基于层次聚类的跨文本中文人名消歧研究

张菲菲[1],李宗海[2],周晓辉[1],李晓戈[1,2]
ZHANG Feifei[1], LI Zonghai[2], ZHOU Xiaohui[1], LI Xiaoge[1,2]

1.西安邮电大学,西安 710121
2.济南中林信息科技有限公司,济南 250100
1.Xi'an University of Posts & Telecommunications, Xi'an 710121, China
2.Jinan Zhonglin Information Technology Co., Ltd, Jinan 250100, China

ZHANG Feifei, LI Zonghai, ZHOU Xiaohui, et al. Cross-document Chinese personal name entity disambiguation based on hierarchical clustering. Computer Engineering and Applications, 2014, 50(6):106-111.

Abstract: Cross-document entity disambiguation is the problem of identifying whether mentions from different documents refer to the same or distinct entities. This paper describes a Chinese information extraction system which involves both document-level IE and corpus-level IE, a pipeline and multi-level modular approach to name entity and Entity Profile extraction. It introduces novel features based on document-level entity profiles and study on the influence of feature selection, parameter selection, parameter validation and analysis on results. Disambiguation is performed based on agglomerative hierarchical clustering using Hadoop. Experiments show that *F*-measure of training set is 91.33% and testing set is 88.73%, using the whole network news corpus dataset from Harbin Institute of Technology.
Key words: entity disambiguation; information extraction; similarity; hierarchical clustering

摘 要:人名消歧已经成为自然语言处理和信息抽取应用中亟待解决的重要问题。运用中文自然语言处理和信息抽取系统识别命名实体和实体关系,生成实体信息对象(Entity Profile),采用实体信息对象(EP)中的个人信息特征,实体关系和上下文相关信息在Hadoop平台上基于凝聚的层次聚类方法解决了实体消歧问题。采用哈尔滨工业大学整理的全网新闻语料作为人名消歧训练和测试数据,着重研究了中文人名消歧特征的选取,参数的确定和验证,在训练集和测试集上分别取得了91.33%和88.73%的*F*值。说明提出的方法具有较好的可行性。
关键词:人名消歧;信息抽取;相似度;层次聚类
文献标志码:A **中图分类号:**TP391.12 doi:10.3778/j.issn.1002-8331.1309-0423

1 引言

在互联网上搜索人名已经十分常见,但人名重名的现象也非常普遍,往往搜索的结果中会出现大量相同名字的网页。曾统计搜索"李静",在结果去重后选取前43个搜索结果,统计网页中的"李静"分别表示了6个不同的人。

命名实体消歧已经成为自然语言处理中亟待解决的重要问题,对问答系统,信息检索[1],网络知识库和复杂信息网络构建有着重要影响。在基本的三大类命名实体中,人名比地名、组织机构名具有更强的歧义性,解决难度也更高。例如,在不同的文本源中,相同的姓名代表不同的人物实体,不同的姓名代表相同的人物实体。这种现象的存在极大地制约着信息抽取应用[2]的可靠性与实用性。本文在自然语言处理和信息抽取技术的基础上,针对由不同文档抽取出来的人物实体信息的相似度矩阵进行聚类,从而实现人名消歧。

2 相关工作研究

人名消歧早期主要是针对新闻类型的文本信息及一些学术中自动处理中人名消歧的问题研究。早在1994年,跨文档指代消解(Cross-Document Co-reference,CDC)作为MUC-6[3]的潜在任务被首次提出。1998年,Bagga

作者简介:张菲菲(1987—),女,硕士,主要研究方向:命名实体消歧和文本数据挖掘;李宗海(1988—),男,主要研究方向:信息抽取、人工智能;周晓辉(1978—),男,博士,教授,主要研究方向:电子商务、并行计算和分布式存储;李晓戈(1962—),男,博士,教授,主要研究方向:自然语言处理、机器学习和文本数据挖掘。
收稿日期:2013-09-27 **修回日期:**2013-11-15 **文章编号:**1002-8331(2014)06-0106-06

和Baldwin[4]提出用向量空间模型(Vector Space Model，VSM)算法，将实体信息的比较转换为空间向量的比较，实现跨文档人名的指代消解。为了对他们的系统进行评估，他们还提出了B-CUBED算法对跨文档指代消解进行性能评估。2007年，WePS(Web People Search)[5]评测研讨会与语义评测研讨会组织了针对英文网页中的人名消歧的评测任务，WePS还分别在2009年与2010年开展了两届关于网络人名消歧的评测会议。2008年，ACE评测会议将GEDR(Global Entity Detection and Recognition)和GRDR(Global Relation Detection and Recognition)作为两项重要的评测内容，并对英语和阿拉伯语两类语种进行了评测。

相比于英文，中文人名消歧研究工作开展较晚。2010年，SIGHAN-CIPS联合学术会议CLP2010[6]开展了首次设置了中文跨文本人名消歧任务评测，其中东北大学的周晓[7]等在实验室开发的领域知识库中，抽取文档人物的属性特征建立不同人物之间的互斥关系，并利用之间的关系进一步聚类。东北大学的丁海波[8]使用了相类似的方法，抽取人物属性进行初步聚类，之后利用局部上下文特征和全局特征依次进行聚类。这些方法都取得了一定的成果，但由于研究过程都非常依赖领域知识库，缺少一定的通用性。哈工大的郎君等[9]依据同名不同人物具有不同网络思想，对搜索结果有重名的人名进行消歧。他们都是从特征选择方面进行人名消歧的研究，而没有对聚类方法进行改进。

3　系统架构与研究方法

3.1　系统架构

跨文本的命名实体消歧任务可分解为命名实体识别，篇章内命名实体融合和跨文本的命名实体消歧。本文提出的人名消歧系统是建立在信息抽取系统之上。图1给出了信息抽取系统的基本架构，其中包含的三大功能模块：(1)基于自然语言处理技术的信息抽取系统；(2)跨文本的实体信息聚合；(3)信息抽取应用系统。为了提高系统运行效率，整个系统运行在由6台服务器组成的Hadoop平台之上，采用了Map Reduce分布式并行计算方式。

信息抽取系统通过对单一文本进行一系列自然语言处理分析，包括实体、实体关系识别，时间、地点归一化分析，别名识别和指代消解，完成文本内的命名实体的信息

对象聚合(Entity Profile merge)，并将结果保存到实体信息库。跨文本信息聚合系统在完成了跨文本命名实体消岐之后，合并相关的实体信息存回信息库。实体信息库为其他上层应用系统，如：问答系统、信息分析系统、信息网络可视化等提供支持。

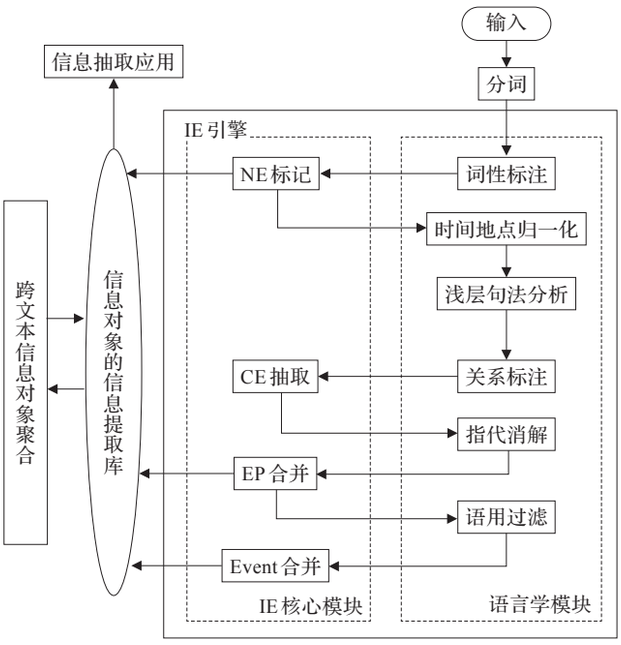

图1　系统框架图

实体信息聚合系统的关键是实体的消岐，在大规模的文本语料中大量地存在着相同的名称表示不同的实体，不同的名称代表相同的实体。跨文本命名实体消岐工作首先需要进行文本中命名实体的识别，篇章内实体消岐。本文所用的基于自然语言处理技术的中文信息抽取系统采用了有限状态转换机(FST)规则和统计机器学习相结合的方法，运用多层模块化设计思想实现了非受限域命名实体(NE)识别(时间、地点、人物、组织机构、产品)，实体关系(Correlated Entity，CE)识别，并利用别名和指代消解实现了命名实体信息对象(Entity Profile)聚合，信息抽取系统的具体技术实现细节将另文介绍。

3.2　实体信息抽取

系统使用最基本的三类命名实体作为信息抽取的中心：人物实体(NePer)，组织实体(NeOrg)，地点实体(NeLoc)。其中，与人名消歧密切相关的是人物实体和组织实体。

人物实体(NePer)在文本中表现形式主要为人物姓名，以及部分常见别名，简称等，如：

曾国藩/NePer谥号是文正，因而也被人称为文正公/NePer.

组织实体(NeOrg)包括组织机构的全名及简称，如：

中国联合网络通信集团有限公司/NeOrg(简称“中国联通/NeOrg”)于2009年1月6日/NeTIME在原中国网通/NeOrg和原中国联通/NeOrg的基础上合并组建而成，是中国/NeLoc唯一一家在纽约/NeLoc、香港/NeLoc、上海/NeLoc三地同时上市的电信运营企业。

实体信息的主要来源是实体间的关系信息，使用规则进行关系实体(CE)的抽取，本质上是模式匹配的过

程。抽取关系实体的规则中主要有两类要件:实体与限定词。定义规则,即是依照行文语法,将目标实体和限定词按照特定的顺序进行排列。当计算机查找到符合这一排列顺序的字符串时,规则生效,关系实体抽取成功。

实体关系定义:R=<ne1/feature1,ne2/feature2>。ne1是关系R中在起始位置的命名实体,feature1是它的词性,实体类别等的特征。同理,ne2是关系R中在结束位置的命名实体,feature2是它的特征。比如:

人物配偶关系:

CePerSPOUSE=<ne1/NeMan,ne2/NeWom>。

起始位置的命名实体特征为男性名,结束位置特征为女性名。

规则:ne1/NeMa(的)[妻子]([是])ne2/NeWom。

规则定义中,()表示此位置限定词允许不出现,[]表示此位置的限定词是一类词。规则中限定词[妻子]位置上允许的词有:妻子、夫人、媳妇、老婆等。()与[]同时出现表示此位置的限定词是一类词且允许其不出现。表1是满足规则的示例。

表1 人物配偶关系的示例

实体关系(A-B)	对象A	对象B	例句
丈夫-妻子	布拉德-琼斯	劳伦斯	在这支球星太太团球队中,我们看到了布拉德-琼斯的妻子劳伦斯。

本文采用随机下载了互联网上新浪新闻80篇,对信息抽取系统进行了命名实体(NE)和实体关系(CE)测试。表2给出了信息抽取系统对于人物,组织机构和地点三类命名实体的测试结果,其准确率达到了89.05%~96.93%,表3为CE关系测试结果,其准确率达到了83.33%~100%。

表2 命名实体测试结果 (%)

类型	P(准确率)	R(召回率)
人物(Per)	89.05	83.11
地点(Loc)	96.93	91.66
组织(Org)	92.77	62.98

表3 实体关系测试结果 (%)

评测指标	员工与所属单位	人与年龄	人与出生日期	人与职位	人与兄弟姐妹	人与配偶	父母与子女	人与籍贯出生地
召回率	71.28	76.32	83.78	69.36	78.57	52.63	59.09	70.00
准确率	98.53	96.67	100.00	96.00	100.00	100.00	92.86	83.33
平均F值	80.25							

3.3 实体信息对象模型

在信息抽取系统中,以命名实体和事件为中心,建立了信息对象模型Entity Profile(EP)。EP可定义为一个属性值矩阵Attribute Value Matrix(AVM),如下:

$$\begin{bmatrix} FEATURE\_1 & VALUE\_1 \\ FEATRURE\_2 & VALUE\_2 \\ \vdots & \vdots \\ FEATRURE\_n & VALUE\_n \end{bmatrix}$$

每一对属性-值通过信息抽取系统的实体关系(CE)表示,实体关系是由实体为核心的属性关系,如:所属机构,出生地点以及实体的修饰语等。在非受限领域里,定义了人物,组织机构,地点,时间,产品5大类基本实体信息对象。表4为一个文章中人物命名实体的Profile例子。

表4 profile结构

Type	PERSON_PROFILE
Name	“李静”
Sentences	“2006年多哈亚运会,香港乒乓球员李静及高礼泽,在男子乒乓球双打赛,先后击败中国的王皓及马龙,以及在决赛击败马琳及陈玘,为香港取得金牌。”
CeLocation	“香港”
CePosition	“乒乓球员”
CeCoexist_PER_PER	“高礼泽”、“王皓”、“马龙”、“马琳”和“陈玘”

在文档中,描述实体特征的关键信息非常重要,比如:人名,别名,组织名,地名,时间,产品名,联系方式(电话号码,电子邮件等)等。本文采用空间向量对profile选取的所有特征进行向量表示,以便每个profile都可以用一组特征向量所表示。一个文档的内容被看成是它含有特征项所组成的集合,对于含有 n 个特征项的文档 profile= $\boldsymbol{P}(t_1, t_2, \cdots, t_n)$,其中 t_k 是特征项,每一个特征项 t_k 都依据一定的原则被赋予一个权重 w_k,表示它们在文档中的重要程度。这样一个profile可用它含有的特征项及其特征项所对应的权重所表示:$\boldsymbol{P}=\boldsymbol{P}(t_1, w_1; t_2, w_2; \cdots; t_n, w_n)$,简记为 $\boldsymbol{P}=\boldsymbol{P}(w_1, w_2, \cdots, w_n)$,$1 \leqslant k \leqslant n$。

3.4 特征选取及相似度计算

本文将人名消歧看作是无监督的层次聚类问题。选取的特征采用权值法和空间向量模型(Vector Space Model,VSM)方法计算profile间的相似度,得到所有profile的相似度,最后,利用凝聚层次聚类算法对相似度矩阵进行聚类。

对相同人名进行消歧,最重要的就是需要选出能够区别不同人名的一些主要特征,然后通过所选的特征及其之间的相似程度,层次聚类算法可以计算出每个人名所属的类。比如,任意两个profile $\boldsymbol{P}_1=\boldsymbol{P}_1(w_{11}, w_{12}, \cdots, w_{1n})$ 和 $\boldsymbol{P}_2=\boldsymbol{P}_2(w_{21}, w_{22}, \cdots, w_{2n})$。本文选取的特征及 $\boldsymbol{P}_1$ 和 $\boldsymbol{P}_2$ 的相似度计算如下:

(1)个人信息特征:个人信息(Personal Information,PI)是识别人物身份特征的重要信息,如姓名,别名,出生日期,出生地点,居住地,Email,职位,家庭成员等。

在计算个人信息的相似度时，根据不同信息对人物特征反映出的程度不同，分别赋予不同的权重。个人信息特征相似度计算公式：

$$psim(\boldsymbol{p}_1,\boldsymbol{p}_2)=\sum_{commonce} w(ce_{1i}\cap ce_{2j}) \tag{1}$$

其中 ce_{1i} 和 ce_{2j} 分别表示 $\boldsymbol{P}_1$ 的第 i 个重要信息和 $\boldsymbol{P}_2$ 的第 j 个重要信息。

(2)关系信息特征：关系信息(Relationship Information，RI)是与人物有关的命名实体(NE)，是指和此人在同一篇文档内共同出现的人，地点，组织机构等。关系信息相似度是指由关系信息构成的空间向量的相似度，计算公式为：

$$rsim(\boldsymbol{p}_1,\boldsymbol{p}_2)=\frac{\sum_{j=1}^{J} w_{1j}\times w_{2j}}{\sqrt{\sum_{j=1}^{J} w_{1j}^2}\sqrt{\sum_{j=1}^{J} w_{2j}^2}} \tag{2}$$

其中 $\boldsymbol{w}_{ij}=tf\times\lg\frac{D}{df}$，表示由NE构成的空间向量。$w_{1j}$ 是特征 t_j 在 $\boldsymbol{P}_1$ 中的权重，w_{2j} 是特征 t_j 在 $\boldsymbol{P}_2$ 中的权重。tf 表示特征 t_j 在 $\boldsymbol{P}$ 中出现的频率，D 表示profile总数，df 表示出现该人名的profile总数。

(3)文档上下文信息特征：文档上下文信息(Document Context Information，DCI)是指在文档内的上下文信息中能够一定程度反映人物特征的信息。文档上下文信息相似度是指由文档上下文信息去除停用词后构成的向量的相似度，计算公式为：

$$dsim(\boldsymbol{p}_1,\boldsymbol{p}_2)=\frac{\sum_{j=1}^{J} w_{1j}\times w_{2j}}{\sqrt{\sum_{j=1}^{J} w_{1j}^2}\sqrt{\sum_{j=1}^{J} w_{2j}^2}} \tag{3}$$

词语构成的空间向量。词组权重的计算同样采用的是TF-IDF方法。

综上，两个人物之间的相似度为：

$$prfsim(\boldsymbol{p}_1,\boldsymbol{p}_2)=\alpha\times psim(\boldsymbol{p}_1,\boldsymbol{p}_2)+\beta\times rsim(\boldsymbol{p}_1,\boldsymbol{p}_2)+\gamma\times dsim(\boldsymbol{p}_1,\boldsymbol{p}_2) \tag{4}$$

然后根据两个profile的相似度值 $prfsim(\boldsymbol{p}_1,\boldsymbol{p}_2)$ 来判断它们是否为共指关系：

$$CO=f(prfsim)=\begin{cases}0, prfsim<threshold\\ 1, prfsim\geqslant threshold\end{cases} \tag{5}$$

其中 $threshold$ 是共指关系的置信度，即类与类之间合并的阈值。若 CO 为1，则它们是共指关系，即 $\boldsymbol{p}_1$ 和 $\boldsymbol{p}_2$ 指相同的实体，否则相反。

3.5　层次聚类算法

根据上述相似度计算方法，计算出两个profile之间的相似度，形成相似度矩阵，然后进行聚类。本文采用的是层次凝聚聚类算法进行处理人名消歧问题，类间距离计算采用的是平均距离法。公式如下：

$$csim(\boldsymbol{c}_m,\boldsymbol{c}_n)=\frac{\sum_{0<i<|\boldsymbol{c}_m|}\sum_{0<j<|\boldsymbol{c}_n|} prfsim(\boldsymbol{p}_i,\boldsymbol{p}_j)}{|\boldsymbol{c}_m||\boldsymbol{c}_n|} \tag{6}$$

聚类初始时，将每个人名对应的profile集 $\boldsymbol{P}=\{\boldsymbol{p}_1,\cdots,\boldsymbol{p}_i,\cdots,\boldsymbol{p}_n\}$ 中的每一个profile $\boldsymbol{p}_i$ 看作是一个具有单个成员的类 $\boldsymbol{C}_i=\{\boldsymbol{p}_i\}$，所以就构成了 $\boldsymbol{P}$ 的一个聚类 $\boldsymbol{C}=\{\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_n\}$，对于类 $(\boldsymbol{c}_i,\boldsymbol{c}_j)$ 之间采用上面的特征向量进行计算其相似度，然后选取相似度值最大的两个簇进行合并，形成一个新的类，即 $\boldsymbol{c}_k=\boldsymbol{c}_i\cup\boldsymbol{c}_j$，从而对于 $\boldsymbol{P}$ 形成一个新的聚类 $\boldsymbol{C}=\{\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_{n-1}\}$；重复上面的步骤，直到所有的簇间的相似度小于某个阈值或全部成为一个簇。伪代码算法如下：

```
Algorithm HAC(P)
  begin
  let each profile be in a singleton cluster {c_i}
  while |C|>1
    maxsim = max_{0<i<|c|, 0<j<|c|} csim(c_i, c_j),(i≠j)
    If maxsim ⩾ threhold
      c_m = c_i ∪ c_j
      remove c_i and c_j
    else
      break
    end if
  end while
  end
```

4　实验及结果分析

4.1　实验数据

本文使用由哈尔滨工业大学整理的基于搜狗全网新闻数据的人名消歧语料作为实验数据[10]，并选取“李静”和“李丽”的文本作为训练集，选取“王磊”和“李明”的文本作为测试集，为了进一步验证训练参数的普遍适用性，对2012年全年人民日报上的“王刚”进行人名消歧，抽取人物profile，对其进行人工标注并以Purity & Inverse Purity Metrics方法对聚类结果进行了评测。

4.2　实验评测标准

本文采用Purity & Inverse Purity评测机制。评测指标有三个：Pur、$InvP$ 及 F 值[10]。公式如下：

准确率(Precision)：

$$Pur=\frac{\sum_{S_i\in S}\max_{R_j\in R}|S_i\cap R_j|}{\sum_{S_i\in S}|S_i|} \tag{7}$$

召回率(Recall)：

$$InvP=\frac{\sum_{R_i\in R}\max_{S_j\in S}|R_i\cap S_j|}{\sum_{R_i\in R}|R_i|} \tag{8}$$

F值：

$$F=\frac{1}{\alpha\frac{1}{P}+(1-\alpha)\frac{1}{R}} \tag{9}$$

为了更好地评估实验结果，取α=0.5，用$F_{\alpha=0.5}$对P和R进行综合评测。$S=\{S_1, S_2, \cdots\}$是将要进行评测的聚类集，$R=\{R_1, R_2, \cdots\}$是人工标注的聚类集。

4.3　实验结果分析

本文对实验数据主要从三个角度分析，即确定最佳参数，不同特征组合的最佳结果对比分析和对最佳参数验证。

(1)参数调整：实验采用语料库中“李静”和“李丽”的数据集作为训练数据，用信息抽取系统对测试集进行处理，共抽取出8 847个人物实体的profile，其中“李静”和“李丽”的profile共641个。实验利用自动测试程序对个人信息、关系信息、文档上下文信息的参数及阈值四者的不同组合进行循环测试，对不同组合下得出的F值进行比较，结果确定最佳一组参数为α=0.36，β=1，γ=0.6，且*threshold*=0.28，其准确率、召回率和F值分别为94.65%、88.24%和91.33%。图2是在最佳参数下准确率、召回率和F值改变的曲线图。

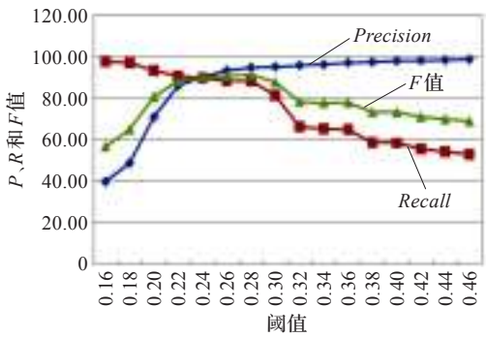

图2　P、R和F随阈值的变化曲线图

图2显示，阈值在0.2~0.3之间时，F值相对比较高，在阈值为0.28时，F值达到最高91.33%。同时准确率在逐渐提高时，召回率在逐渐减小。因为当阈值很低的时候，profile中每两个待消歧的人名就会被聚为一类，所以召回率就比较高。当阈值较高时，使原本应该聚类的profile没有聚类，导致没有正确识别出待消歧人名。

(2)特征分析：根据对个人信息(PI)、关系信息(RI)及文档上下文信息(DCI)特征选取的不同组合进行了不同实验，得到的最佳结果如表5所示。

表5　Purity & Inverse Purity评测机制实验结果统计　(%)

特征	*Pur*	*InvP*	*F*
PI	98.75	40.02	56.87
PI+RI	96.43	84.31	89.97
PI+RI+DCI	94.65	88.24	91.33

以上三种特征组合下的F值如图3所示。

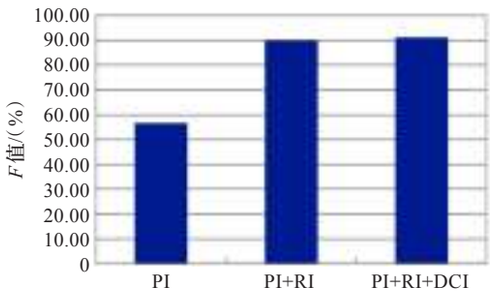

图3　以上三种特征组合下的F值

从表5可以看出，当仅使用个人信息特征时，准确率尚可，但召回率较低，说明个人信息虽能表示一个人的身份，但由于语料中出现的个人信息特征比较少，所以造成召回率比较低。在加入实体关系特征时，召回率提高了44个百分点，说明语料中使用实体关系特征就能够较好地表示一个人的身份，且语料中实体关系较多。同时，在构建社交网络时，实体关系信息特征会起着至关重要的作用。比如：

(1)四川省出席党的十八大代表大会的有丁爱谱、王坚、王志强、李静、刘作明、宋朝华、吴小可等72名成员。

(2)李静、宋朝华和吴小可等出席签约仪式。

此时(1)和(2)中的“李静”并没有明显的个人信息特征，但两次均与“宋朝华、吴小可”两人共同出现，说明两个“李静”是同一个人。

同时使用三种信息特征时，召回率有所提高，准确率稍微下降，是因为在语料中添加能够反映人物特征的信息比较多，但这些特征对于每一个人不具有普遍性，所以造成提高了召回率，准确率下降了1.78个百分点，但总体评测标准F值还是有所提高。而且图2也显示了在三种情况下，F值也是逐渐提高的。

(3)参数验证：对哈尔滨工业大学整理全网新闻数据语料中，选取“王磊”和“李明”进行人名消歧，分别抽取出6 632和14 376个人物实体的profile，对其进行聚类，并采用Purity & Inverse Purity进行评测，同时分别加入维基百科上的“王磊”和“李明”的profile，共6 715和14 407个，用同样的方法进行验证，取得结果分别如表6和表7所示。

表6　两个人名实验测试结果　(%)

消歧人名	*Pur*	*InvP*	*F*
王磊	95.87	82.53	88.70
李明	95.06	80.68	87.28

表7　两个人名加入百科后实验测试结果　(%)

消歧人名	*Pur*	*InvP*	*F*
王磊	95.02	84.58	89.50
李明	94.11	82.56	87.96

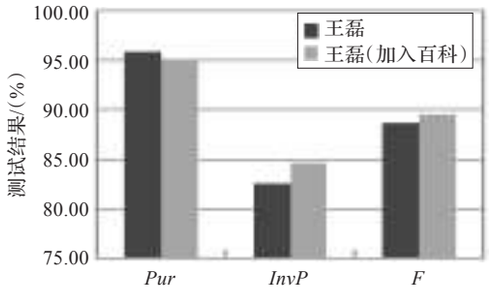

图4(a) 王磊在加入百科前后比较

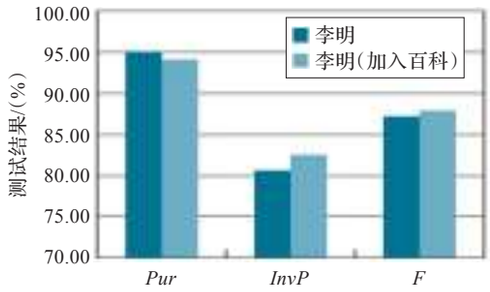

图4(b) 李明在加入百科前后比较

实验结果测试“王磊”和“李明”的*F*值分别为88.7%和87.28%，取得相对比较理想的结果，在加入维基百科数据后，测试*F*值分别为89.5%和87.96%，比未加之前分别提高了一个百分点。维基百科中的数据比较规范，更新比较快，且能够抽取出更为丰富的个人信息和关系信息特征，所以评测的结果显示召回率提高了2个百分点，如图4所示。在加入维基百科数据以后，评测结果说明采用本文系统训练出的这组参数具有普遍适用性。

同时，对2012年全年的人民日报上的“王刚”进行人名消歧，共抽取出54 782个人物的profile，采用同样的方法并对其中533篇“王刚”的profile进行了聚类，聚类结果是6类实体profile，并且对聚类结果进行评测，取得非常好的结果，如表8所示。

表8　人民日报实验测试结果　(%)

消歧人名	*Pur*	*InvP*	*F*
王刚	100.00	99.25	99.62

人民日报的数据集是web数据集上的一个子集，数据源相对比较规范，人物报道相对比较集中，多数profile只通过个人信息和关系信息特征就很容易合并。实验结果表明，本系统在较为规范的数据集下有非常满意的测试结果。

5　结束语

本文主要解决了自然语言处理中的人名消歧问题，采用了基于凝聚层次聚类的方法，通过对个人信息、关系信息及文档上下文信息特征提取，这三个特征基本能够确定一个人的身份，实验通过训练集对部分数据测试，得到一组最佳参数，再用这组参数去测试剩下的数据，为了进一步证明本文方法的适用性，还采用了2012年的人民日报进行测试，均取得比较好的实验结果。

当然，本文的系统还不够完善，在下一步的研究工作中，打算结合互联网上的知识数据库进行进一步研究，改进目前的跨文本实体信息聚合系统。

参考文献：

[1] Gao Liqi，Zhang Yu，Liu Ting，et al.Word sense language model for information retrieval[C]//AIRS，2006.

[2] 李保利，陈玉忠，俞士汶.信息抽取研究综述[J].计算机工程与应用，2003，39(10)：1-5.

[3] McCarthy，Lehnert W.Using decision trees for coreference resolution[C]//Proceedings of the Sixth Message Understanding Conference(MUC-6)，1995.

[4] Bagga A，Baldwin B.Entity-based cross-document coreferencing using the vector space model[C]//Proceeding of the 17th International Conference on Computational Linguistics，Canada，1998：79-85.

[5] WePS-3 workshop program[EB/OL].(2010-07-10).http://nlp.uned.es/weps/.

[6] Task3 Chinese version[EB/OL].(2010-10-16).http://www.clpsc.org.cn/clp2010/task3_ch.htm.

[7] 周晓，李超，胡明涵，等.基于人物互斥属性的中文人名消歧[C]//第六届全国信息检索学术会议(CCIR)，2010：333-340.

[8] 丁海波，肖桐，朱靖波.基于多阶段的中文人名消歧聚类技术的研究[C]//第六届全国信息检索学术会(CCIR)，2010：316-324.

[9] 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(7)：1365-1375.

[10] 王鑫.人名消歧关键技术研究与实现[D].哈尔滨：哈尔滨工业大学，2012.

[11] Shingo O，Issei S，Minoru Y.Person name disambiguation in Web pages using social network，compound words and latent topics[C]//LNAI 5012：PAKDD2008，2008：260-271.