字符分布特征对带有通配符串匹配问题的影响*

王海平, 郭　丹†, 项泰宁, 胡学钢
(合肥工业大学 计算机与信息学院, 合肥 230009)

摘　要: 近年来,字符串匹配问题被不断扩展。其中,具有代表性的是在模式中引入可变长度的通配符,称之为 PMWL 问题。针对此问题,已有工作分析了在不同的模式特征下,匹配数 Ω 随文本长度增加呈指数级增长。同时考虑文本分布特征和模式特征,建立了期望模型 $E(\Omega)=nD\pi(P)$,其中 n 为文本长度,D 为模式中各通配符跨度的乘积,$\pi(P)$ 为基于字符分布的模式出现概率。实验部分,在人工随机数据和 DNA 真实数据上验证了 $E(\Omega)$ 的准确性,得到预测误差率分别为 1.8% ~3.2% 和 4.7% ~7.8%;在不同字符分布中,分析了模式模长和通配符跨度对匹配数 Ω 的影响。$E(\Omega)$ 模型揭示了 Ω 的增长趋势不一定呈指数级,而取决于 $\pi(P)$ 和 D 的共同影响,且 $E(\Omega)$ 模型能够在线性时间内得到近似完备解。

关键词: 字符分布; 串匹配; 期望模型; 通配符; 长度约束
中图分类号: TP391.1　　**文献标志码**: A　　**文章编号**: 1001-3695(2016)02-0356-06
doi:10.3969/j.issn.1001-3695.2016.02.008

Character distribution in string matching with wildcards and length constraints

Wang Haiping, Guo Dan, Xiang Taining, Hu Xuegang
(*School of Computer & Information*, *Hefei University of Technology*, *Hefei* 230009, *China*)

Abstract: Recently, with the development of bioinformatics, information retrieval and other fields, string matching problem was extended with different conditions, especially, the introduction of the problem of pattern matching with wildcards and length constraints (PMWL). Existing results show that the number of occurrences Ω performed the exponential growth with the size of text T by analyzing the pattern characteristic. Based on both character distribution characteristic and pattern characteristic, this paper proposed an expectation model $E(\Omega)=nD\pi(P)$, where n was the length of text T, D was the product of all gap in pattern and $\pi(P)$ was the character distribution probability of P in T. Experiments demonstrate that the model prediction error rate are 1.8% ~3.2% and 4.7% ~7.8% on artificial random texts and DNA sequences respectively; the influence of text length, pattern length and the length of wildcards on Ω in different character distributions. $E(\Omega)$ reveals that Ω not necessarily performs the exponential growth with the size of T, and $E(\Omega)$ can effectively obtain an approximate optimal solution in linear time.

Key words: character distributions; string matching; expectation model; wildcard; length constraints

0　引言

字符串匹配问题作为计算机科学的基本问题之一,是网络安全、信息检索与过滤、计算生物学等重要领域的核心问题[1],同时也是模式挖掘技术的核心和基础[2]。通配符的引入使得匹配问题更加灵活[3],也更加符合实际应用的需要。给定字母表 Σ,通配符 $\varnothing$ 可以匹配 Σ 中任意字符。目前,在带有可变长度通配符的模式匹配问题的研究中[4~12],Chen 等人[6]定义了带有通配符和长度约束的模式匹配问题(pattern matching with wildcards and length constraints, PMWL),模式 P 中任意两个相邻字符间带有一段通配符序列 η_i,且序列长度满足约束 $\eta_i=\eta_i[l(i),u(i)]$,可记为 $P=p_1\eta_1[l(1),u(1)]\cdots p_{m-1}\eta_{m-1}[l(m-1),u(m-1)]p_m$,约束区间长度 $\mathrm{gap}(i)=u(i)-l(i)+1$ 为通配符跨度。例如,给定文本 $T=$ aaggcc,模式 $P=a\eta_1[0,2]g\eta_2[0,2]c$,可得匹配位置为 $\{\{1,3,5\},\{1,3,6\},\{1,4,5\},\{1,4,6\},\{2,3,5\},\{2,3,6\},\{2,4,5\},\{2,4,6\}\}$,匹配数 $\Omega=8$。

字符串匹配的目标是得到 P 在 T 中的匹配位置以及匹配数 Ω。传统字符串匹配问题中,给定文本 $T=t_1t_2\cdots t_n$ 和模式 $P=p_1p_2\cdots p_m$,寻找 P 在 T 中匹配的子序列。传统串匹配的研究不仅关注算法设计,而且着重于分析问题求解的复杂度,而字符的比较次数 C_n 是衡量复杂度的重要标准。C_n 的计算需要基于文本中的字符分布特征[13~16],不同的背景数据往往带有完全不同的分布[17]。例如,人工随机生成的文本数据近似服从伯努利分布,而真实文本数据常采用 K 阶马尔可夫模型模拟得到。针对不同字符分布,文献[18,19]分析了模式匹配数的概率模型及其统计特性。然而,PMWL 问题作为字符串匹配问题的扩展,尚未有工作分析文本字符分布特征对问题匹配数 Ω 的影响。由此,本文认为字符分布是字符串匹配问题的重要参

收稿日期: 2014-10-17; **修回日期**: 2014-12-08　　**基金项目**: 国家自然科学基金资助项目(61305062);国家自然科学基金—港澳学者合作研究基金资助项目(61229301);博士后面上基金资助项目(2012M511403);安徽省自然科学基金资助项目(1308085QF102)

作者简介: 王海平(1986-),男,博士研究生,主要研究方向为模式匹配、机器学习;郭丹(1983-),女,讲师,博士(后),主要研究方向为机器学习、数据挖掘(guodan@hfut.edu.cn);项泰宁(1988-),男,硕士研究生,主要研究方向为模式匹配、数据挖掘;胡学钢(1961-),男,教授,博导,主要研究方向为数据挖掘、人工智能.

数,在 PMWL 问题求解中引入该参数将有助于问题求解和复杂性分析。

基于以上分析,本文针对 PMWL 问题,考虑文本的分布模型,将匹配数 Ω 视为随机变量,研究了关于 Ω 的计数模型及性质。本文给出 Ω 的期望模型 $E(\Omega)=nD\pi(P)$,其中 n 为文本长度,D 为模式中各通配符跨度 gap 的乘积,$\pi(P)$ 为基于字符分布的模式出现概率。本文 4.1 节分析了在不同文本字符分布中的模型误差率。在人工随机文本中,误差率在 1.8% ~ 3.2%;在真实文本中,本文采用 DNA 序列,模型误差率为 4.7% ~7.8%。4.2 和 4.3 节揭示了在不同字符分布下影响 PMWL 问题的因素,包括模式模长和通配符跨度,以及上述因素之间的相互制约关系。由此,本模型的意义在于:a)现有的能够计算得到完备解的算法都为多项式级时间复杂度,在大文本的匹配和挖掘中难以适用,而 $E(\Omega)$ 适用于不同文本,可在线性时间内给出模式匹配次数的估计值;b)该模型可为 PMWL 问题求解的复杂性研究提供依据。

1　相关工作

对传统字符串匹配问题,算法研究关注于模式特征。匹配算法往往利用模式特征提高匹配效率[1],如 KMP 算法利用了模式前缀,BM 算法利用了模式后缀,而 BNDM 算法则是基于模式子串。随后,带有不同的通配符特征的模式引起了人们的广泛关注[4~12]。Navarro 等人[5]致力于解决带有字符组和限长空位(classes of characters and bounded size gap,CBG)的模式匹配问题,模式 P 中任意两个相邻字符间带有通配符序列 $\eta_i=\eta_i[l(i),u(i)]$,最短长度为 $l(i)$,最大长度为 $u(i)$。Chen 等人[6]研究了与之类似的模式。本文称为 PMWL 问题。对于该问题,Min 等人[7]给出完备求解模式匹配数的 PAIG 算法,其算法时间复杂度与文本长度、模式长度和通配符约束呈多项式级关系。目前,对 PMWL 问题的分析往往基于模式特征,即模式模长、通配符间隔约束 η_i 以及通配符跨度 gap_i。此外,对于带有通配符的模式挖掘问题,在给定模式 P 和支持度下,计算 P 在文本中潜在的匹配次数同样仅依赖于模式自身的特征[8,9]。

另一方面,文本特征同样对字符串匹配问题的复杂度分析有着重要作用。考虑文本的概率模型,研究关注两类随机文本[14]:无记忆信源(伯努利模型)生成的文本和 K 阶平稳马尔可夫链(马尔可夫模型)生成的文本。对于不同文本中的字符分布,匹配过程中字符的比较次数 C_n 是衡量算法和问题复杂度的重要指标。基于伯努利模型,Yao[15]通过分析 C_n 证明了单模式串匹配的时间复杂度下界。Baeza-Yates[14]基于模式生成概率 $\pi(P)$,给出了随机文本中模式的匹配数的计算公式及其统计性质。由于实际问题的需要,针对不同字符分布类型的计数模型得到深入研究[17]。然而,以上工作中模式并不带有满足长度约束的通配符,使得以上研究难以直接应用于 PMWL 问题。

对于带有通配符的模式匹配问题,文本特征同样引起了人们的关注,Li 等人[20]基于字符分布研究了带有灵活通配符的字符串匹配算法。Bourdon 等人[21]研究了不同概率模型生成的文本上模式匹配数的变化,但是该工作并不直接针对 PMWL 问题,并且公式复杂,难于实际应用。

综上,现有 PMWL 问题的研究工作缺乏对文本特征作用的衡量,对求解算法的分析往往基于模式自身的特征,而缺少对文本特征的考察,从而忽略了文本特征对算法复杂度的影响。此外,现有算法虽然可以得到模式匹配数 Ω,但算法性能往往受制于 Ω。本文针对 PMWL 问题,通过引入字符分布特征,构建估计模式匹配数 Ω 的期望模型 $E(\Omega)$,降低算法时间复杂度,同时为 PMWL 问题的复杂度分析提供参考。

2　PMWL 问题

定义 1　设字母表 $\Sigma=\{\sigma_1,\sigma_2,\cdots,\sigma_c\}$,序列 $T=t_1t_2\cdots t_n\in\Sigma^*$ 为文本串,其中 $n=|T|$ 是文本长度,$c=|\Sigma|$ 为字母表大小。

定义 2　记 ϕ 表示通配符,匹配 Σ 中任意单个字符。$\eta_j[l(j),u(j)]$ 表示一段通配符序列,最小长度为 $l(j)$,最大长度为 $u(j)$,序列的区间长度 $u(j)-l(j)+1$ 称为通配符跨度,记为 $\text{gap}(j)$。例如,$\eta[2,5]$ 表示匹配任意 2 ~5 个字符的序列,$\eta[3,3]=\phi\ \phi\phi$。

定义 3　二元组 $P=(p,\eta)$ 表示模式。$p=\langle p_1,\cdots,p_m\rangle(p_i\in\Sigma)$ 表示模式字符,m 称为模式模长。$\eta=\langle\eta_1,\cdots,\eta_{m-1}\rangle$ 表示通配符区间序列,其中 $\eta_j=\eta_j[l(j),u(j)]$ 是字符 p_j 和 p_{j+1} 之间的通配符区间,称为长度约束。模式 P 的最大长度为 $L_{\max}=m+\sum_{j=1}^{m-1}|u(j)|$。

定义 4　给定文本 $T=t_1t_2\cdots t_n$,模式 $P=(p,\eta)$。序列 $I=\langle i(1),\cdots,i(m)\rangle(1\leqslant i(1)<\cdots<i(m)\leqslant n)$ 且满足 $l_j\leqslant i(j+1)-i(j)\leqslant u_j$,则 I 称为位置序列。I 为 P 在 T 中的匹配,当且仅当 $t_{i(1)}=p_1,t_{i(2)}=p_2,\cdots,t_{i(m)}=p_m$。此时,序列 I 为 P 在 T 中的匹配位置;P 在 T 中的匹配数记为 Ω。

定义 5　带有通配符和长度约束的模式匹配问题(pattern matching with wildcards and length constraints,PMWL)满足如下条件:a)任意两字符间都带有通配符序列,且彼此独立;b)字符间通配符序列的长度受到长度约束。

3　模式匹配解数目的期望模型

3.1　文本字符分布与模式概率定义

1)文本字符分布模型　设字母表 $\Sigma=\{\sigma_1,\sigma_2,\cdots\sigma_c\}$,文本 $T=t_1t_2\cdots t_n$。字符分布可描述为:T 是一段随机过程序列且 t_i 的概率分布为 $\{Pr_i(\sigma_1),Pr_i(\sigma_2),\cdots,Pr_i(\sigma_c)\}$,即 t_i 是字符 σ_j 的概率为 $Pr_i(\sigma_j)$。

已有文献关注两类随机文本模型[13,16]:a)伯努利模型文本,文本中 t_i 和 $t_{i'}$ 的分布相同且相互独立;b)马尔可夫模型文本,文本是一段由 K 阶平稳马尔可夫链生成的字符序列,即文本中 t_i 的概率分布取决于之前的 K 个字符。真实文本模型中,文本中 t_i 的概率分布无法准确得到。

2)模式出现概率　给定模式 $P=(p,\eta)$,其中 $p=\langle p_1,\cdots,p_m\rangle$ 且 $p_j\in\Sigma$,定义模式 P 的出现概率 $\pi(P)=\prod_{j=1}^{m}Pr_j(p_j)$,其中 $Pr_j(p_j)$ 为字符 p_j 的出现概率,$1\leqslant j\leqslant m$。

3)模式字符出现频率　记 $|\sigma_i|$ 为字符 σ_i 在文本 T 中的出现频数,因此字符 σ_i 在 T 中的出现频率为 $\psi(\sigma_i)=|\sigma_i|/n$。$P$ 在 T 中的字符出现频率 $\psi(T)$ 定义为 $\psi(T)=\prod_{j=1}^{m}\psi(P_j)=\prod_{j=1}^{m}(|P_j|/n)$。

3.2 期望模型 E(Ω)

引理 1 给定字母表 Σ 上的文本 T 和模式 $P=(p,\eta)$，当 P 的最大长度 $L_{\max}$ 相对于文本长度 n 很小时，P 在 T 中匹配数 Ω 的期望 $E(\Omega)$ 为

$$E(\Omega)=D\times\prod_{i=1}^{n}\pi_i(P) \tag{1}$$

其中：$\pi_i(P)$ 是从文本位置 i 开始，模式 P 的出现概率。

证明 记 Ω 表示模式 P 在文本 T 中的匹配次数，令 $\Omega(T,i)$ 表示从文本位置 i 开始的模式匹配次数，由数学期望的性质可得

$$E(\Omega)=\prod_{i=1}^{n}E(\Omega(T,i))$$

对于文本中位置 $i\in[1,n-L_{\max}]$，令 Pr_j^i 表示当 i 为起始位置时，模式 P 中第 j 个字符 P_j 的出现概率，则 Pr_1^i 表示 t_i 是字符 P_1 的概率。由条件期望公式可得：

$$E(\Omega(T,i))=E(\Omega(T,i)\mid(t_i=p_1))Pr_1^i$$

设模式字符 p_j 在文本中匹配的位置是 $i(j)$，即 $p_j=t_{i(j)}$，其中 i 为匹配起始位置。以下分两种情况讨论：

a) 对于伯努利模型文本，由 $p_j=t_{i(j)}$，字符 p_{j+1} 在区间 $[i(j)+l(j)+1,i(j)+u(j)+1]$ 中每个位置的出现为伯努利实验，并且 $\mathrm{gap}_j=u(j)-l(j)+1$ 个伯努利实验相同且独立，因此 p_{j+1} 相对于 $p_j=t_{i(j)}$ 的出现服从二项分布 $\xi\sim B(\mathrm{gap}_j,Pr_{j+1}^i)$。$E(\Omega(T,i))$ 的计数公式为

$$\begin{aligned}
&E(\Omega(T,i))=E(\Omega(T,i)\mid(t_i=p_1))Pr_1^i=Pr_j^i\\
&\sum_{k(1)=0}^{\mathrm{gap}(1)}\cdots\sum_{k(m-1)=0}^{\mathrm{gap}(m-1)}\Big[k(1)\binom{\mathrm{gap}(1)}{k(1)}\\
&(Pr_2^i)^{k(1)}(1-Pr_2^i)^{\mathrm{gap}(1)-k(1)})\cdots k_{m-1}\\
&\binom{\mathrm{gap}(m-1)}{k(m-1)}(Pr_m^i)^{k(m-1)}(1-Pr_m^i)^{\mathrm{gap}(m-1)-k(m-1)}\Big]=\\
&Pr_1^i\times\mathrm{gap}(1)\times Pr_2^i[(1-Pr_2^i)+Pr_2^i]^{\mathrm{gap}(1)-1}\times\\
&\cdots\times\mathrm{gap}(m-1)\times Pr_m^i[(1-Pr_m^i)+Pr_m^i]^{\mathrm{gap}(m-1)-1}=\\
&(Pr_1^i\prod_{j=2}^{m}Pr_j^i)\prod_{j=1}^{m-1}\mathrm{gap}(j)=D\pi_i(P)
\end{aligned}$$

b) 对于非伯努利模型文本包括真实文本，由 $p_i=t_{i(j)}$，则字符 p_{j+1} 在区间 $[i(j)+l(j)+1,i(j)+u(j)+1]$ 上的每个位置的出现概率分别为 $Pr_{j+1,i(j)+l(j)+1}^i$，$Pr_{j+1,i(j)+l(j)+2}^i$，…，$Pr_{j+1,i(j)+u(j)+1}^i$。考虑到 $\mathrm{gap}(j)=u(j)-l(j)+1$，不妨令 $k=i(j)+l(j)+1$，可得 $Pr_{j+1,k}^i,Pr_{j+1,k+1}^i,\cdots,Pr_{j+1,k+\mathrm{gap}(j)-1}^i$，均值记为 $\overline{Pr}_{j+1}^i$。因此给定 p_j 的匹配位置 $i(j)$ 和 $\mathrm{gap}(j)$，p_{j+1} 出现是 $\mathrm{gap}(j)$ 个伯努利实验之和，但是不符合相同且独立的要求，因而不服从二项分布。利用上式中的推导过程可得

$$\begin{aligned}
&E(\Omega(T,i))=E(\Omega(T,i)\mid(t_i=p_1))Pr_1^i=\\
&Pr_1^i(Pr_{2,1}^i+Pr_{2,2}^i+\cdots+Pr_{2,\mathrm{gap}(2)}^i)\cdots\\
&(Pr_{m,1}^i+Pr_{m,2}^i+\cdots+Pr_{m,\mathrm{gap}(m-1)}^i)=\\
&Pr_1^i\prod_{j=2}^{m}\sum_{k=1}^{\mathrm{gap}(j-1)}Pr_{j,k}^i=\\
&Pr_1^i\prod_{j=2}^{m}(\overline{Pr}_j^i\mathrm{gap}(j-1))=\\
&(Pr_1^i\prod_{j=2}^{m}\overline{Pr}_j^i)\prod_{j=1}^{m-1}\mathrm{gap}(j)=D\,\overline{\pi}_i(P)
\end{aligned}$$

其中：$\overline{\pi}_i(P)=Pr_1^i\prod_{j=2}^{m}\overline{Pr}_j^i=Pr_1^i\prod_{j=2}^{m}(\overline{\kappa}_j^iPr_j^i)$ 是从文本位置 i 开始，模式 P 的平均出现概率。

综上，对于文本中位置 $i\in[1,n-L_{\max}]$，有 $E(\Omega)(T,i))=D\pi_i(P)$。因此，$E(\Omega)=D\sum_{i=1}^{n}\pi_i(P)+\Delta$，$\Delta$ 是文本中 $i\in[n-L_{\max},n]$ 位置由于序列长度不够长所造成的误差。当 P 的最大长度 $L_{\max}$ 相对于文本长度 n 很小时，可以忽略 Δ 的影响。证毕。

由引理 1，考虑不同文本的字符分布类型，可得模型(S1)和模型(S2)。

模型 S1 在随机文本中，期望 $E(\Omega)$ 为

$$E(\Omega)=Dn\pi_b(P) \tag{2}$$

其中：$\pi_b(P)=\prod_{j=1}^{m}Pr_j=\psi(P)=\psi(P_1)\prod_{j=2}^{m}\psi(P_j)$，$\psi(P_j)$ 为 p_j 在文本中出现频率。

模型 S2 在真实文本中，期望 $E(\Omega)$ 为

$$E(\Omega)=Dn\pi_m(P) \tag{3}$$

其中：$\pi_m(P)=\sum_i\overline{\pi}_i(P)/n=\sum_{i=1}^{n}(\kappa_1^i\prod_{j=2}^{m}\overline{\kappa}_j^i)\psi(p_1)\prod_{j=2}^{m}\psi(p_j)/n=K\psi(p_1)\prod_{j=2}^{m}\psi(p_j)$，$\pi_m(P)$ 为 P 在文本中的平均出现概率；$K=\sum_{i=1}^{n}(\kappa_1^i\prod_{j=2}^{m}\overline{\kappa}_j^i)/m$。

随机文本模型 $\pi_b(P)$ 中 $K=1$。对比 $\pi_b(P)$ 和 $\pi_m(P)$ 可知，$\pi_m(P)=K\pi_b(P)$，$\pi_b(P)$ 可由模式字符出现频率 $\psi(P)$ 求得，而字符分布类型决定了 K 的取值，从而影响了 $\pi_m(P)$。因此，$\pi_b(P)$ 和 $\pi_m(P)$ 都反映了文本的字符分布特征，统称为 $\pi(P)$；参数 K 称为字符分布参数。

由模型 S1 和 S2，本文对模型公式可进一步演变，得到结论 S3。

结论 S3 综合模型 S1 和 S2，可得本文假设下的期望模型 $E(\Omega)$ 如下：

$$E(\Omega)=nD\pi(P)=nK\psi(p_1)\prod_{j=1}^{m-1}(\mathrm{gap}(j)\psi(p_{j+1})) \tag{4}$$

其中：K 为字符分布参数，n 为文本长度，m 为模式模长，$\mathrm{gap}(j)$ 为模式中通配符跨度，$\psi(p_j)$ 为 p_j 在文本中的出现频率。

由式(4)可知，P 的匹配数 Ω 不仅依赖于模式通配符特征 D，而且依赖于文本特征：由文本字符分布所决定的模式出现概率 $\pi(P)$；文本长度 n。$\pi(P)$ 和 D 相互作用，进一步扩展为 K、$\psi(p_j)$、通配符跨度 $\mathrm{gap}(j)$ 和通配符个数 $m-1$ 的表达式，具体反映了文本特征对 $E(\Omega)$ 的影响，详见 3.3 节性质分析。

3.3 期望性质分析

由模型 S1、S2 及结论 S3 得到如下性质：

性质 1 期望模型 $E(\Omega)$ 的时间复杂度为 $O(n+m)$，其中 n 是文本长度，m 是模式长度。

说明 由 3.2 节分析，期望模型通过扫描一次文本 T 得到文本的字符分布特征，需要时间 $O(n)$，再运用式(4)预测 Ω，需要 $O(m)$ 时间。

性质 2 模式匹配数 Ω 随着文本长度 n 的增加成线性增长。

说明 由式(4)可得此结论，实验部分将进一步验证。

性质 3 Ω 的增长受到文本字符分布 $\pi(P)$ 和模式特征 D 的共同影响，其增长不一定呈指数级。

说明 若不考虑文本字符分布特征 $\pi(P)$，仅考虑模式特征 $D=\prod_{j=1}^{m-1}\mathrm{gap}(j)$，则有匹配数 $\Omega\sim O(\mathrm{gap}(j)^m)$，即随着 $\mathrm{gap}(j)$ 的数目增加 Ω 成指数级增长[7]。然而，$\pi(P)$ 作为模型 $E(\Omega)$ 的重要元素，在以往研究中往往被忽略，影响 $\pi(P)$ 的参数为字符分布参数 K 和模式字符 p_j 在文本中出现的频率 $\psi(p_j)$。一方面，$\pi(P)$ 体现了文本字符的分布特征。参数 $\psi(p_j)$ 能够限制模式通配符特征 D 中 $\mathrm{gap}(j)$ 的作用。由式(4)看出，决定 $E(\Omega)$ 增长的主要部分不再是以往分析中的 $\prod_{j=1}^{m-1}\mathrm{gap}(j)$，而是 $\prod_{j=1}^{m-1}(\mathrm{gap}(j)\psi(p_{j+1}))$。因此字符分布特性使得 $\pi(P)$ 限制 Ω 的增长。本文可得结论，在对 PMWL 已有分析中[7]，认为 Ω 随 $\mathrm{gap}(j)$ 的个数成指数增长是不确切的，上述分

析忽略了文本特征。

性质 4　参数 K 反映了文本模型的不确定程度,体现了文本的内在信息。

K 中 $\overline{\kappa}_j^i$ 反映了模式 P 中字符 p_j 的出现概率 $Pr(p_j)$ 与 p_j 的出现频率 $\psi(p_j)$ 间的偏差,K 度量了 $\pi(P)=\prod_{j=1}^{m}Pr(p_j)$ 与 $\psi(P)=\prod_{j=1}^{m}\psi(p_j)$ 的偏差。给定 (P,T),$\psi(P)$ 为定值,而 $\pi(P)$ 由字符分布类型所决定,是不确定的值,K 描述了 $\pi(P)$ 相对于定值 $\psi(P)$ 的不确定程度。

4　实验

针对 PMWL 问题,现有多种算法有效获取完备匹配数 Ω[5,7],此类算法在获取 Ω 的值上具有一致性。这其中最有代表性的是 PAIG[7],其时间复杂度是 $O(nm^2\times \mathrm{gap}^2)$,本文 4.3 节分析了期望模型 $E(\Omega)$ 的时间复杂度为 $O(n+m)$,其中 n 是文本长度,m 是模式长度,gap 是通配符跨度。由此可知,期望模型可以将时间复杂度从多项式级降为线性。本章将通过实验比较期望模型与 PAIG 算法的匹配数 Ω,验证期望模型的正确性。实验使用了随机文本和真实 DNA 序列文本,DNA 序列下载自美国国家生物计算信息中心(http://www.ncbi.nlm.nih.gov/)。期望模型方面,将运用式(4)计算 $E(\Omega)$。实验不仅验证了期望模型 $E(\Omega)$ 在真实应用环境中有良好性能,还分析了模式长度、通配符跨度等参数对 PMWL 问题中 Ω 的影响。

4.1　模型性能分析

给定字母表 $\Sigma=\{a,c,g,t\}$,生成 Σ 上长为 120 000 的随机文本;真实文本采用 DNA 序列 AC004675.1。所有文本信息如表 1 中所示。

表 1　实验文本构成

文本 T	长度 n	a	c	g	t
随机文本	120 000	30 090	29 930	29 855	30 125
AC004675.1	128 501	34 603	28 343	29 762	35 793

4.1.1　随机文本上模型误差

按照 $m=\{3,4,5\}$ 生成三组模式,每组中 $P_k=p_1\eta_1[0,\mathrm{gap}-1]\cdots p_{m-1}\eta_{m-1}[0,\mathrm{gap}-1]p_m$,对每个 P_k 按照 gap = 10,15,20,…,35 生成全部 $M=|\Sigma|^m=4^m$ 个模式。计算每个模式的误差率 $\varepsilon_k=|E(\Omega_k)-\Omega_k|/\Omega_k$,总平均误差率 $\overline{\varepsilon}=\sum_{K=1}^{M}\varepsilon_k/M$。因此,得到 $m=3$、4、5 时,全体模式在不同 gap 取值时的总平均误差率,实验结果如表 2 所示。由第 3 章模型可知,Ω 的期望模型 $E(\Omega)$ 度量了全体输入实例 (P,T) 的 Ω 的平均值。本实验验证了该结论,总平均误差率 $\overline{\varepsilon}$ 为 1.8% ~3.2%。

表 2　随机文本上模型误差率 $\overline{\varepsilon}$ 实验

m	gap					
	10	15	20	25	30	35
$m=3$	0.019	0.021	0.019	0.019	0.019	0.019
$m=4$	0.026	0.031	0.031	0.031	0.030	0.030
$m=5$	0.043	0.036	0.033	0.033	0.033	0.033

4.1.2　真实文本上模型误差

对于真实文本,设置模型 $E(\Omega)$ 中参数 $K=1$(K 的设置相关将在 4.3 节中进一步说明)。模式设置与随机文本上实验模式相同。实验结果如表 3 所示。由实验结果可知,模型在真实文本上的误差率为 4.7% ~7.9%,因此对于真实文本,$E(\Omega)$ 仍然很好地度量了 Ω 的规模。并且 gap 取值越大时,期望模型的准确率越高;gap 值相同时,$m=3$ 的模式组平均误差率小于 $m=4$ 的模式组平均误差率。

表 3　真实文本上模型误差率 $\overline{\varepsilon}$ 实验

m	gap					
	10	15	20	25	30	35
$m=3$	0.078	0.065	0.058	0.054	0.051	0.048
$m=4$	0.093	0.087	0.077	0.075	0.073	0.070
$m=5$	0.109	0.106	0.096	0.094	0.089	0.079

4.2　期望模型 E(Ω)与模式特征 D、出现概率 π(P)

由式(4)可知,$E(\Omega)=nK\psi(p_1)\prod_{j=1}^{m-1}(\mathrm{gap}(j)\psi(p_{j+1}))$。其中,模式特征 $D=\prod_{j=1}^{m-1}\mathrm{gap}(j)$ 使得模式匹配数 Ω 与模式模长 m 有成指数增长的趋势。然而,模式概率 $\pi(P)$ 会限制 Ω 的增长。本节实验将验证特征参数 D 和 $\pi(P)$ 对模式匹配数 Ω 的影响是否与模型 $E(\Omega)$ 预测值吻合。

由于本节将验证 $\pi(P)$ 中参数 m 和参数 $\psi(p_j)$ 对通配符特征 D 的限制作用,所以设置模式中 $\mathrm{gap}(j)$ 均相同,记为 gap。实验文本采用 DNA 序列 AB008226.1,按照 gap = 4 和 5 随机生成两组模式,如表 4、5 所示。表中从上往下,模式模长 m 依次增加。实验对比了三种结果:随着模式模长 m 的增加,传统分析预测模式的匹配数;本文期望模型预测的模式匹配数;模式的实际匹配数。其中,前两种结果由公式直接计算,无须匹配文本,而最后一种结果需要使用匹配算法,本文采用 PAIG 算法[7]。实验结果如图 1、2 所示,其中第一种的计算结果对应左侧 Y 坐标轴,期望模型计算结果和算法求解得到的模式实际匹配数对应右侧 Y 坐标轴。

表 4　模式组 1 组成

模式模长 m	模式 $P1$(gap = 4)
2	$g\eta_1[0,4]g$
3	$g\eta_1[0,4]g\eta_2[0,4]g$
4	$g\eta_1[0,4]g\eta_2[0,4]g\eta_3[0,4]g$
5	$g\eta_1[0,4]g\eta_2[0,4]g\eta_3[0,4]g\eta_4[0,4]g$
6	$g\eta_1[0,4]g\eta_2[0,4]g\eta_3[0,4]g\eta_4[0,4]g\eta_5[0,4]g$

表 5　模式组 2 组成

模式模长 m	模式 $P2$(gap = 5)
2	$c\eta_1[0,5]g$
3	$c\eta_1[0,5]g\eta_2[0,5]c$
4	$c\eta_1[0,5]g\eta_2[0,5]c\eta_3[0,5]a$
5	$c\eta_1[0,5]g\eta_2[0,5]c\eta_3[0,5]a\eta_4[0,5]t$
6	$c\eta_1[0,5]g\eta_2[0,5]c\eta_3[0,5]a\eta_4[0,5]t\eta_5[0,5]c$

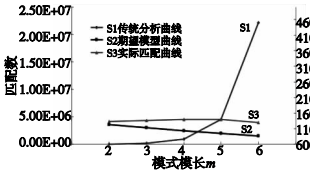

图1　$E(\Omega)$、PAIG 与传统分析的匹配数对照(P_1)

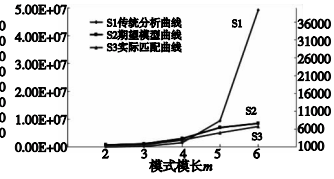

图2　$E(\Omega)$、PAIG 与传统分析的匹配数对照(P_2)

可以看出,本文期望模型得到的变化曲线 S2 与模式实际匹配次数随模长 m 的变化曲线 S3 相吻合,而 S3 并不符合传统模型得到的变化曲线 S1 的变化趋势。这是因为,期望模型基于文本和模式特征,理论上包含了抑制指数增长态势的因子,从而可以较为准确地估计模式的实际增长趋势。

对于表 5 中的模式 P_2,根据 gap = {4,6,7},生成三组期望

模型的变化曲线 S2 和模式实际匹配次数的变化曲线 S3，如图 3 ~ 5 所示。

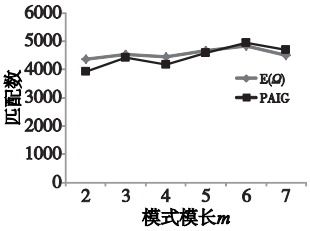

图3　$E(\Omega)$与PAIG匹配数对照(gap=4)

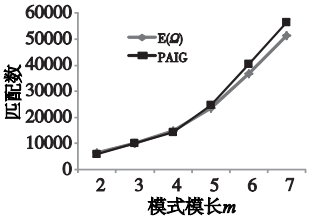

图4　$E(\Omega)$与PAIG匹配数对照(gap=6)

对比期望模型 $E(\Omega)$ 和模式实际匹配次数 Ω 的变化曲线可以看出，期望模型 $E(\Omega)$ 准确地描述了模式实际匹配数 Ω 的变化情况，从而验证了本文模型的正确性和有效性。给定模式 P 和文本 T，本文期望模型可以准确预测 Ω 的增长方式，为问题的分析和求解提供参考。此外，相对于传统分析，本文期望模型准确揭示了文本字符分布对问题求解的影响。由于模式概率 $\pi(P)$ 的作用，使得只有当 $\mathrm{gap} > 1/\sqrt[m-1]{\prod_{j=2}^{m}\psi(p_j)}$ 时，才能使得 Ω 成指数级增长；否则，Ω 的增长规模可能保持不变或者下降。

4.3　期望模型 $E(\Omega)$ 与文本参数 K

由模型性质 4 可知，K 能度量模式概率 $\pi(P)$ 与频率 $\psi(P)$ 的偏差，反映文本模型的不确定度。由 4.1.2 节实验可知，对于真实文本，设置模型参数 $K=1$，模型可以有效求解 gap(j) 较大且文本较长的实例，并且误差率随着模式中 gap(j) 值的增加而逐渐趋近于 1。因此，本节实验进一步对参数 K 的稳定性进行验证。

1）K 随文本长度 n 的变化趋势　采用 DNA 序列 AB008226.1，长度为 7350。由模型理论得 $K=\Omega/E(\Omega)$。随机给定两组模式 $P_1 = a\eta_1[0,4]g\eta_2[0,4]t\eta_3[0,4]c\eta_4[0,4]a$，$P_2 = a\eta_1[0,3]t\eta_2[0,4]g\eta_3[0,5]c$。实验得到两组 K 随 n 变化的曲线，如图 6 所示。观察实验结果可知，随着 n 的增加，K 逐渐增加，并逐渐趋近于稳定值。实验结果表明 K 在真实文本环境中趋近于定值。这个稳定值反映了文本的不确定程度，是文本的内在特征。

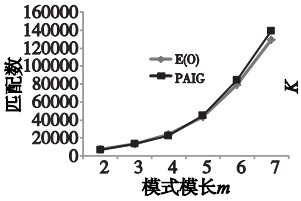

图5　$E(\Omega)$与PAIG匹配数对照(gap=7)

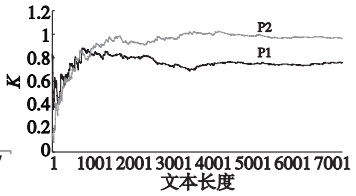

图6　模式P_1和P_2中参数K随文本长度n的变化趋势

2）K 随模式中通配符跨度 gap 的变化趋势　采用 DNA 序列 AC004675.1，给定模式 $P_3 = a\eta_1[0,\mathrm{gap}-1]a\eta_2[0,\mathrm{gap}-1]c\eta_3[0,\mathrm{gap}-1]g\eta_4[0,\mathrm{gap}-1]t$ 和 $P_4 = t\eta_1[0,\mathrm{gap}-1]g\eta_2[0,\mathrm{gap}-1]c\eta_3[0,\mathrm{gap}-1]a\eta_4[0,\mathrm{gap}-1]a$，结果如图 7、8 所示。实验结果表明，随着模式中 gap 值的增大，期望模型预测的匹配数与真实匹配数的误差比率下降，$E(\Omega)$ 中参数 K 也逐渐增大并趋近于 1。

因此在实际应用环境中，可以使用模型 $E(\Omega)$ 准确得到 Ω 的数量级。针对模式中通配符跨度 gap(j) 较大以及文本长度较长时算法求解代价巨大的情况，还可以采用 $K=1$ 的期望模型 $E(\Omega)$ 得到 Ω 的近似值。此外，利用 K 趋于稳定值的特性，可以扩展利用已有的结果评估某类问题的 K 作为特征参数，为以后该类文本的挖掘、检测等提供参考。

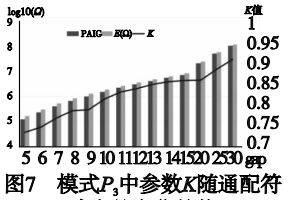

图7　模式P_3中参数K随通配符跨度的变化趋势

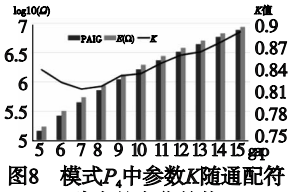

图8　模式P_4中参数K随通配符跨度的变化趋势

4.4　期望模型 $E(\Omega)$ 与文本长度 n

期望模型可以写成 $E(\Omega) = Mn$，其中 $M = K\psi(p_1) \times \prod_{j=1}^{m-1}(\mathrm{gap}(j)\psi(p_{j+1}))$，因此 $E(\Omega)$ 随着文本长度 n 的增加近似成线性增长。本节验证了参数 n 对模式出现次数 Ω 的影响。

实验中文本采用 Σ 上长度 n 为 13000 的随机文本 T 和 DNA 序列 AB038490，长为 131892。实验模式采用模式组如表 6 所示。其中模式为随机给定，模式字符组成相对复杂，约束条件较为灵活。实验结果如图 9、10 所示。实验表明，Ω 与 n 呈线性增长趋势。其中，图 9 中 Ω 与 n 表现出了相对精确的线性关系；而图 10 中 Ω 与 n 的线性关系在局部存在波动。该现象符合模型理论分析。真实文本参数较为复杂，M 表达式中的参数 K 随文本长度 n 和模式通配符跨度的变化而改变，因此系数 M 受到了 $\pi(P)$ 和 D 的波动影响，使得 Ω 与 n 并不是完全精确的线性关系。总体而言，相对于已有分析，模型 $E(\Omega)$ 反映了真实文本的参数复杂性，更为精确地揭示了文本长度 n 在求解 Ω 中的作用。

表 6　模式组组成

序号	模式
P_1	$a\eta_1[0,3]c\eta_2[0,4]c\eta_3[0,5]g\eta_4[0,6]t$
P_2	$g\eta_1[0,6]g\eta_2[0,3]c\eta_3[0,5]a\eta_4[0,6]t$
P_3	$a\eta_1[0,6]c\eta_2[0,3]g\eta_3[0,5]t\eta_4[0,4]t\eta_5[0,5]$
P_4	$a\eta_6[0,5]g\eta_7[0,6]c$
	$c\eta_1[0,3]c\eta_2[0,4]a\eta_3[0,5]a\eta_4[0,13]t\eta_5[0,4]t$

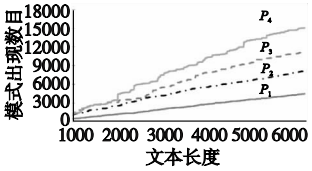

图9　随机文本中参数K随文本长度的变化趋势

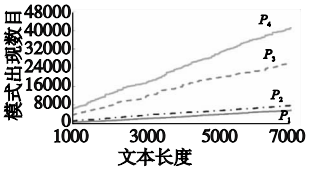

图10　DNA序列中参数K随文本长度的变化趋势

实验表明，Ω 与 n 呈线性增长趋势。其中，图 9 和 10 中 Ω 与 n 表现出了相对精确的线性关系。而图 10 中 Ω 与 n 的线性关系在局部存在波动。该现象符合模型理论分析。真实文本参数较为复杂，M 表达式中的参数 K 随文本长度 n 和模式通配符跨度的变化而改变，因此系数 M 受到了 $\pi(P)$ 和 D 的波动影响，使得 Ω 与 n 并不是完全精确的线性关系。总体而言，相对于已有分析，模型 $E(\Omega)$ 反映了真实文本的参数复杂性，更为精确地揭示了文本长度 n 在求解 Ω 中的作用。

5　结束语

对带有通配符和长度约束的模式匹配问题，本文研究了模式 P 在文本 T 中匹配数 Ω 的计数模型及其性质。基于文本字符分布模型，得到了 Ω 的期望 $E(\Omega) = Dn\pi(P)$，度量了 Ω 的实际规模。其中，$\pi(P)$ 是由字符分布所决定的模式出现概率，D 为模式中各通配符跨度 gap(j) 的乘积，n 为文本长度。本文

研究了在不同文本类型中 $\pi(P)$ 的表达式,并在理论(性质 3)和实验上揭示了在不同文本字符分布影响下,随着模式特征的改变,Ω 不一定呈指数级增长,这有别于已有工作的结论,为今后 PMWL 的复杂性研究提供了依据。在实验部分,本文在随机文本和真实文本上分别验证了期望模型 $E(\Omega)$ 的精确度,说明 $E(\Omega)$ 能够有效估计模式在文本中的出现次数。另外,本文还得到了期望模型中的文本参数 K 的相关性质,进而讨论了模型在真实环境中的应用能力。结果表明,参数 K 具有稳定性,能够评估文本的不确定程度。此外,$E(\Omega)$ 作为一个新的衡量标准,能够在线性时间内评估 Ω 的规模,从而有助于 PMWL 问题在大数据集上的求解。在下一步工作中,将考虑把本文的模型应用于大文本数据中的序列模式挖掘。

参考文献:

[1] Navarro G, Raffinot M. Flexible pattern matching in strings: practical on-line search algorithms for texts and biological sequences [M]. Cambridge: Cambridge University Press, 2002.

[2] Han Jiawei, Cheng Hong, Wu Xindong, *et al*. Frequent pattern mining: current status and future directions [J]. Data Mining and Knowledge Discovery, 2007, 15(1): 55-86.

[3] Fischer M J, Paterson M S. String matching and other products [M]// Karp R M. Complexity of Computation. Cambridge: MIT Press, 1974: 113-125.

[4] Bille P, Gortz I L, Vildhoj H W, *et al*. String matching with variable length gaps [C]//Proc of the 17th International Symposium on String Processing and Information Retrieval. Berlin: Springer, 2010: 385-394.

[5] Navarro G, Raffinot M. Fast and simple character classes and bounded gaps pattern matching, with applications to protein searching [J]. Computational Biology, 2003, 10(6): 903-923.

[6] Chen Gong, Wu Xindong, Zhu Xingquan, *et al*. Efficient string matching with wildcards and length constraints [J]. Knowledge and Information Systems, 2006, 10(4): 399-419.

[7] Min Fan, Wu Xindong, Lu Zhenyu. Pattern matching with independent wildcard gaps [C] //Proc of the 8th IEEE International Conference on Dependable, Autonomic and Secure Computing. 2009: 194-199.

[8] Zhang Minghua, Ben Kao, Cheung D W, *et al*. Mining periodic patterns with gap requirement from sequences [C]//Proc of ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2005: 622-633.

[9] Zhu Xingquan, Wu Xindong. Mining complex patterns across sequences with gap requirements [C]//Proc of International Joint Conference on Artificial Intelligence. 2007: 2934-2940.

[10] Guo Dan, Hong Xiaoli, Hu Xuegang, *et al*. A bit-parallel algorithm for sequential pattern matching with wildcards [J]. Cybernetics and Systems, 2011, 42(6): 382-401.

[11] 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.

[12] Guo Dan, Hu Xuegang, Xie Fei, *et al*. Pattern matching with wildcards and gap-length constraints based on a centrality-degree graph [J]. Applied Intelligence, 2012, 39(1): 57-74.

[13] Knuth D E, Morris J H, Pratt V R. Fast pattern matching in strings [J]. SIAM Journal on Computing, 1977, 6(2): 323-350.

[14] Baeza-Yates R A. String searching algorithms revisited [C]//Lecture Notes in Computer Science, vol 382. 1989: 75-96.

[15] Yao A C. The complexity of pattern matching for a random string [J]. SIAM Journal on Computing, 1979, 8(3): 368-387.

[16] Marschall T, Rahmann S. An algorithm to compute the character access count distribution for pattern matching algorithms [J]. Algorithms, 2011, 4(4): 285-306.

[17] 刘萍,刘燕兵,郭莉,等. 串匹配算法中模式串与文本之间关系的研究[J]. 软件学报, 2010, 21(7): 1503-1514.

[18] Bassino F, Clement J, Nicodeme P. Counting occurrences for a finite set of words: combinatorial methods [J]. ACM Trans on Algorithms, 2012, 8(3): 1-28.

[19] Flajolet I, Pitoura E, Szpankowski W. On pattern occurrences in a random text [J]. Information Processing Letters, 1996, 57(6): 307-312.

[20] Li Zheng, Ge Tianjian. Online windowed subsequence matching over probabilistic sequences [C]//Proc of ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2012: 277-228.

[21] Bourdon J, Vallee B. Pattern matching statistics on correlated sources [C]//Proc of the 7th Latin American Symposium on Theoretical Informatics. Berlin: Springer, 2006: 224-237.