doi:10.3969/j.issn.1673-4785.2010.02.012

# 主题模型LDA的多文档自动文摘

杨 潇[1],马 军[2],杨同峰[2],杜言琦[2],邵海敏[2]
(1.山东经济学院 信息管理学院,山东 济南 250014;2.山东大学 计算机科学与技术学院,山东 济南 250101)

**摘 要**:近年来使用概率主题模型表示多文档文摘问题受到研究者的关注.LDA(latent dirichlet allocation)是主题模型中具有代表性的概率生成性模型之一.提出了一种基于LDA的文摘方法,该方法以混乱度确定LDA模型的主题数目,以Gibbs抽样获得模型中句子的主题概率分布和主题的词汇概率分布,以句子中主题权重的加和确定各个主题的重要程度,并根据LDA模型中主题的概率分布和句子的概率分布提出了2种不同的句子权重计算模型.实验中使用ROUGE评测标准,与代表最新水平的SumBasic方法和其他2种基于LDA的多文档自动文摘方法在通用型多文档摘要测试集DUC2002上的评测数据进行比较,结果表明提出的基于LDA的多文档自动文摘方法在ROUGE的各个评测标准上均优于SumBasic方法,与其他基于LDA模型的文摘相比也具有优势.

**关键词**:多文档自动文摘;句子分值计算;主题模型;LDA;主题数目
**中图分类号**:TP391 **文献标识码**:A **文章编号**:1673-4785(2010)02-0169-08

# Automatic multi-document summarization based on the latent Dirichlet topic allocation model

YANG Xiao[1], MA Jun[2], YANG Tong-feng[2], DU Yan-qi[2], SHAO Hai-min[2]
(1. School of Information Management, Shandong Economic University, Ji'nan 250014, China; 2. School of Computer Science and Technology, Shandong University, Ji'nan 250101, China)

**Abstract**: The representative problem of multi-document summarization using probabilistic topic models has begun receiving considerable attention. A multi-document summarization method was proposed based on the latent dirichlet allocation (LDA) model, itself a model representative of probabilistic generative topic models. In this method, the number of topics in the LDA model was determined by model perplexity, and the probabilistic sentence distribution on topics and the probabilistic topic distribution on words were obtained by the Gibbs sampling method. The importance of topics was determined by the sum of topic weights on all sentences. Two sentence-scoring methods were proposed, one based on sentence distribution and the other on topic distribution. Evaluated by the recall-oriented understudy for gisting evaluation (ROUGE) metrics, results of the both proposed methods surpassed the state-of-the-art SumBasic system and the other two LDA based summarization systems for all the ROUGE scores on the DUC2002 generic multi-document summarization test set.

**Keywords**: multi-document summarization; sentence scoring; topic model; latent dirichlet allocation; number of topics

多文档自动文摘是对内容相关的多篇文本进行分析,并产生可以表达重要信息的摘要文本的过程,其中摘要文本长度需满足指定长度的要求[1].它作为自然语言处理和信息检索中最古老的问题之一,随着移动设备、互联网的广泛应用,用户面临的信息量的激增,在近几年又重新兴起.自动文摘按照摘要目的的不同可以分为通用型文摘(generic summarization)和基于查询的文摘(query-based summarization).通用型文摘提取反映作者意图的总结性文字,而基于查询的文摘则给出与用户查询相关联的摘要[2].按照摘要产生方法的不同自动文摘可以分为抽取式文摘(extract)和理解式文摘(abstract).抽取式文摘计算句子的分值,直接从原文中抽取重要的句子作为文摘句;而理解式文摘则通过对文章进

**收稿日期**:2010-01-05.
**基金项目**:国家自然科学基金资助项目(60970047);山东省自然科学基金资助项目(Y2008G19);山东省科技计划资助项目(2007GG10001002, 2008GG10001026).
**通信作者**:杨 潇. E-mail:yangx@mail.sdu.edu.cn.

行句法、语义和篇章结构的分析获取文档的意义,再通过自然语言生成得到满足要求的文摘[3]. 虽然抽取式的文摘方法经常会产生缺乏连贯性的文摘,但其产生的文摘对人类浏览和判断是有帮助的[4]. 且由于理解式文摘中涉及的多个自然语言处理问题目前没有良好的解决方法,而抽取式文摘则避免了篇章分析、连贯句子的生成等难题,目前的研究大都使用抽取式的文摘方式.

抽取式文摘中的主要问题是句子权重的计算问题. 常用的句子权重计算方法有简单的基于词频的方法[5]、基于主题聚类的方法[6]、基于图的方法[7-8]和基于语言分析的方法[9]等.

以 LDA(latent dirichlet allocation)[10]及其扩展为代表的主题模型广泛应用于文档、图像等信息的建模. 近年来,开始有学者关注其在自动文摘方面的应用[11-13]. 本文基于 LDA 模型,以句子作为处理单元,根据 LDA 模型中主题的概率分布和句子的概率分布提出了 2 种句子权重计算模型.

## 1 相关工作

Chen 等[14]提出了一种结合句子生成概率和先验概率完成句子排序的广播新闻演讲文摘的抽取,其中句子的生成概率的方法来考察了句子主题混合模型(STMM)和词主题混合模型(WTMM)2 种概念匹配形式,混合模型的参数则根据文档的标题由期望最大化(EM)算法训练得到.

Arora 等[12]同样使用 LDA 作为文档的表示模型,但其以文档作为 LDA 的处理单元,提出了基于推论的、半生成性和全生成性的 3 种句子选择形式. 效果最好的是基于推论的方法,其中句子的概率为归一化后的词汇概率加和. 在文献[13]中,Arora 等在使用 LDA 得到单词的权重后,将句子看作单词权重的向量. 每个句子对应一个主题,主题则为所有属于该主题的句子的向量,最终将主题表示为单词的权重矩阵. 然后使用 SVD 求解句子集的正交表示,作为选择文摘句的依据,从而降低文摘中信息的冗余度. Shafiei[15]提出类似于 3 层生成模型 LDA 模型的 4 层模型 Co-Clustering Model,由于该模型表示为词、片段、主题、文档 4 层结构,若将片段选择为句子,则该方法可以为词、句子和文档建立统一的生成模型. 研究者们将该模型应用于文摘中:

Haghighi 等[11]使用层次 LDA 主题模型的变种,将句子、文档和文档集合统一纳入到主题模型中,使用 Gibbs 抽样获得模型参数,同时考虑到文档集的综合主题和特定主题 2 个方面,并以 KL-散度作为文摘评价模型选择句子,使用贪心算法添加句子.

Chang 和 Chien[16]对文档和单个的句子分别执行 LDA,然后通过计算句子语言模型和文档语言模型之间的 KL-散度对句子进行排序. 为了充分地表示词汇、句子和文档之间的关系,又提出了 SLDA,为词汇、句子、主题和文档建立 4 层 LDA 模型,并使用变分推断估计参数,通过计算句子语言模型和文档语言模型之间的 KL-散度对句子进行排序.

## 2 主题模型 LDA

主题模型是一种生成性的概率模型,一般基于如下观点构建:文档是主题上的概率分布;而主题则是词汇上的概率分布. 不同的主题模型做了不同的概率假设;由于主题在词上的概率分布是相关词项上的连贯聚类,因此单个的主题都是可解释的.

### 2.1 LDA 模型简介

LDA 模型是一种常用的主题模型,由 Blei 等人于 2003 年提出. 它是一个生成性的 3 层贝叶斯网络,将词和文档通过潜在的主题相关联. 类似于许多概率模型,LDA 中也做了词袋(bag of words)假设,即在模型中不考虑词汇的顺序而只考虑他们的出现次数.

LDA 模型是一个描述如何基于潜在主题生成文档中词的概率抽样过程,其生成过程如下:

1)从 Dirichlet 先验 $\boldsymbol{\beta}$ 中为每个主题 $k$ 抽取多项式分布 $\phi_k$,共抽取 $K$ 个分布;

2)从 Dirichlet 先验 $\boldsymbol{\alpha}$ 中为每个文档 $w_m$ 抽取多项式分布 $\theta_m$,共抽取 $M$ 个分布;

3)对语料库中所有文档 $w_m$ 和文档中所有词汇 $w_{mn}$:

①从多项式变量 $\theta_m$ 中抽取主题 $z_m$;

②从多项式变量 $\phi_z$ 中抽取词 $w_{mn}$.

其中 $K$ 为主题个数,$M$ 为文档个数. 模型中的主要变量为主题—词分布"$\phi$"和文档—主题分布"$\theta$". 由于直接使用 EM 算法估计 $\phi$ 和 $\theta$ 会存在局部极值的问题,对于给定的观察词 $w_n$,利用 Gibbs 抽样取词汇在主题 $z$ 上后验概率 $P(w_n \mid z)$ 的近似值. 在 Gibbs 抽样中,先固定其他词的主题分配($z_{-n}$),然后估计当前词项 $w_n$ 赋各种主题的概率 $p(z_n = j)$. 边缘化 $\phi$ 和 $\theta$ 间接求得 $\phi$ 和 $\theta$ 的值:

$$P(z_n = j \mid z_{-n}, w_{m,n}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \frac{C^{VK}_{w_{-n}j} + \beta_{w_{n,j}}}{\sum_{v=1}^{V}(C^{VK}_{v_{-n},j} + \beta_{v,j})} \times \frac{C^{MK}_{m_{-n},j} + \alpha_{m,j}}{\sum_{k=1}^{K}(C^{MK}_{m_{-n},k} + \alpha_{m,k})}. \tag{1}$$

式中:$\boldsymbol{C}^{VK}$和$\boldsymbol{C}^{MK}$分别为维数 $V\times K$ 和 $M\times K$ 的数量矩阵,$V$ 为词汇个数. $C^{VK}_{w_{-i}j}$为词 $w$ 赋主题 $j$ 的频数,其中不包含当前记号实例 $n$;$C^{MK}_{d_{-i}j}$为文档 $d$ 中分配给主题 $j$ 的词汇个数,其中不包含当前实例 $n$;由于 $w_n$ 不仅代表词汇 $w$,而且与该词所在的文本位置相关,因此称之为记号. 一旦词的某些记号赋给了主题 $j$,就增加了给任一特定的记号赋予主题 $j$ 的概率;同样地,若主题 $j$ 在一个文档中使用了多次,则该文档中的任意词赋给主题 $j$ 的概率也将增加. 因此,词赋予某一主题的概率不仅与该词跟主题的相近程度有关,而且与文档中该主题的重要程度有关.

### 2.2　Gibbs 抽样

首先为词汇记号赋[1..$K$]之间的一个随机主题,构成初始的 Markov 链;对于文档中的所有词汇记号,根据式(1)给它分配主题,获取 Markov 链的下一个状态;迭代足够次后使得 Markov 链达到稳定状态.

抽样算法为每个单词直接估计其主题 $z$,$\theta$ 值和 $\phi$ 值则由式(2)获得:

$$\hat{\phi}^{(z=j)}_{w_n}=\frac{C^{VK}_{w_{-i},j}+\beta_{w_i,j}}{\sum_{v=1}^{V}(C^{VK}_{v_{-i},j}+\beta_{v,j})},$$

$$\hat{\theta}^{(m)}_{z=j}=\frac{C^{MK}_{d_{-i},j}+\alpha_{d,j}}{\sum_{k=1}^{K}(C^{MK}_{d_{-i},k}+\alpha_{d,k})}. \tag{2}$$

$\phi$ 值为从主题 $j$ 中抽样新词记号 $w_n$ 的预测,而 $\theta$ 为在文档 $w_m$ 中从主题 $j$ 抽取新词的预测.

## 3　基于 LDA 的多文档自动文摘

对于给定的文档集合 $D=\{D_1,\cdots,D_M\}$,各文档 $D$ 中包含句子集合 $D=\{s_1,\cdots,s_k\}$. 为简单起见,本文中将文档集合表示为该集合所有文档中句子的集合,即 $D=\{s_1,\cdots,s_N\}$,其中 $s_i\in D$ 当且仅当 $s_i\in D_j\in D$.

以文档集合中的句子作为 LDA 输入的文档,句子集合作为 LDA 的文档集合,使用 LDA 为句子集合 $D$ 建模,并使用 Gibbs 抽样进行参数估计,得到句子在主题上的分布 $\hat{\theta}^{(s)}_{z=j}$和主题在词汇上的分布 $\hat{\phi}^{(z=j)}_{w}$. 基于这 2 个分布,提出了 2 种不同的句子权重计算方法.

文档中词汇的重要度不仅与该词汇所赋主题的相似度有关,而且与所赋主题的重要度有关. 词汇与所赋主题的相似度由 $\hat{\phi}^{(z=j)}_{w}$ 计算,主题的重要度则由 $\hat{\theta}^{(s)}_{z=j}$得到.

### 3.1　主题的重要度

在 LDA 模型中主题的重要度与其混合成分的比例和超参数 $\boldsymbol{\alpha}$ 有关. 由 Gibbs 算法计算出各主题在句子混合成分中所占的权重后,句子集合中主题的重要度可以使用句子集合包含的所有句子中主题混合成分权重的加和来计算,并在所有主题上进行归一化以保证该值为合适的概率值:

$$P(z_i\mid D)=\frac{\sum_{n=1}^{N}\hat{\theta}^{(s)}_{z}}{\sum_{j=1}^{K}\sum_{n=1}^{N}\hat{\theta}^{(s)}_{z}}. \tag{3}$$

式中:$N$ 为文档集中句子的个数,$K$ 为文档集中主题的个数.

### 3.2　概率生成模型(ProbGenSum)

句子集中词汇的重要程度由词汇与主题的相似度和主题的重要程度共同决定,在概率生成性主题模型中,词汇的概率可以由式(4)计算:

$$P(w\mid D)=\sum_{j=1}^{K}P(w\mid z_j)\times P(z_j\mid D). \tag{4}$$

式中:$K$ 为主题个数,$P(w|z_j)$ 为主题 $z_j$ 在词汇 $w$ 上的概率,在使用 Gibbs 抽样的 LDA 中即为参数 $\hat{\phi}^{(z=j)}_{w}$,而 $P(z_j|D)$ 为主题 $z_j$ 的重要度,由式(3)中的方法获得.

句子的权重可由句子所包含词的权重获得. 由于概率 $P(w|D)$ 为[0,1]之间的值,若使用概率的乘积计算句子的概率即 $P(w\mid D)=\prod_{w\in S}P(w\mid D)^{n(w,S)}$,则短句子占优势,但一般来说在文档中句子越短,其包含的信息量也越少. 本文中使用词汇概率的加和作为句子的权重. 在这种情况下,长句子的概率值将比短句子高,在选择文摘句时占据优势. 且包含具有较高概率值词汇的短句也将获得较高的概率值,选出的句子并非都是长句子. 文摘模型 ProbGenSum 中句子的权重由式(5)计算:

$$P(S\mid D)=\sum_{w\in S}n(w,S)\times P(w\mid D). \tag{5}$$

式中:$n(w,S)$ 为词 $w$ 在句子 $S$ 中出现的次数,$P(w|D)$ 为词 $w$ 的概率值.

### 3.3　句子生成模型(SentGenSum)

在概率生成模型中,文档集 $D$ 中句子 $S$ 的重要性表示为 $P(S|D)$,即给定文档集 $D$ 时句子 $S$ 的后验概率. 根据贝叶斯法则,$P(S|D)$ 可表示为

$$P(S\mid D)=\frac{P(D\mid S)P(S)}{P(D)}. \tag{6}$$

式中:$P(D|S)$ 为句子的生成概率,即文档集 $D$ 由句子 $S$ 生成的可能性,$P(S)$ 为句子 $S$ 重要性的先验概率,$P(D)$ 为文档集 $D$ 的先验概率. 将文档集中的词作为输入观察序列,则句子由预测文档集的分布构成,可以将句子看作文档集的生成模型. 文档集的概率 $P(D)$ 对所有句子都是相同的,不影响句子的排序,因此在计算句子分值时可将其忽略. 本文假设句子的先验概率相同,则句子的分值只与其句子生成概率相关. 根据句子生

成概率 $P(D|S)$ 对文档中的句子进行排序,选出具有最高概率值的句子形成摘要.

在基于主题模型 LDA 的生成性文摘方法中,文档中的句子 $S$ 可以解释为主题的概率混合模型. 在该模型中一个句子可以属于多个主题,同时使用 $K$ 个主题和每个主题在句子中对应的权重预测文档中的单词. 给定句子 $S$ 时文档集 $D$ 的概率表示为

$$P(D\mid S)=\prod_{w\in D}\cdot[\sum_{k=1}^{K}P(w\mid z_k)\cdot P(z_k\mid S)]^{n(w,D)}. \tag{7}$$

式中:$n(w, D)$ 为词 $w$ 在句子集中 $D$ 出现的次数,$P(w|z_k)$ 和 $P(z_k|S)$ 分别为词 $w$ 在潜在主题 $z_k$ 上的概率和主题 $z_k$ 在句子 $S$ 的概率. 在使用 Gibbs 抽样的 LDA 模型中,这 2 个概率值分别通过 $\hat{\phi}_w^{(z=j)}$ 和 $\hat{\theta}_{z=j}^{(s)}$ 估计.

### 3.4 文摘算法

1) 将文档集合中的文本分割为句子,去标题、时间等信息,提取正文中的句子;以文档集中的整句作为 LDA 中的文档,去标点和停用词,并将其转换为 LDA 的输入格式;

2) 为每个文档集合建立一个 LDA 模型,使用 Gibbs 抽样估计句子的主题分布 $\theta$ 和主题的词汇分布 $\phi$;

3) 计算主题重要度,根据提出的 2 种句子权重的计算方法 ProbGenSum 和 SentGenSum 分别计算句子权重;

4) 按照步骤 3) 得到的权重对句子进行排序,相同权重的句子按照非停用词在句子中所占的比例从大到小排序;

5) 从句子序列中由前至后抽取句子作为文摘句,若当前句子与前面句子的主题相同,则过滤当前句子,直到文摘达到长度限制.

## 4 实 验

实验中使用通用型文摘测试集 DUC2002 语料库作为多文档摘要的测试数据. DUC2002 语料库包含 59 个描述同一个主题或相关主题的文档集合,每个文档集合平均包含 10 个文档. 每个文档集合都给出了最大词数分别为 200 和 400 的抽取式专家文摘. 实验中根据提出的文摘算法分别为每个文档集合建立 LDA 模型,生成长度至多为 200 和 400 个词的抽取式文摘,并使用 DUC 评测工具 ROUGE[17] 自动评测文摘结果.

### 4.1 模型参数设置

由于 DUC2002 中各个文档集的词汇数、词汇记号数、句子数各不相同,每个文档集的 LDA 模型参数需要单独设定. 对各文档集来说可变量包括 $\alpha$、$\beta$ 和主题数目 $K$. 一种较好的超参数 $\alpha$ 和 $\beta$ 的选择方式应与主题的数量和词汇表的尺寸相关. 本文中 $\alpha$ 根据主题数目变化,取经验值 $\alpha=50/K$,$\beta$ 取固定的经验值 $\beta=0.01$[18].

LDA 模型的性能受到主题数目的影响,需预先设定主题数目 $K$. 确定主题数目的方法有多种:使用非参数化主题模型 HDP (hierarchical dirichlet process) 的方法[19]、使用层次聚类的方法[20] 和使用模型混乱度分析的方法[21] 等. 本文使用混乱度确定模型主题数目.

文档集上模型的混乱度为文档集中包含的各句子相似性(likelihood)几何均值的倒数[10],模型混乱度随着句子相似性的增加而单调递减:

$$\text{perplexity}(D)=\exp\left\{\frac{\sum_{S=1}^{N}\lg p(S)}{\sum_{S=1}^{N}N_S}\right\}. \tag{8}$$

式中:$N$ 为文档集中句子的个数,$N_S$ 为句子 $S$ 中词项的个数,$p(S)$ 为句子 $S$ 的相似性. LDA 模型中句子的相似性由句子的主题分布和主题的词汇分布计算:$\lg p(S)=\sum_{n=1}^{N}n(w,S)\hat{\phi}_w^{(z=j)}\hat{\theta}_{z=j}^{(s)}$,其中 $n=(w, S)$ 为句子 $S$ 中词 $w$ 的出现次数. 图 1 给出了在 DUC2002 测试集的 59 个句子集上建立的 LDA 模型的混乱度随主题数目变化的曲线.

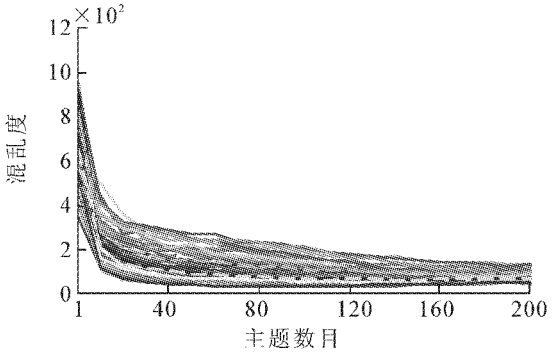



图 1 DUC2002 句子集上的模型混乱度随主题数目变化的趋势

Fig. 1 Variation of perplexity on different number of topics for the LDA model on the DUC2002 data set.

可以看出,随着主题数目的增加,所有句子集合的混乱度都收敛到一个较小的值,实验中当主题数目 $K=170$ 时所有句子集合的平均混乱度达到最小值. 混乱度越低,说明模型的泛化能力越强,因此对于整个 DUC2002 语料库来说,主题数目 $K=170$ 时模型最优.

对于单个句子集合来说,当其使得模型混乱度最低的主题数目小于 170 时,主题集合中会包含一

些在任何句子中都不出现的主题,从而影响了文摘模型的性能. 本文中对各个句子集分别选择使得其模型混乱度最低的主题数目作为各句子集上 LDA 模型的主题数目. 图 2 中给出了最终确定的 59 个句子集各自的主题数目.

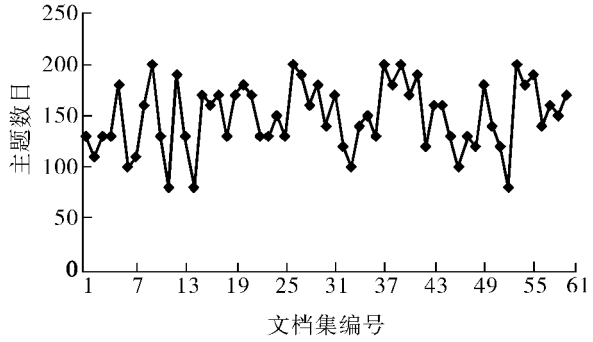



图 2　DUC2002 数据集上由混乱度确定的 LDA 模型的主题数目

Fig. 2　The optimal number of topics for the LDA model determined by perplexity on DUC 2002 data set.

### 4.2　基于 ROUGE 的自动评测

实验中使用 ROUGE-1、ROUGE-2、ROUGE-L、ROUGE-S4(中间有 4 个词间段的词对)和 ROUGE-SU4[17] 5 个评测标准,分别用带停用词和去停用词 2 种计算方式对提出的文摘方法进行评测. 专家文摘和模型生成的文摘都使用 Porter Stemmer 取词干. 对模型在 DUC2002 数据集上生成的长度分别为 200 和 400 的摘要分别进行评测,以考察摘要的长度对摘要质量的影响. 实验中同时给出了用于比较的 SumBasic、Doc-LDA 和 KL-LDA 在 DUC2002 数据集上的 ROUGE 结果.

#### 4.2.1　SumBasic 算法

SumBasic 算法是由 Nenkova 和 Vanderwende 于 2005 年提出的基于词频的多文档抽取式文摘方法[5]. 他们认为文档集合中非停用词的相对频率可以较为准确地反映该词是否出现在专家文摘中. 在 SumBasic 算法中每个句子 $S$ 都赋予一个反映它所包含的词频的权值:

$$\mathrm{Score}(S) = \sum_{w\in S}\frac{1}{|S|}P_D(w). \qquad (9)$$

式中:$P_D(w)$为一元概率观察值,使用最大似然估计计算时近似等于该词在语料库中出现次数占总词数的比例.

根据式(9)计算句子的分值,并按分值将句子由高到低添加到文摘中,直到达到限制的文摘字数,由该方法得到的模型记为 Unigram. 在 SumBasic 算法中已经选为文摘的句子中单词的概率变为原概率的平方:$P_D^{\mathrm{new}}(w) = P_D^{\mathrm{old}}(w)^2$,即选中单词的概率逐渐变小,从而降低文摘的冗余度. 虽然 SumBasic 算法的思想非常简单,但取得了不错的效果[11].

#### 4.2.2　Doc-LDA 算法

Arora 在文献[12]中将 LDA 作为文档的表示模型,属于同一话题的各文档表示为主题的分布,主题表示为词的分布. 文摘算法中根据主题概率大小排序,然后从大到小选择主题,再从主题中选择概率大的句子作为文摘句. 实验中为每个话题按照混乱度估计了最优的主题个数,选取其中性能最好的基于推论的句子权重计算方式作为比较实例,本文中将其称为 Doc-LDA.

Doc-LDA 使用式(10)计算主题的概率值:

$$P(z_j) = \sum_{k=1}^{M} P(z_j \mid D_k) \times P(D_k). \qquad (10)$$

式中:文档概率 $P(D_k)$假设为一常数,因此主题的概率只与主题在各文档中所占的权重有关. 在确定主题后句子的概率值由式(11)计算.

$$P(S_r \mid z_j) = \sum\nolimits_{W_i\in S_r} P(W_i \mid z_j) \times P(z_j \mid D_B) \times P(D_B). \qquad (11)$$

同样地,文档概率 $P(D_B)$为一常数,主题中句子的概率与主题的概率和句子中包含的词在主题下的概率相关.

#### 4.2.3　KL-LDA 算法

Chang 和 Chien[16] 为语料库中的文档集和单个的句子分别使用 LDA 建模,然后计算句子语言模型和文档集语言模型之间的 KL-散度.

其中,句子语言模型表示为

$$p(w_n \mid S) = \sum_{k=1}^{K} P(z_k \mid S) \times P(w_n \mid z_k).$$

文档集语言模型表示为

$$p(w_n \mid D) = \sum_{k=1}^{K} P(z_k \mid D) \times P(w_n \mid z_k).$$

使用式(12)的 KL-散度计算公式估计句子代表文档的能力,对句子进行排序,选择 KL-散度大的句子作为文摘句.

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_i P(i)\lg\frac{P(i)}{Q(i)}. \qquad (12)$$

实验中为每个话题和句子集按照混乱度估计了最优的主题个数,按 KL-散度大小对句子进行排序,选择 KL-散度大的句子作为该文档集的摘要句.

#### 4.2.4　评测结果与分析

表 1 给出了文摘长度为 200 时在 DUC2002 语料库上各模型得到的 ROUGE 值. 可以看出,根据 ROUGE 的 5 个评测标准判断的各模型性能的好坏基本是一致的. 基于 LDA 主题模型的文摘总体上优于基于词频统计的文摘效果. 用 LDA 分别表示句子

和文档进行KL-散度计算的文摘模型的效果在4种基于LDA模型的文摘系统中效果最不理想. 而其他3种模型的效果相当,但本文提出的基于句子LDA概率生成模型的文摘系统的效果要优于基于文档LDA建模的文摘系统.

表1 文摘长度为200个词时在DUC2002语料库上各模型的ROUGE值

Table 1 ROUGE-Scores of models on DUC2002 when summary length limited to 200 words

| 模 型 | ROUGE带停用词 | | | | | ROUGE去停用词 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | R-1 | R-2 | R-L | R-S4 | R-SU4 | R-1 | R-2 | R-L | R-S4 | R-SU4 |
| Unigram | 40.6 | 12.9 | 37.6 | 12.6 | 17.3 | 29.4 | 10.4 | 27.9 | 8.5 | 12.1 |
| SumBasic | 40.9 | 9.8 | 37.5 | 9.7 | 14.9 | 30.9 | 7.4 | 29.6 | 6.0 | 10.2 |
| Doc-LDA | 43.4 | 15.3 | 39.1 | 14.6 | 19.4 | 32.2 | 13.3 | 30.0 | 10.9 | 14.6 |
| KL-LDA | 40.8 | 13.3 | 37.4 | 12.9 | 17.0 | 28.9 | 10.7 | 27.4 | 9.1 | 12.4 |
| ProbGenSum | 44.8 | 17.8 | 41.2 | 17.5 | 22.1 | 33.7 | 15.8 | 31.9 | 13.8 | 17.1 |
| SentGenSum | 45.2 | 17.8 | 41.4 | 17.2 | 21.9 | 34.8 | 15.6 | 32.8 | 13.7 | 17.2 |

表2给出了文摘长度为400个词时在DUC2002语料库上各模型得到的ROUGE值. 从表1和表2中的数据可以看出当文摘长度增加时各模型的ROUGE值都有所提升,这是因为ROUGE是基于召回率的评测,文摘越长分值越有可能会高. 文摘长度为400时SumBasic与Unigram性能的相对好坏与长度为200时大致相同,但ProbGenSum在所有ROUGE值上的结果均优于SentGenSum,说明句子生成模型SentGenSum对于生成短文档摘要比概率生成模型具有优势,而对于较长的摘要则不如概率生成模型. 基于KL-散度的文摘模型的效果仍是效果最差的,而基于文档LDA建模的文摘系统效果优于句子生成模型. 在文摘长度为400个词时,基于主题模型的文摘性能同样明显优于基于词的文摘性能,其中概率生成模型在所有模型中是性能最好的.

表2 文摘长度为400个词时在DUC2002语料库上各模型的ROUGE值

Table2 ROUGE-Scores of models on DUC2002 when summary length limited to 400 words

| 模 型 | ROUGE带停用词 | | | | | ROUGE去停用词 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | R-1 | R-2 | R-L | R-S4 | R-SU4 | R-1 | R-2 | R-L | R-S4 | R-SU4 |
| Unigram | 49.4 | 19.6 | 46.7 | 19.0 | 24.1 | 37.4 | 15.9 | 35.8 | 13.5 | 17.5 |
| SumBasic | 49.2 | 15.6 | 46.4 | 15.2 | 20.9 | 37.8 | 11.9 | 36.4 | 9.5 | 14.3 |
| Doc-LDA | 53.5 | 25.4 | 50.4 | 24.6 | 29.5 | 42.7 | 22.7 | 40.9 | 19.7 | 23.6 |
| KL-LDA | 47.0 | 19.0 | 44.8 | 18.8 | 23.6 | 33.3 | 16.1 | 32.4 | 13.7 | 17.0 |
| ProbGenSum | 53.7 | 26.9 | 50.8 | 26.3 | 30.9 | 43.7 | 24.5 | 41.6 | 21.7 | 25.3 |
| SentGenSum | 52.4 | 23.6 | 49.4 | 23.0 | 27.9 | 41.3 | 20.7 | 39.6 | 18.0 | 21.9 |

实验还考察了各模型对文摘句子长度的偏好. 表3给出了各模型产生的文摘中句子的数量. 可以看出,与Unigram相比,SumBasic倾向于选择短句子,而4个主题模型则都倾向于选择长句子作为文摘句,其中句子生成模型选择的句子长度与专家摘要的句子长度最相近. 基于句子的概率生成模型(ProbGenSum)和基于文档的概率生成模型(Doc-LDA)都选择较长的句子作为文摘句,这是因为文摘系统中没有为句子做归一化所致. 基于KL-散度的文摘模型(KL-LDA)则选择了较短的句子.

表3 各模型产生的文摘中句子的数量

Table 3 Number of sentences in summary generated by each system

| 摘要长度 | Unigram | SumBasic | ProbGenSum | SentGenSum | Doc-LDA | KL-LDA | Reference |
|---|---|---|---|---|---|---|---|
| 200 | 14.80 | 17.15 | 6.15 | 8.03 | 6.15 | 9.58 | 8.53 |
| 400 | 25.14 | 30.69 | 12.41 | 15.59 | 11.98 | 19.42 | 16.59 |

## 5　结束语

本文基于LDA模型中的主题概率分布和句子概率分布提出了2种句子权重的计算方法:1)使用生成性的主题模型计算句子中单词概率的方法;2)将文档集中的句子看作文档集生成模型的方法.在通用型文摘数据集DUC2002上,使用ROUGE评测工具得到的实验结果表明,这2种句子权重计算方法都取得了明显优于传统方法的效果,比其他基于LDA的文摘系统也有优势.LDA模型中做了词袋假设,它没有考虑单词和句子的位置,也没有考虑句子、文档和文档集合之间的结构关系,以后的工作将在主题模型中纳入句法结构信息,并为句子、文档和文档集合建立统一的概率生成框架.另外,目前在句子生成模型中作了句子先验概率相同的假设,以后的工作也将考虑纳入句子先验概率的影响.

## 参考文献:


[1]RADEV D R, HOVY E, MCKEOWN K. Introduction to the special issue on text summarization[J]. Computational Linguistics,2002,28(4):399-408.

[2]LEE J H, SUN P, AHN C M, et al. Automatic generic document summarization based on non-negative matrix factorization[J]. Information Processing and Management, 2009,45(1):20-34.

[3]徐永东,徐志明,王晓龙. 基于信息融合的多文档自动文摘技术[J].计算机学报, 2007,30(11):2048-2054.
XU Yongdong, XU Zhiming, WANG Xiaolong. Multi-document automatic summarization technique based on information fusion[J]. Chinese Journal of Computers, 2007, 30(11):2048-2054.

[4]HIRAO T, ISOZAKI H, MAEDA E, et al. Extracting important sentences with support vector machines[C]//Proc of the 19th International Conference on Computational Linguistics. Taipei, China, 2002: 1-7.

[5]NENKOVA A, VANDERWENDE L. The impact of frequency on summarization: MSR-TR-2005-101[R]. Redmond, USA: Microsoft Research,2005.

[6]LINC Y, HOVY E. The automated acquisition of topic signatures FOR text summarization[C]//Proc of the 18th International Conference on Computational Linguistics. Sarbrficken, Germany,2000:271-278.

[7]ANTIQUEIRA L, Jr OLIVEIRA O N. A complex network approach to text summarization[J]. Information Science, 2009 (179):584-599.

[8]WAN X J, YANG J W. Multi-document summarization using cluster-based link analysis[C]//Proc of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Sheffield, UK, 2008:299-306.

[9]HARABAGIU S, HICKL A, LACATUSU F. Satisfying information needs with multidocument summaries[J]. Information Processing and Management, 2007,43(6):1619-1642.

[10]BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003(3):993-1022.

[11]HAGHIGHI A, VANDERWENDE L. Exploring content models for multi-document summarization[C]//Human Language Technologies: the Annual Conference of the North American Chapter of the ACL Boulder. Colorado, 2009:362-370.

[12]ARORA R, RAVINDRAN B. Latent Dirichlet allocation based multi-document summarization[C]//Proc of the Second Workshop on Analytics for Noisy Unstructured Text data. Singapore, 2008:91-97.

[13]ARORA R, RAVINDRAN B. Latent Dirichlet allocation and singular value decomposition based multi-document summarization[C]//Proc of Eighth IEEE International Conference on Data Mining. Pisa, Italy, 2008:713-718.

[14]CHEN Y T, CHEN B, WANG H M. A probabilistic generative framework for extractive broadcast news speech summarization[J]. IEEE Trans on Audio, Speech, and Language Processing,2009,17(1): 95-106.

[15]SHAFIEI M M, MILIOS E E. Latent Dirichlet co-clustering[C]//Proceedings of the Sixth International Conference on Data Mining (ICDM). Hong Kong, China, 2006: 542-551.

[16]CHANG Y L, CHIEN J T. Latent Dirichlet learning for document summarization[C]//IEEE International Conference on Acoustics, Speech, and Signal Processing. Taipei, China, 2009:1689-1692.

[17]LIN C Y. ROUGE: a package for automatic evaluation of summaries[C]//Workshop on Text Summarization Branches Out. [S.l.],Spain, 2004: 74-81.

[18]STEYVERS M, GRIFFITHS T. Probabilistic topic models [C]//Handbook of Latent Semantic Analysis. Laurence Erlbaum, 2007:1-15.

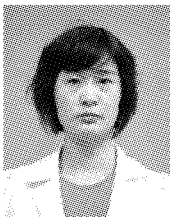
[19]TEH Y W, JORDAN M I, BEAL M J, et al. Hierarchical Dirichlet processes[J]. Journal of the American Statistical Association, 2006, 101(476):1566-1581.

[20]秦　兵,刘　挺,李　生. 基于局部主题判定与抽取的多文档文摘技术[J]. 自动化学报,2004,30(6):905-910.
QIN Bing, LIU Ting, LI Sheng. Multi-document summarization based on local topics identification and extraction [J]. Acta Automatica Sinica, 2004,30(6): 905-910.

[21]石　晶,胡　明,石　鑫,等. 基于LDA模型的文本分

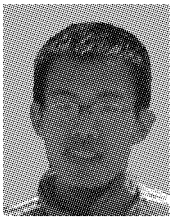
割[J]. 计算机学报, 2008.31(10):1865-1873.
SHI Jing, HU Ming, SHI Xin, et al. Text segmentation based on model LDA[J]. Chinese Journal of Computers, 2008,31(10):1865-1873.

**作者简介:**



杨 潇,女,1981 年生,博士,主要研究方向为自然语言处理. 发表学术论文 10 余篇.

马 军,男,1956 年生,教授,博士生导师,主要研究方向为算法分析与设计、信息检索和并行计算. 曾主持 2 项国家"863"计划金项目课题,1 项国家自然基金课题,2 项教育部基金课题和多项省基金课题. 发表学术论文 80 余篇.



杨同峰,男,1985 年生,博士研究生,主要研究方向为个性化检索和图像标注.

# 第 1 届网络与分布式计算国际会议(ICNDC2010)
# The First International Conference on Networking and Distributed Computing

网络与分布式技术是当前乃至未来的 IT 技术的重要组成部分. 当下一代网络成为现实,移动通信系统将来发展到 3G 甚至 4G 时代时,将会出现传统软件向网络应用转变的趋势. 为了使工业界和学术界共同研讨网络与分布式计算的热点话题和发展趋势,我们将于 2010 年 10 月 21 日至 24 日在中国杭州举办第一届网络与分布式计算国际会议. 会议重点:(1)分布式计算和分布式系统方面,包括集群和网格、服务组合和业务流程、点对点对等系统、云计算等;(2)网络方面,包括 IP 网络、下一代互联网、无线网络、无线网格网络、4G 移动通信等;(3)分布式应用系统方面,包括分布式电子商务应用、分布式电子科学应用、管理应用系统等.

**一、征文范围**

包括分布式计算和布式系统,网络和分布式应用 3 个分支.

1. 分布式计算和分布式系统:集群和网格运算, 协同计算,分布式数据存储,服务组合和编排,分布式多媒体系统,点对点对等系统,云计算,分布式资源管理,多代理系统,中间件技术,服务虚拟化,并行与分布式处理,面向服务的构架,效用计算,面向服务的计算,复杂的分布式系统,面向服务的软件和系统工程,语义网络,社会网络,传感器网.

2. 网络追踪: 主动/可编程网络, 移动/无线网络仿真, 自适应网络, 多跳无线网络, 分布式网络管理, 无线局域网,城域网,广域网,3G 和智能网络, 无线网状网络, 互联网络架构, 多媒体网络, 无线多播,传感器网络, 网络隐私和安全, 无线网络管理, 网络服务质量和性能评价, 无线协议和架构, 下一代网络, 新型网络架构和协议, 按需网络,语境意识网络,3G 和 4G 带宽需求,普及的计算机运作,移动和无线 IP.

3. 分布式应用程序: 业务流程整合,业务流程管理,企业资源规划,企业流程管理,协同电子商务,企业联合会,企业集成,全球企业,虚拟/网络企业,供应链合作,电子物流,代理介导电子商务, B2B, B2C, C2C 模式,电子银行,电子商务,移动商务,商务数据挖掘,业务学习机,自适应业务, 按需电子商务.

**二、征文要求**

1. 论文须未在国内外公开发行的刊物发表或会议上宣读过,内容具体,突出作者的创新与成果,具有较重要的学术价值或应用推广价值.

2. 所有论文必须是英文文稿,全文不能超过 5 页,投稿稿件请用 Word 或 PDF 格式排版. 论文递交的文本格式:请参照 http://www.inetdc.org/meeting/icndc2010/(page submission).

3. 如果论文一经录用,至少有一位作者可以注册和参加本次会议. 所录用的论文将会被 IEEE CPS 出版,EI 和 ISTP 收录引用,由 IEEE CSDL 所存档.

4. 所有论文文稿应提交以下电子提交系统 https://www.easychair.org/login.cgi? conf = icndc2010.

5. 如有任何疑问请联系 icndc2010@inetdc.org

会议网站:http://virgo.sourceforge.net/meeting/icndc2010/