面向查询的多模式自动摘要研究*

李芳[1]，何婷婷[2]

[1](国家数字化学习工程技术研究中心，武汉 430079)，[2](华中师范大学计算机科学系，武汉 430079)

Email: fang_li@mails.ccnu.edu.cn tthe@mail.ccnu.edu.cn

摘要：为了满足用户的个性化需求，提供尽可能丰富、实用、方便的文摘结果，本文设计了面向查询的多文档自动文摘的多种摘要模式。在将查询返回的文档集合表示为以文本、段落为节点的双层复杂网络结构以发现子主题的基础上，除传统的摘要模式外，本文又设计了概括摘要、偏重摘要、全局摘要和详细摘要这四种摘要模式，并给出了各种摘要的生成方法。支持用户以主题为线索自主漫游，按照一定的逻辑顺序浏览信息。

关键词：面向查询的多文档自动文摘，子主题发现，多模式摘要

Research on Query-directed Multi-mode Automatic Summarization

Fang Li[1]; Tingting He[2]

[1]National Engineering Research Center for E-learning, 430079 Wuhan, China

[2]Department of Computer Science, Huazhong Normal University, 430079 Wuhan, China

Email: fang_li@mails.ccnu.edu.cn tthe@mail.ccnu.edu.cn

Abstract: In this paper, we design a variety of summary mode for Query-directed Multi- document Summarization to satisfy the individual requirements and provide convenient results. Firstly, the documents are represented as two-layer complex networks, whose nodes describe text and paragraph respectively, and the thought of network community discovery algorithm is used to cluster text and paragraph. Then, on the basic of network structure of documents, we design four summary modes besides the traditional summary mode and summary element extract strategy. They are document summary, general summary, partial summary, global summary and detailed summary. With the clues of sub topic, users can browse information in certain logical sequence to their own.

Key words: Query-directed Multi-document Summarization; Sub-topic Discovering; Multi-mode Summary

1 引言

随着Internet的飞速发展，越来越丰富的信息出现在网络中，文本数量以指数级的速度增长，这极大方便了人们对信息的获取和使用。但是，随着网络上信息的逐渐增多，在这些海量信息中快速准确地找到所需要的信息也越来越困难[1]。

自动文摘技术是解决当前信息过载问题的一种辅助手段，正日益受到国内外学术界和工业界的密切关注。该技术将文档的主要内容在较短时间内提供给用户，可以提高人们获取信息的效率，给用户判断和浏览感兴趣的内容提供帮助。

面向查询的多文档自动文摘技术是将查询返回的文档集合中的相关内容浓缩为一个包含查询主题各个方面的、内容简洁、组织良好、冗余低、满足个性化需求的摘要。其研究目的在于解

*基金资助：国家自然科学基金重大研究计划(90920005)；国家自然科学基金(60773167)；国家十一五科技支撑计划课题“网络文化安全预警技术研究”(2006BAK11B03)；973国家重点基础研究发展计划(2007CB310804)；教育部/国家外国专家局高等学校学科创新引智计划 (B07042)；湖北省自然科学基金计划项目(2009CDB145)；武汉市晨光计划项目(201050231067).

决从海量数据中获取有用信息的困难，提高信息获取及浏览的速度、适应不同用户对信息的个性化需求，从而提高用户获取和利用信息的效率，提高用户在信息社会中的竞争实力。目前，国内面向查询多文档文摘的研究大多都是围绕DUC(Document Understanding conference)[2]、TAC（Text Analysis Conference）[3]比赛的。Prasad Pingali、Florian Boudin、李素建、何婷婷、滕冲[4-8]等提出了一系列的自动文摘生成方法，这些方法生成的最终摘要都力求既满足用户的查询需要，又尽可能覆盖文档集合的所有内容。但是，有时用户可能只想大致了解文档集合中的信息分类或用户可能只对查询结果集合中的某一个方面的内容感兴趣等等，对于这些个性化的需求，传统的摘要模式就很难给出满意的结果。为此，本文除传统的摘要模式外又设计了概括摘要、偏重摘要、全局摘要和详细摘要四种摘要模式，以更好地满足用户的个性化需求，提供尽可能丰富、实用、方便的文摘结果。

2 文档的复杂网络表示与子主题发现

首先，我们将查询返回的文档集合表示为带权的复杂网络，然后，利用复杂网络中的抱团发现思想来发现文档集合中的子主题。

2.1 文档的复杂网络表示

这里，我们用带权复杂网络来重构查询返回的文档集合，边的权值是两个节点之间的相似度值。首先将每个文本表示成向量，计算相似度矩阵W，再来构建带有权重的网络图：$P=(V, E)$，其中$V=\{S_1, S_2, \cdots S_n\}$是网络中节点的集合，对应所有的文本；$E=\{(S_i, S_j) \mid S_i$和$S_j$的相似度值$w_{ij}$大于某个阈值$\}$是边的集合，边的权重是网络两个节点的相似度值$w_{ij}$。

为了方便后续的子主题发现和降低计算复杂度，文档集合的网络拓扑图表示是通过两个阶段表示的。在第一阶段，网络结构中的每个节点表示一个文本，边的权值为两个文本的相似度，每个网络抱团对应于一个文本类。第二阶段的文档表示是在文本聚类之后，网络结构中包含多个子网，每个子网表示第一阶段中的一个文本类。子网中的每个节点表示一个段落，边的权值为两个段落的相似度。每个网络抱团将对应一个子主题。这样，文档集合的网络拓扑图就形成了上下对应的两层结构，如图1所示。这个结构是文本聚类和段落聚类的基础，也是后面多种摘要模式设计的基础。

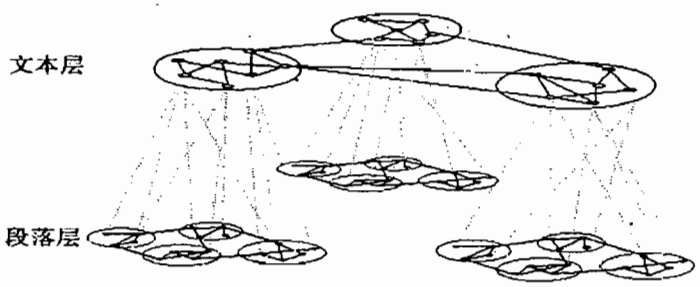

图1 文档集合的网络拓扑图

2.2 子主题发现

子主题发现的基本方法是聚类，每个类是一个子主题。基于聚类的子主题发现过程就是在计算相似度的基础上把聚类单元聚合成不同主题类别。

2.2.1 基于抱团发现的聚类方法

对于复杂网络抱团结构分析，Newman、Clauset等[9,10]提出了基于贪婪算法思想的凝聚算法：Newman快速算法和CNM算法。他们先假设网络中的每个节点为一个抱团，然后再来合并有边相连的抱团对，合并的原则是使模块度Q增大最多或者减少最小，直到整个网络都合并成为一个抱团。此方法是建立在非带权网络上，将每条边看作是相等的。而我们的文档集合表示为一个带权网络，所以我们必须重新定义节点度和模块度。

加权度：$WD_i = \sum_{j \in \Gamma(i)} w_{ij}$，其中$\Gamma(i)$表示节点$i$的邻居节点的集合。

加权模块度：$WQ = \sum_i (we_{ii} - wa_i^2) = \sum_i (\frac{A(V_i,V_i)}{A(V,V)} - (\frac{A(V_i,V)}{A(V,V)})^2)$，其中$A(V_i,V_i)$表示抱团$i$中所有边的权重和，$A(V_i,V)$表示所有与抱团$i$中的节点相连的边的权重和，$A(V,V)$表示整个网络中所有边的权重和。

那么，在文档集合的复杂网络表示基础之上，利用CNM算法思想，就可以得到自动发现社区结构的聚类算法。

表1 自动发现社区结构的聚类算法流程

输入：P （原始网络图）

输出：G_p（P下的抱团组成的集合）

1 初始化$G_p = \{P\}$，即每个节点就是一个独立的抱团。初始的e_{ij}，a_i为：

$$e_{ij} = \begin{cases} \frac{w_{ij}}{\sum w_{ij}} & \text{如果节点 } i\text{和}j\text{相连} \\ 0 & \text{其它} \end{cases} \qquad a_i = \frac{WD_i}{\sum w_{ij}}$$

2 初始化模块性增量矩阵：$\Delta Q_{ij} = \begin{cases} e_{ij} - a_i \cdot a_j & \text{如果节点 } i\text{和}j\text{相连} \\ 0 & \text{其它} \end{cases}$

3 初始化最大堆H为ΔQ每一行的最大元素。

4 从最大堆H中选择最大的ΔQ_{ij}，合并相应的抱团i和j，标记合并后的抱团的标号为j；

更新ΔQ：删除第i行和第i列的元素，更新第j行和第j列的元素：

$$\Delta Q'_{jk} = \begin{cases} \Delta Q_{ik} + \Delta Q_{jk} & \text{如果抱团 } k\text{同时与抱团 } i\text{和抱团 } j\text{都相连} \\ \Delta Q_{ik} - a_j a_k & \text{如果抱团 } k\text{仅与抱团 } i\text{相连，不与抱团 } j\text{相连} \\ \Delta Q_{jk} - a_i a_k & \text{如果抱团 } k\text{仅与抱团 } j\text{相连，不与抱团 } i\text{相连} \end{cases}$$

更新最大堆H：更新ΔQ_{ij}后，就要更新最大堆中相应的行和列的最大元素。

更新向量a：$a'_j = a_i + a_j$；$a'_i = 0$

5 重复步骤4，直到模块性增量矩阵ΔQ中最大的元素由正变到成负后。

6 结束并返回G_p。

算法输出的每个抱团即为一个类，这是一种自适应的聚类方法，可以自动确定类的个数，自动发现文本集合中包含的子主题。它在理论上能够较好地解决一些聚类方法对初始解敏感、易陷于局部最优的缺点，能在全局进行搜索。

2.2.2 文本、段落两阶段聚类策略

在文档集合的双层复杂网络表示的基础上，我们提出文本、段落两阶段聚类的策略来识别文本集合中的潜在子主题。先对文本进行聚类，然后再对每个文本类分别进行段落聚类，每个段落类为一个子主题。这样既可以避免因为聚类单元太小而造成的失去文本内容的许多关联的缺点，又不会因为聚类单元过大而带来太多冗余信息。

复杂网络抱团发现方法的效果会随着网络中节点数目的增加而变差。采用文本、段落两阶段聚类的策略可以有效地控制网络中节点的数目，段落聚类时，只考虑属于某个文本类的段落，而不是所有的段落。这样，网络中的节点就不至于太多，从而在一定程度上提高了聚类的效果。

3 个性化摘要模式

面向查询的多文档自动文摘的一个重要特征是要能满足用户的个性化查询需求。在文档集合的网络拓扑图的支持下，除传统的摘要模式外，又产生了偏重摘要、概括摘要等多种摘要模式。

3.1 个性化摘要模式设计

由于我们对查询返回的文档集合进行了两个层次的复杂网络表示，一层为文本网络拓扑图，另一层为段落网络拓扑图，因此我们能够方便地应用这双层网络结构，以多种策略构成摘要，更好地服务于用户的个性化需求，提供尽可能丰富、实用、方便的文摘模式。

（1）文档摘要

文档摘要以段落类为子主题，把文本集合的所有段落类作为一个整体，从每个段落类中选择文摘句来构成摘要。

文档摘要是最基本、最常见的文摘模式，摘要内容对查询主题覆盖全面且文字简洁、冗余低、可读性强。用户可以在短时间内全面概括地了解查询对象，达到查询目的。

（2）概括摘要

概括摘要以文本类为子主题，从每个文本类分别选择若干句子来生成摘要。

概括摘要是所有文档集合整体上的概述，把每个文本类所描述的内容简要地提供给用户。这样就利于用户对信息作快速浏览和分类，选择自己真正感兴趣的内容，在下面将要介绍的局部摘要的支持下，用户可以作进一步的了解，还可以使用详细摘要模式了解更详细的信息。

（3）局部摘要

局部摘要只单独从一个文本类的各个段落类中提取文摘句，产生一个较完整的摘要。

通常，用户只是对查询结果集合中的某一个方面的内容感兴趣，因此我们设计了局部摘要，这是一种偏重摘要，它只与当前文本类中每个段落子主题密切相关，生成的文摘内容更加集中，能够比较详细地描述当前文本类的主要内容。这样用户就可以只浏览自己感兴趣的内容，从而节省了大量的时间和精力。

（4）全局摘要

所有局部摘要构成整个文本集合的全局摘要。

全局摘要也是对查询返回集合的全面描述，需要从每个段落类中抽取文摘句，但与文档摘要相比，由所有局部摘要合并而成的全局摘要对每个文本类来说，其内容也更加集中，有利于提高文摘句排序的质量，提高文摘的可读性。在技术上，当文本集合很大时，采用这种自底向上的模式可以降低文摘生成的难度。

（5）详细摘要

详细摘要是从每个段落类中提取核心段落，构成以段落为单位的摘要。

这种形式的文摘，可以让用户更准确、更详细地掌握和理解这一个子主题的主要信息，可以弥补单纯追求简洁性的以句子为单位的文摘的一些不足。

图2给出了五种摘要模式的整体示意图。在文档集合的网络拓扑图的支持下，用户可以以主题为线索自主漫游，从文摘句、核心段落、子主题（段落类）到文本类多层次扩展，按照一定的逻辑顺序浏览信息，从而更准确、快速地把握信息，满足查询需要。

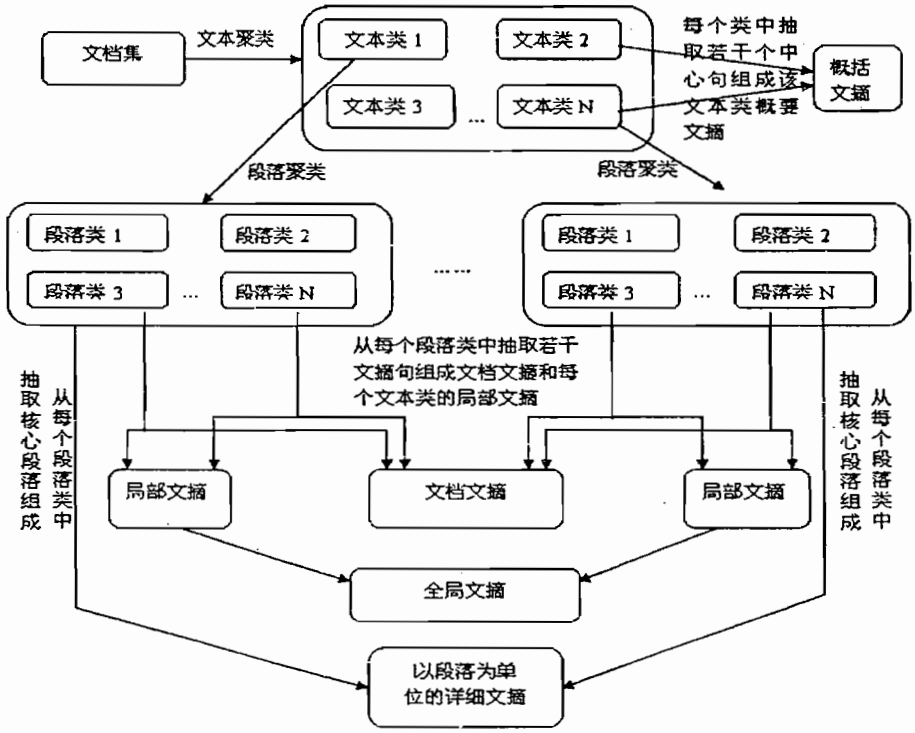

图 2　五种摘要模式示意图

3.2 文摘单元抽取策略

从摘要模式示意图中可以看出，本文抽取的文摘单元有两种：句子和核心段落。对于文摘句的抽取，我们利用了一种基于关键词提取的文摘句提取策略[11]。对于核心段落，可以借鉴复杂网络中挖掘重要节点的理论和方法，下面将重点介绍核心段落的提取方法。

如果某个段落跟当前子主题下的其他很多段落都具有一定的相似度，那么这个段落就可能覆盖了此子主题的大部分内容，因此应该把它作为一个核心段落。在网络中就表现为这个节点与网络中的其他节点都有边相连。我们所构建的网络是带权网络，边的权值是两个段落之间的相似度值，因此，节点加权度也应该纳入考虑。这样当两个节点度相同时，就可以用节点加权度衡量它们的重要性。另外，核心段落必须与查询具有较大的相关性。由此，给出了节点重点性公式：

$$I(x) = \beta_1 \frac{Dx}{2m} + \beta_2 \frac{WDx}{\sum w} + \beta_3 F_X$$

其中 Dx 为节点度，m 为网络中的总边数。WD_x 为节点加权度，$F(x) = \frac{1}{|n|} * \sum_{w_j \in x} Weight\ (w_j)$

为节点 x 所表示的段落包含的关键词得分。$\beta_1, \beta_2, \beta_3$ 是通过实验得出的调整参数，且 $\beta_1 + \beta_2 + \beta_3 = 1$。

在得出每个节点的重要性值后，得分最高的节点所表示的段落就是该段落类的核心段落。

4 实验结果及分析

我们收集了十个查询主题共200篇文章，对于每个主题，把从搜索引擎上返回的前20篇左右的文本作为与查询相关的文档集合，再分别进行聚类发现子主题，生成各种模式的摘要。这十个查询主题分别是奥巴马就任美国总统、奥巴马就任美国总统、台湾选举马萧获胜，“入联公投”失败、“神七”发射，中国航天员首次太空行走、汶川5.12特大地震灾害、中国火车出轨相撞、陈水扁家族海外洗钱案、美国次债问题造成世界金融危机、第29届奥运会在北京拉开帷幕、三鹿奶粉事件、全球华人反“藏独”反暴力。

4.1 聚类效果分析

我们对十个查询主题下的200篇文本进行了文本聚类，正确率是84.5%。然后在此基础上又进行了段落聚类，正确率是92%，可见应用文本、段落两阶段聚类策略可以提高聚类的效果。

实验中，我们是逐步加入每个主题下的文档集合来进行聚类的，我们发现随着文本数即网络中节点数的增加，聚类的效果会有所下降，具体情况如图3所示。

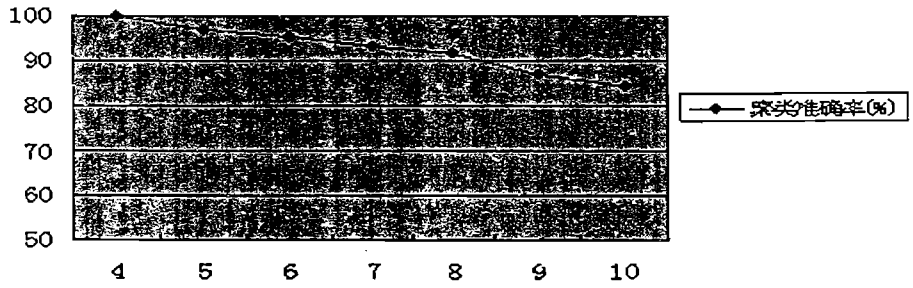

图3 主题数与准确率的关系

在面向查询自动文摘中，查询返回文本的初始类别本身就不会很多，文本聚类效果就有可能比较好，再加上段落聚类对文本聚类效果的弥补，可以得出：基于复杂网络抱团发现的文本、段落两阶段聚类方法用于面向查询自动文摘领域是完全可行的。

4.2 摘要效果分析

我们给出了十个查询主题下每种摘要模式的平均召回率（压缩比为10%），召回率的计算是基于人工标准摘要的。如果人工摘要包含 m 个句子，机器摘要包含 n 个句子，机器摘要和人工摘要重合的句子数为 k，则召回率为 $R = k / m$，这里的重合是指两个句子的相似度值超过某个阈值，并不要求它们完全相同。具体结果如图4所示。

从图中可以看出，所有的召回率都在50%以上，说明了系统得到的文摘效果是可以接受的，基本上能够满足每种摘要模式的要求。

但是，还可以看到，文档摘要和全局摘要的效果稍微要差一些，主要是因为段落类重复，造成文摘结果的冗余偏大，而人工摘要则会选取不同的句子。另外，本文的摘要结果并没有进行文摘句排序，这些都是以后需要改进的地方。

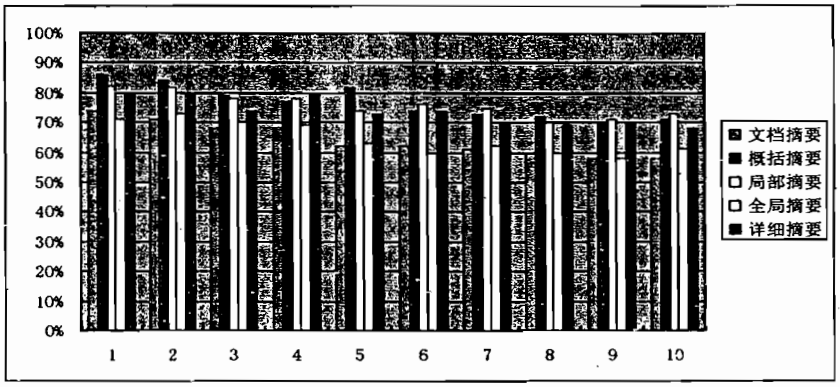

图4 每个查询主题五种摘要模式的平均召回率

5 小结

本文针对面向查询的多文档自动文摘的相关技术展开了讨论。首先，提出了文档集合的复杂网络表示方法，将整个文档集合表示为以文本、段落为节点的双层网络结构。重新定义了模块度增量矩阵，采用CNM算法思想对文本、段落进行自适应聚类，自动发现文本集合中包含的子主题。接着，充分利用了文档的网络拓扑结构，除了传统的摘要模式外，我们又设计了概括摘要、偏重摘要、全局摘要和详细摘要四种摘要模式，更好地满足用户的个性化需求，提供尽可能丰富、实用、方便的文摘结果，支持用户以主题为线索自主漫游，按照一定的逻辑顺序浏览信息。

本文是对面向查询的多文档摘要技术的一个初步探索，还有很多需要进一步研究的问题。摘要模式在一定程度上满足了用户的要求，但是抽取的文摘单元只有句子和段落。而我们认为面向查询的文摘单元也可以是若干个关键词、文本的一个区域等等。另外，我们还将试验利用其他的文摘单元抽取策略，以得到更好的摘要结果。另外，句子和段落的提取只应用了表层信息。为了取得更好的效果，需要对文本或句子进行深入理解，从句子整体的语义信息对相似度和重要性进行更好的刻画。

参 考 文 献

[1] Lynette Hirschman, R.Gaizauskas. Natural Language Question Answering: The View from Here. Natural Language Engineering, 2001; 7(4):275-300.

[2] DUC: http://duc.nist.gov.

[3] TAC: http://www.nist.gov/tac.

[4] Prasad Pingali, Rahul K, Vasudeva Varma. IIIT Hyderabad at DUC 2007. In Proceedings of DUC2007.

[5] B.Florian, B.Fr´ederic, M.El-B`eze,et al. The LIA summarization system at DUC2007. In Proceedings of DUC2007.

[6] Sujian Li, You Ouyang, Bin Sun, Peking University at DUC 2006. In Proceedings of DUC2006.

[7] 邵伟,何婷婷,胡珀等. 一种面向查询的多文档文摘句选择策略. 第九届全国计算语言学学术会议, 2007.8.

[8] 滕冲,何炎祥,刘德喜等.基于基本要素的用户聚焦型文摘内容选择. 2007 年中文信息处理国际会议,2007.

[9] 解伟, 汪小帆. 复杂网络中的社团结构分析算法研究综述. 复杂系统与复杂性科学, 2005.7

[10] Newman MEJ. Fast algorithm for detecting community structure in networks. Phys Rev E, 2004.

[11] 马亮,何婷婷,李芳,陈劲光,邵伟. 以关键词抽取为核心的文摘句选择策略. 中文信息学报,2008.