ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31,No. 1,Jan. 2016,pp. 213-222
DOI:10.16337/j.1004-9037.2016.01.024

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:+86-025-84892742

基于分步聚类的人名消歧算法

阳怡林　周　杰　李弼程　席耀一

(解放军信息工程大学信息系统工程学院,郑州,450001)

摘　要:针对知识库中存在单条实体定义特征稀疏和人工设置相似度阈值适用性不强的问题,本文提出了一种基于分步聚类的人名消歧算法。首先,将知识库中人名实体定义的人物属性特征作为查询特征,利用文本检索的方式实现基于知识库的初次聚类,弥补了知识库中单条实体定义中特征稀疏的问题;然后,利用初次聚类的结果,采用基于自适应阈值的凝聚层次聚类算法实现知识库人名消歧;最后,采用条件随机场进行 Other 类识别,利用基于自适应阈值的凝聚层次聚类完成 S 类聚类,从而实现非知识库人名消歧。在 CLP2012 的中文人名消歧评测语料上进行实验,结果表明本文的算法能够有效地对人名进行消歧。

关键词:人名消歧;特征稀疏;文本检索;凝聚层次聚类;相似度阈值

中图分类号:TP391　　**文献标志码**:A

Name Disambiguation Based on Clustering by Step

Yang Yilin, Zhou Jie, Li Bicheng, Xi Yaoyi

(Institute of Information and System Engineering, PLA Information Engineering University, Zhengzhou, 450001, China)

Abstract: In the knowledge base there exist characteristics of sparse for a single entity, and it is difficult to determine the similarity threshold of clustering. Therefore, this paper presents a name disambiguation algorithm based on cluster by step. Firstly, query features for character attribute are obtained from knowledge base, and the initial clustering based on knowledge base is carried out by text retrieval, which make up characteristics of sparse for a single entity name defined in knowledge base. Then, taking initial clustering results as input, name disambiguation in knowledge base is completed by using hierarchical clustering algorithm based on adaptive threshold. Finally, the other classes are identified by conditional random fields, and the cluster by using hierarchical clustering algorithm based on adaptive threshold is completed. The experiment on data of CLP2012 Chinese person name disambiguation results shows that the proposed algorithm can effectively achieve disambiguation names.

Key words: name disambiguation; characteristics of sparse; text retrieval; hierarchical clustering; similarity threshold

引　　言

随着信息检索技术的日渐成熟,人们可以借助搜索引擎搜索人物信息。然而,现实生活中,不同的

基金项目:国家社会科学基金(14BXW028)资助项目;全军军事研究生课题(2011JY002k-158)资助项目。
收稿日期:2014-06-05;**修订日期**:2014-11-18

人共享同一个人名的情况普遍存在,这种现象称为人名歧义。人名歧义给人物搜索带来了众多不利影响。例如,利用 Google 检索"高峰"的信息时,搜索结果"高峰"会指向几十个不同的人名实体,如相声演员、上海交通大学教授、足球运动员以及广告摄影师。当用户想获取某个特定含义的"高峰"所在页面信息时,需要浏览大量无关的网页,严重影响用户检阅的效率。因此,研究人名消歧技术对人名检索、人物关系挖掘和重点人物舆情分析[1]等具有重要意义。

当前,针对人名消歧的评测会议主要有网页人名检索会议(Web people search, WePS)和中文处理国际会议(Joint conference on Chinese language processing, CLP)。2007~2010 年,连续三届 WePS 均包含了人名消歧的评测任务;2010 年第一届中文处理国际会议(The 1st joint conference on Chinese language processing, CLP2010)[2]首次引入中文人名消歧的任务。以上人名消歧的任务均是将包含同一个人名的网页或文本集合划分成指向同一人名实体的多个类,每个类中描述的仅是同一个人名实体而没有映射到现实中具体的人名实体。2012 年 CLP2012[3]在 CLP2010 人名消歧的基础上,增加了把人名消歧的结果映射到现实中具体人名实体的任务。CLP2012 人名消歧评测语料中包含知识库(Knowledge base, KB)和未标记文本集。在评测任务[3]中,针对每个待消歧人名(假定是"高峰"),提供一个知识库和未标记文本集 T;其中,知识库"高峰_KB"包含共用名字"高峰"的 m 条描述人名实体的定义;未标记文本集中,每个文本 $t \in T$ 均含有词"高峰"。评测任务要求判断 t 中的"高峰"对应于知识库中的哪条定义。当然,如果未标记文本所描述"高峰"为人名实体但在知识库"高峰_KB"没有该实体的定义,则把该类未标记文本归入集合 $S \subset T$。对于 S,还需要按照人名(如"高峰")的指称进一步划分,设划分结果为 Out_XX,其中 XX 为编号,依次为 Out_01, Out_02, …。此外,如果未标记文本所描述"高峰"仅仅是普通词而不是人名实体,就将该类未标记文本归入 Other 类中。本文将描述的待消歧人名属于知识库中人名实体的未标记文本称为"知识库文本",它所包含的待消歧人名称为"知识库人名";将属于 Other 类和 S 类的未标记文本称为"非知识库文本",它所包含的待消歧人名称为"非知识库人名"。知识库文本中,有些文本和与之对应的知识库人名实体定义没有共同的特征,本文将这部分文本称为"第 II 类知识库文本",剩余的知识库文本称为"第 I 类知识库文本"。

当前对人名消歧的研究,主要是把知识库中的每条实体定义当作一篇文本,直接计算它们和未标记文本之间的相似度,实现未标记文本到知识库人名实体的映射。然而,知识库存在单条人名实体定义特征稀疏、覆盖面不全的问题,导致第 II 类知识库文本不能映射到知识库中相应的人名实体,从而使其召回率较低。在人名消歧中,聚类算法通常需要人工预先设定的相似度阈值,使得其不适用于实际情况。针对以上问题,本文提出了一种基于分步聚类的人名消歧算法。

1 相关工作

人名消歧的方法主要分为基于向量空间模型(Vector space model, VSM)的方法和基于社会网络的方法。基于向量空间模型的方法主要是利用人名所在文本上下文之间的相似度来区分文本,采用聚类的方法来实现人名消歧。2008 年,文献[4]利用命名实体的共指特征、关键词特征以及潜在的主题信息特征等实现人名消歧;2011 年 Long 等[5]抽取命名实体特征,并利用特征与检索词间的距离对特征进行加权;2012 年杨欣欣等[6]抽取与网页文本中人名关键字实体相关的依存特征来实现人名消歧。基于社会网络的方法[7-9]主要利用人名实体来构建社会网络,采用图分割的方法实现人名消歧。

针对传统人名消歧不能将消歧的结果映射到现实中的具体人名实体的问题,研究者提出许多有效的方法。Zehuan 等[10]利用训练好的分类器把未标记文本分为知识库文本、S, Other 类,然后对前两类采用凝聚层次聚类的方法进行聚类,从而实现人名消歧;Zong 等[11]重点考虑区分度强的词(别名、人名等),把它们和上下文特征进行一定的融合;Wei 等[12]抽取人物属性特征,然后利用聚类的方法实现人名消歧。李广一等[13]把知识库当作一篇文本,通过训练语料获取 3 个相似度阈值,采用 4 步聚类实现人名消歧。

在传统的人名消歧中,基于 VSM 的方法需要根据人工经验知识或训练语料预先设置聚类的相似度阈值,然而凝聚层次聚类的性能对相似度比较敏感,使得其不适用于实际情况。基于社会网络的方法仅利用实体特征,忽略了大量特征,而且需要预先人工设置图划分的终止条件,适用性不强。目前人名消歧主要是把知识库中单条实体定义当作一篇文本,但知识库单条实体定义简短精炼特征稀疏、覆盖面不全,而未标记文本中内容丰富、冗余特征多,直接计算它们间的相似度将导致部分第 II 知识库文本不能正确映射到相应的实体定义。针对以上问题,本文利用知识库实体中区分度强的特征(人物属性)作为查询特征,利用文本检索方法对知识库中单条实体定义进行扩展,通过分析文本间相似度矩阵的分布自动获取相似度阈值,提出了基于分步聚类的人名消歧算法。

2 研究方法

基于分步聚类的人名消歧算法流程如图 1 所示。第 1 步为知识库人名消歧,包括为基于知识库的初次聚类和基于自适应阈值的凝聚层次聚类。第 2 步为非知识库人名消歧,包括 Other 类识别和 S 类聚类。

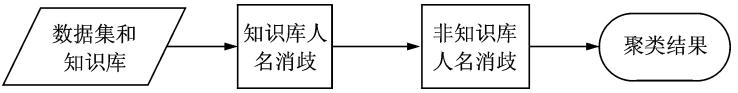

图 1 基于分步聚类的人名消歧算法流程图

Fig. 1 Flow chart of names disambiguation algorithm based on clustering by step

2.1 知识库人名消歧

在未标记文本中,知识库文本所占比重较大,它的聚类性能直接影响整个算法的性能,然而,知识库中单条实体定义特征稀疏、覆盖面不全。现有的方法主要将知识库中的每条实体定义当作一篇文本,直接计算它们和未标记文本之间的相似度导致第 II 类知识库文本不能映射到知识库中的人名实体。凝聚层次聚类需要预先设置相似度阈值,而现有的方法相似度阈值的设置主要依懒于人工经验或训练语料,适用性不强。

本文充分利用知识库中实体定义特征区分性强的特点,借鉴文本检索的思想实现未标记文本到知识库中实体定义的初次映射,从而实现对知识库中实体定义的扩展,弥补知识库中单条实体定义特征稀疏、覆盖面不全的问题。针对相似阈值需要人工预先设定,本文通过分析文本间相似度矩阵的分布自动获取相似度阈值。

2.1.1 基于知识库的初次聚类

知识库中实体定义的特征区分度强,基于此,本文设计了基于知识库的初次聚类。首先从知识库的实体定义中抽取部分人物属性特征作为查询特征,把未标记文本集合作为待检索文本集合,然后通过文本检索的方式实现未标记文本到知识库中实体定义的初次映射。

对于某一查询特征,利用文本检索的方式从未标记文本集合中查询得到包含该查询特征的子集合。由于不同人物通常具有不同的属性特征,因此每一子集合内的文本相似度一般大于子集合间文本的相似度,即各子集合分别对应了不同的类别。同时,由于每一子集合分别对应一个查询特征,使得每一子集合与知识库中的人名实体实现了一一对应,从而实现了未标记文本到知识库中实体定义的初次映射。初次映射后,把知识库中每条实体定义以及与之一一对应的子集合当作一个类,从而实现知识库人名实体特征的扩展。

基于知识库的初次聚类的关键是构建查询特征表,查询特征要简短且区分度高。本文查询特征的构建主要基于以下两个假设:(1) 具有相同人名(别名、昵称和社会关系)或者相同作品的同名人指向同一人名实体。(2) 在假设(1)不满足的情况下,具有相同的毕业院校(学校和专业)或者相同职业(单位和职业)

的同名人指向同一人名实体。大部分未标记文本和知识库的实体定义中都包含人名、作品名、职业和院校(具体关键特征分布情况如表 1),说明人名、作品名、职业和院校等人物属性特征具有普遍性。

表 1 人物属性特征分布情况

Tab. 1 Distribution of the character attributes

类型	未标记文本集(5 503)		知识库的人名实体定义(473)	
	文本数	所占百分比/%	文本数	所占百分比/%
人名	4354	79.12	133	28.12
作品	2049	37.23	97	20.51
职业	2525	45.88	293	61.95
毕业院校	1628	29.58	265	56.02

未标记文本中不满足假设(1,2)文本所占文本的比例较少(具体情况如表 2 所示),说明本文假设(1,2)在错误率允许的范围内成立。

表 2 不满足假设(1,2)的文本分布情况

Tab. 2 Text distribution without satisfied assumptions (1,2)

类 型	未标记文本集(5503)	
	文本数	所占百分比/%
含相同的人名但指向不同实体	17	0.31
含相同的作品但指向不同实体	21	0.38
含相同的职业但指向不同实体	54	0.98
含相同的院校但指向不同实体	31	0.56

基于假设(1,2),从知识库的实体定义中抽取人物属性特征,构建查询特征表,如表 3 所示。

表 3 查询特征

Tab. 3 Query feature

级别	人物属性特征
1 级	人名(别名、昵称、社会关系)、作品名
2 级	职业(单位和职业)、院校(学校和专业)
3 级	其他命名实体(地名、机构名、专有名词等)

根据查询特征区分度的大小,将查询特征分为 3 级:第 1 级为人名(别名、昵称)、作品名;第 2 级为职业和毕业院校;第 3 级为其他命名实体(地名、机构名和专有名词等)。其中,第 1 级查询特征的区分度最高,第 2,3 级查询特征的区分度依次降低。查询特征区分度越高,检索结果越准确。为了保证文本检索结果具有较高正确率,第 1 级、第 2 级查询结果只要非空,将直接跳出本条实体定义的查询,由于第三级查询特征的区分度较低,所以只取检索结果中排名第一的文本作为返回结果。具体步骤如下:

步骤 1 用 1 级查询特征进行或查询,如果返回结果不为空,则跳到步骤 4;

步骤 2 用二级查询特征进行或查询,如果返回结果不为空,则跳到步骤 4;

步骤 3 从三级查询特征中依次选择三个查询特征进行与查询,如果返回结果不为空,则选取排名第一的未标记文本作为返回结果;

步骤 4 把返回的文本集映射到知识库中相应的人名实体;

步骤 5 对知识库中的每条实体定义重复步骤 1~4;

步骤 6 取消重复文本(被知识库中 1 个以上人名实体定义指向的未标记文本)到知识库实体的映射。

2.1.2 基于自适应阈值的凝聚层次聚类

由于知识库的单条实体定义特征稀疏、覆盖不全面,导致第2类知识库文本不能通过初次聚类与知识库中实体定义形成对应关系。考虑到同一人名实体所处的上下文语义环境相似的可能性较大,不同实体所处的上下文语义环境相似的可能性较低,例如,描述足球运动员"高峰"的文本用语与描述上海交通大学教授"高峰"的文本用语相同的可能性较小,因此本文通过凝聚层次聚类的方法对初次聚类的结果进行二次聚类。然而凝聚层次聚类需要设置相似度阈值,现有的相似度阈值设置主要依懒于人工经验或训练语料,适用性不强。

本次聚类充分利用初次聚类的结果来扩展知识库中单条实体定义的特征。如果知识库"高峰_KB"包含 m 条实体定义,则初次聚类后,m 条实体定义和与之聚到一起的未标记文本构成 m 类文本,再加上未标记文本中除去映射到知识库的文本所剩下 n 篇文本,总共构成 $m+n$ 类文本;然后从每类文本中选取特征,计算文本类之间的相似度;最后通过基于自适应阈值的凝聚层次聚类方法实现知识库人名消歧。

(1) 特征选择与相似度计算

命名实体一直是人名消歧的重要特征,Chen 等[14]通过组合使用4种文本特征进行人名消歧时发现,仅使用命名实体便可让整体的 F 值达到75%。本文具体特征如表4所示。

表4 本文具体特征

Tab. 4 Specific features of the paper

命名实体	人名、地名、机构名以及其他专名等
作品名	书名、电影名和曲名等
职业名	通过互联网构建常用的职业词典
其他特征	名词、动名词、形容词名词等

从待消歧人名的上下文本中抽取特征,虽然能避免抽取许多无关的特征,但会导致部分有效的特征被丢弃。尤其是当文本中存在待消歧人名的指示代词时,部分本是描述待消歧人名的特征将被丢失。本文特征加权基本原则是:距离待消歧人名近的特征与待消歧人名的相关性要比距离待消歧人名远的特征的相关性要大。Long 等[5]利用下式来衡量待消歧人名距离对特征权重的影响

$$f(u)=\left(1-\frac{d(u)}{d_{\max}}\right)^2 \tag{1}$$

式中:$d(u)$ 为待消歧人名和特征 u 句间距离的最小值;$d_{\max}$ 为 $d(u)$ 允许的最大值,若待消岐人名和 u 出现在同一句子中,则 $d(u)=0$;若 $d(u)\geqslant d_{\max}$,则 $f(u)=0$。本文在式(1)的基础上结合词频、逆文本频率和词类型等信息,作出以下改进

$$W(u)=\mathrm{IDF}(u)\times\sum_{u\in d}f(u)\times\mathrm{TYPE}(u) \tag{2}$$

式中:$W(u)$ 为特征权重;$\mathrm{IDF}(u)$ 为逆文本频率;$\sum_{u\in d}f(u)$ 为包含特征的词频信息以及与待消歧人名的距离信息;$\mathrm{TYPE}(u)$ 为特征类型的权重,不同类型的特征对待消歧人名的区分度不一样的。本文根据特征的类型设置权重,具体如表5所示。

表5 特征权重

Tab. 5 Weight of features

特征类型	人名	作品名	职业名	机构名	地名	其他专有名词	其他词
权重	2.0	2.0	1.5	1.3	1.3	1.1	1.0

通过把每篇文本转化为一个特征向量,利用余弦相似度来计算文本之间的相似度,即

$$Sim(d_i,d_j)=\frac{\sum_{k=0}^{n}d_{ik}d_{jk}}{\sqrt{\sum_{k=0}^{n}d_{ik}^2}\sqrt{\sum_{k=0}^{n}d_{jk}^2}} \tag{3}$$

式中：d_i，d_j 为文本的向量化表示；$Sim(d_i,d_j)$ 为文本 d_i，d_j 之间的相似度。

(2) 凝聚层次聚类

凝聚层次聚类一种全局最优的聚类算法，每次从当前类中选择相似度最高的两个类，如果该两类之间的相似度大于预先设定的阈值，则把它们合并为一个类，否则结束聚类。具体步骤如下：

步骤 1　把初次聚类后知识库中的每条实体定义以及与之映射到一起的未标记文本、剩下的未标记文本集中的每篇文本分别看作一类构成文本类集合 $C=\{C_1,C_2,\cdots,C_i,\cdots,C_n\}$。

步骤 2　计算当前文本类集合中两两之间的相似度，选择相似度最大的两个类 C_i，C_j。

步骤 3　如果类 C_i 和 C_j 之间的相似度 $Sim(C_i,C_j)$ 大于预先设定的相似度阈值，则将相似度最高的 C_i 和 C_j 合并为 C_{ij}，把 C_{ij} 添加到 C，同时去除 C_i 和 C_j，否则跳到步骤 5。

步骤 4　如果类集合中类的个数大于 1，则重复步骤 2，步骤 3。

步骤 5　结束聚类。

计算类 C_i 和 C_j 之间的相似度 $Sim(C_i,C_j)$ 时，将知识库中的每条实体定义当作一篇文本，具体步骤如下：

步骤 1　为了避免一个类中含有两条不同实体定义，若两个类中均含有知识库的实体定义，则该两类之间的相似度为 0。

步骤 2　为了弥补知识库单条实体定义特征稀疏的不足，若两类中仅有一个类 C_i 含知识库中的实体定义，则分别计算 d_{Eid}，D_i 和 C_j 类的相似度，然后取这两个相似度的最大值作为 C_i 和 C_j 的相似度。

步骤 3　若两个类中都不含有知识库的实体定义，则直接计算 C_i 和 C_j 之间的相似度，即有

$$Sim(C_i,C_j)=\begin{cases}0 & C_i,C_j \text{ 包含知识库的实体定义}\\ \max\left(\frac{1}{|D_i||C_j|}\sum_{d_x\in D_i}\sum_{d_y\in C_j}Sim(d_x,d_y),\frac{1}{|C_j|}\sum_{d_y\in C_j}Sim(d_{Eid},d_y)\right) & C_i \text{ 包含知识库的实体定义，} C_j \text{ 不包含知识库文档}\\ \frac{1}{|C_i||C_j|}\sum_{d_x\in C_i}\sum_{d_y\in C_j}Sim(d_x,d_y) & C_i,C_j \text{ 都不含知识库的实体定义}\end{cases} \tag{4}$$

其中如果类 C_i 包含知识库中的实体定义，则 C_i 由知识库中实体定义 d_{Eid} 和与之映射到一起未标记文本子集合 $D_i=\{d_{i1},d_{i2},\cdots,d_{ij},\cdots,d_{in}\}$ 组成。$Sim(d_{Eid},d_y)$ 为知识库的实体定义 d_{Eid} 和未标记文本 d_y 的相似度，$Sim(d_x,d_y)$ 为未标记文本之间的相似度。

(3) 自适应阈值计算

凝聚层次聚类需要预先设定相似度阈值，且该阈值决定聚类的结束时刻。Artiles 等[15]指出，凝聚层次聚类的性能比较依懒于相似度阈值，而且不同的待消歧人名对应的最佳相似度阈值往往不同。针对该问题，本文利用文本类之间相似矩阵的分布情况自动获取相似度阈值。首先，通过计算文献[16]两两文本类之间的相似度获得该文本类集合的相似度矩阵；然后，利用曲线拟合的方法拟合出相似度分布曲线，最后搜索曲线的拐点，把拐点对应的相似度作为聚类的相似度阈值。

计算任意两个类之间的相似度得到相似度矩阵

$$\boldsymbol{S}_M=\begin{bmatrix}S_{11} & S_{21} & \cdots & S_{n1}\\ S_{12} & S_{22} & \cdots & S_{n2}\\ \vdots & \vdots & \vdots & \vdots\\ S_{1n} & S_{2n} & \cdots & S_{nn}\end{bmatrix} \tag{5}$$

式中 S_{ij} 表示类 C_i 和 C_j 之间的相似度。将该矩阵中所有样本构成的分布作为该文本类集合的相似度分布。以 x 轴表示相似度,y 轴表示相似度对应的类的对数,画出 y 随 x 的变化曲线作为相似度分布曲线。如图 2 所示,待消歧人名“吉祥”大部分文本对的相似度值在[0,0.1]之间,待消歧人名(除华明85%外)相似度值在[0,0.1]范围内的类的对数占到了文本类对总数的 90%以上,直接在[0,1]范围搜索相似度阈值时,由于曲线中含大量的奇异点,很难获得有效的拐点,所以将拐点的搜索范围限制在[0,0.1]。

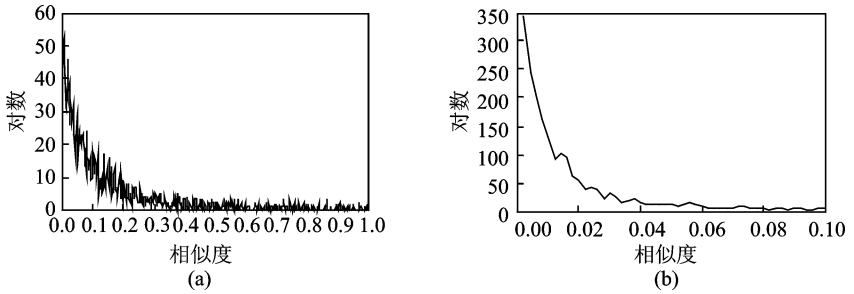

图 2 待消歧人名“吉祥”的相似度分布曲线

Fig. 2 Similarity distribution curve of "Ji Xiang"

虽然限定相似度的范围提高了搜索拐点的准确度,相似度分布曲线仍不平滑,如果直接搜索仍难准确地定位有效的拐点。为了准确地定位拐点,先利用平滑函数对相似度分布曲线进行平滑,在保留曲线变化趋势的情况下去除无效的拐点。曲线平滑后,采用基于最小二乘法的曲线拟合算法搜索曲线的拐点:利用 m 次多项式 $Y(x) = a_m x^m + a_{m-1}x^{m-1} + \cdots + a_0$ 拟合相似度分布曲线,其中 $a_m, a_{m-1}, \cdots, a_0$ 用下式求解

$$\begin{bmatrix} \sum_{i=1}^{n} 1 & \sum_{i=1}^{n} x_i & \cdots & \sum_{i=1}^{n} x_i^m \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \cdots & \sum_{i=1}^{n} x_i^{m+1} \\ \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^{n} x_i^m & \sum_{i=1}^{n} x_i^{m+1} & \cdots & \sum_{i=1}^{n} x_i^{2m} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ \vdots \\ \sum_{i=1}^{n} x_i^m y_i \end{bmatrix} \tag{6}$$

式中:(x_i, y_i) 表示相似度分布曲线上的点,n 为点的数量。本文采用三次多项式曲线拟合算法得到 $Y(x)$,然后通过解方程 $Y(x)' = 0$ 得到曲线的一阶拐点,把较小的拐点对应的相似度作为相似度阈值。

2.2 非知识库人名消歧

经过知识库人名消歧后,未标记文本中只剩下非知识库文本,它们所占的比重较小,非知识库人名消歧分为 Other 类识别和 S 类聚类。

评测任务中的待消歧人名大多在汉语中通常作为普通词,命名实体识别算法对该类人名的识别效果一般,且 Other 类文本数占总文本的比例较小。如果先进行 Other 类识别,将导致大量实体被误判为 Other 类。所以,本文首先进行知识库人名消歧,然后进行 Other 类识别,最后进行 S 类聚类。本文采用条件随机场(Conditional random field,CRF)来识别 Other 类。CRF 的训练语料选用北京大学的人民日报命名实体识别的训练数据集,选择词和词性作为特征;分词和词性标注采用北京理工大学张华平的 NLPIR 汉语分词算法。经过知识库中人名聚类消歧和 Other 类识别后,剩下部分就是 S 类,对该部分文本,本文利用 2.1.2 节的基于自适应阈值的凝聚层次聚类算法实现。

3 实验结果及分析

3.1 实验语料与评测标准

CLP2012[3]提供了人名消歧任务的测试语料，该语料总共包含 32 个待消歧人名，每个待消歧人名对应一个未标记文本集合和知识库，其中未标记文本集共 5 503 篇文本。与每个待消歧人名对应的未标记文本集的大小从 47 到 502 之间不等，其中属于知识库文本占主要部分，知识库文本消歧的效果直接影响整体的性能；非知识库文本的比重较小。本文以该语料作为实验语料。本文采用 CLP2012 会议提供的评测标准，整个实验数据的综合评价的定义如下

$$\mathrm{Pre}=\frac{\sum_{n}\mathrm{Pre}(n)}{|N|} \tag{7}$$

$$\mathrm{Rec}=\frac{\sum_{n}\mathrm{Rec}(n)}{|N|} \tag{8}$$

$$F=\frac{2*\mathrm{Pre}*\mathrm{Rec}}{\mathrm{Pre}+\mathrm{Rec}} \tag{9}$$

式中：N 为待消歧人名的集合，$n\in\mathbf{N}$；$\mathrm{Pre}(n)$ 为单个人名的正确率；$\mathrm{Rec}(n)$ 为单个人名的召回率，详细的计算方法参考文献[3]。

3.2 实验结果分析

在未标记文本中知识库文本占主要部分，它的实验结果直接影响算法的整体性能，此外凝聚层次聚类算法的性能比较依懒于相似度阈值[15]，而且不同的待消歧人名对应的最佳相似度阈值往往不同。基于此，本文实验结果分析包括知识库人名消歧性能分析、相似度阈值影响分析以及综合性能分析。

3.2.1 知识库人名消歧性能分析

如表 6 所示，初次聚类的准确率(Pre)比较高(0.979 0)，表明基于知识库的初次聚类是有效的，能较准确地区分不同的人名实体，但知识库中单条人名实体定义所含有的属性特征较少，导致第 II 知识库文本不能通过初次聚类与知识库中的实体定义形成对应关系。导致召回率较低(0.668 1)。采用基于自适应阈值的凝聚层次聚类后，召回率(Rec)有较大的提高(0.945 6)，这是因为经过初次聚类，知识库中实体定义的特征得到了丰富，使第 II 类知识库文本能够映射到知识库中相应的实体定义。经过聚类后，虽然牺牲了一定的准确率(Pre)，但是召回率(Rec)提高明显，从而提高了 F 值。

表 6 知识库人名消歧结果

Tab. 6 Results of names disambiguation in knowledge base

方法	准确率(Pre)	召回率(Rec)	F 值
初次聚类	0.979 0	0.668 1	0.794 2
两次聚类	0.843 3	0.945 6	0.891 5

3.2.2 相似度阈值影响分析

在[0,0.1]范围内等间隔选取 10 个值和本文自动获取的相似度阈值分别进行实验。每个待消歧人名最高 F 值、最低 F 值和采用自动获取相似度阈值所得 F 值如图 3 所示。相似度阈值对 F 值的影响较大，同一待消歧人名，最高的 F 值和最低的 F 值相差较大(如：高雄 24.75%)；同时本文采用自动获取相似度阈值的方法能较准确地逼近甚至等于最高 F 值，说明本文算法能较准确地逼近最佳相似度阈值。

如图 4 所示，当相似度阈值为 0.04 时，F 值达到了最大值 0.862 7；而本文采用自动设置相似度阈值方法取得的 F 值为 0.861 5，仅比人工设定阈值最好的结果低 0.12%，这进一步说明本文提出的自适应阈值获取算法能较准确地逼近最佳相似度阈值。

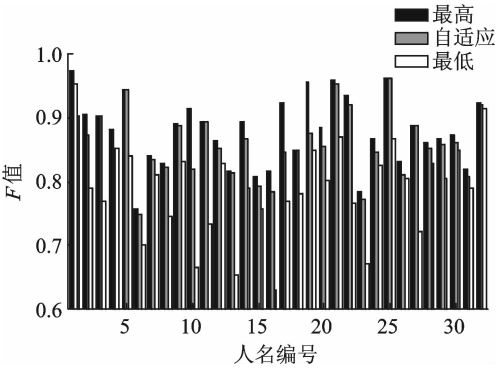

图 3 待消歧人名 F 值

Fig. 3 F of pre-prepared disambiguation names

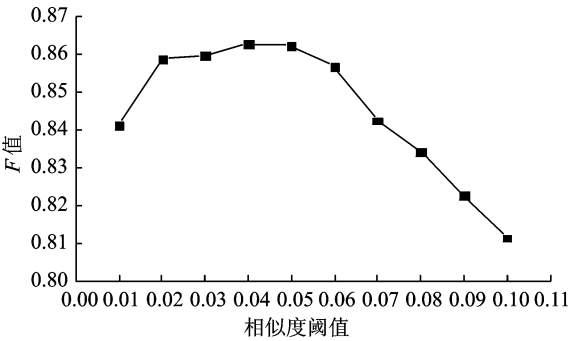

图 4 不同相似度阈值下的 F 值

Fig. 4 F in different similarity threshold

3.2.3 综合性能分析

本文选取 CLP2012 中人名消歧评测任务中前 3 名的方法[10-12]进行对比实验。其中,Zehuan 等[10]把知识库中的所有实体定义当作一篇文本,首先利用训练好的分类器把未标记文本分为知识库文本、S、Other 类,然后对前两类采用凝聚层次聚类的方法进行聚类,从而实现人名消歧;Zong 等[11]把知识库中的每条实体定义当作一篇文本,采用关键词抽取算法抽取其中的关键词作为特征;Wei 等[12]把知识库中的每条实体定义当作一篇文本,从每条实体定义和未标记文本中抽取 19 个人物属性特征形成特征向量。文献[10]把知识库中所有的实体定义当作一篇文本,一定程度上解决了知识库中单条实体定义特征稀疏的问题,其实验性能明显优于文献[11,12]。本文算法利用文本检索的方法来弥补知识库单条实体定义特征稀疏的不足,在没有利用训练语料的情况下 F 值比 CLP2012[2]人名消歧任务中评测的第一名高 5.93%,说明本文算法能够有效地对人名进行消歧。详细实验结果如表 7 所示。

表 7 本文算法与其他算法对比实验结果

Tab. 7 Comparison of results of different algorithms

算法	准确率(Pre)	召回率(Rec)	F 值
本文算法	0.821 9	0.905 2	0.861 5
文献[10]	0.794 8	0.809 8	0.802 2
文献[11]	0.725 6	0.792 3	0.757 5
文献[12]	0.671 8	0.856 2	0.752 9

4 结束语

目前人名消歧的方法主要把知识库中单条实体定义当作一篇文本,然而知识库中单条实体定义特征稀疏、覆盖面不全,直接计算实体定义与未标记文本的相似度将导致部分第 II 知识库文本不能正确映射到相应的实体定义。凝聚层次聚类中相似度阈值常常依懒于人工经验或训练语料。针对以上问题,本文提出基于分步聚类的人名消歧算法。第 1 步为知识库人名消歧:首先,利用知识库获取人物属性特征作为查询特征,通过文本检索的方式丰富知识库中实体定义的特征;然后利用文本类之间相似度矩阵的分布情况自动获取聚类的相似度阈值;最后利用基于自适应阈值的凝聚层次聚类算法完成知识库人名消歧。第 2 步非知识库人名消歧,首先利用 CRF 对第一步中未被映射到知识库的未标记文本进行 Other 类识别,然后再利用基于自适应阈值的凝聚层次聚类完成 S 类的聚类。下一步工作,主要是通过引入外部知识,深层次挖掘文本的语义信息,从而进一步提高人名消歧的性能。

参考文献：

[1] 周耀明，李弼程. 一种自适应网络舆情演化建模方法 [J]. 数据采集与处理，2013，28(1)：69-76.
Zhou Yaoming, Li Bicheng. An adaptive evolution modeling method of Internet public opinions [J]. Journal of Data Acquisition and Processing, 2013, 28(1): 69-76.

[2] Chen Ying, Jin Peng, Li Wenjie, et al. Exploration of personal name disambiguation in Chinese news [C]// CIPS-SIGHAN Joint Conference on Chinese Language Processing. Bejing, China: ACL, 2010: 20-26.

[3] He Zhengyan, Wang Houfeng, Li Sujian. The task 2 of CIPS-SIGHAN 2012 named entity recognition and disambiguation in Chinese bakeoff [C]// CIPS-SIGHAN Joint Conference on Chinese Language Processing. Tianjin, China: ACL, 2012: 108-114.

[4] Ono S, Sato I, Yoshida M, et al. Person name disambiguation in web pages using social network, compound words and latent topics [C]// Advances in Knowledge Discovery and Data Mining. [S. l.]: Springer Berlin Heidelberg: 2008: 260-271.

[5] Long C, Shi L. Web person name disambiguation by relevance weighting of extended feature sets[C]// 11th Workshop of the Cross-Language Evaluation Forum. Padua: ACL, 2010: 1-13.

[6] 杨欣欣，李培峰，朱巧明. 基于网页文本依存特征的人名消歧[J]. 计算机工程，2012，38(19)：133-136.
Yang Xinxin, Li Peifeng, Zhu Qiaoming. Name disambiguation based on dependency feature in web page text [J]. Computer Engineering, 2012, 38(19): 133-136.

[7] Fan Xiaoming, Wang Jianyong, Pu Xu, et al. On graph-based name disambiguation [J]. Journal of Data and Information Quality, 2011, 2(2): 1-23.

[8] 郎君，秦兵，宋巍，等. 基于社会网络的人名检索结果重名消解[J]. 计算机学报，2009，32(7)：1365-1374.
Lang Jun, Qin Bing, Song Wei, et al. Person name disambiguation of searching results using social network[J]. Chinese Journal of Computer, 2009, 32(7): 1365-1374.

[9] 陈晨，王厚峰. 基于社会网络的跨文本同名消歧[J]. 中文信息学报，2011，25(5)：75-82.
Chen Chen, Wang Houfeng. Social network based cross-document personal name disambiguation [J]. Journal of Chinese Information Processing, 2011, 25(5): 75-82.

[10] Peng Zehuan, Sun Le, Han Xianpei. A Chinese named entity recognition and disambiguation system using a two-stage Method[C]// The 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing. Tianjin, China: ACL, 2012: 115-120.

[11] Hao Zong, Wong Derek F, Chao Lidia S. A template based hybrid model for Chinese personal name disambiguation[C]// The 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing. Tianjin, China: ACL , 2012: 121-126.

[12] Han Wei, Liu Guang, Mao Yuzhao, et. al. Attribute based Chinese named entity recognition and disambiguation[C]// The 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing. Tianjin, China: ACL , 2012: 127-131.

[13] 李广一，王厚峰. 基于多步聚类的汉语命名实体识别和歧义消解[J]. 中文信息学报，2013，27(5)：29-34.
Li Guangyi, Wang Houfeng. Chinese named entity recognition and disambiguation based on multi-stage clustering[J]. Journal of Chinese Information Processing, 2013, 27(5): 29-34.

[14] Chen Ying, Martin J. Cu-coMem: Exploring rich features for unsupervised web personal name disambiguation [C]// Proceedings of the 4th International Workshop on Semantic Evaluations. [S. l.]: Association for Computational Linguistics, 2007: 125-128.

[15] Artiles J, Gonzalo J, Sekine S. Weps 2 evaluation campaign: Overview of the web people search clustering task [C]// Proceedings of 2nd Web People Search Evaluation Workshop. Madrid, Spain: [s. n.], 2009: 9-16.

[16] 张猛，王大玲，于戈. 一种基于自动阈值发现的文本聚类方法[J]. 计算机研究与发展，2004，41(10)：1748-1753.
Zhang Meng, Wang Daling, Yu Ge. A text clustering method based on auto-selected threshold[J]. Journal of Computer Research and Development, 2004, 41(10): 1748-1753.

作者简介：

阳怡林(1987-)，男，硕士研究生，研究方向：人名消歧，E-mail：yangyilinyx@163.com。

周杰(1984-)，男，博士研究生，研究方向：实体消歧。

李弼程(1970-)，男，教授，博士生导师，研究方向：语音信号处理与智能信息处理。

席耀一(1987-)，男，博士研究生，研究方向：信息抽取。