第 30 卷 第 6 期 自 动 化 学 报 Vol.30, No.6
2004 年 11 月 ACTA AUTOMATICA SINICA Nov., 2004

基于局部主题判定与抽取的多文档文摘技术[1)]

秦 兵 刘 挺 李 生

(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

(E-mail: qinb@ir.hit.edu.cn)

摘 要 提出了一个通过对同一主题的多文档集合内局部主题的判定和抽取生成多文档文摘的方法. 首先在对多文档集合中句子依存分析和语义分析的基础上进行相似度计算, 将相似句子经过聚类形成多文档集合内不同的局部主题, 然后进行每个局部主题中质心句的抽取和排序, 生成多文档文摘. 该方法实现了文摘长度随文档内容自动确定, 从而保证了文摘中包含的信息的全面和简洁. 最后文中还给出了多文档文摘的评价方法和实验结果, 文摘的平均精确率和平均压缩率分别为 71.4% 和 25.2%.

关键词 多文档文摘, 局部主题, 聚类

中图分类号 TP391

Multi-document Summarization Based on Local Topics Identification and Extraction

QIN Bing LIU Ting LI Sheng

(*School of Computer Science and Tecchology, Harbin Institute of Technology, Harbin* 150001)

(E-mail: qinb@ir.hit.edu.cn)

Abstract This paper describes a multi-document summarization method based on local topics identification and extraction. The similarity of sentences is measured by analysis of dependency and semantics. Local topics are found by sentence clustering. The centroid sentence is extracted from each local topic and is ordered to generate summarization. The size of summarization is determined according to content of multiple documents, as a result, the summarization becomes general and concise. Finally, the evaluation and experiment are given, the average precision of summarization and the average ratio of compressibility are 71.4% and 25.2%, respectively.

Key words Multi-document summarization, local topic, clustering

1 引言

互联网的普及使人们的生活方式发生了巨大的变化, 在网络带给人们大量信息的同

1) 国家自然科学基金 (60203020) 及国家 "863" 高科技项目基金 (2001AA114041) 资助
Supported by National Natural Science Foundation of P.R. China (60203020) and National "863" High Technology Research and Development program of P.R. China (2001AA114041)
收稿日期 2003-10-09 收修改稿日期 2004-06-23
Received October 9, 2003; in revised form June 23, 2004

时，人们的需求也随着网络信息的急剧增长不断地发生着变化，从而促进了许多新技术诞生和发展．多文档文摘成为新的研究热点．同时，在新一代搜索引擎问答系统 (Q&A)、话题的监测与跟踪技术 (Topic Detection and Tacking, TDT)、国家安全部门的非法信息监测、特殊信息的定制与融合等方面发挥重要作用．

国际上一些比较权威的会议，例如 ACL 都有自动文摘专题，并且有关多文档文摘的文章也在逐年增多．值得关注的是，由 NIST 的系列会议之一 TIDES(DARPA's Translingual Information Detection, Extraction, and Summarization program) 赞助发起文本理解会议 (Document Understanding Conference, DUC), 使研究者共同参与到大规模文本测试中来，促进了自动文摘包括多文档文摘的发展．DUC 会议自 2001 年起每年举办一次，其中有一项任务就是对多文档文摘系统进行评测，表明了多文档文摘的研究正在向规范化、统一化迈进．还有一些国际上著名的会议，如 TREC 和 TDT 等一些评测子任务也都涉及到了多文档文摘技术．

以往对文摘的研究大多集中在单文档文摘方面，跨文档文摘的研究很少，网络的普及使跨文档信息融合正在成为新的研究热点．在该领域比较著名的方法是哥伦比亚大学 Goldsdein 提出的基于 MMR(Maximal Marginal Relevance) 的多文档自动文摘方法[1]，密歇根大学 Redev 提出基于质心的多文档自动文摘方法[2] 等．近来也有一些学者通过聚类来生成多文档文摘，通过聚类生成的文摘可以更好地提高文摘的覆盖率[3]．同时，一些多文档文摘系统被开发出来[4]．目前多文档文摘的研究方法尽管不同，但存在一些共同的问题，即文摘的长度是根据规定的字数或压缩比确定的．而一般情况下文档的内容都是未知的，如果按照统一的字数或压缩比生成文摘，会使文摘过于冗余或覆盖的信息不够全面．本文提出了一种新的多文档文摘方法．该方法在对多文档集合句子进行相似度计算基础上将相似的句子聚类形成局部主题，通过对局部主题中质心句的抽取和排序，生成多文档文摘．该方法有效地解决了文摘中信息不够全面或者信息过于冗余的问题．方法的框架如图 1 所示．

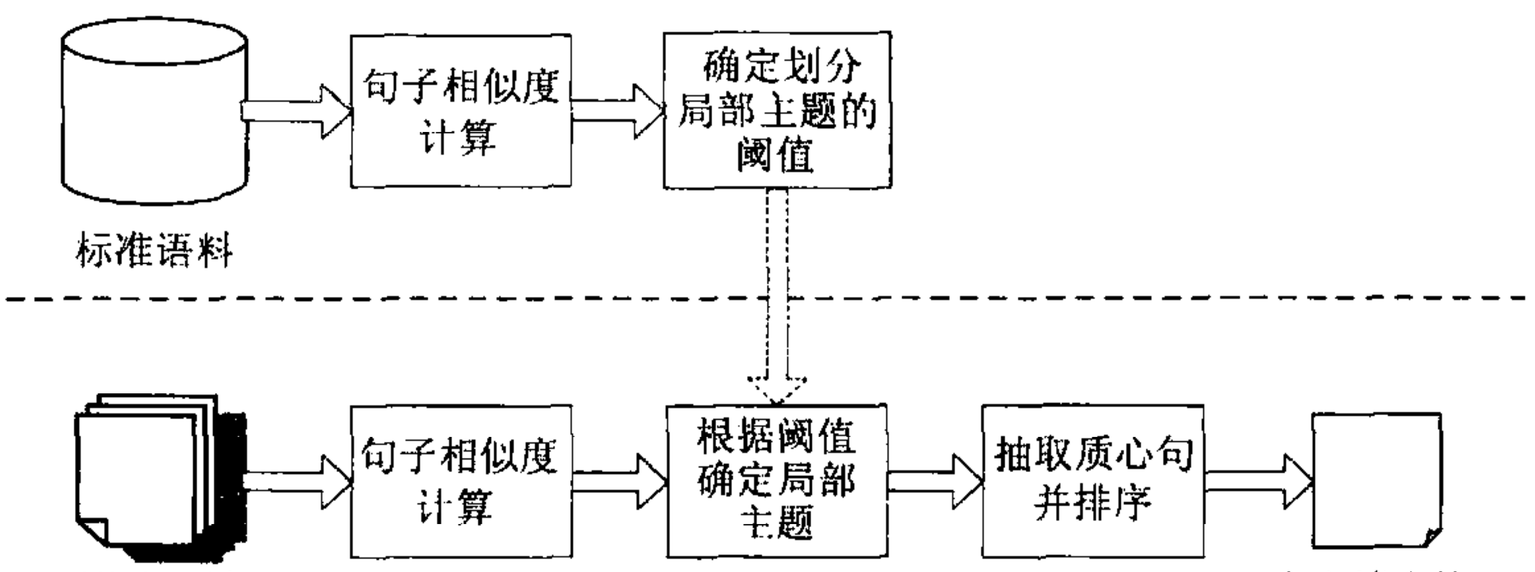

图 1 基于局部主题判定与抽取的多文档文摘系统

Fig. 1 The system of Multi-Document Summarization based on local topics identification and extraction

2 局部主题的概念及判定

2.1 局部主题的概念的提出及意义

对于任意的多文档集合可以从物理结构和逻辑结构进行划分．从物理结构来看，多

文档集合可以理解为文本单元组成的集合，即 $D = \{d_i | i = 1, 2, \ldots, n\}$，每一篇文档可以表示为文本单元的集合，文中提到的文本单元是句子，每一个文本可以表示为 $d_i = \{s_{i\,k} k = 1, 2, \ldots, m_i\}$，因此多文档集合可以看作为句子的集合，即 $D = \{s_{i,l} | i = 1, \ldots, n; l = 1, \ldots, m_i\}$，其中 i 表示集合中文本号，l 表示 $s_{i,1}$ 在所在文本中的位置. 从逻辑结构来看，一个主题是由对其不同侧面的描述组合而成的，每一个侧面信息称之为一个局部主题 (T_i). 因此，多文档集合又可以理解为由多个局部主题构成的，$D = \{T_i | i = 1, 2, \ldots\}$，每个 T_i 是一个句子集合. 实际上就是将多篇相同主题的文档界限打破，按照句子表达意思的相近程度进行重新组合.

2.2　局部主题的判定

局部主题的判定首先需要进行句子的聚类和相似度计算. 本文利用文献 [5] 的方法计算句子的相似度. 首先对句子进行依存分析得到依存树. 从计算效率和相似度准确程度两方面因素综合考虑，在利用句子的依存结构进行相似度计算时，只考虑那些有效搭配对之间的相似程度. 同时在依存树的节点上加入了语义信息，对有效搭配对的比较从词转化到语义级，从而提高匹配的精确度. 相似度的计算公式如下：

$$\mathrm{SIM}(\mathrm{sen1}, \mathrm{sen2}) = \lambda^* \mathrm{SIM1}(\mathrm{sen1}, \mathrm{sen2}) + (1 - \lambda)^* \mathrm{SIM2}(\mathrm{sen1}, \mathrm{sen2}) \tag{1}$$

式中第一项是基于依存分析的句子相似度，第二项是基于语义分析的句子相似度，经过合并成为两个句子最后的相似度，λ 为经验值，本文中取值为 0.5. 通过采用基于依存分析及语义分析的方法进行句子相似度计算，准确率为 81.4%[5].

为了能够判定任何一个文本集合所包含的局部主题，需要通过标准语料确定划分局部主题的参数及阈值. 借助于半偏相关系数 η 来观察标准语料中每个文档集合聚类过程，从而确定能够区分这些标准文档局部主题的阈值.

首先对每一个文档集合根据句子相似度，采用系统聚类法将最相似的两个子类合并成一类，最初是每个句子自成一类.

计算每次合并后的半偏相关系数 η，计算公式为

$$\eta = W_M - W_K - W_L \tag{2}$$

式中 $W_M - W_K - W_L$ 表示类 C_K 和类 C_L 合并为下一层次的类 C_M 时引起的类内离差平方和的增量，$W_I = \sum_{i \neq j\ x_i, x_j \in C_I} (1 - SIM(x_i, x_j))^2$ 表示类内离差平方和，x_i 和 x_j 为集合中任意两个文本单元.

按照上述方法，通过将每一组训练文档都做一次完整的聚类 (直到聚到标准的类数)，记录每次聚类的 η 值，η 值在聚类过程中的变化曲线见图 2 所示，图中纵轴是 η 的值，横轴数值代表合并次序，曲线上的点代表第 n 次合并的半偏相关系数的值.

观察所有训练语料在合并的类别与标准的类别数 (G) 相等时半偏相关系数以及类别数为 G-1 时半偏相关系数，取每一个文档集合这两个值的平均值作为划分该集合的局部主题的阈值，然后取训练语料所有集合的划分局部主题的阈值的平均值作为统一划分类别的阈值. 测试语料在进行聚类过程中，当 η 的值大于阈值时，聚类停止，此时每个类别就判定为一个局部主题.

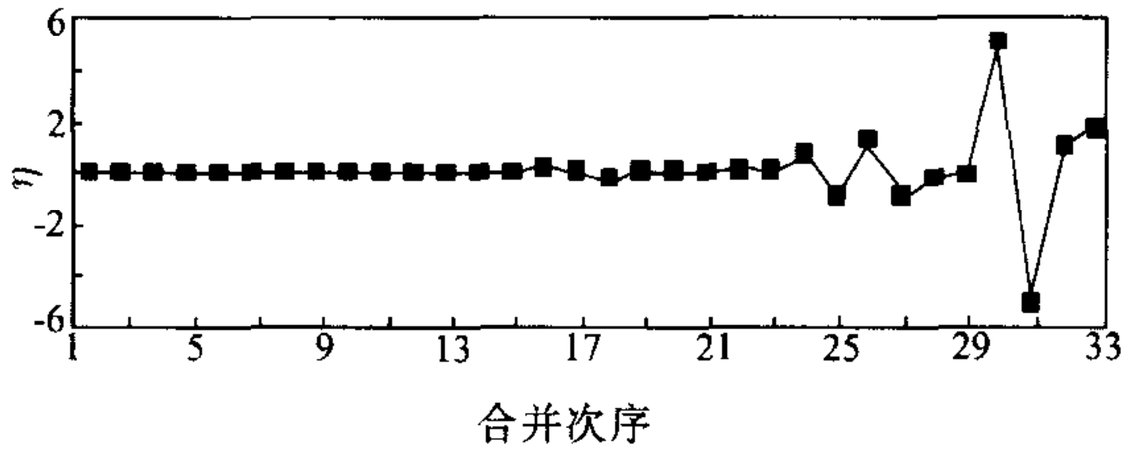

图 2 文档聚类过程中半偏相关系数 (η) 曲线

Fig. 2 The curve of half-deflection correlation coefficient (η) of clustering process

3 局部主题的抽取及文摘的生成

3.1 局部主题的抽取

聚类结束后，每类代表一个局部主题，实际上相当于获取了该主题文档集合各个侧面信息. 但每类内的句子之间表达的意思接近，为了去除这些冗余信息，需要在每类中找到最代表本类信息的句子来表示该类的信息. 每个局部主题中最代表本类信息的句子称为质心句，即

$$C_i = \arg\min_{s_i} \sum_{\substack{j=1 \\ i \neq j}} SIM(S_i, S_j) \tag{3}$$

根据式 (3) 对每个局部主题质心句抽取，得到多文档集合的质心句集合 $C = \{C_i | C_i \in T_i, i = 1, 2, \cdots k\}$, 这是组成多文档文摘主要内容.

3.2 文摘的生成

文摘的生成实际上是对质心句的排序. 首先在多文档集合中找到包含质心句最多的文本 d_c 作为参考框架，将质心句按该文本框架排序，最终生成句子顺序合理的文摘. 该文本 d_c 获得方法如下：

$$d_c = \arg\min_i \{|d_i \cap C|\} \tag{4}$$

通过上式，找到最符合条件的文本 d_c, 以 d_c 为框架，将质心句插入到该文本中进行排序，根据质心句 C_i 与文本 d_c 中的句子相似度，将与质心句 C_i 最相似的句子位置信息作为 C_i 在文摘中的参考位置信息.

4 实验

4.1 评价方法

在参考 DUC 评测标准的基础上结合多文档文摘的应用，我们提出了面向应用的多文档文摘相应的评价标准. 通过文摘的精确率和压缩率来对文摘的全面性和简洁性进行评价. 信息覆盖率表示文摘的包含原文档信息的多少，压缩率表示文摘对原文档去冗余的程度.

定义. $F(t_i)$: 第 i 个主题文摘的精确率; $C(t_i)$: 第 i 个主题文摘压缩率; $StaSen(t_i)$: 第 i 个主题标准文摘的句子集合; $GenSen(t_i)$: 系统生成第 i 个主题文摘的句子集合;

$TopicSen(t_i)$: 第 i 个主题所有文档的句子集合，

$$F(t_i) = \frac{|StaSen(t_i) \cap GenSen(t_i)|}{|GenSen(t_i)|} \times 100\% \tag{5}$$

注. 此句子匹配的含义是自动文摘的句子和标注文摘的句子相同或处于同一类别中.

$$C(t_i) = \frac{|GenSen(t_i)|}{|TopicSen(t_i)|} \times 100\%$$

4.2 实验结果

通过搜索引擎获得 20 个不同主题新闻语料的集合，经过整理每类集合包括来自同一事件的报道 (不包括重复的文档) 大约 6~7 篇，根据训练得到的阈值参数曲线，实验中选择 $\eta = 1.65$ 作为划分局部主题的标准. 按照前述方法得到局部主题，对各局部主题中代表句进行抽取和排序后，得到多文档文摘，实验结果如表 1 所示.

表 1 多文档文摘评价结果

Table 2 The result of evaluation of multi-document summarization

多文档集合编号	信息覆盖率	压缩率
1	71.2%	16.9%
2	68.9%	64.0%
3	80.7%	21.2%
4	72.6%	22.2%
5	81.4%	24.0%
6	68.2%	25 7%
7	57 2%	27.8%
8	68 3%	28 0%
9	74 9%	23 3%
10	72 8%	21 9%
11	45 2%	22.7%
12	76 8%	12.2%
13	63.3%	36.8%
14	50 0%	35.0%
15	87 6%	18.9%
16	74.7%	15.8%
17	72.4%	25.0%
18	83.3%	18.8%
19	77.8%	14.3%
20	80.1%	29.6%
平均	71.4%	25.2%

实验表明，由本文方法得到的多文档文摘信息覆盖率较高，压缩率反映出压缩幅度较大，去冗余信息的效果很好，文摘的质量是令人满意的，并且实现了文摘的长度随内容的变化而变化.

同时将本文方法与传统的抽取首句方法比较，即从上述多文档集合的每个原始文档抽出首句，组成一个多文档文摘，与基于局部主题的文摘相比较，通过比较包含有效词 (去除停用词) 个数和高频词个数 (在文档集合中出现两次以上词) 这两个指标，本文提出的方法比抽取首句的多文档文摘方法覆盖的有效词平均高出 14.73%, 高频词覆盖率平均高出 7.59%.

通过对生成的文摘分析，可以看到，由于在聚类过程中训练的阈值是平均值，对个别文档集合的类别划分不一定很准确，会影响文摘的效果．系统也存在其他一些文摘系统中的共性问题，对于个别描述多个地点和人物文档的文摘，由于指代不明，容易造成混乱，有时会影响文摘的可读性．下一步，我们将命名实体识别和指代消解技术应用在文摘中，并在文摘句的排序问题上作进一步研究，以便提高文摘的可读性．

5 结论

本文提出的基于局部主题识别与抽取的多文档文摘方法，克服了传统文摘中由于给定压缩比造成文摘有时冗余，有时由于字数限制表达不够全面的问题，实现了文摘长度随文档内容自动确定；同时通过聚类的方法实现局部主题信息的发现，并由局部主题的质心句生成文摘，对存在于少数文档中的个性信息很好地挖掘出来，使这些信息不会由于多文档中相同信息太多而淹没．文中对多文档文摘的评价提出的一些建议和方法，希望会对多文档文摘的研究和发展带来裨益．

References

1 J Goldstein, M Kantrowitz, V Mittal, J Carbonell. Summarizing text documents: Sentence selection and evaluation metrics. In: Proceedings of SIGIR-99, Berkeley, CA: 1999. 121~128

2 Dragomir R Radev, Hongyan Jing, Malgorzata Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In: ANLP/NAACL 2000 Workshop. Seattle, Washington, USA: 2000. 21~29

3 Pascale Fung, Grace Ngai. Combining Optimal Clustering and Hidden Markov Model for Extractive. In: Proceedings of the ACL 2003 workshop on multilingual summarization and question answering. Sapporo, Japan: 2003. 21~28

4 Chin-Yew Lin, Eduard Hovy. From Single to Multi-document Summarization: A Prototype System and its Evaluation. In: Proceeding of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL-02). Philadelphia: 2002. 457~462

5 Li bin, Liu Ting. Qin Bing Chinese Sentence Similarity Computation Based on Semantic Dependency Relationship Analysis. *Application Research of Computers*, 2003, **20**(12): 15~17 (in Chinese)

秦 兵 哈尔滨工业大学计算机学院副教授．主要从事中文信息处理，信息抽取，多文档文摘等研究工作．

(**QIN Bing** Associate professor of Department of Computer Science and Engineering, Harbin Institute of Technology. Her main research interests include Chinese information processing, information extraction and multi-document summarization.)

刘 挺 哈尔滨工业大学计算机学院教授．主要从事自然语言处理，信息检索等研究工作．

(**LIU Ting** Professor of school of Computer Science and Technology, Harbin Institute Technology. His main research interests include natural language processing and information retrieval.)

李 生 哈尔滨工业大学计算机学院教授，博士生导师．主要从事然语言处理，机器翻译等研究工作．

(**LI Sheng** Professor and Ph.D. director of school of Computer Science and Technology, Harbin Institute of Technology. His main research interests include natural language processing and machine translation.)