文章编号：1000-5013(2016)02-0201-06 **doi**：10.11830/ISSN.1000-5013.2016.02.0201

采用改进最长公共子序列的人名消歧

林翠萍，吴扬扬

（华侨大学 计算机科学与技术学院，福建 厦门 361021）

摘要： 将名词、形容词、动名词和命名实体作为文本特征，考虑词序与词频，结合特征项的语义，提出一种基于改进最长公共子序列的文本聚类(LCSC)方法. 实验结果表明：相对于传统的余弦值聚类方法，LCSC 方法在人名消歧的 P-IP 指标上，F 平均值由 74.2%提高到了 84.9%；相对于最长公共子序列方法，总体性能也提高了 3.7%.

关键词： 人名消歧；文本相似度；最长公共子序列；层次聚类

中图分类号： TP 391 **文献标志码**： A

据统计，在 Google 或 Yahoo 上搜索人名的量达到了 30%[1]，作为互联网检索的一个子任务，人名搜索返回结果往往是相关重名人的网页. 目前人名消歧的主流方法是基于向量空间模型的聚类方法，该方法的研究主要集中在特征提取和表示. Bagga 等[2]用向量空间模型解决跨文档人名的共指消解问题. Mann 等[3]自动提取了出生地、出生年月、职务等的人物传记信息，构成丰富的特征空间. Pedersen 等[4]抓住文档中的共现词，以前词为行、后词为列的矩阵经过奇异值分解后得到表示文档的特征. Chen 等[5]把特征系统地划分为基于名词和基于命名实体的特征，用 SoftTFIDF 计算特征权重，最后进行层次聚类. Ikeda 等[6]在以人名实体、混合关键词和网络链接为特征的基础上，提出两阶段聚类方法. 在中文人名消歧方面，Yang 等[7]把特征分为命名实体特征和普通词特征，通过引入同义词词林和词语相似度来降低数据的稀疏性. 一方面，传统的向量空间模型[8]把特征词或短语组成一个集合；另一方面，特征空间的稀疏性将会限制文本相似度计算的精度. 针对上述问题，本文提出了一种改进最长公共子序列的聚类方法(longest common subsequence clustering，LCSC).

1 相关工作

1.1 知网词语相似度

知网是一个网状知识库，描述了概念与概念之间的关系[9]. 每一个词汇可以有多个概念，每一个概念都用一系列的义原来描述. 这些义原用树状结构组织起来，义原根据义原之间的属性关系分为多棵义原树，这些存在一定关系的义原树就形成了网状知识结构. 刘群等[10]提出了一种计算语义相似度的方法，该方法实际上是获取两个词汇的最大概念相似度. 特征项 w_1 有 m 个概念：$s_{1,1}$，$s_{1,2}$，…，$s_{1,m}$，特征项 w_2 有 n 个概念：$s_{2,1}$，$s_{2,2}$，…，$s_{2,n}$，则 w_1 和 w_2 的语义相似度为

$$\mathrm{sim}(w_1,w_2)=\max_{i=1,\cdots,m;j=1,\cdots,n}\mathrm{sim}(s_{1,i},s_{2,j}). \tag{1}$$

文献[10]对实词概念语义的表达式分成了 4 个部分：第一独立义原描述式，记为 $\mathrm{sim}_1(s_1,s_2)$；其他独立义原描述式，记为 $\mathrm{sim}_2(s_1,s_2)$；关系义原描述式，记为 $\mathrm{sim}_3(s_1,s_2)$，符号义原描述式，记为 $\mathrm{sim}_4(s_1,s_2)$. 因此，两个概念的语义表达式的整体相似度记为

$$\mathrm{sim}(s_1,s_2)=\sum_{i=1}^{4}\beta_i\prod_{j=1}^{i}\mathrm{sim}_j(s_1,s_2). \tag{2}$$

收稿日期： 2014-08-31

通信作者： 吴扬扬(1957-)，女，教授，博士，主要从事数据库技术和数据挖掘的研究. E-mail：wuyy@hqu.edu.cn.

基金项目： 福建省科技计划重大项目(2011H6016)；福建省科技计划重点项目(2011H0028)

式(2)中:β_i(1≤i≤4)是可调节的参数,且满足 $b_1+b_2+b_3+b_4=1$,$\beta_1\geqslant\beta_2\geqslant\beta_3\geqslant\beta_4$.

1.2 最长公共子序列算法描述

最长公共子序列(longest common subsequence,LCS)最初是 Wagner 等[11]在 1974 年提出来的,即一个数列 S,如果分别是两个或多个已知数列的子序列,且是所有符合此条件序列中最长的,则 S 称为已知序列的最长公共子序列.

Hirschberg[12]用动态规划有效地解决了此问题.假设有两个字符串 X,Y,其分别表示为 $X=\{a_0,a_1,\cdots,a_{m-1}\}$和 $Y=\{b_0,b_1,\cdots,b_{n-1}\}$. 用一个二维矩阵 $\boldsymbol{C}_{m\times n}$存储迭代过程中当前的最长公共子序列长度. 其中:$c[i][j]$记录 a_0 到 a_i 和 b_0 到 b_j 的最长公共子序列的长度,即原始问题的一个子问题的解.当 $i=0$ 或 $j=0$ 时,空序列是 a_i 和 b_j 的最长公共子序列,故 $c[i][j]=0$.其他情况下,结合语义相似度可建立递归关系为

$$c[i][j]=\begin{cases}0, & i=0,\text{ or } j=0,\\ c[i-1][j-1]+1, & i,j>0,\quad a_i=b_j,\\ \max\{c[i][j-1],c[i-1][j]\}, & i,j>0,\quad a_i\neq b_j.\end{cases}\tag{3}$$

2 LCSC 方法

从计算机角度看,人名消歧是将多个重名人的文档集合划分为若干个子集合,即给定包含同一人名 n 的文档集合 D,背景知识 K,求 D 的划分 $p=\{D_1,D_2,\cdots,D_m\}$,并使划分中一个子集合对应一个人物 ρ_i(1≤i≤m).人名消歧步骤,如图 1 所示.

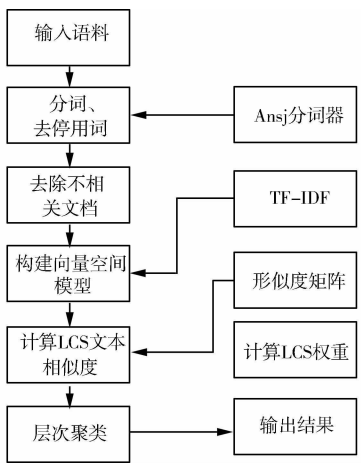

图 1 基于 LCSC 方法的人名消歧步骤

Fig. 1 Person name disambiguation steps based on LCSC

2.1 文本预处理及特征表示

假设一个人物实体对应一篇文档,先对每篇文档进行分词、词性标注、命名实体识别,并去除不相关文档.文中所采用的分词器是孙健开发的 Ansj(http://www.ansj.org/),以人民日报 1998 年 1 月的语料库为测试的结果准确率高达 98%,召回率为 96%,被广泛运用于自然语言处理中的命名实体识别、多级词性标注、关键词提取等.

特征提取的目的是降维,并得到有区分度的特征词.对于人名消歧而言,其作用可归纳为:找到能区分不同人物的重要词,即对相似度计算重要的词.在文本相似度计算上,以最长公共子序列为依据;在特征提取上,需要尽可能保留较全面的文本信息.因此,文中依次抽取文中出现的名词(n)、形容词(a)、动名词(vn)和命名实体(nr,ns,nt),按其在原文的顺序组成一个有序的词语序列表示文本,即 $d=\{w_1,w_2,\cdots,w_n\}$,其中,w_i 即为所抽取的特征项.

采用经典的 TFIDF(term frequency,inverse document frequency)方法来计算特征权重.其核心思想是:如果某个词或短语在一篇文章中出现的频率高,并且在其他文章中很少出现,则认为此词或短语具有很好的类别区分能力[13].加权函数为词频乘以反文档频率,即

$$W_{i,k}=\mathrm{TF}\times\mathrm{IDF}=f_{i,k}\times\log\frac{N}{n_k},\tag{4}$$

式(4)中:TF 即为词频,指特征 i 在文档 k 中出现的频率;IDF 为反文档频率;N 为所有类别中的文档的总数;n_k 为包含特征 i 的文档数.

利用词语分类的重要程度为后续的文本相似度服务,并非要利用 TFIDF 进行特征的选择.

2.2 改进最长公共子序列的文本相似度

为了充分利用文本自身的词序和词频信息,提出一种基于最长公共子序列的文本相似度计算方法.

2.2.1 词语相似度 为了弥补向量空间模型中特征项相互独立正交的缺陷,文中借助知网(Hownet)的词汇描述方式,根据文献[10]的语义相似度,建立特征项相似度矩阵.二维矩阵 $\boldsymbol{S}_{L(A),L(B)}$ 用于存储特征项之间的相似度.特征项 a_i 和 b_j 若是完全匹配,它们的词语相似度 $\mathrm{sim}[i][j]$将被置为 1.0;否则,根据知网词语相似度计算方法返回相应的值.

2.2.2　结合语义知识的 LCS 算法　文本去除停用词以后，抽取其中的名词、形容词、动名词和命名实体组成一个有序的特征词语序列. 当结合语序考虑文档相似度时，与经典的最长公共子序列的问题极为类似. 设文档 A 和文档 B 的特征项序列分别表示为 $a_1,a_2,\cdots,a_m$ 和 $b_1,b_2,\cdots,b_n$. 用一个二维矩阵 $\boldsymbol{C}_{(m+1)(n+1)}$ 存储当前的最长公共特征子序列长度. 其中：$c[i][j]$ 记录 a_1 到 a_i 和 b_1 到 b_j 的最长公共特征子序列的长度.

考虑表达的多样性，将原始的 LCS 算法结合词语之间的语义信息，即添加特征项相似度，提出一种结合语义的 LCS 算法，即

$$c[i][j]=\begin{cases}0, & i=0,\text{或 } j=0,\\ c[i-1][j-1]+1, & i,j>0,\ \mathrm{sim}[i-1][j-1]>\varepsilon,\\ \max\{c[i][j-1],c[i-1][j]\}, & \mathrm{sim}[i-1][j-1]<\varepsilon.\end{cases}\tag{5}$$

式(5)中：$\mathrm{sim}[i-1][j-1]$ 是 A 文档特征序列中第 i 个特征项和 B 文档序列中第 j 个特征项的相似度，若两个特征项的相似度超过了一个阈值 ε，认为它们是匹配的，最长公共特征子序列长度动态增加一个单位.

2.2.3　基于 LCS 的文本相似度　自 Hirschberg[14] 提出基于 LCS 的文本相似度方法之后，不少研究人员在此基础上提出改进并优化. 比较常见到的计算方法有文献[15]提到的 LCS 与较长的文本长度的比值，即

$$\mathrm{sim}(A,B)=\frac{\mathrm{LCSL}(A,B)}{\max\{L(A),L(B)\}}.\tag{6}$$

2 倍的 LCS[16] 除以两文本的长度之和，即

$$\mathrm{sim}(A,B)=\frac{2\times\mathrm{LCSL}(A,B)}{L(A)+L(B)}.\tag{7}$$

式(7)中：$\mathrm{LCSL}(A,B)$ 为文档和文档的最长公共特征子序列长度；$L(A)$，$L(B)$ 分别为文档 A 和文档 B 的特征向量的长度.

对于两篇描述同一人物的长文本和短文本来说，如果采用文献[15]的方法计算文本相似度，将会得到较小的值. 通过加入特征项的权重，提高文本相似度的精度. 改进的文本相似度为

$$\mathrm{sim}(A,B)=\frac{2\times(\mathrm{LCSL}(A,B)+\sum_{k=1}^{\mathrm{LCSL}}T_k)}{L(A)+L(B)}.\tag{8}$$

$$T_k=\begin{cases}1, & w_{i,k}>\delta,\quad w_{j,k}>\delta,\\ 0, & \text{其他.}\end{cases}\tag{9}$$

式(9)中：$w_{i,k}$，$w_{j,k}$ 分别是两篇文档的最长公共特征子序列中对应特征项的权重，包括了两个特征序列中的完全匹配特征项和不完全匹配特征项；δ 是一个平衡参数. 由于文档特征向量普遍较长，而最长公共特征子序列的长度则较小，所以加入 T_k 进行适当调节. 只有当两个对应特征的权重都超过 δ 时，T_k 才增加 1 个单位. 因此，该方法不仅考虑到了词序与词频，而且在 LCS 算法中结合了特征项之间的语义相关度，最终达到提高具有相同含义但使用不同词汇的文本相似度的目的.

2.3　聚类算法

对于人名消歧，由于重名者个数的不确定性，采用层次聚类算法比较合适. 文中采用自底向上的单链的层次聚类算法.

3　评价指标

实验选用 CIPS-SIGHAN 提供的两种人名消歧评价方法：P-IP 和 B-Cubed 指标. 两种方法分别计算了聚类结果的正确率，召回率和 F 值. P-IP 指标为

$$\mathrm{Pur}=\frac{\sum_{s_i\in s}\max_{R_j\in R}|S_i\cap R_j|}{\sum_{s_i\in s}|S_i|},\tag{10}$$

$$\mathrm{InvP}=\frac{\sum_{R_i\in R}\max_{s_j\in s}|R_i\cap S_j|}{\sum_{R_i\in R}|R_i|},\tag{11}$$

$$F\text{-measure}=\frac{2\times\mathrm{Pur}\times\mathrm{InvP}}{\mathrm{Pur}+\mathrm{InvP}}.\tag{12}$$

B-cubed 指标的公式为

$$\mathrm{Pre}=\frac{\sum_{s_i\in s}\max_{R_j\in R;d\in R_j}\frac{|S_i\cap R_j|}{|S_i|}}{\sum_{s_i\in s}|S_i|},\tag{13}$$

$$\mathrm{Rec}=\frac{\sum_{R_i\in R}\max_{s_i\in s;d\in s_j}\frac{|R_i\cap S_j|}{|R_i|}}{\sum_{R_i\in R}|R_i|},\tag{14}$$

$$F\text{-measure}=\frac{2\times\mathrm{Pre}\times\mathrm{Rec}}{\mathrm{Pre}+\mathrm{Rec}}.\tag{15}$$

式(13)～(15)中:$S=\{S_1,S_2,\cdots\}$是系统输出的聚类结果;$R=\{R_1,R_2,\cdots\}$是人工标注的聚类结果.

通常情况下,为了验证人名消歧系统的整体性能,取各个人名消歧效果的平均表现,即

$$\mathrm{Presion}=\frac{1}{n}\sum_{i=1}^{n}\mathrm{Presion}_i,\tag{16}$$

$$\mathrm{Recall}=\frac{1}{n}\sum_{i=1}^{n}\mathrm{Recall}_i,\tag{17}$$

$$F\text{-measure}=\frac{1}{n}\sum_{i=1}^{n}F\text{-measure}_i.\tag{18}$$

4 实验结果与分析

为了检验文中提出方法的有效性,进行了对比实验,将提出的 LCSC 方法与 Baseline、LCS 及文献[17]中的 AE 方法进行对比. 其中:Baseline 是传统的基于向量空间模型的聚类方法,以全文除停用词外的所有词为文本特征,采用 TFIDF 为特征权重计算公式,以特征向量的夹角余弦值作为文本相似度,再进行单链层次聚类. LCS 方法中 LCS 和文本相似度分别采用式(3),(7)计算. 文献[17]中的 AE 方法通过抽取人物属性信息作为特征来进行人名消歧.

4.1 数据集

采用的数据集是搜狗全网新闻人名消歧语料[17],该语料选取了国内最常用的 50 个人名,抽取含有这 50 个人名串的新闻报道. 对其中新闻报道最多的 12 个人名的总共 11 876 篇文档进行了人工标注.

4.2 实验结果分析

通过对搜狗全网人名消歧语料中 12 个人名进行实验,结果表明:提出的基于改进的 LCS 的文本相似度的聚类算法在两个评测指标上都表现出了良好的效果. P-IP 的 F 值对比,如图 2 所示. B-Cubed 的 F 值对比,如图 3 所示.

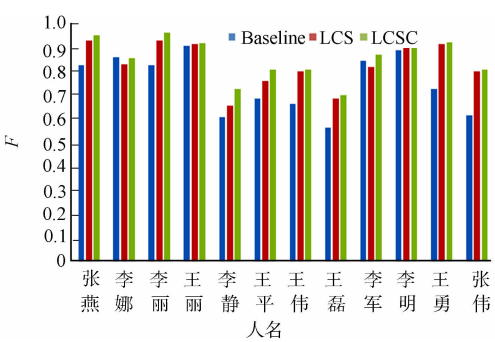

图 2 P-IP 的 F 值对比

Fig. 2 Comparison of P-IP F-measure

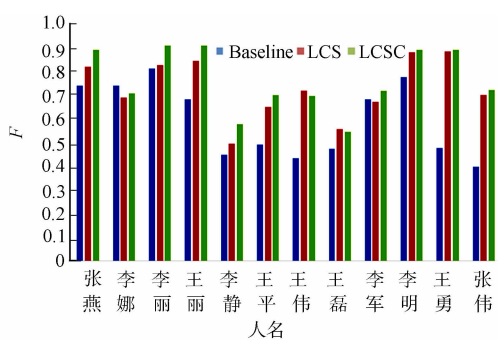

图 3 B-Cubed 的 F 值对比

Fig. 3 Comparison of B-Cubed F-measure

图 2,3 结果表明:除“李娜”以外的其他人名,LCS 方法和文中提出的 LCSC 方法的 F 值都比 Baseline 有所提高.

在进行 LCSC 方法的实验过程中需要调节 3 个参数,分别是聚类停止阈值 l,式(5)的词语相似度阈值 ε 及式(9)计算文本相似度时的权重平衡参数 δ.结合多个重名人的实验结果,l 在 0.25 左右取得总体较高的 F 值.在 l 保持一定的情况下,分别对 ε 和 δ 进行控制变量法获取最优值.

“张伟”的聚类阈值 4 在调整过程中对结果的影响,如图 4 所示.由图 4 可知:ε 和 δ 的最优值分别为 0.9 和 0.01.12 个人的平均 F 值的对比,如图 5 所示.由图 5 可知:P-IP 的 F 值从 Baseline 的 74.2%提高到 84.9%;B-Cubed 的 F 值从 55.0%提高到 75.7%.与 LCS 相比,LCSC 方法也分别高出 3.7%和 3.5%;与 AE 方法相比,文中方法在 P-IP 指标上体现出一定的优势.可见,加入了特征项的权重信息对文本相似度计算起到了一定的作用,使得 LCSC 方法体现出了较好的性能,与人物属性抽取方法比较也略胜一筹.

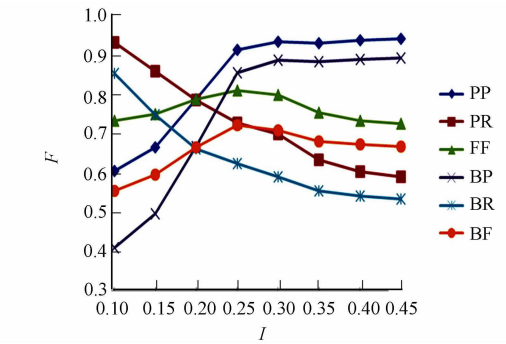

图 4　“张伟”的聚类阈值对结果的影响

Fig. 4　Effect of Zhang wei′s clusterin threshold to result

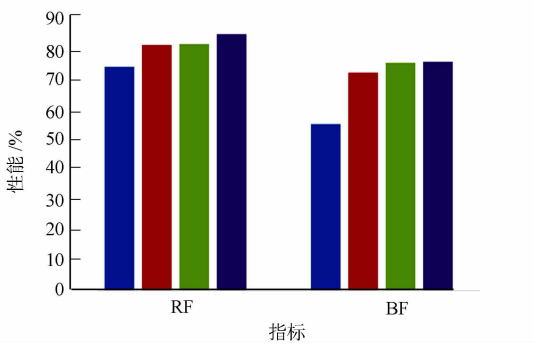

图 5　12 个人的平均 F 值的对比

Fig. 5　Comparison of average F-measure for 12 persons

5　结束语

通过引入知网的词语相似度计算,弥补了向量空间中特征项之间相互独立的缺陷.在最长公共子序列的计算中加入了权重的平衡参数,避免了传统余弦相似度导致的特征稀疏性,从而提高了文本相似度计算的准确率.针对 LCSC 方法的不足,后续工作将从预处理和语义分析两方面入手.此外,提取与 LCSC 方法结合的文本特征,也是需要进一步深入的问题.

参考文献:

[1] ARTILES J,GONZALO J,VERDEJO F. A testbed for people searching strategies in the WWW[C]// Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Piscataway:ACM,2005:569-570.

[2] BAGGA A,BALDWIN B. Entity-based cross-document coreferencing using the vector space model[C]// Proceedings of the 17th International Conference on Computational Linguistics. Boston:Association for Computational Linguistics,1998:79-85.

[3] MANN G S,YAROWSKY D. Unsupervised personal name disambiguation[C]// Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL. Edmonton:Association for Computational Linguistics,2003:33-40.

[4] PEDERSEN T,PURANDARE A,KULKARNI A. Name discrimination by clustering similar contexts[C]// Computational Linguistics and Intelligent Text Processing. Berlin:Springer Berlin Heidelberg,2005:226-237.

[5] CHEN Y,MARTUB J. Towards robust unsupervised personal name disambiguation[C]// EMNLP-CoNLL. Washington D C:IEEE Press,2007:190-198.

[6] IKEDA M,ONO S,SATO I,et al. Person name disambiguation on the web by two-stage clustering[C]// 2nd Web People Search Evaluation Workshop. New York:Association for Computing Machinery,2009:33-38.

[7] YANG Xia, JIN Peng, XIANG Wei. Exploring word similarity to improve Chinese personal name disambiguation

[C]// Web Intelligence and Intelligent Agent Technology. Washington D C:IEEE Press,2011:197-200.
[8] SALTON G,WONG A,YANG C S. A vector space model for automatic indexing[J]. Communications of the ACM,1975,18(11):613-620.
[9] 董振东,董强. 知网简介[EB/OL][2014-03-16]. http://www.keenage.com.
[10] 刘群,李素建. 基于《知网》的词汇语义相似度计算[J]. 中文计算语言学,2002,7(2):59-76.
[11] WAGNER R A,FISCHER M J. The string-to-string correction problem[J]. Journal of the ACM (JACM),1974,21(1):168-173.
[12] HIRSCHBERG D S. A linear space algorithm for computing maximal common subsequences[J]. Communications of the ACM,1975,18(6):341-343.
[13] 施聪莺,徐朝军,杨晓江. TFIDF 算法研究综述[J]. 计算机应用,2009,29(B6):167-170.
[14] HIRSCHDERG D S. Algorithms for the longest common subsequence problem[J]. Journal of the ACMWeb Intelligence and Intelligent Agent Technology. Washington D C:IEEE Press,1977,24(4):664-675.
[15] 全方磊. 数据特征提取在高铁车地传输中的应用研究[D]. 杭州:浙江大学,2013:39-40.
[16] 牛永洁,张成. 多种字符串相似度算法的比较研究[J]. 计算机与数字工程,2012,40(3):14-17.
[17] 张鑫. 人名消歧关键技术研究与实现[D]. 哈尔滨:哈尔滨工业大学,2012:32-33.

Person Name Disambiguation Based on Revised Longest Common Subsequence

LIN Cuiping, WU Yangyang

(College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China)

Abstract: This paper uses nouns, adjectives, gerunds and named entities as text features, and also considers the word order and word frequency when computing the text similarity. A text clustering method based on revised longest common subsequence (LCSC) is proposed. The experimental results show that the LCSC method can significantly improve the overall performance in person name disambiguation compared with traditional clustering method and make the average *F*-measure increase from 74.2% to 84.9%. The overall performance also improved by 3.7% when compared with the longest common subsequence method.

Keywords: person name disambiguation; text similarity; longest common subsequence; hierarc

(**责任编辑**:陈志贤 **英文审校**:吴逢铁)