利用 D-S 证据理论进行特征融合的同义实体识别*

何晶晶，蔡德胜，介 飞，吴共庆
（合肥工业大学 计算机与信息学院，合肥 230009）

摘 要：针对现实中同一实体存在不同表象的问题，提出一种基于 D-S 证据理论特征融合的同义实体识别方法。以搜索引擎为外部知识库获取实体特征信息，利用相似函数计算特征值，由 D-S 证据理论融合一组特征值，经阈值判断完成同义实体的识别。特征融合识别算法在医疗机构数据集上的识别精度、召回率和 F 值分别达到了 85.80%、81.18%、83.43%，比单纯利用实体名的算法分别提高了 4.09%、4.30% 和 4.21%。实验表明 D-S 证据理论将多特征融合，对同义实体识别具有更好的识别效果。

关键词：D-S 证据理论；特征融合；同义实体识别；搜索引擎；相似函数
中图分类号：TP301.6 **文献标志码**：A **文章编号**：1001-3695(2018)05-1429-05
doi:10.3969/j.issn.1001-3695.2018.05.032

Synonymous entity recognition based on D-S evidence theory for feature fusion

He Jingjing, Cai Desheng, Jie Fei, Wu Gongqing
(*School of Computer & Information, Hefei University of Technology, Hefei* 230009, *China*)

Abstract: As the same entity has different expressions in the real world, this paper proposed a synonymous entity recognition method based on the D-S evidence theory for feature fusion. First the recognition method obtained the entity features from a search engine, an external knowledge base, and calculated feature values using a similarity function. Then it identified synonymous entities through a threshold value after fusing a group of features using the D-S theory. The recognition accuracy, recall and F value of using feature fusion on the medical institution dataset were 85.80%, 81.18% and 83.43%, respectively, which were 4.09%, 4.30% and 4.21% higher than the method simply using the entity name. Experiments show that the D-S evidence theory has better recognition effect on synonymous entities by multi-feature fusion.

Key words: D-S evidence theory; feature fusion; synonymous entity recognition; search engine; similarity function

0 引言

互联网信息呈爆炸性增长的同时，数据、信息的质量问题受到了人们的广泛关注。劣质数据很大程度上降低了信息的可用性，提供给用户陈旧的、缺损的、冗余的、甚至错误的信息。有关统计表明，劣质数据每年给美国企业带来数以亿计的经济损失，其在国内也已成为一个不可忽视的存在[1]。

冗余信息中最小冗余单位就是冗余的实体，冗余实体是指同一实体在现实生活中存在不同的表象。如番茄的另一种表象为西红柿，将番茄和西红柿两个实体识别为相同实体的过程称为同义实体识别。同义实体识别是提高数据质量的一个重要步骤，其任务就是寻求数据中描述同一实体的若干元组并将其合并[1]，使得融合后数据的可信度更高。

基于相似函数的方法是目前比较通用的同义实体识别方法。姜孟晋等人[2]利用文本的内部信息对信息元素进行分类，用不同的相似函数度量不同的信息元素，但在实体识别前期，需要对不同信息元素作不同的处理。信息量有限，即使充分利用文本内部信息，相似函数再好，也难以很好地度量实体特征，如利用文本编辑距离去度量电话号码、邮编等。徐喆昊等人[3]利用 Web 动态知识库，以 Web 搜索返回的页面摘要数衡量相似度，能一定程度上弥补不同相似函数的误差，但没有利用实体特征。

为利用 Web 动态知识库，同时结合实体特征，本文在文献[3]的基础上，将实体特征与实体名组合作为查询词搜索，获取 Web 搜索返回的页面摘要数。再结合相似函数，可获取实体的特征值，为识别出同义实体，还需将多个特征值融合，作出最终判断。特征融合的方法有很多，Khaleghi 等人[4]从各角度对常用融合方法进行比较，其中，针对冗余互补类的信息，D-S 证据理论采用逻辑推理的技术进行融合。本文获取的特征值为一组冗余互补的数据，可运用 D-S 证据理论进行处理。

本文的主要贡献如下：

a）提出了结合实体特征的同义实体识别思路；

b）针对同义实体识别问题，提出了利用 D-S 证据理论进行特征融合的策略；

c）以搜索引擎为工具，实现了基于 D-S 证据理论进行特征融合的同义实体识别框架和算法。

经过实体统一后，能够更好地维护数据的实体同一性，提高数据质量，向用户提供更高层的服务。如在情报分析、数据挖掘等方面，用户可通过单数据源快速、高效、准确地获取多个

收稿日期：2017-01-06；**修回日期**：2017-02-28 **基金项目**：国家“863”计划资助项目(2012AA011005)；国家自然科学基金资助项目(61273297)

作者简介：何晶晶(1990-)，女，湖北随州人，硕士研究生，主要研究方向为 Web 智能、数据挖掘(he_jingjing@163.com)；蔡德胜(1991-)，男，安徽宣城人，硕士研究生，主要研究方向为 Web 智能、数据挖掘；介飞(1991-)，男，山西运城人，博士研究生，主要研究方向为数据挖掘、社交网络；吴共庆(1975-)，男，安徽安庆人，副教授，主要研究方向为 Web 智能、数据挖掘.

异构数据源上的数据。因此,该项研究具有重要意义。

1 相关工作

命名实体识别(named entity recognition,NER)作为自然语言处理(natural language processing,NLP)的一个基本任务,由第六届消息理解会议(message understanding conference-6,MUC-6)于 1995 年第一次引入[5]。2003 年我国首次将中文命名实体识别作为分词标注的子任务引入。实体识别的别名有重复探测、记录链接、实体链接、实体解析、名称匹配、数据消重等。识别技术在信息抽取、信息检索、机器翻译、问答系统中均有广泛应用[6]。

传统同义实体识别方法主要有基于相似函数的方法[2,7,8]、基于统计的方法[9~11]和基于规则的方法[12,13]。相似函数是利用定义在特征值上的相似性判定实体间的关系。Bilenko 等人[7]提出两种文本相似函数进行实体识别,短字符串用编辑距离,长字符串用向量空间模型;从字符串长短方面考虑用不同的相似函数,但没有针对实体特征的不同类型作相应改进。Cohen 等人[8]总结了运用不同距离函数进行实体匹配的方法,指出字符串距离能够识别实体,但不适用非平凡结构的实体比较。基于统计的方法具有较好的鲁棒性和灵活性,难点在于特征选择上,且需要大量的标注集。Christen[9]提出两步法,即最近邻和 SVM 分类器方法,将同义实体的识别视为分类问题,相比用阈值或统计方法处理分类问题,自监督学习方法更方便,但需耗费额外代价进行训练。怀宝兴等人[10]利用维基百科和概念主题模型将词和命名实体映射到同一个主题空间,由实体的位置向量准确定位其同义实体,此过程需要对训练数据进行大量的歧义标注。Wang 等人[11]用统计方法估计实体对的特征值和聚合值,对记录对进行匹配。基于语义规则的方法需要制定规则,准确率较高,但依赖于特征领域,可移植性差。Fan 等人[12]首次提出记录匹配规则的概念,运用已知规则推理有用规则,为利用语义规则进行实体识别奠定了基础。Gupta 等人[13]结合最大信息论和共享语境进行大规模处理,识别同义词短语,该方法目前只适用单语言环境,不适用多语言环境。

将 Web 动态知识库作为语料库[3,14~17],可避免传统文本信息的有限性问题,更方便准确地进行实体识别。Chen 等人[14]基于 Web 搜索复查模式,提出五种相似函数度量实体间的关系。利用 Web 搜索返回的页面数,能更方便度量实体间的关系,但没有利用实体特征。从一个词集合中抽取出与给定词最相似的词,称为同义词抽取[15]。Hu 等人[16]以 Web 知识库为词集,结合维基百科和启发式规则自动标记样本,从 Web 中抽取尽可能多的同义词,只简单处理了候选词,没有用到文本特征和语义关系。Banerjee 等人[17]针对订阅新闻中信息过载现象,将同义实体的识别视为聚类问题,从维基百科中生成短文本特征,用提高短文本聚类的准确率达到实体识别的目的,但没有针对不同聚类算法给出不同的特征。

利用多特征进行同义实体识别,需要将多特征值融合。特征融合[4,18~22]既保留了多特征的有效鉴别信息,又消除了主客观因素的影响,关键是对特征参数的优化组合。Sun 等人[18]将关联规则特征融合方法运用于图像识别,对抽取的两组特征向量建立尺度函数,再利用关联规则抽取典型相关特征构成判别向量;但该融合方法只适用于高维空间和小样本。Han 等人[19]运用统计学特征抽取方法从现实和合成模板中学习步态特征,并对学习到的有效特征进行融合,达到人体识别的目的;步态特征的抽取是难点之一,抽取质量对识别效率有较大影响。徐从富等人[20]指出 D-S 方法为不确定信息的表达和合成提供了强有力的理论依据。基于 D-S 证据理论,Che 等人[21]将图像的颜色和纹理等多特征融合;Wu 等人[22]将 Web 新闻中抽取的标签路径特征进行融合,表明 D-S 方法适用于冗余互补类的信息融合。

2 基于搜索引擎的特征值计算

本章主要给出基于搜索引擎的实体特征值计算方法。实体的特征值计算主要分为实体特征信息的获取和实体特征值的计算。

同义实体识别的方法包含基于相似函数、统计、语义规则和搜索引擎的方法。基于搜索引擎的同义实体识别方法能够避免基于相似函数方法传统文本信息的局限性问题;不必依赖基于语义规则方法的具体语言和必须具备领域知识的局限性;不需要如基于统计方法对数据集进行专门训练和对于特征的高要求性。本文提出的同义实体识别方法,在已知实体名称和实体特征的前提下,只需通过搜索引擎进行查询处理,即可获取返回的页面摘要信息,从而判断两个实体是否为同义实体。

2.1 基于搜索引擎的实体特征信息获取

对于任意一个命名实体,记为实体 A,实体 A 有 n 种特征,第 $i(1\leqslant i\leqslant n)$ 种特征记为 $f_{A,i}$,则 n 种特征分别记为 $f_{A,1}$,$f_{A,2},\cdots,f_{A,n}$。在实体特征信息的获取阶段,主要是将实体名和实体特征组合,作为查询词搜索。当查询词为"实体 A + 特征 $f_{A,i}$"时,本文称为实体 A 的第 i 种特征组合。查询返回的结果是页面摘要信息,简称页面信息,包含摘要的标题、内容和相应的链接。实体 A 有 n 种特征,对应 n 种特征组合情况,返回 n 类页面信息。

经过搜索引擎查询,可以获取实体 A 和任意特征组合得到的页面信息。笔者发现,在实体 A 的页面信息中,除了显示 A 的相关信息外,还会显示 B 的相关信息(A、B 的实体名称不同)。基于搜索引擎的工作原理,笔者认为 B 和 A 存在某种关联。这种关联程度可以由返回 A 的页面信息集合中包含 B 的页面摘要条数来衡量。

对于实体 A 的第 i 种特征组合,统计 A 的页面信息中包含 B 的页面摘要条数,记为 $c_{A,i}^{B}$。同理,B 的第 i 种特征组合返回的页面信息中包含 A 的页面摘要数为 $c_{B,i}^{A}$。将经搜索引擎处理的 $c_{A,i}^{B}$ 和 $c_{B,i}^{A}$ 分别称为实体 A、B 结合第 i 种特征组合的实体特征信息,则共有 n 类实体特征信息。

2.2 实体特征值的计算

基于 Web 动态知识库,以搜索引擎为工具,得到第 i 种特征组合下,A、B 的实体特征信息分别为 $c_{A,i}^{B}$ 和 $c_{B,i}^{A}$。将 $c_{A,i}^{B}$ 和 $c_{B,i}^{A}$ 相加,得到第 $i(1\leqslant i\leqslant n)$ 种特征下,实体对 (A,B) 的实体特征信息记为 $c_{(A,B),i}$,$c_{(A,B),i}=c_{A,i}^{B}+c_{B,i}^{A}$。

然而,以 $c_{(A,B),i}$ 代表实体对 (A,B) 的同义关联程度,衡量实体对的相似关系,显得不够严谨。因为搜索引擎中获取的页面摘要总数越大,$c_{A,i}^{B}$ 和 $c_{B,i}^{A}$ 就越大,$c_{(A,B),i}$ 也越大。为了更好地衡量实体对 (A,B) 的关系,引入页面摘要总数,N_A 表示 A 的页面摘要总数,即搜索引擎中检索实体 A 的结果总数,提出特征

下实体对(A,B)的相似度计算公式。对于第 i 种特征,实体对(A,B)的相似度 $\mathrm{sim}(A,B,i)$ 如式(1)所示。

$$\mathrm{sim}(A,B,i)=\frac{c_{(A,B),i}}{N_A+N_B} \tag{1}$$

其中:N_A 表示 A 的页面摘要总数;N_B 表示 B 的页面摘要总数。

任意实体有 n 种特征,故任意实体对有 n 种特征,得到 n 种相似度值,作为特征值。那么,实体对(A,B)的 n 个特征值构成了 n 维特征向量 $\boldsymbol{F}_n$,$\boldsymbol{F}_n=(\mathrm{sim}_1,\mathrm{sim}_2,\cdots,\mathrm{sim}_n)$。

对于实体对(A,B)的 n 维特征向量 $\boldsymbol{F}_n$,每种特征对应一个特征值,每一维特征值进一步反映了此种特征下的同义关联程度。对该特征值作阈值判断,可以初步得到该种特征下的同义实体识别结果。

3　基于 D-S 证据理论特征融合的同义实体识别

基于搜索引擎的实体特征值计算方法,可以获取实体对(A,B)的 n 个特征值。将多个特征融合成一个有效特征的过程称为特征融合。

特征融合是信息融合(information fusion)的一种,起初被称为数据融合(data fusion)。基于 D-S 证据理论的信息融合方法应用层次广泛,以证据的方式表征不确定性,不需要知道先验概率,具有其他信息融合方法所不具备的优势。D-S 证据理论在专家系统、决策分析、故障诊断、目标识别等信息融合领域得到了广泛应用。

基于 D-S 证据理论,Wu 等人[22]设计了一系列的标签路径抽取特征进行 Web 新闻的抽取,为综合各个特征的优势,将多个标签路径特征进行融合,达成信息融合的目的。Che 等人[21]将图像的颜色和纹理特征融合,进而完成图像识别。Liu 等人[23]将灰色关联与 D-S 证据理论相结合,建立推理决策模型,提出一种快速有效的决策方法与模式进行决策分析。无论是信息融合,还是目标识别,或者是决策分析,采用的模型都是将获取的多个不确定信息进行有效融合,从而给出最终的判决结果。

当利用多特征(属性)进行同义实体识别时,亦是将多个实体特征的不确定信息进行融合,只是最后作出的判决是判定实体对是否为同义实体对,适用于利用 D-S 证据理论进行目标识别的应用场景。故采用 D-S 证据理论的特征融合策略,进行同义实体识别。

3.1　D-S 证据理论

Dempster-Shafer(简称 D-S)理论,也称为 Dempster 合成规则,可综合不同数据源的数据,在信息融合等领域得到了广泛应用。

设 U 为一识别框架,U 上的基本概率赋值是一个 $2^U\rightarrow[0,1]$ 的函数 m,称为 mass 函数,m 满足

$$m(\varnothing)=0;\sum_{A\subseteq U}m(A)=1 \tag{2}$$

其中:命题 A 是识别框架 U 的非零子集;$m(A)$ 表示对 A 的信任程度。$\forall A\subseteq U$ 上的有限个 mass 函数 $m_1,m_2,\cdots,m_n$ 的 Dempster 合成规则如式(3)所示。

$$m_1\oplus m_2\oplus\cdots\oplus m_n(A)=\begin{cases}\dfrac{\sum_{A_1\cap A_2\cap\cdots\cap A_n=A}m_1(A_1)\times m_2(A_2)\times\cdots\times m_n(A_n)}{K} & A\neq\varnothing\\ 0 & A=\varnothing\end{cases} \tag{3}$$

其中:$K=\sum_{A_1\cap A_2\cap\cdots\cap A_n\neq\varnothing}m_1(A_1)\times m_2(A_2)\times\cdots\times m_n(A_n)$,$K$ 是归一化常数。

3.2　基于 Dempster 规则的特征融合

根据要解决的同义实体识别问题,假设 $U=\{P_1、P_2\}$,其中,P_1 表示判定实体对(A,B)是同义实体对,P_2 表示判定(A,B)是非同义实体对。根据 Dempster 合成规则,第 i 个特征值对应第 i 个 mass 函数 m_i,n 维特征向量 $\boldsymbol{F}_n$ 对应 n 个 mass 函数。当集合 $A=\{P_1\}$时,即判定实体对为同义实体对时,对应的 n 个 mass 函数$(m_1,m_2,\cdots,m_n)=(\mathrm{sim}_1,\mathrm{sim}_2,\cdots,\mathrm{sim}_n)$;同理,当集合 $A=\{P_2\}$时,对应的 n 个 mass 函数$(m_1,m_2,\cdots,m_n)=(1-\mathrm{sim}_1,1-\mathrm{sim}_2,\cdots,1-\mathrm{sim}_n)$。对应的分布情况如表 1 所示。

表 1　识别框架 U 上的 n 个 mass 函数分布情况

	$m_1()$	$m_2()$	$m_i()$	$m_n()$	$m()$
P_1	sim_1	sim_2	sim_i	sim_n	
P_2	$1-\mathrm{sim}_1$	$1-\mathrm{sim}_2$	$1-\mathrm{sim}_i$	$1-\mathrm{sim}_n$	

根据 D-S 证据理论特征融合时,会导致“Zadeh 悖论”的发生。但是,对于实体对(A,B),本文假定,如果它们不是同义实体对,那么一定就是非同义实体对。当识别框架 $U=\{P_1、P_2\}$时,即 U 上只有两个集合元素时,P_1 出现小概率时,P_2 必然出现大概率,这样就避免了“Zadeh 悖论”的发生。

对于任意实体对(A,B),利用 Dempster 规则将 $m_1,m_2,\cdots,m_n$ 进行融合,可以按照以下两步完成:

a)计算归一化常数 K,K 的计算结果如式(4)所示。

$$K=\mathrm{sim}_1\times\mathrm{sim}_2\times\cdots\times\mathrm{sim}_n+(1-\mathrm{sim}_1)\times(1-\mathrm{sim}_2)\times\cdots\times(1-\mathrm{sim}_n) \tag{4}$$

b)计算 P_1、P_2 的组合 mass 函数,P_1、P_2 的组合 mass 函数如式(5)(6)所示。

$$m_1\oplus m_2\oplus\cdots\oplus m_n(\{P_1\})=\frac{\mathrm{sim}_1\times\mathrm{sim}_2\times\cdots\times\mathrm{sim}_n}{K} \tag{5}$$

$$m_1\oplus m_2\oplus\cdots\oplus m_n(\{P_2\})=\frac{(1-\mathrm{sim}_1)\times(1-\mathrm{sim}_2)\times\cdots\times(1-\mathrm{sim}_n)}{K} \tag{6}$$

按照 Dempster 合成规则,完成了对任意实体对(A,B)的 n 个特征值的融合,融合后的特征值如式(7)所示。

$$m_1\oplus m_2\oplus\cdots\oplus m_n(\{P_1\})=\frac{\mathrm{sim}_1\times\mathrm{sim}_2\times\cdots\times\mathrm{sim}_n}{\mathrm{sim}_1\times\mathrm{sim}_2\times\cdots\times\mathrm{sim}_n+(1-\mathrm{sim}_1)\times(1-\mathrm{sim}_2)\times\cdots\times(1-\mathrm{sim}_n)} \tag{7}$$

融合后的特征值,经阈值比较,便可以进行同义实体的识别。依据 D-S 证据理论特征融合的同义实体识别算法 SER-DS(synonymous entity recognition Dempster-Shafer)如下所示。

算法 SER-DS

输入:命名实体对(A,B),阈值 δ。

输出:1/0

```
1   featureValList = null。
2   for each label in featureLabelList{
3         s_A←extractSnippets(A,label);
4         s_B←extractSnippets(B,label);
5         c_A←getFeatureInfo(s_A,B);
6         c_B←getFeatureInfo(s_B,A);
7         featureVal←sim(c_A,c_B,label);
8         featureValList. add(featureVal);
9    }
10   fusedVal←getFusedFeatureValue(featureValList);
11   if(fusedVal≥δ)
```

```
12          return 1;
13      else
14          return 0;
```

其中,特征标签集合(featureLabelList)中的每个特征标签(label)表示实体对(A,B)的每种特征,遍历特征标签集合,获取相应的特征值,存放在特征值列表(featureValList)中。

对于实体对(A,B),算法第 3 ~ 4 行获取 label 对应的页面信息;第 5 ~ 6 行获取实体特征信息 c_A 和 c_B;第 7 行利用式(1)计算 label 对应的特征值;第 8 行将特征值存放到 featureValList 中。第 10 行利用式(7)将 featureValList 中的多个特征值融合,得到融合后的特征值 fusedVal。第 11 行将 fusedVal 与阈值 δ 比较,第 12 行判断为同义实体,则返回 1;第 14 行判断为非同义实体,返回 0。其中,阈值的设置在 4.3 节中详细讨论。

4 实验结果

4.1 实验数据集

实验数据源于国家"863"计划课题"多源异构数据集成与挖掘的关键技术"的示范应用系统"普适医疗系统"。选取部分北京地区的数据,命名实体为医疗机构实体,选择医疗机构的名称(name)、电话(tel)、地址(addr)、电话地址(tel&addr)组合特征共四个特征。其中,添加实体的名称特征是指将实体名作为查询词搜索,添加电话特征是将实体名和电话组合作为查询词搜索,添加地址特征是将实体名和地址组合,添加电话地址特征是将实体名、电话和地址组合。由必应搜索引擎完成页面信息的获取。所有数据均来自真实网站,其中命名实体共有 316 个,由笛卡尔积组合成命名实体对,经过人工整理出 186 对同义实体对。

实验数据集中共四种实体特征,基于单个特征的同义实体识别方法记为 SER-特征,则四种基于单一特征的同义实体识别方法分别为 SER-name、SER-tel、SER-addr 和 SER-tel&addr。基于特征融合的同义实体识别算法为 SER-DS。

4.2 评价标准

本文采用精度 P(precision)、召回率 R(recall)和 F 值作为同义实体识别的性能评估指标。记 S_e 为算法获取的同义实体对集合,S_l 为标准同义实体关系对集合,$S_e \cap S_l$ 为正确识别的同义实体对集合。$|S_e|$ 表示 S_e 中实体对的对数,$|S_l|$ 表示 S_l 中实体对对数,$|S_e \cap S_l|$ 表示 $S_e \cap S_l$ 中实体对对数。相关计算如式(8) ~ (10)所示。其中,精度 P 表示正确识别的同义实体对占识别出的同义实体对的比例,召回率 R 表示正确识别的同义实体对占标准同义实体对的比例,F 值是一个综合指标。

$$P = \frac{|S_e \cap S_l|}{|S_e|} \tag{8}$$

$$R = \frac{|S_e \cap S_l|}{|S_l|} \tag{9}$$

$$F = \frac{2 \times P \times R}{P + R} \tag{10}$$

4.3 阈值的设置

经阈值 δ 对特征值进行判断,能够将实体分为同义实体对和非同义实体对。由 P、R、F 的定义可知,P 随着阈值的增加而增加,R 随着阈值的增加而降低。一个合适的阈值是均衡精度和召回率,使得 F 值最高的关键。因此,如何选取合适的阈值是精确识别同义实体的关键因素之一。

常见的阈值选取方法有平均值、中间值和标准差等。本文随机抽取医疗机构数据集上 30% 的同义实体对和 30% 的非同义实体对,构成数据集 D_1。以 FSE 算法为例,图 1 给出了该算法在 D_1 上实验时,P、R、F 随阈值变化而变化的趋势。

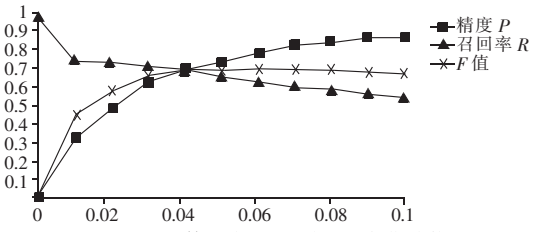

图 1 FSE 算法在 D_1 上随 δ 的变化趋势

可以看出,$\delta = 0$ 时,召回率为 100%,而精度仅有 0.3%。这是因为 $\delta = 0$ 时,所有的实体都被识别为同义实体,同时所有的噪声数据也被识别为同义实体。随着 δ 的增加,精度增加,而召回率降低。在 δ 取 0.06 附近时,召回率和精度达到均衡,此时 F 值最高。本文的同义实体对和非同义实体对的特征值之间差别不大,直接取中间值为经验值。FSE 算法的阈值设置为 0.06,SER-DS 算法的阈值为 4.0E-05,基于单一特征 SER-name、SER-tel、SER-addr 和 SER-tel&addr 方法的阈值分别为 0.09、0.09、0.08、0.08。

4.4 实验结果与分析

从不同特征和不同实体识别算法两方面对实验结果进行分析。

4.4.1 不同特征对实验结果的影响

为比较不同特征对同义实体识别的影响,将基于单个特征和基于特征融合的同义实体识别方法进行对比,如表 2 所示。

表 2 不同特征的同义实体识别 P、R、F 值比较 /%

识别方法	P	R	F
SER-name	81.71	76.88	79.22
SER-tel	81.29	74.73	77.87
SER-addr	80.53	**82.26**	81.38
SER-tel&addr	75.69	73.66	74.66
SER-DS	**85.80**	81.18	**83.43**

观察比较基于单个特征的同义实体识别性能发现:SER-addr 方法在精度和 F 值上优于 SER-name 方法。表明添加医疗机构实体的地址特征,相较于搜索单一实体名,对于同义实体的识别具有促进作用。而 SER-tel、SER-tel&addr 方法低于 SER-name 方法,表明添加医疗机构实体的电话特征、电话地址特征,相较于搜索单一实体名,对于同义实体的识别不具有促进作用。

观察比较 SER-DS 算法和单个特征的同义实体识别性能发现:SER-DS 算法在精度和 F 值上优于任意一个单一特征,比 SER-name 方法分别增长了 4.09% 和 4.3%,而 SER-DS 的召回率稍低于 SER-addr,比 SER-name 高 4.21%。实验结果表明:基于 D-S 证据理论特征融合的同义实体识别方法 SER-DS,能够有效融合实体对的各个特征值,综合了各特征的不同影响,相比基于单一特征进行同义实体识别,同义实体的识别效果提高了。

4.4.2 对比实验

为验证 SER-DS 算法的有效性,以文献[3]的 FSE(find synonymous entities)算法为对比算法。FSE 算法是利用相似度式(11),基于单一实体名进行同义实体识别的。同时,为体现本文相似函数的优势,本文将 SER-name 方法,即利用相似度

式(1),与基于单一实体名的同义实体识别方法同时进行对比。

$$\mathrm{sim}(A,B)=\frac{\sqrt{c_A^B\times c_B^A}}{\sqrt{N_A\times N_B}} \tag{11}$$

其中:c_A^B 表示 A 的页面信息中包含 B 的页面摘要条数;c_B^A 表示 B 的页面信息中包含 A 的页面摘要数;N_A 表示 A 的页面摘要总数;N_B 表示 B 的页面摘要总数。

表 3 给出了 FSE、SER-name 和 SER-DS 算法在医疗机构数据集上的对比结果。发现在精度、召回率和 F 值上,SER-DS 方法优于 SER-name,SER-name 方法优于 FSE。

表 3　不同算法的同义实体识别 P、R、F 值比较　/%

误别方法	P	R	F
FSE	79.87	63.98	71.04
SER-name	81.71	76.88	79.22
SER-DS	**85.80**	**81.18**	**83.43**

实验结果表明:FSE 方法能够正确识别出的同义实体对的比例较低,说明识别结果中噪声较多,误识率较高;同时 FSE 方法能够识别出的同义实体对的比例较低,说明漏识了同义实体对,漏识率较高。两方面共同导致综合指标 F 值低于 SER-name 和 SER-DS 方法。

SER-DS 方法在精度、召回率和 F 值上优于 FSE 方法,主要原因是 SER-DS 方法结合了实体特征,并运用 D-S 证据理论将各个特征值融合,既保留了多特征的有效鉴别信息,又消除了主客观因素带来的冗余信息,对同义实体的识别具有促进作用。

5　结束语

现实中同一个实体存在很多不同的表象,在 Web 信息融合过程中,需要对数据库中不同数据源的数据进行统一化,以达到去除冗余数据的目的。而冗余实体的识别,即同义实体的识别,是本文亟待解决的问题。本文结合搜索引擎、相似度度量和特征融合的技术,提出一种基于 D-S 证据理论进行特征融合的同义实体识别算法。该算法在真实医疗机构数据集上进行实验,比基于单一特征的同义实体识别方法有不同程度的提高。

本文提出的同义实体识别方法,只是对于机构实体得到了较好的实验结果,要想算法具有更好的通用性,后期会实验于不同类型的实体,这是今后需要改进的地方。

参考文献:

[1] 刘显敏,李建中. 实体识别问题的相关研究[J]. 智能计算机与应用,2013,3(2):1-5.

[2] 姜孟晋,周雅倩,黄萱菁. 基于同义实体扩展的冗余信息去重[J]. 中文信息学报,2012,26(1):42-51.

[3] 徐喆昊,吴共庆,胡学钢. 基于同义实体识别的 Web 信息集成[J]. 计算机系统应用,2015,24(9):35-42.

[4] Khaleghi B, Khamis A, Karray F O, *et al*. Multisensor data fusion: a review of the state-of-the-art[J]. Information Fusion, 2013, 14(1): 28-44.

[5] Grishman R, Sundheim B. Message understanding conference- 6: a brief history[C]//Proc of COLING. Stroudsburg, PA: Association for Computational Linguistics, 1996: 466-471.

[6] Lei Jianbo, Tang Buzhou, Lu Xueqin, *et al*. A comprehensive study of named entity recognition in Chinese clinical text[J]. Journal of the American Medical Informatics Association, 2014, 21(5): 808-814.

[7] Bilenko M, Mooney R J. Adaptive duplicate detection using learnable string similarity measures[C]//Proc of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2003: 39-48.

[8] Cohen W, Ravikumar P, Fienberg S. A comparison of string metrics for matching names and records[C]//Proc of KDD Workshop on Data Cleaning and Object Consolidation. Palo Alto, CA: AAAI Press, 2003: 73-78.

[9] Christen P. Automatic record linkage using seeded nearest neighbour and support vector machine classification[C]//Proc of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2008: 151-159.

[10] 怀宝兴,宝腾飞,祝恒书,等. 一种基于概率主题模型的命名实体链接方法[J]. 软件学报,2014,25(9):2076-2087.

[11] Wang Xin, Sun Ang, Kardes H, *et al*. Probabilistic estimates of attribute statistics and match likelihood for people entity resolution[C]//Proc of IEEE International Conference on Big Data. Washington DC: IEEE Press, 2014: 92-99.

[12] Fan Wenfei, Jia Xibei, Li Jianzhong, *et al*. Reasoning about record matching rules[J]. Proceedings of the VLDB Endowment, 2009, 2(1): 407-418.

[13] Gupta D, Carbonell J G, Gershman A, *et al*. Unsupervised phrasal near-synonym generation from text corpora[C]//Proc of AAAI. 2015: 2253-2259.

[14] Chen H H, Lin Mingshun, Wei Yuchuan. Novel association measures using Web search with double checking[C]//Proc of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2006: 1009-1016.

[15] Turney P D. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL[C]//Proc of the 12th European Conference on Machine Learning. London: Springer-Verlag, 2001: 491-502.

[16] Hu Fanghuai, Shao Zhiqing, Ruan Tong. Self-supervised synonym extraction from the Web[J]. Journal of Information Science and Engineering, 2015, 31(3): 1133-1148.

[17] Banerjee S, Ramanathan K, Gupta A. Clustering short texts using Wikipedia[C]//Proc of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2007: 787-788.

[18] Sun Quansen, Zeng Shenggen, Liu Yan, *et al*. A new method of feature fusion and its application in image recognition[J]. Pattern Recognition, 2005, 38(12): 2437-2448.

[19] Han Ju, Bhanu B. Statistical feature fusion for gait-based human recognition[C]//Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Press, 2004: 842-847.

[20] 徐从富,耿卫东. Dempster-Shafer 证据推理方法理论与应用的综述[J]. 模式识别与人工智能,1999,12(4):424-430.

[21] Che Chang, Yu Xiaoyang, Yu Guang. The research of image retrieval based on multi feature DS evidence theory fusion[J]. International Journal of Signal Processing, Image Processing and Pattern Recognition, 2016, 9(1): 51-62.

[22] Wu Gongqing, Li Lei, Li Li, *et al*. Web news extraction via tag path feature fusion using DS theory[J]. Journal of Computer Science and Technology, 2016, 31(4): 661-672.

[23] Liu Sifeng. 基于灰色关联分析和 D-S 证据理论的区间直觉模糊决策方法[J]. 自动化学报,2011,37(8):993-998.