

对专业搜索引擎中未登录词的识别研究

张 赢, 万仲保

(华东交通大学 信息工程学院, 江西 南昌 330013)

摘 要:未登录词的识别是中文分词要面对的数个难题之一, 解决好了这个问题就能够有力提升中文分词的效率。对此, 文中简要介绍了专业搜索引擎的概况, 并具体分析了对电影名进行中文分词的特点和介绍了传统的未登录词的识别策略, 最后在此基础上提出了一种电影名未登录词的识别策略并简要分析了这种策略未来的优化方向。

关键词:电影名; 中文分词; 未登录词; 识别策略

中图分类号: TP393

文献标识码: A

文章编号: 1673-629X(2009)05-0134-03

Professional Search Engine Unknown Word of Recognition

ZHANG Ying, WAN Zhong-bao

(School of Information Engineering, East China of Jiaotong University, Nanchang 330013, China)

Abstract: Identification of unknown words is the Chinese word segmentation to face one of a number of problems, to resolve this problem will be able to effectively enhance the efficiency of the Chinese word segmentation. Gave a briefing on the profile of professional search engine, and a detailed analysis of the film were conducted at the Chinese word on the characteristics and traditions of the unknown word identification strategies, the last on the basis of this presents a film who did not login term strategy for the identification and brief analysis of this strategy of optimizing future direction.

Key words: film name; Chinese segmentation; unknown words; identification strategy

0 引 言

垂直搜索引擎, 即专业或专用搜索引擎, 是针对某一个行业或某一主题的专业搜索引擎, 是搜索引擎的细分和延伸, 是对网页库中的某类专门的信息进行一次整合, 定向分字段抽取需要的数据进行处理后再以某种形式返回给用户^[1]。

传统的综合搜索引擎^[2]一次搜索的结果有成千上万条, 而在这些庞大的信息中, 有用的信息只是其中的一小部分, 这就使用户需求和市场服务间产生巨大反差, 形成所谓的“搜索噪音”。而垂直搜索引擎则不同, 它是更有针对性的搜索引擎, 它只搜索特定主题的信息。由于覆盖的学科领域少, 信息量相对较少, 这就大大降低了收集信息的难度, 提高了信息的质量。搜索引擎的出现, 整合了互联网上众多的网页资源, 并提供信息导航和信息查询服务, 使信息的价值得到了网民和厂商的普遍认可。但是, 众多专业性网站、行业网站

独立于互联网的成功, 恰恰证明了互联网的格局应该是多方面的。通用搜索引擎的性质^[3], 决定了其不能满足特殊领域、特殊人群的精准化信息需求服务。市场需求多元化决定了搜索引擎的服务模式必将出现细分, 针对不同行业提供更加精确的行业服务模式。可以说通用搜索引擎的发展为垂直搜索引擎的出现提供了良好的市场空间, 势必将出现垂直搜索引擎在互联网中占据部分市场的趋势, 也是搜索引擎行业细分化的必然趋势。

垂直搜索引擎在行业门户中的应用价值很大。“内容门户”+“搜索门户”, 有效提高行业门户的竞争力。进入2006年以来, 垂直搜索发展迅速, 购物搜索、影视搜索、旅游搜索、政府搜索、大学搜索等花样翻新、层出不穷, 这些都给行业门户带来了一定的冲击。其中多媒体电影搜索发展最为迅猛, 如何在搜索“大行其道”的今天有效把握电影行业门户的竞争力, 在电影行业内容门户的基础上建立垂直搜索服务, 无疑是关键所在。垂直搜索引擎可以更好地整合电影行业相关资源、增强多媒体内容资源的相关性、挖掘电影行业资源的价值。同时, 垂直搜索还可以为用户提供更加个性化的信息服务, 并通过把握和分析用户的搜索行为, 改

收稿日期: 2008-08-26

基金项目: 江西省自然科学基金(0411035)

作者简介: 张 赢(1982-), 男, 湖南常德人, 硕士研究生, 研究方向为信息安全、搜索引擎; 万仲保, 副教授, 研究方向为信息安全、网络工程。

进信息服务质量,提高服务水平。

1 电影名中文分词的特点分析

1.1 电影名的基本特点

(1)直接用数字形式或是数字的组合作为电影名,比如王家卫的电影《2046》等;

(2)具有一定含义的单字作为电影名,比如徐克的电影《刀》等;

(3)常见的多字词作为电影名,比如人名、地名、特有名词、有特色的动词都可以作为电影名;

(4)中国常见的成语作为电影名,比如陈可辛的电影《金枝玉叶》。

1.2 电影名未登录词的简要分析

电影名除了一部分满足以上一些基本特点外,大部分都属于各大词典难以查询的未登录词,这些词都是根据影片剧情的需要同时能够反应影片的梗概而出现的,具有一定的偶然性和难以预测性。但是,通过分析大量的电影名,也从中找到了一些规律:

(1)两个或多个常见词组成的词组作为电影名,日常生活中很少遇到,但是经过一定的组合就可以精妙的反映影片的内容,比如电影《风车与火车》;

(2)对中国的成语稍加改造作为电影名以求适应影片需要,比如电影《龙爷虎孙》。

1.3 电影名中文分词的难题

中文电影名的分词由于两个原因使其比英文名实体识别困难得多:

(1)在中文句子中,词与词之间没有空格,导致分词的精确度不高;

(2)中文电影名没有明显的特征可以区分开(如:英文人名第一个字母大写),这对正确地识别中文电影名提出了挑战。

所以,如何解决这些电影名中文分词的难题对于其分词系统非常关键。

2 中文分词中传统的未登录词识别策略

分词是中文信息处理的第一步,也是中文信息处理中的瓶颈部分。中文以字为书写单位,以词为概念单位,只有将一个个概念单位准确划分出来,才能让计算机进一步理解概念及概念之间的关系。因此分词的准确性直接影响到后继的句法分析、语义理解和语用解析等工作的顺利进行。

目前所有的分词系统中都存在两大疑难问题:歧义切分和未登录词的识别。如何解决这两大难题成为众多中文信息处理工作者的攻关对象。而未登录词识别是中文信息处理中的一个难点,在大规模中文文本

的自动分词中,未被识别的未登录词是造成分词错误的一个重要原因。未登录词识别技术的突破对提高汉语自动分词和句法分析的准确性都有很重要的意义。

2.1 隐马尔可夫模型

隐马尔可夫模型(Hidden Markov Model, HMM)是由马尔可夫过程扩充而来的一种随机模型,它的基本理论是由数学家 Baum 及其同事在 20 世纪 60 年代初建立起来的,70 年代的中后期应用于语音处理,而在 80 年代逐渐广泛应用于文本处理中^[4]。

利用基于隐马尔可夫模型并结合概率估值公式来评价在真实文本中构成未登录词的能力,可以实现未登录词的自动识别系统,其准确率和召回率可以接受。但是仍存在一些需要解决的问题:由于语料库的规模有限,未登录词在真实文本中的覆盖率不完全,使得计算结果不够客观;对于那些小概率稀疏事件,系统没有较好考虑,造成系统识别错误。

2.2 基于决策树的汉语未登录词识别

这种识别策略把未登录词识别问题看成一种分类问题,即分词程序处理后产生的分词碎片分为‘合’(合成未登录词)和‘分’(分为两单字词)两类,然后用决策树的方法来解决这个分类的问题^[5]。

从语料库及现代汉语语素数据库中统计出六类知识:前字前位成词概率、后字后位成词概率、前字自由度、后字自由度、互信息、单字词共现概率,用这些知识作为属性构建了训练集,最后用相应算法生成了决策树。但是其准确率和召回率不是很令人满意,相信随着可利用的语料资源的增加,识别效果会得到相应改善。

2.3 基于统计的中文姓名未登录词识别

此识别策略的基本思想:首先扫描常规切分后得到的汉语句子,根据姓氏字表和名字用字字表建立潜在姓名链;其次计算每一个潜在姓名的可信度,再根据可信度处理潜在姓名链;最后,比较按姓名切分和常规切分两种情况下句子的可信度的大小来决定取哪一种分词结果。此识别策略的主要特点在于在训练过程中加入了奖惩机制。所谓奖惩机制,就是指在训练过程中,对正确识别出的姓名,分别对其姓用字和名用字及其同现词进行奖励,而对于识别错误的姓名,则要进行相应的惩罚。实践证明,此方法对提高识别中文姓名的召回率和精确率是行之有效的。

3 探索一种电影名未登录词的识别策略

对于电影名未登录词的识别主要的方法有:隐马尔可夫模型(如 HMM Tool)、角色定义、语料库训练等。它们都利用对语料库的学习,提出专用字的规律

和统计数据对未登录词的识别,但是识别准确率都不高。此识别策略则是使用词加权算法^[6]来识别电影名未登录词。

3.1 识别模型

本识别模型由输入、词典、最大匹配算法、词加权算法、输出等几部分组成,具体流程图见图 1。

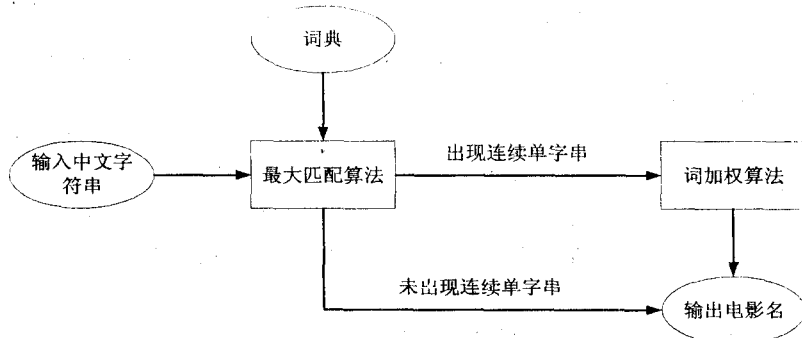


图 1 中文电影名分词流程图

输入的中文字符串通过正向最大匹配或逆向最大匹配进行切分,就可以将中文信息划分成为比较准确的电影名。在最大匹配分词中出现的单字词长串按照词加权算法识别出是否为未登录多字词。其中最大匹配是指从左向右或从右向左,每次取最长词得到切分结果的方法,词典是目前所统计的电影名的专业词典^[7]。

最大匹配方法是最基础的分词方法^[8],仅靠最大匹配就可以达到 80% 以上的分词正确率,但是最大匹配无法识别出组合歧义,对交叉歧义也有遗漏。在最大匹配的基础上加入其他的分词算法来进一步提高分词的准确率。对输入的中文信息通过最大匹配就可以得到比较正确的分词划分,所以将输入作为研究对象,进行正向或逆向最大匹配。匹配的过程中如果出现连续单字串,那么运用词加权算法识别其是否为未登录多字词;如果未出现连续单字串,就不需要使用词加权算法来识别未登录词而是直接作为分词结果输出为电影名。

3.2 词加权算法

通过采用全局统计信息等级划分与局部统计信息相结合的方式定义词加权计算公式。根据各类专有词库和各类已有电影名词库总结出统计知识,即中文电影名常用字,作为全局统计量并赋予这些字一个词性系数使其跟普通用字的等级区分开来^[9]。再利用一段中文输入中连续单字词出现的频率作为局部统计量,频率越高,表示在局部情况下这几个字作为一个词的几率越大。对单字串长度的数值进行乘方运算,来加大等级层次距离,减少误差。通过最大匹配算法识别出所有的单字词和多字词,对于连续出现的单字词

串,有可能是未登录的中文电影名。系统将识别出的连续单字串作为一个单位按输入的先后顺序存入数组,并记录每个单字串的内容、词性特征、字节长度 l 、出现频次 r 、词性系数 b 和词加权 q 。

词加权的大小由公式 $q = rl^2 \sum_{i=1}^n b_i$ 来确定,其值

直接决定单字串是否作为一个未登录多字词处理。可以对词加权大小 q 设定一个阈值,如果超过这个阈值,则表明单字串是一个未登录多字词,否则作为一个一个的单字词划分开处理。

对于每次输入的单字串系统都按顺序存储到数组中,并且查找数组中是否已有相同的记录,如果有相同记录,那么这个单字串和已有记录的出現频次都加 1,词加权也做出相应变动;如果已有记录中有与输入单字串中两个以上连续相同的字串,那么将长单字串进行分割,相同字串作为一个单位,剩余部分如果大于等于两个字的长度也分别作为一个单位,按前后顺序存储在数组中,否则将剩余部分删除,并且修正各个单位中的参数值。输入结束时,确定各单位中单字串的词加权,如果词加权大于之前设定的阈值,那么这个单位中的单字串被作为一个词划分。

该策略提出的电影名分词系统,基于最大匹配算法进行综合处理,对于单字串使用词加权算法识别其是否为未登录词^[10]。通过大量测试发现系统达到了比较好的准确率和召回率,使用较小的存储空间实现快速分词功能,针对小型系统具有良好的使用性能,可以面向小存储需求用户提供快速有效的服务。

4 结束语

文中笔者创新点:创造性地分析了对电影名进行中文分词的特点,并从中找出了电影名未登录词的一些规律,为之后电影名识别策略的提出做了准备;通过介绍隐马尔可夫模型和基于决策树的汉语未登录词识别两种识别策略并分析其不足之处,从而引出了新的电影名未登录词的识别策略;在电影名未登录词的识别模型中由于加入了最大匹配算法,仅靠最大匹配就可以达到 80% 以上的分词正确率,故使得识别策略的稳定性很强,同时有了词加权算法的加入,使得识别策略有了更好的准确性和灵活性。

需要进一步解决的问题是:

(1) 此策略针对小型系统具有良好的使用性能,可

(下转第 139 页)

当对比度低于 β , 则新开辟一个端口, 把这次分类结果作为一个新的类别。再利用改进的分类算法, 就可以把聚类中心控制在一个比较合理的范围内。这样就避免了分类过程中因为模式漂移把不同分类的信息合为一个类别。

3 仿真结果

文中采用文献[9]中的实验, 用天煌 THJ-3 型高级过程控制对象系统实验装置上采得的关于水箱液位控制的数据 30 组, 每组数据为 8 维。在进行 ART2 神经网络的训练和仿真的时候参数选择 $\rho = 0.98, a = 10, b = 10, c = 0.1, d = 0.9, \theta = 1/18$ 。

把训练好的神经网络用来进行分类时, 传统 ART2 网络由于多种原因产生了 8 个分类, 把同一类别的样本分为多类, 产生了较大分错率; 在改进的 ART2 网络下, 产生了 5 个分类, 而且产生的错分率要小的多。如表 1 所示, 改进的网络比传统的网络有更好的分类效果。

表 1 两种神经网络的分类比较

	样本数	实际类别数	网络的分类数	正确分类的样本数	错分率
标准 ART2	30	6	8	20	33.3%
改进 ART2	30	6	5	26	13.3%

4 结束语

针对 ART2 神经网络分类特点引入对胜利节点权

值对比进行增减分类调整的算法, 保证了分类的精度, 提高网络利用率, 而且有效控制聚类中心, 解决了权值模式漂移。

通过对冗余节点进行合理的归并和剔除, 使得保留下来的节点准确记录了输入样本的类别模式, 分类结果稳定可靠。

参考文献:

[1] Carpenter G A, Gossberg S. The ART of adaptive paten recognition by a self - organizing neural network[J]. IEEE, Trans on computer, 1988, 37(3): 77 - 88.

[2] Davenport M P, Tius A H. Multilevel category structure in the ART2 network[J]. Neural Network, 2004, 15(1): 145 - 158.

[3] Frank T, Kraiss K F, Kuhlen T. Comparative - analysis of fuzzy art and ART2A network clustering performance [J]. IEEE, Trans on Neural Network, 1998, 9(3): 544 - 549.

[4] 杨 兴, 朱大奇, 桑庆兵. 一种改进的 ART 型神经网络学习算法[J]. 计算机技术与发展, 2006, 16(9): 27 - 29.

[5] 郑君里, 杨行峻. 人工神经网络[M]. 北京: 高等教育出版社, 1992.

[6] Carpenter G A, Gossberg S. ART2: self - orgnization of stable category recognition codes for analog input pattern[J]. App () ptics, 1987, 26(23): 4919 - 4930.

[7] 谭锦华, 邝献涛. 基于 ART2 神经网络与动态聚类的分类器[J]. 控制工程, 2005(12): 64 - 66.

[8] 贾 鹏. 引入遗忘机制的 ART2 改进模型[J]. 计算机工程与应用, 2006(9): 60 - 62.

[9] 姚光顺. ART-2 神经网络的缺点及其改进[J]. 仪器仪表用户, 2008, 15(2): 112 - 113.

(上接第 136 页)

以面向小存储需求用户提供快速有效的服务, 但是对于较大的系统则还不尽人满意;

(2) 词加权算法的公式还有待改善, 各个参数还有待调整, 以求提高策略的识别效率。

参考文献:

[1] 俊 英. 垂直搜索引擎的研究与实现[D]. 哈尔滨: 哈尔滨工业大学, 2004: 30 - 35.

[2] 杜光芹, 张化祥, 赵瑞东. 主题 Web 挖掘研究[J]. 计算机技术与发展, 2008, 18(2): 94 - 97.

[3] 耿亚玮, 熊桂喜. 一种用于数据库搜索引擎的数据采集模型[J]. 微计算机信息, 2007, 33: 136 - 137.

[4] 郑家恒, 张 辉. 基于 HMM 的中国组织机构名自动识别[J]. 计算机应用, 2002, 22(11): 1 - 2.

[5] 秦 文, 苑春法. 基于决策树的汉语未登录词识别[J]. 中文信息学报, 2004, 18(1): 14 - 19.

[6] 苏 菲. 基于标记的规则统计模型与未登录词识别算法[J]. 计算机工程与应用, 2004(15): 43 - 45.

[7] 邱明明, 吴国新. 一种个性化垃圾邮件识别系统的设计[J]. 计算机技术与发展, 2007, 17(1): 136 - 138.

[8] Islam A, Inkpen D, Kiringa I. Applications of corpus - based semantic similarity and word segmentation to database schema matching[J]. The VLDB Journal, 2008, 17(5): 1293 - 1320.

[9] Pekar V, Mitkov R. Finding translations for low - frequency words in comparable corpora[J]. Machine Translation, 2006, 20(4): 247 - 266.

[10] Stegmann J, Grohmann G. Hypothesis generation guided by co - word clustering[J]. Scientometrics, 2003, 56(1): 111 - 135.