北京大学学报(自然科学版) 第 56 卷 第 4 期 2020 年 7 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 56, No. 4 (July 2020)
doi: 10.13209/j.0479-8023.2020.049

# 基于多特征融合的同名专家消歧方法研究

曾健荣[1] 张仰森[1,2,†] 王思远[1] 黄改娟[1,2] 崔佳[3] 马欢[3]

1. 北京信息科技大学智能信息处理研究所, 北京 100101; 2. 国家经济安全预警工程北京实验室, 北京 100044;
3. 国家计算机网络与信息安全管理中心, 北京 100029; † 通信作者, E-mail: zhangyangsen@163.com

**摘要** 针对专家库构建过程中出现的同名歧义现象, 提出一种基于多特征融合的同名专家消歧方法。从中国知网(CNKI)数据源中获取专家的论文信息, 抽取论文的标题、摘要、关键词、作者单位和合作者等关键信息, 并将其作为属性特征, 构建特征表示模型, 进而定义同名专家之间的相似度计算函数。根据计算得到的相似度, 将同名消歧问题转化为聚类问题。利用近邻传播聚类算法进行聚类, 解决同名消歧问题。在采集的专家论文数据上的实验表明, 基于多特征融合的同名专家消歧方法的准确率可达92%, 取得良好的消歧效果。

**关键词** 多特征融合; 同名消歧; 专家库; 聚类算法; 数据采集

# Research on Expert Disambiguation of Same Name Based on Multi-feature Fusion

ZENG Jianrong[1], ZHANG Yangsen[1,2,†], WANG Siyuan[1], HUANG Gaijuan[1,2], CUI Jia[3], MA Huan[3]

1. Intelligent Information Processing Laboratory of Beijing Information and Technology University, Beijing 100101;
2. Beijing Laboratory of National Economic Security Early-warning Engineering, Beijing 100044; 3. National Computer Network and Information Security Management Center, Beijing 100029; † Corresponding author, E-mail: zhangyangsen@163.com

**Abstract** According to the expert ambiguity with the same name in the process of building expert database, an expert disambiguation method based on multi-feature fusion is proposed. The paper information of experts is obtained from data sources such as CNKI. Key information (title, abstract, keyword, affiliation and collaborator) is extracted. The feature representation model is constructed with these information as attribute features. The similarity calculation function between experts of the same name is defined. According to the similarity, the problem of disambiguation of the same name is transformed into clustering problem. Affinity propagation clustering algorithm is used to solve the problem of homonymy disambiguation. Experiments on the collected expert papers show that the accuracy of the same-name expert disambiguation method based on multi-feature fusion can reach 92%, and good disambiguation results are achieved.

**Key words** multi-feature fusion; homonymy disambiguation; expert database; clustering algorithm; data collection

近年来, 随着国家对科研投入的力度越来越大, 各类科研项目的申请数量也越来越大, 使得遴选项目评审专家的难度加大, 因此对项目评审专家库的需求也愈加强烈。在构建一个全面而准确的专家库过程中, 极大的可能存在不同文本中含有相同专家姓名的现象, 即跨文本同名歧义现象。如果对跨文本同名歧义现象不做处理, 就无法对同一专家实体的信息进行融合, 势必对最终的专家推荐结果的科学性与准确性造成极大的影响。

同名消歧问题旨在将记载同名专家的多个文档进行区分, 将拥有相同姓名的文档映射到现实世界中的专家实体, 用以消除相同姓名造成的歧义。同

国家自然科学基金(61772081)和促进高校内涵发展–研究生科技创新项目(5121911044)资助
收稿日期: 2019–07–17; 修回日期: 2019–11–26

名消歧问题的形式化定义如下: 给定一个待消歧文档集合 $D=\{d_1,d_2,\ldots,d_n\}$, 每篇文档 $d_i(1\leqslant i\leqslant n)$都与同一个姓名 $N$相关, 并且具有一系列的属性(专家姓名、工作单位、研究方向、摘要和关键词等), 姓名 $N$由现实世界中的 $k$($k$值不确定)个人物实体集合 $E=\{e_1, e_2, \ldots, e_k\}$共享, 同名消歧的任务就是要建立文档簇集合 $C=\{c_{e1}, ce_2, \ldots, c_{ek}\}$, 其中 $c_{ei}=\{d_{i1}, d_{i2}, \ldots, d_{ix}\}(1\leqslant x\leqslant n)$, 是描述人物实体 $e_j(1\leqslant j\leqslant k)$的文档集合, $d_{ix}$是待消歧文档集合 $D$中的一个文档元素。因此, 经过同名消歧后的文档就与对应的人物实体划分到一起, 完成同名消歧任务。

## 1 相关工作

随着自然语言处理应用的日益广泛, 同名消歧的研究也受到越来越多的关注。针对同名消歧问题的学术会议主要有数字图书馆国际会议(Joint Conference on Digital Libraries, JCDL)[1]、web 人名搜索会议(Web People Search, WePS)[2]和中文处理国际会议(Joint Conference on Chinese Language Processing, CLP)[3]。JCDL 针对数字参考文献检索系统中的作者同名问题进行研究。WePS 设立针对英文人名消歧的评测任务, 从互联网网页上获取含待消歧人名的文档集合, 然后将具有相同指称的名字聚集在一起, 完成同名消歧任务。CLP 提供一个关于实体名的知识库, 对于人物姓名, 先判定其在知识库中是否有定义以及是知识库中的哪一条定义, 对于不属于知识库中定义的名字进行聚类, 有相同指称的名字聚为一类。

目前用于解决同名消歧的方法主要有 3 种: 基于向量空间模型(vector space model, VSM)的聚类消歧方法、基于社会网络的聚类消歧方法和基于实体链接的聚类消歧方法。

基于向量空间模型的聚类消歧方法首先选择并优化一些人名的属性特征, 如工作单位、研究方向和邮箱等, 然后用这些属性特征构建人物表征模型, 最后定义一个相似度函数来刻画同名人物之间的相似程度。Bagga[4]提出一种跨文档的人名消歧方法, 首先对文档进行指代消解分析, 然后基于 VSM 模型计算文档摘要之间的相似度, 最后将同一人物实体的文档聚类在一起。辛涛等[5]提出一种基于组合特征的 web 人名消歧方法, 首先提取与待消歧人名相关的不同特征集, 然后基于 VSM 构造人物特征实体的组合, 最后利用层次聚类算法对相似度高的文档优先进行聚类。

基于社会网络的聚类消歧方法一般先建立与待消歧人名相关的社会网络图, 利用图的相关理论求解图中的拓扑距离, 从而判断同一人名是否指向同一人物实体。陈晨等[6]先使用谱聚类对社会网络中的人名进行聚类, 然后根据不同社会网络边权值以及不同图划分准则对人名消歧效果的影响, 引入模块度阈值作为社会网络划分的停止条件, 在 CLP 2010 的中文人名消歧数据集上验证了社会网络分析方法的有效性。

基于实体链接的聚类消歧方法首先利用维基百科等知识库, 构建一个语义实体库, 然后通过相似度计算, 将提取的待消歧人名与语义实体库进行匹配, 从而实现同名人物与真实世界中实体的链接问题。Wang 等[7]通过计算待消歧人物文本与知识库实体文本的相似度, 实现实体链接的映射达到消歧的目的。宁博等[8]从中文维基百科中抽取包含人物信息和实体关系的实体信息对象, 与待消歧对象进行链接, 完成同名消歧。

本文提出基于多特征融合的同名专家消歧方法(homonymous disambiguation method based on multi-feature fusion, HDMMF), 该方法充分利用同名人物的属性特征, 通过综合多种属性特征, 从多侧面匹配的角度计算同名专家的相似度, 弥补单一特征的不足, 提高同名消歧的准确性。

## 2 基于多特征融合的同名专家消歧方法

基于多特征融合的同名专家消歧方法分为 4 个步骤: 专家数据的获取和预处理、同名专家的特征选择和表示、多特征融合的专家相似度计算以及基于聚类算法进行同名消歧, 整体框架如图 1 所示。

### 2.1 专家数据的获取与预处理

本文选择中国知网(CNKI)学术论文数据库作为数据源。CNKI 是集期刊杂志、博士论文、硕士论文、会议论文、报纸、工具书、年鉴、专利、标准、国学和海外文献资源为一体, 具有国际领先水平的网络出版平台[9]。采用网络爬虫框架(图 2)采集知网的数据。

在对 CNKI 页面结构的分析中发现, 从知网主页的搜索框查找文献时, 返回的所有搜索结果都嵌套在一个 iframe 中, 并且发送与返回的请求过于繁杂, 各请求之间的数据还相互纠缠掺杂, 不便于获取想要的数据。经过多次尝试后, 发现 http://search.

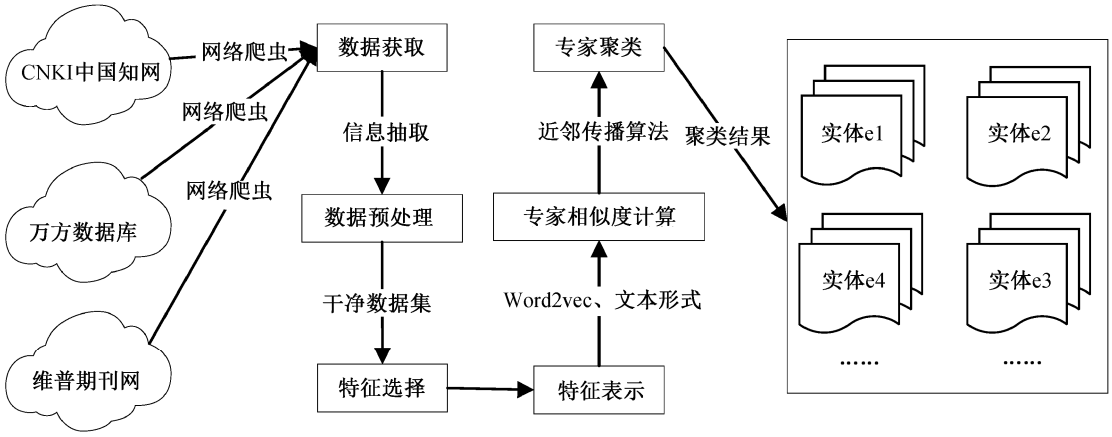


**图 1　基于多特征融合的同名专家消歧方法整体框架**

Fig. 1　Overall framework of expert disambiguation based on multi-feature fusion

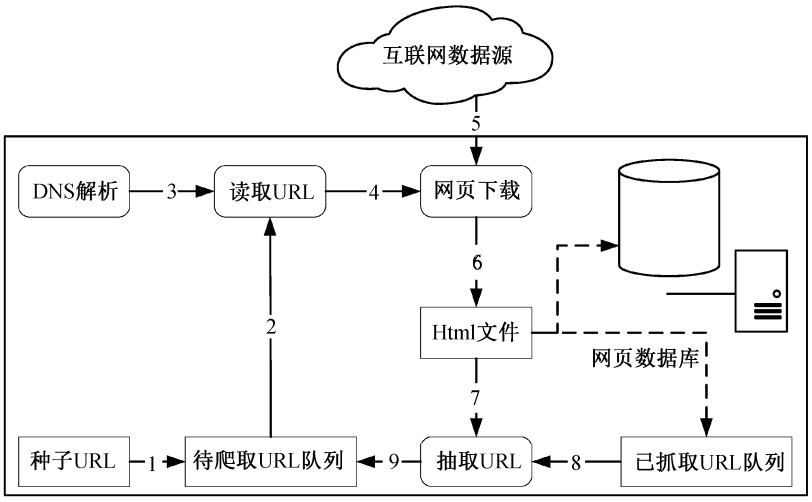


**图 2　网络爬虫框架**

Fig. 2　Web crawler framework

cnki.com.cn/default.aspx 这个接口符合我们的期望，请求简单，搜索结果与知网主页几乎没有差别，因此将其作为知网爬虫的接口。

对爬取的网页进行预处理。首先过滤一些空网页、广告页和不相关网页等噪声数据，然后进行信息抽取，抽取的内容包含论文的标题、作者、摘要、关键词、作者单位、被收录的期刊/会议、发表时间和分类号等信息，如表 1 所示。将这些数据以结构化的形式存储在数据库中。有些同名消歧的方法中还利用人物的传记式信息(例如职业、出生日期、E-mail 地址和电话等)[10]，虽然这些特征在同名消歧上具备较强的区分能力，但是由于网页内容的不确定性，这些传记式信息并不易获取，效果比较差，因此本文不考虑将人物传记式信息引入消歧方法。

**表 1　爬取的论文信息**

Table 1　Paper information crawled

| 序号 | 字段名称 | 字段描述 |
|---|---|---|
| 1 | title | 论文的标题 |
| 2 | author | 论文的作者 |
| 3 | abstract | 论文的摘要 |
| 4 | keyword | 论文的关键词 |
| 5 | affiliation | 作者的工作单位 |
| 6 | Journal_conference | 论文被收录的期刊/会议 |
| 7 | time | 论文发表的时间 |
| 8 | class_number | 论文的中图分类号 |

## 2.2 同名专家的特征选择与表示

### 2.2.1 特征选择

同名消歧的关键在于融合利用多种属性特征，

因此需要分析每一种特征消歧能力的强弱, 据此对每一种属性特征相似度赋予不同的权重, 得到同名专家的最终相似度。本文定义如下 5 种特征。

论文标题特征(title)。论文标题是论文内容的高度凝练与概括。根据论文作者不同的研究方向, 标题中往往包含能够在一定程度上反映作者研究方向的短语, 可以将这些关键性短语作为作者研究方向的特征。此外, 一般同一专家实体的研究领域与研究方向相对固定, 很少有专家的研究领域出现“振荡”。因此, 本文假设若同名专家的论文标题相似度较高, 则粗略地认为这些论文的同名专家可能是同一专家实体, 因为同一领域内还可能存在同名同姓的专家。

论文摘要特征(abstract)和论文关键词特征(keyword)。与论文标题类似, 论文摘要重点概括论文所做的工作, 也在一定程度上反映作者的研究方向。论文关键词一般是几个指明研究领域和方向的短语, 因此可以直接作为作者的研究方向。类似地, 若同名专家的论文摘要具有较高的相似度, 可以粗略地认为这些同名专家可能是同一专家实体。

作者单位特征(affiliation)。在本文采集的专家数据集中, 作者单位主要指论文作者任职的高校和科研院所等, 偶尔也有少量的公司和企业。若同名专家的单位具有较高的相似度, 可以粗略地认为同名专家是同一专家实体, 并且这种可能性比对前 3 种特征的判断更大, 因为同一单位的范围比同一研究领域的范围小, 同一单位中存在同名专家的概率低于同一研究领域存在同名专家的概率, 因此本文假设作者单位特征对同名专家的区分能力比前 3 种特征更强。

合作者特征(co-author)。对于具有多个署名作者的论文, 认为所有署名作者两两之间均是合作者关系, 将这种合作关系作为论文的合作者特征。假设两篇论文 $P_1$ 和 $P_2$ 的署名作者集合分别是

$$A_{P_1}=\{a_1,a_2^1,\ldots,a_m^1\}(m\geq 1), \tag{1}$$

$$A_{P_2}=\{a_1,a_2^2,\ldots,a_n^2\}(n\geq 1), \tag{2}$$

其中, $a_1$ 为同名专家, 则 $a_1$ 在论文 $P_1$ 中的合作者关系 $\text{co-author}_{a_1}^{P_1}$ 为集合 $A_{P_1}$ 除去元素 $a_1$ 后的子集, 即

$$\text{co-author}_{a_1}^{P_1}=\{a_2^1,\ldots,a_m^1\}。 \tag{3}$$

同理有:

$$\text{co-author}_{a_1}^{P_2}=\{a_2^2,\ldots,a_n^2\}。 \tag{4}$$

若式(5)成立,

$$\left|\text{co-author}_{a_1}^{P_1}\cap\text{co-author}_{a_1}^{P_2}\right|>0, \tag{5}$$

即表示同名作者 $a_1$ 在论文 $P_1$ 和 $P_2$ 中拥有共同的合作者, 则认为论文 $P_1$ 和 $P_2$ 的同名作者 $a_1$ 是同一专家实体, 并且拥有的共同合作者越多, 这种可能性越大。还有一种情况是拥有的共同合作者也存在同名歧义的现象, 这种情况比较复杂, 需要先对同名的共同合作者进行消歧, 但现实世界中这种情况出现的概率极小, 因此本文对这种复杂且出现概率低的情况不予考虑。

综上所述, 本文针对各属性特征对于同名专家区分力强弱程度的定性分析结果如表 2 所示。

### 2.2.2 特征表示

为了准确地计算属性之间的相似度, 进而得到同名专家之间的相似度, 本文对专家的论文标题、摘要和关键词这 3 个属性特征采用词向量模型来表示, 借助 Word2vec[11]来训练词向量; 对作者单位和合作者特征这 2 个属性, 直接用其原始的文本形式来表示。获得词向量的主要步骤如下。

1) 生成语料库。根据获取并经过预处理的专家论文数据, 选择每篇论文的标题、摘要和关键词, 按顺序以字符串拼接的形式组成一段文本, 用来表示这篇论文。这样, 从知网上爬取的论文库就形成一个大型的训练语料库。

2) 训练词向量。采用 Word2vec 的 CBOW 模型(continuous bag-of-words model, 连续词袋模型), 设置词向量维度为 200, 进行训练并输出训练结果, 保存到文件。

3) 生成专家特征向量。假设专家特征向量用一个五元组表示:

$$\boldsymbol{P}=\{\boldsymbol{T},\mathbf{Ab},\boldsymbol{K},\mathrm{Af},C\}, \tag{6}$$

**表 2 专家属性特征的强弱程度**

Table 2 Strength degree of expert attribute features

| 序号 | 属性特征名称 | 区分力强弱程度 |
|---|---|---|
| 1 | 论文标题特征(title) | 弱 |
| 2 | 论文摘要特征(abstract) | 弱 |
| 3 | 论文关键词特征(keyword) | 弱 |
| 4 | 作者单位特征(affiliation) | 中 |
| 5 | 合作者特征(co-author) | 强 |

其中，$\boldsymbol{T}$表示论文标题的词向量，**Ab**表示摘要的词向量，$\boldsymbol{K}$表示关键词的词向量，Af表示作者单位，$C$表示合作者。前3项均是向量化的表示，通过从词向量训练结果中找到对应词的词向量相加得到；后两项是原始的文本。这样，就生成表征一个专家的特征向量，用来计算同名专家之间相似度。

## 2.3 多特征融合的专家相似度计算

对于两个同名专家的特征向量$\boldsymbol{P}_1=\{\boldsymbol{T}_1, \mathbf{Ab}_1, \boldsymbol{K}_1, \mathrm{Af}_1, C_1\}$和$P_2=\{\boldsymbol{T}_2, \mathbf{Ab}_2, \boldsymbol{K}_2, \mathrm{Af}_2, C_2\}$，采用基于余弦相似度[12]的方法来计算前3项向量化的特征$\boldsymbol{T}$，**Ab**和$\boldsymbol{K}$之间的相似度:

$$\mathrm{Sim}\left(F_1, F_2\right)=\frac{\sum_{i=1}^{n} F_{1i} \times F_{2i}}{\sqrt{\sum_{i=1}^{n}\left(F_{1i}\right)^2} \times \sqrt{\sum_{i=1}^{n}\left(F_{2i}\right)^2}}, \tag{7}$$

式(7)中，$\mathrm{Sim}(F_1, F_2)$表示一组特征对$(F_1, F_2)$之间的相似度，$F_{1i}$和$F_{2i}$分别表示特征向量$F_1$和$F_2$的各分量。

作者单位和合作者属性都是长度较小的短语，并且计算它们的相似度不需要上、下文语义信息，只需考虑它们字符级别的相似度。因此，对这两类特征采用编辑距离(Levenshtein距离)[13]来计算其相似性。编辑距离表示从一个字符串转化为另一个字符串所需要的最少编辑次数(此处，编辑指替换、插入和删除字符的操作)。计算公式如下:

$$\mathrm{Sim}\left(S_1, S_2\right)=1-\frac{\mathrm{LevenshteinDistance}(S_1, S_2)}{\max\{\mathrm{Length}(S_1), \mathrm{Length}(S_2)\}}, \tag{8}$$

其中，$\mathrm{LevenshteinDistance}(S_1, S_2)$表示字符串$S_1$和$S_2$之间的编辑距离，$\max\{\mathrm{Length}(S_1), \mathrm{Length}(S_2)\}$表示$S_1$和$S_2$字符串长度中的最大值。

由向量化特征相似度和文本特征相似度可得到一对同名专家<$\boldsymbol{P}_1$, $\boldsymbol{P}_2$>相似度计算公式:

$$\mathrm{Sim}(\boldsymbol{P}_1, \boldsymbol{P}_2)=\boldsymbol{W}\cdot\mathbf{F\_Sim}, \tag{9}$$

$$\boldsymbol{W}=(\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3, \boldsymbol{w}_4, \boldsymbol{w}_5), \tag{10}$$

$$\mathbf{F\_Sim}=(\mathrm{Sim}(\boldsymbol{T}_1, \boldsymbol{T}_2), \mathrm{Sim}(\mathbf{Ab}_1, \mathbf{Ab}_2), \mathrm{Sim}(\boldsymbol{K}_1, \boldsymbol{K}_2), \mathrm{Sim}(\mathrm{Af}_1, \mathrm{Af}_2), \mathrm{Sim}(C_1, C_2)), \tag{11}$$

其中，相似度权重$\boldsymbol{W}$是一个5维行向量，$w_i(1\leqslant i\leqslant 5)$表示对应属性特征相似度的权重；**F_Sim**也是一个5维行向量，各分量分别表示各个属性特征的相似度。

## 2.4 基于近邻传播聚类算法的同名消歧方法

本文的目标是将具有相同专家姓名的论文记录正确地做出区分，使这些论文归属于正确的专家实体。以本质上讲，这是一个聚类问题。因此，本文采用聚类思想来解决本文的同名消歧问题。

常见的聚类算法主要有层次聚类算法、划分式聚类算法、基于网格和密度的聚类算法等[14]。层次聚类算法[15]在每次合并簇之后都需要重新计算样本点之间的距离，当数据样本点比较大时，计算量和算法复杂度也比较大；划分式聚类算法中最著名的就是K-means[16]，其优点是速度快且计算简便，缺点是需要事先知道数据要分成多少组，但是在同名消歧问题中，我们事先并不知道同名专家到底对应现实世界中多少个专家实体，因此K-means 算法不适用。基于网格和密度的聚类算法是对聚类数据空间进行网格规划，范围过大或过小都会影响结果。本文采用近邻传播算法(Affinity Propagation Algorithm, AP)对同名专家进行聚类。

近邻传播算法[17]的基本思想是，首先通过消息传递机制，搜索各个数据点的聚类中心以及数据点与聚类中心之间的隶属关系，然后根据隶属关系，对待聚类数据集进行划分，形成若干具有特定意义的子集。该算法不需要指定聚类个数，聚类中心是待聚类数据的某个确切的数据点，算法的输入可以是对称或非对称的相似度矩阵。

近邻传播算法需要定义一个吸引度矩阵$\boldsymbol{R}$和一个归属度矩阵$\boldsymbol{A}$。吸引度矩阵$\boldsymbol{R}$中的元素$r(I, k)$表示点$k$适合作为数据点$i$的聚类中心的程度，归属度矩阵$\boldsymbol{A}$中的元素$a(i, k)$表示点$i$选择点$k$作为其聚类中心的适合程度。在迭代过程中不断更新这两个矩阵，具体步骤见算法1。

**算法1** 近邻传播算法

输入：数据的相似度矩阵$\boldsymbol{S}$;

输出：数据聚类列表;

步骤1. 输入数据的相似度矩阵$\boldsymbol{S}$，初始化吸引度矩阵$\boldsymbol{R}$和归属度矩阵$\boldsymbol{A}$中的值均为0;

步骤2. 计算吸引度矩阵$\boldsymbol{R}$的值;

步骤3. 计算归属度矩阵$\boldsymbol{A}$的值;

步骤4. 迭代更新$\boldsymbol{R}$值和$\boldsymbol{A}$值;

步骤5. 若迭代满足以下条件之一:

① 超过迭代的最大次数,

② 信息量更新低于阈值,

③ 聚类中心保持稳定，则停止循环，否则继续重复

步骤 2~4;

步骤 6. 根据求出的聚类中心对数据进行分类, 并输出分类结果。

相似度矩阵 $S$ 中的元素根据 2.3 节中各属性特征相似度的计算方法得到, 吸引度矩阵 $\boldsymbol{R}$ 和归属度矩阵 $\boldsymbol{A}$ 的更新公式分别为式(12)和(13)。

$$r(i,k)=s(i,k)-\max_{k'\neq k}\{a(i,k')+s(i,k')\}, \quad (12)$$

$$a(i,k)=\begin{cases}\min\{0,r(k,k)\}+\sum_{i'\in\{i,k\}}\max\{0,r(i',k)\}, i\neq j,\\ \sum_{i'\neq k}\max\{0,r(i',k)\}, i=j。\end{cases} \quad (13)$$

## 3 实验结果及分析

### 3.1 评价标准

本文定义准确率(precision)、召回率(recall)和 $F_1$ ($F$-measure)值作为同名消歧方法的评价指标, 这 3 个指标计算公式如下:

$$\text{precision}=\frac{|A\cap B|}{|B|}\times 100\%, \quad (14)$$

$$\text{recall}=\frac{|A\cap B|}{|A|}\times 100\%, \quad (15)$$

$$F\text{-measure}=\frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}\times 100\%。 \quad (16)$$

假设 $D$ 是待消歧专家的论文集合, 在式(14)和(15)中, 集合 $A$ 表示 $D$ 中与特定专家实体相关的真实论文簇, 集合 $B$ 表示通过同名消歧方法得到的特定专家实体的论文簇。

### 3.2 结果与分析

无论是基于单特征的同名消歧方法, 还是基于多特征融合的同名消歧方法, 都采用 2.4 节中的 AP 算法进行聚类, 不同之处在于 AP 算法的输入——相似度矩阵。对于单特征的同名消歧方法, 相似度矩阵只含有该单一特征的相似度, 而对于多特征融合的同名消歧方法, 相似度矩阵包含所有 5 种特征的相似度。

我们对表 2 中的每一类属性特征进行消歧测试, 得到的结果如图 3 所示。可以看出, 本文提出的基于多特征融合的同名专家消歧方法的消歧效果比任何单特征消歧的方法都好, 准确率、召回率和 F1 值都有较明显的提升, 从而验证了基于多特征融合的消歧方法的有效性。此外, 论文的标题、摘要、关键词、作者单位和合作者这 5 种特征对同名消歧的效果依次增强, 证明我们关于各个属性特征对于同名专家区分力强弱的分析是正确的。

## 4 结语

针对目前专家库构建过程中出现的同名歧义现象, 本文提出一种基于多特征融合的同名专家消歧方法。从 CNKI 等数据源中获取专家的论文信息, 抽取论文的标题、摘要、关键词、作者单位和合作者等关键信息作为属性特征, 构建特征表示模型, 进而定义同名专家之间的相似度计算函数, 根据计算得到的相似度, 将同名消歧问题转化为聚类问题, 再利用近邻传播聚类算法进行聚类, 解决了同名消歧问题。在专家论文数据上的实验表明, 与利用单特征进行消歧的其他方法相比, 基于多特征融合的同名消岐方法的效果有明显提升。我们目前主要针对中文姓名进行研究, 下一步将考虑含有拼音和英文的同名专家消歧问题。

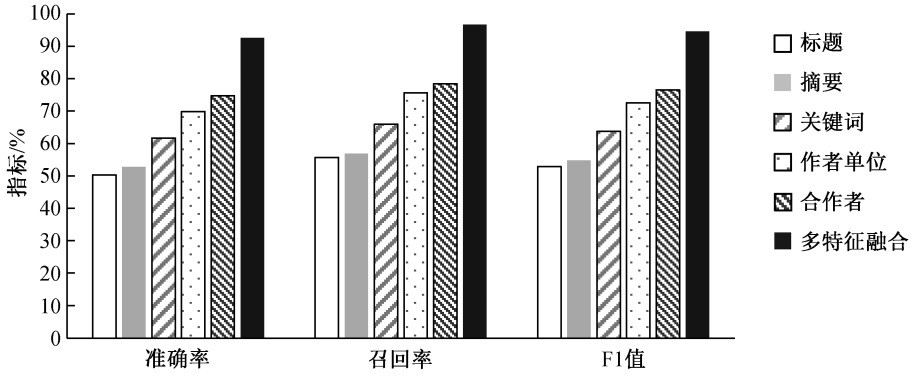


**图 3 测试结果**

Fig. 3 Test results

## 参考文献


[1] Nelson M. About JCDL [EB/OL]. (2013) [2019–01–01]. http:// www.jcdl.org/index.php

[2] Artiles J, Gonzalo J, Sekine S. The SemEval-2007 WePS evaluation: establishing a benchmark for the web people search task // Proceedings of the 4th International Workshop on Semantic Evaluations. Association for Computational Linguistics. Prague, 2007: 64–69

[3] 王厚峰, 李素建. 汉语命名实体识别与歧义消解 [EB/OL]. (2012–02–08)[2019–01–01]. http://www.cipsc.org.cn/clp2012/task2-cn.html

[4] Bagga A. Coreference, cross-document coreference, and information extraction methodologies. Durham, NC: Duke University, 1998

[5] 辛涛, 程绍银, 蒋凡. 基于组合特征的Web人名消歧方法. 计算机系统应用, 2015, 24(11): 162–166

[6] 陈晨, 王厚峰. 基于社会网络的跨文本同名消歧. 中文信息学报, 2011, 25(5): 75–83

[7] Wang L, Li S, Wong D F, et al. A joint Chinese named entity recognition and disambiguation system // Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing. Tianjin, 2012: 146–151

[8] 宁博, 张菲菲. 基于异构知识库的命名实体消歧. 西安邮电大学学报, 2014, 19(4): 70–76

[9] 谭捷, 张李义, 饶丽君. 中文学术期刊数据库的比较研究. 图书情报知识, 2010(4): 4–13

[10] Mann G S, Yarowsky D. Unsupervised personal name disambiguation // Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. Association for Computational Linguistics. Edmonton, 2003: 33–40

[11] Goldberg Y, Levy O. Word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method [EB/OL]. (2014–02–15)[2019–01–01]. https://arxiv.org/abs/1402.3722

[12] 武永亮, 赵书良, 李长镜, 等. 基于 TF-IDF 和余弦相似度的文本分类方法. 中文信息学报, 2017, 31(5): 138–145

[13] Ristad E S, Yianilos P N. Learning string-edit distance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(5): 522–532

[14] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究. 软件学报, 2008, 19(1): 48–61

[15] Johnson S C. Hierarchical clustering schemes. Psychometrika, 1967, 32(3): 241–254

[16] Hartigan J A, Wong M A. Algorithm AS 136: a k-means clustering algorithm. Journal of the Royal Statistical Society: Series C (Applied Statistics), 1979, 28(1): 100–108

[17] Frey B J, Dueck D. Clustering by passing messages between data points. Science, 2007, 315: 972–976