基于属性信息的中文人名消歧*

李 丽，孙甲申，王小捷，李 江，宋占江
北京邮电大学 计算机学院 智能科学技术中心，北京 100876
诺基亚北京研究院，北京 100176
E-mail:wbg111@126.com,b; bigart911@gmail.com; xjwang@bupt.edu.cn
li.jiang84@gmail.com; zhanjiang.song@nokia.com

摘 要：本文针对中文人名消歧任务提出了一个人名属性抽取的方案，对无法获取属性的文本，利用《知网》进行属性推理。为了更准确地计算文档间属性值的相似度，利用《同义词词林》扩展了职业属性值。通过分析不同属性的作用，采用信息增益对属性进行差异化处理。针对人名属性消歧的特点，提出了“双阈值”的聚类算法。实验结果表明，本文提出的方法的消歧性能超过了现有的最好方法。

关键词：属性抽取；属性扩展；信息增益；双阈值；人名消歧

Chinese Personal Name Disambiguation Based on Attribute Information

Li Li, Sun Jiashen, Wang Xiaojie, Li Jiang, Song Zhanjiang
Center for Intelligence Science and Technology, School of Computer, Beijing University of Posts and Telecommunications, Beijing 100876
Nokia Research Center, Beijing 100176
E-mail:wbg111@126.com; b.bigart911@gmail.com; xjwang@bupt.edu.cn
li.jiang84@gmail.com; zhanjiang.song@nokia.com

Abstract: This paper proposes a Chinese personal name disambiguation method based on personal attribute information. This method employs Hownet to deduce attributes for texts without attributes. It expands attribute values using TongYiCiCiLin to get more accurate similarity of documents' attribute values. By analyzing influence of different attributes, we use information gain to address attributes differentiation. A "double thresholds" clustering algorithm is developed for addressing the characteristic of personal name disambiguation. Experimental results show that our method outperforms the-state-of-the-art methods.

Keywords: attributes extraction; attributes expansion; information gain; double thresholds; name disambiguation

1 引言

随着网络信息的爆炸式增长，人类对网络信息搜索的诉求迅速增长，特别是对人名的搜索。据统计，在 google 或者 yahoo 搜索引擎上人名搜索量占 30%之多[1]。在人名搜索中，同名问题（不同的人具有相同姓名）是影响搜索性能的一个主要因素。由于同名问题的存在，在搜索引擎上搜索一个人名，得到的结果中很可能包含多个不同人的信息。对于中文人名搜索，除了同名人的问题，还存在异名人和非人名的问题，例如，搜索人名“高军”时，会返回“高军红”等异名人信息或“最高军事法庭”等非人名的信息。这些问题和同名问题纠结在一起，增加人名消歧的复杂性。

目前，国外对于人名消歧做了很多的研究工作。WePS(Web People Search)是针对英语人名消歧任务的评测，已分别在 2007 年、2009 年和 2010 年举办了三届。前两届只要求抽取属性信息，不要求基于属性信息进行消歧，而可以采用其他信息消歧。2010 年则要求将属性抽取和聚类相结合

* 本文得到了国家自然科学基金(90920006)，高等学校博士点基金(20090005110005)以及 BUPT-Nokia 合作项目的支持。

进行消歧。在 WePS 中定义了 18 类属性信息。针对属性信息抽取和聚类出现了不少的方法。

抽取属性信息的方法有：1）人工定义模板。例如，was born in<.*>作为出生地的模板。2）基于 NER(name entity recognition)抽取人名、地名、机构名。3）字典匹配算法，建立字典（如职业，学校等），在文本中匹配得到属性值。[2][3][4]将三种方法都应用到属性抽取中，将 18 类属性分成三类或四类，分别采用上述三种方法进行属性值的获取。采用人工定义的模板来获取属性数量较少且片面，采用 NER 获取的地名和机构名不一定和当前人名相关，利用字典匹配得到的属性值会出现同样的问题。[5][6]将获取属性值分为两步：第一步是候选属性值获取，第二步是候选属性验证。第一步和上述方法一致，第二步采用分类器，对于每个候选属性值，用训练好的分类器判断该属性属于哪个对象。这种方法对分类器的性能要求很高。

属性信息聚类的方法有层次凝聚聚类（HAC：Hierarchical Agglomerative clustering）和基于图的聚类。[6]将属性信息分为三类，每一类的属性采用不同的权重，采用自底向上的层次聚类方法进行聚类。这种方法对不同属性进行了粗略的区分，但是仍然忽略了不同属性对于人名消歧结果的影响。[7]采用图模型进行聚类，将属性值表征为图节点，边为属性值的共现，计算最小类之间的连接强度，如果达到阈值则聚成一类，直到遍历完所有的最小类。这种方法没有考虑无属性信息的文本。最近，[8]提出两步聚类算法，第二步采用 bootstrapping 算法，这种方法对命名实体抽取的性能要求高。

CIPS-SIGHAN 于 2010 年举办了首次中文人名消歧（PND: Personal name disambiguation）的评测。其中融合属性信息进行人名消歧的方法取得了较好的性能。[9]融合属性信息和共现词等信息进行多步消歧，属性信息获取方法也采用人工模板和 NER 工具等。在聚类过程中，首先基于大量属性相关的规则进行消歧，然后采用 HAC 的方法利用共现词进行消歧。该聚类方法中忽略了不同属性对于人名消歧结果的重要性，同时大量的规则使得这种方法对于语料的依赖性很大。

可以看到，人物属性信息对于人名消歧具有重要作用，但是属性的获取质量、各种属性在消歧中的不同作用以及如何利用属性信息进行聚类还有待进一步研究。

本文提出了一个人名属性信息抽取的方案，对无法获取属性信息的文本，利用《知网》进行属性的推理。为了更准确的计算文档间属性值的相似度，提出了利用《同义词词林》扩展职业属性值，使用同义词或相关词集合代替单个属性值。通过对不同属性信息的作用进行了分析，提出了利用信息增益进行属性的差异化处理，衡量了不同属性对于人名消歧的影响度。在基于属性的聚类算法中，针对基于属性信息的人名消歧的特点，提出了“双阈值”的判定方法。实验结果表明，本文提出方法的消歧性能超过了现有的最好方法。

2 方法

2.1 属性获取

本文定义了 17 类人名属性，这 17 类属性是基于 WEPS 的 18 类属性，并结合中文自身的特点而设定的，见表 1(第二列)。

为了克服人工定义模板获取属性的片面性和数量少的缺点，以及避免字典匹配和 NER 得到无关信息，本论文依次采用如下四种属性获取的方法。

一、基于 bootstrapping 算法的属性抽取。本方法基于无标注语料自动获取属性模板。本文利用 Bootstrapping[10]方法，自动获取属性模板。算法如下：

i＝1；i 代表循环次数；

{

1. 利用新种子模板得到属性值；

2. 计算新的候选属性值的可信度；
3. 将可信度最高的前三个属性值放入属性值字典；
4. 利用这三个属性值，遍历所有的文本，得到属性值的上下文（取上文和下文的五个字），将这些上下文作为候选模板。
5. 计算新的候选模板的可信度。
6. 将可信度最高的前三个模板加入模板列表中。
7. $i:=i+1$
8. 返回第一步。

}

停止条件是：$i>10$ 或者无新的模板或者属性信息产生。之后，将得到的模板遍历所有的文本，抽取属性值。

二、含有限制条件的字典匹配方法。本文所用字典含 458 个职业称谓。方法执行步骤如下：

1. 在人名紧邻上下文中匹配职业字典，得到该人名的候选职业。

2. 在候选职业前查找机构名或者机构名后缀（(.*)公司，(.*)大学等）。例如："北京邮电大学教授张三参加了此次研讨会。""教授"存在于职业字典中，则职业的属性值为"教授"，同时，在职业紧邻的上文中出现了公司名属性的后缀"大学"，则"北京邮电大学"作为公司名的属性值。该方法在人名的上下文中寻找属性信息，使得属性值都与当前人名错误匹配的概率降低。

三、通过 NER 工具抽取文本中的其他人名，作为关系属性的属性值。该方法基于这样一种假设：人名上下文中的其他人名能够较好的反映该人名所对应个体的领域和环境，从而更好的区别不同的个体。

四、属性值推理。对于无法获取属性的文本，借助《知网》词典中概念与其相关概念场之间的关系，在《知网》中查找该人名下其他文本获取的职业属性，将其相关概念场作为该职业属性值的扩展值（单字词除外），然后将这些扩展值在无属性文本中进行匹配，如果匹配上的扩展值个数超过一定阈值，则此属性值作为无属性文本的职业属性值。例如："影星"的相关概念有"扮相、扮演、饰演、影院"等，将这些相关概念及"影星"在无属性文本中进行匹配，如果"饰演、娱乐圈、影院"在无属性文本中出现，则该无属性文本的职业属性值为"影星"。

2.2 属性扩展

由于语言中用词的灵活性，为了更准确地衡量属性值的匹配程度，本文利用《同义词词林》[11]作为语义体系，对职业属性进行词扩展，使用同义词或相关词集合代替单个属性值，从而提高文档间属性值的匹配度。

本文将与查询词具有相同的第四级级别的词集作为该词的扩展词集，单字词除外。例如："演员"所在行的编码为"Ae17A01"，将具有编码"Ae17A"的所有词作为"演员"的扩展词集，扩展词集的编码行有"Ae17A01"、"Ae17A02"、"Ae17A03"、"Ae17A04"、"Ae17A05"等等。之后通过扩展词集来计算词之间的语义相似度。如果两者扩展词集的交集达到一定阈值，则认为两词的语义相近。

2.3 属性聚类

在已有的属性权重计算中，均忽略了或者没有全面考虑不同属性对于人名识别的差异性。而实际上，在人名识别时，不同的属性在区分不同的人时的重要性是不一样的。例如，职业提供的信息量多于星座。本文认为，属性的权值由其对不同人的区分度来决定，利用（1）计算的信息增益作为每个属性的权重。S_i 代表每个人名集合，A_j 代表每种属性，N 代表人名数 Gain (S_i, A_j 为属性

A_j 相对于子集 S_i 中 A_j 的信息增益，Entropy (s_i)是 S_i 相对于分类的熵。

$$\text{Gain}(S, A_j)' \equiv \left(\sum_{i=N}^{i=1} \frac{\text{Gain}(s_i, A_j)}{\text{Entropy}(s_i)} \right) / N \quad (1)$$

基于上述权重对属性向量进行加权后，进行聚类。我们首先利用 CRF_CLUSTER 工具[12]进行人名识别，将所有异名人划分开，这样只需处理同名人的消歧问题。

然后，采用“双阈值”（“兴奋”和“抑制”）方法进行聚类，其中“兴奋”指文档间相同属性值的权重阈值，“抑制”指文档间不同属性值的权重阈值。算法描述如下：

1. 将所有文档中的属性存入含有 17 列的矩阵中，无属性则对应列为空，第一列为标志位（初始值为行号），标志当前文档是否被合并（如果标志位为行号则表示未合并，否则代表此文档被合并到标志位指向的行），其余的列表示属性，行表示每篇文档。

2. i=1；合并从第一行开始。

3. j=1；j 代表被比较行。

4. 如果 i=j 则跳到第 6 步，否则，比较第 i 行和第 j 行的属性值，如果第 k 列属性值相同，则 right+=A_k (right 代表所有相同属性的权重之和，A_k 代表第 k 列属性所占的权重)；如果第 k 列属性值不相同，则 wrong+=A_k (wrong 代表所有不同属性的权重之和)。由于对职业属性进行了扩展，在比较职业词集相似度时，如果交集的词个数大于等于 5，则两职业属性值相同。

5. 如果 right>P1&&wrong<P2（P1 为“兴奋”阈值，P2 为“抑制”阈值），则代表 i 和 j 是关于同一个人的文档，判断第 i 行的标志位是否为 i 行行号，如果是的话，将 j 行属性合并到 i 行中去，并将 j 行的标志位改为 i。如果不是的话，将 j 行属性合并到 i 行标志位指向的行 k 行，并将 j 行的标志位改为 k。如果不满足阈值，则进行下一步。

6. j++；如果 j<=row_num（row_num 为所有行数），则跳转至第 4 步，否则跳转至第 7 步。

7. i++，如果 i<=row_num（row_num 为所有行数），则跳转至第 3 步，否则程序结束。

本聚类算法不同于传统的 HAC 算法，采用“双阈值”更能体现基于属性信息的消歧方式，即使在两个文档间大多数属性值相同，只要个别重要属性值不同也能将其区分出来。

3 实验设计

为了评价本文的人名消歧方法，本文采用了两种不同的实验数据：PND 语料[13]和网络语料，PND 语料是 CIPS-SIGHAN 评测提供的训练和测试语料。网络语料是从百度抓取的 34 个人名的搜索结果网页，先对网络语料进行预处理（去除了网页中的 HTML 标签），并人工标注属性信息以及所属个体类别。实验采用 CIPS-SIGHAN[13]提供的两种评价方法：B-cubed 和 I IP。其中 B-cubed 评测方法是目前最佳的一种评测方式[13]。实验中利用 PND 提供的评测程序进行评测。

3.1 PND 语料

在 PND 语料的测试语料中，通过上述四种获取属性的方式获取属性信息，各种属性所占比重见表 1 第三列（PND 中各属性所占比重）。测试语料中文本总数为 5416 篇，含有属性的文本数为 5401 篇，属性的文档覆盖率为 99.72%。覆盖率=$\frac{\text{含有属性的文本数}}{\text{文本总数}} \times 100\%$。

针对 PND 语料进行了下列实验。为了与其他消歧系统做比较，本系统将 CIPS-SIGHAN 提供的初步聚类结果作为统一基准（初始结果）。为了评测属性扩展和属性权值的影响，分别评测了未加入属性扩展和信息增益（未属性扩展+无信息增益），加入信息增益（+信息增益），加入属性扩展和信息增益（+属性扩展+信息增益）的结果，同时和 CIPS-SIGHAN 评测中第一名的测试结果

(Top1 in CIPS-SIGHAN)作比较，具体结果见表2。在聚类过程中，本文采用“单阈值”和“双阈值”进行对比实验，结果见表3，通过在训练语料中训练得到最佳阈值，双阈值为“0.900/0.850”，单阈值为“0.900”。

实验结果显示，本文方法比top1高0.04个百分点。从属性所占的比例可以看出，职业、公司名和关系属性占据了98%以上，其他属性信息非常稀少，实验表明，职业、公司名和关系对人名消歧的影响非常重要，同样说明基于属性的人名消歧方法的合理性和可行性。采用属性扩展后，性能提升了1.08个百分点，说明属性扩展的有效性，但是效果不显著的原因在于PND语料是新闻类语料，语言表达很统一，较少出现多词一义的现象。在进行属性权值差异化处理后结果没有变化，因为职业、公司名和关系都对消歧具有很大的影响力，在其他属性几乎没有的前提下，无法体现其他属性的辅助作用。在聚类过程中采用“双阈值”更能区分不同的人，即使文档间大多数属性值相同，存在个别重要属性值不同，也能分别出来。这种情况下，“单阈值”无法区分。

表1　属性列表

属性	各属性的权重	PND中各属性所占比重	网络语料中各属性所占比例
外文名	0.323	0.00%	0.68%
别名	0.677	0.00%	0.88%
性别	0.842	0.04%	7.88%
出生日期	0.988	0.09%	6.52%
血型	0.226	0.00%	0.28%
星座	0.420	0.00%	0.95%
身高	0.644	0.42%	2.17%
出生地	0.990	0.04%	6.41%
民族	0.659	0.00%	2.32%
国籍	0.385	0.00%	0.54%
政治面貌	0.792	0.08%	2.44%
学校	0.950	0.04%	8.18%
学历	0.821	0.08%	5.41%
职业	0.908	42.40%	31.78%
公司名	0.994	34.68%	19.45%
现居地	1.000	0.00%	4.09%
关系	0.870	21.71%	0.00%

表2　PND语料的测试结果

评测	B-cubed		
	正确率	召回率	F值
初始结果	71.11%	100%	80.92%
未属性扩展+无信息增益	95.74%	87.60%	91.10%
+信息增益	95.74%	87.60%	91.10%
+属性扩展+信息增益	95.33%	89.76%	92.18%
Top1 in CIPS-SIGHAN	95.6%	89.74%	92.14%

表3　阈值比较

评测	B-cubed		
	正确率	召回率	F值
双阈值	95.33%	89.76%	92.18%
单阈值	85.36%	94.44%	88.73%

3.2 网络语料

如上所述，PND 语料是规范的新闻语料，本文方法的一些特点并没有完全体现，为此，我们在实际网络语料中进行了实验。网络语料的文本总数为 4573 篇，人工标注含属性文本数为 3417 篇，利用上述前两种属性抽取方式得到的含属性文本数为 2566 篇。标准属性的文本覆盖率为 74.72%，系统得到属性的文本覆盖率为 56.11%。系统抽取属性的正确率为 28.12%，召回率为 25.95%，F-1 值为 26.99%。表 1 第四列（网络语料中各属性所占比例）显示了各种属性所占的比例。抽取属性的性能较低，一方面是噪声数据多等自身特点，抽取到的属性值不一定与当前人名相关，另一方面是在评价抽取属性时，采用完全匹配的方式，忽略了属性多词一义的现象。

从获取属性的结果可以发现，有 2007 篇网页是未抽取到属性值。本文采用第四种方式这 2007 篇网页进行属性值的推理，实验结果显示，其中 842 篇网页获得属性值。对 2566 篇含有属性值的网页单独进行聚类，结果见表 4 的第二行（2566 行）；推理前，将 2566 篇和 2007 篇网页合在一起进行聚类，结果见表 4 的第三行（2566+2007 行）；推理后，将 2566 篇、842 篇和 1165 篇网页合在一起进行聚类后的结果见表 4 第 4 行（2566+842+1165 行）。实验结果表明，利用《知网》进行属性的推理减少了无法抽取属性的文本数量，同时提高了性能。

在上述实验的基础上（Baseline）进行了权重的差异化（+信息增益）；对属性值进行了扩展和差异化（+属性扩展），实验结果见表 5。

表 4　无属性文本的影响

评测	B-cubed		
	正确率	召回率	F 值
2566	78.76%	73.80%	73.36%
2566+2007	77.66%	59.59%	61.26%
2566+842+1165	73.40%	61.67%	63.16%

表 5　网络语料的测试结果

评测	B-cubed		
	正确率	召回率	F 值
Baseline	73.40%	61.67%	63.16%
+信息增益	79.65%	62.38%	68.52%
+属性扩展	82.52%	66.07%	70.70%

网络语料相比 PND 语料而言，对属性权重进行差异化处理的效果明显。原因在于，虽然“职业”等属性对于消歧的影响度很大，但是其他属性信息的存在具有很大的辅助的作用，体现差异化处理的优势。同样，对属性进行扩展也提高了系统性能，这是因为在网络语料中，文章内容书写格式自由，使用词汇更为自由，会造成语言表达的不一致性，对属性进行扩展在一定程度上缓解了这种不一致性。但是提高幅度不大，主要原因职业属性值在《同义词词林》的覆盖率不高。

比较两种语料可以发现，虽然网络语料含的属性信息很丰富，但是网络语料的测试结果明显低于 PND 语料，原因是网络语料本身的一些特点：第一点，网页内容包含该人名，可能内容并不与此人相关，或者存在多个人的描述，使得抽取到的属性值不一定与当前人名相关。第二点，网页中含有其他无关的信息，例如广告，链接锚文本等，使得网页清洗的质量不高。第三点，网络语料的语言表达风格更为自由多样，而且存在信息不精确现象，例如，有些网页显示该人名身高为 182cm，另有一些网页却显示为 180cm。

4 总结与展望

本文提出了一个进行人名属性信息抽取的方案，对无法获取属性信息的文本，利用《知网》进行属性的推理，提高了属性的获取量以及精确度。为了更准确的计算文档之间属性值的相似度，提出了利用《同义词词林》扩展职业属性值。通过分析不同属性信息对人名消歧的作用，提出了利用信息增益进行属性的差异化处理，衡量了不同属性对于人名消歧的影响度。针对基于属性信息的人名消歧的特点，提出了采用“双阈值”的聚类算法。实验结果表明，本文提出的方法能够提升消歧的性能，并且超过了现有的最好方法。同时，基于属性信息的人名消歧能够获得人名的结构化信息，使得人名消歧具有可解释性。

虽然目前的工作提升了人名消歧的效果，但是人名消歧仍然是具有挑战性的研究任务，特别是针对真实网络语料。在未来的工作中，将在如下三个方面开展进一步的研究：第一点是属性的抽取，目前属性抽取的精度还较低，也不够全面，在 WePS-3 国际评测中，最优系统的 F 值只有 18%，足以体现该任务的难度。第二点是对于无法获取属性网页，扩大属性推理的范围，使得一般词也可以作为特征以指导聚类；第三点是属性扩展的算法，目前采用的字典还不够全面，今后需要利用更加有用的知识库或者广泛的语料库来提高算法的精度和适用程度。

参 考 文 献

[1] J. Artiles, J. Gonzaloand F. Verdejo, “A testbed for people searching strategies in the www” in Proceedings of the 28th annual International ACM SIGIR conference on Research and Development in Information Retrieval(SIGIR2005), 2005, pp.569-570.

[2] Man Lan and YezheZhang. Which who are They? People Attribute Extraction and Disambiguation in Web Search Results*. In 2nd Web People Search Evaluation Workshop(WePS 2009), 18th WWW Conference, 2009.

[3] Keigo Watanabe, Danushka Bollegala al. A Two-Step Approach to Extracting Attributes for People on the Web, In 2nd Web People Search Evaluation Workshop(WePS 2009), 18th WWW Conference, 2009.

[4] KrisztianBalog,Jiyin He et al.The University of Amsterdam at Weps2, In 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference, 2009.

[5] Xianpei Han, Jun Zhao. CASIANED: People Attribute Extraction Based on Information Extraction, In 2nd Web People Search Evaluation Workshop(WePS 2009), 18th WWW Conference, 2009.

[6] István T. Nagy, Richárd Farkas.Person attribute extraction from the textual parts of Web pages.In Third Web PeopleSearch Evaluation Forum (WePS-3), CLEF 2010, 2010.

[7] Lili Jiang and JianyongWang. GRAPE:A Graph-Based Framework for Disambiguating People Appearances in Web Search. In 2009 Ninth IEEE International Conference on Data Mining, 2009.

[8] Minoru Yoshida, Masaki Ikeda et al. Person name disambiguation on the web by two-stage clustering. In *WWW*, April 2009.

[9] Huizhen Wang and Haibo Ding. A multi-stage Clustering Framework for Chinese Personal Name Disambiguation.In Chinese News. In Proceedings of CIPS-SIGHAN, Beijing, China, 2010.

[10] Michael Thelen and Ellen Riloff. A Bootstrapping Method for Learning Semantic Lexicon using Extraction Pattern Contexts. In Proceedings of the Conference on Empirical Methods in Natural language Processing. 2002.

[11] 《同义词词林》，哈尔滨工业大学.

[12] Jiashen Sun, Tianmin Wang and Li Li. Person Name Disambiguation based on Topic Model. In Chinese Iformation Processing Society of China and SIGHAN(CIPS-SIGHAN), 2010.

[13] Ying Chen, Peng Jin al.The Chinese Persons Name Disambiguation Evaluation: Exploration of Personal Name Disambiguation in Chinese News. In Chinese Information Processing Society of China and SIGHAN(CIPS-SIGHAN), 2010.