带有通配符和长度约束的模式匹配问题求解模型

汪 浩 王海平 吴信东
(合肥工业大学计算机与信息学院 合肥 230009)

摘 要 讨论了带有通配符和长度约束的模式匹配(PMWL)问题,其中模式由子模式序列集组成,两个相邻子模式的间隔在一定长度范围内。针对PMWL问题,已有工作包括设计启发式求解算法和对特殊情况进行完备性分析,然而还需要构建问题的基础求解模型。借鉴约束可满足问题框架,构建了由变量、值域和约束组成的三元组求解模型,对PMWL问题的基本概念和基本性质给出了形式化描述。最后,给出了算法求解PMWL问题的特定条件下的完备解。

关键词 长度约束,通配符,求解模型,模式匹配
中图法分类号 TP301 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.4.057

Models for Pattern Matching with Wildcards and Length Constraints

WANG Hao WANG Hai-ping WU Xin-dong
(School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China)

Abstract For the problem of pattern matching with wildcards and length constraints (PMWL), the patterns are composed of a sequence of sub-patterns, where any two adjacent sub-patterns with flexible gaps are in a specified range of the text. Existing work includes a heuristic strategy and its completeness analysis with constraints, but the PMWL problem still needs systematic studies. We drew on the experience of constraint satisfaction problems(CSPs) and set up a 3-tuple model consisting of variables, domains and constraints. We then derived formal descriptions for the basic concepts and properties. Also, a tree-based matching algorithm was presented to solve the PMWL problem under certain conditions.

Keywords Length constraints, Wildcards, Matcing model, Pattern matching

1 引言

在模式识别问题中引入通配符这种特殊字符以匹配字母表中的任意字符,从而形成了带有通配符的模式匹配这一新的研究方向[1],它在计算生物学[2]、信息检索[3]等研究领域得到了广泛关注。然而,上述领域的大量研究表明,频繁共现在一段文本区域内的多个模式之间表现为某种模式形式。比如在DNA序列中,TATA序列作为内含子的起始标志常出现在CAATCT序列下游30-50 bp的位置,这两个子序列共同组成的模式可提高序列特异性,可以标记为"…CAATCT[30,50]TATA…"。

上述情况可进一步推广为子模式序列集,其中两个相邻子模式的间隔在一定长度范围内,为表示这种灵活的位置间隔,将通配符从指代单个字符扩展为指代一定长度的子串,称此为长度约束。另外,通过引入one-off约束来保证子模式序列集的稳定性,避免了个别子模式异常的出现次数影响子模式集的匹配。由此得到了带有通配符和长度约束的模式匹配(PMWL)问题。

PMWL问题的求解目标是使得满足约束条件的解集最大。因此PMWL既是模式匹配问题,也是最优化问题,这使得问题求解需要平衡算法时间复杂度和解的质量[4-7]。在算法设计上,已有的算法大多采用启发式搜索策略,其中SAIL算法[8]使用滑动窗口技术,先根据可变跨度标记候选匹配位置,再反向扫描选择较小的候选位,从而尽可能将更多匹配资源留给之后的匹配;MOTW算法[9]在位并行算法的基础上扩展转移函数,并通过非确定性有限自动机模拟带有长度约束模式串的匹配过程,该策略降低了时间复杂度,其匹配结果与SAIL算法的相同。当PMWL问题满足某些限定条件时,王等人证明了在某特定模式特征下PMWL问题可解;Chen等人说明了使用left-most贪心策略能得到在线情况下的完备解,而在离线情况下不完备;在不考虑one-off约束下,PAIG算法[10]和SETS算法[11]能在多项式时间内得到带有PMWL问题的完备解,然而PAIG与SETS算法只能得到匹配数,而不返回匹配位置。

综上,已有的工作主要包括:1)设计启发式算法求近似解;2)寻找一种限定条件以求得完备解,然而还缺少描述PMWL问题的求解模型,这可作为PMWL问题的算法设计和理论分析的基础。因此,本文将借鉴约束可满足问题框架对PMWL问题建立求解模型。模型由变量、值域和约束条件构成三元组,并运用形式化语言描述了问题的约束条件和解

到稿日期:2015-10-12 返修日期:2016-01-24 本文受国家自然科学基金项目(31100956,61173117)资助。
汪 浩(1980—),男,博士生,主要研究方向为模式识别、机器学习;王海平(1986—),男,博士,主要研究方向为模式识别、数据挖掘,E-mail:halewall@163.com;吴信东(1963—),男,教授,博士生导师,主要研究方向为数据挖掘、多源海量信息处理。

空间等主要概念。基于此模型，本文说明了 PMWL 问题的基本性质，其中包括问题在 8 种特殊条件下的完备性；之后，采用 FIN 算法可以得到特定条件下的完备解。

2 PMWL 的约束可满足问题模型

2.1 PMWL 问题

带有通配符和长度约束的模式匹配(PMWL)问题定义如下[8]。

定义 1(PMWL 问题)

(1)输入为字母表 Σ，文本 T，模式 P 和可变跨度的约束区间，输出为 P 在 T 中所有出现位置的集合，即匹配解集；

(2)可变跨度：模式 P 相邻两个字符之间存在通配符 ϕ，它可以指代一个字符串 s，其中 $s\in\Sigma^*$，且长度受到区间约束；

(3)one-off 约束：文本 T 中的字符至多只能被一个匹配解使用。

例如，给定 $T=ACATTTTCTTCA$，$P=A\,\phi[0,2]T\,\phi[0,2]T\,\phi[0,1]C$，可得到匹配解集$\{\{1,4,6,8\},\{3,6,9,11\}\}$。

2.2 PMWL 问题的求解模型

约束可满足问题(Constraint Satisfaction Problems, CSPs)框架作为求解模型，可以揭示组合优化问题在约束条件下的解结构特征[12]，目前已应用于带有约束条件的模式匹配问题[13]。在已知的工作中，文献[4-8]对 PMWL 问题的基本概念给出了定义，然而上述描述大多采用自然语言。本节将借助 CSPs 框架，建立 PMWL 问题的求解模型，对问题的约束条件和解空间等基本概念用形式化语言作更精确的描述。

定义 2 给定字母表 Σ，文本 $T=t_1t_2\cdots t_n$ 和模式 $p=p_1\,\phi_1\,p_2\,\phi_2\cdots\phi_{m-1}\,p_m$，其中通配符 ϕ_i 可以指代字符子串 s_i，且满足 $s_i\in\Sigma^*\wedge|s_i|\in[l_i,u_i]$，可构造 PMWL 问题的求解模型：

$$PMWL=\{X,D,C\}$$

其中，X 为变量集，D 为作用域集，C 为约束集，且满足：

$$X=\{(\alpha_1,\alpha_2,\cdots,\alpha_m)\,|\,t[\alpha_i]=p_i,\alpha_i\in D,1\leqslant i\leqslant m\}$$

其中，$D=\{1,2,\cdots,n\}$，称$\{p_1,p_2,\cdots,p_m\}$为模式 P 的子模式集，变量 $\alpha_1,\alpha_2,\cdots,\alpha_m$ 为各子模式在文本 T 中的匹配位置，则 m 为子模式数目，称为模式 P 的模长。从而可得初始解空间

$$H=D^m=\{(\alpha_1,\alpha_2,\cdots,\alpha_m)\,|\,\alpha_i\in D,1\leqslant i\leqslant m\}$$

考虑约束条件 $C=\{LC,AC\}$，其中，GC(Length Constraints)表示限制单个变量的长度约束，AC(Association Constraints)表示变量之间的约束。

定义 3 在模式匹配问题中，通配符可用于改变模式 P 相邻字符的匹配位置，形成了限制单个变量的长度约束 LC。LC 的一般形式为：

$$LC=\{LC(X_i),i=1,2,\cdots,m\}$$

其中，

$$LC(X_i):\text{for } X_i=(\alpha_1^i,\alpha_2^i,\cdots,\alpha_m^i),\text{s.t. } f(\alpha_j^i,\alpha_{j+1}^i),1\leqslant j<m$$

$f(\alpha_j^i,\alpha_{j+1}^i)$可用于描述模式中字符 p_j 与字符 p_{j+1}在文本中匹配位置的函数关系，如位置相邻时满足 $f(\alpha_j^i,\alpha_{j+1}^i)=\alpha_{j+1}^i-\alpha_j^i=1$。在 PMWL 问题中，通配符$\phi_1,\phi_2,\cdots,\phi_{m-1}$存在于相邻子模式之间，其中$\phi_j$ 可指代字符子串 $s_j,s_j\in\Sigma^*$，且满足$|s_j|\in[l_j,u_j],1\leqslant j<m$，称$[l_j,u_j]$为长度约束区间，记为 LC_{gap}，可描述为：

$$LC_{gap}(X_i):X_i=(\alpha_1^i,\alpha_2^i,\cdots,\alpha_m^i),\text{s.t. } \alpha_{j+1}^i-\alpha_j^i-1\in[l_j,u_j],1\leqslant j<m$$

定义 4 在 PMWL 问题中，one-off 约束是指任意一个文本字符至多只能被一个匹配解使用，记为 AC_{oneoff}，可描述为：

$$AC_{oneoff}(X_i,X_j)=R(X_i,X_j):X_i=(\alpha_1^i,\alpha_2^i,\cdots,\alpha_m^i),X_j=(\alpha_1^j,\alpha_2^j,\cdots,\alpha_m^j),\text{s.t. }\alpha_k^i\neq\alpha_g^j,1\leqslant k,g\leqslant m,i\neq j$$

由定义 3 和定义 4 可知，长度约束是对单变量取值的约束，而 one-off 约束则是对多个变量关系的约束。在传统匹配问题中，不同匹配解的查找是独立的；但在 one-off 约束下，文本中的字符是有限资源，匹配解之间存在竞争关系。因而求解 PMWL 问题需要在搜索过程中考虑如何合理分配文本资源，这使得 PMWL 问题具有难解性，且目前仍为开放性问题[8]。

定义 5 在 PMWL 问题中，匹配时满足长度约束 LC_{gap} 的变量称为候选解，可得：

$$\Gamma=\{\Gamma_i\,|\,\Gamma_i=(\alpha_1^i,\alpha_2^i,\cdots,\alpha_m^i),\text{s.t. }LC_{gap}(X_i)\}$$

其中，Γ 是所有候选解组成的集合，构成解空间，且满足 $\Gamma\subseteq H$。

定义 6 在 PMWL 问题中，匹配解集是解空间中满足 one-off 约束的匹配解集合，可记为：

$$S=\{\Gamma_i\,|\,\Gamma_i\in\Gamma,\text{s.t. }AC_{oneoff}(\Gamma_1,\Gamma_2,\cdots,\Gamma_k)\}$$

其中，S 为 PMWL 问题的输出，且满足 $S\subseteq\Gamma$，集合 S 的大小(cardinality)可表示为 $|S|$，称为匹配数。其中，解 $\Gamma_i=(\alpha_1^i,\alpha_2^i,\cdots,\alpha_m^i)$是字符匹配位置的序列，因而可对解集 $\Gamma_1,\Gamma_2,\cdots,\Gamma_K$ 进行字典排序，使得解集 S 为匹配解的序列集。

定义 7 在 PMWL 问题中，给定多组解集，构成集合$U=\{S_i\}$。考虑到 U 是有限集，则匹配数存在最大值

$$|S|_{\max}=\max\{|\{S_i\}|\},i=1,2,\cdots,N$$

称匹配数$|S|=|S|_{\max}$的解集是最优解集，$|S|_{\max}$为完备匹配数。注意到，解集 S 存在且不唯一。因而可以打造完备解集的集合：

$$U_{complete}=\{S_i\,||S_i|=|S|_{\max}\}$$

例如，给定文本 $T=GGAAAA$，$P=G\,\phi[0,1]A\,\phi[0,1]A$，候选解集为$\{\{1,3,4\},\{1,3,5\},\{2,3,4\},\{2,3,5\},\{2,4,5\},\{2,4,6\}\}$，在 one-off 约束下，完备解集为$\{\{1,3,5\},\{2,4,6\}\}$，完备匹配数为 2。

定义 8 PMWL 问题的最优解可记为：

$$OPMWL\langle Y,g,PMWL\rangle$$

其中，决策变量是：

$$Y=(Y_1,Y_2,\cdots,Y_K)\subseteq\Gamma$$

效用函数是：

$$g(Y):\Gamma\to\mathcal{R}$$

在 one-off 约束下有：

$$g(Y)=\max|Y|,\text{s.t. }AC_{oneoff}(Y_1,Y_2,\cdots,Y_K)$$

定义 9 在不考虑约束条件下，模式 $P=p_1p_2\cdots p_m$ 的解空间是 P 中各子模式值域的笛卡尔积，记为 H：

$$H=D^m=\{(\alpha_1,\alpha_2,\cdots,\alpha_m)\,|\,\alpha_i\in D,1\leqslant i\leqslant m\}$$

在引入匹配的定义后，设 $D_1,D_2,\cdots,D_m$ 是 P 中各字符在 T 的匹配位置组成的集合：

$$D_i=\{k\,|\,P_i=T_k,k=1,2,\cdots,n\}$$

此时，由值域 $D_1,D_2,\cdots,D_m$ 的笛卡尔积可构成初始解空间：

$$H=D_1\times D_2\times\cdots\times D_m=\{(\alpha_1,\alpha_2,\cdots,\alpha_m)\,|\,\alpha_i\in D_i\wedge T_{\alpha(i)}=P_i,1\leqslant i\leqslant m\}$$

进一步，引入长度约束后，解空间即为候选解集，可表述为：

$\Gamma=\{(\alpha_1,\alpha_2,\cdots,\alpha_m)\mid \alpha_i\in D_i \wedge T_{\alpha(i)}=P_i \wedge \alpha_{j+1}^i-\alpha_j^i-1\in [l_i,u_i],1\leqslant i\leqslant m\}$

3 PMWL 问题的基本性质

以下 5 条性质说明了 PMWL 问题在 8 种限定条件下的完备性，这些特定条件下的完备性都已被证明或求解，通过本文提出的求解模型，可以将这些性质统一表述。

性质 1 PMWL 问题的长度约束区间都为$[0,\infty)$时，问题可解。

证明：由三元组模型，长度约束可表示为：

$LC_{gap}(X_i):X_i=(\alpha_1^i,\alpha_2^i,\cdots,\alpha_m^i)$, s. t. $\alpha_{j+1}^i-\alpha_j^i-1\in[0,\infty),1\leqslant j<m$

即满足

$LC_{gap}(X_i):X_i=(\alpha_1^i,\alpha_2^i,\cdots,\alpha_m^i)$, s. t. $\alpha_{j+1}^i>\alpha_j^i,1\leqslant j<m$

则问题中通配符可以匹配任意长度的模式串，Kucherov 等人[14]求解了此问题。

性质 2 PMWL 问题的长度约束区间的上下界相同时，问题可解。

证明：由三元组模型，长度约束可表示为：

$LC_{gap}(X_i):X_i=(\alpha_1^i,\alpha_2^i,\cdots,\alpha_m^i)$, s. t. $\alpha_{j+1}^i-\alpha_j^i-1\in[l_j,u_j],1\leqslant j<m$

即满足

$LC_{gap}(X_i):X_i=(\alpha_1^i,\alpha_2^i,\cdots,\alpha_m^i)$, s. t. $\alpha_{j+1}^i-\alpha_j^i=C_j,1\leqslant j<m,C_j$ 是常数

则子模式之间的距离为固定长度，因而问题转化为带有通配符的模式匹配，其中通配符仅匹配单个字符，问题可解[15]。

例如，模式 $P=a\notin\notin gg\notin c\notin t$。

性质 3 PMWL 问题若不考虑 one-off 约束，则问题可解。

证明：由三元组模型，问题可表述为

$PWML=\{X,D,LC_{gap}\}$

Min 等人[10]给出了此问题的求解算法。

性质 4 PMWL 问题若不考虑长度约束和 one-off 约束，问题转化为多模式匹配问题，此问题可解。

证明：由三元组模型，模式可表述为 $P=\{p_1,p_2,\cdots,p_m\}$，这里 P 为集合，而不是序列集，则问题转化为多模式匹配问题，SBOM[16]算法可以求解。

性质 5 PMWL 问题中，若模式满足：(1)子模式长度都为 1，即满足$|p_1|=|p_2|=\cdots=|p_m|=1$；(2)子模式 $p_1,p_2,\cdots,p_m$ 两两不同，则问题可解。

证明：该问题等价于在单个模式中两个相邻字符之间插入带有长度约束的通配符，且模式中无重复字符。此问题已被王等人[17]证明可解。

例如，合$\notin[1,2]$工$\notin[0,4]$大。

4 PMWL 问题求解

4.1 FIN 算法思路

对于 PMWL 问题，在不考虑 one-off 条件下，PAIG 算法可以得到完备匹配数[10]，但目前还缺少可以得到完备的匹配位置的算法，本文将给出 FIN 算法以得到此特定条件下完备的匹配位置。FIN 算法将通过树结构表示 PMWL 问题的匹配解，其中节点对应为子模式的匹配位置，每条路径则表示相邻子模式满足长度约束。一条自根节点到叶子节点的路径即为匹配解。

4.2 FIN 算法流程

FIN 算法的主要思路是将 PMWL 问题表示为多叉树，其中节点表示模式在文本中的匹配位置，且第 i 层节点对应为字符 p_i 的匹配位置；相邻层次 i 与 $i+1$ 的节点间可能存在边，需满足字符 p_i 与 p_{i+1} 的匹配位置在长度约束范围内。由此，一条自根节点到叶子节点的路径即为一个匹配解，上述路径的集合即为解集，由此可将 PMWL 问题转化为多叉树的路径搜索问题。

1)主函数的流程

```
Procedure Main
输入:文本 T=t1t2…tn,模式 P=p1p2…pm,长度约束[lk,uk],1≤k<m
输出:候选匹配解集 Γ
(1)for begin←1 to n do
(2)if t[begin]=p[1] then
(3)BuildTree(T,P,[lk,uk],begin,Tree); //构建多叉树
     //先序遍历多叉树,输出所有从根到叶子节点的路径
(4)GetOccurreces(T,P,Tree,Γ);
```

2)BuildTree 函数的流程

```
Procedure BuildTree(T,P,[lk,uk],begin,Tree)
输入:文本 T=t1t2…tn,模式 P=p1p2…pm,可变跨度[lk,uk],1≤k<m,
     匹配起始位置 begin,root 为根节点,Q 为存放节点的队列
输出:以 root 为根节点的多叉树
(1)postInP ← 1
(2)currentPost ← begin
     // 一个节点对应一个匹配对〈p[i],t[j]〉
(3)root. postInT ← begin
(4)root. postInP ← 1
(5)Q. push(root)
(6)while not Q. empty() &&. currentPost < n do
(7)TreeNode newNode ← Q. top()
(8)currentPost ← Q. top(). postInT
(9)Q. pop()
(10)if newNode. postInP=m-1 then
(11)continue
(12)for i ← p[postInP]. min to p[postInP]. max do
(13)if T[i]=p[postInP+1]. letter then
(14)TreeNode childNode
(15)childNode. postInT ← i
(16)    childNode. postInP ← newNode. postInP + 1
(17)newNode. childrenQueue. push(childNode)
(18)Q. push(childNode)
```

3)GetOccurreces 函数的流程

```
Procedure GetOccurreces(T,P,Tree,Γ)
输入:文本 T=t1t2…tn,模式 P=p1p2…pm,多叉树 Tree
输出:候选匹配解集
(1)while T. childrenQueue. size() > 0 do
(2)occurrence. push(T. childrenQueue. front(). postInT)
(3)    GetOccurrences(T. childrenQueue. front())
(4)if not T. childrenQueue. empty() then
(5)T. childrenQueue. pop()
(6)if T. childrenQueue. empty() then
(7)    if occurrence. size()=m then
(8)result. push(occurrence)
(9)    occurrence. clear()
```

例 1 $T=acctcctgtt$，$P_1=act$，$P2=a\notin\notin c\notin t$，$P_3=a\notin[0,3]c\notin[1,3]t$，所形成的匹配解结构如图 1 所示。

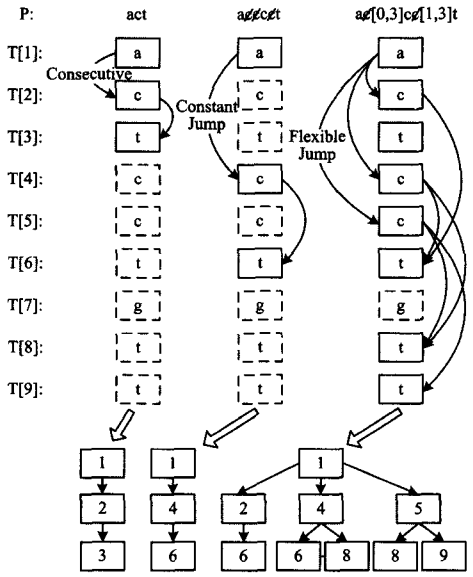

图 1　PMWL 问题的解结构示例图

所得到的匹配解分别为 $Occurrences_1=\{1,2,3\}$，$Occurrences_2=\{1,4,6\}$，$Occurrences_3=\{\{1,2,6\},\{1,4,6\},\{1,4,8\},\{1,5,8\},\{1,5,9\}\}$。

4.3　FIN 算法完备性证明

性质 6　PMWL 问题可等价表示为树结构。

证明：该转化可分为两步。

(1)PMWL 问题的输入可转化为树结构搜索问题的输入。

PMWL 问题的输入是文本 $T=t_1t_2\cdots t_n$、模式 $P=p_1p_2\cdots p_m$ 和长度约束 $[l_k,u_k]$，其中 $1\leqslant k<m$。由此可以构造树：

$$Tree=(V,E)$$

其中，

$$V=\{v_{ij}\mid v_{ij}=(p_i,t_j),\text{ s. t. } p_i=t_j,1\leqslant i\leqslant m,1\leqslant j\leqslant n\}$$

$$E=V^2=\{e_{ij,i'j'}\mid e_{ij,i'j'}=(v_{ij},v_{i'j'}),\text{ s. t. } i'-i=1\wedge j'-j-1\in[l_i,u_i],1\leqslant i<i'\leqslant m,1\leqslant j<j'\leqslant n\}$$

因此，节点 v_{ij} 表示子模式 p_i 在文本 T 中匹配位置是 t_j。$e_{ij,i'j'}=(v_{ij},v_{i'j'})$ 是从节点 v_{ij} 到节点 $v_{i'j'}$ 的有向边，表示子模式 p_i 与 $p_{i'}$ 在文本 T 中的匹配位置满足长度约束。此外，称节点 v_{ij} 的层次是 i，根节点的层次为 1。

(2)树结构搜索问题的输出可转化为 PMWL 问题的输出。

在树结构中，一条自顶层的根节点到底层的叶子节点的路径记为 top-down 路径，且构成集合 Q，需要证明集合 Q 与候选解集 Γ 是双射关系。

证明：在树结构图中，一条 top-down 路径可表示为

$$q=\{v_{1,j(1)},e_{1,j(1),2,j(2)},v_{2,j(2)},\cdots,v_{m-1,j(m-1)},e_{m-1,j(m-1),m,j(m)},v_{m,j(m)}\}$$

其中，$v_{1,j(1)}$ 为根节点，$v_{m,j(m)}$ 为底层叶子节点。为简化，可忽略路径中的边，表示为顶点序列：

$$q=\{v_{1,j(1)},v_{2,j(2)},\cdots,v_{m,j(m)}\mid \text{s. t. } j(1)\neq j(2)\neq\cdots\neq j(m)\wedge t_{j(k+1)}-t_{j(k)}-1\in[l_k,u_k],1\leqslant k<m\}$$

由此可构造匹配解

$$\Gamma_j=\{j(1),j(2),\cdots,j(m)\}$$

因此，存在映射 $f:Q\to\Gamma$。

另一方面，假设存在匹配解 $\Gamma_j=\{j(1),j(2),\cdots,j(m)\}$，且在树结构图中不存在对应的路径，可分两种情况讨论：

①存在匹配位置 $j(k)$ 在树结构上没有对应的节点；

②存在节点 $v_{k,j(k)}$ 不与节点 $v_{k-1,j(k-1)}$ 或节点 $v_{k+1,j(k+1)}$ 连接。

由问题输入的等价性进行分析，对于情况①，必然存在节点 $v_{k,j(k)}$ 用于表示匹配位置 $j(k)$；对于情况②，考虑 Γ_j 是候选解，则满足 $j(k)-j(k-1)\in[l_k,u_k]$，因而在节点 $v_{k-1,j(k-1)}$ 与 $v_{k,j(k)}$ 之间一定存在边。同理，在节点 $v_{k,j(k)}$ 与 $v_{k+1,j(k+1)}$ 之间一定存在边。因而假设错误，存在映射 $g:\Gamma\to Q$。因此，集合 Q 与集合 Γ 为双射关系，即 PMWL 问题可等价表示为树结构。

由 FIN 算法流程可知，给定多叉树，FIN 可对树中所有节点遍历，以得到树中的所有从根到叶子节点的路径作为匹配解，由性质 6，这些匹配解即为 PMWL 问题在不考虑 one-off 约束时的匹配解集。因此 FIN 算法在此特定条件下完备。进一步，由定义 9，在不考虑 one-off 条件下 FIN 所得的完备解即为 PMWL 问题的解空间。

4.4　FIN 算法的时间复杂度分析

由算法描述可知，

$$O(FIN)=O(BuildTree)+O(GetOccurreces)$$

$$O(FIN)=k*O(GetOccurreces)+O(GetOccurreces)$$

$$O(FIN)\sim O(GetOccurreces)$$

由性质 6，树的节点对应于一个匹配位置，树从根到叶子节点的路径对应于一个匹配解。因此匹配解的规模等价于节点数，考虑到访问一个节点的时间是 O(1)，有：

$$O(GetOccurreces)\sim O(N*m)$$

其中 N 为模式的匹配数，m 为子模式数目。因此有

$$O(FIN)\sim O(N*m)$$

综上，FIN 算法的时间复杂度与模式的匹配数和子模式数目呈线性关系。

5　实验结果及分析

本节使用真实 DNA 序列和基因序列作为实验文本，模式选用随机生成的满足特征的模式和拟南芥的模体序列，其中 m 表示子模式数，gap 表示模式通配符跨度，N 表示每组特征下的模式数目。实验环境为：奔腾双核 CPU，2.2GHz 主频，2GB 内存，Windows 7 操作系统。

1. DNA 序列测试

(1)选用文本 AX829174，其包含 10011 个字符，构造随机模式，满足 $N=10$，$m=\{2,3,4\}$，$gap=\{5,6,\cdots,14\}$。模式跨度 gap 和子模式数 m 对匹配数的影响如图 2 所示。

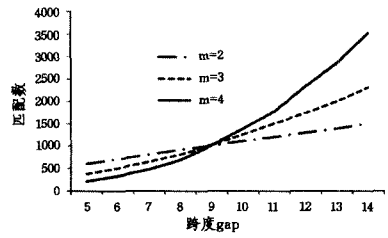

图 2　模式跨度 gap 和子模式数 m 对匹配数的影响

(2) 选用文本 AB008226_1，其包含 131892 个字符，构造随机模式，满足 $N=10$，$m=4$，$gap=\{10,20,\cdots,70\}$。随着模式跨度的增加 PMWL 问题匹配数的变化趋势如图 3 所示。

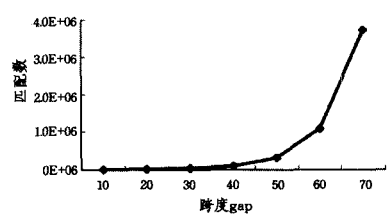

图 3　随着模式跨度的增加 PMWL 问题匹配数的变化趋势

结果表明：

(1)PMWL 问题中，模式跨度 gap 和子模式数 m 越大，匹配解规模越大，且随 gap 的增加匹配解呈指数级增长；

(2)FIN 算法的时间复杂度取决于匹配解的规模，因此 FIN 算法更适用于模式跨度较小的情况。

2. 完备性对照实验

采有 PAIG 算法[10]可以得到 PMWL 问题的完备匹配数，本实验在真实 DNA 序列集中对照 FIN 算法与 PAIG 算法，结果如表 1 所列。

表 1　FIN 算法在基因序列中的测试结果

模式	文本		
	AX829174	AB008226_1	AB038490_1
a[0,3]t[0,3]a[0,3]t[0,3]a[0,3]t[0,3]a[0,3]t[0,3]a[0,3]t[0,3]a	27232	30841	457853
g[1,5]t[0,6]a[2,7]g[3,9]t[2,5]a[4,9]g[1,8]t[2,9]a	98292	69998	1570341
g[1,5]t[0,6]a[2,7]g[3,9]t[2,5]a[4,9]g[1,8]t[2,9]a[1,9]g[1,9]t	480155	309938	7823457
c[1,5]g[0,6]a[1,7]g[3,5]t[2,5]a[1,9]t[1,8]c[2,9]t[4,9]a	93338	70916	1243020
a[0,4]g[0,4]g[0,4]t[0,4]a[0,4]g[0,4]a[0,4]g[0,4]a[0,4]g[0,4]a[0,4]a[0,4]a	182381	58849	1484873
t[1,5]t[0,6]t[2,7]a[3,4]a[2,5]g[4,9]g[1,8]t[2,9]t[1,5]a[4,5]t[1,8]t[2,4]a	106812	101852	2834853
g[1,6]t[1,6]a[1,6]g[1,6]t[1,6]a[1,6]g[1,6]t[1,6]a[1,6]g[1,6]t	156328	118059	2627907

表 1 结果表明：

(1)FIN 算法与 PAIG 算法的匹配数一致，都是完备解，而 FIN 可同时得到完备匹配位置。

(2)当模式的跨度较长时，例如 g[1,9]t[1,9]a[1,9]g[1,9]t[1,9]a[1,9]g[1,9]t[1,9]a[1,9]g[1,9]t，该模式在 AX829174 上的匹配数约为 9610000，使用 FIN 算法无法返回结果。考虑到 PMWL 匹配数随着模式跨度呈指数级增长，而 FIN 的计算时间与匹配数线性相关，FIN 算法适用于模式跨度较小的情况。

3. 基因序列测试

在生物信息学中，基因序列由{A,C,G,T} 4 种含氮碱基构成，而模体(motif)是蛋白质中具有特定构象和功能的结构成分，可通过基因中一段特定序列表达，字母表由 IUPAC 标准定义。在研究中，给定一组基因序列，一个常见的需求是从中寻找出现频率较高的模体，并以此作为基因的描述特征[18]。一个模体可能由多个基因片段表达，从而形成了多个不连续的子序列。如序列数据库 PROSITE 中模体可表示为 $[RK]-x[2,3]-[DE]-x[2,3]-Y$，其中 $[RK]$可匹配字符 R 或 K，$x[2,3]$可指代长度为 2～3 的字符串，即为长度约束。因此，在基因序列中搜索模体与 PMWL 问题定义一致。

本实验数据来自 PROSITE 数据库，文本选用拟南芥基因序列 chr1AT、chr2AT、chr3AT、chr4AT 和 chr5AT；模式为 TNGA·ϕ[12,14]·TWNYTNNA·ϕ[19,21]·TNTMYRT·ϕ[4,6]·WNCCNNNNRG·ϕ[72,95]·TGNNA·ϕ[100,125]·TNTANRTNRAYGA[19]。实验结果如表 2 所列。

表 2　FIN 算法在基因序列中的测试结果

文本序列	大小(MB)	匹配数	逆转录因子	运行时间(s)
chr1AT	29	19	7	82
chr2AT	19	3	1	59
chr3AT	22.7	8	0	73
chr4AT	16.9	7	2	60
chr5AT	25.7	8	7	89

由表 2 可知，在文本 chr1AT－chr5AT 上，FIN 算法都可以在 100s 内得到基因序列中模体的匹配位置，其中 37.8%被确认为是逆转录因子[19]。该结果表明，模体在基因序列中的匹配过程可以抽象为 PMWL 问题，通过 FIN 算法可以发现有潜在生物学意义的匹配结果。

结束语　针对带有通配符和长度约束的模式匹配问题(PMWL)，已有工作研究了其在特定条件下的完备性，并对问题的一般情况通过设计启发式算法得到近似最优解。本文基于上述研究，借鉴约束可满足问题框架，给出了 PMWL 问题的三元组求解模型，模型对问题的约束条件和解空间等基本概念作出了形式化描述，并且将问题已知的 5 种特殊情况统一表述为问题的基本性质。之后，本文给出了算法 FIN，可得到 PMWL 问题在特定条件下的完备解集，并从理论上说明了 FIN 算法的完备性。实验结果表明，当模式跨度较大时匹配解将呈指数级增长，因此 FIN 算法更适用于模式跨度较小的情况。

参 考 文 献

[1] Fischer M J, Paterson M S. String matching and other products [R]. Cambridge, MA: Massachusetts Institute of Technology, 1974

[2] Manber U, Baeza-Yates R. An Algorithm for String Matching with a Sequence of Don't cares[J]. Information Processing Letters, 1991, 37(3): 133-136

[3] Califf M E, Mooney R J. Bottom-up Relational Learning of Pattern Matching Rules for Information Extraction[J]. Journal of Machine Learning Research, 2003, 4(6): 177-210

[4] Guo Dan, Hu Xue-gang, Xie Fei, et al. Pattern Matching with Wildcards and Gap-length Constraints Based on a Centrality-degree Graph[J]. Applied Intelligence, 2013, 39: 57-74

[5] Dahiya A, Garg D. Maximal pattern matching with flexible wildcard gaps and one-off constraint[C]// International Conference on Advances in Computing, Communications and Informatics. Arras, France, 2014: 1107-1112

[6] Wu Xin-dong, Qiang Ji-peng, Xie Fei. Pattern Matching with Flexible Wildcards[J]. Journal of Computer Science and Technology, 2014, 29(5): 740-750

[7] Xiang Tai-ning, Guo Dan, Wang Hai-ping, et al. Characteristic Analysis of Pattern Matching with Wildcards Problem and its Solution Space[J]. Computer Science, 2014, 41(9): 269-273(in Chinese)
项泰宁，郭丹，王海平，等. 带通配符的模式匹配问题及其解空间特征分析[J]. 计算机科学，2014，41(9)：269-273

[8] Chen Gong, Wu Xin-dong, Zhu Xing-quan, et al. Efficient string matching with wildcards and length constraints[J]. Knowledge and Information Systems, 2006, 10(4): 399-419

[9] Qiang Ji-peng, Xie Fei, Gao Jun, et al. Pattern Matching with Arbitrary-length Wildcards[J]. Acta Automatica Sinica, 2014, 40(11): 2499-2511(in Chinese)

(下转封三)

目标检测算法的基础上，通过构建类属超图(CSHG)模型，提出了基于 Adaboost-CSHG 的目标跟踪识别算法，通过 RSOM 聚类树，改善了单独利用 Adaboost 目标检测算法训练出的特定类目标坦克模型级联分类器的检测效果，有效滤除了大量的目标虚警，并同时完成了对坦克目标的跟踪；最后在目标"精检测"的基础上，利用基于类属超图的目标识别原理实现了对目标的识别，实验结果均表明算法在简单背景和复杂背景图像条件下具有可行性。但是本文的研究仍存在一些不足，通过 Adaboost 目标检测算法只训练得到特定的坦克分类器，使得目标检测对象受限，因此该方法仅适用于同一特定类型的多个感兴趣目标的跟踪与识别，后期可对目标检测进行更多的探究摸索。

参 考 文 献

[1] Liu Jian-jun. Research on Local Invariant Features Based Class Specific Hyper Graphs Learning and Object Recognition[D]. Changsha: National University of Defense Technology, 2010: 110-112(in Chinese)
刘建军. 基于图像局部不变特征的类属超图构建与目标识别技术研究 [D]. 长沙：国防科学技术大学，2010：110-112

[2] Li Jie. Human Detection Based on Adaboost Algorithm[D]. Beijing: North China University of Technology, 2010: 17-20(in Chinese)
李杰. 基于 Adaboost 算法的人体目标检测[D]. 北京：北方工业大学，2010：17-20

[3] Ai Juan. Implement of Face Detection and Study of Eye Location [D]. Shanghai: Fudan University, 2008: 25-26(in Chinese)
艾娟. 人脸检测实现及眼睛定位算法研究[D]. 上海：复旦大学，2008：25-26

[4] Sivic J, Russllb C, Efros A A, et al. Discovering Objects and Their Location in Images [J]. International Conference on Computer Vision, 2005, 1(1): 872-877

[5] Csurka G, Dance C R, Fan L, et al. Visual categorization with bags of keypoints [C]// Workshop on Statistical Learning in Computer Vision. ECCV, 2004: 1-22

[6] Torralba A, Fergus R, Weiss Y. Small Codes and Large Image Databases for Recognition [C]// International Conference on Computer Vision. 2008

[7] Bonev B, Escolano F, Lozano M A, et al. Constellations and the Unsupervised Learning of Graphs[J]. Proceedings of the Graph-Based Representations in Pattern Recognition, 2007, 14(1): 340-350

[8] Torsello A, Hancock E. Learning Shape-Classes Using a Mixture of Treeunions [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(6): 954-967

[9] Jiang X, Munger A, Bunke H. On Median Graphs: Properties, Algorithms and Applications [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(10): 1144-1151

[10] Ferrari V, Tuytelaara T, Van-Cool L. Simultaneous Object Recognition and Segmentation from Single or Multiple Model Views [J]. International Journal of Computer Vision, 2006, 67(2): 159-188

[11] Lowe D. Local Feature View Clustering for 3d Object Recognition [J]. International Conference on Computer Vision and Pattern Recognition, 2001, 2(1): 1682-1688

[12] Chung F. Spectral Graph Theory [C]// CBMS Regional Conference Series in Mathematics. Conference Board of the American Mathematical Science, Washington D C, 1997: 92

[13] Zheng Jun-jun, Xia Sheng-ping, Li Xin-guang, et al. K nearest neighbors detecting algorithm based on a RSOM tree[J]. Journal of Shangdong University, 2011, 41(2): 80-84(in Chinese)
郑君君，夏胜平，李新光，等. 基于 RSOM 聚类树的图像 K 近邻求解算法[J]. 山东大学学报，2011，41(2)：80-84

[14] Liu Jian-jun, Zhu Yi-wei, Li Xin-guang, et al. Imaging Object Recognition Based on Hyper Graph Model[J]. Computer Engineering, 2010, 36(21): 181-184(in Chinese)
刘建军，祝一薇，李新光，等. 基于超图模型的图像目标识别[J]. 计算机工程，2010，36(21)：181-184

[15] Xia Sheng-ping, Song Rui, Liu Jian-jun, et al. Learning Large Scale Class Specific Hyper Graphs forNon-Cooperative Object Recognition[J]. Acta Electronica Sinica, 2011, 39(6): 1399-1404 (in Chinese)
夏胜平，宋锐，刘建军，等. 面向非合作目标识别的大规模类属超图建模[J]. 电子学报，2011，39(6)：1399-1404

(上接第 283 页)

强继朋，谢飞，高隽，等. 带任意长度通配符的模式匹配[J]. 自动化学报，2014，40(11)：2499-2511

[10] Min Fan, Wu Xin-dong, Lu Zhen-yu. Pattern matching with independent wildcard gaps[C]// Proceedings of the 8th IEEE International Conference on Dependable, Autonomic and Secure Computing. 2009: 194-199

[11] Wu You-xi, Liu Ya-wei, Guo Lei, et al. Subnettrees for Strict Pattern Matching with General Gaps and Length Constraints [J]. Journal of Software, 2013, 24(5): 915-932(in Chinese)
武优西，刘亚伟，郭磊，等. 子网树求解一般间隙和长度约束严格模式匹配[J]. 软件学报，2013，24(5)：915-932

[12] Brailsford S C, Potts C N, Smith B M. Constraint satisfaction problems: Algorithms and applications[J]. European Journal of Operational Research, 1999, 119: 557-581

[13] Bala S. Regular language matching and other decidable cases of the satisfiability problem for constraints between regular open terms[J]. Theory of Computing Systems, 2004, 39: 596-607

[14] Kucherov G, Rusinowitch M. Matching a set of strings with variable length don't cares[C]// Proceedings of the 6th symposium on Combinatorial Pattern Matching. 1995: 230-247

[15] Kalai A. Efficient pattern-matching with don't cares[C]// Proceedings of the 13th ACM-SIAM Symposium on Discrete Algorithms. 2002: 655-656

[16] Allauzen C, Raffinot M. Factor oracle of a set of words: Technical Report 99-11[R]. Institut Gaspard-Monge, 1999

[17] Wang Hai-ping, Hu Xue-gang, Xie Fei, et al. Impact of Pattern Feature on Pattern Matching Problem with Wildcards and Length Constraints[J]. Pattern Recognition and Artificial Intelligence, 2012, 25(6): 1013-1021(in Chinese)
王海平，胡学钢，谢飞，等. 模式特征对带有通配符和长度约束的模式匹配问题的影响[J]. 模式识别与人工智能，2012，25(6)：1013-1021

[18] Morgante M, Policriti A, Vitacolonna N, et al. Structured motifs search[J]. Journal of Computational Biology, 2005, 12(8): 1065-1082

[19] Initiative T A G, Copenhaver G P. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana[J]. Nature, 2002, 408(6814): 796-815