基于非完全吸收马尔科夫链的多文档自动文摘算法

高 晶 房 俊
(北方工业大学云计算研究中心 北京 100144)

摘 要 吸收马尔科夫链模型在自动文摘领域的有效性已经证实。然而,此模型中的平均期望历经次数需要通过矩阵求逆得到,所以模型的时间复杂度很高。此外,由于自身的局限性,它也无法利用除句子间相互关系以外的其它信息。针对此问题建立了一个新的模型:非完全吸收马尔科夫链;并以此为基础提出了一个新的多文档文摘算法。证明了吸收马尔科夫链的平均期望历经次数与对应的非完全吸收马尔科夫链的稳态概率分布的等价性,而后者可通过迭代求解。同时,这个新的模型还可以引入除句子间相互关系以外的其它信息,从而生成更准确的文摘。在TAC2011上的实验证实了该模型的有效性。

关键词 非完全吸收马尔科夫链,LexRank,面向主题的先验分布,多文档自动文摘
中图法分类号 TP391 **文献标识码** A

Partial Absorbing Markov Chain Based Multi-document Summarization

GAO Jing FANG Jun
(Research Center for Cloud Computing, North China University of Technology, Beijing 100144, China)

Abstract Absorbing Markov Chain has been proven to be effective in text summarization. However, the algorithm based on Absorbing Markov Chain is not only time-consuming due to matrix inversion but also inept to integrate other information except relationships among sentences because of the limitation of the model. This paper presents a novel multi-document summarization approach based on Partial Absorbing Markov Chain. The equivalent relationship between the average expected visits in Absorbing Markov Chain and the stationary probability in the corresponding Partial Absorbing Markov Chain was demonstrated. Then, the stationary probability in Partial Absorbing Markov Chain which is easily calculated serves as a criterion to rank sentences. In addition, other kinds of information are incorporated together to generate a more accurate solution of the stationary probability. Experiments on TAC2011 main task are performed.

Keywords Partial absorbing markov chain, LexRank, Topic-oriented prior distribution, Multi-document summarization

1 介绍

近年来,自动多文档文摘在文档管理与检索系统中展现出了很强的实用性,吸引了众多研究者的关注。多文档文摘能够简洁地描述文档集的内容,便于用户理解。

至今,由于文本生成技术[1]不够成熟,很多文摘系统还是基于抽取式文摘框架。它们抽取包含文档集重要内容并与指定话题相关的句子构成文摘。而面向话题的多文档文摘研究存在两个主要困难:首先是对不同文档中的内容进行整合,从而得到既完整又冗余的文摘;其次是有目的地选取与用户查询有关的内容。

为了解决上述问题,本文建立了一个新的模型——非完全吸收马尔科夫链模型(PAMC),并在此基础上提出了一个可以融合多种信息的文摘算法。在这个算法中,已经入选到文摘中的句子被标记为非完全吸收状态,然后对所有句子建立PAMC模型,最后通过LexRank[2]计算剩余句子成为文摘句的可能性。为了考虑更多的信息,本文定义了查询相关的先验分布,并将其引入到LexRank中。

2 相关工作

研究者们提出了一些方法来解决上述面向用户查询的多文档文摘中存在的两个问题。

一方面,为了去除文摘中的冗余内容,Carbonell和Goldstein[3]提出了Maximal Marginal Relevance (MMR)算法;其他的还有基于句子互信息[4]、子主题新颖性[5]、新颖度惩罚[6]、子主题划分[7]等。这些算法的基本思想是如果某个句子与已入选的文摘句内容相似,则对它进行冗余度惩罚。然而,这些算法把重要性和新颖性的计算过程分裂开来,忽略了二者之间的联系。GrassHopper[8]通过引入吸收马尔科夫链,可以以一种统一的方式来考虑句子的重要性和新颖性,并通过实验证实了自身的有效性。但是,GrassHopper由于需要进行矩阵求逆的计算,因此时间复杂度很高。此外,它也无法考虑除句子间相互关系以外的其他有用的信息。

另一方面,为了生成面向用户查询的文摘,NeATS[9]使用查询关键词、句子位置和词频等信息辅助抽取重要内容;Harabagiu和Lacatusu[10]提出了一种基于查询主题的新的查

到稿日期:2012-10-29 返修日期:2013-03-02 本文受国家自然基金重点项目(61033006),国家自然基金项目(60970131)资助。
高 晶(1982—),女,硕士,助理研究员,CCF会员,主要研究方向为云计算、实时数据处理、文本挖掘等,E-mail:gaojing@ncut.edu.cn;房 俊(1976—),男,博士,副研究员,主要研究方向为服务计算、云数据管理、文本挖掘等。

询表示方法；Saggion[11]直接计算每个句子与查询的相似性；CLASSY[12]利用了查询中的词与命名实体。实际上，有很多信息可以应用于文摘生成，如基于词的特征、句子全局特征（位置、长度、类型等）、句子间相互关系及句子与用户查询的相关性[7,13]。就本文所知，上述算法并不能将这些信息很好地整合起来。

3 吸收马尔科夫链

吸收马尔科夫链是一种马尔科夫链：它至少包含一个吸收状态——到自身的转移概率为1的状态。而非吸收状态称为瞬时状态。在一个吸收马尔科夫链中，从任一一个瞬时状态出发，经过有限步跳跃（状态转移），都可到达某个吸收状态。吸收马尔科夫链的转移概率矩阵表示为：

$$P=\begin{bmatrix} R & Q \\ I & 0 \end{bmatrix} \tag{1}$$

式中，I 是一个 $r*r$ 单位阵，Q 是一个 $(n-r)*(n-r)$ 方阵（$n-r$ 表示瞬时状态的个数），0 是各元素全为 0 的矩阵，R 是 $(n-r)*r$ 矩阵。Q 定义了各瞬时状态间的转移概率，而 R 指定了瞬时状态到吸收状态的转移概率。

引理 1[8] 在一条吸收马尔科夫链中，任意一次随机行走将中止（进入某个吸收状态）的概率是 1（即 $Q^n\to 0$ as $n\to\infty$，Q^n 表示各瞬时状态间的 n 步转移概率）。

引理 2[8] 吸收马尔科夫链满足以下性质：

1. $(I-Q)$ 存在逆矩阵 N 且 $N=I+Q+Q^2+\cdots$；

2. N 的第 i 行 j 列的元素表示从瞬时状态 s_i 出发的随机行走历经瞬时状态 s_j 的期望次数。

基于引理 2，对每个瞬时状态 s_i 的所有期望历经次数求取平均值，用矩阵表示为：

$$v^{\mathrm{T}}=\frac{e^{\mathrm{T}}N}{n-r} \tag{2}$$

式中，e 表示所有元素全为 1 的列向量。

从直观上看，一方面，那些与吸收状态紧密相连的瞬时状态的平均期望历经次数比较小，因为如果某次随机行走经过这些瞬时状态后，将更易于转移到吸收状态中而终止行走。反之，远离吸收状态的瞬时状态将享有更多机会被历经。因此，平均期望历经次数可以反映瞬时状态与吸收状态的差异性。另一方面，对于与其他瞬时状态关系紧密的瞬时状态，在进入吸收状态前，由于任意一次随机行走会使它更容易转移，因此它同样可以获得更多的被历经的机会。从而，平均期望历经次数也反映了瞬时状态间的关系。

4 非完全吸收马尔科夫链

GrassHopper[8]将平均期望历经次数这个概念引入自动文摘领域，作为一个文摘句选取的依据，并用实验证实了它的有效性。然而，从式(2)可以看出，此方法要求 $(I-Q)$ 计算的逆矩阵 N，时间复杂度很高；此外，N 只依赖于 Q，或更明确地说，N 仅取决于句子间的相互关系。这表明了吸收马尔科夫链的局限性。为此，本文提出了一个新的模型——非完全吸收马尔科夫链，其中任意一个吸收状态到所有瞬时状态都有一个相同的转移概率。相应地，非完全吸收马尔科夫链的转移概率矩阵表示为：

$$\hat{P}=\begin{bmatrix} R & Q \\ \alpha I & \dfrac{1-\alpha}{n-r}U \end{bmatrix} \tag{3}$$

式中，I 表示一个 $r*r$ 单位阵；α $(0\leqslant\alpha\leqslant 1)$ 称为吸收因子，表示每个吸收状态到自身的转移概率；U 表示一个所有元素全为 1 的 $r*(n-r)$ 矩阵，将每个吸收状态到任意一个瞬时状态的转移概率设置为 $(1-\alpha)/(n-r)$；R 和 Q 与式(1)中的含义相同。

下面本文将证明，非完全吸收马尔科夫链的稳态概率等价于吸收马尔科夫链的平均期望历经次数。

定理 1 假设一个吸收马尔科夫链 C 有 n 个状态，其中 r 个是吸收状态。对于某个瞬时状态 s_j，v_j 表示它的平均期望历经次数。$\hat{C}$ 表示与 C 相对应的非完全吸收马尔科夫链，p_j 表示 s_j 在 $\hat{C}$ 中的稳态概率。则：

$$p^{\mathrm{T}}=\frac{v^{\mathrm{T}}}{\dfrac{1}{1-\alpha}+v^{\mathrm{T}}e_{n-r}} \tag{4}$$

式中，v 是一个 $(n-r)$ 维列向量，它的第 i 个元素表示状态 s_i 的平均期望历经次数；p 也是一个 $(n-r)$ 维列向量，它的第 i 个元素表示状态 s_i 的稳态概率；e_{n-r} 是一个元素全为 1 的 $(n-r)$ 维列向量；α 是吸收因子。

证明：令 p_{abs} 表示一个 r 维列向量，它的每个元素对应一个吸收状态的稳态概率；$\hat{P}$ 表示非完全吸收马尔科夫链的转移概率矩阵，则：

$$[p^{\mathrm{T}}\quad p_{abs}^{\mathrm{T}}]=[p^{\mathrm{T}}\quad p_{abs}^{\mathrm{T}}]\hat{P} \tag{5}$$

且

$$p_{abs}^{\mathrm{T}}e_r+p^{\mathrm{T}}e_{n-r}=1 \tag{6}$$

式中，e_r 是一个元素全为 1 的 r 维列向量。将式(3)引入到式(5)，可以得到：

$$p_{abs}^{\mathrm{T}}=\alpha p_{abs}^{\mathrm{T}}+p^{\mathrm{T}}R \tag{7}$$

$$p^{\mathrm{T}}=\frac{1-\alpha}{n-r}p_{abs}^{\mathrm{T}}U+p^{\mathrm{T}}Q \tag{8}$$

如果 α 小于 1，由式(7)得：

$$p_{abs}^{\mathrm{T}}=\frac{p^{\mathrm{T}}R}{1-\alpha} \tag{9}$$

将式(9)代入式(8)得：

$$p^{\mathrm{T}}=p^{\mathrm{T}}RU\frac{(I-Q)^{-1}}{n-r} \tag{10}$$

实际上，U 可以表示为 $U=e_re_{n-r}^{\mathrm{T}}$，而 $N=(I-Q)^{-1}$。则将式(2)引入式(10)得：

$$p^{\mathrm{T}}=(p^{\mathrm{T}}Re_r)\frac{e_{n-r}^{\mathrm{T}}N}{n-r}=(p^{\mathrm{T}}Re_r)v^{\mathrm{T}} \tag{11}$$

注意到，$p_r^{\mathrm{T}}Re_r$ 是一个标量，因此 p 与 v 成正比。将式(9)和式(11)代入式(6)得：

$$p^{\mathrm{T}}Re_r=\frac{1}{\dfrac{1}{1-\alpha}+v^{\mathrm{T}}e_{n-r}} \tag{12}$$

将式(12)代入式(11)得：

$$p^{\mathrm{T}}=\frac{v^{\mathrm{T}}}{\dfrac{1}{1-\alpha}+v^{\mathrm{T}}e_{n-r}} \tag{13}$$

上述结论在 $\alpha<1$ 时成立。实际上，如果 $\alpha=1$，即吸收状态到自身的转移概率为 1，则从式(12)得到 $p=0$。而这也是正确的，因为根据引理 1，在吸收马尔科夫链上的任意一次随机行走必将终止于某个吸收状态，则所有瞬时状态的稳态概率都为 0。经过简单的数学变换，可以得到：

$$v^{T}=\frac{p^{T}}{(1-\alpha)(1-p^{T}e_{n-r})} \tag{14}$$

即非完全吸收马尔科夫链的稳态概率等价于吸收马尔科夫链的平均期望历经次数。

而非完全吸收马尔科夫链的稳态概率可以通过 LexRank 方法迭代求解。PageRank[13] 提供了一种计算网页重要度的方法。Erkan 和 Radev 将 PageRank 引入到自动文摘领域，称为 LexRank[2]。LexRank 的复杂度等同于 PageRank。而 PageRank 算法的复杂度在最坏的情况下为 $O(k*n^2)$[14]。Haveliwala 对 PageRank 算法进行了复杂度分析，事实上，迭代次数 k 的最大值不会超过 100，即它的算法复杂度为 $O(100*n^2)$[15]，其中数字 100 指算法需要迭代 100 次才能达到可接受的等级值。而吸收马尔科夫链的算法复杂度取决于矩阵求逆的复杂度，即为 $O(n^3)$。由此可以得出，在一个较小的范围内（$n<100$），基于非完全吸收马尔科夫链的算法复杂度和基于吸收马尔科夫链的算法复杂度相当；而当 $n>100$ 时，则要优于基于吸收马尔科夫链的算法。在实际的多文档文摘中，矩阵元素 n 往往是远大于 100 的。

此外，本文将面向用户查询的先验概率引入到 LexRank 中，从而更准确地计算稳态概率。

5 面向用户查询的先验概率

LexRank 没有考虑除句子间相互关系以外的其它信息对句子的 LexRank 得分的影响。本文引入了面向用户查询的先验概率的概念，并在实验中证实了它的有效性。

给定一个文档集 D 和一个用户查询 q，D 中的某个句子 s 的用户查询的先验概率定义为 $p(s/q)(s\in D)$，它表示在 q 已知的条件下，s 入选为文摘句的可能性。由贝叶斯法则得：

$$p(s|q)=p(q|s)p(s)/p(q) \tag{15}$$

对任意一个句子 s，$p(q)$ 都相同，则式(15)可以表示为：

$$p(s|q)\propto p(q|s)p(s) \tag{16}$$

式中，$p(q/s)$ 表示根据 s 得到 q 的可能性；$p(s)$ 表示 s 在 D 中的重要性。因此，为了得到 $p(s/q)$，需要分别计算 $p(q/s)$（相关性模型）和 $p(s)$（重要性模型）。$p(s)$ 与查询无关，普遍适用于任何用户查询；而 $p(q/s)$ 要根据指定用户查询计算。

5.1 重要性模型 $p(s)$

本文利用句子全局特征，根据朴素贝叶斯模型计算 $p(s)$。

5.1.1 句子全局特征

本文中，句子全局特征主要考虑句子的长度、位置及类型，具体定义为以下 4 种：

(1) 段落特征 x_1：单一段落（一片文档仅有一段），首段，尾段，其它；

(2) 段内特征 x_2：单一语句（一个段落仅有一句），首句，尾句，其它；

(3) 语气特征 x_3：陈述句和引用句；

(4) 长度特征 x_4：本文划分了若干长度区间，每个区间对应一个特征值。

5.1.2 朴素贝叶斯模型

句子可以用上述 4 个全局特征形式化表示为 $s=\{x_1,x_2,x_3,x_4\}$，其中 $x_i(i=1,2,3,4)$ 是第 5.1.1 节中定义的全局特征。假设句子全局特征相互独立，根据朴素贝叶斯模型得：

$$\begin{aligned}p(s)&=p(s\in S|x_1,x_2,x_3,x_4)\\&=\frac{p(s\in S)p(x_1,x_2,x_3,x_4|s\in S)}{p(x_1;x_2,x_3,x_4)}\\&=\frac{p(s\in S)\prod_{i=1}^{4}p(x_i|s\in S)}{\prod_{i=1}^{4}p(x_i)}\end{aligned} \tag{17}$$

式中，S 表示文摘句的集合；$p(s\in S|x_1,x_2,x_3,x_4)$ 表示全局特征已知时，s 被选为文摘句的可能性；假设当任何信息都未知时，所有句子入选的可能性相同，即 $p(s\in S)$ 为常数；$p(x_i)$ 全局特征 i 取值为 x_i 的可能性，可以进行如下计算：

$$p(x_i)=\frac{n(x_i)}{n} \tag{18}$$

式中，n 和 $n(x_i)$ 分别表示文档集中的所有句子数目和其中全局特征 i 取值为 x_i 的句子数目。$p(x_i|s\in S)$ 表示在文摘句集合中，全局特征 i 取值为 x_i 的可能性，可以用类似方法计算：

$$p(x_i|s\in S)=\frac{n_S(x_i)}{n_S} \tag{19}$$

式中，n_S 和 $n_S(x_i)$ 分别表示文摘句集合中的所有句子数目和其中全局特征 i 取值为 x_i 的句子数目。

5.2 相关性模型 $p(q|s)$

一个文档集中既包含与某个用户查询直接相关的句子，也包含与它关系不密切的句子。如何对二者进行区别对于生成高质量的文摘有很大影响。在 TAC2011 数据集中，每个用户查询都包含一个标题和一段叙述。标题通常是一个与用户查询紧密相关的词组或短语；而叙述具体给出了用户关注的几个方面。本文利用用户查询中的命名实体来评价任意一个句子与它相关与否。已知一个文档集 D 和一个用户查询 q，$p(q|s)$ 可以形式化表示为：

$$p(q|s)=\begin{cases}\dfrac{1}{c}, & NE_q\cap NE_s=\varnothing\\[2ex]\dfrac{1+\ln\frac{L}{L_q+1}}{c}, & NE_q\cap NE_s\neq\varnothing\end{cases} \tag{20}$$

式中，N 是 D 中的句子总数，L_q 是包含 q 中某个命名实体的句子的个数；NE_q 是 q 中的命名实体的集合；NE_s 是 s 中的命名实体的集合；c 是一个足够大的常量。从式(20)中可以看出，如果 L_q 接近 L，则 $p(q|s)$ 近似为常数。注意到，$p(q|s)$ 与 c 无关。实际上，根据式(15)可以得到：

$$\begin{aligned}p(s|q)&=\frac{p(q|s)p(s)}{\sum_s p(q|s)p(s)}\\&=\frac{\left(I_{(NE_q\cap NE_s=\varnothing)}*1+I_{(NE_q\cap NE_s\neq\varnothing)}*(1+\ln\frac{L}{L_q})\right)p(s)}{\sum_s\left(I_{(NE_q\cap NE_s=\varnothing)}*1+I_{(NE_q\cap NE_s\neq\varnothing)}*(1+\ln\frac{L}{L_q})\right)p(s)}\end{aligned} \tag{21}$$

式中，$I(condition)$ 为判别函数。若 *condition* 为真，则 $I(condition)$ 取值为 1，反之为 0。从式(21)可以看出，$p(s|q)$ 与 c 无关。

这样，最终可得到面向用户查询的先验分布，然后可以利用 LexRank 对句子进行最后评分。

6 算法描述

已知一个文档集 $D=\{s_1,s_2,\cdots,s_n\}$ 和一个用户查询 q。$f:D\to[0,1]$ 表示一个评分函数，可以为 $s_i(1\leqslant i\leqslant n)$ 被选为

文摘句的可能性给一个分值 f_i，分值越高，意味着越可能入选。f 可以表示为向量形式 $f=[f_1,f_2,\cdots,f_n]^T$。面向用户查询的先验分布 $r=(r_1,r_2,\cdots,r_n)^T(r_i=p(s_i|t))$已经计算完毕。$S$ 表示已入选的文摘句的集合，初始设为空集。则算法描述如下：

(1)计算所有句子的面向用户查询的先验概率 $r=(r_1,r_2,\cdots,r_n)^T(r_i=p(s_i|q))$；

(2)计算句子间的余弦相似度，句子的词权重用 $tf*isf$ 方式，其中，tf 为句子中的词频，isf 定义为 $1+\log(n/n_t)$，n 是句子总数，n_t 是包含词 t 的句子数；

(3)计算转移概率矩阵 P：$P_{ij}=\frac{\mathrm{sim}(s_i,s_j)}{\sum_{j=1}^{n}\mathrm{sim}(s_i,s_j)}$，其中 $\mathrm{sim}(s_i,s_j)$表示 s_i 和 s_j 的余弦相似度；

(4)将 S 中的所有句子定义为非完全吸收状态，相应地将 P 改为$\hat{P}=\begin{bmatrix}R & Q\\ \alpha I & \frac{1-\alpha}{n-r}U\end{bmatrix}$；

(5)利用 LexRank 算法对 $f(t+1)=\lambda r+(1-c)\hat{P}^T f(t)$ 迭代求解，其中 λ 是置信度因子，定义区间为[0,1]。如果 λ 设为 0，则 r 将不被考虑；如果 λ 设为 1，则 f 等于 r。λ 体现了我们对 r 的信任程度；

(6)选择分值最高的句子加入到 S 中；

(7)如果文摘长度达到要求则停止，否则转到(4)。

7 实验

7.1 数据集

TAC2011 是面向用户查询的多文档文摘。它包含 44 个文档集，其中每个文档集包含 20 篇文档和 4 篇作为标准答案用于评测的人工文摘。本文使用 TAC2011 的主任务作为本算法的评测集。在预处理步骤中，我们做了分句，去除了停用词并作了词性还原。

7.2 评测方法

TAC 使用 ROUGE 工具进行机器文摘质量的自动评测。ROUGE 通过计算机器文摘与人工文摘间的 n-gram、词序列和词对的重叠度进行评价。其中两个最重要的评价指标分别为 ROUGE-2 和 ROUGE-SU4。本文使用这两个指标进行评价，并将其与他方法进行对比。

7.3 实验结果

7.3.1 模块分析

我们的系统由 3 个独立模块构成：

(1)重要度模型 $p(s)$：在这个模型中，用句子全局特征计算 $p(s)$。如果不计算 $p(q|s)$，仅用 $p(s)$代替 $p(s|q)$ 作为面向用户查询的先验概率，且不引入非完全吸收马尔科夫链，评测结果为：ROUGE-2 得分 0.10887，ROUGE-SU4 得分 0.14770，其中 λ 设为 0.3；

(2)相关性模型 $p(q|s)$：在这个模型中，利用用户查询中的命名实体计算句子与查询的相关性。如果同时使用 $p(q|s)$和 $p(s)$，但不引入非完全吸收马尔科夫链，评测结果改进为：ROUGE-2 得分 0.11277，ROUGE-SU4 得分0.15190，其中 λ 设为 0.3；

(3)非完全吸收马尔科夫链模型：这个模型可以将句子的重要性和它与用户查询的相关性作为一个整体进行计算。加入这个模型，得到了最好的结果：ROUGE-2 得分0.13011，ROUGE-SU4 得分 0.16171，其中 λ 设为 0.3。

结果显示随着新的模型的引入，我们的测试结果也越来越好。

7.3.2 系统比较

在 TAC2011 主任务中，将 TAC2011 中得分最高的系统，得分评价为很好的两个系统以及 TAC 的两个 BASELINE 与我们的系统进行对比，结果如表 1 和表 2 所列。我们的系统的得分远高于两个 BASELINE，高于其他的两个系统，略低于得分 R 的最高系统。

表 1 各系统在 TAC 2011 主任务上的 ROUGE-2 得分

Systems	ROUGE-2
best	0.13447
ours	**0.13011**
NUS2	0.12994
irlab2011	0.11887
BASLINE1	0.06410
BASLINE2	0.08682

表 2 各系统在 TAC 2011 主任务上的 ROUGE-SU4 得分

Systems	ROUGE-SU4
best	0.16519
ours	**0.16171**
NUS2	0.15984
irlab2011	0.14794
BASLINE1	0.09934
BASLINE2	0.11749

结束语 本文提出了一个新的数学模型——非完全吸收马尔科夫链，它可以将句子的重要性和它与用户查询的相关性的计算融入到一个统一的框架下，同时还可以利用很多有用的信息。与马尔科夫链相比，它不但具有更小的时间复杂度，而且可以更灵活地整合除句子间相互关系以外的其他许多信息。此外，这个模型也使我们对吸收马尔科夫链有了更深入的理解。以后，我们希望设计一些生成训练数据的自动方法，并引入更多更有效的句子全局特征，同时尝试将这个模型应用到更多领域。

参 考 文 献

[1] Yu P S, Tsotras V J, Fox E A. et al. A system for query-specific document summarization[C]//Proceedings of the 2006 International Conference on Information and Knowledge Management. New York: ACM Press, 2006: 622-631

[2] Erkan G, Radev D. LexRank: prestige in multi-document text summarization[C]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Morristown, NJ: ACM Press, 2004: 365-371

[3] Carbonell J, Goldstein J. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries [C]//Proceedings of the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval. New York, NY: ACM Press, 1998: 335-336

[4] Radev D. A common theory of information fusion from multiple text sources, step one: Cross-document structure[C]//Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue. Morristown, NJ: ACL, 2000: 74-83

[5] Zhai C X, Cohen W W, Lafferty J. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval[C]// Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY: ACM Press, 2003: 10-17

[6] Zhang B Z, Li H, Liu Y, et al. Improving web search results using affinity graph[C]// Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY: ACM Press, 2005: 504-511

[7] Zhang J, Xu H B, Cheng X Q. GSPSummary: A Graph-based Sub-topic Partition Algorithm for Summarization[C]// Proceedings of 2008 Asia Information Retrieval Symposium. AIRS, 2008: 321-334

[8] Zhu X J, Goldberg A, Gael J V. et al. Improving diversity in ranking using absorbing random walks: Human Language Technologies[C]// The Annual Conference of the North American Chapter of the Association for Computational Linguistics. Rochester, NY, NAACL-HLT, 2007

[9] Lin C Y, Hovy E H. From Single to Multi-document Summarization: A Prototype System and its Evaluation[C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Morristown, NJ: Association for Computational Linguistics, 2002: 25-34

[10] Harabagiu S, Lacatusu F. Topic themes for multi-document summarization[C]// Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY: ACM Press, 2005: 202-209

[11] Saggion H, Bontcheva K, Cunningham H. Robust generic and query-based summarization[C]// Proceedings of 10th Conference of the European Chapter of the Association for Computational Linguistics. Morristown, NJ: ACL, 2003: 235-238

[12] Conroy J M, Schlesinger J D. CLASSY query-based multi-document summarization[C]// Proceedings of Document Understanding Conference. Vancouver, Canada, 2005

[13] Page L, Brin S, Motwani R, et al. The PageRank Citation Ranking: Bringing Order to the Web [R]. Stanford, California: Stanford University Database Group, 1998

[14] Motwani R, Raghavan P. Randomized algorithms [M]. Cambridge Press, 1996, 28: 33-37

[15] Haveliwala T. Efficient computation of pagerank[R]. Stanford, California: Database Group, Computer Science Department, Stanford University, 1999

(上接第 183 页)

同时增加了状态分层，使模型语义更加精确，同时为今后的 UML 状态机的操作语义研究打下基础。

未来的工作包括：

(1)对本文定义的 UML 状态机的模型元素的形式化描述进行验证，同时考虑验证过程中的性能开销。

(2)在 UML 状态机模型元素语义形式化定义的基础上，进行操作语义的形式化定义。

(3)在 UML 状态机完整语义定义的基础上，进行 UML 图的一致性的自动检测[4]。

参 考 文 献

[1] OMG. UML2.0 Infrastructure Specification [OL]. http://www.omg.org/cgi-bin/doc? formal/2005-07-05.pdf, 2005

[2] OMG. Object Constraint Language. Version 2.3.1 [OL]. http://www.omg.org/cgi-bin/doc? formal/2009-02-02.pdf, 2009

[3] 蒋慧，林东，谢希仁. UML 状态机的形式语义[J]. 软件学报，2002, 13(12): 2244-2250

[4] Egyed A. Automatically Detecting and Tracking Inconsistencies in Software Design Models[J]. IEEE Transactions on Software Engineering, 2011, 37(2): 188-204

[5] Bjorner D. 软件工程卷 1：抽象与建模[M]. 刘伯超，向剑文，译. 北京：清华大学出版社，2010: 18-21

[6] Woodcock J. Formal Methods: Practice and Experience[J]. ACM Computing Surveys, 2009, 41(4): 19: 1-19: 36

[7] 郭峰，姚淑珍. 基于 Petri 网的 UML 状态图的形式化模型[J]. 北京航空航天大学学报，2007, 33(2): 248-252

[8] 董威，王戟，齐志昌. UML Statecharts 的模型检验方法[J]. 软件学报，2003, 14(4): 750-756

[9] 朱雪阳，唐稚松. Statecharts 的组合语义与求精[J]. 软件学报，2006, 17(4): 670-681

[10] Jin Yan, Esser R. A method for describing the syntax and semantics of UML statecharts[J]. Software System Model, 2004, 3(2): 150-163

[11] 单黎君，朱鸿. UML 的形式化描述语义[J]. 计算机工程与科学，2010, 32(3): 96-103

[12] 郭亮，缪淮扣，王晢，等. UML 模型到 FSM 模型的转换[J]. 计算机科学，2009, 36(7): 113-116

[13] Sun Meng, Zhang Nai-xiao. The Formalization for UML statechart Diagrams[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2005, 41(3): 344-356

[14] 李明，杨海波，张其文，等. 基于时序描述逻辑的 UML 状态图语义[J]. 计算机工程，2010, 36(23): 76-78

[15] 李留英，王戟，齐治昌. UML Statechart 图的操作语义[J]. 软件学报，2001, 12(12): 1865-1868

[16] 曾一. 基于形式化规格说明的 UML 状态图提取[J]. 计算机应用研究，2011, 28(5): 1767-1769

[17] Bendraou R. A Comparison of Six UML-Based Language for Software Process Modeling[J]. IEEE Transactions on Software Engineering, 2010, 36(5): 662-675

[18] Whittle J. Synthesizing Hierarchical State Machines from Expressive Scenario Descriptions[J]. ACM Transactions on Software Engineering, 2010, 19(3): 8: 1-8: 40

[19] Bjorner D. 软件工程卷 2：系统与语言规约[M]. 刘伯超，向剑文，译. 北京：清华大学出版社，2010: 406-413

[20] Ershov A P, Itkin V E. Correctness of mixed computation in Algol-like programs[J]. Mathematical Foundations of Computer Science, 1977, 53: 59-77