第 34 卷 第 22 期 计 算 机 工 程 2008 年 11 月
Vol.34 No.22 Computer Engineering November 2008
·软件技术与数据库· 文章编号：1000—3428(2008)22—0049—03 文献标识码：A 中图分类号：TP311

基于局部主题关键句抽取的自动文摘方法

徐 超[1]，王 萌[2]，何婷婷[3]，张 勇[3]

(1. 福建师范大学软件学院，福州 350007；2. 广西工学院计算机工程系，柳州 545006；3. 华中师范大学计算机科学系，武汉 430079)

摘 要：自动文摘是语言信息处理中的重要环节。该文提出一种基于局部主题关键句抽取的中文自动文摘方法。通过层次分割的方法对文档进行主题分割，从各个局部主题单元中抽取一定数量的句子作为文章的文摘句。通过事先对文档进行语义分析，有效地避免了数据冗余和容易忽略分布较小的主题等问题。实验结果表明了该方法的有效性。

关键词：自动文摘；主题分割；局部主题单元

Automatic Summarization Method Based on Extracting Sentences from Local Topics

XU Chao[1], WANG Meng[2], HE Ting-ting[3], ZHANG Yong[3]

(1. Faculty of Software, Fujian Normal University, Fuzhou 350007; 2. Department of Computer Engineering, Guangxi University of Technology, Liuzhou 545006; 3. Department of Computer Science, Huazhong Normal University, Wuhan 430079)

【Abstract】 Automatic summarization is an important issue in natural language processing. This paper proposes a new method for automatic summarization of Chinese text based on extracting sentences from subtopics. The document is segmented into several units in terms of the subtopics in the document. The most representative sentences in each subtopic unit are selected as the summary sentences. By analyzing semantic structure of the documents in advance, the summary sentences are not redundancy and the coverage of each subtopic is balanced. Experimental results show that the method is effective.

【Key words】 automatic summarization; topic segmentation; local topic unit

1 概述

随着信息技术的迅速发展，人们容易得到海量的数据。这些数据通常都是未经处理的，需要经过加工处理才能得到有用的信息，对这些数据进行加工处理越来越重要。自动文摘技术通过计算机自动提取文档中的内容摘要，通过对数据的浓缩可以从少量的数据中获取丰富的信息。

自从 1958 年Luhn提出自动文摘这一概念以来，国内外很多学者都在从事这方面的研究，并取得了相应成果。自动文摘的方法大体上可以分为 2 类：(1)基于统计的机械文摘[1]通过统计的方法，直接从原文中抽取句子作为文摘句。它利用各种统计信息如位置信息、频率统计等信息找出最能代表文章主题的句子作为文摘句。此类方法具有较高的实用性。(2)基于意义的理解文摘[2]需要对文章进行句法分析和语义分析，在理解的基础上产生文摘句。其优点是文摘质量高，缺点是领域严重受限、实用性较低。按文章结构划分，基于统计的机械文摘又可以分为无结构分析型和有结构分析型[3]自动文摘。前者通过对文章中的句子进行权重计算，找出权重较大的句子作为文摘句。该类方法容易忽略文章中分布较小的主题，而且冗余度大。后者首先对文章的语义结构进行分析，找出文章的主题结构，然后从各个主题中分别抽取句子构成文摘，有效地避免了无结构分析型方法的不足。

2 局部主题关键句抽取的算法

2.1 句子的表示

文档都由若干个段落组成，每个段落又可以细分为若干个句子。为了对句子进行自动处理，必须对句子进行形式化的表示。向量空间模型(Vector Space Model, VSM)是由 Salton 等在 20 世纪 60 年代提出的，它把文本表示成特征项和特征项权重组成的向量。对于汉语来说，词语是比较好的特征项。在完成对句子的分词处理后，以句子中的非功能性词汇为特征项对句子进行向量化。

在表达某一局部主题时，同一句子中可能多次用到相同的词汇，而该词语并不是全文主题的关键词。相反，某些出现频率不是特别高，但在全文中分布比较广的词语，才是表达全文词汇的关键词。此外，如果文章有标题，其标题一般都是对文档主题的归纳，出现在文档标题中的词语即使出现频率不高、分布不广，也应该是比较重要的词语。本文通过综合考虑词语在句子中出现的频率、词语在全文中的分布情况、词语在标题中出现过等多个因素来计算特征项权重。

本文中，句子中词汇权重的计算如下：

$$Weight(w) = \ln M \cdot \ln N \cdot Pos(w) \quad (1)$$

其中，$Weight(w)$表示词语 w 的权重；M 表示词语 w 在整篇文档中出现的次数；N 表示出现该词汇的段落数；$Pos(w)$是该词语的位置系数。在该方法中，把曾经出现在标题中的词汇的位置系数设为 1.5，曾经出现在每个段落的首句或者尾句的词汇的位置系数设为 1.3，其他词语的位置系数设为 1。

基金项目：国家自然科学基金资助项目(60773167, 60673040)

作者简介：徐 超(1981－)，男，助教、硕士，主研方向：自然语言理解，智能处理技术；王 萌，讲师、硕士；何婷婷，教授、博士；张 勇，助教、硕士

收稿日期：2007-12-22 **E-mail：**zhangpu@cqupt.edu.cn

2.2 局部主题的判定

一般来说，一篇文档包含一个大的主题，大的主题又由若干局部主题构成。在对文档进行自动文摘时，如果仅仅只考虑句子的重要程度，抽取出的文摘句会大量围绕文档所表达的分布较广的局部主题，会把一些分布不是很广的局部主题信息遗漏。因此，在选择文摘句时，除了考虑句子的重要性之外，还要综合考虑句子所在的局部主题。文档的摘要应该能较全面地概括各个局部主题的信息。

因此，本方法在从文档中选择文摘句时，首先根据局部主题的不同，对文档进行了主题划分。在选择主题划分的方法时，采用了层次分割法[4]。

在同一篇文档中，为了阐述某个局部主题，作者一般都会用连续的段落来表达该局部主题。在对主题进行划分时，假设属于某个局部主题的段落通常都是连续的。利用层次分割法对文档的局部主题进行切分时，本文把属于同一局部主题的所有段落设为一个局部主题单元。首先，将每个段落看作是一个局部主题单元，然后，反复地把相邻且最相似的 2 个局部主题单元合并为一个局部主题单元，直到局部主题单元的个数满足要求。

在对局部主题单元进行合并时，关键问题是局部主题单元之间的相似度计算。一个局部主题是通过属于该局部主题的段落来表达的。计算 2 个局部主题单元之间的相似度可以通过计算这 2 个局部主题单元中段落之间的相似度来实现。

在利用向量空间模型计算段落之间的相似度时，可以通过将一个段落形式化为一个向量的方式来计算 2 个段落的相似度。但是，仅仅将一个段落形式化为一个向量的话，该段落中所有的噪声都会出现在这个向量中。利用这种方式得到的相似度会给计算结果带来一定的偏差。本方法对 2 个句子之间的相似度作了适度的扩展来计算 2 个段落的相似度。段落是由若干个句子组成的，可以把一个段落看作是由若干个句子组成的句子集合。2 个段落的相似度可以利用这 2 个组成段落的句子集合相似度来计算。

在利用向量空间模型对句子进行形式化后，对句子之间的相似度的计算可以通过式(2)来计算：

$$Sim(S_1,S_2)=Sim(\boldsymbol{V}_1,\boldsymbol{V}_2)=\frac{\boldsymbol{V}_1\cdot\boldsymbol{V}_2}{\sqrt{\boldsymbol{V}_1\cdot\boldsymbol{V}_1}\cdot\sqrt{\boldsymbol{V}_2\cdot\boldsymbol{V}_2}} \tag{2}$$

其中，$Sim(S_1, S_2)$表示句子S_1和S_2的相似度，$\boldsymbol{V}_1$和$\boldsymbol{V}_2$ 表示S_1和S_2形式化所得到的向量。

这里把一个句子和一个段落的相似度定义为该句子与段落中所有句子中相似度最大的值，即$Sim(S, P)=\mathrm{Max}\{Sim(S, S_i)\}$，其中，$Sim(S,P)$表示句子和段落的相似度，$Sim(S, S_i)$表示句子$S$与段落$P$中第$i$个句子的相似度。

在计算 2 个段落之间的相似度时，首先从第 1 个段落中选出一个句子，计算该句子与另一个段落的相似度。这样，可以得到第 1 个段落中的句子与第 2 个段落的相似度的集合。利用同样的方法可以计算出第 2 个段落中的句子与第 1 个段落之间的相似度。把这些句子与段落的相似度的算法平均值作为这 2 个段落之间的相似度，即利用式(3)计算出 2 个段落之间的相似度：

$$Sim(P_1,P_2)=\frac{\sum_{i=1}^{M}Sim(S_i,P_2)+\sum_{i=1}^{N}Sim(S_i,P_1)}{M+N} \tag{3}$$

其中，$Sim(P_1, P_2)$表示 2 个段落P_1和P_2之间的相似度；$Sim(S_i, P_2)$表示第 1 个段落中的第i个句子与第 2 个段落的相似度；$Sim(S_i, P_1)$表示第 2 个段落中的第i个句子与第 1 个段落的相似度；M和N 分别表示第 1 个段落和第 2 个段落中的句子数。

基于同样的思想，可以计算 2 个主题单元之间的相似度。首先，通过式(4)计算段落和局部主题单元之间的相似度。在此基础上，通过式(5)计算 2 个主题单元之间的相似度。

$$Sim(P,T)=\mathrm{Max}\{Sim(P,P_i)\} \tag{4}$$

其中，$Sim(P,T)$表示段落P与局部主题单元T之间的相似度；P_i表示局部主题单元T中的第i个句子。

$$Sim(T_1,T_2)=\frac{\sum_{i=1}^{M}Sim(P_i,T_2)+\sum_{i=1}^{N}Sim(P_i,T_1)}{M+N} \tag{5}$$

其中，$Sim(T_1, T_2)$表示 2 个局部主题单元T_1和T_2之间的相似度；$Sim(P_i, T_2)$表示第 1 个局部主题单元中的第i个段落与第 2 个局部主题单元T_2的相似度；$Sim(P_i, T_1)$表示第 2 个局部主题单元中的第i个段落与第 1 个局部主题单元T_1的相似度；M和N 分别表示第 1 个局部主题单元和第 2 个局部主题单元中的段落数。

2.3 局部主题数目的计算

在利用层次分割法对文章进行主题划分时，需要确定局部主题的个数。一个好的划分应该满足：同一个局部主题单元内的段落之间具有较高的相似度，不同局部主题单元段落之间具有较低的相似度。因此，通过定义目标函数 $Object(K)$来衡量文档被划分为 K 个局部主题时的合理性。目标函数$Object(K)$的计算公式如下：

$$Object(K)=\begin{cases}\dfrac{Intern\_sim(K)}{Outer\_sim(K)} & K>1\\ 1 & K=1\end{cases} \tag{6}$$

其中，$Object(K)$表示局部主题被划分为 k 个时的合理程度；$Intern_sim(K)$表示当局部主题为 k 个时，局部主题单元内部的相似度；$Outer_sim(K)$表示局部主题为 k 个时局部主题单元之间的相似度。

在本文中，局部主题单元内部的相似度被定义为每个段落与该局部主题单元所有段落间的相似度之和，局部主题单元之间的相似度定义为各个局部主题单元与其他局部主题单元之间相似度的平均值之和，计算方法如式(7)和式(8)所示：

$$Intern\_sim(K)=\prod_{i=1}^{K}\sum_{j=1}^{Num(i)}\sum_{l=1}^{Num(i)}Sim(T_{i,j},T_{i,l})) \tag{7}$$

$$Outer\_sim(K)=\prod_{i=1}^{K}\frac{\sum_{j=1,j\neq i}^{K}Sim(T_i,T_j)}{K-1} \tag{8}$$

其中，$Intern_sim(K)$表示当划分为K个主题时局部主题单元内部的相似度；$Outer_sim(K)$表示当划分为K个主题时局部主题单元之间的相似度；$T_{i,j}$表示第i个局部主题单元中的第j个段落；T_i表示第i个局部主题单元；$Num(i)$表示第i个局部主题单元中的段落数。

如果一篇文章的段落数为 N，则局部主题的数目可能取 1~N 之间的值。比较局部主题个数取不同值时目标函数的值，使得 $Object(K)$最大的 K 即为文章的局部主题数。

2.4 局部主题关键句的选择

在对文档进行局部主题划分后，从每个局部主题单元中选择一定数量的句子作为文档的文摘句。文摘句的句子数由文档中总的句子数和文摘的抽取率得到。为了解决从每个局部主题单元中抽取多少个句子作为文摘句的问题，定义了局部主题单元的重要度，即该局部主题单元与整篇文档之间的相似度，在这里整篇文档被看作是一个大的主题单元。从某

个局部主题单元中选择的文摘句数量取决于它的重要度，通过式(9)计算从一个局部主题单元中抽取的句子数：

$$Number(T_i)=\frac{Sim(T_i,D)}{\sum_{j=1}^{K}Sim(T_j,D)} \tag{9}$$

其中，$Number(Ti)$表示从第 i 个局部主题单元中抽取的作为文摘句的句子数；D 表示将整篇文档作为一个大的局部主题单元。

从局部主题单元中抽取出来的文摘句应该能概括该局部主题单元的主要内容。由于文摘句的存在，因此一个局部主题内各个句子会变得更难区分，即该局部主题单元内部会变得更加模糊。直观上，如果 2 个句子完全一样，这时句子之间的相似度为 1，不能区分。反之，如果它们完全不一样，当句子相似度为 0 时，就能清楚地区分。因此，定义了一个句子对某个局部主题单元内部模糊程度的贡献，计算公式如下：

$$Ind(S_i,T)=\frac{\sum_{j=1}^{N}\left|Sim(S_i,S_j)-1\right|}{N} \tag{10}$$

其中，$Ind(S_i, T)$表示局部主题单元中的第i个句子S_i对该局部主题单元内部模糊程度的贡献。

在计算了一个局部主题内所有句子对该局部主题单元内部模糊程度的贡献后，根据它们的贡献大小进行排序。如果 2 个句子对该局部主题单元内部模糊程度的贡献相同，则比较该句子与整个文档的相似度，把与整个文档相似度较大者排在前面。

最后，选择排名靠前的 N 个句子作为该局部主题单元的文摘句。

3 实验结果与评估

对自动文摘的评估[5]一般有 2 种方式：外部评测和内部评测。外部评测方式通过比较自动文摘的结果对信息检索等其他工作的影响来判断文摘的质量。内部评测是在提供参考摘要的前提下，以参考摘要为基准评价系统摘要的质量。

这里采用内部评测的方式对所设计的方法进行验证。通过比较所抽取的文摘句与人工抽取的文摘句，利用召回率、准确率以及对召回率和准确率进行综合考虑的 *F-Score* 3 个指标对文摘质量进行评估。召回率、准确率和 *F-Score* 的计算如式(11)~式(13)所示：

$$\text{准确率}=\frac{S\cap R}{S} \tag{11}$$

$$\text{召回率}=\frac{S\cap R}{R} \tag{12}$$

$$F-Score=\frac{2\times\text{准确率}\times\text{召回率}}{\text{准确率}+\text{召回率}} \tag{13}$$

其中，S 表示通过本方法得到的文摘句；R 表示人工得到的文摘句。

通过如下的方法获取实验数据：首先，在网络上随机下载 100 篇不同类型的文章，并请中文系的学生通过人工的方式标明不同抽取率情况下的摘要。将这样的语料作为评测语料，然后将该方法生成的摘要与评测语料进行比较。实验得到的数据如表 1 所示。

表 1 实验结果 (%)

类别	抽取率	准确率	召回率	*F-Score*
文学类	10	41	42	41
	15	43	46	44
	20	46	47	46
	25	51	49	50
	30	54	53	53
新闻报道类	10	66	62	64
	15	68	64	66
	20	70	67	68
	25	73	69	71
	30	75	72	74
其他类	10	58	54	56
	15	60	56	58
	20	63	61	62
	25	65	63	64
	30	69	65	67

从实验数据中可以发现，本方法得到的文摘效果随文章题材的不同差异很大。对于文学类等语义结构并不十分明显的文章，文摘的效果较差。对于新闻报道等语义结构相对明显的文章，实验值则高于平均值。这与文章体裁不同对局部主题的划分产生的影响有直接的关系。同时，本方法所抽取得到的文摘句的分布几乎都包括文章的全部主题，文摘的冗余度较小。这与本方法先对文章进行主题划分，然后从各个局部主题单元中抽取文摘句有很大的关系。

4 结束语

本文提出了一种对文章进行主题划分后从局部主题单元中抽取关键句作为文摘句的自动文摘方法。该方法抽取的文摘比较全面地包含了文章的各个主题，而且冗余度小。通过实验进一步验证了该方法的有效性。

文档主题的划分对自动文摘的结果有着明显的影响，同时相似度的计算也影响文摘的结果。今后的工作中将从这 2 个方面入手改进划分方法，并将所采用的思想进一步扩展，开展对多文档自动文摘方法的研究。

参考文献

[1] Gong Y, Liu X. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis[C]//Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New Orleans, Louisiana, USA: [s. n.], 2001.

[2] 杨晓兰, 钟义信. 基于文本理解的自动文摘系统研究与实现[J]. 电子学报, 1998, 26(7): 155-158.

[3] 胡 珀, 何婷婷, 姬东鸿. 基于主题区域发现的中文自动文摘研究[J]. 计算机科学, 2005, 32(1): 177-181.

[4] Yaari, Yaacov. Segmentation of Expository Texts by Hierarchical Agglomerative Clustering[C]//Proceedings of the RANLP'97. Tzigov Chark, Bulgaria: [s. n.], 1997.

[5] Mani I. Summarization Evaluation: An Over Overview[C]// Proceedings of the NTCIR Workshop Evaluation of Chinese and Japanese Text Retrieval and Text Summarization. Tokyo, Japan: National Institute of Informatics, 2001.