基于多层最大熵模型的句子主干分析

葛 斌 封孝生 谭文堂 肖卫东
(国防科技大学C4ISR技术国防科技重点实验室 长沙 410073)

摘 要 句子主干分析的主要任务是自动识别句子的主干成分。鉴于汉语句子之间成分的相关性,提出一种多层最大熵模型,它的底层最大熵利用句子的上下文特征识别主干词候选项,高层最大熵利用底层最大熵模型的计算结果,结合句子内的远距离特征和句子之间的关系,对底层最大熵模型识别出的主干词候选集进行分析。实验证明,该模型对于简单的主干成分识别正确率较高,对训练语料有一定的依赖;随着语料规模的增长,模型性能缓慢提升。

关键词 最大熵,多层最大熵模型,主干词,主干分析,自然语言理解
中图法分类号 TP311 **文献标识码** A

Skeleton Parsing Based on Multi-layer Maximum Entropy Model

GE Bin FENG Xiao-sheng TAN Wen-tang XIAO Wei-dong
(C4ISR Technology National Defense Science and Technology Key LAB, National Univ. of Defense Technology, Changsha 410073, China)

Abstract The main task of Skeleton Parsing is to identify the skeleton of a sentence automatically. Chinese Skeleton Parsing is a key problem in NLP. Because of the interrelation of the skeleton in the same context, a Multi-layer Maximum Entropy Model(MMEM) for the skeleton parsing was proposed. The low-layer ME analyzed skeleton by the context features while the high-layer ME analyzed skeleton by both the result of the low-layer ME and the features between sentences. The experiment showed that MMEM was efficient for Chinese skeleton parsing. A high precision was achieved under a small corpus while it was dependable on the scale of corpus. With the increasing of the corpus, the precision of MMEM improves slowly.

Keywords Maximum entropy, Multi-layer maximum entropy model, Skeleton word, Skeleton parsing, Natural language processing

句子主干分析是自然语言处理中的关键性问题之一,其主要任务就是自动识别句子的主干成分。汉语句子主干分析问题的解决对于机器翻译、信息抽取、自动文摘和语义检索等问题有着极其重要的意义。

最大熵作为一种统计方法已经在自然语言处理领域中得到广泛的应用。该模型具有简洁、通用的优点,能够灵活地选择语言特征。它把语言模型和计算模型分开,可以不用关心语言内部的细节,经过分析句子主干的语法、语义特征,选择适合于主干分析的特征集合。Adwait Ratnaparkhi 首次将最大熵模型应用到句子边界识别、词性标注等方面[1];Kamal 等使用最大熵模型进行文本分类[2];李荣陆等使用最大熵模型进行中文文本分类[3];李素建把最大熵用于组块识别并取得比较好的效果[4]。这些研究结果表明,最大熵模型在 NLP 的应用中表现出了良好的性能。

鉴于汉语句子之间成分的相关性,本文在最大熵模型的基础上提出一种多层最大熵模型。

1 基本思想

最大熵模型主要用于识别和分类功能。当一个句子的主语候选词多于一个时,最大熵只能简单地从中选择出概率最大的主语选项,因此一般采取阈值过滤的方法选出概率较大的几个事件。在句子主干分析中,若仅简单地过滤小概率事件,则不能有效利用主、谓、宾 3 种句法成分的关系。

多层最大熵模型(Multi-layer Maximum Entropy Model, MMEM)分析算法,通过两层最大熵模型的叠加来实现句子主干分析。底层最大熵模型利用句子里的上下文特征识别主干词候选项,根据训练结果从候选集中选出概率超过某个阈值的词,加入到高层最大熵模型的候选集中;高层最大熵模型利用底层最大熵模型的计算结果,结合句子内的远距离特征和句子之间关系,对底层最大熵模型识别到的主干词候选集进行分析,利用主谓宾的语法关系识别出概率最大的句子主干;最后经过基于语境相似度的平滑处理得出最终的识别结果。

多层最大熵模型的训练分为两个部分:第一部分是底层最大熵模型的训练,这一层的模型训练主要根据词的上下文特征集来进行,没有利用远距离的词的信息;第二层是高层最大熵模型的训练,这一层的训练利用句中和句子间的远距离

到稿日期:2010-01-29 返修日期:2010-04-06 本文受国家自然科学基金项目(60903225,60172012),湖南省自然科学基金项目(03JJY3110)资助。
葛 斌 博士生,主要研究方向为文本分析和语义检索,E-mail:gebin1978@gmail.com;封孝生 副教授,主要研究方向为信息资源管理;谭文堂 博士生,主要研究方向为社会化网络分析;肖卫东 教授,博士生导师,主要研究方向为信息管理、信息系统集成。

特征，主要是句中主干候选词的词和位置等。

2 基于 MMEM 的句子主干分析方法

2.1 最大熵模型的特征选择

分析句子主干，首先要从文本中选择对该项任务有影响的各种特征，并把这些特征按一定格式嵌入到最大熵模型中。要充分考虑影响句子主干分析的各种因素和规则，使选择的特征集合尽可能地符合分析任务的要求。同时，如果不对特征进行选择过滤，可能导致模型的特征空间变得很大。由于汉语词的数量巨大，手工选择每个词的特征是不现实的，因此需要引入一定方法进行特征选择，常用的包括阈值过滤[5]和改进的基于最大似然增益的特征选择算法[6]，MMEM 采用阈值过滤算法。

基于阈值的特征选择基于这样的假设：不常出现的特征是噪声或不相关的，只有那些出现次数比较大的特征才真正代表了数据的特性。因此，通常选择候选特征空间中所有在训练数据中出现次数大于某一阈值的特征，作为特征选择的结果。阈值的最优值与任务和数据相关，通常通过多次实验来确定。文献[5]通过实验证明一般取阈值为 5，模型性能比较好。MMEM 的阈值取为 5。

2.1.1 底层最大熵模型的特征获取

根据可能影响主干分析的因素得到一个特征选取的空间，并定义特征模板。通过分析实际的标注语料库，对模板进行实例化，自动构建一个可能的候补特征集合。对于一些特殊的不能通过模板获取的特征，可以根据特征的定义格式，手工编写后加入到候补特征集中。

首先定义如下特征空间，目的是找出影响句子主干分析的因素。

1）词性信息：当前词及其前后词的词性。

2）词：当前词及其前后上下文的特定词。

根据特征空间，定义模型中的模板，每个模板只考虑一种因素，是当前词上下文的一个函数，称之为原子模板。

仅仅用原子特征不能完全体现当前词上下文的特征。李素建在组块分析中使用复合特征，取得了很好的效果[4]。于浚涛认为，原子模板只能表现出单个位置的词或者词性信息，是片面和单调的，容易造成事件的期望与实际结果偏差巨大，从而导致权值 λ 反常的结果[7]。根据主干分析的特点，制定包含两个以上原子特征的复合特征。特征由以下 3 部分组成：

1）当前词的上下文特征函数，如(PosCurr，WordCurr-2)表示当前词的词性标注结果，并且当前词的前面第二个词为某个特殊的词。

2）当前词的主干标注结果。

3）特征函数值。

用二值函数表示复合特征，如特征 $PosCurr(N)$ And $WordCurr-1(的)=na$，表示当前词为名词，而且前一个词是“的”，当前词为主语，用二值函数表示为

$$f_i(x,y)=\begin{cases}1, & \text{if } POSCurr=N \text{ And } WordCurr-1=的 \\ & \quad \text{And } y=subject \\ 0, & \text{otherwise}\end{cases}$$

确定好特征模板之后，可以根据模板自动从语料中获取特征。MMEM 采用以下算法来自动获取特征：

1）建立一个空特征集合。

2）分析已经标注主干的语料库，对每个主干词依照模板给相应的模板函数赋值，对模板进行实例化。每个模板都可以自动产生一个特征。

3）对生成的特征进行分析，如果特征集合里已存在该特征，则对该特征计数加 1；若不存在，则把该特征加入到特征集合中去，并对该特征计数为 1。

4）反复执行 2)，3)，直到处理完语料库中的所有主干词。

2.1.2 高层最大熵模型的特征获取

高层最大熵模型利用两方面的特征，一是底层最大熵模型计算出来的概率；二是利用主谓宾的关系和句中的远距离特征，识别底层最大熵计算出来的主干候选项，底层最大熵把概率高于某个阈值的词加入到高层最大熵的候选集，高层最大熵模型利用底层最大熵模型的结果，包括计算出来的概率，综合考虑主语、谓语、宾语候选集中各个词的关系来识别主干，所以高层最大熵考虑的特征空间主要是：

1）底层最大熵计算出来的概率。底层最大熵计算出来的概率仍然是高层最大熵识别的主要根据，底层特征中出现的某些特征可能就直接说明了词的语法功能。如“漂亮”后面跟着词“吸引”，就说明“漂亮”是主语。所以高层最大熵必须把底层最大熵模型的概率作为主要的特征，并且直接把这个特征加入最终特征集。

2）各个候选集中的词。句子中所有可能充当句子主干的词对当前词有很大的表征作用。对于候选词“解决”，如果宾语候选词中有“问题”，则“解决”就是谓语。所以高层最大熵模型需要考虑同一句子中的其它候选词，把其它词作为重要的特征。

3）前一个句子的候选集。通过分析文本，容易发现相邻的句子谈论的话题往往是一致的，前一句的主语可能是当前句子的主语，也可能是当前句子的宾语。如“网络是重要的媒体，媒体还包括很多其它类型”，这里“媒体”是前一个句子的宾语，在后面一个句子充当主语。由于相邻句子有相关性，因此本文把前后句子的候选词作为当前句子的特征。

根据以上特征空间为高层最大熵定义特征模板。先根据特征空间定义原子特征，这些特征都是句子中或句子间远距离特征。定义远距离特征需首先引入 3 个定义：

定义 1(句首)　句子的前半部分称为句首。

定义 2(句尾)　句子的后半部分称为句尾。

定义 3(句中)　去掉句首、句尾，剩下的句子中间部分称为句中。

最大熵的一大优点就是特征选择机制灵活，可以方便地把一些跨距离的特征加入到模型中。前述对于模板的设定都选在当前词左右不超过 3 个词的距离，然而对于特殊的语言现象，对当前词发生影响的因素可能不在这个距离范围之内。通过对语料进行观察，定义了远距离特征加入候选特征集，如 InForSentence 特征主要是面向主语候选项和宾语候选项，表示词在句首位置；InMidSentence 特征主要面向谓语候选项，表示词在句中位置；InEndSentence 特征主要面向宾语候选项，表示词在句尾位置。和底层最大熵一样，需要定义复合特征模板来表征复杂的语言现象。

2.2 高层最大熵模型的参数训练

高层最大熵的参数估计与底层最大熵模型有所不同，高

层最大熵模型必须考虑底层最大熵的计算结果和句子间的远距离特征，所以高层最大熵模型必须先得到前后句子的主干词候选集和当前句子的主干候选集才能对参数进行训练。同时，高层最大熵在语料中无法利用底层最大熵模型计算的概率来进行训练，所以底层最大熵模型计算的概率的权重必须通过实验来确定。MMEM 使用 IIS 算法[5] 对高层最大熵进行参数估计。

首先确定一个权值 $\lambda=1.0$，然后进行参数估计和训练，得到高层最大熵模型。根据得到的模型对文本进行测试，通过分析测试结果调整权值。在测试语料为 10000 句的情况下，从实验结果中得到在 $\lambda=2.0$ 时系统的整体性能比较好，MMEM 最后确定 $\lambda=2.0$ 为底层最大熵模型概率的权重。

2.3 句子主干分析

分析给定的句子时，首先对句子进行预处理，然后提取句子的特征。通过底层最大熵模型得到 3 个候选集：主语候选集、谓语候选集、宾语候选集；高层最大熵模型利用从句子中和句子间提取的远距离特征对底层最大熵模型得到的候选集做进一步分析，得到最终识别结果。

基于多层最大熵模型的句子主干分析框架如图 1 所示。

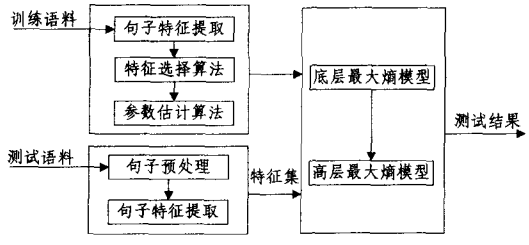

图 1 多层最大熵模型句子主干分析系统框架

2.3.1 底层最大熵模型句子主干分析

底层最大熵根据前述的训练方法得到，训练的特征是根据前述定义的特征集，主要考虑当前词及其上下文。根据前述分析，最大熵方法就是在特征矩受限的分布族中找到使条件熵最大的那个分布，可形式化表示为

$$p=\arg\max_{p\in C}-\sum_{x,y}\hat{p}(x)p(y|x)\log\hat{p}(x)p(y|x)$$

可以求出：

$$p=\frac{\exp(\sum_i\lambda_i f(x,y))}{Z_\lambda(x)}$$

根据训练结果得到 λ_i，可以计算出一个词充当句子主干的概率 p 和不充当句子主干的概率 $\bar{p}$。如果 $p>\bar{p}$，则该词可能充当句子主干。若只把 $p>\bar{p}$ 的词加入到高层最大熵的候选集中，概率太低的候选词会成为噪声，导致高层模型的计算量很大，因此引入基于频率的阈值过滤。如果整个句子比较复杂，可能导致整个句子的主干词的概率都比较低，所以阈值不能简单地定义为一个常数，必须动态计算。

经过底层最大熵识别得到的主干词集合，一般为 3 个集合 $S_{subject}$，$S_{Predicate}$，S_{object}。

定义 4（概率阈值 CP） 当主干候选词的概率低于这个阈值时，从所属的主干词候选集中过滤掉，$CP=\frac{\sum_{s\in S}p_s}{count(S)}\times 0.5$。

具体算法如下：

1）把 $S_{subject}$，$S_{Predicate}$，S_{object} 置为空。

2）通过底层最大熵模型识别文本中的句子的主干，得到的词加入 $S_{subject}$，$S_{Predicate}$，S_{object} 中。

3）对 3 个集合 $S_{subject}$，$S_{Predicate}$，S_{object} 分别计算 $CP_{subject}$，$CP_{predicate}$，CP_{object}。

4）对 3 个集合 $S_{subject}$，$S_{Predicate}$，S_{object} 中的每个词，比较其概率与概率阈值的大小。如果小于阈值，则从集合中删除该词。

5）若集合中所有的词的概率都大于概率阈值 CP，退出。

2.3.2 高层最大熵模型句子主干分析

高层最大熵模型利用句子和句子间的远距离特征分析当前候选词，得到当前词可能充当主干的概率。通过比较候选词之间概率的大小，得到概率最大的词。具体的识别过程和底层最大熵模型是一致的，此处不再说明。

高层最大熵模型分析完后，鉴于汉语句子句型的复杂性，为提高分析效果，需要对识别选项进行处理。

2.3.3 高层最大熵模型句子主干分析结果处理

分析最大熵模型后，仍然有可能得到概率相近的候选项。如果简单地把概率最大的项作为结果，可能会丢失本来也是主干的词。

识别在最大熵模型后，对其中的紧缩句进行处理。紧缩句是一种用单句形式表达复句内容的特殊句式。这样的句子一般只有一个主语，但是有多个谓语，谓语之间存在转折、条件、因果等复句关系。紧缩句结构灵活，具有比较明显的结构特征，一般句中存在以下关联词语：

“一…就”，如“他一甩手就走了”。

“非…不可”，如“这件事非他不可”。

“再…也”，如“困难再多也吓不倒我们”。

“越…越”，如“他越想越气”。

单个副词关联，主要有就、也、才、再、都、还等。

根据上述特征判断当前的单句是否为紧缩句。如果是紧缩句，则把关联词之间或者其后的动词都作为谓语。

2.4 数据稀疏问题及平滑算法

由于最大熵模型没有完全体现先验知识，因此 MMEM 必须进行平滑。虽然最大熵模型自身具有平滑机制，但模型平滑机制比较简单，特别是当训练的语料较小时，数据稀疏会严重影响模型的性能。

目前，针对最大熵模型的平滑技术研究较少，但对于 N-gram 的平滑技术研究则相对较多，主要包括加法平滑（Additive Smoothing）[8]、线性插值平滑（Linear Interpolation Smoothing）、Good-Turing 估计[9]、折扣参数平滑（Discounting Smoothing）[10]、Katz 回退平滑[11] 等。现有研究通常是把对 N-gram 的平滑技术引入到最大熵模型的平滑中，然而这些方法都是简单地给未出现的词或者特征赋予一定概率，以消除零概率，不能从有限的语料库中得到未出现的词或者特征的任何信息。MMEM 针对上下文特征和主干词采取两种不同的平滑算法，从两方面对最大熵进行平滑。

1）上下文特征平滑。主要对词的上下文特征进行平滑。由于语料库的限制，导致有些特征并没有在语料库中出现，而在实际应用中可能出现，因此不能简单地认为其概率为 0。对于主干词的上下文特征，使用高斯先验平滑算法。

2）主干词的平滑。目前对于句子主干分析没有相应的大型语料库，很多实际应用中大量出现的词并没有出现在语

料库中或者在语料中出现的次数很少，简单地给予这些词以0概率与现实不符。本文采用一种基于语境相似度的最大熵模型平滑算法，通过统计和比较词的上下文得到汉语词的语境相似度，利用最相似的词对训练语料中未出现的词进行平滑。

3 实验与分析

3.1 实验系统构建

基于MMEM的句子主干分析原型系统结构如图2所示。

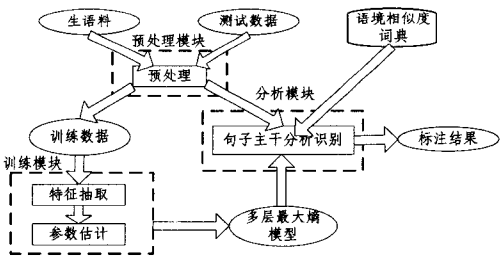

图2 实验系统结构图

系统主要由预处理模块、训练模块和分析模块3部分组成，各模块功能描述如下：

1）预处理模块主要对文本进行预处理，即文本分词、断句、单复句的识别和词组的处理。

2）训练模块主要根据最大熵模型的特征获取算法从标注语料中获取特征、对最大熵模型进行参数估计和参数选择。训练分为底层最大熵模型和高层最大熵模型的训练。

3）分析模块利用训练模块得到的最大熵模型，对汉语句子进行主干分析，同时利用语境相似度计算平滑最大熵模型。通过两层最大熵模型的识别，最后输出句子标注的结果。

3.2 主干标注语料库构建

主干标注语料库是指标注了主干词的语料。实验中训练所用的语料来自3个部分：

1）对兰开斯特汉语标注语料库(LCMC)[12,13]进行处理和转换。由于兰开斯特汉语标注语料库是一个平衡语料库，没有标注句子的主干，因此需要对该语料进行预处理。实验中通过一个编辑程序辅助该语料的主干标注。

2）计算所机器翻译树库。该树库是中科院计算技术研究所与北京大学计算语言学研究所联合开发汉英机器翻译系统时制作的，树库规模为3082个句子。该树库标注了主要短语类型。

3）生语料整理。由于上面两个语料规模都不是很大，实验搜集了大量的生语料，其中包括人民日报语料库中1998年1月份的标注语料。对这些文本进行分词和词性标注，然后人工标注句子主干。

在收集语料的过程中，尽量使用现有的语料，避免重复工作。在处理生语料时，尽量用计算机辅助标注，以减少人工标注的工作量。最后整理得到主干标注语料库约100多万词、14000多个句子。

3.3 实验结果

本文把语料分为两个部分，其中10000个句子作为训练语料、4000个句子为测试语料。

实验采用自然语言处理领域广泛使用的评估标准：召回率、精确率及综合评价函数F，其中

$$F=\frac{2\times\text{精确率}\times\text{召回率}}{\text{精确率}+\text{召回率}}$$

实验得到的结果如表1所列。

表1 MMEM模型试验结果

句子主干类型	召回率(%)	精确率(%)	F(%)
主语	85.81	86.16	85.98
谓语	84.6	85.49	85.04
宾语	86.17	87.22	86.69

为测试模型性能对于语料的依赖程度，以及语料规模对于模型性能的影响，采用不同容量的训练语料分别训练模型。分别采用2000句、5000句、8000句、10000句进行训练，仍然采用原来的4000个句子作为测试语料，得到如表2所列的结果。

表2 不同语料规模对MMEM模型性能的影响测试结果

测试结果	句子主干类型	召回率(%)	精确率(%)	F(%)
2000句	主语	75.8	75.16	75.47
	谓语	74.37	72.94	73.64
	宾语	76.58	75.02	75.79
5000句	主语	80.37	79.56	79.96
	谓语	79.36	80.49	79.92
	宾语	80.17	80.22	80.19
8000句	主语	83.83	84.61	84.22
	谓语	82.15	83.64	82.89
	宾语	83.73	85.80	84.75
10000句	主语	85.81	86.16	85.98
	谓语	84.6	85.49	85.04
	宾语	86.17	87.22	86.69

图3显示了语料规模对模型性能的影响测试结果，从图中可以看出，模型的性能在语料较小时性能较差，但是随着语料规模的增长，系统性能缓慢提升。

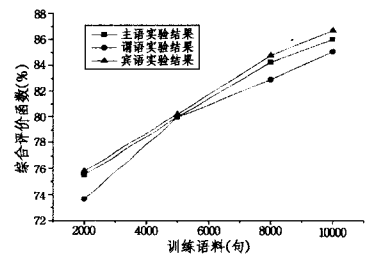

图3 语料规模对模型的影响试验结果

3.4 实验结论分析

上述实验证明，采用多层最大熵模型分析汉语句子的主干词，能够达到较高的识别率。但由于采用的训练语料比较小，限制了模型的性能。同时，从实验结果看出，系统对于复杂的句子，主干词识别的效果不是很好，主要存在以下问题：

1）词组和特殊结构的识别准确度不高。

2）单复句的判断精确度不高，有时把句首状语当作单句。

3）对于句子中出现的连动式谓语，识别效率比较低。

结束语 利用多层最大熵模型识别汉语句子，实验证明是有效可行的。在训练语料较小的情况下，模型取得了比较高的识别准确率。在本文方法的基础上可进一步分析组块，提出的平滑算法也可应用到隐马尔科夫模型等统计语言模型中。

参考文献

[1] Adwait R. Maximum entropy models for natural language ambi-

guity resolution[D]. Pennsylvania:Pennsylvania,1998
[2] Nigam K,Laferty J,McCallum A. Using Maximum Entropy for Text Categorization[C]// Workshop on Machine Learning for Information Filtering. 1999:61-67
[3] 李荣陆,王建会,陈晓云,等. 使用最大熵模型进行中文文本分类[J]. 计算机研究与发展,2005,42(1):94-102
[4] 李素建. 汉语组块计算的若干研究[D]. 北京:中国科学院计算技术研究所,2002
[5] Berger A L,Pietra S A D,Pietra V J D. A Maximum Entropy Approach to Natural Language Processing[J]. Association for Computational Linguistics,1996,22(1):39-71
[6] Jelinek F, Mercer R L. Interpolated Estimation of Markov Source Parameters from Sparse Data[C]// Proceedings of the Workshop on Pattern Recognition. Practice,1980:381-398
[7] 于浚涛. 基于最大熵的汉语介词短语自动识别[D]. 大连:大连理工大学,2006
[8] Zhai C,Lafferty J. A Study of Smoothing Methods for Language Models Applied to Information Retrieval[J]. ACM Transactions on Information Systems,2004,22(2):179-214
[9] Gao J,Goodman J,Li M,et al. Toward a Unified Approach to Statistical Language Modeling for Chinese[J]. ACM Transactions on Asian Language Information Processing,2002,1(1):3-33
[10] 黄永文,何中市. 基于全局折扣的统计语言模型平滑技术[J]. 重庆大学学报:自然科学版,2005,28(8):51-55
[11] 黄建中,王肖雷. Katz 平滑算法在中文分词系统中的应用[J]. 计算机工程,2004,增刊(1):370-372
[12] 许家金. 兰开斯特汉语语料库介绍[EB/OL]. http://nlp.org/,2006
[13] Yang Xiaojun. Survey and Prospect of China's Corpus-based Researches[C]// The Corpus Linguistics Conference. Lancaster University(UK). 2003

(上接第 137 页)

[6] Dou D, LePendu P. Ontology-based Integration for Relational Databases[C]// Proceedings of the 2006 ACM Symposium on Applied Computing. Dijon,France,2006:461-466
[7] Wang Jin-peng, Lu Jian-jiang, Zhang Ya-fei, et al. Integrating heterogeneous data source using ontology[J]. Journal of Software,2009,4(8):843-850
[8] 吕艳辉,马宗民,王玉喜. 基于关系数据库的 OWL 本体构建方法研究[J]. 计算机科学,2009,36(7):153-156
[9] Laborda C P, Conrad S. Bringing Relational Data into the Semantic Web using SPARQL and Relational. OWL[C]// Proceedings of the 22nd International Conference on Data Engineering Workshops. Atlanta,USA,2006:55
[10] Erling O,Mikhailov I. Integrating Open Sources and Relational Data with SPARQL[C]// The 5th European Semantic Web Symposium and Conference. Tenerife,Spain,2008:838-842
[11] Weiske C, Auer S. Implementing SPARQL Support for Relational Databases and Possible Enhancements[C]// Proceedings of the 1st Conference on Social Semantic Web. Leipzig,Germany,2007:69-80
[12] Halevy A. Answering queries using views:A survey [J]. Very Large Data Bases Journal,2001,10(4),270-294
[13] Abiteboul S,Duschka O M. Complexity of answering queries using materialized views[C]// Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems. Seattle,US,1998:254-263
[14] Beeri C,Levy A,Rousset M C. Rewriting queries using views in description logics[C]// Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems(PODS). Tuscon,Arizona,1997:99-108
[15] Baader F,Hollander B. Kris:Knowledge representation and inference system[J]. SIGART Bulletin,1991,2(3):8-14
[16] Levy A,Rajaraman A,Ordille J. Querying heterogeneous information sources using source descriptions[C]// Proceedings of the 22th International Conference on Very Large Data Bases. Bombay,India,1996:251-262
[17] Pottinger R,Halevy A. MiniCon:A Scalable Algorithm for Answering Queries Using Views[J]. The VLDB Journal,2001,10(2/3):182-198

(上接第 133 页)

图 2 显示的则是两个关系数据流分布不协调一致的情况,BSJ 算法能显示出其优越性。这是由于两个关系流分布不协调一致,也就是说在 R_1(或者 R_2)中可能存在一些关键字的大量取值,在 R_2 中与此相等的关键字含量过低,因而通过算法能够准确地找到关系流中出现概率较低的元组及其相应分区。

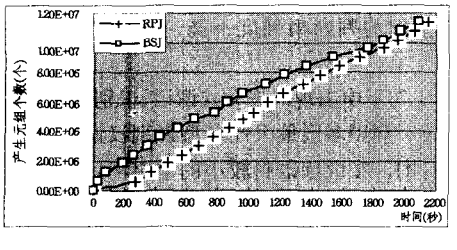

图 2 数据分布不协调时的性能

结束语 提出了一个新的内存刷新策略,再对数据流上的数据频率进行近似的统计分析,将分析结果应用于关系连接的输出流上,很好地反映了输入流中的数据分布情况,提高了刷新策略的准确性。

参 考 文 献

[1] Lawrence R. Early Hash Join:A configurable algorithm for the efficient and early production of join results[C]// VLDB. 2005:841-852
[2] Wilschut A N,Apers E M G. Pipelining in query execution[C]// Proc. of the International Conference on Databases,Parallel Architectures and their Applications. Miami,USA,March 1990
[3] Urhan T,Franklin M J. XJoin:Getting Fast Answers from Slow and Burst Networks[R]. CS-TR-3994. Computer Science Department,University of Maryland,1999
[4] Mokbel M F,Lu Ming,Aref W G. Hash-Merge Join:A Non-blocking Join Algorithm for Producing Fast and Early Join Results[C]// Proceeding of the 20th International Conference on Data Engineering. Washington:IEEE Computer Society,2004:251-263
[5] Tao Yufei,Yiu M L,Papadias D,et al. RPJ:producing fast join results on streams through rate-based optimization[C]// Proceedings of the 24th ACM SIGMOD International Conference on Management of Data. New York:ACM press,2005:371-382