基于隐主题马尔科夫模型的多特征自动文摘

刘江鸣[1]　　徐金安[1,†]　　张玉洁[1]

1. 北京交通大学计算机与信息技术学院，北京 100044；

† 通讯作者，E-mail: jaxu@bjtu.edu.cn

摘要　本文基于隐主题马尔科夫模型，消除 LDA 主题模型的主题独立假设，使得文摘生成过程中充分利用文章的结构信息，并结合基于内容的多特征方法提高文摘质量。此外本文提出在不破坏文章结构的前提下，从单文档扩展到多文档的自动文摘策略。最终实现完善的自动文摘系统。本文通过在 DUC2007 标准数据集上的实验结果，证明了隐主题马尔科夫模型和文档特征的优越性，并且所实现的自动文摘系统 ROUGE 值有明显提高。

关键词　隐主题马尔科夫模型；　多特征；多文档自动文摘

中图分类号　TP391

Summarization Based on Hidden Topic Markov Model with Multi-Feature

LIU Jiangming[1]，XU Jinan[1,†]，ZHANG Yujie[1]

1. School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100871

†Corresponding Author, E-mail: jaxu@bjtu.edu.cn

Abstract　Based on hidden topic markov model, this paper eliminates assumption limitation in LDA to exploit the structure information during generating summary, and uses multi-features based on document content to improve the summary quality. Furthermore, this paper proposes the method for developing single-document summarization to multi-document summarization without breaking document structure, to achieve the perfect automatic summarization system. Meanwhile, experimental results on the standard dataset DUC2007 show the advantage of HTMM and multi-feature. The ROUGE values are improved by using HTMM and multi-document comparing with those of LDA.

Key words　Hidden Topic Markov Model; Multi-Features; Multi-Document Summarization

自动文摘是指自动地从大量的文本信息中抽取重要的信息代表文本的主要思想，人们通过阅读文摘能够快速且准确地明白大量文本的主要信息，不仅能够提高阅读效率，并且能够过滤冗余信息。随着互联网快速发展，信息爆炸的时代来临，准确且快速地自动生成大量文本信息的文摘尤为重要。文本理解会议(Document Understanding Conference, DUC)努力促进构建多文档自动文摘系统，并提供人工文摘用以评测机器自动文摘的效果。

自动文摘分为两类，理解性文摘和机械性文摘[1-4]。两种文摘方法的主要区别在于，机械性文摘的生成来自于文本中的句子，然而理解性文摘则不然，其文摘句子通过语法知识自动生成。理解性文摘需要利用领域语义语法等知识进行判断、推理，得到文摘的意义表示，最后从意义表示中生成文摘。理解性文摘缺点在于领域适应能力弱。因此在理解性文摘上的突破成为目前自动文摘的难点。目前，基于机器学习的方法是机械性文摘的主流方法，其分为两类：基于有监督学习的方法和基于无监督的学习方法。有监督的学习方法视自动文摘为分类问题，文本中的句子被分成两类是文摘句和非文摘句，并按照归属类别的程度选择文摘句子[5]。无监督的学习方法通过文本的语义信息对句子打分[6-7]。主题模型是一种有效的文档浅层语义表示模型。

主题可以看成是词项的概率分布。主题模型的起源是隐性语义索引主题模型（Latent Semantic Indexing,

基金项目：国家自然科学基金项目(批准号:61370130)、科技部国际科技合作计划（K11F100010）、中央高校基本科研业务费专项资金（2010JBZ2007)
中国科学院计算技术研究所智能信息处理重点实验室开放课题(IIP2010-4)和北京交通大学人才基金(2011RC034)资助

LSI)。Hofman 在 LSI 基础上提出了概率隐性语义索引（probabilistic Latent Semantic Indexing, pLSI）该模型被认可为真正意义上的主题模型。Blei 在 pLSI 的基础上提出了 Latent Dirichlet Allocation (LDA)[8]，LDA 对 pLSI 进行了更加完美的扩展[9]。之后出现了许多在 LDA 模型上扩展的工作，大多数是引入其他相关信息，例如 Syntactic Topic Model(STM)[10]选择主题时考虑句法信息，Author Topic (AT)[11]考虑作者信息，Dynamic Topic Model (DTM)[12]考虑时间信息，Topic-Sentiment Model (TSM)[13]考虑情感信息。无论是在 LDA 还是 LDA 扩展模型中，主题独立假设很强地限制了主题模型的表现能力，忽略了文本结构信息。因此 Gruber 提出隐主题马尔科夫模型(Hidden Topic Markov Model, HTMM)[14]，句子间的主题关系满足马尔科夫性质，并且主题转移服从二项式分布。由于两个现象：上述的主题模型都受词袋假设限制；文本的主要内容（文摘）可以通过文本特征获得。因此本文提出基于 HTMM 结合文档内容多特征的多文档自动文摘方法。此方法的优越性体现为两个方面：有效利用文档结构信息和有效利用文档内容信息提高自动文摘质量。

本文第 1 节简要介绍和比较 LDA，HTMM 两种主题模型；第 2 节介绍利用主题模型生成文本文摘方法；第 3 节介绍分析文档内容特征；第 4 节分析实验结果；第 5 节给出本文的总结和未来工作。

1 主题模型

Latent Dirichlet Allocation (LDA) 在 pLSI 的基础上被提出，用服从 Dirichlet 分布的 K 维隐含随机变量表示文档的主题概率分布，并且文档中词项的先验概率服从 Dirichlet 分布。LDA 是一个完整的文档生成模型，通过主题和词项的概率分布模拟文档的产生过程（图 1(a)）。在文档的生成过程中，LDA 有两个假设：主题独立和词袋假设。然而，Hidden Topic Markov Model (HTMM) 在模型的构建上打破主题独立假设。HTMM 生成文档过程中，词由不同主题生成，并且主题与主题之间服从二项式分布（图 1(b)）。

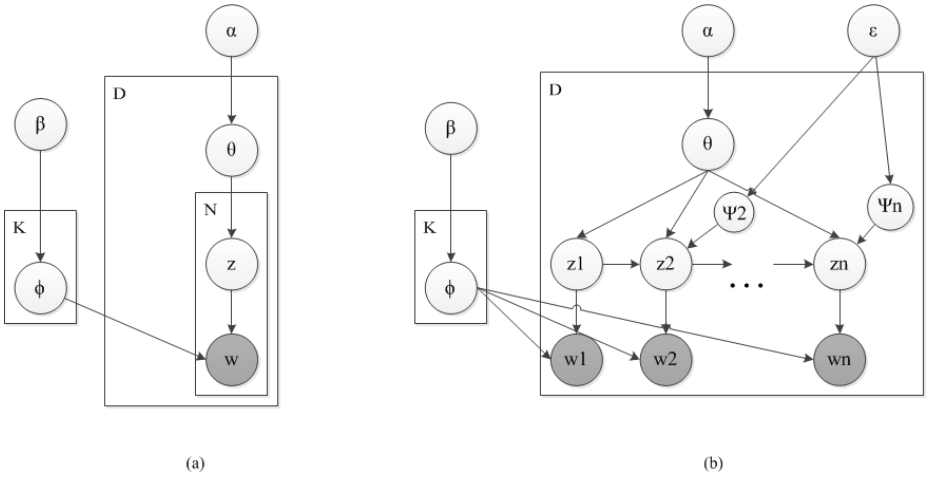

图 1 LDA 与 HTMM 的图结构

Fig. 1 graph structure of LDA and HTMM

1.1 LDA 框架

图 1(a)为 LDA 的图结构，也是各种 LDA 改进模型的基础框架。α为主题先验分布的 Dirichlet 参数；β为词先验分布的 Dirichlet 参数。θ 是一个$|D| \times K$维矩阵，$\theta_{ij} = \mathrm{p}(z_j|d_i)$，即$\theta$是文档产生主题的概率；$\phi$是一个$K \times V$维矩阵，$\phi_{ij} = \mathrm{p}(w_j|z_i)$，即$\phi$是主题产生词的概率。$|D|$为文档数，$K$为主题数，$V$为语料中不同词数。LDA 模型文档的生成过程如下：

Step 1. 为每个主题 k 抽取多项式分布$\phi_k \sim Dirichlet(\beta)$，总共$K$个分布

Step 2. 语料库中的所有文档d = 1…$|D|$

(1) 为文档d抽取主题多项式分布$\theta_d \sim Dirichlet(\alpha)$，

(2) 为文档d抽取主题$z \sim Multinomial(\theta_d)$

(3) 为文档d抽取词$w \sim Multinomial(\phi_z)$

因此模型中主要参数为θ（文档-主题概率分布）和ϕ（主题-词概率分布）并且使用 gibbs 方法[15]进行参数估计。

1.2 HTMM 框架

HTMM 在 LDA 的基础上进行改进，其图结构为图 1(b)。其中参数α、β、θ、ϕ与 LDA 中意义相同。ε为二项式分布参数，表示句子间的主题转移概率，因此ψ取值只可能是 0 或 1。HTMM 假设一个句子中的所有词属于相同主题，并且主题转移仅出现在每个句子的首词。因此 HTMM 文档的生成过程如下：

Step 1. 为每个主题 k 抽取多项式分布$\phi_k \sim Dirichlet(\beta)$，总共 K 个分布

Step 2. 语料库中的所有文档d = 1…$|D|$

(1) 为文档d抽取主题多项式分布$\theta_d \sim Dirichlet(\alpha)$，

(2) $\psi_1 = 1$

(3) 文档d中所有词w_n，n= 2 … N_d 其中N_d表示文档d中不同词的个数

如果句子的开始位置： $\psi_n \sim Binomial(\varepsilon)$

否则：$\psi_n = 0$

(4) 文档d中所有词w_n，n = 1 … N_d

如果$\psi_n == 0$ ： $z_n = z_{n-1}$

否则：$z_n \sim Multinomial(\theta_d)$

$w_n \sim Multinomial(\phi_{z_n})$

不同于 LDA 模型，HTMM 的文档生成过程中需要考虑文档中词的顺序，通过二项式分布描述句子间主题的转移。

HTMM 相对于 LDA 来说更具有广泛性，当所有ψ都置为 1 时，HTMM 模型将退化成 LDA 模型。当所有ψ都置为 0 时，HTMM 模型就退化成了混合 unigram model，即文档中所有的词具有相同主题。

2 基于主题模型的自动文摘

主题模型将主题表示为隐性变量，仅文档内容可观测到，最主要的参数为各主题下的词项概率分布（ϕ）和各文档下的主题概率分布（θ）。

2.1 单文档自动文摘

基于主题模型的抽取式文摘（机械文摘）生成，文摘句来自于文档中的句子。主要利用主题模型中的θ与ϕ概率分布，计算句子的主题概率分布和文档的主题概率分布。文摘句具有与文档相似的主题概率分布[16]。因此本文选取与文档具有高相似主题概率分布的句子作为文摘句。因此生成文摘可分为三个过程：(1) 训练主题模型参数，得到θ与ϕ概率分布；(2) 计算句子主题分布及文档主题分布；(3) 比较句子主题分布与文档主题分布，从而抽取候选文摘句。

本文使用 HTMM 获取θ与ϕ概率分布，这与 LDA 参数意义相同，不同之处在于参数估计的迭代公式中，HTMM 以句子为单位考虑主题转移概率。之后计算文档的主题概率分布和文档中所有句子的主题概率分布。一方面，主题模型的参数估计可以直接获得文档的主题概率分布$P(Z|D)$；另一方面，句子的主题概率分布$P(Z|S)$来自于两部分：文档-主题的概率分布ϕ和主题-词项的概率分布θ。本文使用文献[16]提出的公式(1)计算句子S的产生主题概率分布。

$$P(Z|S) = \frac{\sum_{w \in S} P(w|Z) \times P(Z|D) \times P^2(D)}{P(S)} \tag{1}$$

其中D表示句子S所在的文档，$P(w|Z)$为主题Z下词项 w 的概率，$P(Z|D)$为文档D下主题Z的概率。公式(1)假设词与词相互独立。本文假设文档出现概率均等，句子出现概率也均等，并且消除长句优势，将公式(1)简化为公式(2)

$$P(Z|S) = \frac{\sum_{w \in S} P(w|Z) \times P(Z|D)}{len(S)} \tag{2}$$

其中$len(S)$表示以词为单位的句长。

我们使用 KL 散度值计算，两种概率分布的相似程度。计算公式(3)如下

$$D_{KL}(P||Q) = \sum_i P(i) \times \log \frac{P(i)}{Q(i)} \tag{3}$$

其中P与Q为概率分布，KL 散度值越小，P与Q概率分布之间的差异就越小。因此本文选取句子主题概率分

布与文档主题概率分布 KL 散度值最小的句子集合作为候选文摘句，即$D_{KL}(P(Z|S)||P(Z|D))$最小。

2.2 多文档自动文摘

基于主题模型的文摘方法，从单文档自动文摘扩展到多文档自动文摘，比较常见的方法是将多篇文档句子组合成一篇文档，然后按照单文档自动文摘的方法生成文摘[17]。但是将多篇文档组合成一篇文档，文档结构被破坏是无法避免的。HTMM 缓解了主题独立假设，句子间主题的转移概率表现了文档的结构信息。因此，基于 HTMM 的多文档文摘不适合按照传统的方法，相反传统方法会引入不必要的噪音。本文提出以下策略：

(1) 对各文档从先验分布函数中抽取各自的主题分布。

(2) 多文档中的句子集合，分别对各自原文档的主题分布计算KL散度值，将其作为文档集合的主题KL散度值。

一方面，此策略的假设前提是文档集中的所有文档所描述的主题一致或类似。另一方面，本文考察的内容特征在保留原文档集的前提下，能够正确获取。内容特征在第三节有详细说明。因此本文以此策略为前提，进行 LDA 与 HTMM 模型实验及内容特征实验。

3 文档内容特征

传统方法进行自动摘要，通过使用文档句子级特征和词级特征，计算句子作为文摘句的权重，从而生成文档文摘，但由于无法深入获取文档的语义信息，单纯基于特征的传统方法遇到了瓶颈。之后，基于主题模型的自动文摘是一个突破，利用了浅层的语义特征。然而传统方法的自动摘要仍然有其独有的优势[18-19]。因此本文通过文档内容特征对候选文摘句（基于主题模型的文摘）进行重排序，最终得到文档文摘。

本文使用的文档特征包括句子级特征（表 1）和词级特征（表 2）。句子级特征有：

1. 位置特征　　结构性文档中开头句或结尾句具有总结意味。使用公式(4)计算位置特征得分。

$$score_s(SP) = |\frac{pos(S,D)}{|sent(D)|} - \frac{1}{2}| \times 2 \tag{4}$$

其中$pos(S,D)$表示句子S在文档D的位置（句子序号），$sent(D)$表示文档D的句子数

2. 长度特征　　使用权重值约束文摘句的长度，避免摘要偏向于长句子。使用公式(5)，计算长度权重值。

$$score_s(SL) = 1 - \frac{\min(|len(S)-arglen(D)|,arglen)}{arglen(D)} \tag{5}$$

其中$len(S)$为句子长度，$arglen(D)$为文档中句子的平均句长。

3. 相似度特征　　标题是对文档中心内容的概括。因此，与标题具有高相似度的句子更可能成为文摘句。本文使用余弦相似度公式(6)，计算句子与标题的相似度。

$$score_s(ST) = \cos(SV, TV) \tag{6}$$

其中句子向量SV与标题向量TV为$|V|$维向量，$|V|$为文档中的词个数。

4. 命名实体特征　　文本文摘不仅是对文档的总结，更具有很强的信息量。命名实体包括人名、地名和组织机构名，并且命名实体具有很强的信息标识作用。本文使用公式(7)计算命名实体特征对文摘句的贡献。

$$score_s(SNE) = \frac{NE(S)}{len(S)} \tag{7}$$

其中$NE(S)$表示句子S中命名实体的个数。

词级特征有：

1. 位置特征　　结构性文档中多次出现在文章开头或者文章结尾的词具有总结意味，使用公式(8)计算词位置特征得分。

$$score_S(WP) = \frac{\sum_{W \in S} \frac{|\frac{pos(W,D)}{|D|} - \frac{1}{2}|}{|W|}}{len(S)} \tag{8}$$

其中$pos(W,D)$表示词W出现在文档D的位置（词序号），$|W|$表示词W在文档出现的次数。

2. 词频特征　　使用TF-IDF值作为词频特征得分，见公式(9)。

$$score_S(WF)=\frac{\sum_{W\in S} normal(TF\_IDF(W))}{len(S)} \tag{9}$$

其中$normal(*)$为归一化函数。

表 1　句子级特征

Table 1　Features in sentence-level

特征	特征描述
位置	所在文档的位置
长度	句子长度
相似度	句子与标题的相似度
命名实体	命名实体所占比率

表 2　词级特征

Table 2　Features in word-level

特征	特征描述
位置	所在的文档的位置
词频	TF-IDF 值

其中公式(4)-(9)中 $score(*)\in[0,1]$。最后使用公式(10)，将各特征加权组合，根据组合得分和候选文摘句的 KL 离散值，对主题模型生成的候选文摘句序列进行重排序。

$$\begin{aligned} score_s(Multifeature) &= \lambda_{SP}\cdot score_s(SP)+\lambda_{SL}\cdot score_s(SL)+\lambda_{ST}\cdot score_s(ST)+\lambda_{SNE}\cdot score_s(SNE) \\ &+\lambda_{WP}\cdot score_s(WP)+\lambda_{WF}\cdot score_s(WF) \end{aligned} \tag{10}$$

其中λ为各特征的参数。且$0\le\lambda\le 1$。参数设定在 4.2 节有详细论述。

由于D_{KL}值需要最小，而$score(Multifeature)$需要最大。因此使用公式(11)进行计算生成文摘句的花费值

$$cost(S)=D_{KL}-score_s(Multifeature) \tag{11}$$

最终在文摘长度的限制下，通过抽取花费值$cost$最小的句子集合生成文本文摘。

4　实验及结果分析

4.1 实验数据及评测方法

本文采用 DUC2007 自动评测任务中的数据集作为实验数据，该数据包含 45 个文档集合，每个集合包含 25 篇具有相关主题的文档，并且每个文档集合都给出了 4 个最大词数为 250 的人工生成的专家文摘作为正确摘要。本实验按照评测要求，对每个文档集进行不超过 250 个词的多文档摘要生成。

实验结果的分析与评测使用 ROUGE[20]自动评测工具，ROUGE 的评测指标结合 n-gram 及 WordNet，比较生成文摘与专家文摘的相似程度，评价生成文摘的质量。ROUGE 使用基于召回率的自动评测方法，并有效计算相关 F 值。DUC2007 使用 ROUGE-2 和 ROUGE-SU4 两个指标。因此本文的实验也按此两个标准便于比较分析

4.2 参数设定

实验中，主题模型先验分布参数α设置为$\frac{K+50}{K}$，K 为主题数；β取经验值 1.01。通过预处理实验确定 K 值，图 2 显示了不同主题下的 SU-4 值。因此本文最终确定 K 值为 50 为作为主题数。

实验中通过贪心算法（greedy Algorithm），将特征权重参数以 0.1 为步进长度，值域为[0,1]上调节参数，最终得到优化参数（见表 3）。由于特征参数较少，且最终 *cost* 值由 KL 值所约束，而 KL 值是通过无监督

学习方法的参数计算得到，因此参数对数据的依赖程度不明显。此外，由于命名实体特征依靠于命名实体识别系统的性能。鉴于识别精度的限制，我们初期对其赋予较小的值。后期我们将对命名实体特征做深入研究探索。

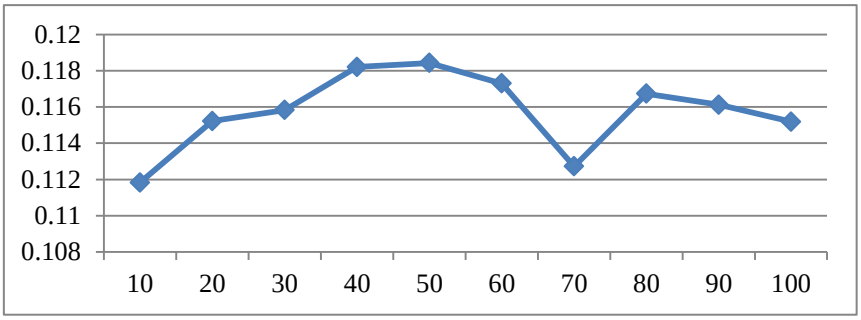

图 2　不同主题数下基于 HTMM 模型系统的 SU-4 值

Fig. 2　the SU-4 values given by summarization system based on HTMM in different topic number

表 3　特征参数设定

Table 3　Features in word level

参数	λ_{SP}	λ_{SL}	λ_{ST}	λ_{SNE}	λ_{WP}	λ_{WF}
参数值	0.3	0.2	0.4	0.1	0.3	0.3

4.3 实验结果及分析

表 4 给出了实验结果。在同等条件下，HTMM 利用文档结构信息获取文档主题分布的效果优于 LDA 模型。并且通过实验表明，基于 HTMM 生成的文摘，无论是关键信息覆盖评测指标（ROUGH-1），还是文摘可读性评测指标（ROUGE-2, ROUGE-3, ROUGE-4）都优于 LDA 模型，并且有显著的提高。同时，实验表明在同等主题模型下，加入基于文档内容的特征，可以有效提高文摘质量。ROUGE-n 指标以 n-gram 为基础，比较生成文摘与专家文摘中 n 个连续出现的公共子串的 F 值；ROUGE-L 为最长公共子串的 F 值；ROUGE-SU 为任何公共子串的 F 值。LDA 模型加入内容特征后，ROUGE-2 值提高了 0.24%，ROUGE-SU4 值提高了 0.17%。HTMM 加入内容特征后，ROUGE-2 值提高了 1.18%，ROUGE-SU4 值提高了 0.829%。单独使用特征效果不是很理想，原因在于无法准确地获取文章隐藏的语义信息。而主题模型则填补了未使用浅层语义的空缺。表 4 的实验结果可以看出，不仅 n-gram 评测指标最长公共子串评测指标都说明，使用 HTMM 所得到的主题概率分布，相对于 LDA 更能表现文档内容，从而选取的文摘句更贴近专家文摘。

表 4　实验结果

Table 4　results of experiment

	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4	ROUGE-L	ROUGE-SU4
Features	0.32272	0.05056	0.01309	0.00582	0.29022	0.10133
LDA	0.26861	0.02523	0.00343	0.00130	0.24962	0.07454
LDA+Features	0.27142	0.02767	0.00406	0.00139	0.25165	0.07626
HTMM	0.36076	0.06505	0.01652	0.00626	0.32947	0.11843
HTMM+Features	0.37125	0.07686	0.02231	0.00940	0.34096	0.12672

同时，本文使用 DUC2007 提供的 SCU-marked 文档对各文档内容特征进行考察。SCU-marked 文档来自于部分 DUC2007 评测文档集，其对句子进行标注，标注内容为评测系统中选取某句子作为文摘句的系统个数 $count(S)$。本文使用公式(12)对文档集的句子划分为文摘句集 $(summaries)$ 和非文摘句集 $(non\_summaries)$。

$$\begin{cases} S \in summaries & if\ count(S) \geq \delta \\ S \in non\_summaries & if\ count(S) < \delta \end{cases} \tag{12}$$

本文取δ为 3，同时被 3 个系统选取为文摘句更具有说服力。分别考察句子级和词级特征(见图 3 和表 5)。文摘句与非文摘句的内容特征得分中，句子位置特征、主题相似度特征、命名实体特征和词频特征有明显的不同，文摘句的得分高于非文摘句，尤其是在主题相似度特征得分上，文摘句得分为非文摘句的接近一倍。相对于句长特征，则表现相反。由此看出，大部分文摘系统普遍倾向选取长句对文摘进行概述，而句子位置特征、主题相似度特征等是普遍认同的。由于多文档文摘中文章的书写涉及作者个人意识及习惯，并且词位置特征和词频特征仅使用词形信息，因此，词级特征所提供信息较少。为突出词级特征的优越性，我们在未来工作将考虑词间的同义性和相关性。

表 5　不同内容特征的得分差异率

Table 5　the difference rate of scores of different features

特征	SP	SL	ST	SNE	WP	WF
差异率	0.2887	-0.3022	0.9140	0.6708	-0.0390	0.7512

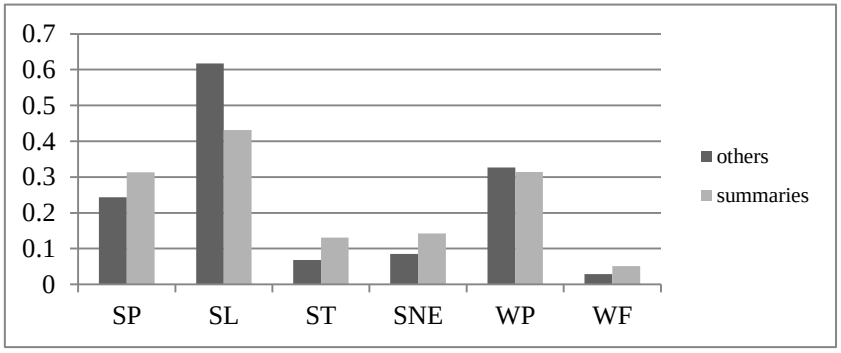

图 3　不同特征在文摘句和非文摘句的得分

Fig. 3　the scores of different features in summaries and non-summaries

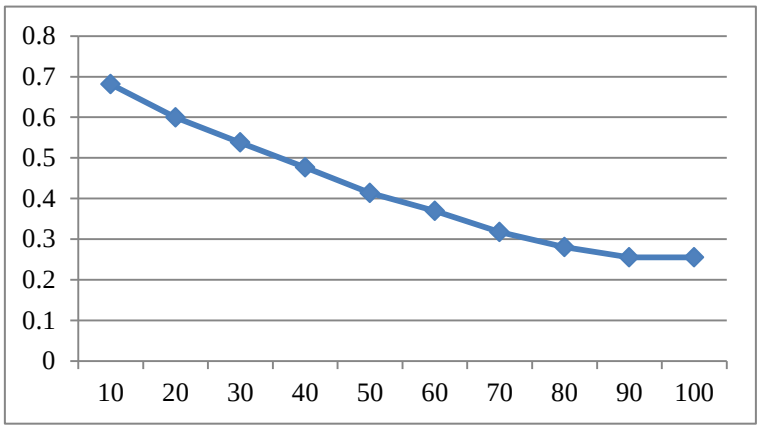

图 4　不同主题数文档间主题分布的 KL 散度值

Fig. 4　the KL value of topic distribution between documents in different topic number

5　总结与未来工作

本文提出基于 HTMM 的多特征自动文摘方法，优越性体现在三个方面：一方面，HTMM 缓解了 LDA 对主题独立的假设，更加宽泛地表现了文档的主题分布，并利用到句子级的文档结构信息；一方面，主题模型特征结合传统文档内容特征，对候选文摘句的重排序，提高文摘质量；一方面，提出单文档到多文档

自动文摘策略，并以此解决单文档自动文摘生成到多文档自动文摘生成的结构破坏问题。该方法在DUC2007 评测数据上有显著的效果。

由于 HTMM 在文档生成上的优越性只表现在相邻句子间的主题转移概率分布，即延续前个句子主题或重新生成主题，而非真正意义上的主题间的转移概率分布。因此将其二项式分布扩展为多项式分布成为今后研究的重点，以此探究文档结构在结构化多文档自动摘要的重要性。其次在传统特征的权值参数设定上，可采取半监督机器学习的方法进行调参优化，达到与理想文摘更加接近的目的。

同时我们考究同一主题下的文档集中，文档间在不同主题数下的主题分布 KL 散度值见图 4。如图 4 可见 KL 散度值随着主题数增加，散度值下降。此现象说明，主题数决定了主题模型描述文档潜在语义内容的能力。因此，对于主题数的研究是必要的，也正是今后我们未来的工作。

参考文献

[1] 刘挺, 王开铸. 自动文摘的四种主要方法. 情报学报, 1999, 18(1): 11-19.

[2] Arora R, Ravindran B. Latent dirichlet allocation based multi-document summariza-tion. Proceedings of the second workshop on Analytics for noisy unstructured text data. ACM. USA, 2008: 91-97.

[3] Gong Y, Liu X. Generic text summarization using relevance measure and latent se-mantic analysis. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM. USA, 2001: 19-25.

[4] Bhandari H, Shimbo M, Ito T, et al. Generic text summarization using probabilistic latent semantic indexing. Proceedings of The Third International Joint Conference on Natural Language Processing. India. 2008: 133-140.

[5] Shen D, Sun J T, Li H, et al. Document summarization using conditional random fields. Proceedings of the 20th international joint conference on Artifical intelli-gence. India, 2007, 7: 2862-2867.

[6] 王红玲, 张明慧, 周国栋. 主题信息的中文多文档自动文摘系统. 计算机工程与应用, 2012, 48(25): 132-136.

[7] Titov I, McDonald R. A joint model of text and aspect ratings for sentiment summarization. Proceedings of 46th Annual Meeting of the Association for Computational Linguistics (ACL'08), USA, 2008.

[8] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. the Journal of machine Learning research, 2003, 3: 993-1022

[9] 徐戈, 王厚峰. 自然语言处理中主题模型的发展. 计算机学报, 2011, 34(8): 1423-1436.

[10] Boyd-Graber J, Blei D M. Syntactic topic models. Neural Information Processing Systems. 2009.

[11] Steyvers M, Smyth P, Rosen-Zvi M, et al. Probabilistic author-topic models for in-formation discovery. Proceedings of the tenth ACM SIGKDD international con-ference on Knowledge discovery and data mining. ACM, USA, 2004: 306-315.

[12] Blei D M, Lafferty J D. Dynamic topic models. Proceedings of the 23rd international conference on Machine learning. Montreal ACM, USA, 2006: 113-120.

[13] Mei Q, Ling X, Wondra M, et al. Topic sentiment mixture: modeling facets and opinions in weblogs. Proceedings of the 16th international conference on World Wide Web. Alberta : ACM, USA, 2007: 171-180.

[14] Gruber A, Weiss Y, Rosen-Zvi M. Hidden topic Markov models. International Conference on Artificial Intelligence and Statistics, 2007: 163-170.

[15] Casella G, George E I. Explaining the Gibbs sampler. The American Statistician, 1992, 46(3): 167-174.

[16] 张明慧, 王红玲, 周国栋. 基于 LDA 主题特征的自动文摘方法. 计算机应用与软件, 2011, 28(10): 20-22.

[17] 秦兵, 刘挺, 李生. 多文档自动文摘综述. 中文信息学报, 2005, 19(6): 13-20.

[18] 吴晓锋, 宗成庆. 一种基于 LDA 的 CRF 自动文摘方法. 中文信息学报, 2009, 23(6): 39-45.

[19] Galanis D, Lampouras G, Androutsopoulos I. Extractive Multi-Document Summarization with Integer Linear Programming and Support Vector Regression. Mumbai:COLING, India, 2012: 911-926.

[20] Dang H T, Owczarzak K. Overview of the TAC 2008 update summarization task. Proceedings of text analysis conference, 2008: 1-16.