基于主题概念的多文档自动摘要研究[1)]

刘德荣[1,2] 王永成[1] 刘传汉[1]

(1. 上海交通大学计算机科学与工程系，上海 200030；2. 上海海事大学，上海 200135)

摘要 文章叙述了一种针对大规模文档集的综合性自动摘要的研究与实践。首先利用 HOWNET 来计算文献主题概念的内聚度，在此基础上，处理文档之间的相关度以及各自在整个文档集中的主题重要度等特征；其次阐述了基于文档综合主题辞和综合优先度的多文档自动摘要生成原理。实验结果表明，该系统经过对新闻多文档集进行综合性分析，生成的摘要能有效地反映重要的主题内容。

关键词 概念内聚度 主题重要度 综合优先度 多文档摘要

Study of Multiple Documents Summarization Based on Subject Concept Cohesion

Liu Derong[1,2], Wang Yongcheng[1] and Liu Chuanhan[1]

(1. *Department of Computer Science & Engineering, Shanghai Jiaotong University, Shanghai* 200030;
2. *Shanghai Maritime University, Shanghai* 200135)

Abstract In this paper we present a method to automatically generate an integrated summary from large document sets. Firstly we calculate the subject concept cohesion according HOWNET. Based on the above condition, some features are obtained, such as the similarity of documents, the subject importance in the set. Then we describe a CMDIS system designed specifically to summarize multi-documents by synthesis priority and subject string. The results show that this model can generate an effective satisfied summarization by focusing on document sets of news.

Keywords concept cohesion, subject importance, synthesis priority, multi-documents summarization.

1 前 言

自动摘要系统可以对一篇电子文档进行自动分析，判断文献的主题信息，从而获取基本反映主题内容的摘要信息。人们在实际的工作业务或学术研究中，往往需要了解和阅读一系列的相关文档材料并进行综合性处理，因此对大规模的相关文档集进行自动分析处理成为信息浓缩的一个有实用价值的研究内容[1]。Internet 正成为全球最大和传播范围最广的信息库，其中大量的 Web 页面文档缺乏组织条理性。例如，搜索引擎对某一个新闻事件的查找可能返回几十甚至上百篇文献，即使有结果文档的相关性排序，用户仍然需要阅读大量的文档，并查阅各文档的相关部分，从而取得目标信息。自动提炼、浓缩所有文献的相关主题信息，组成一个综合性的信息概要，可以使用户快速了解大文档集的主要内容，从而节省大量的阅读时间，提高阅读电子信息的效率。本文利用反映文档信息的主题概念的关联分析，判断多文档间的相关度以及全集中的主题重要度，利用文档综合优先度、综合主题辞并结合统计与语义处理的方法，实现多文档的自动摘要。

收稿日期：2004 年 4 月 28 日

作者简介：刘德荣，男，1972 年生，博士生，主要从事智能信息处理、自动摘要等研究。

1)基金项目：国家 863 计划项目(编号 2002AA119050)。

2 主题概念内聚度(Concept Cohesion)

人们在查阅大量文献时,往往不是通读所有原文,而是通过摘要、关键词及其他索引信息来了解文章的大意。我们可以首先对文献进行自动标引处理得到一组主题辞(词、词组或短语),并用它们来标识文献的信息内容。对文献的自动分析比较,可以用主题辞的简化对比来分析文章的大致主题内容。我们发现,人们在识别文献的主题时,头脑中获得的是文献的主题概念,而不仅仅是主题辞。例如,人们在分析文献间的"试卷"和"考试"主题辞时,通常是把它们作为相关的概念进行语义比较处理。

我们用概念内聚度来表示概念的相互语义关联度。由于自然语言的灵活性与复杂性,准确度量概念间的相关度是比较困难的。HOWNET是一个描述有关概念及其属性之间的关系的知识库[2],所以我们将借助它来粗略计算概念的内聚度。

HOWNET的概念词典中,概念的内涵由相关的义原来表达,概念与概念的关系主要体现在每个记录的概念定义项(DEF项)中。为了方便计算处理,每一个概念我们可以用一组组成定义项的义原集合来表示:$Concept_1 = Def(Atom_1, Atom_2, \cdots, Atom_k)$。

则两个概念p, q的内聚度可以用义原的关系组合来表达Concept_Cohesion(p, q)。由于计算过程中义原对是非固定类型搭配的情况,我们采用动态优化的方法处理。

Function Concept_Cohesion(p, q);
Begin
Setp = Def(A_1, A_2, $\cdots$, A_m) = { $A_1 \cdots, A_m$ };
q = Def(B_1, B_2, $\cdots$, B_n) = { $B_1 \cdots, B_n$ };

$$\{(A_i, B_j)_k\} = \arg \max_{\substack{A_i \in p-\{\cdots,(A)_{k-1}\} \\ B_j \in q-\{\cdots,(B)_{k-1}\}}} (Atom\_Cohesion(A_i, B_j))$$

// 判断两概念的最佳比较匹配义原对;

$$\text{return} \quad \frac{\sum_{k=1}^{\min(m,n)} \beta_k * Atom\_Cohesion(A_i, B_j)_k}{\max(m,n)}$$

// 返回概念内聚度结果; (1)
End

HOWNET对义原的关系与组织,主要由以下几个侧面来描述义原之间的逻辑与语义关系,分别为Event, Entity, Quantity, Attribute, Secondary Feature, Syntax, Event Role & Feature, Antonym, Converse,其中每个侧面表示为树型关系。我们可以利用义原在树型中节点属性位置的路径关系来表达它们的内聚度[3]。设有两个义原A、B,则义原的内聚度表示为:

$$Atom\_Conhesion(A, B) = \sum_{i=0}^{i=9} p_i * hnc(A, B)$$

$$\sum_{i=0}^{i=9} p_i = 1 \qquad (2)$$

其中,$hnc(mc_1, mc_2)$为每一个侧面分量关系的义原内聚度,p_i为每一个侧面分量的权重系数。

义原可能是树的节点或节点的属性,则每一个分量的内聚度为

hnc(A, B) = Node_Cohesion(A, B);
Function Node_Cohesion(A, B)
Begin
N_1 = node(A) //查找A所在节点,如果是属性取所属的节点
N_2 = node(B)
father = near_father(N_1, N_2) //查找共同N_1、N_2最近的父节点
d = distance(father, root) //查找父节点father与节点root之间的距离
h_1 = distance(N_1, root)
h_2 = distance(N_2, root)
h = max(h_1, h_2)
设i为深度层次,令$P_i = (1+i)^2$ $(0 <= i <= h)$
return (P_d/ P_h)
End

3 文档特征数据挖掘处理

对于通过搜索引擎得到的一组文献或很多网站的新闻专题这样的一个文档集,它们可能是某一新闻事件的大致相同的报道或者是从不同的侧面进行叙述。为了反映每一个文档描述的主题信息和它在文档集中的重要度,我们需要对文档的特征数据进行提取分析处理。首先是对文档的格式进行处理(如HTML、DOC等转化为标准文本格式),然后进行主题的识别与抽取,由每一组主题辞构成文档特征矢量,根据主题概念的关系算法计算文档特征矢量之间的距离,评估相互之间的关联度。最后计算每

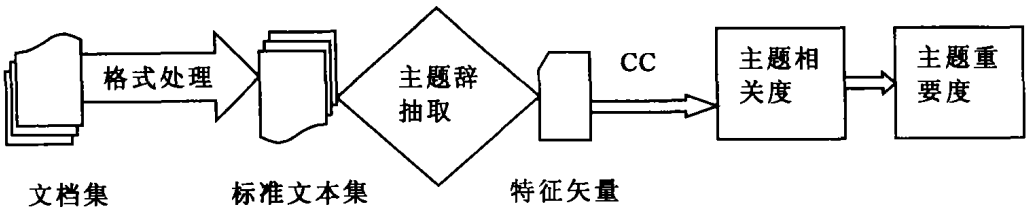

图 1　文档特征数据挖掘过程

一个文档在数据集中的重要度及系统综合主题辞。如图1所示。

3.1　文档主题辞表示

由于中文词的表示没有具体的标志，我们针对文档的 TFIDF 特征，考虑系统处理的效率与准确性，而不采用分词处理的方法。我们主要是利用字同现原理和马尔科夫统计模型进行高频字串统计，同时考虑标题位置及指示信息，得到一组主题辞（反映文献主题的词、词组或短语），用他们来表示文档的主题特征[4]。为了提高主题的覆盖度，我们取主题辞的个数为 8～10 个，即：

$$\vec{D}_i = (w_{T1}, w_{T2}, \cdots, w_{Tn}) \tag{3}$$

其中，$\vec{D}_i$ 表示指定的文档主题特征矢量，T_n 标识文档的主题辞，w 表示主题辞的权重。

3.2　基于主题概念的文档相关度计算

对于每一个分析文档我们用一组特征矢量来标识它的主题，通过矢量的距离计算来确定文档之间的主题相关度。我们是根据主题概念的关系及其在文档中的权重来对矢量进行计算的。文档间主题相关度 R 计算函数为

Function Relation_ Docu(D_X, D_Y)

$$\vec{D}_X = (w_{T1}, w_{T2}, \cdots w_{Tn})$$

$$\vec{D}_Y = (w'_{T1}, w'_{T2}, \cdots w'_{Tm})$$

$$R = \sum_{i=1}^{n}\sum_{j=1}^{m} w_i * w'_j * Concept\_Cohesion(T_i, T'_j) \tag{4}$$

3.3　文档主题重要度标识

数据集中的每一个成员文档根据与其他文档的主题相关度来确定它在文档集中的主题重要度。令集合 $M = \{D_1, D_2, \cdots, D_n\}$ 代表多文档集，下面的矩阵表示文档集中所有成员之间的两两相互相关度关系。

$$\begin{array}{c} 1 \\ 2 \\ 3 \\ \cdot \\ n \end{array}\begin{pmatrix} R_{11} & R_{12} & R_{13} & \cdots & R_{1n} \\ & R_{22} & R_{23} & \cdots & R_{2n} \\ & & R_{33} & \cdots & R_{3n} \\ & & & \cdots & R_{n-1,n} \\ & & & & R_{nn} \end{pmatrix} \tag{5}$$

其中，R_{ii} 表示文档 D_i 与自身的主题相关度，在矩阵中不予计算。R_{ij} 表示文档 D_i 与 D_j 的主题相关度。由于 $R_{ij} = R_{ji}$，在处理时仅考虑右上矩阵元素的计算。

我们定义文档 D_i 在文档集 M 中的主题重要度为：

$$V_i = \sum_{p=1}^{i-1} R_{pi}\ (i>1) + \sum_{q=i+1}^{n} R_{iq}\ (i<n) \tag{6}$$

为了后述的系统计算，我们取主题重要度的相对值：

$$V_{Ri} = V_i / \max_{i=1}^{n}(V_i) \tag{7}$$

3.4　文档集综合主题辞

对文档集中所有成员文档主题辞进行合并，根据阈值 L 选取一组主题辞 A_i，满足 $\forall a \in A_i$，Weight (a, M) > L。这样抽取的 $A = \sum_{i=1}^{m} A_i$ 作为文档集的综合主题辞，用向量表示为：

$$\vec{A} = (w_{A1}, w_{A2}, \cdots, w_{Am}) \tag{8}$$

4　多文档自动摘要生成系统（CMDIS）原理

为了有效地对一个大规模文档集的所有文档进行综合性处理，我们对人工专家完成多文档综合性摘要的过程与方法进行了研究。我们发现多数人处理文档集时，首先是浏览整个文集的内容，通过比较分析，然后抓住其中主要几篇覆盖全集主题的文献，浓缩文献信息，同时补充其他文档的特别的内容，最后润色加工成一篇摘要文档。我们的多文档摘要处

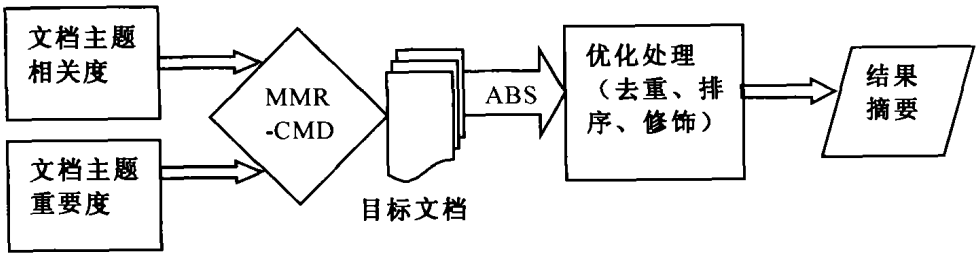

图 2 多文档自动摘要原理流程

理系统，也是参照人工处理的方法，经过文档主题特征的识别，计算文档主题覆盖全集的重要度，以其中能复合反映全集主题的文档为优先处理对象，进行抽取、压缩、排序、修饰完成最后的摘要结果。本系统的组成框图如图 2 所示。

4.1 文档综合优先度 U_i 的确定

在一个反映相关主题的多文档集中，每一个成员文档表达的主题内容可能相互交叉重叠，也可能没有直接关联。我们在进行综合处理时，应选取最重要、主题覆盖尽可能宽的文档内容，同时应去除冗余，保持文档子主题的多样性。Carbonell 曾采用最大边际相关（MMR）的方法来进行检索文档与用户查询相关性的重排序[5]。我们也采用类似的方法，以文档主题重要度及其他的影响因素来确定文档综合优先度 U_i。我们用 MMR-CMD 的算法来抽取最能集中反映全集主题的文档，同时尽量去除它们的重复性。

MMR-CMD 的算法：

$$U_i = \max_{D_j \in M \setminus S} \left[\lambda (Sim_1 (D_i , V_{Ri})) - (1 - \lambda) \max_{D_j \in S} Sim_2 (D_i , D_j , S) \right] \quad (9)$$

$$Sim_1 (D_i , V_{Ri}) = p_1 * V_{Ri} + p_2 * Type (D_i) + p_3 * Time\_sequence (D_i , M) \quad (10)$$

$$Sim_2 (D_i , D_j , S) = q * Relation\_Docu (D_i , D_j) \quad (11)$$

$$Type (D_i) = \sum_{W \in D_i} Kspec (W) \quad (12)$$

$$Time\_sequence (D_i , M) = \frac{timestamp (D_{max\ time}) - timestamp (D_i)}{timestamp (D_{max\ time}) - timestamp (D_{min\ time})} \quad (13)$$

其中，M 表示大规模的文档全集；S 表示已选的文档集合；D_i 表示待选的文档成员；D_j 表示已选的文档成员；Sim_1 表示在整个文档集中主题的覆盖度；Sim_2 用来处理主题冗余的文档；W 表示文档中含有的特定的字串，如“综合报道”、“…社综述”等；Timestamp 用来标识文档的报道或出版日期；p_1，p_2，p_3，q 为系数。

在上述算法中，我们用 Sim_1 来选取整个多文档集中最重要、综合度较高的文档，它除了主要考虑文档主题重要度 V_{Ri} 外，还参考文档类型（如文档本身就是复合或综述性的文献）。另外我们也认为日期越近的文档，在摘要中的重要性也相对较高。另一方面，为了避免抽取结果仅偏重于某一个主题最重要的侧面，去除摘要中的冗余，保持一定的主题内容的多样性，我们用 Sim_2 来处理。为了重要性及多样性之间取得平衡，使用参数 λ 来进行调整。

4.2 文档集目标句综合权重计算

文档集中所有句子根据切割标记进行句子切分。对于文档 D_i 中的第 j 个句子 S_{ij}，它的综合权重值为：

$$Y (S_{ij}) = f_1 * A_{ij} * U_i + f_2 * P_{ij} \quad (14)$$

式中，A_{ij} 为句子与文档集综合主题辞向量的相关度；U_i 是为了考虑文档的综合优先度；P_{ij} 是考虑句子在各文档中对主题、位置、指示等单文档中计算的权值；f_1，f_2 为系数。

4.3 摘要的输出

根据要求的输出长度和文档集中目标句子的综合权重的大小，选出 M 句组成摘要文本。

摘要句子去重。我们的摘要以摘录型为主，M 个摘要句子中可能还有一定的句子重复，所以我们采用一定的去重算法进行处理。

排序处理。文档集中的成员文档并没有特定的先后次序，为了在摘要中对 M 句文本进行一个合理排序，我们设计了一个摘要输出的排序算法。该算法主要考虑时序，即摘要句所在的文档的内容发生日期。如无法确定日期，则以文档的主题重要度大小来排序。

5 实验分析

5.1 评价方法

利用信息检索中的召回率和准确率的评价指标,可以对多文档自动摘要进行有效的评估[6]。设待测试的多文档系统摘要为 S,标准参考摘要为 M,在 S 中包含的摘要句子单元数为 N_S,在 M 中包含的摘要句子单元数为 N_M,同时出现在 S 和 M 中的共同句子单元数为 N_K,则定义:

准确率 Precision = N_K / N_S

召回率 Recall = N_K / N_M

召回率和准确率的值,可以反映自动摘要与标准摘要的一致度,从而来评价多文档自动摘要的性能。考虑到多文档集中目标句子数量较大,同一语义内容可能有多个形态不同的句子表示,如果仅仅简单采用相同句子数来计算评价指标,实际上并不科学。因此可以设定重合度 C 来表示目标摘要与标准摘要中的单元句子的内容关系度。为了评测方便,我们使用四级判断来度量重合度 C 的值,如句子完全一致 C = 1,大致相同为 2/3,部分相同为 1/3,完全不一致则为 0。则考虑了内容重合度的评价指标摘要准确率和召回率为:

$$Precision\_C = \frac{\sum C_S}{N_S} \tag{15}$$

$$Recall\_C = \frac{\sum C_M}{N_M} \tag{16}$$

5.2 结果分析

测试数据是从搜索引擎和新闻网站的专题中下载的新闻文档,包括 8 个新闻专题集共 200 余篇文档。我们请相关语言学专业人员从每个专题集各抽取一个约 400 字的综合摘要作为测试的基准摘要。同时我们的多文档摘要系统(CMDIS)和人工(两位研究生完成)分别生成相应字数的比较摘要。

如从某网站新闻专题栏目中下载的一组有关"911 后美国反击及各方反应"的新闻文档集(18 篇文档)。我们的系统 CMDIS 实验结果中,其成员文档主题重要度分布如表 1 所示。

图 3 表示我们的多文档摘要系统(CMDIS)产生的摘要和人工摘要分别与基准摘要比较的相关测试结果,其中 PC_CMDIS 表示 CMDIS 系统的准确率值,PC_Human 是人工摘要的准确率。

表 1　主题重要度分布

DocNo.	1	2	3	4	5	6	7	8	9
V_R	0.65	0.83	1.00	0.35	0.00	0.09	0.33	0.15	0.35
	10	11	12	13	14	15	16	17	18
	0.52	0.41	0.22	0.30	0.22	0.87	0.74	0.26	0.76

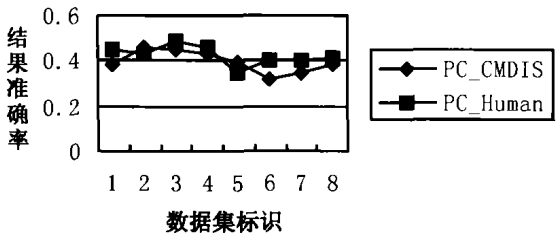

图 3　系统和人工的测试结果数据比较

测试集中 PC_CMDIS 的平均值为 39.5%,基本接近 PC_Human 的均值 42.4%,说明本系统的性能接近于人工抽取的摘要,摘要结果基本上能反映多文档集的主题内容。实验表明本系统在篇章级上来处理多文档自动摘要抽取,考虑综合主题辞和文档优先度的策略是可行和有效的。经过一定的语料测试,为平衡多文档集的综合摘要结果重要性与多样性分布,系统的 MMR-CMD 算法中参数 λ 取值 0.65 较好。

同时测试数据显示人工摘要与基准摘要也有较大的差异,反映多文档摘要本身有很大的个体差异。如何建立一个更加合理和客观的多文档摘要系统评估机制也是一个需要进一步研究的领域。

6 结　论

本文提出了一个利用文献主题概念相关性进行多文档自动摘要的系统原型。针对新闻文献领域,实验表明,能有效抽取大规模文档集中重要文档内容,满足读者对多文档集一定的预览需求。该系统中文档主题提取的准确性还需提高,同时系统合成摘要的连贯性处理有待进一步研究。

参　考　文　献

1 Udo Hahn, Inderjeet Mani. The challenges of automatic summarization. Computer, 2000,33 (11):29 ~ 36

2 董振东,董强. 知网 http://www.keenage.com, http://www.how-net.com

3 刘功申,王永成等. 基于概念粘合度(CC)的多主题分析. 情报学报,2002(1)

4 韩客松. 中文文本主题自动提取和标引若干关键技术研究:[博士学位论文].上海交通大学, 2001

5 Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summarization. In:Proceedings of SIGIN-98, Melbourne, Australia, August 1998

6 DUC 2001. The Document Understanding Workshop 2001. http://www-nlpir.nist.gov/projects/duc/2001.html

(责任编辑 许增棋)