基于组合特征的 Web 人名消歧方法①

辛 涛, 程绍银, 蒋 凡

(中国科学与技术大学 计算机学院, 合肥 230027)

摘 要: 重名问题在Web人物搜索过程中是很普遍的现象. 研究了Web人名消歧相关问题, 提取与待消歧人名相关的不同特征集, 运用向量空间模型构造人物实体的组合特征, 最后通过层次聚类算法将相似度高的文档优先聚类, 由此实现人名消歧. 在 WePS 数据集上的实验结果表明, 提出的方法具有良好的消歧效果.

关键词: 重名问题; 人名消歧; 特征提取; 组合特征; 层次聚类

Web Name Disambiguation Approach Based on Combined Features

XIN Tao, CHENG Shao-Yin, JIANG Fan

(School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China)

Abstract: Name ambiguity is a common phenomenon when one tries to search for someone's information in the Internet. In this paper, we have studied the web name disambiguation issue in detail. After extracting different features related to the name and then creating combined features by vector space model, we give priority to cluster the documents with high similarity by hierarchical clustering algorithm. Evaluated on the WePS data set, the proposed method showed its effectiveness in solving name disambiguation problem.

Key words: name ambiguity; person disambiguation; feature extraction; combined features; hierarchical clustering library

在互联网上搜索人物信息是十分常见的事情, 据统计[1]约有 15%-17%的搜索引擎查询涉及到人名. 然而由于人物重名现象的存在, 通用搜索引擎的结果往往是许多同名不同人物的网页混合集, 例如 Google 搜索"Michael Jordan", 结果会涉及若干个不同人物实体, 比如篮球明星、大学教授、电影演员等. 对此, 用户需要耗费大量时间筛选特定目标人物信息, 且有遗漏重要信息的风险.

人名消歧就是用来解决此类由于人物重名而导致的歧义问题, 其可以形式化描述成: 给定文档集 $D=\{d_1,d_2,...,d_n\}$ 是从搜索引擎中通过搜索某人名 A 得到的重名人物网页混合集, 假设人名 A 由现实中 k (k 未知)个实体 $E=\{e_1,e_2,...,e_n\}$ 所共享, 人名消歧的作用就是建立一个网页簇集合 $C=\{c_1,c_2,...,c_n\}$, 其中 $c_i=\{d_{i1},d_{i2},...,d_{ix}\}(1\le i\le k)$, d_{ix} 是D中一个网页, c_i 是描述人物个体 e_i 的网页集合. 因此, 经过人名消歧后的网页集按照现实人物个体进行划分, 可以方便用户查找目标人物信息. 同时, 人名消歧是进行人物信息抽取集成的基础, 在人物搜索、热点人物追踪、问答系统等领域有着重要应用.

本文对 Web 人物搜索中的的人名消歧进行重点研究, 提出一种基于组合特征的 Web 人名消歧方法. 实验证明, 该消歧方法可行且效果良好.

1 相关研究

人名消歧最初作为实体共指问题进行研究, 主要思路是判断两篇文档中人名是否指向同一个体. 文献[2]首先对文档进行指代消解, 选取人名周围句子重组成概要信息, 然后使用向量空间模型对概要进行建模, 通过向量间相似度比较进行消歧. 该方法的缺陷在于比较仅在两文档之间进行, 没有考虑全局特点.

随着互联网的发展, Web 人名消歧开始受到更多关注. WePS(Web People Search)研讨会于 2007 年开始连续三届开展了针对 Web 人名消歧的评测活动, 该评

① 收稿时间:2015-03-09;收到修改稿时间:2015-05-08

测活动把人名消歧定义为聚类问题, 并建立了统一的评测平台, 取得了显著成果[3-5].

不同的人名消歧方法选取不同的聚类特征进行消歧, 而人名消歧的效果与所使用的特征高度相关. 文献[6]提出基于个人传记特征(例如生日、职位等)的人名消歧方法, 但由于传记特征的稀有性其消歧效果并不好. 文献[7]使用单一的命名实体特征进行消歧, 另有许多方法使用网页中的词汇标记作为聚类特征, 比如普通词汇[8]、单词 n 元语法(n-gram)[8,9]、短语[10]等. 一些方法尝试引入其他语料帮助人名消歧, 例文献[11]通过搜索引擎获取更多的摘要信息以增进文档间的衔接; 文献[12]利用维基百科知识库帮助识别网页中的专有名词、短语等; 文献[13]使用查询扩展文档中的实体特征, 外部语料的引入在提高人名消歧效果的同时也大大增加了系统的消耗.

本文方法的特点是充分利用本地特征, 通过对多种特征信息进行组合, 弥补了单一特征的不足之处, 提高人名消歧最终效果.

2 人名消歧框架及方法

在总结前人相关研究的基础上, 本文把人名消歧过程分为四个步骤: 预处理、特征提取、特征组合及层次聚类. 如图 1 显示了本文人名消歧方法框架图.

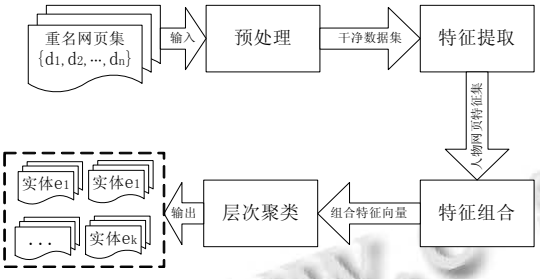

图 1 人名消歧方法框架图

(1) 预处理: 对重名人物网页集进行清洗工作, 去除其中与人物信息无关的噪音, 得到干净的数据集;

(2) 特征提取: 对经过预处理之后的网页数据集, 利用不同的处理策略, 提取其中的人物网页特征信息;

(3) 特征组合: 采用向量空间模型对提取的网页特征进行建模, 通过线性组合构造组合特征向量;

(4) 层次聚类: 根据组合特征向量间的相似度, 运用层次聚类算法进行聚类消歧, 生成针对每个重名不同实体的网页聚簇

2.1 预处理

预处理的目标是去除 HTML 网页源码中的标签、script 脚本和对人名消歧无用的导航菜单, 广告等噪音信息, 提取网页的正文信息. 原始 HTML 网页首先借用 HTMLParser 工具包去除其中的标签和 script 脚本. 采用和文献[14]中相似的方法, 对于块级层次的标签(例如 , <p>)只有超过 10 个单词的文本块内容被保留, 以便移除导航菜单、广告等. 预处理的质量对之后的聚类算法效果有着很大影响.

2.2 特征提取

人物个体均有其自身的特征, 同名人物之所以可被区分, 就是因为其与众不同的特征, 例如生日、居住地等; 人物实体网页中词汇信息同样可以反映人名归属. 聚类算法可以通过这些特征及特征间相似度判断同名人物网页所属类别.

本文分别对重名人物网页的 URL、标题、摘要、正文等不同的信息源采用了不同的处理策略, 抽取其中的文本特征信息.

(1) 网页 URL

网页的 URL 中包含了一些指示信息. 例如 http://www.cs.berkeley.edu/~jordan/ 指向的是'berkeley'大学"cs"(计算机学院)的教授"jordan". 对 URL 的处理过程如下: 以 URL 中的分隔符(: / .)为边界把其切分成独立的小字串单元, 对比自建字典移除其中常见的对消歧无意义的字符串, 例如 http、www 等. 纯数字串、特殊符号同样会被移除. 上文举例 URL 经过处理后得到的关键字有 cs | berkeley | edu | jordan.

(2) 网页正文

由于单复数、时态语态等语法的存在, 英语具有丰富的词形变化. 对网页正文的处理目的是将其中的词汇转化为较为规整的格式. 对此, 本文采用通用的文本处理方法, 主要包括文本分词、归一化、去除停用词、词干提取, 最终得到含有重复词汇的词干集合. 例如 computes、computing、computed 经过处理后均得到 comput, comput 为表征网页正文的一个关键字.

(3) 网页标题及摘要

标题是对网页正文总结性的文字, 摘要是由搜索引擎返回的概要信息, 是对查询主题词信息的高度概括, 因此两者对人名消歧的指示作用相比正文更准确. 对标题和摘要的处理与正文类似, 均需做词干提取.

(4) 命名实体

命名实体是区分人物身份很重要的特征, 人物居住地点、工作单位名称和周围人物姓名都可以很好地标识一个人物. 本文使用 The Stanford Named Entity Recognizer1 识别网页正文中出现的人名、地名、公司组织等命名实体. 因为命名实体大多为专有名词, 因此只需统一转化为小写, 并不做词干提取. 另外, 本文通过定义正则表达式规则帮助识别文本中出现的邮箱、数字串等其他特殊字串信息, 由于其出现频率一般较少但消歧效果更佳, 本文通过加倍其频率值的方式增加其影响.

2.3 特征组合

对于提取的人物网页特征集, 需要对其进行建模, 将其转化为计算机可以理解的数学模型, 才方便作进一步的处理.

本文首先采用向量空间模型对上文提取的特征信息分别进行建模, 即特征集被表示成具有一定权重值的关键字组成的多维向量, 关键字的权重设为其在特征集中的出现频率. 由于网页 URL、标题和摘要的篇幅比较短小, 在建模过程中把三者合并在一起, 简称为网页基本信息. 假设建模情况如下:

网页基本信息生成的特征集 f1:

$$\vec{v}_1=\left((k_{11},tf_{11}),(k_{12},tf_{12}),...,(k_{1m},tf_{1m})\right)$$

网页正文生成的特征集 f2:

$$\vec{v}_2=\left((k_{21},tf_{21}),(k_{22},tf_{22}),...,(k_{2m},tf_{2m})\right)$$

命名实体生成的特征集 f3:

$$\vec{v}_3=\left((k_{31},tf_{31}),(k_{32},tf_{32}),...,(k_{3m},tf_{3m})\right)$$

其中, tf 是特征集中的关键字 k 的出现频率.

本文采用线性加权的方式将三个独立特征向量融合在一起构造组合特征向量 $\vec{v}=\lambda_1\vec{v}_1+\lambda_2\vec{v}_2+\lambda_3\vec{v}_3$, 加权系数 λ_1、λ_2、λ_2 依照特征集对人名消歧指示作用的贡献程度设定, 具体数值是通过实验人工给出的相对较优的结果. 例如相比较正文信息, 网页基本信息有着更强的消歧指示作用, 因此 λ_1 值设置相对高于 λ_2.

对于新生成的组合特征向量 $\vec{v}$ 使用 TF-IDF 统计方法对其权重值进行重新优化, 主要思想是: 如果关键字在一篇文档中出现的频率比较高, 且在其他文档中很少出现, 则该关键字具有很好的类别区分能力, 适合用来消歧. 因此, 最终生成的代表某个人物相关网页的组合特征向量为:

$$\vec{v}'=\left((k_1,w_1),(k_2,w_2),...,(k_{m'},w_{m'})\right)$$

其中, $w_i=log(tf_i+1)\times log(\frac{N_d}{df_i}),1\le i\le m'$, tf_i 是 $\vec{v}$ 中某关键字的出现频率, N_d 是待消歧人物网页文档总数, df_i 表示出现关键字 k_i 的文档数.

2.4 层次聚类

假设同一篇网页文档中出现的待消歧人名仅对应现实中一个人物个体, 则人名消歧可以看成是硬聚类问题, 聚类结果没有重叠; 同时, 由于重名人物的数目未知且不固定, 此类问题又属于非监督类问题, 适用于聚类算法.

本文采用凝聚层次聚类算法, 两文档之间的相似度由文档特征向量之间的夹角余弦来表示, 类间相似度采用的是平均距离法. 聚类初始时, 每个人名对应的文档集 $D=\{d_1,d_2,...,d_n\}$ 看作是一个具有单个成员的类, 因此构成初始聚类 $C=\{c_1,c_2,...,c_n\}$; 计算类(c_i,c_j)之间的相似度, 选取相似度最大的两个类进行合并, 生成新的类 $c_m=c_i\cup c_j$, 从而构成 D 的一个新的聚类 $C=\{c_1,c_2,...,c_{n-1}\}$; 重复上述步骤, 直到所有类之间相似度小于给定的相似度阀值 β 或者全部聚为一类为止.

聚类算法伪代码如下: Input:$D=\{d_1,d_2,...,d_n\}$

```
Output:C={c1,c2,...,cn}
Algorithm HAC(D)
  begin
  let each document be in a singleton cluster{ ci }
  while |C|>1
     maxsim= max_{0<i<|C|, 0<j<|C|} sim(ci,cj),(i≠j)
     if  maxsim>= β
         cm = ci ∪ cj （m≠i,m≠j）
         remove ci , cj from C
     else
          break;
     endif
  endwhile
```

3 实验设计及分析

3.1 实验数据

本文实验数据集采用 SemEval WePS[2] 提供的数据集, 该数据集包含 49 个人名的训练集、30 个人名的测试集, 每个人名对应搜索引擎返回的前 100 个网页以及相关的 URL、标题、摘要等信息. 该数据集还提供了人工标注好的正确结果供评测使用.

3.2 评测标准

针对每个人名对应的文档集进行评测, 采用 SemEval WePS 通用评测标准: 纯度 P(Purity)、逆纯度

IP(Inverse Purity)及综合评测 F 值. 公式如下:

$$P=\sum_i \frac{|C_i|}{n} max\ precision(C_i, L_j)$$

$$IP=\sum_j \frac{|L_j|}{n} max\ precision(L_j, C_i)$$

$$F=\frac{1}{\alpha\frac{1}{P}+(1-\alpha)\frac{1}{IP}}$$

其中, $precision(C_i, L_j)=\frac{|C_i\cap L_j|}{|C_i|}$, C 为待评测人名聚类的集合, L 为人工标注的正确聚类结果集, C_i 、L_j 分别为 C、L 中元素, n 为被聚类的元素总数. 评测中取 α 为 0.5, 用 $F_{0.5}$ 对 P 和 IP 进行综合评测.

3.3 实验分析

本文对实验数据主要从三个角度分析, 即层次聚类相似度阀值设定, 聚类算法有效性验证和不同特征组合的结果对比分析.

3.3.1 相似度阀值设定

使用 WePS 数据集中的训练集, 完成对层次聚类相似度阀值的设定. 实验用枚举方法对相似度阀值 β 从 0.0 到 1.0, 步长为 0.1 进行了反复测试.

图 2 给出了层次聚类在相似度阀值 β 不同取值情况下人名消歧的实验结果, 结果均以宏平均(macro-average accuracy)的方式给出,即每个人名的评测结果相加之后取平均值(下同).

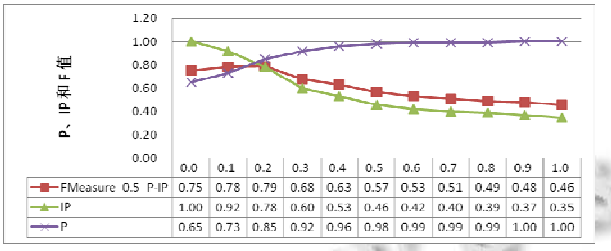

图 2 P、IP 和 F 值随阈值变化曲线图

由上图 2 可看出, 随着聚类阀值增加, 聚类对文档相似程度的要求增高, 则文档最终聚类的结果愈加偏向分散, 因此纯度 P 值越高, 逆纯度 IP 的值则逐渐减少. 而 $F_{0.5}$ 值在 β=0.2 时取值最大. 层次聚类相似度阀值设定为使聚类效果最优的数值, 因此在以后的实验中相似度阀值取 0.2.

3.3.2 聚类算法有效性验证

聚类是人名消歧的一个关键因素. 为了验证层次聚类方法解决人名消歧问题的有效性,本文选择了三个基准方法进行对比:

(1) All-in-one:待消歧人名的所有网页文档被归为一个类.

(2) One-in-one:待消歧人名下的每个网页文档被划分为单独的类.

(3) Kmeans 聚类方法: Kmeans 聚类算法中 k 的值定义为 WePS 数据集中提供的官方聚类数目, 即人工标注的重名人物个数. 如此就避开了 k 因素的影响, 仅考虑聚类方法的有效性.

实验在 WePS 测试集上进行, 特征组合中加权系数 λ_1、λ_2 、λ_2 分别设置为 2、1、2, 人名消歧效果评测结果如下表 1 所示.

表 1 测试语料集上评测结果

聚类特征	ALL_IN_ONE			ONE_IN_ONE		
	$F_{0.5}$	IP	P	$F_{0.5}$	IP	P
无	0.40	**1.00**	0.29	0.61	0.47	**1.00**
聚类特征	层次聚类			Kmeans 聚类		
	$F_{0.5}$	IP	P	$F_{0.5}$	IP	P
f1	0.72	0.67	0.85	0.70	0.64	0.78
f2	0.76	0.75	0.79	0.70	0.66	0.78
f3	0.75	0.70	0.83	0.68	0.63	0.76
f1+f2	0.78	0.77	0.83	0.72	0.67	0.80
f1+f3	0.79	0.74	0.89	0.72	0.67	0.79
f1+f2+f3	**0.82**	0.76	0.91	0.72	0.68	0.80

对比三个基准方法: 显然, All-in-one 方法的逆纯度 IP 值为 1.00, One-in-one 的纯度 P 值为 1.00. 基于现实中有许多人物实体仅对应一个网页文档的事实, One-in-one 的 $F_{0.5}$ 高于 All-in-one, 能取得相对较优的结果. 对于 Kmeans 聚类算法, 在给定精确 k 值的情况下, 针对 6 种不同的特征组合, 其人名消歧效果均不如层次聚类算法. 由此可见, 凝聚层次聚类的方法更加适用于解决人名消歧问题.

3.3.3 特征组合分析

图 3 显示了层次聚类算法在不同特征组合情况下的消歧效果, 具体数值见表 1.

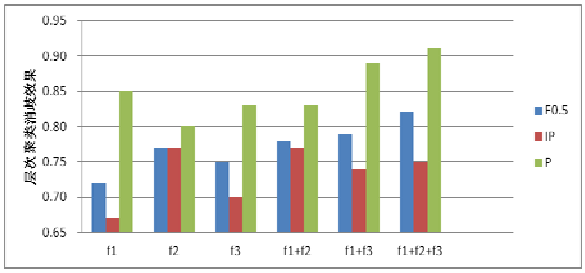

图 3 层次聚类在不同特征组合下消歧效果

由上图 3 可以看出: 当仅使用特征 f1 进行人名消歧时, 纯度 P 尚可, 但逆纯度 IP 较低, 说明虽然特征 f1 中的个人信息较准确, 但由于信息量少, 数据比较稀疏, 造成其逆纯度较低. 基于特征 f2 的消歧在逆纯度 IP 有所提高的同时其纯度 P 降低, 主要原因是正文中含有大量个人信息的同时引入了一些其他噪音. 命名实体有助于准确分辨人物实体, 因此基于特征 f3 的消歧结果纯度 P 也比较高, 其逆纯度 IP 比较低, 可能是由于受命名实体识别工具准确度的影响.

不同的特征有着不同的特点, 通过组合可以弥补单独使用某一类特征的不足之处, 最终提高人名消歧的效果. 基于 f1+f2、f1+ f3 的消歧在保证聚类结果纯度适当的同时其逆纯度均有所提高. 基于 f1+f2+f3 的聚类其纯度、逆纯度都有增加, 取得了最佳的消歧效果: P=0.91, IP=0.76, $F_{0.5}$ =0.82.

同时在实验中发现, 基于特征 f1 的消歧效率最高, 消歧效果可接受, 适用于在线搜索消歧. 基于 f3 的消歧, 因借助于命名实体识别工具, 效率相对较低, 但其在消歧之后的 Web 个人信息集成中有很大的实用性. 因此, 在实际问题中可根据实际需求选择不同的特征组合进行人名消歧.

4 结语

本文研究了 Web 人物搜索中的人名消歧问题, 重点关注了聚类特征的提取及组合问题, 并在 WePS 数据集上做了实验, 实验结果表明使用组合特征的凝聚层次聚类算法取得了较好的人名消歧效果. 但在实验中发现基于命名实体的人名消歧效果不是太突出, 可能原因是在命名实体识别的准确度, 因此在下一步研究中可以重点探讨如何提高命名实体识别准确度. 同时, 其他的一些特征也可以考虑加入帮助解决人名歧异问题, 例如网页间超链接、图表信息等.

参考文献

1 Guha R, Garg A, Guha R, et al. Disambiguating people in search. Thirteenth International World Wide Web Conference, 2004.

2 Bagga A, Baldwin B. Entity-based cross-document coreferencing using the vector space model. Proc. of the 17th International Conference on Computational Linguistics. [S.l.]. IEEE Press. 1998. 75–85.

3 Artiles J, Gonzalo J, Sekine S. The semeval-2007 weps evaluation: establishing a benchmark for the web people search task. Proc. of the 4th International Workshop on Semantic Evaluations, Association for Computational Linguistics. 2007. 64–69.

4 Artiles J, Gonzalo J, Sekine S. Weps 2 evaluation campaign: overview of the web people search clustering task. 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference. 2009. 9.

5 Artiles J, Borthwick A, Gonzalo J, et al. Overview of the web people search clustering and attribute extraction tasks. CLEF Third WEPS Evaluation Workshop. 2010.

6 Mann GS, Yarowsky D. Unsupervised personal name disambiguation. Proc. of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003. Association for Computational Linguistics. 2003, 4. 33–40.

7 Popescu O, Magnini B. Irst-bp: Web people search using name entities. Proc. of the 4th International Workshop on Semantic Evaluations (SemEval-2007). 2007. 195–198.

8 Chen Y, Lee SYM, Huang CR. Polyuhk: a robust information extraction system for web personal names. 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference. 2009.

9 Delgado AD, Martlnez R, Fresno V, et al. A data driven approach for person name disambiguation in web search results. COLING 2014, the 25th International Conference on Computational Linguistics. Dublin, Ireland. 2014.

10 Bollegala D, Matsuo Y, Ishizuka M. Extracting key phrases to disambiguate personal name queries in web search. Proc. of the Workshop on How Can Computational Linguistics Improve Information Retrieval? Association for Computational Linguistics. 2006. 17–24.

11 Rao D, Garera N, Yarowsky D. JHU1: an unsupervised approach to person name disambiguation using web snippets. Proc. of the 4th Int. Workshop on Semantic Evaluations. Association for Computational Linguistics. 2007. 199–202.

12 Long C, Shi L. Web person name disambiguation by relevance weighting of extended feature sets. CLEF (Notebook Papers/LABs/Workshops). 2010.

13 杨欣欣,李培峰,朱巧明.基于查询扩展的人名消歧.计算机应用,2012,32(9): 2488–2490.

14 Liu Z, Lu Q, Xu J. High performance clustering for web person name disambiguation using topic capturing. Ratio, 2011.