基于错误驱动的翻译模板自动获取

张春祥[1] 梁颖红[2] 于林森[3]

(哈尔滨理工大学软件学院 哈尔滨 150080)[1] (苏州市职业大学计算机系 苏州 215104)[2]

(哈尔滨理工大学计算机科学与技术学院 哈尔滨 150080)[3]

摘 要 翻译模板自动获取是提高MT系统译文输出质量和领域快速移植能力的关键性因素。利用Tree-to-String方法来抽取等价对,使用错误驱动的学习方法来获取翻译模板。将获取的模板用于MTS2005中,同时对其译文质量进行开放测试。实验结果表明:所提出的模板获取方法的性能要好于传统方法,当新模板加入系统原模板库后,开放测试语料的3元Nist评测分数提高了3.41%。

关键词 翻译模板,错误驱动,Nist评测

中图法分类号 TP391.2 **文献标识码** A

Automatic Acquisition of Translation Templates Based on Error-driven

ZHANG Chun-xiang[1] LIANG Ying-hong[2] YU Lin-sen[3]

(School of Software, Harbin University of Science and Technology, Harbin 150080, China)[1]

(School of Computer Engineering, Suzhou Vocational University, Suzhou 215104, China)[2]

(College of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China)[3]

Abstract Automatic acquisition of translation templates is important for MT system to improve its translation quality and its ability of adapting to new domain. In this paper, Tree-to-String method was applied to extract translation equivalents. Error-driven learning method was used to acquire templates. These templates were applied to MTS2005, and open test experiments were conducted for its translation quality. The experiment results show that the performance of new method in this paper is better than that of old method. Combination of new acquired templates and original ones makes assessment score of open test corpus improved by 3.41% under 3-gram Nist assessment metric.

Keywords Translation templates, Error-driven, Nist assessment

1 引言

翻译模板是一种能够实现译文选择与调序的翻译知识,在基于句法分析的EBMT系统[1]和统计机器翻译系统[2]中有着广泛的应用。因而,翻译模板自动获取一直是提高系统译文输出质量和领域快速移植能力的关键性因素。目前,获取翻译模板的主要方法是:对双语句对中的源语言与目标语句子分别进行句法分析,依据词对齐结果实现两棵句法树内部结点的对齐,即分析-分析-匹配;将两棵句法树中对齐的子树视为翻译等价对,并以此为基础获取翻译模板[3]。屈刚以"有效"句型和"翻译中相对不变准则"为基础解决了两棵句法树结构对齐过程中的歧义问题[4]。但分析-分析-匹配方法在对齐过程中要受到双语语法体系一致性的制约,因而很难将两棵句法树完全实现结构对齐。同时对齐过程会受到目标语句法分析精度的影响,因而在获取的翻译模板中存在着大量的噪声。

Fai Wong提出了源语言-目标语译文等价子树对标注体系[5]。该体系仅对源语言进行句法分析,以双语句对词的对齐结果为依据确定源语言句法树中每个子树在目标语句子中对译片断的边界,从而记录了每个源语言结构短语对应的目标语译文片断,即Tree-to-String方法。尽管目标语部分不完全合乎语法要求,但其等价对的提取过程不受目标语句法分析精度与双语语法体系一致性的制约,因而表现出了较好的性能。本文采用Tree-to-String方法从汉、英双语语料中抽取等价对,通过查汉-英机器翻译词典实现等价对中英语部分的译文动作映射,并依据汉语短语的句法、词性、词法及词特征获得学习实例。利用错误驱动的机器学习方法从学习实例中提取翻译模板,并将其应用于MTS2005中。实验结果表明:本文所提出的模板获取方法的性能要好于传统方法,将新获取的模板加入系统原模板库后,3元Nist评测分数提高了3.41%。

2 翻译模板自动获取

2.1 汉、英翻译等价对提取

通常,翻译等价对的获取主要有两种方式:分析-分析-匹配方法和Tree-to-String方法。

2.1.1 汉、英句法分析树结构对齐

分析-分析-匹配方法依据双语句对词的对齐来结果实现汉、英句法树内部结点的对齐。汉、英句法树结构对齐的过程

到稿日期:2009-08-18 返修日期:2009-10-26 本文受国家自然科学基金(60903082),黑龙江省教育厅科学技术研究项目基金(11541045),哈尔滨理工大学青年科学研究基金(2008XQJZ017),江苏省现代企业信息化应用支撑软件工程技术研究开发中心项目(SX200907)资助。

张春祥(1974—),男,博士生,副教授,主要研究方向为自然语言处理、机器翻译及机器学习,E-mail:z6c6x6@yahoo.com.cn。

如下：

(1)分别遍历汉、英句法树提取除树根外的所有非叶结点，获取汉语句法子树森林$\{ST_1^c, ST_2^c, \cdots, ST_m^c\}$和英语句法子树森林$\{ST_1^e, ST_2^e, \cdots, ST_n^e\}$；

(2)对汉语句法子树ST_i^c，使用式(1)确定与其对齐的英语句法子树ST_j^e；

$$ST_j^e = \max_{u \in \{ST_1^e, ST_2^e, \cdots, ST_n^e\}} \frac{Link(ST_i^c, u)}{Num(ST_i^c) + Num(u)} \quad (1)$$

式中，$Link(ST_i^c, u)$为汉语句法子树ST_i^c与英语句法子树u之间的词对齐数目，$Num(x)$为子树x包含的单词个数。

汉、英句法树结构对齐过程如图1所示，对齐的内部结点为VO(1)-VP(1′)，NP(3)-NP(3′)和VO(4)-PP(4′)，将对齐的内部结点视为翻译等价对。

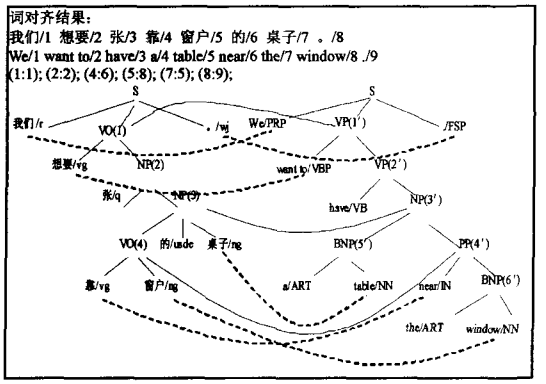

图1 汉、英句法树结构对齐

2.1.2 基于Tree-to-String的汉、英翻译等价对获取

Tree-to-String方法以双语句对词的对齐结果为依据，建立汉语句法树与英语句子之间的对应关系。

对齐区间：对于汉、英双语句对$(c_1, c_2, \cdots, c_m \langle - \rangle e_1, e_2, \cdots, e_n)$，$T$为汉语句子$c_1, c_2, \cdots, c_m$对应的句法树。若$ST$为$T$的叶结点(即汉语词)，且$ST$与英语句子中位置$j$的英语单词对齐(有多个对齐仅取第一个)，则$ST$在英语句子中的对齐区间记为$[j,j]$；若$ST$为非叶结点，且$ST$的孩子结点为$ST_{i1}^{[j1,k1]}, ST_{i2}^{[j2,k2]}, \cdots, ST_{im}^{[jm,km]}$，则$ST$在英语句子中的对齐区间为$[j,k]$($j=\min(j_1, j_2, \cdots, j_m)$，$k=\max(k_1, k_2, \cdots, k_m)$)。后序遍历汉语句法树$T$，获得每个结点$ST$的英语对齐区间$[j,k]$，因而$T$中的结点表示为$ST^{[j,k]}$。

利用Tree-to-String方法对齐汉语句法树与英语句子，前序遍历汉语句法树获得所有翻译等价对，如图2所示。尽管目标语不完全合乎语法要求，但整个对齐过程不受目标语句法分析精度与双语语法体系一致性的制约，因而可以覆盖更多的语言现象。

本文使用Tree-to-String方法来提取汉、英翻译等价对。翻译等价对左部为包含句法信息的汉语短语，右部是英语译文串。汉语短语可以分为嵌套短语和非嵌套短语两种形式。因而，所提取的翻译等价对也包括非嵌套和嵌套两种形式。

非嵌套翻译等价对：$Phrtype(W_0/pos_0 + W_1/pos_1 + \cdots + W_m/pos_m) -> e_1, e_2, \cdots, e_n$，其中$e_1, e_2, \cdots, e_n$为汉语短语对应的英语译文。

嵌套翻译等价对：$Phrtype(PHRASE_0, PHRASE_1, \cdots, PHRASE_m) -> e_1, e_2, \cdots, e_n$，其中$e_1, e_2, \cdots, e_n$为汉语短语对应的英语译文。

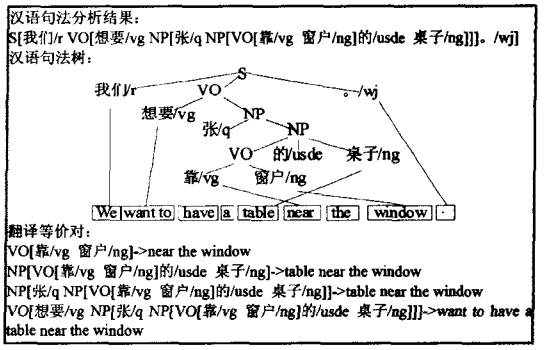

图2 使用Tree-to-String方法抽取翻译等价对

2.2 学习实例获取及翻译模板学习

翻译模板是变量化的翻译表达式，左部是条件，右部是翻译动作[6]。模板类型与汉语句法树中短语句法标注完全一致。在翻译过程中，每一类模板仅作用于汉语句法树中具有相同类型的短语(例如：VO类型的模板只用于实现VO类型短语的翻译)。只有模板条件被匹配时，才会执行其翻译动作。在进行模板条件匹配时，要求模板左部各结点的变量值完全被匹配。变量可以是句法特征$Phrtype$、词性特征pos、词法特征$Head$和词特征W。等价对左部的汉语短语用作学习翻译模板的条件，因而在生成学习实例时，应获取翻译等价对左部汉语短语的句法、词性、词法和词特征。

对于非嵌套翻译等价对$Phrtype(W_0/pos_0, W_1/pos_1, \cdots, W_m/pos_m) -> e_1, e_2, \cdots, e_n$，通过查汉-英机器翻译词典提取汉语短语的各词单元$W_i$的词法特征$Head$(例如：在“窗户”的词典词条中，其词法特征为$Head=Object$)。对汉语短语$Phrtype(W_0/pos_0, W_1/pos_1, \cdots, W_m/pos_m)$的各结点进行编号，获取其序号、词性、词法及词特征四元组列表$\{\langle i, Cate=pos_i, Head=head_i, W=W_i\rangle, i=0,1,2,\cdots,m\}$。从汉语短语$VO$[靠/$vg$ 窗户/ng]中提取的特征四元组列表为$\{\langle 0, Cate=vg, Head=Svn, W=$靠$\rangle, \langle 1, Cate=ng, Head=Object, W=$窗户$\rangle\}$。

对于嵌套翻译等价对$Phrtype(PHRASE_0, PHRASE_1, \cdots, PHRASE_m) -> e_0, e_1, \cdots, e_n$，$PHRASE_i$可为非嵌套或嵌套短语。为了能获取更加抽象的翻译模板，提取短语$PHRASE_i$核心结点的词特征及该词的词法特征用于翻译模板的学习。确定$PHRASE_i$核心结点的具体过程为：

(1)建立$PHRASE_i$的句法树；

(2)后序遍历该句法树，每次遇到非叶结点时，将其右孩子设置为它的核心结点，每次遇到叶结点时，将其自身设置为核心结点；

(3)遍历结束时，树根中记录了$PHRASE_i$的核心结点。

对汉语短语$Phrtype(PHRASE_0, PHRASE_1, \cdots, PHRASE_m)$的各结点进行编号，获取其序号、句法、核心结点词法及核心结点词特征四元组列表$\{\langle i, Cate=Phrtype_i, Head=head_i, W=W_i\rangle, i=0,1,2,\cdots,m\}$。从$NP$[$VO$[靠/$vg$ 窗户/ng]的/$usde$ 桌子/ng]中提取的特征四元组列表为$\{\langle 0, Cate=VO, Head=Svn, W=$窗户$\rangle, \langle 1, Cate=usde, Head=NULL, W=$的$\rangle, \langle 2, Cate=ng, Head=Object, W=$桌子$\rangle\}$。

等价对中的英语译文用于学习模板的翻译动作。在将汉语句子译成英语译文的过程中，通常存在着3种情况：选译、

增译与省译。选译:即选择汉语词在词典词条中合适的译项作为译文;增译:英语译文中的某些词不能直接由汉语词翻译过来,需要在翻译时根据其上下文的词法、句法信息添加相应的英语单词;省译:某些汉语词在翻译过程中不必进行翻译。因此,模板的翻译动作包括:选译操作 $j:*$,第 j 个汉语单词的译文;插入译文操作 $I:e$,插入英语单词 e。

译文动作映射:通过查汉-英机器翻译词典确定等价对右部英语词中,哪些属于选译操作、哪些属于插入译文操作,从而确定翻译动作。

对于非嵌套翻译等价对 $Phrtype(W_0/pos_0, W_1/pos_1, \cdots, W_m/pos_m) -> e_1, e_2, \cdots, e_n$,其译文动作映射过程如图 3 所示。

1.初始化译文动作映射数组 A[n]=""。
2.对 $e_i \in (e_1, e_2, \ldots, e_n)$
a.若 e_i 在 W_j 的词典义项中出现,且 $W_j/pos_j \in (W_0/pos_0, W_1/pos_1, \ldots, W_m/pos_m)$,则英语词 e_i 的译文动作映射为 j:*,A[i]=j:*;
b.若 e_i 不能在 $(W_0/pos_0, W_1/pos_1, \ldots, W_m/pos_m)$ 中任何一个 W_j 的词典义项中找到,则该英语词 e_i 的译文动作映射为 I:e_i,A[i]=I:e_i;
3.对 $\forall$ i,j(A[i]=A[j]=ord:*,具有相同结点序号的选译操作)之间的插入译文操作 I:进行删除,并去掉连续重复的项,使用联结符'+'将 A 中所有非空项连接起来输出。

图 3 非嵌套翻译等价对的译文动作映射

$VO[靠/vg\ 窗户/ng] -> near\ the\ window$ 执行译文动作映射算法后,其右部译文动作映射结果为 0:* +I:the+1:*。

对于嵌套翻译等价对 $Phrtype(PHRASE_0, PHRASE_1, \cdots, PHRASE_m) -> e_1, e_2, \cdots, e_n$,其翻译动作映射过程如图 4 所示。

1.初始化译文动作映射数组 A[n]=""。
2.对 $e_i \in (e_1, e_2, \ldots, e_n)$
a.若 e_i 在 $PHRASE_j$ 中某个词的词典义项中出现,则将 e_i 映射为 j:*,A[i]=j:*;
b.若 e_i 不能在任何一个 $PHRASE_j$ 中词的词典义项中找到,则将 e_i 映射为 I:e_i,A[i]=I:e_i;
3.对 $\forall$ i,j(A[i]=A[j]=ord:*,具有相同序号的选译操作)之间的插入译文操作 I:进行删除,并去掉连续重复的项,使用联结符'+'将 A 中所有非空项连接起来输出。

图 4 嵌套翻译等价对的译文动作映射

$NP[VO[靠/vg\ 窗户/ng]的/usde\ 桌子/ng] -> table\ near\ the\ window$ 执行译文动作映射算法后,其右部译文动作映射结果为 2:* +0:*。

将翻译等价对左部汉语短语的特征四元组列表作为源模式,右部译文动作映射结果作为目标模式构成学习实例。

学习实例 E:$E=SPat -> TPat$,其中 $SPat=\{\langle i, Cate=pos_i/Phrtype_i, Head=head_i, W=W_i\rangle, i=0,1,2,\cdots,m\}$ 为 E 的源模式;当左部汉语短语的第 i 个结点为词单元时,$Cate=pos_i$,当左部汉语短语的第 i 个结点为短语时,$Cate=Phrtype_i$;$TPat$ 为目标模式,例如 $TPat=0:*+2:*+\cdots+n:*$。

学习实例 E 的抽象源模式记为 $G(E)=0:Cate=pos_0/Phrtype_0+1:Cate=pos_1/Phrtype_1+\cdots+m:Cate=pos_m/Phrtype_m$,即各结点仅使用句法特征或词性特征。$E$ 对应的抽象翻译模板为 $G(E) -> TPat$。

学习实例与翻译等价对一一对应,将获取的学习实例,按左部汉语短语的句法标注进行归类共获得 31 类学习实例集。

本文使用错误驱动的学习方法从等价对中学习翻译模板,其算法的核心思想是:首先获取抽象翻译模板,抽象翻译模板的条件部分仅使用句法特征或词性特征;若学习实例与抽象模板相匹配,但目标模式与抽象模板的翻译动作不同,即利用抽象模板对该等价对的汉语短语进行翻译时会产生错误,则使用词法特征或词特征对抽象模板的条件进行特化,同时使用学习实例的目标模式作为翻译动作以生成新的模板[6]。

3 实验

3.1 汉、英翻译等价对及翻译模板获取

本文从 30000 句来自旅游领域的汉、英双语句对中获取翻译模板。使用的词对齐工具和汉、英句法分析器是哈尔滨工业大学语言语音教育部-微软重点实验室开发的。词对齐工具使用了基于词典、基于语义相似度、基于统计和基于语言学知识相融合的策略[7]。句法分析器使用了词汇化的统计分析技术。其性能如表 1 所列。

表 1 词对齐与句法分析器性能分析

	精确率(Precision)	召回率(Recall)
词对齐工具	86%	89%
英语句法分析器	77%	80%
汉语句法分析器	78%	79%

分别使用分析-分析-匹配与 Tree-to-String 方法从双语句对中获取翻译等价对。通过查汉-英机器翻译词典对等价对右部英语单词进行译文动作映射,同时获得左部汉语短语的句法、词性、词法和词特征以生成学习实例。使用错误驱动方法从学习实例中获取翻译模板。在两种方法中,获得的翻译等价对及模板数目如表 2 所列。

表 2 两种方法中获得的翻译等价对及模板数目

	非嵌套等价对	嵌套等价对	翻译模板
分析-分析-匹配	30741	60781	10917
Tree-to-String	35321	74930	12361

3.2 翻译模板的性能评价

将自动获取的模板用于 MTS2005[8] 中,以检测其翻译性能。该系统原始模板库中共包含 5092 条翻译模板,这些模板是由 20 名语言工程师经过 2 年时间编写调试出来的。另外,搜集 1200 句对旅游领域的汉、英双语句对作为测试语料和标准答案,进行开放测试。为了比较本文所提方法的性能,共进行了 3 组对比实验。实验 1 在 MTS2005 中使用原始模板库对开放测试语料进行翻译;实验 2 使用分析-分析-匹配方法来获取的模板对开放测试语料进行翻译;实验 3 使用 Tree-to-String 方法获取的模板对开放测试语料进行翻译。利用 Nist[9] 和 Bleu[10] 方法对机器译文进行评测,其译文评测分数如表 3 所列。3 组实验中使用的模板数目如图 5 所示。

表 3 4 组实验中开放测试语料的译文评测分数

	Nist3	Bleu3
实验 1	4.0012	0.1539
实验 2	3.7898	0.1437
实验 3	3.8918	0.1486
实验 4	4.1377	0.1599

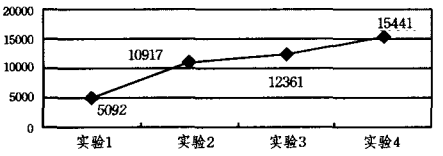

图 5 4 组实验中的翻译模板数目

从表 3 中可以发现实验 3 的译文评测分数超过了实验

2。这是因为:使用 Tree-to-String 方法获取翻译等价对时,由于不受目标语句法分析精度及双语语法体系一致性的制约,因此等价对中包含的噪声较小,进而从中获取的翻译模板质量较高。

从表 3 中可以发现,实验 3 的译文评测分数接近实验 1。这是由于模板自动获取要受到词对齐工具和汉语句法分析器精度的制约,因而自动获取的模板其质量还没有达到手工书写的。但从模板获取的时间和人工劳动强度而言是可取的,同时随着词对齐工具和汉语句法分析器性能的逐步提升,以及翻译等价对、翻译模板优化方法研究的深入,自动获取的模板质量会进一步提高。

从图 5 中可以发现,实验 3 使用的模板数目要远超过实验 1,这是由于利用机器学习方法还不能获取像语言工程师所编写的具有高度抽象概括的翻译模板。但从模板获取的代价和译文质量提升方面来讲,本文所提出的自动模板获取方法是可取的。

为了比较手工书写的模板与自动获取模板之间的一致性,在实验 4 中,将实验 3 所获取的 12361 条模板加入系统原始模板库中,去掉其中冲突和冗余的部分,并对开放测试语料进行翻译。其译文评测分数如表 3 所列,实验 4 使用的模板数目如图 5 所示。

实验 4 与实验 1 相比,开放测试语料的 3 元 Nist 评测分数提高了 0.1365,上升了 3.41%。这表明:利用本文的方法所获取的模板与手工书写模板之间具有较好的一致性。

结束语 本文使用 Tree-to-String 方法从汉、英双语句对中抽取翻译等价对,利用错误驱动的机器学习方法获取翻译模板。将自动获取的模板用于 MTS2005,并使用 1200 句旅游领域的测试语料进行开放测试。实验结果表明:本文所提的模板获取方法其性能要好于传统方法,当新获取的模板加入系统原始模板库后,表现出了较好的一致性。

参 考 文 献

[1] Imamura K, Okuma H. Example-based Machine Translation Based on Syntactic Transfer with Statistical Models[C]// The 20th International Conference on Computational Linguistics. 2004:99-105

[2] Zettlemoyer L S, Moore R C. Selective Phrase Pair Extraction for Improved Statistical Machine Translation[C]// Proceedings of NAACL HLT 2007. 2007:209-212

[3] Meyers A, Yangarber R. Deriving Transfer Rules from Dominance-Preserving Alignments[C]// The 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. 1998:843-847

[4] 屈刚,陈笑蓉,陆汝占. 基于有效句型的英汉双语短语对齐[J]. 计算机研究与发展,2003,40(2):143-149

[5] Wong Fai, Hu Dong-cheng. A Flexible Example Annotation Schema: Translation Corresponding Tree Representation[C]// The 20th International Conference on Computational Linguistics. 2004:1079-1085

[6] 张春祥. 基于短语评价的翻译知识自动获取研究[D]. 哈尔滨:哈尔滨工业大学,2007

[7] 吕雅娟,赵铁军,李生,等. 统计和词典方法相结合的双语语料库词对齐[C]// 第六届计算语言学联合学术会议. 2001:108-115

[8] Xue Yongzeng, Zhao Tiejun. Research On Sports News Oriented Skeleton Machine Translation[C]// The 7th International Conference for Young Computer Scientists. 2003:310-312

[9] Doddington G. Automatic evaluation of machine translation quality using n-gram cooccurrence statistics[C]// Proceedings of ARPA Workshop on Human Language Technology. 2002

[10] Papineni K, Roukos S. BLEU: a method for automatic evaluation of machine translation[C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002: 311-318

(上接第 178 页)

[12] Breiman L. Bagging predictor [J]. MachineLearning, 1996, 26 (1):5-24

[13] 谢志鹏,刘宗田. 概念格节点的内涵缩减及其计算[J]. 计算机工程,2001,27(3):9-10

[14] 齐红,刘大有,胡成全,等. 基于搜索空间划分的概念生成算法[J]. 软件学报,2005,16(12):2029-2035

[15] 李云,刘宗田,陈崚,等. 多概念格的横向合并算法[J]. 电子学报,2004,32(11):1849-1854

[16] 刘贵龙. 粗糙集的粗糙度[J]. 计算机科学,2004,31(3):140-141,153

[17] 王玲芝. 粗糙模糊集的贴近度[J]. 四川师范大学学报:自然科学版,2002,25(5):476-478

[18] 丁卫平,管致锦,石振国. 基于扩展粗糙集模型近似概念格规则挖掘研究[J]. 南京邮电大学学报:自然科学版,2009,29(2):10-15

[19] 王志海,胡可云,胡学钢,等. 概念格上规则提取的一般算法与渐进式算法[J]. 计算机学报,1999,22(11):66-70

[20] UCI 机器学习数据库. ftp://ftp.ics.uci.edu/pub/machine-learning-databases

[21] 丁卫平. 关联规则挖掘 Apriori 算法的改进及其应用研究[J]. 南通大学学报:自然科学版,2008,7(1):50-53

[22] 丁卫平,顾春华,石振国,等. 基于形式概念分析的不完备电子病历系统粗糙挖掘研究[J]. 计算机科学,2009,36(10):230-233