基于主题模型与信息熵的中文文档自动摘要技术研究

李 然 张华平 赵燕平 商建云
(北京理工大学计算机学院 北京 100081) (北京理工大学管理与经济学院 北京 100081)
(北京理工大学软件学院 北京 100081)

摘 要 提出了一种基于LDA模型以及信息熵的文档自动摘要技术,即通过LDA模型对文档进行浅层语义分析,得到文档的主题分布以及不同主题下的词语分布;通过对主题的分析,可以得到最能代表文档中心思想的主题,以及该主题下的词语分布。同时,提出了一种新的基于信息熵的度量句子重要性的方法,并将该方法应用于文档的关键句抽取过程中。该方法将文档中句子的出现看成一个随机变量,通过对随机变量建模并度量它的信息熵来选取文档中的关键性语句。实验结果表明,应用主题模型与信息熵摘取的文档摘要能有效地从文档中摘出中心句。

关键词 摘要,LDA模型,主题,信息熵
中图法分类号 TP391 **文献标识码** A

Automatic Text Summarization Research Based on Topic Model and Information Entropy

LI Ran ZHANG Hua-ping ZHAO Yan-ping SHANG Jian-yun
(School of Computer Science, Beijing Institute of Technology, Beijing 100081, China)
(School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China)
(School of Software, Beijing Institute of Technology, Beijing 100081, China)

Abstract This paper presented a method for automatic summarization based on LDA model and information entropy for Chinese document. It uses LDA model to do shallow semantic analysis work on documents and gets the distribution of topics under each document. Through analyzing the topics of document, we got the topic which has the best expression of central idea for document. Meanwhile, this paper proposed a new method to compute the sentence weight and extract the most important sentence based on measuring the information entropy for each sentence. It treats the sentence as a random variable and calculates the information entropy for every random variable. Experimental results show that this method can pick out the most important sentence in the document.

Keywords Summarization, LDA, Topic, Information entropy

1 引言

摘要,是指按照文档的中心思想以简洁的形式准确地表达文档的主要内容。根据自动摘要是否来源于原文,可将自动摘要分为抽取型摘要和概括性摘要。抽取型摘要是指从原文中直接摘取有代表性的句子作为文档的摘要。通常抽取型摘要是将文档看成是一个句子的集合,通过一系列算法选取这个句子集合中的句子作为文档的摘要。抽取型摘要的结果主要依赖于算法的选择,好的算法通常可以准确地找出文档中的主旨句来生成文档的摘要。此外抽取型摘要通常有不受领域限制的特点。概括型摘要主要是通过对原文进行深层次分析以及根据领域知识库进行信息抽取,再利用自然语言生成技术对句子进行语义分析,然后通过语言学知识记忆自然语言技术生成文档摘要。

文档自动摘要的研究开始于50年前,当时Luhn通过统计词频来计算词语的权重,通过词语权重来计算句子权重,并按照权重选取特定的句子作为文档的摘要[1]。国外对文档摘要技术的研究具有很长的历史并取得了较大的发展。Edmundson[2,3],Pollock,Joseph,Antonio Zamor[4],Paice[5]等选取字词的不同特征作为提取摘要的关键。但该方法依赖于特定文体的自此特征选取,通用性较差。后来的学者又尝试引入文档的篇章结构特征以及相似性分析等手段来选取文档的摘要。Salton和Gerard[6]通过文章的结构,以段落为单位对文档进行分析,通过度量段落间的相似度来获得度量段落的重要性。但该方法依赖于文章的篇章结果,对简单篇章结构的文档则适用性较差。Sasha,Blair-Goldensohn等在DUC2004上采用了一种叫做SC的方法[7],该算法的核心思想是通过句子聚类的结果来度量不同类别的重要性。包含句子越多的类别被认为是越重要的类别,然后通过抽取类中的有代表性的句子作为文档的摘要,其中聚类的过程中用VSM模型表征句子,用向量间的Cosine值来度量相似性。该方法

李 然(1987—),男,硕士生,主要研究方向为自然语言处理、机器学习;张华平(1978—),男,副教授,硕士生导师,主要研究方向为大数据搜索与挖掘、自然语言处理、信息检索与信息安全;赵燕平(1956—),女,教授,硕士生导师,主要研究方向为网络数据计算与挖掘;商建云(1965—),女,副教授,硕士生导师,主要研究方向为网络数据计算与挖掘。

对文档的分析停留在词法分析结构，且 VSM 模型表示句子的过程中会造成维度灾难，训练代价过高。

国内对中文自动摘要的研究起步较晚，开始于上世纪 80 年代。1988 年上海交大研制了汉语文献自动编制实验系统，该系统已能对科技文献进行摘要并取得了一定效果，经过这些年的发展，中文自动摘要技术已经取得了长足的进展。王继成等[8]提出了一种基于篇章结构的中文 Web 文档自动摘要技术，即依次通过篇章结构分析、词语权重计算、关键词提取并统计句子的权重来最终生成摘要。张奇等[9]提出基于句子相似度方法得到文档摘要，他们在度量句子相似度的时候考虑了一元、二元和三元的信息，并通过一种回归的方法将这几种相似度结合起来。该方法引用了统计机器学习的方法，考虑了词语的位置信息，有利于关键词语的挖掘，但是对一些有价值的出现次数较少的词语，如人名、地名等不能很好地识别其重要性，从而抽取到和这些词语相关的主题句的概率会降低。尹存燕等[10]将传统的抽取型摘要方法应用于 Web 文本上并取得了不错的效果。张云涛等[11]提出了基于各个主题的摘要抽取技术，即根据每个主题的重要性选取不同数量的代表该主题的句子作为文档摘要。但该方法依赖于主题句的选取，由主题句确定主题的个数，选取的主题需要人工指定，需要人工干预。纪文倩等[12]提出了一种基于 LexRank 改进算法的自动文摘系统，她们提出了基于 LexRank 和指示性词语特征以及句子位置的句子权重计算方法，通过计算句子的权重，得到文档的摘要。罗文娟等[13]通过有监督的机器学习的方式得到文档的摘要，他们选取熵和相关度等性质作为句子的特征。这种方法属于有监督训练，需要训练样本和测试样本在相同领域才能有较好的效果，领域通用性较差。任昭春等[14]通过对论坛文档的动态建模来分析文档的摘要。文章引入了主题模型，但是由于要考虑论坛的回帖等相关信息，对文档的属性有一定要求，无法应用到传统的普通文档上。刘平安[15]引入了 HLDA 模型作为多文档的主题摘要，在选取主题句的过程中它使用句子中所包含的主题的数量和质量来度量句子的重要性。此方法虽然考虑了主题信息，但是抽取的中心句多为长句。对于中心主旨句为短句的情况，如新闻等文体，抽取性较差且需要设定额外的规则来进行过滤匹配。

本文为了克服一些方法在提取摘要时需要的规则多、通用性差等缺点，提出了一种基于主题模型的无监督的文档摘要算法。通过 LDA 模型获取出文档集中每一篇文档的主题分布和每个主题对应的词语分布。同时根据主题分布的权重，选取与文档最相关的主题来挖掘文本的浅层语义。同时为了将主题信息应用到摘要句的提取工作中，文本提出了一种基于信息熵的方法度量句子重要程度的度量方法。该方法对句子这一随机变量进行概率模型建模，并根据该模型计算出句子出现的概率值，以此计算句子的信息熵。最终根据信息熵来度量句子的重要程度。同时，考虑到句子字数对句子权重的影响，本文设定句子最低字数的阈值，过滤低于该阈值的句子。

2 基于主题模型与信息熵的文档自动摘要技术

2.1 主题模型

主题模型(Topic Model)在机器学习和自然语言处理等领域是用来在一系列文档中发现抽象主题的一种统计模型。直观来讲，如果一篇文章有一个中心思想，那么一些特定词语会更频繁地出现。比方说，如果一篇文章是讲狗的，那么“狗”和“骨头”等词出现的频率会高些。如果一篇文章是讲猫的，那么“猫”和“鱼”等词出现的频率会高些。而有些词例如“这个”、“和”大概在两篇文章中出现的频率会大致相等。但真实的情况是，一篇文章通常包含多种主题，而且每个主题所占比例各不相同。因此，如果一篇文章 10% 和猫有关，90% 和狗有关，那么和狗相关的关键字出现的次数大概会是和猫相关的关键字出现次数的 9 倍。一个主题模型试图用数学框架来体现文档的这种特点。主题模型自动分析每个文档，统计文档内的词语，根据统计的信息来断定当前文档含有哪些主题，以及每个主题所占的比例各为多少[17,18]。

隐含狄利克雷分布(Latent Dirichlet allocation，LDA)是目前一种比较流行的主题模型，也是一种典型的词袋模型，即它认为一篇文档是由一组词构成的一个集合，词与词之间没有顺序以及先后的关系。一篇文档可以包含多个主题，文档中每一个词都由其中的一个主题生成。

即一篇文档的多个主题之间是假设服从多项式分布，而一个主题之间的所有词语也是假设服从主题分布。此外，采用贝叶斯估计的方法，假设文档的主题分布的先验分布是服从狄利克雷分布，主题的词语分布的先验分布同样是服从狄利克雷分布，如图 1 所示。主题模型的文档生成过程如下[17,18]：

a)从狄利克雷分布 α 中取样生成文档 i 的主题分布 θ_i；

b)从主题的多项式分布 θ_i 中取样生成文档 i 第 j 个词的主题 $z_{i,j}$；

c)从狄利克雷分布 β 中取样生成主题 $z_{i,j}$ 的词语分布 $\phi_{z_{i,j}}$；

d)从词语的多项式分布 $\phi_{z_{i,j}}$ 中采样最终生成词语 $\omega_{i,j}$。

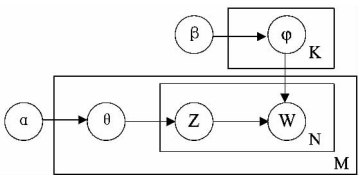

图 1 LDA 模型生成图

因此，整个模型中的所有可见变量以及隐含变量的联合分布是：

$$p(\omega_i,z_i,\theta_i,\Phi|\alpha,\beta)=\prod_{j=1}^{N}p(\theta_i|\alpha)p(z_{i,j}|\theta_i)p(\Phi|\beta)p(w_{i,j}|\theta_{z_{i,j}})$$

Gibbs Sampling 的具体过程如下：

1. 首先对所有文档中的所有词遍历一遍，为其都随机分配一个主题，即 $z_{m,n}=k\sim Mult(\frac{1}{K})$，其中 m 表示第 m 篇文

档，n 表示文档中的第 n 个词，k 表示主题，K 表示主题的总数，之后对应的 $n_m^{(k)}+1$，n_m+1，$n_k^{(t)}+1$，n_k+1 分别表示在 m 文档中 k 主题出现的次数、m 文档中主题数量的和、k 主题对应的 t 词的次数、k 主题对应的总词数。

2. 之后对下述操作进行重复迭代。

3. 对所有文档中的所有词进行遍历，假如当前文档 m 的词 t 对应主题为 k，则 $n_m^{(k)}-1$，n_m-1，$n_k^{(t)}-1$，n_k-1 即先拿出当前词，之后根据 LDA 中 topic sample 的概率分布 sample 出新的主题，在这个新主题 k 上所对应的各种计数 $n_m^{(k)}$，n_m，$n_k^{(t)}$，n_k 分别做加一操作。topic sample 的概率分布的计算公式如下：

$$p(z_i=k\mid z_{-i},\omega)\propto(n_{k,-i}^{(t)}+\beta_t)(n_{m,-i}^{(k)}+\alpha_k)/(\sum_{t=1}^{V}n_{k,-i}^{(t)}+\beta_t)$$

4. 迭代结束后根据所得主题的分布情况，对模型的参数进行估计，参数的估计的公式为：

$$\phi_{k,t}=(n_k^{(t)}+\beta_t)/(\sum_{t=1}^{V}n_k^{(t)}+\beta_t)$$

$$\theta_{m,k}=(n_m^{(k)}+\alpha_k)/(\sum_{k=1}^{K}n_m^{(k)}+\alpha_k)$$

2.2 信息熵

为了度量句子的重要性，我们引入了信息熵这一度量单位。在信息论中，熵用来衡量一个随机变量出现的期望值。对于一个值域为 $\{\chi_1,\cdots,\chi_n\}$ 的随机变量 X 的熵值 H 定义为：

$$H(X)=E(I(X))$$

其中，$I(X)$ 为随机变量 X 的自信息。同时，根据期望的定义以及自信息的公式展开，得到熵值 H 的另一种表现形式为：

$$H(X)=\sum_{i=1}^{n}p(\chi_i)I(\chi_i)=-\sum_{i=1}^{n}p(\chi_i)\log_b p(\chi_i)$$

在本文中，信息熵用来度量句子以某种词语组合方式出现这个随机变量的平均期望值，该随机变量在建模时采用指示型随机变量建模方法，即设定它的值域为二值，即{出现，不出现}。并根据该随机变量在上述值域上取值的概率计算句子的信息熵。

2.3 句子信息熵的计算方法

文本用信息熵度量句子的权重，对文档的每一个句子，本文对文档中的词语做独立性假设，认为每个词语的出现与其他词语的出现无关。因此，对于一个句子，它在一篇文档中出现的概率为：

$$p(sentence\mid topic_{(j)})=\prod_{i=1}^{m}p(token_{(i)}\mid topic_{(j)})$$

其中，$token_{(i)}$ 为句子的第 i 个词语，m 为当前句子中词语的个数，$topic_{(j)}$ 则是当前文档主题分布中概率最高的主题。$p(token_{(i)}\mid topic_{(j)})$ 为 LDA 模型训练获得的在当前主题下特定词语出现的概率值，即 $\phi_k=topic_{(j)}$，$t=token_{(i)}$。

而本文将句子以某种词语组合的方式出现看成是一个随机变量，该随机变量的值域为{出现，不出现}，则该随机变量的信息熵的计算公式为：

$$E(sentence)=$$

$$p(sentence\mid topic_{(j)})\cdot\log(\frac{1}{p(sentence\mid topic_{(j)})})+$$

$$\bar{p}(sentence\mid topic_{(j)})\cdot\log(\frac{1}{\bar{p}(sentence\mid topic_{(j)})})$$

其中，$E(sentence)$ 为某个句子的信息熵，$p(sentence\mid topic_{(j)})$ 为当前主题下句子出现的概率值，$\bar{p}(sentence\mid topic_{(j)})$ 为当前主题下该句子不以当前词语组合出现的概率值。

2.4 算法介绍

2.4.1 算法提出

传统的文档摘要系统通常计算词语的权重以及句子的相似性，忽略了文档的主题信息。而考虑文档的主题信息的摘要算法通常是基于主题句判断一篇文档的主题，而主题的判断依赖于主题句的确定。而对于一个文档集合，可将它看成是一个由不同的主题生成文档的过程，这里的主题类似于文档的类别。而对于文档集合中的一篇文档，它的文章中心思想通常就来源于一个或者少数几个的主题，因此我们在文档摘要的过程中的工作应主要聚焦在这一个或者几个主题上。同时，针对不同的主题，每个主题下面的词语的权重又是不相同的，如某个体育主题下的词语与体育相关的词语权重就应该高。除此之外，一些体育类的生僻词语，如人名、赛事名称等也应该能被这个体育类的主题识别出来并赋予比属于其他主题词语更高的生成概率，而 LDA 模型恰能完美地解决这些问题，它可以准确地得到文档的主题分布，并以概率的形式展示出主题的优先程度，同时又能正确地得到每个主题下面的词语分布，使一些隶属于该主题的生僻词语不因出现次数稀少而失去它的权重与辨识度。在标识某主题下词语的优先程度的时候，同样以词语生成概率的形式进行表征，概率高的词语有高的权重。这样，对于一些类别明确且较生疏的词语，也能很好地对其权重进行估计。

基于此，本文采用 LDA 模型对文档集合及文档做浅层语义分析，并得到文档集合中每篇文档的主题分布以及相应主题的词语分布。通过过滤主题，得到文档中词语的权重。同时考虑到用信息熵度量句子权重时对超短句赋予的权重过高，与现实情况并不完全吻合，本文对文档主旨句的字数做阈值设定，只选取字数多于该阈值的句子作为文档的主旨句。

2.4.2 算法流程

综上，基于主题模型及信息熵的摘要算法流程如下，流程图如图 2 所示：

1. 对文档集合中的文档进行中文分词，并将文档转化为词语空间向量。

2. 对上述的空间向量运行 Gibbs sampling 工作，得到文档的主题分布。

3. 对于每篇文档选取主题分布中概率最高的主题，并根据选取的主题获得对应的词语概率分布。

4. 对于文档中的句子，计算每个句子的信息熵，并得到句子的权重。

5. 根据句子的权重，过滤句子的字数限制，得到每篇文档的摘要。

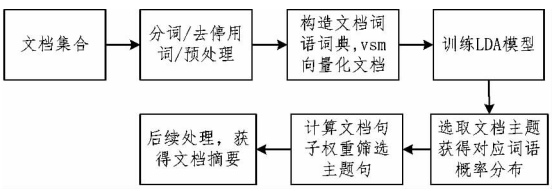

图 2 算法流程

（下转第 332 页）

Rare Events from Spatial Data Sets[J]. Geoinformatica, 2006, 10(3): 239-260

[16] Xiao X, Xie X, Luo Q. Density-based co-location pattern discovery[C]// Proc. of the 16th ACM International Conference on Advances in Geographic Information Systems. Irvine, California, 2008: 11-20

[17] Huang Yan, Pei J, Xiong H. Mining Co-location Patterns with Rare Events from Spatial Data Sets [J]. Geoinformatica, 2006, 10(3): 239-260

[18] Wang Li-zhen, Wu Ping-ping, Chen Hong-mei. Finding Probabilistic Prevalent Co-locations in Spatially Uncertain Data Sets [J]. IEEE Transactions on Knowledge and Data Engineering (TKDE), 2011, 25(4): 790-804

[19] Sheng C, Hsu W, Lee M. Discovering Spatial Interaction Patterns [M]. Berlin: Springer, 2008: 95-109

[20] 欧阳志平，王丽珍，陈红梅. 模糊对象的空间 Co-Location 模式挖掘研究[J]. 计算机学报, 2012(10): 1947-1956

[21] Wang Li-zhen, Chen Hong-mei, Zhao Li-hong, et al. Efficiently Mining Co-Location Rules on Interval Data [C]// ADMA 2010, Part I. LNCS 6440, 2010: 477-488

(上接第 300 页)

3 实验结果

文本采用中科院 ICTCLAS 对中文文档进行分词处理，针对自然语言处理面向真实语料与面向实例化的趋势，文本的测试基于 300 篇真实文本。这些文本是由设计爬虫从新浪新闻频道随机爬取的各类新闻。文本的 LDA 模型的主题个数 k 设定为 200，超参数根据经验[17] 设定为 α 为 0.25，β 为 0.01，迭代次数设定为 200。并设定句子字数的阈值为 10，不选取低于该阈值的句子作为最终文档摘要的候选句。同时设定摘要字数的上限值为 200，在计算句子权重之后，根据权重选取权重从高到低的一个或多个句子作为文档的摘要，选取句子的个数依赖于已经选取的句子的字数，使最终的文档摘要总字数小于我们设定的上限值。最后将每篇文档的摘要与文档的内容与文档标题进行对比，并判断摘要与文档内容的相关程度。

文本采用人工打分对摘要结果进行评测，打分分为 3 个标准，分别是准确反映主题、基本反映主题、没有很好反映主题。同时，本文为了减少人为差异对最终统计结果的影响，最终的结果为去掉最高和最低项之后的均值。具体的测试结果如表 1 所列。

表 1 摘要测试结果

分类标准	评价结果	比例
准确反映主题	184	61.33%
基本反映主题	90	30.00%
没有很好反映主题	26	8.67%

结束语 文本提出了基于主题的文档摘要算法，通过主题得到文档中不同词语的生成概率。同时在得到了词语生成概率之后，本文对句子进行了概率建模，从而引入了信息熵来对句子的权重进行度量。为了验证该方法的效果，本文随机爬取了新浪新闻频道的若干新闻，并进行了实验。实验结果表明，在合适的模型参数的情况下，该方法抽取的摘要能较好地概括文档的主要内容。

参 考 文 献

[1] Luhn, Hans P. The automatic creation of literature abstracts [J]. IBM Journal of research and development, 1958, 2(2): 159-165

[2] Edmundson, Harold P, Wyllys R E. Automatic abstracting and indexing—survey and recommendations[J]. Communications of the ACM, 1961, 4(5): 226-234

[3] Edmundson, Harold P. New methods in automatic extracting [J]. Journal of the ACM(JACM), 1969, 16(2): 264-285

[4] Pollock, Joseph J, Zamora A. Automatic abstracting research at chemical abstracts service[J]. Journal of Chemical Information and Computer Sciences, 1975, 15(4): 226-232

[5] Paice, Chris D. The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases [C]// Proceedings of the 3rd Annual ACM Conference on Research and Development in Information Retrieval. Butterworth & Co., 1980

[6] Salton, Gerard, et al. Automatic text structuring and summarization[J]. Information Processing & Management, 1997, 33(2): 193-207

[7] Blair-Goldensohn, Sasha, et al. Columbia university at duc 2004 [C]// Proceedings of the Document Understanding Conference, DUC-2004. Boston, USA, 2004

[8] 王继成，武港山. 一种篇章结构指导的中文 Web 文档自动摘要方法[J]. 计算机研究与发展, 200340(3): 398-405

[9] 张奇，黄萱菁，吴立德. 一种新的句子相似度度量及其在文本自动摘要中的应用[J]. 中文信息学报, 2005, 19(2): 93-99

[10] 尹存燕，戴新宇，陈家骏. Internet 上文本的自动摘要技术[J]. 计算机工程, 2006, 32(3): 88-90

[11] 张云涛，龚玲，王永成. 基于综合方法的文本主题句的自动抽取[J]. 上海交通大学学报, 2006, 40(5): 771-774

[12] 纪文倩，等. 一种基于 LexRank 算法的改进的自动文摘系统[J]. 计算机科学, 2010, 37(5): 151-154

[13] 罗文娟，等. 权衡熵和相关度的自动摘要技术研究[J]. 中文信息学报, 2011, 25(5): 9-16

[14] 任昭春，马军，陈竹敏. 基于动态主题建模的 Web 论坛文档摘要[J]. 计算机研究与发展, 2013, 49(11): 2359-2367

[15] 刘平安. 基于 HLDA 模型的中文多文档摘要技术研究[D]. 北京：北京邮电大学, 2013

[16] http://zh.wikipedia.org/wiki/隐含狄利克雷分布

[17] Blei, David M, Ng A Y, et al. Latent dirichlet allocation[J]. the Journal of machine Learning research, 2003, (3): 993-1022

[18] Wei X, Croft W B. LDA-based document models for ad-hoc retrieval[C]// Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2006: 178-185