于社团网络聚类的方法中谱分割算法的复杂度相对较低, 一般情况下, 计算一个 $n \times n$ 矩阵的全部特征向量的算法复杂度为 $O(n^3)$. 但是这种方法的缺陷就是它每次只能将网络二分, 如果需要一个网络分成两个以上的社团, 就必须对子社团多次重复该算法. 每次如何选取子社团进行二分, 以及将整个网络分成几个子社团才算最好? 这是谱分割算法在进行多类划分时存在的问题. 为此, 本文引入模块度来解决这个问题.

2.2.2 模块度

Newman 等人提出模块度 (Modularity) 评价指标来衡量网络划分的质量^[9]. 考虑某种划分形式, 将网络划分成 k 个社团. 定义一个 $k \times k$ 的对称矩阵 $E=(e_{i,j})$, 其中元素 $e_{i,j}$ 表示网络中连接两个不同社团的节点的边在初始网络中所有边中所占的比例, 这两个节点分别位于第 i, j 个社团.

设矩阵中对角线上各元素之和为 $Tr e = \sum_i e_{i,i}$, 表示网络中连接某一社团内部各节点的边在所有边数目中所占的比例. 定义每行 (或者列) 中各元素之和为 $a_i = \sum_j e_{i,j}$, 表示与第 i 个社团中的节点相连的边在所有边中所占的比例. 在此基础上, 模块度 Q 定义为

$$Q = \sum_i (e_{i,i} - a_i^2) = Tr e - \|E^2\| \quad (5)$$

其中 $\|x\|$ 表示矩阵 x 中所有元素之和. 上式的物理意义是: 网络中连接两个同种类型的节点的边 (即社团内部边) 的比例, 减去在同样的社团结构下任意连接两个节点的边的比例的期望值. 如果社团内部边的比例不大于任意连接时的期望权值, 则有 $Q=0$. Q 的上限为 1, 而 Q 越接近 1, 说明社团结构越明显, 即图划分的效果越好. 实际网络中, 该值通常位于 0.3~0.7 之间.

2.2.3 基于谱分割和模块度的图聚类算法

基于上面介绍的谱分割和模块度, 可以采用贪心方法来实现图的自动确定类别数量的聚类. 具体方法是: 将某一步谱分割得到的全部子图中选定一个子图进行试探性二分, 同时其它子图固定不变, 计算试探性二分得到的模块度. 这样轮流将全部子图进行试探性二分, 选取模块度增加最大或者减少最小的一个试探性二分结果来作为下一步的谱分割结果. 重复这个过程直到整个网络不能再被划分为止, 记录其中每一步的分割方式以及得到的模块度, 然后选取模块度最大的一种分割方式来将网络进行划分. 算法描述如下, 其中第 5 行表示采用 2.2.1 节谱分割算法将子图 V_i 分割为两个不交叉子集 $V_{i,1}$ 和 $V_{i,2}$.

算法 1. 基于谱分割和模块度的图聚类算法.

Input: $G=(V,E)$

Output: The best partition of V , $P_{\text{best}} = \{V_1, V_2, \dots, V_n\}$, which satisfies $\forall i, j \in \{1, \dots, n\}, V_i \cap V_j = \emptyset$, and $V_1 \cup V_2 \cup \dots \cup V_n = V$

1. $P_{\text{best}} \leftarrow \emptyset, P_{\text{current}} \leftarrow \{V\}, M_{\text{best}} \leftarrow 0, M_{\text{current}} \leftarrow 0$
2. while $|P_{\text{current}}| < |V|$
3. $P_{\text{tempbest}} \leftarrow \emptyset, M_{\text{tempbest}} \leftarrow 0, M_{\text{differ}} \leftarrow -\infty$
4. for each V_i in P_{current}
5. $\{V_{i,1}, V_{i,2}\} \leftarrow \text{Spectral-Partition}(G, V_i)$
6. $P_{\text{temp}} \leftarrow \{V_{i,1}, V_{i,2}\} \cup P_{\text{current}} \setminus \{V_i\},$
 $M_{\text{temp}} \leftarrow \text{Modularity}(G, P_{\text{temp}})$
7. if $(M_{\text{temp}} - M_{\text{current}}) > M_{\text{differ}}$ then
8. $P_{\text{tempbest}} \leftarrow P_{\text{temp}}, M_{\text{tempbest}} \leftarrow M_{\text{temp}},$
 $M_{\text{differ}} \leftarrow M_{\text{temp}} - M_{\text{current}}$
9. endif
10. endfor
11. $P_{\text{current}} \leftarrow P_{\text{tempbest}}, M_{\text{current}} \leftarrow M_{\text{tempbest}}$
12. if $M_{\text{current}} > M_{\text{best}}$ then
13. $M_{\text{best}} \leftarrow M_{\text{current}}, P_{\text{best}} \leftarrow P_{\text{current}}$
14. endif
15. endwhile
16. return P_{best}

对于 2.1 节中构建好的社会网络, 可以在每个

连通子图上应用上述算法得到每个子图的最佳划分,然后将全部的最佳划分合并起来形成最终的社会网络的最佳划分。

2.3 社会网络聚类结果映射回初始检索 Snippet

整个系统的目标是将初始检索的 Snippet 集合划分成互不相交的子集,每个子集中检索人名对应的都是现实世界中的同一人物。2.2 节中生成的社会网络的划分结果还需要映射到初始 Snippet 中。采用的方法是依次查看每个 Snippet,将其中的人名列表分配到 2.2 节得到的人名社团集合,分配最多的人名社团作为当前 Snippet 所属的类别,如果有多个分配最多的人名社团,就随机选择其中一个,从而实现了对初始 Snippet 的划分。

3 评价方法

人名检索结果的重名消解本质上就是跨文档共指消解。本文采用共指消解领域常用的 B-Cubed 评价方法^[10]。描述如下:

假设 Snippet 集合 S 的大小为 N ,其人工划分结果为 Key ,系统输出的划分结果为 $Response$,某个 Snippet s 对应的检索人名在现实世界中的人物记为 P_s 。对于每个 Snippet s , $Key(s)$ 表示人工划分中包含 s 的子集, $Response(s)$ 表示系统输出划分中包含 s 的子集。Snippet s_i 的精确率 (precision) 和召回率 (recall) 分别定义为

$$Precision_{s_i} = \frac{|Key(s_i) \cap Response(s_i)|}{Response(s_i)} \tag{6}$$

$$Recall_{s_i} = \frac{|Key(s_i) \cap Response(s_i)|}{Key(s_i)} \tag{7}$$

整体系统输出划分的精确率和召回率分别为

$$Precision = \frac{1}{N} \sum_{i=1}^N Precision_{s_i} \tag{8}$$

$$Recall = \frac{1}{N} \sum_{i=1}^N Recall_{s_i} \tag{9}$$

采用 $F-measure$ 来评价系统的整体性能,计算公式如下。

$$F-measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{10}$$

4 语料准备及实验结果上下界分析

4.1 语料准备情况

为了评价上述方法,本文选取中国人名中重名现象较为突出的 4 个人名的检索结果进行标注。4 个人名分别是“王刚”、“张伟”、“王伟”、“刘波”。根据

相关报道,4 个人名在中国的重名覆盖人数分别是:“王刚”15 万,“王伟”28 万,“张伟”29 万^①,“刘波”130 万^②。

对于每个人名,本文选取 Baidu 返回的前 100 个有效 Snippet 构成语料,然后人工查阅各个 Snippet 相关的网页进行人物划分。

对于语料标注,不同人之间会存在差异,为了后续分析算法性能可能达到的上界 (Upper bound),我们找了两位经常上网浏览新闻的大学生分别标注四份语料^③。语料标注情况如表 1 所示。

表 1 4 个人名语料的两组人工标注类别数

人名	标注 1	标注 2
王刚	11	9
张伟	65	61
王伟	44	41
刘波	23	28

4.2 上下界分析

假设采用程序来实现同一人名检索得到的 Snippet 的划分的性能上界 (Upper bound) 就是两组标注的一致性水平。对于两组标注,本文定义一致性水平为将一人标注结果作为 Key ,另一人标注结果作为 $Response$,得到的 $F-measure$ 。从式 (6) ~ (10) 可知交换 Key 和 $Response$ 得到的 $F-measure$ 也是一样的。具体数据如表 2 所示。

表 2 系统性能的 Upper bound 分析

人名	Upper bound
王刚	0.9603
张伟	0.9230
王伟	0.9643
刘波	0.9696
平均	0.9543

从表 2 显示的平均 Upper bound 显示,两人标注的结果具有很高的-致性。这说明不同人对同一人名检索得到的网页对应的人物进行区分的能力是相似的。为此,本文在后续的实验中都采用标注 1 进行评价。

对于性能下界 (Lower bound),可以默认两种 baseline 结果:将待划分 Snippet 集合不作任何划分,或者划分成每个 Snippet 单独形成一个类别。两

① 参见文章:《中国重名最多 50 姓名公布“张伟”居首近 30 万人》(根据中国公安部全国公民身份号码查询服务中心公布的数据获得),<http://news.cctv.com/society/20070725/105888.shtml>

② 参见文章:《中国十大最俗名字》,<http://www.dianping.com/group/ent/topic/52704?pageno=1>

③ 语料下载及系统演示网址:<http://ir.hit.edu.cn/demo/findyou>

种划分的 F -measure 中较大的一个可以作为最终的 Lower bound. 具体分析数据如表 3 所示.

表 3 系统性能的 Lower bound 分析

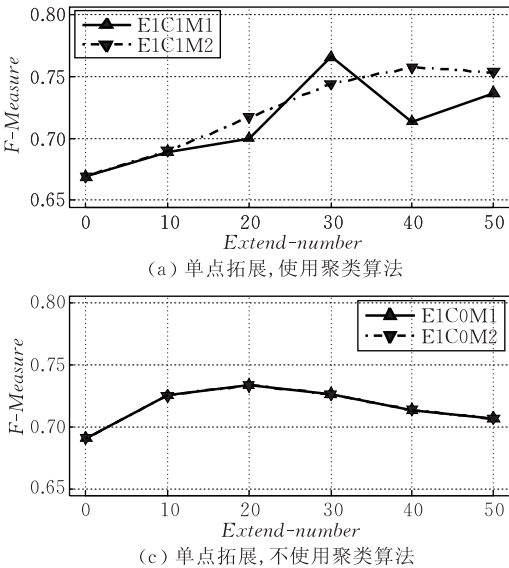
人名	类数	不划分	完全划分
王刚	11	0.6609	0.1981
张伟	65	0.1529	0.7878
王伟	44	0.3609	0.6111
刘波	23	0.2623	0.3770
平均值	35.75	0.3593	0.4935

从表 3 可知,人工标注的类数越多完全划分时的 F 值越高,例如“王刚”;类数越少不划分的 F 值越高,例如“王伟”;当类别数处于中间水平时,两种 F 值均偏低,例如“刘波”. 从表 3 最后一行得知,系统的平均 Lower bound 值为 0.4935.

5 实验部分

5.1 实验设计

系统选用 Baidu 搜索引擎,并采用语言技术平台(LTP)^[11]对 Snippet 进行命名实体识别,提取其中的人名. 系统需要考察的因素有如下 4 项:



(1)子图节点的两种选定方法:单点拓展(E1), 两点拓展(E2).

(2)拓展检索的两种处理方法:平均拓展(M1), 最大拓展(M2).

(3)是(C1)否(C0)采用基于谱分割和模块度的聚类算法. 这里不采用图聚类算法就是默认社会网络图中各个连通子图形成单独的类别,采用图聚类算法使得可以进一步将连通子图内的社会圈子挖掘出来.

(4)拓展检索有效 Snippet 数(Extend-number): 每次选取子图中节点进行拓展检索时需要的有效 Snippet 数,这里选取拓展数分别为 0,10,20,30,40,50.

针对以上 4 种因素,本文对每个人名的语料详细进行了以上各种情况 $2 \times 2 \times 2 \times 6 = 48$ 组实验. 下面介绍详细的实验结果和分析.

5.2 实验结果

按照 5.1 节的实验设计,4 个人名总体的结果如图 2 所示. 图中有两条曲线,对应每种组合下 F -measure 值随着拓展有效 Snippet 数的变化曲线.

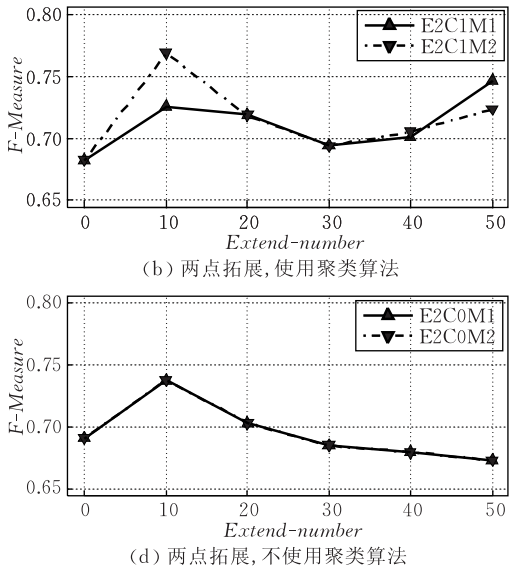


图 2 总体实验结果曲线图

5.3 实验结果分析

实验中涉及到的参数较多,针对各种情况,分析如下:

(1) 总体分析

从图 2 可以看出,总体实验达到的最好水平是 0.7689,其参数配置为 E2C1M2 拓展数为 10,即两点拓展、采用图聚类算法、最大拓展、每个拓展点选取 10 个有效 Snippet. 这个结果在 Upper bound

(0.9543) 和 Lower bound (0.4935) 之间处于 0.5977 的水平,即达到理论最好效果的 60%,说明本文的方法还存在值得改善的地方. 事实上,在实验过程中发现,本文采用的命名实体识别模块对于 Snippet 进行的人名识别效果不是非常好,例如会把“组图”、“宝典”等识别为人名,也会把“孙俪”等不识别为人名. 这样会对社会网络的构建以及聚类产生负面影响.

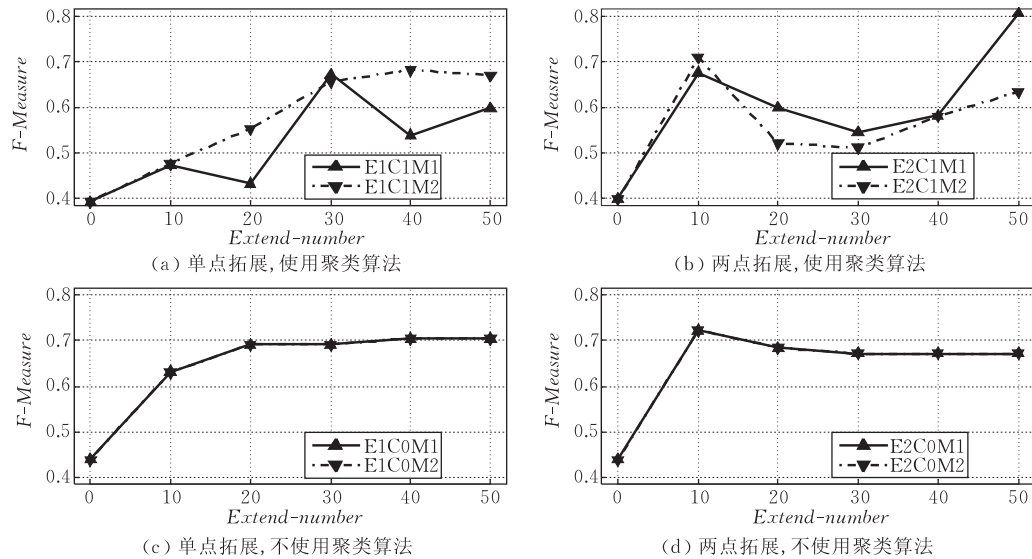


图 3 “王刚”实验结果曲线图

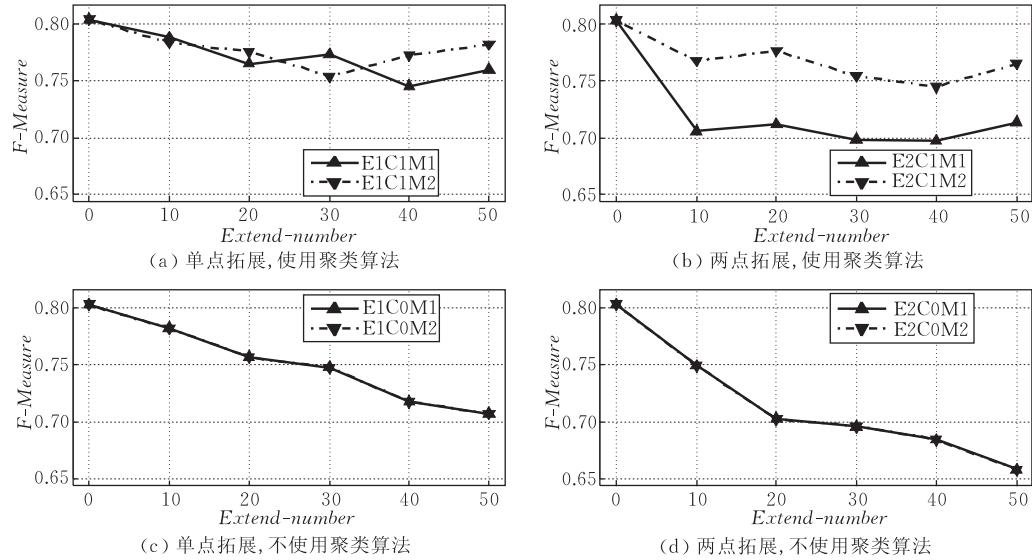


图 4 “张伟”实验结果曲线图

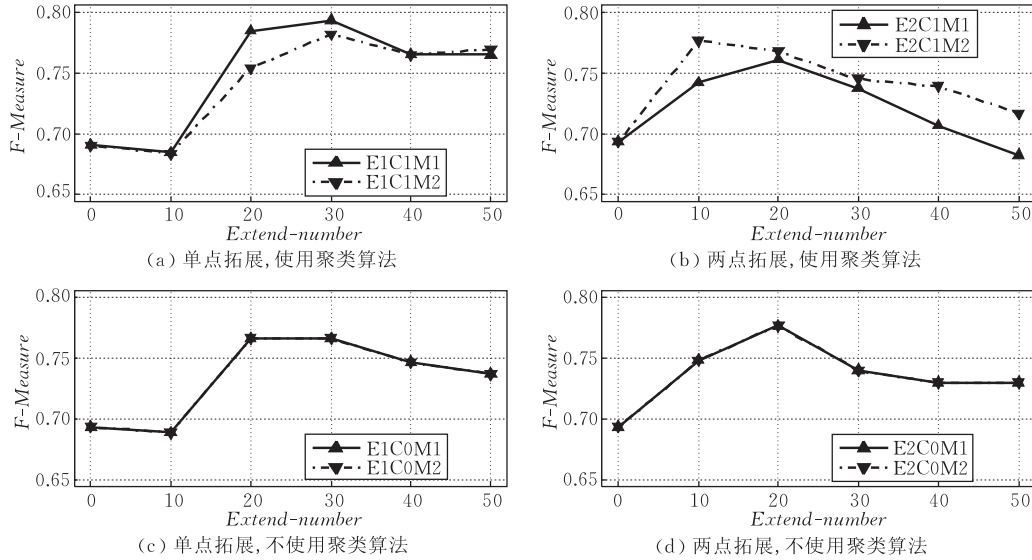


图 5 “王伟”实验结果曲线图

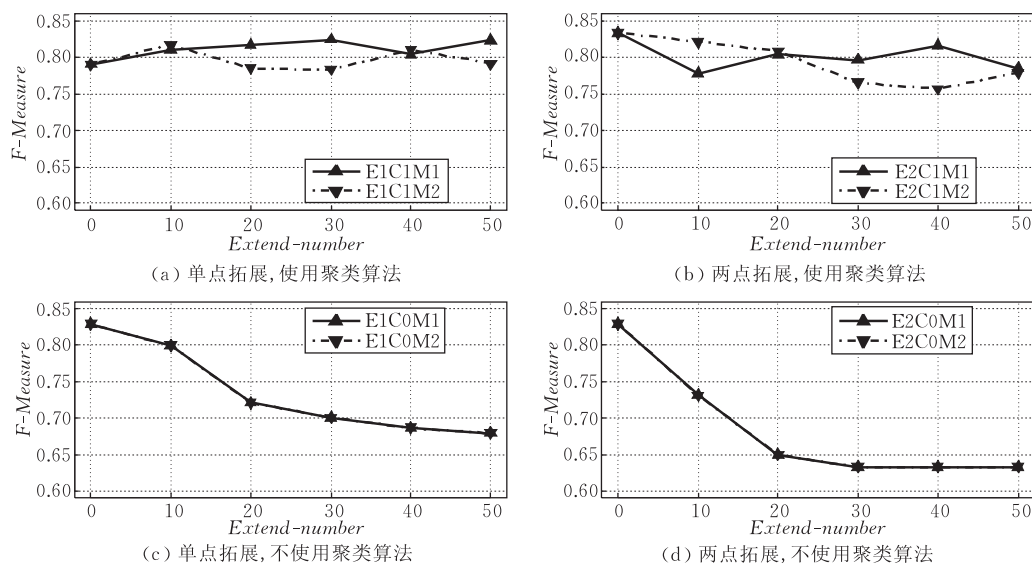


图6 “刘波”实验结果曲线图

另外实验中发现现在拓展后的社会网络图中存在为数众多的单点. 这种单点产生的原因是检索得到的有效 Snippet 中, 除了检索人名外仅含有一个人名, 而这个人名在拓展检索时没有引入新的人名或者没有能找到和其它子图连接起来的人名. 最终在进行图聚类的时候注定了只能单独成为一个社团. 这种情况如果得以进一步处理的话, 会使得实验效果得到提高.

(2) 标注人名类数和实验结果的关系

在比较 4 个人名单独的实验曲线时可以发现: 各自能够达到的最好结果都在 0.8 左右, 但当人名对应的人工标注类数较少时结果随着拓展有效 Snippet 数的增加, 结果曲线的变动较为剧烈. 这说明本文提出的算法对于类别数少的人名鲁棒性不够.

(3) 图聚类算法的效果分析

图聚类算法的目的是将拓展后网络中的连通子图进行自动划分, 在每幅图的 (c)、(d) 和其上的 (a)、(b) 相对比时可以发现, 采用图聚类算法整体上能够提高实验效果. 这说明图聚类算法起到了很好的效果.

(4) 单点拓展、两点拓展的区别

综合比较每幅结果图中的左右两列, 可以发现, 两点拓展可以使得结果曲线能较快达到平缓的阶段. 观察发现, 一般两点拓展能够比单点拓展在拓展数目上提前 10 个有效 Snippet.

(5) 拓展检索处理方法的影响

实验结果的每幅图中的 (c)、(d) 图中两条曲线重合在一起. 这是因为, 不采用图聚类算法的时候,

平均拓展和最大拓展本质上只是关系矩阵中非 0 关系系数不完全相同. 对应的社会网络图中的连通子图的拓扑结构是完全一致的, 因此在不采用图聚类算法来处理连通子图内部分布的时候, 两种拓展的结果就应当是完全一致的.

在每幅结果图中观察 (a)、(b) 中可以发现, 最大拓展的曲线相对平均拓展而言变化舒缓一些, 即最大拓展比平均拓展具有更好的鲁棒性. 这验证了前文提出最大拓展时的考虑.

(6) 拓展数的影响

从每幅图的 (c)、(d) 可以发现随着拓展数目增加到 20 后, 实验结果曲线开始变得平缓. 另外, 随着拓展数目的增加, 不使用图聚类算法的曲线变得平缓, 使用图聚类算法的曲线仍然变化甚至出现效果变好. 这说明拓展数目增加到一定程度后, 已经不能将一些连通子图连接起来, 但是仍然能够将一些连通子图内的网络结构进行更新, 同时也可能引入更多的噪音反而影响实验效果.

6 结论和展望

人名检索结果重名消解可以采用社会网络的方法来实现. 本文结合检索结果中共现人名之间隐含的社会网络自动聚类的方法实现了检索结果的重名消解. 值得一提的是, 这种聚类方法不需要设定参数, 能够自动确定最终的类别数量. 从实验效果来看, 整体性能达到较好水平. 图聚类算法的采用能帮助连通社会网络的进一步划分, 从而提高聚类效果. 当消解的目标人名对应的人物数量较少时, 本文的

方法鲁棒性不够. 每次选取连通子图中带权度最大的两点来进行拓展比选取单点拓展能使系统性能较快提高. 拓展检索对于连通子图的完善比连接不同连通子图更加明显. 实验还证明通过中间人判定另外两人之间的关系密切程度时, 多个中间人的判定结果选取最大值比平均值更为合理.

下一步, 我们计划采用更好的人名识别模块来提高系统性能, 在算法效率以及图聚类算法上进行优化, 可以考虑抽取其它类型的命名实体以及结合文本聚类的思想来更好地完成检索结果的重名消解.

参 考 文 献

- [1] Wang Houfeng, Mei Zheng. Chinese multi-document personal name disambiguation. *High Technology Letters*, 2005, 11 (3): 280-283
- [2] Wang Houfeng. Cross-document transliterated personal name coreference resolution//Wang Lipo, Jin Yaochu eds. *Proceedings of the Fuzzy Systems and Knowledge Discovery*. Changsha, China: Springer, 2005: 11-20
- [3] Bollegala D, Matsuo Y, Ishizuka M. Disambiguating personal names on the Web using automatically extracted key phrases//Brewka Gerhard, Coradeschi Silvia, Perini Anna, Traverso Paolo eds. *Proceedings of the 17th European Conference on Artificial Intelligence*. Riva del Garda, Italy: IOS Press, 2006: 553-557
- [4] Yu Man-Quan. Research on knowledge mining in person tracking [Ph. D. dissertation]. Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 2006 (in Chinese with English abstract)
(于满泉. 面向人物追踪的知识挖掘研究 [博士学位论文]. 中国科学院计算技术研究所, 北京, 2006)

- [5] Ron Bekkerman, McCallum Andrew. Disambiguating Web appearances of people in a social network//Ellis Allan, Hagino Tatsuya eds. *Proceedings of the 14th International Conference on World Wide Web*. Chiba, Japan: ACM Press, 2005: 463-470
- [6] Bradley Malin. Unsupervised name disambiguation via social network similarity//Kargupta Hillol, Srivastava Jaideep, Kamath Chandrika, Goodman Arnold eds. *Proceedings of the Workshop on Link Analysis, Counterterrorism, and Security*, in Conjunction with the SIAM International Conference on Data Mining. Newport Beach, California, USA: SIAM, 2005: 93-102
- [7] Wang Xiao-Fan, Li Xiang, Chen Guan-Rong. *Complicated Network Theory and Its Application*. Beijing: Tsinghua University Press, 2006 (in Chinese)
(汪小帆, 李翔, 陈关荣. 复杂网络理论及其应用. 北京: 清华大学出版社, 2006)
- [8] Mohar B. Some applications of Laplace eigenvalues of graphs. *Graph Symmetry: Algebraic Methods and Applications*, 1997, 497: 225-275
- [9] Newman M E J, Girvan M. Finding and evaluating community structure in networks. *Physical Review E*, 2004, 69 (2): 26113
- [10] Bagga A. Evaluation of coreferences and coreference resolution systems//Rubio A, Gallardo N, Castro R, Tejada A eds. *Proceedings of the 1st International Conference on Language Resources and Evaluation*. Granada: European Language Resources Association, 1998: 789-793
- [11] Lang Jun et al. LTP: An XML-based open language technology platform//Cao You-Qi, Sun Mao-Song eds. *Proceedings of the 25th Anniversary of the Chinese Information Processing Society of China*. Beijing: Tsinghua University Press, 2006: 561-572 (in Chinese with English abstract)
(郎君等. 基于 XML 的开放式语言技术平台: LTP//曹右琦, 孙茂松主编. 中国中文信息学会成立 25 周年学术年会论文集. 北京: 清华大学出版社, 2006: 561-572)



LANG Jun, born in 1981, Ph. D. candidate. His research interests include information extraction, coreference resolution, and machine learning.

QIN Bin, born in 1968, Ph. D., professor. Her research interests include natural language processing, and information retrieval.

SONG Wei, born in 1983, Ph. D. candidate. His re-

search interests include natural language processing, and information retrieval.

LIU Long, born in 1985, M. S. candidate. His research interests focus on natural language processing.

LIU Ting, born in 1972, Ph. D., professor, Ph. D. supervisor. His research interests include natural language processing, and information retrieval.

LI Sheng, born in 1943, Ph. D., professor, Ph. D. supervisor. His research interests include natural language processing, and machine translation.

Background

Searching people information is one of the major activities of Internet users. However, in the real world, a number of people share one name is a very common phenomenon. This has led to the results of searching a person name are usually a mixture of pages about the namesakes. Although some systems can handle searching results clustering, they deal with the person name as general terms. Moreover, the generating labels of the clusters are common terms. These systems have not directly distinguished the namesakes’ searching results.

Person name disambiguation of searching results can utilize the context of person names and adopt similar methods to word sense disambiguation. The common approaches extract the searched snippets or corresponding Web pages, and extract the key context phrases for vector space model, then use the vector similarity for final searching results clustering. The better solution is extracting the people related information records for calculating people similarity and judging person identity, such as gender, nation, native place, birth date, family ties, home address, position, and so on.

To person name disambiguation of searching results, text-clustering approaches have considered many useless words, and require manual setting threshold or class number. The personal information extraction and person similarity based approach is very dependent on the personal information extraction. All kinds of extraction errors are easy to cause cascading errors.

To solve these problems, this paper proposed using social network for person name disambiguation of searching results. The method is mainly based on the saying “Birds of a feather flock together”. That is different people with same name have the distinction of their own social networks. For example, the social network for “Wang Gang” of entertainment circle is significantly different from that for “Wang Gang” of political circle. This paper utilizes the hidden social networks of the searching results for person name disambiguation. It aims at Chinese person names’ searching results, and combines spectral partition and modularity evaluation for automatically clustered into different social communities.