文章编号:1001 - 9081(2010)11 - 2928 - 04

音乐命名实体识别方法

佘 俊[1],张学清[2]
(1. 中国南方电网调峰调频发电公司 信息通信中心,广州 510630; 2. 电子科技大学 计算机科学与工程学院,成都 611731)
(sjshej@163.net)

摘 要:为了能快速、准确地将分散在 Web 网页中的音乐实体抽取出来,在全方位了解音乐领域中命名实体的特征的基础上,提出了一种规则与统计相结合的中文音乐实体识别方法,并实现了音乐命名实体识别系统。通过测试发现,该系统具有较高的准确率和召回率。

关键词:命名实体识别; 音乐命名实体;隐马尔可夫模型
中图分类号: TP18;TP311.13 **文献标志码**:A

Musical named entity recognition method

SHE Jun[1], ZHANG Xue-qing[2]
(1. *Information and Communication Centre, China Southern Power Grid Power Generation Company, Guangzhou Guangdong* 510630, *China*;
2. *School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu Sichuan* 611731, *China*)

Abstract: In order to extract musical entities from different Web pages quickly and correctly, this paper presented a hybrid approach based on rules and statistics for Chinese named entity recognition in music domain based on the characteristics in music domain, and implemented the musical named entity recognition system. The experimental results show that this system has a higher precision and recall rate.

Key words: named entity recognition; musical named entity; Hidden Markov Model (HMM)

0 引言

命名实体识别(Named Entity Recognition, NER)是目前文本信息自动化处理中一个基础而关键的技术。命名实体(Named Entity, NE)是文本中基本的信息元素,是文本中的固有名称、缩写及其他唯一标识,它往往指示了文章的主要内容。命名实体识别目前已经发展成一个独立的研究分支,COLING2002 就有专门的命名实体识别专题[1]。

国外对于英文命名实体识别的研究开始较早,英文命名实体的识别已经达到了较高的水平,MUC(Message Understanding Conference)会议测试的准确率和召回率可以达到 97% 左右[2]。英文命名实体的识别主要采用基于统计模型和机器学习的方法,如基于隐马尔可夫模型(Hidden Markov Model, HMM)[3-4]、最大熵模型[5-6]和支持向量机[7]等。这些方法利用英文命名实体的词频、词缀等统计信息,并结合一定的句法信息和语义特征来识别命名实体及确定其类别。

中文命名实体识别的研究起步比较晚,国内外关于中文命名实体识别的准确率和召回率的报告一般在 90% 左右。近年来,基于统计的中文命名实体识别方法是研究的主流。已有的中文命名实体研究方案,可以根据研究范围的不同,分为两种:

一种是个别解决方案,只识别某一类命名实体。这种方案针对某一类命名实体的特点,提出了有效的识别方法,如中文人名识别[8-9]、地名识别[10-11]、机构名识别[12]。但是这种方案忽视了不同种类命名实体间的歧义问题。

另一种是一体化解决方案,可以同时识别多种命名实体。大多数的方案采用机器学习的方法,如基于层叠隐马尔可夫模型的词法分析方法[13-14]、基于决策树[15]和基于组块[16]的方法等。但是它不能充分分析不同命名实体间的差异性,制约了整体的识别性能。

1 音乐命名实体识别的难点

本文主要研究中文文本中音乐命名实体的识别问题,音乐命名实体包括:

歌手名 是人名中特殊的一类,指特定歌手的固有名称、别名、英文名和译名,如“刘德华”、“华仔”、“周渝民”、“仔仔”、“布兰妮”、“小甜甜”等。

音乐组合名 指特定乐队、乐团和歌唱组合的固有名称,如“小虎队”和“Beyond”等。

歌曲名 指特定歌曲的固有名称,如:“梦醒时分”和“月亮代表我的心”等。

专辑名 指特定专辑的固有名称,如“永远的邓丽君”和“旷世情歌全记录”等。

目前,无论是中文还是英文命名实体识别都会遇到以下难题:首先,命名实体是一个开放的类,数量庞大,难以以列表或词典形式完全列举。以歌曲名为例,世界上有不计其数的歌曲,一一列举出来并不现实。其次,命名实体并非一个稳定的类,随着时间的推移,不断会有新的命名实体产生。例如一个新歌手的出道、一个新专辑的发布都会影响到音乐命名实体的内容和数量。

音乐命名实体的特殊性以及各类音乐命名实体的自身特

收稿日期:2010 - 04 - 30;**修回日期**:2010 - 07 - 16。
基金项目:国家自然科学基金资助项目(60973069);华为高校技术合作基金资助项目(YBIN2008126)。
作者简介:佘俊(1973 -),男,四川彭州人,高级工程师,主要研究方向:商业智能、知识管理; 张学清(1984 -),女,硕士研究生,主要研究方向:Web 文本挖掘、命名实体识别。

点,给识别带来了一定的难度。

1)歌手名和音乐组合名识别的难点。

歌手名和音乐组合名与上下文成词或自身成词的现象严重,容易产生歧义。例如:“天娱旗下艺人组合/至上励合/作客湖南卫视”可能被误分为“天娱旗下艺人组合/至上励/合作客湖南卫视”。如果不对这些情况进行特别的处理,许多歌手名和组合名就会被切开,降低识别的准确率和召回率。

歌手名和音乐组合名的长短不一,而且构成形式多样。歌手名的常见形式如下:姓和名组成,如张学友、任贤齐等;前缀和姓组成,如阿杜,小柯等;姓或名和后缀组成,如华仔、周董等;别名或艺名,如老狼、刀郎等。而音乐组合名,不像歌手名有规律,不仅用词比较随意,而且长度从一到十几个不等,如羽泉、南拳妈妈、凤凰传奇等。

区分歌手名和普通人名比较困难。在一个文本中,并非所有的人名都是歌手名。歌手名是人名的一个种类,具有相同的特性,因此,如何从人名中正确识别出歌手名是一个难题。

2)歌曲名和专辑名识别的难点。

歌曲名和专辑名的组成方式非常随意。许多歌曲名和专辑名都是很常见的词、短语或句子,如周华健的《朋友》、满文军的《懂你》、宋祖英的《长大后我就成了你》等。而且近来流行歌曲名中还出现了标点符号,如张韶涵的《亲爱的,那不是爱情》,这给区分歌曲和普通句子带来了很大难度。歌曲名、专辑名中含有歌手名,如安又琪的《你好,周杰伦》。这很可能会导致歌曲名、专辑名的漏选,以及歌手名的错选。歌曲名和专辑名的长度极其不固定。这导致歌曲名和专辑名的边界很难确定。

很多情况下,一首歌曲和一个专辑使用相同的名称,如许秦的专辑《又见溜溜的她》,其主打曲为《又见溜溜的她》。这将导致很难辨别一个实体到底是歌曲名还是专辑名。

经过对大量音乐领域文本的分析发现,在音乐命名实体自身及其上下文中,存在着许多规则和统计信息。

2　规则与统计相结合的音乐实体识别方法

该方法的核心思想为:首先,利用各种启发式规则信息,在原始音乐文本(即未进行分词)上,采用基于规则的音乐实体识别算法来识别部分明显的音乐命名实体。然后,对含有音乐实体的音乐文本进行分词。在分词之后,引入 HMM 识别音乐命名实体。最后,利用音乐实体库和修正规则,去除部分错误识别的歌手名和组合名,并修正歌曲名和专辑名的类型。该方法的实现框架如图 1 所示。

2.1　基于规则的 MNE 识别

在音乐文本中,音乐命名实体的前后位置常常会出现一些特定的符号或词。这些符号和词可能指示一个实体的开始,或一个实体的结束。比如,“歌手”这个词的后面很可能是一个歌手名,而“组合”的后面则可能是一个组合名;同样的“演唱”的后面可能是歌曲名,其前面则可能是歌手名或组合名,如此等等。基于上述原因,本文提出一种基于规则的 MNE 识别算法,即在分词和基于 HMM 的识别之前,利用音乐实体的出现规则来识别部分明显的音乐实体。

2.1.1　边界规则

规则的构建是基于规则的 MNE 识别算法的关键,规则越完善,识别的效果越好。由于音乐实体构成形式多样,且长度不一,很难在音乐的内部组成结构中提出通用性的规则。而在音乐实体的上下文中存在着大量的边界规则。虽然人工构造的规则准确率比较高,但比较耗时费力,且很难全面覆盖。因此,先由机器自动提取规则,再由人工进行筛选。

本文定义 8 种边界规则(Boundary Indicator dictionary, BI)来表示和存储音乐实体的边界,如表 1 所示。

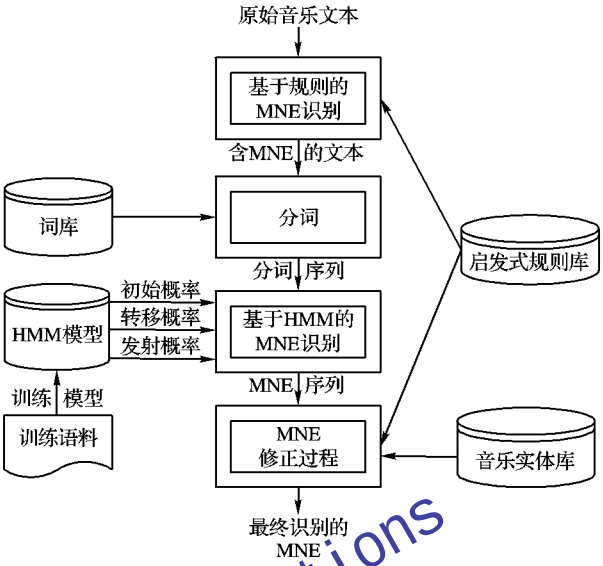

图 1　规则和统计相结合的音乐实体识别方法框架

表 1　音乐实体边界规则表

规则类型	缩写	组成部分
歌手名左边界	SIN_LBI	经常出现在歌手名左边的词或短语
歌手名右边界	SIN_RBI	经常出现在歌手名右边的词或短语
组合名左边界	BAN_LBI	经常出现在组合名左边的词或短语
组合名右边界	BAN_RBI	经常出现在组合名右边的词或短语
歌曲名左边界	SON_LBI	经常出现在歌曲名左边的词或短语
歌曲名右边界	SON_RBI	经常出现在歌曲名右边的词或短语
专辑名左边界	ALB_LBI	经常出现在专辑名左边的词或短语
专辑名右边界	ALB_RBI	经常出现在专辑名右边的词或短语

2.1.2　自动提取规则

以歌手名边界规则为例,规则提取的具体步骤:

1)遍历分词序列,收集出现在歌手名前后的词或短语,认为它们可能是歌手名的边界规则:

a)左边界规则。如果歌手名的前一个词为单字,则收集由其前两个词合并而成的短语;否则,收集该词。

b)右边界规则。如果歌手名的后一个词为单字,则收集由其后两个词合并而成的短语;否则,收集该词。

2)统计可能为歌手名左(右)边界规则的各词或短语出现在歌手名前(后)面的次数,并计算其作为左(右)边界规则的概率,其值为该词或短语出现在歌手名前(后)面的次数与该词或短语的出现总次数的比值。

3)将可能为歌手名左(右)边界规则的所有词或短语按上述概率从大到小排序,然后选取前 50% 个词或短语作为歌手名的候选左(右)边界规则,并将其保存至歌手名左(右)边界规则中。

2.1.3　识别

基于规则的 MNE 识别就是利用音乐实体的边界规则来识别音乐实体。在实现该算法时,将除了引号、书名号之外的标点符号作为所有实体的左边界规则和右边界规则。由于各类音乐实体的识别过程是类似的,因此,以歌手名为例来介绍

基于规则的识别算法：

1）读入未分词的音乐文本。

2）通过匹配的方式，从文本中提取出所有位于歌手名左边界和歌手名右边界之间的文本作为候选歌手名。

3）判断候选歌手名是否符合歌手名的要求，如长度等。若不符合，则丢弃。剩下的就是识别出来的歌手名。

2.2 基于统计的音乐实体识别

本文采用 HMM 作为统计模型，其主要原因是该模型能较好地捕获命名实体的特征现象和位置信息，具有高效的基于动态规划思想的解码算法，并且易于训练。

2.2.1 定义 HMM 参数

通过对音乐命名实体的分析，本文采用二元隐马尔可夫模型来识别各种音乐命名实体，即每个词的出现概率仅依赖于它前面一个词的出现概率。HMM 的输入文本是含有音乐实体的分词序列。那么，HMM 的观察值应该包括词和 *SIN*、*BAN*、*SON*、*ALB* 标识串，则定义模型的观察值集合 $\boldsymbol{O} = \{W_i, SIN, BAN, SON, ALB\}, 1 \leqslant i \leqslant m$，其中，$W_i$ 表示词库中的一个词，m 为词库中词的数目，*SIN* 表示已识别的歌手名，*BAN* 表示已识别的组合名，*SON* 表示已识别的歌曲名，*ALB* 表示已识别的专辑名。若用 M 表示所有可能观察值的数目，则 $M = m + 4$。

应用于命名实体识别的 HMM 的状态集合是那些能够反映命名实体组成的属性集合，即每一个状态都能代表某一类命名实体的内部组成成分、上下文或无关成分。分析歌手名、音乐组合名、歌曲名和专辑名的内部组成结构，发现以下主要特点：歌手名的首部一般是姓氏；歌曲名和专辑名的中部可能由多个词组成；歌手名和组合名常被嵌套到歌曲名和专辑名中。另外，歌曲名和组合名通常具有相同的上下文，与歌曲名、专辑名的上下文不同，且歌曲名和专辑名的上下文也不一样。因此，本文的 HMM 的状态集合 $\boldsymbol{S}$ 包含 $N = 29$ 个状态，$\boldsymbol{S}$ 集合的具体定义在表 2 中。

那么，HMM 的状态转移矩阵 $\boldsymbol{A} = (a_{ij})_{29\times29}, a_{ij} = \boldsymbol{P}(s_j \mid s_i)$ 表示从状态 s_i 转移到状态 s_j 的概率；发射矩阵 $\boldsymbol{B} = (b_{ij})_{29\times M}, b_{ij} = \boldsymbol{P}(o_j \mid s_i)$ 表示状态 s_i 对应观察值 o_j 的概率；初始状态概率为 $\boldsymbol{\pi} = \{\pi_1, \pi_2, \cdots, \pi_{29}\}, \boldsymbol{\pi}_i = \boldsymbol{P}(s_i)$ 表示初始状态是 s_i 的概率。

表 2 HMM 的状态集合

状态 s_i	含义	例子	状态 s_i	含义	例子
RB	歌手名的首部词	[郭]富城	QE	歌曲名的尾部词	月亮代表我的[心]
RI	歌手名的中部词	郭[富]城	QS	歌曲名自身成词	[烟火]
RE	歌手名的尾部词	郭富[城]	JB	专辑名的首部词	[星光]依旧灿烂
RS	歌手名自身成词	[高锋]	JI	专辑名的中部词	星光[依旧]灿烂
RU	上文与歌手名的首部成词	[为何]润东	JE	专辑名的尾部词	星光依旧[灿烂]
RV	歌手名的尾部与下文成词	张学[友好][吗]	JS	专辑名自身成词	[寓言]
RN	被嵌套的歌手名、组合名	你好[周][杰][伦]	KA	歌手名和组合名的上文	艺人
ZB	组合名的首部词	[至上]励合	LA	歌手名和组合名的下文	演唱
ZI	组合名的中部词	至上[励]合	KB	歌曲名的上文	插曲
ZE	组合名的尾部词	至上励[合]	LB	歌曲名的下文	这首
ZS	组合名自身成词	[零点]	KC	专辑名的上文	发行
ZU	上文与组合名的首部成词	[做梦]之旅的专访	LC	专辑名的下文	收录
ZV	组合名的尾部与下文成词	飞轮[海边]唱	ML	音乐实体间联结词	和、与
QB	歌曲名的首部词	[月亮]代表我的心	WL	与实体组成及上下文无关的词	
QI	歌曲名的中部词	月亮[代表][我][的]心			

2.2.2 训练得到状态转移矩阵和发射矩阵

在使用 HMM 识别实体之前，必须进行 HMM 训练，即获得初始状态概率、状态转移概率矩阵 和发射概率矩阵这 3 个关键参数。HMM 自动训练过程的基本流程如图 2 所示。

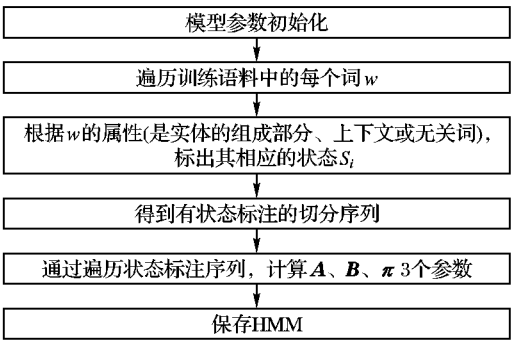

图 2 HMM 训练过程的基本流程

从图 2 可以看出，自动训练过程主要包括两个部分：状态标注过程和模型参数计算过程。其中，状态标注过程就是将训练语料中包含实体标注的切分序列的词性标注转换为状态标注；模型参数计算过程就是通过遍历带有状态标注的切分序列，计算出初始状态概率 $\boldsymbol{\pi}$、状态转移概率矩阵 $\boldsymbol{A}$ 和发射概率矩阵 $\boldsymbol{B}$ 的取值。

2.2.3 基于动态规划思想的过滤解码算法识别

通过训练得到 HMM 之后，便可以进行音乐命名实体的识别工作。识别过程就是从输入分词序列的所有可能状态序列中寻找最佳状态序列，然后将最佳状态序列中与音乐实体组成相关的状态进行组合，便得到各种音乐实体。假定有一个分词序列 $\boldsymbol{O} = (o_1, o_2, \cdots, o_m)$，音乐实体识别问题就相当于求解下式：

$$\boldsymbol{S}^* = \arg\max_{\boldsymbol{S}} \boldsymbol{P}(\boldsymbol{S} \mid \boldsymbol{O}) \tag{1}$$

其中：$\boldsymbol{S}$ 表示隐藏在 $\boldsymbol{O}$ 之后的状态序列，$\boldsymbol{S}^*$ 表示最佳状态序列。

$$\boldsymbol{S}^* = \arg\max_{\boldsymbol{S}} \prod_{i=1}^{m} \boldsymbol{P}(o_i \mid s_i)\, \boldsymbol{P}(s_i \mid s_{i-1}) \tag{2}$$

通常对式(2)取负对数，则：

$$\boldsymbol{S}^* = \arg\min_{\boldsymbol{S}} \left[-\sum_{i=1}^{m} (\ln \boldsymbol{P}(o_i \mid s_i) + \ln \boldsymbol{P}(s_i \mid s_{i-1})) \right] \tag{3}$$

在获得最佳状态序列之后,需要根据状态组合模板来查找音乐实体。根据各种状态的含义,本文定义了以下 4 种类型的状态组合模板。

歌手名:RB + RI* + RE,RI + RE,RE + RE,RU + RI* + RE,RB + RI* + RV,RS。

组合名:ZB + ZI* + ZE,ZI^ + ZE,ZE^ + ZE,ZU + ZI* + ZE,ZB + ZI* + ZV,ZS。

歌曲名:QB + QI* + RN* + QE,QB + QI* + RN^,RN^ + QI* + QE,RN^,QI^ + QE,QE^ + QE,QS。

专辑名:JB + JI* + RN* + JE,JB + JI* + RN^,RN^ + JI* + JE,JI^ + JE,JE^ + JE,JS。

例如有一状态序列:"老牌/WL 歌手/KA 林/RB 忆/RI 莲/RE 首/LA 唱/WL 新/WL 歌/ WL《/KB 柿子/QS 》/LB",通过匹配模板"RB + RI + RE"和"QS",可以识别出歌手名"林忆莲",歌曲"柿子"。

2.3　MNE 修正

经过基于规则的和基于 HMM 的音乐实体识别过程,已经识别出大部分音乐实体,但其中仍存在一些错误识别的实体,如非音乐实体、实体类型错误等。通过分析,总结出引起这些错误的主要原因有:

1)歌手名和普通人名的区分比较困难,可能会将普通人名误识别为歌手名;

2)歌手名和音乐组合名具有相似的上下文,歌曲名和专辑名常相同,可能会导致实体类型错误。

为此,针对歌手名和音乐组合名:由于歌手名和音乐组合名的数量比较少且容易收集,所以采用通过音乐实体库(包含歌手名库、组合名库)来过滤各种错误实体的方法。对于歌曲名和专辑名:由于歌曲名和专辑名的数量非常多,若采用同样的方法,则效率比较低。而且,通过分析发现已识别的歌曲名和专辑名的边界比较准确,只是还存在一些实体类型上的错误,如把歌曲名识别成专辑名等。因此,利用实体上下文的启发式规则来更正实体类型。

3　实验

实验所使用的语料有两个:一个是训练语料,另一个测试语料。从新浪、网易、搜狐等几大网站上收集的大量音乐领域文本,从中挑选了 1046 KB 的包含音乐实体的文本,其中包含了 156 KB 的歌手名、110 KB 的组合名、285 KB 的歌曲名和 207 KB 的专辑名,对这些文本进行标注,得到最终的训练语料。测试语料是从人民网、音乐在线、一听等网站上收集的音乐领域文本,大约为 572 KB,其中包括 80 KB 的歌手名、62 KB 的组合名、154 KB 的歌曲名和 105 KB 的专辑名。

实验 1　测试基于 HMM 的音乐实体识别方法的识别性能。

实验 2　测试基于规则的前处理和基于 HMM 相结合的音乐实体识别方法的识别性能。

实验 3　测试本文提出的规则和统计相结合的音乐实体识别方法的识别性能。该方法先使用基于规则的方法识别部分音乐实体,再使用基于 HMM 的方法来识别,最后对前两步识别出进行修正处理。

在实验中,采用了 3 个评测指标:召回率、准确率和 F 指数。各实验的结果如图 3 ~ 5 所示。

从结果数据以及曲线图可以看出,本文研究的规则和统计相结合的音乐实体识别方法具有良好的性能,无论在召回率、准确率还是 F 指数上,都高于其他两种方法。

4　结果分析

从实验结果中,可以看出仅用 HMM 的方法进行音乐实体识别的准确率和召回率都不太理想。首先,这是因为音乐实体识别自身存在的难点,如歌曲名和专辑名的组成方式非常随意,都是很常见的词、短语甚至句子;其次,该方法依赖于分词的结果,但分词仍然是个难题,其准确率不高;最后,二元隐马尔可夫模型只考虑词本身的发生概率而忽视了上下文对当前词的影响和词之间的联系,很容易因为词的错误切分而造成错误识别。

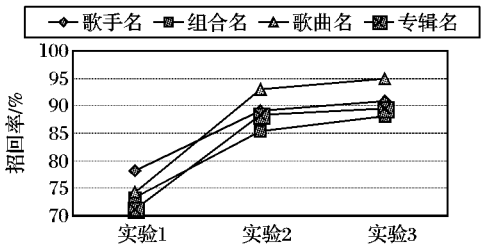

图 3　不同方法召回率的比较

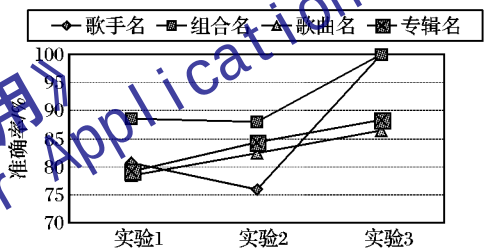

图 4　不同方法准确率的比较

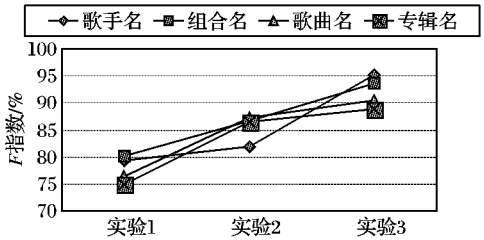

图 5　不同方法 F 指数的比较

因此,需要在基于 HMM 的方法之前,引入基于规则的前处理过程,来弥补 HMM 的不足。基于规则的 MNE 识别是在分词前进行的,它能够识别出一些基于 HMM 的方法无法识别出的实体。例如,句子"在演唱会上,至上励合演唱了",若"至上励合"在训练语料中很少出现,则基于 HMM 的方法很难将该组合名识别出来,但可以通过基于规则的方法将其识别出来。从实验结果中也可以看出,基于规则的前处理过程提高了音乐实体识别的召回率。

但是,召回率的提高同时,却带来准确率降低的问题。主要原因有:歌手名和普通人名的区分比较困难;歌手名和音乐组合名具有相似的上下文,歌曲名和专辑名常相同,可能会导致实体类型错误。而基于规则的后处理过程,即音乐实体修正过程刚好是用来解决这类问题。这使得识别的召回率和准确率都有一定的提高。

可见,基于规则的 MNE 识别、基于 HMM 的 MNE 识别和 MNE 修正 3 种方法相结合,使得识别的准确率、召回率和 F 指数都比较理想,证明本文所提出的规则和统计相结合的方案具有实用价值。

(下转第 2948 页)

的基础上提出相关实体子树集和语义相关实体子树集的概念,并在 LISA 算法基础上进行改进,实现了求解语义相关实体子树集的算法。最后将几种查询方式在 3 种数据集上进行查询,并以 XQuery 方式查询结果为标准,对几种查询方式的查准率进行了比较,实验结果验证了本文方法能够有效地提高查询结果的质量。未来的工作主要在于研究既能使用户表达出对目标 XML 片段所应有的结构信息,又能避免用户必须了解实际的 XML 数据组织结构的查询方式以及实现的方法。

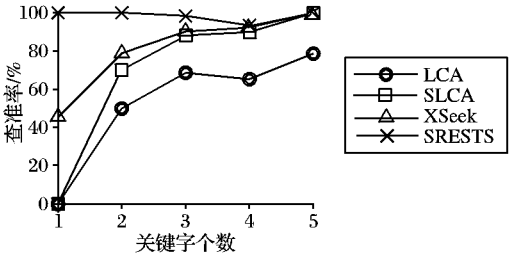

图 2 SIGMOD Record 上查准率

参考文献:

[1] 李颖. 关于 XML 检索和索引技术的研究[D]. 济南: 山东大学, 2009.

[2] SCHMIDT A, KERSTEN M, WINDHOUWER M. Querying XML documents made easy: nearest concept queries [C]// Proceedings of the 17th International Conference on Data Engineering. Heidelberg, Germany: IEEE Computer Society, 2001: 321 - 329.

[3] XU Y, PAPAKONSTANTINOU Y. Efficient keyword search for smallest LCAs in XML databases [C]// Proceedings of the ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2005: 537 - 538.

[4] LI Y Y, YU C, JAGADISH H V. Schema-free XQuery [C]// Proceedings of the 30th International Conference on Very Large Data Bases. Toronto, Canada: Morgan Kaufmann Press, 2004: 72 - 83.

[5] LI G L, FENG J H, WANG J Y, *et al*. Effective keyword search for valuable LCAs over XML documents [C]// Proceedings of the 16th ACM Conference on Information and Knowledge Management. Lisbon, Portugal: ACM Press, 2007: 31 - 40.

[6] LIU Z Y, CHEN Y. Identifying meaningful return information for XML keyword search [C]// Proceedings of the ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2007: 329 - 340.

[7] LIU Z Y, WALKER J, CHEN Y. XSeek: A semantic XML search engine using keywords [C]// Proceedings of the 33rd International Conference on Very Large Data Bases. New York: ACM Press, 2007: 1330 - 1333.

[8] TATARINOV I, VIGLAS S, BEYER K S, *et al*. Storing and querying ordered XML using a relational database system [C]// Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2002: 204 - 215.

[9] 李辛. 基于语义相关性的 XML 关键字查询的研究与实现[D]. 北京: 北京交通大学, 2009.

[10] 孔令波, 唐世渭, 杨冬青, 等. XML 信息检索中最小子树根节点问题的分层算法[J]. 软件学报, 2007, 18(4): 919 - 932.

(上接第 2931 页)

5 结语

经过实验证明,中文音乐实体识别系统已经表现出良好的性能,但仍存在一些不足,需要进一步的完善,主要包括以下方面:

1)本文提出的基于规则的音乐实体识别算法对规则冲突问题的考虑不够全面,即不同实体的边界库中可能存在相同的词或短语。可以结合词或短语作为某类边界规则的概率来解决冲突问题。

2)基于 HMM 模型的音乐实体识别算法未考虑各词的词性信息,这些信息对音乐实体的识别起重要作用。因此,下一步要融合词性信息到 HMM 模型中。

3)本文对歌曲名和专辑名的准确率还有待进一步提高,今后需要进一步讨论它们的识别方法。

总的来看,音乐命名实体识别技术已经取得了一定的成果,但对其研究还远没有结束。今后除了更一步改善音乐实体识别的效果外,还要结合应用来进行音乐命名实体识别,例如音乐搜索、音乐个性化推荐、音乐趋势分析等。

参考文献:

[1] 张晓艳, 王挺, 陈火旺. 命名实体识别研究[J]. 计算机科学, 2005, 32(4): 44 - 48.

[2] 愈鸿魁, 张华平. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报, 2006, 27(2): 87 - 95.

[3] RICHMAN A E, SCHONE P. Mining Wiki resources for multilingual named entity recognition, ACL-08[EB/OL]. [2009 - 12 - 12]. http://aclweb.org/anthology-new/P/P08/P08-1001.pdf.

[4] FU GUOHONG, LUKE K-K. Chinese named entity recognition using lexicalized HMMs [J]. ACM SIGKDD Explorations Newsletter, 2005, 7(1): 19 - 25.

[5] DAVID N, SATOSHI S. A survey of named entity recognition and classification [J]. Linguisticae Investigationes, 2007, 30(1): 3 - 26.

[6] XIONG DEYI, LIU QUN, LIN SHOUXUN. Maximum entropy based phrase reordering model for statistical machine translation [C]// Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. Morristown, NJ: Association for Computational Linguistics, 2006: 521 - 528.

[7] BENAJIBA Y, DIAB M, ROSSO P, *et al*. Arabic named entity recognition: An SVM-based approach [EB/OL]. [2009 - 12 - 12]. http://eref.uqu.edu.sa/files/eref2/folder6/f131.pdf.

[8] 张腾飞, 王晓磊, 王保云. 基于场景信息融合的中文姓名识别方法研究[J]. 计算机工程与应用, 2009, 45(34): 147 - 151.

[9] 贾宁, 张全. 基于最大熵模型的中文姓名识别[J]. 计算机工程, 2007, 33(5): 31 - 33.

[10] 唐旭日, 陈小荷, 许超, 等. 基于篇章的中文地名识别研究[J]. 中文信息学报, 2010, 12(2): 36 - 41.

[11] 李丽双, 黄德根, 岳广玲, 等. SVM 与规则相结合的中文地名自动识别[J]. 中文信息学报, 2006, 20(5): 77 - 82.

[12] 周俊生, 戴新宇, 尹存燕, 等. 基于层叠条件随机场模型的中文机构名自动识别[J]. 电子学报, 2006, (5): 16 - 21.

[13] 张华平, 刘群. 基于角色标注的中国人名自动识别研究[J]. 计算机学报, 2004, 27(1): 85 - 91.

[14] 张铭, 银平, 邓志鸿, 等. SVM + BiHMM: 基于统计方法的元数据抽取混合模型[J]. 软件学报, 2008, 19(2): 358 - 368.

[15] 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.

[16] ZHANG CHENG-ZHI, WANG HUI-LIN, LIU YAO, *et al*. Automatic keyword extraction from document using conditional random fields [J]. Journal of Computational Information Systems, 2008, 4(3): 1169 - 1180.