基于子主题和用户查询的多文档摘要系统①

徐晓丹

(浙江师范大学数理与信息工程学院，金华 321004)

摘 要： 文章描述了一种基于子主题划分和查询相结合的多文档自动摘要系统的设计：首先利用同义词词林计算句子语义相似度，通过对句子的聚类得到子主题，然后根据用户的查询对子主题进行重要度排序，在此基础上，采用一种动态的句子打分策略从各个主题中抽取句子生成摘要。实验结果表明生成的摘要冗余少，信息全面。

关键词： 多文档自动文摘；子主题；句子相似度；用户查询

Multi-Document Summarization System Based on Sub Topic Partition and User's Query

XU Xiao-Dan

(Physics and Information Engineering College, Zhejiang Normal University, 321004, China)

Abstract: A multi-document summarization method based on sub topic partition and user's query is described in this paper. The similarity of sentences is measured by a thesaurus dictionary. Sub topics are found by sentence clustering and sorted by user's query. Then sentences from all sub topics are selected by using a dynamic strategy of scoring the sentences. The experiment result indicates that the summarization has less redundancy and more information.

Keywords: multi-document summarization; sub topic; sentence similarity; user's query

1 引言

网络的普及使人们在线可获取的资源呈指数级增长，如何快速的浏览信息，获取所需要的信息已成为大家关注的一个焦点。多文档文摘能对同一主题进行汇总和压缩,将包含这些文档的重要、全面的信息提供给用户，成为自然语言处理的一个研究热点。

近年来,研究者们在多文档文摘的研究上取得了一定的成绩，主要使用的方法可以归纳为两类：一是基于单文档摘要技术的方法,例如 Lncy 的基于单文档结构信息和主题概念特征生成的多文档摘要[1]，Radev 和 McKeown 的基于信息抽取技术生成的多文档摘要[2]。另一类是基于多文档集合特征的方法，该方法主要是利用多文档集合的信息,将多文档集合作为一个整体进行研究，通过对多文档集合中的句子按其表达意思的相近程度重新组合聚类，从不同的类别中抽取摘要句。例如 Radev 的基于质心距离的多文档自动文摘系统[3]。其中基于子主题聚类的方法[4]主要是根据文档中句子的相似度对句子进行聚类以划分文档主题，在此基础上生成文摘，这是种理论上有效的方法。在实际的处理中，有几个关键问题亟待解决：一是句子相似度计算，最常用的是基于词形建立向量空间模型的句子相似度计算方法，但该方法忽略了词之间的语义联系，影响了相似度的计算；二是如何从各个主题中抽取句子，以达到最大化的主题覆盖率和最小的冗余也是一直没有得到很好解决的一个问题。

为此，本系统从文本内容出发，以同义词词林计算句子的语义相似度为基础，首先形成多文档集合的各个子主题，并根据用户查询和子主题的相关性对主题的重要性进行排序，在此基础上根据用户的查询和句子相似度计算句子权重，并选取句子生成文摘,使生成的摘要冗余少，信息覆盖率高。下面的内容组织如下：第 2 节介绍子主题的划分和排序，第 3 节介绍从子主题中抽取摘要句的方法，第 4 节分析实验结果，第 5 节给出结论。

① 收稿时间:2010-08-14;收到修改稿时间:2010-08-30

2 子主题的划分和排序

多文档集合内的文档通过某个共同的主题关联起来,这个共同的主题称为中心主题。一般来说，多文档包含一个中心主题，而中心主题又由若干子主题构成，这些子主题分别代表多文档集合中的各个局部信息。在对文档进行自动文摘时，除了考虑句子的重要性之外，还要综合考虑句子所在的子主题。文档的摘要应该能较全面地概括各个局部主题的信息。

句子相似度的计算是子主题形成的必要环节。本文借助句子中词与词之间的语义关系计算句子之间的语义相似度，从而实现主题的划分。

2.1 句子相似度计算

本文所涉及的句子相似度的计算是基于有效词的集合。文档通过分词后获得词的集合，文档通过分词后获得词的集合，在这基础上再进行停用词的处理，所剩下的所有词看作候选有效词，由于在这些词中存在着许多价值较小的高频词和低频词，我们通过采用文档频率方法来抽取高、低频词。具体方法是：对语料库中的每个文本集合中的候选有效词计算文档频率FW，即指出现了该词的文档个数，从候选词中去除FW低于某个预定阈值的低频词或高于某个预定阈值的高频词，剩下的候选词作为文本的有效词。一个句子由若干个有效词组合，句子之间的相似度通过计算句子所包含词汇之间的语义距离获得。句子相似度的计算公式如下所示：

$$Sim(A,B)=\frac{\sum_{i=1}^{m}Sim(a_i,B)+\sum_{j=1}^{n}Sim(b_j,A)}{m+n} \quad (1)$$

其中 A 和 B 表示两个句子， 句子 A 包含的有效词为 a_1， a_2……a_m， 句子 B 包含的有效词为 b_1,b_2……b_n,a_i 和 b_j 之间的语义相似度用 Sim(A,B)表示，Sim(a_i，B)表示词 a_i 和句子 B 的相似度，Sim(b_j，A)表示词 b_j 和句子 A 的相似度，Sim(a_i，B)=Max(Sim(a_i,b1),Sim(a_i, b_2),……Sim(a_i， b_n))， Sim(b_j ， A)=Max(Sim(b_j, a_1),Sim(b_i， a_2),……Sim(b_i， a_m)),Sim(a_i,b_j)表示两个词 a_i,b_j之间的语义相似度。

为了计算词语之间的语义相似度，我们采用哈尔滨工业大学信息检索研究室提供的《同义词词林扩展版》作为语义知识资源。《同义词词林》[5]是 20 世纪 80 年代出版的一部对汉语词汇按语义全面分类的词典，它采用层级体系，按照树状的层次结构把所有收录的词条组织到一起。该词典收录词语近 7 万，把词汇分成大、中、小三类，小类以下再以词义的远近和相关性划分词群，每个词群中的词语又进一步分成了若干个行，同一行的词语要么词义相同(有的词义十分接近)，要么词义有很强的相关性。《同义词词林》对里面的词语提供了三层编码，即大类用大写英文字母表示，中类用小写英文字母表示，小类用二位十进制数表示。哈工大的扩展版在此基础上，新增了第四级和第五级的编码，第四级用大写英文字母表示，第五级用二位十进制整数表示。新增的第四级和第五级的编码与原有的三级编码合并构成一个完整的编码，唯一的代表词典中出现的词语。例如:Ba01A02=3 物质 质素，Ba03A11= 2 证物 信物。这就为语义距离的计算提供了方便。

两个词之间的相似度是和它们之间的语义距离相关的，语义距离越小，相似度越高。反之亦然。因此可以通过对词的语义距离计算得到相似度。其公式可以写成：

$$Sim(a_i,b_j)=\frac{\alpha}{\alpha+D(a_i,b_j)} \quad (2)$$

在上式中，D(a_i,b_j)表示词 a_i,b_j之间的语义距离，它是根据两个词的语义编码来得到的，具体见公式(3)；α 是一个可调节的参数，本文中 α 的取值为 4。

D(a_i,b_j)=2*(6-n) (3)

其中，n(2<=n<=6)为它们之间的语义代码从第 n 层开始不同，当 n=6 时，表示前面的 5 层全部相同，语义距离为 0,说明两个词为同义词，它们的相似度就为 1。例如“桃子” Bh07A28，“西瓜”Bh07A56，“可惜” Gb17A01，“惋惜” Gb17A01。用公式计算可知 D(桃子，西瓜)=2，D(可惜，惋惜)=0。

如果两个词的语义代码从第一层就开始不同，考虑到《同义词词林》大类之间的相关性，如第一至第四大类多为名词，第六至第十大类多为动词等等，本系统中将语义代码从第一层就开始不同但同属于A、B、C、D 大类或者同属于 F、G、H、I、J 大类的词语之间的语义距离设为 D(a_i，b_j)=12，否则就设为+∞。

2.2 子主题划分与排序

计算出句子的相似度后，就可以进行子主题的划分。本文采用系统聚类的方法，初始状态假设每个句子自成一类，将相似度超过一定阈值的句子聚成一类，最终获得多文档文摘的多个子主题。其方法如下：

1）将句子集合 S={S_1, …, S_i, …, S_n}中的每个句子看作一个类 C_i={s_i}，这些类构成了集合 C={C_1, …, C_i, …, C_n}；

2）计算 C 中每两类（C_k, C_l）之间的距离 D_{kl}；

3）选具有最小距离的类对 $\text{argmin}D_{kl}$,并将 C_k,C_l 合并为一个新的类 $C_M=C_k \cup C_L$,集合 C 变为为 C={C_1, …, C_{n-1}}

4）重复 2,3 直至达到聚类的阈值为止。

其中阈值的确定是关键环节，本文采用文献[6]的方法，通过标准语料的训练来确定划分子主题的阈值[6]。在该方法中，通过一个半偏系数 HDCC 来观察聚类结果。其计算方法为：HDCC=(W_M-W_K-W_L)/T,其中，分子 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，SIM(x_i, x_j) 表示任意两个句子的相似度，分母 $T = \sum_{i\neq j} (1 - SIM(x_i, x_j))^2$ 表示所有类中文本单元总的离差平方和。对训练语料中每一个文档集合聚类,当达到标准的聚类数时，会得到两个值,即到达标准聚类时的 $HDCC_i$ 以及超过标准聚类一次的 $HDCC_{i+1}$，通过对多个标准文档集的 $HDCC_i$ 和 $HDCC_{i+1}$ 的比较,得到统一的阈值 HDCC。

子主题划分后，多文档集合 D 可以描述成由多个子主题 T_i 构成的：D={T_i|i=1,2…k}，每个 T_i 是一个句子集合:Ti={$S_{i,k}$|k=1,..m}。子主题是对多文档的各个层面信息的描述，为了更好的获取摘要，在子主题划分后还需要进行重要性排序，以确定文摘句的抽取顺序。在本系统中，我们使用下面的公式计算子主题 Ti 和查询的相似度：

$$Sim(Q,Ti) = \frac{1}{m}\sum_{k=1}^{m} Sim(Q, S_{i,k}) \quad (4)$$

其中， Sim(Q,T_i)表示查询 Q 和某个子主题 T_i 之间的相似度，Sim(Q,$S_{i,k}$)表示查询 Q 和主题 Ti 中的句子 $S_{i,k}$ 的相似度,其计算公式采用 2.1 节的公式(1)，Sim(Q,T_i)取 Q 和主题内所有句子之间的相似度的平均值。

3　摘要生成

3.1 句子评分策略

子主题的重要性排序确定后，就可以依据优先次序从各个主题中抽取重要的句子组成摘要。为了保证抽取出来的句子包含全面的信息，应该抽取主题中包含信息最多且和查询相关最大的句子，因此，我们从两个方面给出句子的评分：

1）句子 Si 和主题内其他句子的相似度，用于衡量句子在一个主题中的重要性。

$$C_1(S_i) = \frac{1}{m-1}\sum_{Sj\in T\text{-}Si} Sim(S_i, S_j) \quad (5)$$

2）句子 Si 和查询 Q 的相似度。其公式如下所示：

$$C_2(S_i) = Sim(S_i, Q) \quad (6)$$

综合考虑上述因素，得到句子 Si 的评分公式为：

$$Score(S_i) = tC_1(S_i) + (1-t)C_2(S_i) \quad (7)$$

3.2 去冗余的句子选择方法

同一文档中的句子之间存在着信息冗余，这种情况在来自不同文档的句子之间尤为突出，常常会存在一些在内容上非常接近的句子。为了尽可能减少信息冗余，在从各个子主题中抽取句子时，要考虑句子间的信息冗余情况，进行冗余去除。

参考 MMR 的信息冗余处理策略[7]，在考察摘要侯选句时，要选择包含重要信息且与已选句子重复信息少的句子，根据这一原则，我们使用下面的公式来更新句子 Si 的当前 Score 值：

$$Score(S_i) = \lambda Score(S_i) - (1-\lambda)Sim(S_i, S) \quad (8)$$

其中 λ 为两者之间的权值参数，$\lambda \in [0,1]$，Sim(Si,S)表示句子 Si 与当前摘要句子集合 S 的相似度，计算公式为：

$$Sim(S_i, S) = \frac{1}{m}\sum_{j=1}^{m} Sim(S_i, S_j) \quad (9)$$

我们采用的下面步骤动态的抽取句子：

1）对子主题排序，得到按权重大小排序的 n 个子主题{T_1, T_2, ……T_n}；

2）对每个主题内的句子按照 Score 值大小降序排序；

3）按顺序从子主题中抽取 Score 值最大的句子加入摘要。如果摘要达到指定长度，终止；否则对其余的句子根据公式(8)重新计算 Score 值；

4）转到第 2 步，重复该过程。

4　实验与评估

对多文档自动摘要的评价目前还没有统一的评价[8]标准。我们采用 DUC 会议使用的评价指标进行人

工评价。自 2005 年开始，DUC 的文摘任务设定为面向查询的多文档文摘任务，在评价摘要时，主要有以下几个指标：

1) 响应度(Responsiveness):评价一篇摘要能在多大程度上满足查询描述中提出的信息需求。分值为 1-5。

2) 语言质量(Linguistic Quality)：评价一篇文摘的可读性和流畅性。主要有 5 个方面：语法性、非冗余性、指代明晰、焦点、结构和连贯性。

我们建立的多文档自动摘要系统能对网页和普通文本进行处理。实验的语料库来自 2001 年的人民网的原始网页，包含军事、国际、经济等八个大类别共 5096 个网页。切词使用的实用词词典是对大规模的网页库统计后得到的。在实用词词典的生成过程中，要使用两个词典库：标准主题词词库和平凡词词库。其中标准主题词词库来源于国内通用的汉字词库。平凡词词库即汉字虚词词库。实用词词典是在标准主题词词库的基础上，经平凡词库过滤后的词库。

在实验中从语料库的军事、经济、生活类别中分别抽取若干个主题的文章，每个主题包含 5-10 篇文章。根据响应度和语言质量由专家给出得分，摘要得到摘要的统计结果如下：

表 1 各个类别的平均响应度和平均质量得分

	平均响应度(1-5 分)	平均质量(1-5 分)
军事类	4.1	3.75
经济类	3.75	3.5
生活类	3.41	3.1

同时，我们也建立了一个基于质心的多文档自动摘要系统原型（F1）：首先根据 tfidf 方法从某主题的文档集中选取重要的词，然后和查询的词合在一起作为质心，计算质心和每个句子的相似度，将相似度高的句子提取出来，经过适当冗余处理后生成摘要。为了比较，首先给出专家的标准结果，两个系统在 15%摘要比下的平均准确率,平均召回率和平均 F-measure 值如下表所示：

表 2 两个系统的召回率、准确率和 F-measure

系统	召回率 R	精确率 P	F-measure
F1	0.658	0.649	0.653
F2	0.708	0.689	0.699

由表中数据可以看出，F1 基于质心的方法从整个文档集合的角度对句子排序，没有考虑子主题之间的信息。本文提出的基于子主题排序和查询的句子选择方法不仅包含了重要信息，同时由于考虑到各个主题之间的相关性，有效的减少了句子的冗余。

5 小结

本文论述了一种基于子主题和用户查询的多文档自动摘要方法，在该方法中，基于同义词词典的句子相似度计算方法较客观的描述句子之间的语义关系，基于查询的子主题抽取句子策略能较好的把每个主题中重要的句子抽取出来，并减少句子之间的冗余。实验表明本文提出的方法是有效。在今后的工作中，我们将针对摘要句中的指代不明以及句子信息不一致等情况做进一步研究，以改善摘要的可读性，提高摘要的质量。

参考文献

1 Lncy HE. From single to multi-document summarization: a prototype system and its evaluation. Proc. of the 40th Annual Metting of the Association for Computational Linguistics. Philadelphia: ACL, 2002: 457－464.

2 Dragomir RR, Kathleen RM. Generation Natual Languages Summaries from Multiple Online Sources. Computational Linguistics, 1998,24(3):21－29.

3 Radev R, Hongyan J, Malgorzata B. Centroid-based summarizaiton of multiple documents: sentence extraction, utility-based evaluation, and user studies. ANLP/NAACL 2000 Workshop C. 2000: 21－29.

4 Boros E, et al. A clustering based approach to creating multi-document summaries. Proc. of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New Orleans, LA, 2001: 34－42.

5 梅家驹.同义词词林仁.上海:上海辞书出版社, 1983.

6 秦兵,刘挺,陈尚林,等.多文档文摘中句子优化选择方法研究.计算机研究与发展, 2006,43(6):1129－1134.

7 Elhadad NJ. The use of MMR,diversity-based reranking for recording documents and producing summaries. SIGIR-98. Australia, 1998:335－336.

8 魏继增,孙济舟,秦兵.多文档文摘评价标准的研究.计算机工程与应用, 2007,43(2):180－183.