面向冗余度控制的中文多文档自动文摘*

王红玲，黄超超，张明慧，周国栋
苏州大学 计算机科学与技术学院，江苏 苏州 215002
江苏省计算机信息处理技术重点实验室，江苏 苏州 215002
E-mail:hlwang@suda.edu.cn

摘 要：多文档自动文摘能够帮助人们自动、快速地获取信息，是目前的一个研究热点。相比于单文档自动文摘，多文档自动文摘需要更多考虑文档之间的相关性，以及文档信息之间的冗余性。因此如何控制信息冗余是多文档自动文摘的一个关键所在。本文在考虑文摘特性的基础上提出了一个冗余度控制模型，该模型通过计算文本单元在主题概率分布之间的相似度来决定句子的选择，从而达到控制冗余的目的。实验结果表明，该方法能够有效降低冗余度，且总体性能优于现有的自动文摘系统。

关键词：冗余度控制；多文档自动文摘；中文自动文摘

Chinese Multi-document Summarization Based on Redundancy Control

Wang Hongling, Huang Chaochao, Zhang Minghui, Zhou Guodong
School of Computer Science & Technology, Soochow University, Suzhou 215002
Jiangsu Provincial Key Laboratory of Computer Information Processing Technology, Suzhou 215002
E-mail: hlwang@suda.edu.cn

Abstract: Multi-document summarization can help people to access information automatically and fast. Compared to single-document summarization, multi-document is more to consider the correlation and redundancy between documents. Therefore, how to control information redundancy is a key problem to multi-document summarization. This paper proposes a model of redundancy control based on considering the features of summary. In this model, various similarities among the text units over topic's probability distribution are used to determine the choice of a sentence. Experimental results show that this method can effectively reduce redundancy, and overall better performance than existing automatic summarization system.

Keywords: reduandancy control; multi-document summarization; Chinese automatic summarization

1 引言

多文档自动文摘是指从一组文档集合中提取出重要信息组成代表该文档集合的摘要，该文档摘要可以帮助人们快速、高效地获取信息。通常多文档自动文摘可分成三步：文本分析，文本内容选择和文摘生成。和单文档自动文摘相比，多文档自动文摘需要考虑文档之间的相关性，以及文档信息之间的冗余性。因此如何控制信息冗余和如何选择文摘句来代表文档内容是多文档自动文摘的关键所在。本文在充分考虑文摘特性的基础上，提出了一个冗余度控制模型，该模型主要在文摘生成阶段通过综合考虑文本单元之间的相似度来选择句子作为文摘。在相似度计算上，本文通过计算文本单元的主题概率分布之间的相似性来获得。

文章第 2 部分简述了中文多文档自动文摘的相关研究。第 3 部分介绍了基于冗余度控制模型。第 4 部分则对总体介绍了面向冗余控制的中文多文档自动文摘系统。第 5 部分对实验结果进行了分析和比较。最后第 6 部分对本文进行了总结，并对后期工作进行了展望。

* 基金资助：国家自然科学基金(60873150)，江苏省高校自然科学基金(10KJB520016)

2 相关工作

中文多文档自动文摘相比于英文而言起步较晚，从技术上看，采用的主要技术手段大致相同。同时在这些技术使用过程中，需要利用的一些中文的资源和测试平台还不够成熟，例如中文多文档文摘缺乏统一的标注语料和评测方法，一些中文信息处理技术还不够成熟，在某种程度上制约了中文多文档自动文摘的发展。近阶段的相关研究包括基于句子抽取的策略（刘德喜等 2006），基于规则和统计的策略（傅间莲等 2006），基于图的策略（马慧芳等 2007，宋锐等 2009）和基于篇章的文摘策略（徐永东等 2007）等。

其中宋锐等（2009）通过抽取中文多文档集合中的主-述-宾三元结构构建文档语义图，再对语义图中的节点利用编辑距离进行语义聚类，并应用排序算法进行权重计算，选取包含权重较高的节点和链接关系的三元组生成多文档摘要。徐永东等（2007）受到 Radev(2000)交叉文本结构理论 CST 的启发，提出了一个用于多文本结构分析式文摘的多文本结构 MDF ,并在该结构的基础上进行候选文摘句的抽取、文摘句排序及文摘生成等一系列工作。

常用的冗余识别方法通常有两种：聚类法和排序法。聚类法通过测量所有句子对之间的相似性，用聚类的方法识别公共信息的主题，并从每个类别中抽取中心句子作为文档摘要。排序法相比于聚类法更加常用，其基本方法是根据某种打分规则，对文档中的所有句子打分并排序，选择高分值的句子作为文档摘要，典型的工作如最大边缘相关法 MMR(Goldstein 等 1999)和文档间信息包含法 CSIS。在 MMR 方法中，系统首先测量候选文摘与已选文摘之间的相似度，仅当候选段含有足够的新信息时才将其入选，其主要根据句子在文档中的相关性和已选中句子之间的冗余性的权值组合来选择合适的句子，相关性和冗余性都使用余弦相似度来计算。而 CSIS 方法（Radev 等 2004）则通过一个句子是否被包含在已在文摘中的另一个句子中来决定是否选择该句作为文摘句，该方法中的句子包含关系需要人工标注。Haghighi & Vanderwende(2009)则通过判断文档集合与候选文摘之间的相关度来判断冗余信息。

3 冗余度控制模型

图 1 给出了冗余度控制模型，该模型既可以面向通用型文摘(Generic Summarization)也可以面向基于查询的文摘(Query-based Summarization)。当面向基于查询的文摘时，需要考虑图中给出的用户查询部分与当前句子和扩充文摘之间的相似度，其中的用户查询可使用信息检索术中的查询扩展技术来扩充查询内容。本文只考虑通用型文摘，故忽略用户查询与句子和扩充文摘之间的相似度。

在此模型中，通过使用文本单元之间的相似性来反映文摘的各类特性，包括代表性、信息性和多样性等，其中候选文摘与文档集合之间的相似性可反映文摘的代表性，句子与文档集合之间的相似性可反映文摘的信息性，句子与文摘之间的相似性可反映文摘的多样性。

本文中该模型的评价函数定义为：

$$score(s_i) = \sum \lambda_i * f_i \tag{1}$$

其中的 f_i 为衡量各文本单元的相似度值，λ_i 为权值，即 f_1 为扩充文摘与文档集合之间的相似度；f_2 为当前句子 S 与文档集合之间的相似度；f_3 为当前句子 S 与当前文摘之间的相似度。句子通过该评价函数计算句子的得分，得出的分值越高说明与原多文档集合相似度越高、与当前摘要相似度越小，可以有效地减少冗余。

传统的根据特征的句子打分方法实际上只考虑了 f_2 值，即只反映信息性，在此模型中只需要设置 $\lambda_1 = 0$，$\lambda_3 = 0$；而在Haghighi & Vanderwende (2009)则使用了基于代表性的动态模型，即 $\lambda_2 = 0$，

$\lambda_3=0$；当$\lambda_1=0$，$\lambda_2=0$时，本模型考虑多样性。而MMR模型则同时考虑了信息性和多样性，此时$\lambda_1=0$。因此，本模型综合考虑了文档文摘所应具有的三种特性。

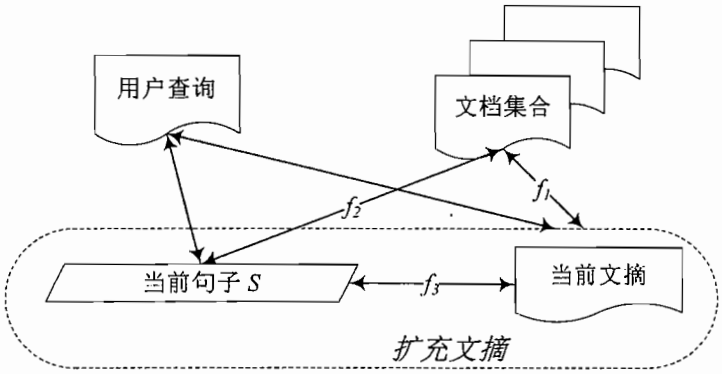

图1 冗余度控制模型

4 面向冗余度控制的多文档自动文摘

4.1 相似度计算

在图 1 所示的冗余度控制模型中，我们使用各文本单元之间的相似度来评价句子的得分，并由此来判断该句子的取舍，因此文本单元相似度的计算是该模型的一个重要组成部分。在此文本单元包括句子、文摘、文档和文档集合。本文通过计算各文本单元在文档主题上的概率分布之间的相似性来计算他们之间的相似性。

给定任意两个文本单元，a 和 b，其相似度值为：

$$TSim(a,b)=-(D_{KL}(P_a\|P_b)+D_{KL}(P_b\|P_a)) \tag{2}$$

其中 Pa 和 Pb 是文本单元 a 和 b 在主题上的概率分布，$D_{KL}(P_a\|P_b)$ 是两个概率分布 P_a 和 P_b 之间的 KL 散度，即

$$D_{KL}(P_a\|P_b)=\sum_i P_a(i)\log\frac{P_a(i)}{P_b(i)} \tag{3}$$

由于 KL 散度具有不对称性，我们同时包含 $D_{KL}(P_a\|P_b)$ 和 $D_{KL}(P_b\|P_a)$ 来保证相似度的对称性。

对于给定文本单元 a，其文档主题的概率分布 P_a 可以使用主题模型 LDA 的输出：文档 d 在主题 z 上的分布 $p(z|d)$ 和主题 z 在词汇 w 上的分布 $p(w|z)$ 来计算得到，具体的计算方法参见 Wang & Zhou (2010)。

4.2 文摘生成

传统的文摘生成方法是根据句子的分值，从高到底抽取句子组成文摘，这是一种静态的文摘生成方式。而采用冗余度控制模型后，需要根据得分动态计算当前句子与其他文本单元之间的相似度，逐渐扩充摘要。换句话说，判断一句话是否要作为文摘句加入到当前文摘中，不仅要计算句子与当前文摘的相似度，还要计算扩充文摘与给定多文档之间的相似度，这个过程是一个动态过程。

因此，使用冗余度模型产生文摘的具体过程如下：

1）运行 LDA 模型，得到 $p(z|d)$ 和 $p(w|z)$，计算句子得分并排序；

2）挑选得分最高的句子作为当前文摘；

3）对集合中的每个句子，将该句子与当前文摘组合形成扩充文摘，按照相似度计算方法，计算各文本单元之间的相似度；

4）选中使评价函数得分最高的句子加入到当前文摘中，形成新的当前文摘；

5）重复第 3 和第 4 步，直到文摘达到指定长度。

5 实验结果及分析

5.1 实验设置

由于目前中文自动文摘没有一个公认的标注语料，为了便于性能的比较，我们选用徐永东等（2007）描述的多文档数据作为实验语料。这些数据来自于网络上的新闻报道，覆盖的主题有运动、经济、事故等等，整个数据被分成 19 个文档集合，每个文档集合含有 5~10 篇文档，并且同样的文档集合有同一个中心的主题。

本文评价方法采用对每个主题采用模糊标注的方法，标注过程中，除了在源文档集合中标注出标准文摘句，还标注出在源文档中可替换标准文摘句、且不能与标准文摘句在文摘中同现的句子，我们称之为候选文摘句。每个候选文摘句根据可替换程度赋予一个取值在(0，1]之间的权值。这样得到的评测语料库就可以采用准确率、冗余度和总体质量三项指标来评估文摘系统质量，以解决传统多文档自动文摘评测出现的无法顾全文本集合中存在多个可替换文摘句的问题。在此基础上，采用准确率、冗余度和综合质量等几方面指标来评估待测系统：

$$precision = (\sum_{i=1}^{k_1} \omega_i) / K$$

$$redundancy = (\sum_{i=1}^{k_1} \sum_{j=i+1}^{k_1} \phi(s_i, s_j)) / K$$

$$total = precision - redundancy$$

其中，K 是待评测文摘的句子总数。k1 是标准文摘的句子在待评测文摘中出现的句子总数，$(\omega_1, \omega_2, ..., \omega_k)$ 是每个句子的权值，该权值由上述手工标注方法得到；$\phi(s_i, s_j)$ 是一个二元判别函数，当 si，sj 为同类文摘句时，$\phi(s_i, s_j)$ =1；否则为 0。

5.2 实验结果及分析

在预处理阶段，本文使用了 ICTCLAS 2009 系统[1]进行中文分词处理，然后根据停用词表去除停用词，另外根据文档特征去掉了对文摘作用不大的介词、虚词、数词等词语，提高系统准确率。

- 冗余度控制模型实验

为评价冗余度模型的性能，我们进行了对比实验，来验证冗余度控制模型的有效性。表 1 给出了我们系统分别在 5 句、10 句、20 句文摘情况下的系统性能。表中的静态方法表示只根据句子得分来抽取句子形成最终文摘，不使用冗余度控制模型的方法。该方法中句子得分使用了句子与文档集合在主题分布上的相似度值。而动态方法是指使用冗余度控制模型抽取文摘的方法。从表 1 的结果可看出，使用冗余度控制模型后，系统的准确率和冗余度总是优于静态方法（不使用冗余度控制模型），特别是冗余度有明显的降低，这说明了冗余度控制模型的有效性。

表 2 给出了徐永东等（2007）一文中给出的在相同评价体系下、同一语料库上的上限系统性能和其使用的 MDF 框架的性能，其中上限系统中的文摘是指根据人工标注信息抽取的摘要，而 MDF 中的所有信息是自动生成的。比较表 1 和表 2 的结果，可看出除 5 句文摘的冗余度值，动态方法的性能在系统的准确率和冗余度方法都明显好于 MDF 的性能，；但相比于上限系统，我们系

[1] http://ictclas.org

统的准确率还有很大的差距，不过在冗余度方面，两者的性能已经比较接近，这进一步说明了冗余度控制模型的有效性。需要说明的是，从理论上讲 5 句文摘的冗余度不可能为 0。

表 1　系统性能

文摘长度(句子个数)	静态方法			动态方法		
	P(%)	R	T(%)	P(%)	R	T(%)
5	68.73	4.38	64.35	76.84	1.25	75.59
10	71.89	10.46	61.43	74.47	5.94	68.53
20	76.16	9.75	66.41	78.90	6.67	72.23

表 2　上限系统性能和 MDF 的性能

文摘长度(句子个数)	上限系统			MDF		
	P(%)	R	T(%)	P(%)	R	T(%)
5	88.125	0	88.125	70.31	0	70.31
10	90.625	4.68	85.945	68.13	9.37	58.76
20	86.17	5.94	80.23	72.66	7.81	64.85

- 主题数目实验

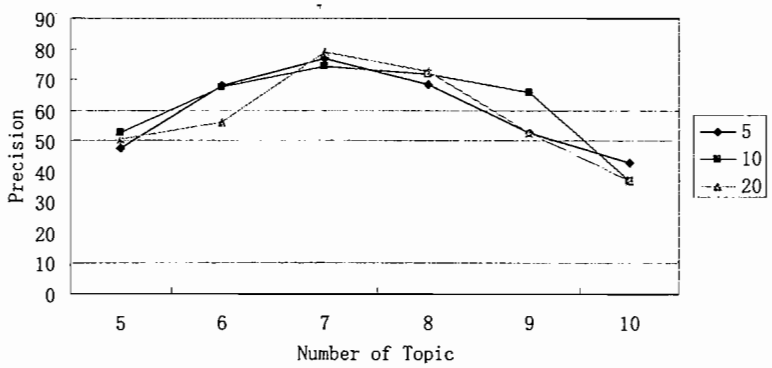

图 2　主题数目对准确率的影响

由于 LDA 训练时的主题数目会影响系统性能，我们对不同的主题数目进行实验。图 2 展现了中文语料在不同主题数目下文摘的准确率，图中表明当主题数目 k 设为 7 的时候，系统能获得最好的性能。这与我们的最初判断是一致的，即，尽管每个文档集合都有一个中心主题，但其中的每个文档都有自己的主题，也就是每个文档至少有一个主题。基于此观察，我们发现每个文档集合平均有 7 个文档，所以，当 $k=7$ 时，我们得到最好的结果，随着主题个数的增长，数据稀疏性增大，性能降低。图中不同的线型代表不同文摘长度的准确率，这同样表明不同长度的摘要有相似的准确度曲线。本文其余实验中，主题数目均设置为 K=7，迭代次数为 2000 次。

6　总结

本文针对中文多文档自动文摘中的信息冗余问题，提出了一个冗余度控制模型，该模型从考虑文摘的特性出发，综合考虑各文本单元之间的相似度，包括句子与文档集合之间的相似度，句子与文摘之间的相似度和文档集合与文摘之间的相似度。本文使用各文本单元在文档主题概率分布上的 KL 散度值来表示相似度。实验结果表明，应用冗余度控制模型后能有效减少自动文摘的冗余度。

抽取关键句子及计算文本单元之间的相似度有较多的方法，因此在下一步工作中，我们将继续探索有效的句子打分方法和相似度计算方法，以进一步提高系统性能。

参 考 文 献

[1] 刘德喜, 何炎祥, 姬东鸿等. 2006. 一种基于演化算法进行句子抽取的多文档自动摘要系统 SBGA. 中文信息学报, 20(6): 14-20.

[2] 傅间莲, 陈群秀. 2006. 基于规则和统计的中文自动文摘系统. 中文信息学报, 20(6): 10-16.

[3] 马慧芳, 祁云平, 杨小东. 2007. 一种基于文本关系图的多文档自动摘要技术. 情报学报, 23(3): 67-69.

[4] 宋锐, 林鸿飞. 2009. 基于文档语义图的中文多文档摘要生成集中. 中文信息学报, 23(3): 110-115.

[5] 徐永东, 徐志明, 王晓龙. 2007. 基于信息融合的多文档自动文摘技术. 计算机学报. 30(11): 2048-2054.

[6] Hongling Wang, Guodong Zhou. Topic-driven Multi-document Summarization. IALP'2010.

[7] Radev, DR., H. Jing and M. Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. ANLP/ NAACL 2000: 21-29.

[8] Radev, D.; Jing, H.; Sty´s, M.; and Tam, D. 2004. Centroid-based summarization of multiple documents. Information Processing and Management 40: 919-938.

[9] Haghighi A. and Vanderwende L. 2009. Exploring Content Models for Multi-Document Summarization. NAACL'2009: 362-370.