生物事件触发词识别方法研究

魏小梅[1,2] 黄 钰[2] 陈 波[1] 姬东鸿[1]
(武汉大学计算机学院 武汉 430072)[1] (华中农业大学信息学院 武汉 430070)[2]

摘 要 从生物文献中抽取生物事件对于生物领域的知识挖掘起着重要的作用,而事件触发词的识别是生物事件抽取的一个关键步骤。系统分别采用词汇及其上下文特征、短语标记特征、词聚类特征以及统计的词典特征构造不同的基于词级的CRF模型,用于生物事件触发词的标记。然后针对不同的触发词类型选择对应最优的标记模型,构造了一个混合CRF模型。在BioNLP 2009 ST语料库上进行了实验评估,结果表明提出的方法取得了很好的性能,为生物事件的抽取建立了良好的基础。

关键词 生物事件,触发词,CRF模型,Brown Cluster,特征
中图法分类号 TP391.1 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.10.049

Research on Tagging Biomedical Event Trigger

WEI Xiao-mei[1,2] HUANG Yu[2] CHEN Bo[1] JI Dong-hong[1]
(College of Computer, Wuhan University, Wuhan 430072, China)[1]
(College of Informatics, Huazhong Agricultural University, Wuhan 430070, China)[2]

Abstract Event extraction from biomedical literature plays an important role in the knowledge mining in biomedical domain. The trigger identification is the key step in biomedical event extraction. We used rich features including lemma, context, phrase label, word cluster and learned trigger dictionary to build several kinds of CRF models. Then we chose the best model for each type of triggers to combine a hybrid model. The evaluation on the BioNLP 2009 ST data set shows that our approach achieves good performance, which lays foundation for biomedical event extraction.

Keywords Biomedical event, Trigger, CRF model, Brown Cluster, Feature

1 前言

高通量技术的应用,产生了大量的生物数据和生物研究课题,使得越来越多的研究者投入到生物技术研究中,从而生物文献也海量增长。从生物文献中抽取数据也成为生物自然语言处理(BioNLP)领域的重要研究课题。其中,在生物事件抽取方面,影响越来越大的是从BioNLP 2009[1]开始的系列GE share task工作。该任务定义了9类分子生物事件,并提供了训练、开发和测试数据集,数据集中的生物文献来自于MEDLINE。所有数据集都提供了蛋白质标注,训练和开发数据集提供了生物事件标注,任务要求对测试数据集抽取生物事件。图1显示一个语料中生物事件标注的例子,上面方框中是来自生物文本的句子,并做了蛋白质标注(以T*i*标识);下面方框中是根据上面方框中的标识识别的触发词(以T($i+j$)标识)和事件(以E*i*标注)。事件关系图如图2所示,句子中粗体标识的词是触发词,虚线框中是触发词的类型;句子中斜体标识的是蛋白质,连线表示的是触发词和蛋白质构成一个生物事件,其中箭头方向是事件的触发词。

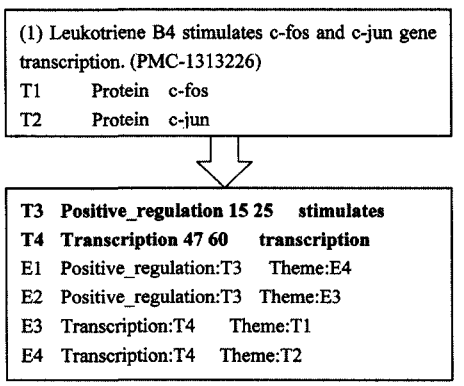

图1 语料中的事件标注

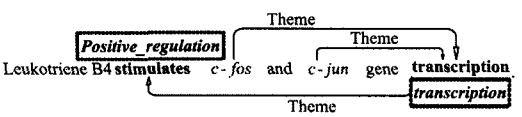

图2 一个句子中的生物事件实体关系描述

这几年,针对该share task已经有一些生物事件抽取方法被提出,其中很重要的一种方法是串行分步法[2]。该方法

到稿日期:2014-05-03 返修日期:2014-08-01 本文受国家自然科学基金(61202304,61173095,61202193),国家哲学社会科学重大计划招标项目(11&ZD189)资助。

魏小梅(1973-),女,硕士,副教授,主要研究方向为自然语言处理、生物信息,E-mail:may@mail.hzau.edu.cn;黄 钰(1975-),男,博士,讲师,主要研究方向为生物信息;陈 波(1976-),女,博士,主要研究方向为自然语言处理;姬东鸿(1967-),教授,主要研究方向为自然语言处理(通信作者)。

的基本策略就是先识别出事件的触发词(图 1 中 T3 和 T4)，然后确定触发词的论元(事件 E1 中的触发词 T3 所对应的论元是 E4，事件 E2 的触发词 T3 对应的论元是 E3，而事件 E3 和 E4 所对应的触发词是 T4，对应论元则分别是 T1 和 T2)。从图 1 中可以看到，论元可以是标注蛋白质，也可以是已抽取的其它事件。很显然，在串行事件抽取方法中，触发词的识别性能直接影响到下一步的事件关系抽取过程的性能，所以触发词的准确识别对整体性能的提高有至关重要的作用。

2 相关工作及系统框架

从相关的文献来看，触发词的识别方法主要分为 3 种：基于规则、基于词典匹配和基于机器学习方法。基于规则的方法应用一套泛化的语言规则来抽取触发词，例如，文献[3]中用词干处理后的规则来判断触发词，这种方法的性能非常依赖于规则的覆盖率，规则没有覆盖到的，触发词识别不到。基于词典匹配的方法，通过收集训练数据中所有作过触发词的词，将它们构成一个词典。需要识别触发词时，将文本中的词语与词典中的词语匹配。实际上，这种方法很不可靠。研究中发现，在生物事件中充当触发词的这一类词尽管会反复作为事件触发词在不同的句子中出现，但更多的时候会以普通词的形式出现。文献[4]指出，“activates”这个词，在语料中只有 28% 的出现是触发词，还有 72% 的出现是普通词；而“overexpression”这个词，则以 Gene_expression，Positive_regulation和 Negative_regulation 等多种类型的触发词出现。所以，单凭统计和词典方法，会引入大量的错误的触发词实例。而机器学习方法则会应用大量的特征和数据，建立统计模型，实现对样本的确定。

通用的机器学习模型有支持向量及模型(SVMs)、最大熵模型(MEMs)和隐马尔可夫模型(HMMs)等。在所有的机器学习模型中，CRF(条件随机域)模型通常被用来作序列标注[5]。由于 CRF 在序列标注中的卓越表现，我们决定采用 CRF 作为机器学习模型来实现触发词的标注。文献[6]的实验结果显示，词聚类特征在序列标记中起着明显的积极作用。由于生物事件中的触发词重现的频率很高，因此根据训练语料构建了生物触发词词典，并将其作为特征应用到 CRF 模型中。因此，实际上我们采用了 CRF 结合词聚类以及词典统计的方法来识别生物文献中的触发词。整个识别系统的流程如图 3 所示。

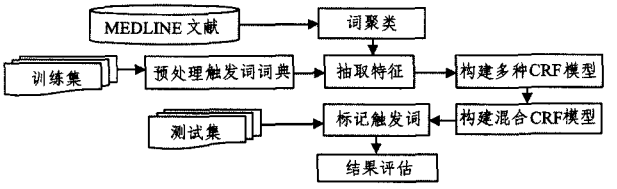

图 3 触发词识别系统框架

3 算法

3.1 CRFs 模型

条件随机域是一种无向图模型。研究表明，它特别适合于名称识别等序列标注问题。模型的算法如下：

设 $x=(x_1,x_2,\cdots,x_n)$ 表示输入的观察数据序列，$y=(y_1,y_2,\cdots,y_n)$ 表示对应的状态序列，那么对于给定的观察输入序列 X，CRF 模型定义的状态序列 Y 的联合条件概率为：

$$p(y|x,\lambda)=\frac{1}{Z(x)}\exp\left(\sum_j \lambda_j F_j(y,x)\right) \quad (1)$$

其中，λ_j 是需要从训练数据中学习的参数，用来表示相关特征的权重；$Z(x)$ 是只依赖于观测序列的归一化函数；$F_j(y,x)$ 的计算方法如下：

$$F_j(y,x)=\sum_{i=1}^{n} f_j(y_{i-1},y_i,x,i) \quad (2)$$

式中的每一个 $f_j(y_{i-1},y_i,x,i)$ 都是一个状态函数。

3.2 词聚类

研究[7,8]表明，在实体名称识别和序列标记问题中加入词聚类方法，对系统的性能都有明显的提高。Brown Cluster[9]就是一种普遍采用的词聚类方法，已经成功地应用于自然语言处理的各个领域[10]。它是一种层次聚类算法，它将词聚类以最大化 Bigram 的互信息，因此它是一种基于分类的 Bigram 语言模型，运行时间开销是 $O(V\cdot K^2)$，其中 V 是参加聚类的词汇个数，K 是聚类数。

本文聚类词汇来自于生物医药文献。从 PubMed[1] 下载了生物医药领域的 74000 篇摘要，将这些摘要分句、分词并用 Brown Cluster 算法进行词聚类，获得 1000 个词聚类类别，每个类别获得一个基于 Huffman 编码的二进制类别码。图 4 显示了部分聚类结果和它们的二进制聚类标记。

10011011100	looking	31
10011011100	contaminating	32
10011011100	worthy	33
10011011100	famous	33
10011011100	immunocytochemical	33
10011011100	plotted	34
10011011100	chemiluminescent	36
10011011100	competes	38
10011011100	neutralize	38
10011011100	enters	38
10011011100	Obese	38
10011011100	myelodysplastic	38
10011011100	decabromodiphenyl	40

图 4 聚类结果示例

3.3 词典构建

同时，在研究中发现，有些词作为生物事件触发词的频率很高，例如“expression”、“regulation”、“induce”等，这些词反复作为生物事件触发词出现。根据标注的训练数据，用统计的方法构建了一个生物事件的触发词词典，词典中包含作为触发词出现过的那些词的词干和它们出现在各种事件类型中的频率。将其作为特征引入 CRF 模型。

4 触发词识别

4.1 预处理

4.1.1 实体名称处理

由于生物文本的特殊性，其中的生物实体名称比较多，而且通常这些实体名称是多词构成的。同时，已经标注的蛋白质名称也经常是多词的，在语义上，这些多词的实体名称表达的是一个概念。因此，在预处理中对蛋白质名称做了替换处理。所有标注的蛋白质用名称“PRO”替换，其它的关键词实

1) www.ncbi.nlm.nih.gov/pubmed

体在用 Gdep[1] 句法分析器解析时会被识别出来。这样的操作实际上是生物文本中概念的一种泛化。经过这样处理后的图 1 中的句子结果如下：

例 1：(PMC-1313226) Leukotriene B4 stimulates PRO and PRO gene transcription

4.1.2 复合词"-"分隔线处理

触发词在事件结构中通常是独立于 PRO 单独成词，如例 1 所示。但是很多时候，它们会与 PRO 构成一个词，成为一个合成词，如"PRO-induced"中的"induced"是一个事件触发词。但是，英文的分词是以空格为单位，该类合成词经过分词处理后是一个整体，基于词级的 CRF 模型不能标出包含在合成词中的触发词，因此，预处理中必须将这些词以空格分开。前面提到的合成词"PRO-induced"在文中被处理成"PRO"，"-"，"induced" 3 个独立的词。

4.1.3 句法解析

经过上面的预处理步骤，还需要对文本进行分句、分词、词性标记等预处理操作。经过这些预处理操作后，可以抽取相应的词法、句法特征和语义特征。这里采用 Gdep 来处理文本，例 1 中的句子经过解析后的结果如图 5 所示。

1	Leukotriene	Leukotriene	B-NP	NN	O	2	NMOD
2	B4	B4	I-NP	NN	O	3	SUB
3	stimulates	stimulate	B-VP	VBZ	O	0	ROOT
4	PRO	PRO	B-NP	NN	B-protein	8	NMOD
5	and	and	I-NP	CC	O	8	NMOD
6	PRO	PRO	I-NP	NN	O	8	NMOD
7	gene	gene	I-NP	NN	O	8	NMOD
8	transcription	transcription	I-NP	NN	O	3	OBJ
9	.	.	O	.	O	3	P

图 5 例 1 的句子的解析结果

4.2 特征选择

本事件触发词标记系统采用的是基于词级的 CRF 标记模型。在模型训练中，采用的标记符号根据 9 种事件的类型标记{Gene_expression, Transcription, Protein_catabolism, Phosphorylation, Localization, Binding, Regulation, Positive_regulation, Negative_regulation}，加上"B-"前缀和"I-"前缀，并增加一种标记"N"，构成 CRF 模型中的标记集。其中"B-"加上类型标记，标识一个类型触发词的开始；"I-"加上类型标记，则表示一个类型触发词的延续，"I-"前缀是为了标记一个触发词由多个单词组成的情况；"N"标记则是标识非触发词。

特征的选择影响到分类的性能。在选择特征时，除了考虑到触发词本身的词法特征，包括单词、词干、词性、短语标记和前后缀等，还加入了前述的单词的词聚类特征和词典特征。由于在自然语言处理中，词的上下文特征对于确定其语义特征非常重要，因此在模型中对不同词法特征做了不同的窗口设置，例如词干特征窗口设置为{-2,2}，这样通过特征函数生成的特征就会包括每个词前后两词的特征。除此以外，还在特征模板中设置了它们的交叉特征。

4.3 后处理

触发词标记完成后，遍历所有的句子，将不含有"PRO"实体的句子删除，因为没有实体的句子不可能有事件，也就不可能有事件触发词。通过这种方式，可以去掉一些明显错标的事件触发词。

5 实验和讨论

5.1 实验语料和评价方法

采用 BioNLP 2009 GE share task 的训练集作为训练数据，构建机器学习模型。训练集有 BioNLP 2009 组织方提供的 800 篇摘要。采用了 CRF++工具[2]，引入了 Brown Cluster[3] 方法生成的词聚类特征，以及触发词的统计词典，实现了生物事件中 9 类触发词的识别。在开发集上进行了触发词的识别和评估，开发集包含 150 篇摘要。评估方法采用了广泛使用的精度(Precision)、召回(Recall)和 F 值(F-score)等评价指标，对 9 类事件进行分别评价和总体评价。

这些评价指标的具体计算方式如下：

$$P=\frac{TP}{TP+FP}, R=\frac{TP}{TP+FN}, F=\frac{2P\times R}{P+R}$$

其中，TP 是系统识别为正的正样本的数目，FP 表示系统识别为正的负样本的数目，FN 是指被系统预测为负的正样本数。因此，P 指示的是系统识别出来的触发词中真正的触发词的比率；R 则是所有的真正的触发词被系统识别出来的比率；F 则是前两个值的融合。

5.2 实验结果和评估

进行了多组实验，根据实验性能选择了一些对实验结果有积极影响的特征进行比较实验，选取的特征描述如下。

Lemma：词干，句子中每个词的词干；

Phrase：短语标记，句子中每个词的短语标记，如 NP，VP，ADJP 等；

Suffix：触发词的后缀，取每个词的后面 3 个字母作后缀；

Dict：词典特征，按是否出现在词典中，取值为 Y 或者 N；

Cluster：聚类特征，每个词标上的 Brown 聚类标记。

首先以词干特征为基础，分别加入词典、短语标记、聚类，实验结果如表 1 所列。从实验结果看出，加入词典后的总体性能优于其它特征，而且在召回上也稍好于其它特征，精度却是加入短语特征后表现最好。从实验结果也可以看出，每个特征对各种类型触发词的识别贡献不同。比如，从总体性能上看，后缀特征的表现最差，但是在 Protein_catabolism，Phosphorylation，Binding，Negative_regulation 这 4 类事件的抽取性能上表现却是最好的。因此，如果对不同类型的事件采用不同的特征，将会得到整体性能上的提升。

对于这些选择的特征，进一步对它们的组合特征做了实验。由于词典特征和短语特征在前面实验中的表现良好，以它们的组合为基础，分别加入其它特征，实验结果如表 2 所列。表 2 中实验结果显示，它们的特征组合取得了更好的效果，F 值达到了 60.2%，多个事件也达到了较好的性能。这个结果说明，选择好的特征，它们的组合特征和交叉特征能够提高性能。但是，观察整体情况仍然可以发现，不同的事件对不同特征的敏感程度不同。因此，根据前面多次的实验结果，针对不同的事件触发词的类型，采用不同的特征，构造了一个混合 CRF 标记模型(hybrid-CRF model)，以获得最好的整体性能。根据它们最好表现的统计结果，获得的整体性能如表 3 所列。

---

[1] http://people.ict.usc.edu/~sagae/parser/gdep/

[2] http://sourceforge.net/projects/crfpp/files/

[3] https://github.com/percyliang/brown-cluster

表 1　Lemma 加上表中特征后的触发词抽取性能

触发词类型	+dict			+phrase			+cluster			+suffix		
	P	R	F	P	R	F	P	R	F	P	R	F
G_exp.	82.6	71.7	76.8	85.8	70.9	**77.7**	85.3	70.2	77	80.7	67.9	73.8
Trans.	72.2	59.1	**65**	75	50	60	77.1	56.1	64.9	62.1	54.5	58.1
Prot.	100	84.2	**91.4**	100	73.7	84.8	100	73.7	84.8	100	84.2	**91.4**
Phos.	89.3	65.8	75.8	88.9	63.2	73.8	92	60.5	73	92.6	65.8	**76.9**
Loca	75.9	55	63.8	90.9	50	**64.5**	72.4	52.5	60.9	67.7	52.5	59.2
Bind.	81	46	58.7	81.9	43.8	57	82.1	44.3	57.6	77.7	49.4	**60.4**
Reg.	54.1	24.3	**33.5**	59.6	20.6	30.6	52	19.1	28	46.3	22.8	30.5
Pos.	67.6	45.1	**54.1**	71.9	42.3	53.3	71.7	43	53.8	64.1	44.9	52.8
Neg.	63.7	40.3	49.4	63.7	40.3	49.4	69.4	34.7	46.3	61.9	41.7	**49.8**
Total	73.4	**50.1**	**59.5**	**76.9**	47.5	58.7	76.6	47.2	58.4	69.6	49.4	57.8

表 2　组合特征对触发词识别的性能

触发词类型	Dict+Phrase			Dict+Phrase+Cluster			Dcit+Phrase+Cluster+Suffix		
	P	R	F	P	R	F	P	R	F
G_exp.	83.9	70.9	76.9	83.6	69.4	75.9	83.6	71.3	**77**
Trans.	80	54.5	64.9	78.7	56.1	**65.5**	77.8	53	63.1
Prot.	100	73.7	84.8	100	73.7	84.8	100	84.2	**91.4**
Phos.	89.3	65.8	75.8	89.7	68.4	**77.6**	88.9	63.2	73.8
Loc.	91.3	52.5	66.7	71.4	50	58.8	**95.7**	55	69.8
Bind.	84	44.9	**58.5**	83.2	44.9	58.3	81.9	43.8	57
Reg.	55.1	19.9	29.2	52.7	21.3	30.4	55.8	21.3	**30.9**
Pos.	71.2	44.2	54.5	71.8	44.4	54.9	71.5	45.6	**55.7**
Neg.	64.1	41	50	66.7	37.5	48	64	44.4	**52.5**
Total	76.4	48.6	59.4	75.9	48.2	59	76.5	49.6	**60.2**

表 3　混合 CRF 模型的触发词识别性能

触发词类型	P	R	F
G_exp.	85.8	70.9	77.7
Trans.	78.7	56.1	65.5
Prot.	100	84.2	91.4
Phos.	89.7	68.4	77.6
Loc.	95.7	55	69.8
Bind.	79.3	50	61.3
Reg.	46.8	26.5	33.8
Pos.	71.5	45.6	55.7
Neg.	64	44.4	52.5
Total	75.1	51.2	60.9

最终的混合 CRF 模型 F 值为 60.9%。将该结果与 Zhang[11] 和 Miwa[12] 的实验结果进行了比较，发现我们的实验结果在整体性能上有一定程度的优势，在精度上有明显的优势，如表 4 所列。精度的提升主要源于选择了合适的特征组合。

表 4　系统性能比较

系统	精度(P)	召回(R)	F 值
Miwa[12]	70.2	52.6	60.1
Zhang[11]	65.0	30.2	41.2
Ours	75.1	51.2	60.9

5.3　讨论

从实验结果来看，选择合适的特征，用 CRFs 模型对词序列标记可以获得较好的性能，词聚类特征有利于提高序列标记系统的精度，从而提高整个触发词识别系统的性能。但是从实验数据也明显可以看出，整个识别系统的召回率偏低，说明总体性能还有提升的空间。今后可以从以下几方面进行改进：1)可以对标记序列进行更充分的预处理，实现序列的语义更大程度的泛化，有利于提高模型的使用效率；2)可以在模型上进行改进，比如采用半监督的 CRFs 模型，目前它在相关领域取得了比较好的效果；3)我们的标记模型是基于词级的，事实上，触发词本身的词构成特征也有助于提高触发词的识别准确率。例如“express”和“coexpress”这两个词通常都作为 Gene_expression 类型的事件，它们的后缀相同，因此模型可以构建在更小粒度的单元上；4)由于识别触发词的实验采用的语料来自于 BioNLP 2009 share task，这些语料是用关键词“human”，“blood”和“cancer”从 PumMed 下载的 MEDLINE 摘要，因此认为与该语料库相关度越大的文本，其词聚类结果越能反映这些语料中的词汇特征。为了比较聚类文本的领域因素对触发词识别的差异，可以用与标注语料领域相同的文献进行词聚类，并将之应用于标记模型。

结束语　本文采用了基于 CRF 模型结合 Brown Cluster 方法以及统计词典的方法，实现了对生物事件触发词的识别。Brown Cluster 采用大量生物文献，生成 1000 个类别标记，实现了对生物文本中的词汇特征的泛化。词典特征的加入，提高了模型的整体性能。最后，根据不同特征对不同类型的事件触发词的影响差异，对不同类型的事件触发词选择不同的特征，构造了一个混合的 CRF 模型，与 Baseline 相比，所提识别系统显示了一定的优势。生物事件的触发词的识别是生物事件抽取的第一步，其性能的提高必将对生物事件的抽取产生积极的影响。

参 考 文 献

[1] Kim Jin-dong, Ohta T, Pyysalo S, et al. Overview of BioNLP'09 Shared Task on Event Extraction[C]// BioNLP Shared Task 2009 Workshop, 2009. Boston, MA, USA, 2009: 1-9

[2] Miwa M, Saetre R, Jin-dong K, et al. Event extraction with complex event classification using rich features[J]. Journal of bioinformatics and computational biology, 2010, 8(1): 131-146

[3] Casillas A, de Ilarraza A D, Gojenola K, et al. Using Kybots for extracting events in biomedical texts[C]// BioNLP Shared Task 2011 Workshop, 2011. Portland, Oregon, USA, 2011: 138-142

[4] Bjǒrne J, et al. Extracting Complex Biological Events with Rich Graph-Based Feature Sets[C]// Proceedings of the Workshop on BioNLP: Shared Task, 2009. Boulder, Colorado, 2009: 10-18

[5] MacKinlay A, Martinez D, Baldwin T. Biomedical event annotation with CRFs and precision grammars[C]// Workshop on Current Trends in Biomedical Natural Language Processing, 2009. Boulder, Colorado, 2009: 77- 85

[6] Lu Ya-nan, Yao Xiao-yuan, Wei Xiao-mei, et al. CHEMDNER System with Mixed Conditional Random Fields and Multi-scale Word Clustering[J]. Journal of Cheminformatics, 2015, 7(Suppl 1): S4

[7] Miller S, Guinness J, Zamanian A. Name Tagging with Word Clusters and Discriminative Training[C]// HLT-NAACL. 2004: 337-342

[8] Turian J, Ratinov L, Bengio Y. Word representations: a simple and general method for semi-supervised learning[C]// Procee-

dings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010. 2010:384-394

[9] Brown P F, deSouza P V, Mercer R L, et al. Class-based n-gram models of natural language[J]. Computational Linguistics, 1992, 18(4):467-479

[10] 刘远超,王晓龙,徐志明,等. 文档聚类综述[J]. 中文信息学报, 2006,20(3):55-62
Liu Yuan-chao, Wang Xiao-long, Xu Zhi-ming, et al. A survey of document clustering[J]. Journal of Chinese Information Processing, 2006, 20(3):55-62

[11] Zhang Y, Lin H, Yang Z, et al. Biomolecular event trigger detection using neighborhood hash features[J]. Journal of Theoretical Biology, 2013, 318(2):22-28

[12] Miwa M, Sætre R, Kim J-D, et al. Event extraction with complex event classification using rich features[J]. Journal of Bioinformatics & Computational Biology, 2010, 8(1):131-146

(上接第 210 页)

再交给 Reduce 规约为 1000。而基于 MapReduce 和 Bigtable 技术的 BM-Apriori 并行算法是将时间戳相同的数据项检索出来生成事物列项,扫描一次数据库即可完成候选项集与事务模式匹配,同时自动计算生成事务项对应的支持度,显著提升了运行效率。

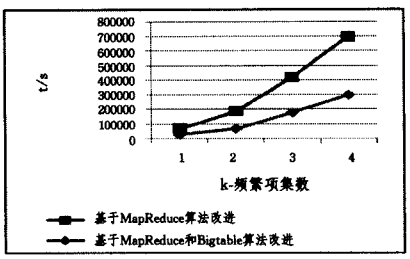

图 1 K-项频繁项集的时间消耗

图 2 为基于 MapReduce 和 Bigtable 技术的 BM-Apriori 算法和基于 MapReduce 技术的 Apriori 改进算法在不同数目节点上运行的结果。从中可以看出,节点数目越多,计算所需时间越短。这是因为,Bigtable 会根据行键自动划分为片 Tablet,主服务器具有自动均衡节点功能,可以高效有序地使所划分的片分布到不同的服务器节点上,并行执行算法,提高运行效率。但是随着节点数增加,运行时间也会增长,最终会使系统性能达到瓶颈。当节点数达到 6 时,再增加节点数对减少运行时间已没有多大帮助。从图中可以看出,在相同节点的情况下,基于 MapReduce 和 Bigtable 技术的 BM-Apriori 算法比基于MapReduce 技术的 Apriori 改进算法运行时间短,体现了改进算法的优越性。

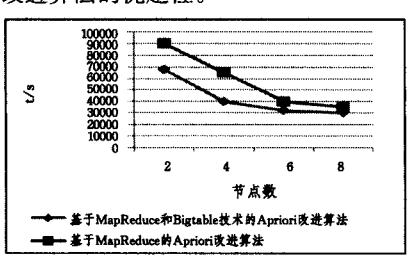

图 2 2 种算法在不同节点下的运行时间

结束语 本文对基于 MapReduce 的 Apriori 改进算法进行了研究整合,在此基础上首创引入 Bigtable 技术实现的基于 MapReduce 与 Bigtable 技术并行的 BM-Apriori 算法,并在 Hadoop 集群环境进行了实验与测试。实验证明,该算法只需扫描一次数据库,即可完成候选项集的模式匹配,并避免了大量的⟨itemset,1⟩键/值对的产生,并行分布技术又进一步提升了 Apriori 算法的运行效率。逐个击破了 Apriori 算法时间消耗的症节点。在数据爆炸式增长的时代,BM-Apriori 算法将分布资源的计算能力整合,对大数据具有高效的处理挖掘功能,这为进一步研究 Apriori 数据挖掘算法的高效执行提供了有力的支持,具有良好的研究价值和应用前景。

参 考 文 献

[1] Hajian S, Domingo-Ferrer J. A methodology for direct and indirect discrimination prevention in data mining [J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(7):1445-1459

[2] Lara J, Lizcano D, Martinez A, et al. A UML profile for the conceptual modeling of structurally complex data: Easing human effort in the KDD process [J]. Information and Software Technology, 2014, 56(3):335-351

[3] Agrawal R, Imielinski T, Swami A. Database mining: a performance perspective [J]. IEEE Transactions on Knowledge and Data Engineering, 1993, 5(6):914-925

[4] 张震,汪斌强,陈庶樵,等. 基于多维计数型布鲁姆过滤器的大流检测机制[J]. 电子与信息学报, 2010, 32(7):1608-1613
Zhang Zhen, Wang Bin-qiang, Chen Shu-qiao, et al. A Mechanism of Identifying Heavy Hitters Based on Multi-dimensional Counting Bloom Filter[J]. Journal of Electronics & Information Technology, 2010, 32(7):1608-1613

[5] Wang B L, Shen Y G. Improvement of Apriori algorithm based on boolean matrix [J]. Advanced Materials Research, 2011, 159:144-148

[6] 罗丹,李陶深. 一种基于压缩矩阵的 Apriori 算法改进研究[J]. 计算机科学, 2013, 40(12):75-78
Luo Dan, Li Tao-shen. Research on improved Apriori algorithm based on matrix compression [J]. Computer Science, 2013, 40(12):75-78

[7] 李晓虹,尚晋. 一种改进的新 Apriori 算法[J]. 计算机科学, 2007, 32(4):196-197
Li Xiao-hong, Shang Jin. An improved Apriori algorithm[J]. Computer Science, 2007, 32(4):196-197

[8] Grudzinski P, Wojciechowski M. Integration of candidate hash trees in concurrent processing of frequent itemset queries using Apriori[J]. Control and Cybernetics, 2009, 38(1):47-65

[9] Jongwook W. Market Basket Analysis algorithms with MapReduce [J]. Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery, 2013, 3(6):445-452

[10] Karim R, Hossain A, Rashid M, et al. A MapReduce Framework for Mining Maximal Contiguous Frequent Patterns in Large DNA Sequence Datasets [J]. IETE Techical Review, 2012, 29(2):162-168

[11] Chang F, Dean J, Ghemawat S, et al. Bigtable: A distributed storage system for structured data [J]. ACM Transactions on Computer Systems, 2008, 46(2):205-218

[12] Kim W. Web data stores (aka NoSQL databases): a data model and data management perspective [J]. International Journal of Web and Grid Services, 2014, 10(1):100-110