一种基于关联聚类的汉语共指消解方法*)

李元龙　周俊生　陈家骏
(南京大学计算机软件新技术国家重点实验室　南京 210093)
(南京大学计算机科学与技术系　南京 210093)

摘　要　指代消解是自然语言处理领域中的一个重要问题。本文引入图对汉语名词短语的指代消解问题进行建模，将指代消解问题转化为图划分问题，并应用关联聚类算法来实现对图的自动划分。相对于传统的 link-first 和 link-best 聚类机制，该方法并不是孤立地针对每一对名词短语分别进行共指决策，而是充分考虑了多个名词短语之间的相关性，且不需事先给出聚类的数量以及距离阈值。通过在 ACE 中文语料上名词短语消解的实验结果表明，该方法是一个有效的指代消解算法。

关键词　指代消解，共指，关联聚类，线性规划

Applying Correlation Clustering to Chinese Noun Phrase Coreference Resolution

LI Yuan-Long　ZHOU Jun-Sheng　CHEN Jia-Jun
(National Laboratory of Novel Software Technology, Nanjing University , Nanjing 210093)
(Deptartment of Computer Science and Technology, Nanjing University , Nanjing 210093)

Abstract　Coreference resolution plays an important role in natural language processing. In this paper, coreference resolution is converted to a graph clustering problem firstly, and then correlation clustering is used for automatic graph clustering. Compared with the traditional clustering approaches: link-first and link-best, the proposed algorithm takes the relations among the NPs into account sufficiently. In addition, it does not need to specify the desired number of clusters and a distance threshold. The experimental results on the ACE Chinese training corpus demonstrate that the proposed method of coreference is an effective one.

Keywords　Reference resolution, Coreference, Correlation clustering, Linear programming

1　引言

指代是自然语言中一种非常普遍和常见的语言现象，文本的概念关联性在很大程度上就是通过指代关系来刻画的。指代消解的过程实际上是建立概念关联的过程，是文本处理的核心问题之一[1]。随着自然语言处理应用的日益广泛，特别是对文本处理需求的进一步增加，指代消解的重要性越来越突出，在信息抽取、问答系统、文本摘要、机器翻译以及对话解释系统等方面都有很大的作用。指代一般可分为两种情况：回指(Anaphora)和共指(Coreference)。所谓回指，是指当前的指示语与上文出现的词、短语或句子(句群)之间存在语义关联性；共指则主要是指两个名词(包括名词短语、代名词)指向现实世界中的同一参照物。本文所讨论的指代消解概念属于共指消解。

Soon 等曾提出通过一种经典的有监督机器学习方法进行名词短语的共指消解[2]，该方法首先对于训练语料中的每个名词短语对都用属性特征去描述，然后使用一个学习算法学出一个分类器，把测试文档中按一定策略结合的名词短语对提交给该分类器，由分类器给出共指决策，根据最近优先策略(link-first)将共指的名词短语合并成一个聚类。Ng and Cardie 又给出了改进方法[3]，该方法根据最佳优先策略(link-best)把共指的名词短语并入聚类中并对与知识源相关的特征集进行了增强和选择。在应用有监督机器学习方法解决中文指代消解的研究方面，李国臣等[4]提出了一种采用优先选择策略的汉语人称代词指代消解方法。这些方法存在的不足是：在聚类过程中每次都是分别针对一对名词短语进行共指决策，而实际上各对短语的共指决策之间并不是相互独立的。针对上述问题，Luo 等提出了通过自底向上搜索 Bell 树给出聚类决策的方法[5]将聚类过程转化成对树的搜索，但搜索过程是局部的和启发式驱动的，而且不能处理逆照应现象。

为此，本文提出了一种基于图的汉语指代消解算法，采用图来对名词短语的指代消解问题进行建模，将指代消解看成图聚类过程，从而将指代消解问题转化为图划分问题，并使用关联聚类算法对图进行划分。本文所提出的基于图划分的指代消解算法并不是孤立地对每一对名词短语分别进行共指决策，而是充分考虑了多个待消解项之间的相关性，从全局的角度实现对共指等价类的划分。

2　关联聚类

关联聚类算法就是一种对图划分，即对图中顶点聚类的方法[6]。给定一个有 n 个顶点的图，图中每一条边 (u,v) 都标有〈+〉或〈−〉，〈+〉表示顶点 u 和顶点 v 之间相似，〈−〉表示

*)国家自然科学基金项目(60673043)、国家 863 高技术研究发展计划(2006AA01Z143、2006AA01Z139)。**李元龙**　硕士研究生，研究方向为自然语言处理；**周俊生**　博士研究生，主要研究方向为自然语言处理、信息抽取；**陈家骏**　教授，博士生导师，研究方向为自然语言处理、机器翻译、软件工程。

u 和 v 之间不相似。图中每一条边有一个权值 $c_e \in [0,\infty]$。我们定义标记〈+〉的边为“正边”，标记〈−〉的边为“负边”。这种定义不涉及权值，权值永远是与标记无关的非负值。用 $e(u,v)$ 表示 (u,v) 边的标记(〈+〉或〈−〉)，用 $E^{(+)}$ 表示“正边”的集合以及 $G^{(+)}$ 用表示由 $E^{(+)}$ 生成的仅有“正边”的图，$E^{(+)}=\{(u,v)|e(u,v)=\langle+\rangle\}$，$C^{(+)}=(V,E^{(+)})$。类似定义 $E^{(-)}$ 和 $G^{(-)}$。

图中权值大(表示顶点之间的关联度强)且标记为〈+〉的边促使它们的顶点属于同一个聚类，而权值大且标记为〈−〉的边促使它们的顶点属于不同的聚类。目的就是要得到一种顶点的划分，该划分尽可能地与边的标记一致。关联聚类就要求最后的聚类结果满足一致性权值最大原则或不一致性权值最小原则。一致性权值是指，聚类内部标记〈+〉边的权值与聚类间标记〈−〉边的权值的总和。同理，不一致性权值是指，聚类内部标记边的权值与聚类间标记边的权值的总和。例如，图 1 中我们得到了一个已优化过的聚类结果，其中发生错误的是两条标记为〈+〉的边和一条标记为〈−〉的边，它们的总权值为 5。

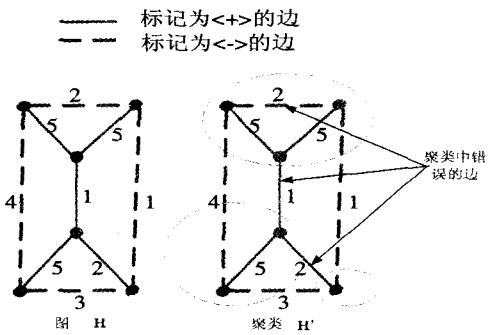

图 1　带权图中的聚类实例

由于关联聚类既不要求事先给出聚类的数量，也不需选择距离阈值，再加上充分考虑了图中顶点之间的相关性，因此是一种有效的聚类方法。

3　基于关联聚类的共指消解算法

3.1　基本思想

我们把测试语料中的各个待消解项作为图中的顶点，将两个待消解项之间的置信度值看成是连接这两个顶点的边所对应的权值，这样就构成了图 $G(V,E)$；然后使用关联聚类算法对图 G 进行自动划分，这样将指代消解的过程转化成对该图 G 的划分过程。本文选择最小化不一致性权值作为关联聚类的目标，在求解关联聚类问题时，首先将整数规划问题松弛为线性规划问题，再应用一个取整算法给出最终的聚类决策。

这种基于图划分所进行的指代消解并不是孤立地对每一对名词短语分别进行共指决策，而是充分考虑了多个待消解项之间的相关性。当确定将图中某个顶点(待消解项)分配至某个聚类时，不是单单地依赖该顶点与某一个其它顶点的置信度，而是依赖于该顶点与这个目标聚类中的所有顶点甚至其它聚类中所有顶点的置信度关系。

3.2　学习算法

上述图中的对应各边权值的共指置信度值是通过 C4.5 决策树分类算法获取的[7]。我们首先通过一个预处理器扫描训练语料和测试语料[8]，抽取消解项和待消解项所具有的特征信息。消解项之间按照某种策略形成训练例，每个训练例都是以 $i\{NP_j,NP_i\}$ 的形式表示。在训练时，对于训练文本中每一个指示语 NP_j，它与离它最近的先行语 NP_i 构成一个正例，它与 NP_j 和 NP_i 之间的所有 NP 构成反例集，依此构成训练例的集合。随后就用该集合训练出一个决策树分类器。在消解时，测试语料中的每个待消解项 NP_j 与其前面的所有待消解项 NP_i 构成测试例集，由分类器给出每个测试例的置信度值(测试例共指的可能性)，该值是已经过平滑的比例值 $\frac{p+1}{t+2}$(p 代表决策树上给出决策信息的叶子结点中正例的数目，m 代表该结点中训练例的总数目)。当此值大于 0.5 时，说明叶子结点中正例的数目多于反例的数目，这时认为该测试例共指，否则就认为不共指。因此，我们把置信度值减去 0.5，大于 0 的就当作图中的正边，否则为负边。

本文在 C4.5 决策树模型中所使用的剪枝方法是规则后减枝技术，即决策树中从根结点到叶结点的每条路径抽取形成规则。对于每一条规则，如果删除了其上的某一个结点不会导致精度的降低，就保留此次删除；相反，就取消此次删除。依次修剪所有的规则，对于修剪后的规则根据精度对它们进行排序，并按照这样的顺序应用这些规则来分类测试实例。

3.3　特征表示

为了能使上述的学习算法得到尽可能全面且正确的信息，以致使其能给出尽可能正确的共指决策，我们引入下列 9 个特征值来描述中文文本中的每一对名词短语对。

(1) 短语特征。直接取当前待消解项名词短语本身作为该属性值。

(2) 中心词特征。一般取当前待消解项名词短语中的最后一个词作为中心词。

(3) 同位语特征。如果当前待消解项属于专有名词特征，与之相邻的前一个语言单位也是待消解项而且不属于专有名词特征，那么就是同位语。

(4) 代词特征。用于指定当前待消解项是否为代词。

(5) 专有名词特征。用于指定当前待消解项是否为命名实体类型，主要包括人名、地名和机构名三种。

(6) 性别特征。主要分为男性、女性和未知三种。性别属性的确定主要是根据代词信息(如“他”、“她”等)、明显的性别特征词信息(如“先生”、“太太”，“小姐”等)以及配偶的信息来进行识别的，对于其他无明显性别特征的普通名词，则将其识别为未知类别。

(7) 单复数特征。主要分为单数、复数和未知三种。这一属性同样根据明显的单复数搭配词语进行识别，例如含有“们、一大群、许多、每个”等等，表示并列关系的名词短语归入复数类别。另外，人名默认为第一人称单数，地名和机构名默认为复数，职务名且没有复数特征的名词也识别为单数。对于无法直接确定单、复数的名词短语，将其归为未知类。

(8) 语义类别特征。依据知网中所定义的词汇语义层次，我们将名词短语的语义类别分为粗粒度的五种类型：时间(TIME)、城市(CITY)、动物(ANIMAL)、人(HUMAN)和对象(OBJECT)。对于每一个名词短语的中心词，我们通过查询知网获取相应的语义类别属性。对于多义词，目前我们还没有引入词义消歧算法，因而对每个词目前只是简单地选择其在知网中的第一个词义作为其语义类别。

(9) 指示语特征。跟在“这”、“那”、“这些”和“那些”之后的名词或名词短语被称为指示语。

3.4 关联聚类的近似计算

关联聚类问题实际上是一个整数规划问题。由于它是NP-hard问题,因而需要采用近似求解方法。

3.4.1 线性规划(LP)

我们给每一对顶点赋上一个0-1变量,因此 $x_{uv}=x_{vu}$。当$(u,v)\in E$,因为 $e=(u,v)$,为方便起见,可以把 x_{uv} 写成 x_e。对于一个给定的聚类,如果 u 和 v 在同一个聚类中,就让 $x_{uv}=0$,否则就让 $x_{uv}=1$;计算不一致性权值 $w(s)$的公式如下:

$$W(S)=\sum_{e\in E^{\langle -\rangle}} c_e(1-x_e)+\sum_{e\in E^{\langle +\rangle}} c_e x_e$$

其中当边 e 在聚类中,$1-x_e$ 就是1;当 e 在聚类间,则 $1-x_e$ 就是0。我们的目标就是要能找到一组对 x_{uv} 有效的赋值,使得 $W(S)$的值最小。如果 $x_{uv}\in\{0,1\}$且所有 x_{uv} 之间满足三角不等式,那么 x_{uv} 的这一组赋值是有效的(对于某一种聚类)。我们首先将上述的整数规划松弛为如下的线性规划:

$$\min \sum_{e\in E^{\langle -\rangle}} c_e(1-x_e)+\sum_{e\in E^{\langle +\rangle}} c_e x_e$$

s.t. $x_{uv}\in[0,1]$

$$x_{uv}+x_{vw}\geqslant x_{uw}$$

$$x_{uv}=x_{vu}$$

3.4.2 区域增长技术(Region growing)

在得到 x_{uv} 的值之后,接下来用区域增长技术对上述线性规划产生的分数解进行取整,得到时间复杂度为 $O(\log n)$的可以近似求解关联聚类问题的算法[6]。

我们假设有一个球,该球的半径按照某些固定的数值(根据 x_{uv} 的值计算得来的)不断地增长,每次增长形成的新球都能够包含图中的某些点,直到最后得到的球能包含图中所有的点才停止增长半径,这组同心球(只是半径不同)就构成了最终近似解的所有聚类。固定的半径保证了聚类内部不一致性的近似率,同时区域增长技术本身保证了聚类间不一致性的近似率。

我们先给出定义取整算法所需要的一些符号。一个球 $B(u,r)$就是以顶点 u 为球心、r 为半径、所有满足 $x_{uv}\leqslant r$ 的顶点 v 的集合。一个顶点集合 S 的切,用 $cut(S)$来表示,就是有且仅有一个端点在 S 内的"正边"的权值的和,其计算公式表示如下:

$$cut(S)=\sum_{|\{v,w\}\cap S|=1,(v,w)\in E^{\langle +\rangle}} c_{vw}$$

一个球的切就是由这个球所包含的顶点集合的切。

一个顶点集合 S 的卷,用 $vol(S)$来表示,就是两个顶点都在 S 内的"正边"的加权距离的和,其计算公式表示如下:

$$vol(S)=\sum_{\{V,W\}\subset s,(v,w)\in E^{\langle +\rangle}} c_{vw}x_{vw}$$

最后,一个球的卷就是该球内"正边"的加权距离(小数值)的和。换句话说,如果某个球有一条正切边,即$(v,w)\in E^{\langle +\rangle}$,$v\in B(u,r)$且 $w\notin B(u,r)$,那么边(v,w)可以给球的卷贡献 $c_{vw}\cdot(r-x_{uv})$的权值。我们还包括给每个球的卷赋上一个初始卷值 I(如球 $B(u,0)$有卷值 I)。取整算法如下所示:

(1)在图 G 中任意取一个点 u;

(2)把半径 r 初始化为0;

(3)按照 $\min\{(x_{uv}-r)>0:v\notin B(u,r)\}$增加半径 r,以至于让一条未被包含的正边进入球 $B(u,r)$内。如此反复,直到 $cut(B(u,r))\leqslant c\ln(n+1)\times vol(B(u,r))$;

(4)输出 $B(u,r)$内包含的顶点集作为聚类 S 中的某一个类;

(5)从图 G 中删除球 $B(u,r)$内的所有顶点和边;

(6)重复步骤1~5,直至图 G 为空。

4 实验结果

我们用类似文[2,3]中提出的共指消解算法所实现的系统作为本文研究的 Baseline 系统。文[2]中的系统用来协调分类矛盾的聚类机制是最近优先策略(link-first),最近优先策略选择离指示语最临近的置信度大于0.5的候选先行语作为最终先行语。文[3]中又提出了对上述系统的改进,它用的聚类机制就是最佳优先策略(link-best),最佳优先策略则是从置信度大于0.5的候选先行语中选择置信度最大的作为最终先行语。

在本实验中,使用了查准率(precision)、查全率(recall)和综合 F 值(F-measure)来评测实验结果。定义如下:

查准率 P = 正确识别出的共指对数/ 识别出的共指对数×100%.

查全率 R = 正确识别出的共指对数/ 实际共指对数×100%

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

我们选择了在 ACE 中文语料上进行训练和测试,其中由229篇文本组成了训练语料,由57篇文本组成了测试语料。本实验在所有的待消解项已被正确识别的假设得到满足的基础上进行的,用预处理器抽取这些待消解项所具有的词性、实体类型以及子类型等特征信息,构成特征向量。

我们设计了一个实验系统,输入的是带有切分标注以及共指信息的 ACE 语料,经过系统的预处理器得到特征向量。然后根据训练语料中的共指信息,将这些特征向量进行比对,形成表征共指关系和非共指关系的向量集合,用训练语料中的向量集合训练出一个 C4.5 的分类器,由经过规则后剪枝的分类器给出测试语料中的向量集合的共指置信度值,最后使用关联聚类的方法确定共指关系。整个处理完全自动进行,没有人工干预。

表1中第一行就是 Baseline 系统使用 link-first 聚类机制所得到的评测结果,第二行则是由 Baseline 系统使用 link-best 聚类机制的评测结果,第三行也就是我们的使用关联聚类近似算法的评测结果。

表1 各种方法的测试结果比较

Experiments	Precision(%)	Recall(%)	F(%)
link-first	75.42	73.96	74.68
link-best	75.72	74.26	74.98
Our Approach	76.65	76.25	76.45

不管是 link-first 还是 link-best 都只是针对一对名词短语进行共指决策,完全没有考虑到它们与其他名词短语之间的相关性,甚至与各个聚类之间的联系。其实这些名词短语之间、名词短语与聚类之间都不是相互独立的,所以这两种决策是相对片面的。而基于图划分的聚类算法就不是孤立地对每一对名词短语分别进行共指决策,而是充分考虑了多个名词短语之间以及名词短语与形成的聚类之间的相关性,从全局的角度给予共指决策。

结束语 本文提出了用图的划分来聚类待消解项的思想,该思想充分考虑了各个待消解项之间以及待消解项与它

(下转第256页)

对水印图像进行局部的修改，修改效果如图 6(a)，对提取的水印和原始水印每四位组成一个数。图 6(b)是提取的水印和原水印的效果比较图，此时 $\lambda(W,W_t)=0.9648$。通过比较，容易确定修改的图像块。同样，图 6(c)是对含有水印的图像添加椒盐噪声后原水印和提取水印的效果图，此时有 $\lambda(W,W_t)=0.9250$。

4.4 水印同步攻击实验

假定水印嵌入算法和提取算法被窃取，但是密钥没有公开，这时图像可能受到其它一些水印嵌入的攻击。这种情形下，我们随机选择了 1000 组合原始水印同样长度的水印进行相关度测试(其中有一个和原始水印相同)，得到的结果如图 7。

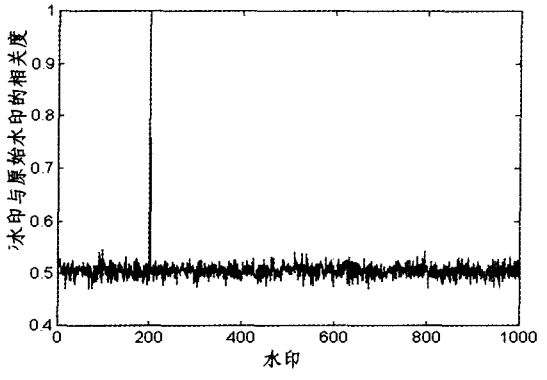

图 7 水印同步实验结果

从上面的图形可以看出，即使受到假冒水印的攻击，算法和相关度也能够对水印进行识别，只要适当选取阈值即可。本文我们选择 $T=0.85$，即可较好地区别对水印图像的各种篡改和完成对图像的认证。

结束语 本文提出了一种基于混沌系统的图像分块算法，该算法将图像分成两个大小相等的部分，其中一部分用来提取基于图像特征的水印，另外一部分则用来嵌入提取的特征水印。水印嵌入在小波变换的逼近子带中。水印提取算法不需要原始图像，而且通过比较原始水印和水印图像中的特征水印与提取的水印的两个相似度，能够区分图像受到攻击的种类。算法对恶意的篡改攻击能够较好地进行定位。实验结果表明，该算法在保持对常见的小波压缩稳健的同时，能有效地区分偶然失真与恶意篡改。

参 考 文 献

1 Chao H M, Hsu C M, Miaou S G. A Data Hiding Technique with Authentication, Integration and Confidentiality for Electronic Patient Records. IEEE Trans Inf Technol Biomed, 2003, 6 (1): 46～53

2 Kundur D, Hatzinakos D. Digital watermarking for telltale tamper-proofing and authentication. Proc. IEEE, 1999, 87(7): 1167～1180

3 Lin E T, Podilchuk C I, Delp E J, Detection of image alterations using semi-fragile watermarks. Proc. SPIE, 2000(3971): 152～163

4 张静，张春田. 用于 JPEG2000 图像认证的半脆弱性数字水印算法. 电子学报，2004，32(1)：157～160

5 李春，黄继武. 一种抗 JPEG 压缩的半脆弱图像水印算法. 软件学报，2006，17(2)：315～32

6 陈生潭，侯振华，王虹现. 双重认证的变换域图像半脆弱数字水印算法. 计算机辅助设计与图形学学报，2005，17(5)：1114～1119

7 王兴元，石其江. 基于图像特征和超混沌迭代的图像认证算法. 计算机研究与发展，2005，42(11)：1896～1902

8 Zhao D W, Chen G R, Liu W B. A chaos-based robust wavelet-domain watermarking algorithm. Chaos, Solitons and Fractals, 2004, 22: 47～54

9 Hassan M H, Gilani S A M. A Semi-fragile signature based scheme for ownership identification and color image authentication. Transactions on Engineering. Computer and Technology, 2006, 13: 308～311

10 Celik M U, Sharma G, Saber E, et al. Hierarchical Watermarking for Secure Image Authentication with Localization. IEEE Trans Image Process, 2002, 11 (6): 585～595

11 黄达人，刘九芬，黄继武. 小波变换域图像水印嵌入对策和算法. 软件学报，2002，13(7)：1290～1296

12 Hu J Q, Huang J W, Huang D R, et al. Image fragile watermarking based on fusion of multi-resolution tamper detection. Electronics Letters, 2002, 38(24):1512～1513

(上接第 218 页)

们形成的聚类之间的相关性，而且无需事先给出无从知晓的聚类的数量以及聚类的距离阈值，因此易于实现，并取得了良好的实验效果。当然，本方法对于内存的需求量较大，待消解项过多的测试语料可能难以一次性评测，所以，我们下一步拟采用 LP Chunking 的方法来评测待消解项过多的测试语料。

参 考 文 献

1 王厚峰，梅铮. 鲁棒性的汉语人称代词消解. 软件学报，2005，16(5)：700～707

2 Soon W M, Ng H T, et al. A machine learning approach to coreference resolution of noun phrases. Computational Linguistics, 2001, 27 (4): 521～544

3 Ng V, Cardie C. Improving machine learning approaches to coreference resolution. In: Proc. of the ACL, Philadelphia, 2002

4 李国臣，罗云飞. 采用优先选择策略的中文人称代词的指代消解. 中文信息学报，2005，19 (4)：24～30

5 Luo X, Ittycheriah A, et al. A mention～synchronous coreference resolution algorithm based on the Bell tree. In: Proc. of the ACL, Barcelona, 2004

6 Demaine E D, Emanuel D, et al. Correlation Clustering in General Weighted Graphs. Theoretical Computer Science, 2006, 361 (2-3): 172～187

7 Quinlan R J. C4. 5: Programs for Machine Learning. San Francisco, CA: Morgan Kaufmann, 1993

8 Wang H F, Mei Z. An empirical study on pronoun resolution in Chinese. In: Gelbnkh A, ed. Proc. of the 5th CiCLing Conf., Heidelberg, 2004