基于反比概率模型和规则的中文姓名自动辨识系统*

季姮 罗振声

计算语言学研究室

清华大学人文学院，北京 100084

E-mail: jiheng00@mails.tsinghua.edu.cn

摘 要：中文姓名的辨识是自动分词、自动文摘的基础。我们提出了基于语料库统计的反比姓名概率模型，并结合上下文模式、邻接链、特殊姓、位置依存信息等四个辨识模块，设计了一个中文姓名辨识系统。本文描述了本系统的算法，，测试结果表明系统有较高的召回率和精确率，召回率达到93.75%，精确率达到83.95%。

关键词：反比姓名概率模型，数据稀疏，上下文模式，特殊姓，邻接链

Inverse Name Frequency Model and Rules Based Chinese Name Identifying

Ji Heng, Luo Zhensheng

Computational Linguistics Lab

School of Humanities and Societies Science, Tsinghua University, Beijing 100084

Email: jiheng00@mails.tsinghua.edu.cn

Abstract: The processing of Chinese names is important to the approach of Chinese word segmentation and automatic abstraction. In this paper we put forward an inverse name frequency model. Based on this model, context pattern, adjacent chain, special name table and position dependent information, we designed an effective system for automatically identifying Chinese names in texts. This paper describes the algorithm of this system, and the experiment result shows its upper recall and precision rate. Its recall rate reaches 93.75% and precision rate reaches 83.95%.

Keywords: inverse name frequency model, data sparsity, context pattern, special surname, adjacent chain

1 引 言

自动分词是自动文摘的基础，目前存在很多有效的自动分词方法，但当待处理文本中存在大量未登录词，如人名、地名、机构名时，分词效果便很难达到令人满意的程度。据统计，中文姓名在待处理文本中一般只占1%—2%，但姓名的切分错误高达50%以上。对所有分词错误进行统计，姓名错误占了将近90%。所以中文姓名确辨识是提高自动文摘精确性重要组成部分。

中文姓名识别的主要难点在于，中文姓名用字具有很大的任意性，中文姓名不像印欧语言那样可以通过大写字母来辨识，也不像英文译名那样用字局限在较小的字库内，而且姓名用字还可以与上下文成词。见诸报道的中文姓名辨识系统中运用的主要方法有基于语料库【文献3，5】、统计方法【文献2、6】、规则【文献2、4、5、6】等，尤其是【2】和【6】在姓名识别的召回率和精确率方面都达到了较高水平。但这些系统运用的方法也存在一些不足和可以改进的地方：(1)自动分词歧义引起姓名漏报；（2）未能充分挖掘真实语料中姓名用字的统计信息；(3)规则库偏小，误报率较高。

本论文尝试着挖掘语料库中姓名的统计信息，提出了新的姓名概率统计方法——反比姓名概率模型（INF），并结合多种规则，采用四个规则辨识模块融合的方法筛选姓名，系统处理的文本不需要经过自动分词，可以避免自动分词歧义带来的缺陷，取得了较高的召回率和精确率。

*国家自然科学基金，批准号【69972025】

2 算法

2.1 反比姓名概率（INF）模型

2.1.1 INF 定义

为了考察任一字串在语料中出现时充当姓名的可能性，我们设计了基于真实语料的反比姓名概率（Inverse Name Frequency，以下简称 INF）模型。

语料库{TOTAL}分为姓名字串库 NAME 和非姓名字串库 NONNAME，以二字字串为例：

设任一二字字串 S=<S1S2>，令：NS：表示 S 是姓名；P(S)：表示 S 在语料中出现的概率；

P(N)：表示姓名在 TOTAL 的概率；P(NS|N)：表示 NAME 中字串 S 的概率

P(NS|S)：表示 S 在语料中作为姓名出现的概率；则：

$$P(NS \mid S) = \frac{P(NS)}{P(S)} = \frac{P(NS \mid N) \times P(N)}{P(S)}$$ ——公式 **2.1**

由于在特定的待处理语料中 P(N)为常数。我们定义反比姓名概率 P(INFS)如下：

$$P(INFS) = \frac{P(NS \mid N)}{P(S)}$$ ——定义 **2.1**

则：$P(NS \mid S) = P(INFS) \times P(N)$ ——公式 **2.2**

因此对 P(NS|S)的考察就转化为对反比姓名概率 P(INFS)的考察。根据 P(INFS)我们可以设定合适的阈值来构造候选姓名表。

2.1.2 INF 计算

姓名用字概率表

本系统中，姓名用字概率 P(NS|N) 根据 1982 年全国人口普查资料【文献 1】，对 174，900 个中文姓名进行抽样综合统计，得到了单姓概率表（436 个）、复姓概率表（11 个）和名字用字概率表（1397 个），分别记为 SSF，DSF 和 FF。

总字频表

我们应用清华大学中文系计算语言研究室研制的 TH 大型语料库系统(微机版本)【文献 7】计算 P(STRi)，对 700 百万字语料进行了统计，得到含 20902 个汉字的字频表，记为 TCF。

由于复姓的数量较少且相对集中，本论文只描述单姓的 INF 计算。

当 S 为二字字串时，可以表示为 C_1C_2 (C_1 ,C_2 均为汉字)，定义如下参数：

$P(SURC_1|SUR)$：表示 C_1 在 SSF 中的概率；$P(FIRC_2|FIR)$：表示 C_2 在 DSF 中的概率；

$P(C_1)$，$P(C_2)$分别表示 C1，C2 在 TCF 中的概率；

则字的反比姓名概率可以定义如下：

$$P(INFC_1) = \frac{P(SURC_1 \mid SUR)}{P(C_1)}; \quad P(INFC_2) = \frac{P(FIRC_2 \mid FIR)}{P(C_2)}$$ ——定义 **2.2**

在本系统中我们假定字串中的字之间的概率相互独立，所以：

$$P(INFS) = \frac{P(NS \mid N)}{P(S)} = \frac{P(SURC_1 \mid SUR) \times P(FIRC_2 \mid FIR)}{P(C_1) \times P(C_2)} = P(INFC_1) \times P(INFC_2)$$ ——公式 **2.3**

同理，当 **STRi** 为单姓双名时，设可以表示为 $C_1C_2C_3$ (C_1,C_2,C_3 均为汉字)，则有：

$$P(INFS) = P(INFC_1) \times P(INFC_2) \times P(INFC_3)$$ ——公式 **2.4**

2.1.3 数据稀疏问题

语料库中出现的中文姓名有6479个，定义为ACF，显然$SSF, FF \subset ACF \subset TCF$由于姓名用字（SSF，FF）只占了总字库（TCF）很小的一部分（姓氏6.44%，名字20.66%），ACF中的字也只占TCF的30.99%。若将不在姓名用字库中的字的概率都视为0，则会出现严重的数据稀疏问题。

我们根据奈和埃森(H.Ney 和 U.Essen)在【文献 9】中提出的线性减值法来处理数据稀疏问题。基本思想是样本的概率为原来的值乘以(1－α)，其中α是正好出现一次的样本数目除以样本大小。不在样本中的对象则平分α。以姓氏为例，具体算法如下：

设$P(SURC_i|SUR)$, $P(FIRC_i|FIR)$和$P(C_i)$经数据稀疏处理后的值，分别记为$P'(SURCi|SUR)$, $P'(FIRCi|FIR)$ and $P'(C_i)$。令：$\alpha = n_1/N_{ACF}$，其中N_{ACF}是ACF的大小，n1是ACF中正好出现1次的字的数目，本系统中$\alpha = 203/86405823 = 2.35\times10^{-6}$　$\beta = m1/N_{SSF}$，其中N_{SSF}是姓氏样本数据的大小，m_1是在姓氏样本中正好出现1次的姓氏数目。本系统中$\beta = 81/25000 = 0.00324$。令$C_i$为SSF中的字，$C_j$是不在SSF中但在ACF中的字，$C_k$是不在SSF也不在ACF中的字。则：

$$\begin{cases} P'(SURC_i \mid SUR) = (1-\beta)\times P(SURC_i \mid SUR) \\ P'(SURC_j \mid SUR) = P'(SURC_k \mid SUR) = \dfrac{\beta}{N_{TCF} - N_{SSF}} \end{cases} \quad \begin{cases} P'(C_i) = P'(C_j) = (1-\alpha)\times P(C_{i/j}) \\ P'(C_k) = \dfrac{\alpha}{N_{TCF} - N_{ACF}} \end{cases}$$ 故：

$$\begin{cases} P(INFC_i) = \dfrac{(1-\beta)\times P(SURC_i \mid SUR)}{(1-\alpha)\times P(C_i)}, P(INFC_j) = \dfrac{\beta}{(1-\alpha)\times(N_{TCF} - N_{SSF})\times P(C_j)} \\ P(INFC_k) = \dfrac{\beta\times(N_{TCF} - N_{ACF})}{\alpha\times(N_{TCF} - N_{SSF})} \end{cases}$$ ——公式 **2.5**

名字用字反比概率同理可得。

2.1.4 阈值

根据公式2.3，2.4，2.5我们可以计算二字和三字字串的反比姓名概率如下。

姓名字串 **INF** 表和非姓名字串 **INF** 表

我们对清华大学近8年14457名学生的姓名的概率进行了统计，构造了包含二字姓名字串（含6553个字串）概率表和三字姓名字串（含7730个字串）概率表。

我们从TH语料库中抽取了166640个非姓名字串，构造了非姓名二字字串（含108169个字串）概率表和非姓名三字字串（含93281个字串）概率表。得到计算结果曲线如图2.1,2.2：

根据贝叶斯决策理论【文献 8】，我们可以根据曲线划定阈值，使大于该阈值的字串作为候选姓名。在本系统中阈值取lg后的值：二字字串阈值R_{newtwo}=10.013，三字字串阈值R_{newthr}=8.168

二字字串概率曲线——反比姓名概率(INF)模型

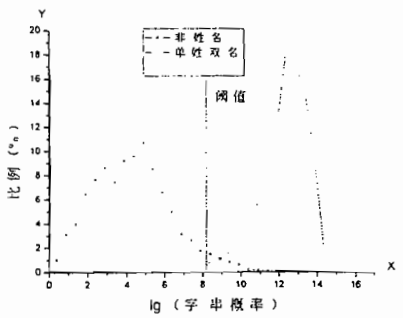


图2.1

三字字串概率曲线——反比姓名概率（INF）模型

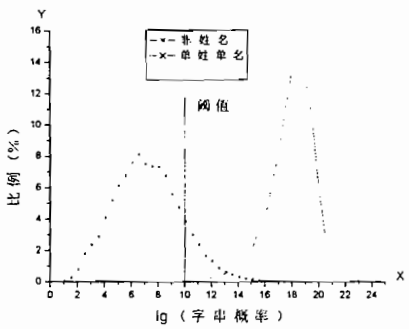


图2.2

2.1.5 INF 模型与传统模型比较

在传统的姓名辨识算法中，姓名用字概率表是构造候选姓名表的唯一依据。为了便于根传统的方法相比较，我们同样画出了运用传统方法的到的姓名字串和非姓名字串的相应概率曲线：

二字字串概率曲线—传统的姓名用字概率模型

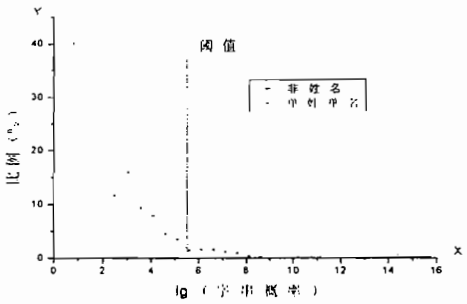


图 2.3

三字字串概率曲线—传统的姓名用字概率模型

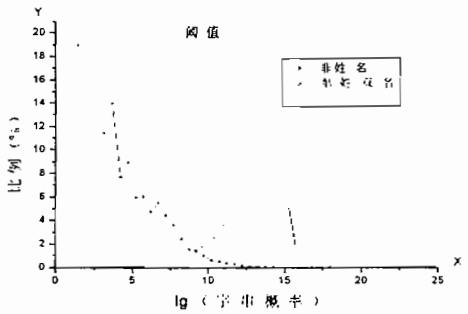


图 2.4

从不同的曲线（图 2.1 和 2.3，2.2 和 2.4）我们可以看出，和传统的姓名用字概率模型相比，反比姓名概率模型使辨识结果得到了一定程度的改善：

1. 两个模型得到的非姓名概率曲线趋势不同。由传统的模型得到的非姓名字串概率曲线分布近似指数分布，而 INF 模型得到的曲线近似正态分布，后者更为辨识算法所乐见。

2. 由于采用传统的模型得到的姓名概率较多集中于低频范围，反映在曲线上是产生了很多抖动点，对划定阈值不利。

3. INF 模型更能排除一些含高频字的非姓名字串，提高了精确率。试看下面的句子：

博士段恩和导师王松茂教授正在讨论问题。

采用传统的模型，“段恩和”充当姓名的概率为 12.8871> R_{oldthr} 辨识为候选姓名，“博士”、“导师”作为称谓词符合上下文模式，起定界作用，因此“段恩和”被误判为姓名。在 INF 模型中，由于“和”属高频字，因此“段恩和”反比概率仅为 8.1189<R_{newthr} 不被列为候选姓名。

4. INF 模型更能召回真实姓名。对于在姓名用字表（SSF，FF）中的频度较低的汉字，采用传统的模型时，包含这些字的姓名往往不被列为候选姓名。但由于他们在 TCF 中的总字频也较低，因此反比姓名概率较高，即这一类字串一旦在语料中出现，极有可能就是姓名，在下面的例句中：全国政协委员和各界人士邓季惺、平杰三、李默庵、沈醉、徐萌山、蔡端、鲜恒。

这句中的两个比较生僻的姓名“沈醉”和“鲜恒”，运用传统的模型，字串概率分别为 4.725、5.0136 均小于 R_{oldtwo}，而由 INF 模型得到的概率值分别为 9.787125、10.088952，均大于 R_{newtwo}，被正确识别。因此采用新模型 INF 后，真实语料中的部分偏僻真实姓名从阈值线的左边移到了右边，提高了系统的召回率。

2.2 规则筛选

对于运用上述反比姓名概率模型得到的候选姓名表，需要通过规则进行筛选。在本系统中，我们设计的规则有上下文模式，特殊姓，邻接链和位置依存信息等，具体算法如下：

2.2.1 上下文模式

姓名在文本中出现任意性很大，但也不是无规律可循。通过语料的分析，我们发现一些特殊词，如称谓词、指界动词、连词、时间词、成语及习惯用语等，常常紧接在姓名前后，一些特殊的结构短语也对辨识姓名有指示作用，由此我们归纳了确认姓名的五种上下文模式：

模式	上文	候选姓名		下文
1	称谓词 or 连词 or 时间词 or 左指界动词 or 非汉字	姓氏及双名末字为非常用字		称谓词、多字指界动词、时间词、成语、习惯用语
2			末字为非语气词	非汉字
3			双名	单字指界动词
4	以			“为”+称谓词
5	……字串＋“、”＋字串＋“、”至少一个字串为姓名			字串＋“、”＋字串＋“、”＋……至少一个字串为姓名

若候选姓名符合表中上下文模式之一，则可被确认为姓名。

2.2.2 特殊姓辨识

一些特殊的姓，在特定的上下文限制条件下不充当姓，对这类姓要视上下文情况单独处理。

（1）张、项、周、章、万、段

这类姓前面若是数词或阿拉伯数字，或“多”、“个”，则认为是作为量词出现；

（2）于：若前面是“关”、“由”、“用”等可以和“于”成词的字，“于”不被辨识为姓氏。

2.2.3 邻接链回溯检查

中文姓名辨识的传统作法是在筛选候选姓名之前先进行自动分词，不但影响了系统的执行效率，还由于分词的歧义降低了召回率。试看例句： 李白天天喝酒。

若用正向最大匹配法分词：李 白天 天 喝酒 。将导致姓名姓名“李白”漏报。

本系统处理的是不需要自动分词处理的原始文本，我们设计了规则：

设一候选姓名的上下文环境为：…B2 B1 B0　C0…Cn　F0 F1 F2…

其中 C0…Cn 是候选姓名，…B2B1B0 是上文，F0F1F2…是下文，若（**C0** 与上文成词 且 **B0** 不与上文成词）或（**Cn** 与下文成词 且 **F0** 不与下文成词），则否认 **C0…Cn**

运用这条规则，上例中“白天”和“天天”都是词，所以“李白”不会由于分词不当被漏报，也提高了执行效率。

2.2.4 位置依存信息筛选

在真实文本中，任一候选姓名 CN 包含了诸多信息：

{ start_pos: 在文中的起始位置； end_pos: 在文中的结束位置； com: 构成的具体字符

lg(p)：反比姓名概率对数　info=lg(p)/(end_pos－start_pos)：平均信息量 }

它们之间通过位置信息它们相互制约、相互影响，对筛选姓名起着重要的辅助作用。经过多次测试，我们归纳了以下几种位置信息模式及其具体操作。任意一对候选姓名对 CN_i 和 CN_j

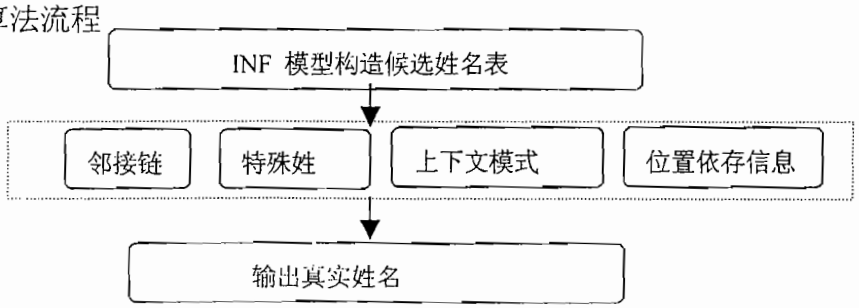
(1) 若(CNi→com＝CNj→com)且 (CNi→start_pos≠CNj→start_pos) 则确认 CNi<=>确认 CNj

(2) 若 (CNi→start_pos≤CNj→start_pos≤CNi→end_pos+1) 则：

a. 确认 Cni<=>否认 CNj　b. 若(CNi→info<CNj→info) 否认 CNj

c. 若(CNi 符合上下文模式) 否认 CNj　d. 若(CNi 为孤立字或低频词)否认 CNj

2.3 算法流程



上图说明了本系统的算法流程。首先运用 INF 模型，筛选出待处理文本中反比姓名概率超过阈值的字串构成候选姓名表。接着用四个规则辨识模块辨识候选姓名表中的真实姓名。由每个模块我们得到各候选姓名的中间权值，最后将这些权值相加，权值较高的候选姓名得到确认。

3 实验结果和今后的研究工作

评测一个中文姓名辨识系统的优劣主要有两个参数：

召回率＝文本中的中文姓名被辨识出的比例

精确率＝辨识为中文姓名者真正为中文姓名的比例

为了总体上评测本系统的辨识姓名效果，我们从 TH 语料库中选取了不同体裁的文章 104 篇，含 848 个中文姓名。系统辨识出“中文姓名” 947 个，其中 795 个为真正正确的姓名。则本系统的召回率和精确率分别为：召回率＝795/848=93.75%　　精确率＝795/947=83.95%

中文姓名辨识是难度很大的课题，我们认为以后还可以在以下几方面作研究和努力：

1. 共现矩阵

(1)姓名内部用字共现矩阵:我们在系统设计过程中发现，姓氏和名字、名字首字和名字末字出现的概率并非完全独立，而是有一定的邻接关系。如果能对大量语料标注基础上统计邻接关系，我们可以引入 n 元语法，对姓名字串的总概率重新加以计算。

(2)姓名和上下文共现矩阵:即在一定长度范围内其他字串和该候选姓名字串同时出现的概率，能排除很大一部分成词的候选姓名。例如在“对华永久正常贸易关系的解决”一句中，“华永久”由于上文模式被识别为姓名，但“对华”经常与“贸易”、“关系”等字串同时出现，如果能通过构造这样的共现矩阵来筛选，将能提高精确率。

2. 高频字串表

上例中，也可构造诸如“对华永久”这样的高频字串（不是词但是在真实语料中频繁出现的字串），出现在这样的字串里的候选姓名被排除。也是一种有效的辨识方法。

致谢

本 INF 模型设计和系统实现过程中得到了博士生万敏的热情指点和帮助，在此深表感谢！

参考文献

【1】中国社会科学院语言文字应用研究所汉字整理研究室，《姓氏人名用字分析统计》，语文出版社，1990

【2】 孙茂松， 黄昌宁， 高海燕， 方捷 “中文姓名的自动辨识”，《中文信息学报》Vol.9, No.2 1994

【3】张俊盛 ，陈舜德等，“多语料库作法之中文姓名辨识”，《中文信息学报》，第 6 卷，第 3 期，1992

【4】郑家恒，刘开瑛，“自动分词系统中姓氏人名处理策略探讨”，陈力为主编《计算语言研究与应用》，北京语言学院出版社，北京，1993

【5】宋柔，朱宏，潘维佳，尹振海，“基于语料库和规则库的人名识别法”，陈力为主编《计算语言研究与应用》，北京语言学院出版社，北京，1993

【6】王省，黄德根，杨元生“基于统计和规则相结合的中文姓名识别”，黄昌宁，董振东主编《计算语言学文集》，清华大学出版社，北京，1999

【7】韩兆兵，罗振声，“清华语料库的设计和实现”，清华大学本科毕业论文，1999

【8】边肇祺等，《模式识别》，清华大学出版社，1988

【9】H.Ney and U.Esser, Estimating ‘Small’ Probabilities by leaving-One-Out.In Eurospeech,pages2239~2242,1993