改进的最小最大聚类方法在新闻主题来源追踪中的应用

周序生

（湖南工业大学，湖南 株洲 412008）

摘 要：在分析新闻文档的特殊结构、内容特点以及常用聚类算法优缺点的基础上，提出了一种基于改进的最小最大聚类方法的主题来龙去脉生成方法。实验结果证明，该方法生成的摘要能有效地覆盖主题的内容，较为准确地体现主题的演变过程。

关键词：新闻主题；多文档自动文摘；IMMC

中图分类号：TP393 **文献标识码**：A **文章编号**：1673-9833(2009)01-0066-05

The Application of Improved MMC Method in News Topics' Sources Tracing

Zhou Xusheng

(Hunan University of Technology，Zhuzhou Hunan 412008，China)

Abstract: By analysing the special structure and contents of news documents as well as the advantages and disadvantages of clustering algorithm, a new generating method based on the improved min-max clustering method is put forward.. The experimental results prove this method can cover the subject matter effectively and embody the evolutional process of the topic correctly.

Key words: news topic；multi-document automatic summarization；IMMC

0 引言

网络信息的爆炸式增长给信息处理技术带来巨大的挑战，新闻报道则是其中主要的信息类型之一[1]。人们十分希望能方便快捷地了解某一热点事件或感兴趣事件的内容及其变化发展情况。在新闻主题检测技术的基础上，结合目前流行的多文档摘要技术，根据新闻信息的特点，笔者提出一种有效的新闻主题来龙去脉生成方法。它将多篇同一主题的文档进行汇总整理，将其中多次重复的相关信息以简洁的方式一次性表达在文摘中，解决冗余信息给人们带来的困扰，为用户提供高层次服务。

新闻主题来龙去脉生成是指采用多文档自动摘要技术对某个特定主题生成摘要，使人们方便快捷地了解主题的内容及其变化发展情况。其提取的对象是一个主题，也就是一个核心事件及其相关活动的报道，目标是对一个主题形成一种目的性的摘要。

多文档自动摘要技术是随着互联网上的信息急剧膨胀而发展起来的新的文本信息处理技术，是将多文档集合中的多次重复信息一次性呈现于文摘中，并将其它与主题相关的信息根据重要性依次抽取的文本集合压缩技术[2]。多文档自动文摘技术的提出是继单文档自动文摘之后对文本压缩技术的又一挑战，并在近几年随着DUC（Document Understanding Conference）等

收稿日期：2008-11-25

基金项目：湖南省教育厅科研基金资助项目（08C285）

作者简介：周序生（1970-），男，湖南常德人，湖南工业大学副教授，硕士，主要研究方向为智能信息处理和网络安全，E-mail：zxslxy@163.com

国际评测会议的连续举办有了较大突破[3]。

在多文档自动摘要技术中，采用聚类方法区分主题，达到去冗和全面提取内容的目的，即将每篇新闻的文摘聚类，算法收敛后形成的 K 个聚类中心作为最后的多文档摘要输出。经典 K-means 法[4]是目前使用较多的快速聚类方法，该方法首先将样本进行粗略分类，然后再按照某种原则进行修正，直到分类结果比较合理为止。但这类方法的不足之处是首先需要确定分类的个数和选择聚类点，而这 2 个初始值的选择对于聚类结果的影响较大。传统最小最大聚类算法[5]可以解决初始聚点的选择问题，但仍需假设聚类数目 K，而根据经验简单定义 K 值的方法势必影响系统性能。研究发现，对最小最大聚类方法作进一步改进，可以有效解决初始聚点和 K 值的选择问题。

1　新闻文档结构

新闻文档的 3 个主要结构如图 1 所示。

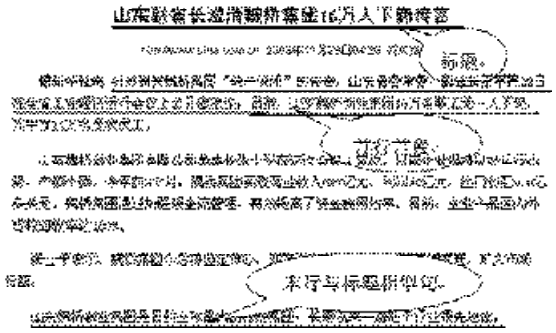

图 1　新闻文档的结构

Fig. 1　Structure of news report

1）标题。在自动文本摘要技术中，文档的标题被视为提示主题的关键特征。基于统计的自动文摘方法就十分重视标题中出现的有效词。

2）首段首句。一般，新闻的第一句话常以据……社报道、……电、……消息等开头，之后的语句则总括性介绍这篇新闻主要讲述的内容。因此，新闻的首段首句十分重要，具有很强的主题提示作用。很多文摘系统都将多文档集合中每篇文档的首句提取出来，形成文摘作为基准摘要（baseline），供各系统评估评测，如 DUC。

3）末段。在新闻报道中，正文末段往往会对整篇新闻的内容作总结或评价。所以，选择末段中句子作为文摘句将具有较强的概括性。

根据美国的 P.E.baxendale 调查结果显示，段落的主题在段落句首的概率为 85 %，在段落末句的概率为 7 %[6]。根据新闻文档在结构上的上述特点，可以抽取每篇新闻首段的首句以及末段中与新闻标题最相似的句子组合作为该篇新闻文摘句。

2　最小最大聚类原理

2.1　最小最大聚类原理简介

基于最小最大原则的聚类中心选择方法的基本原理是：假设要将样本分成 K 个类别，则先把相距最远（余弦相似度最小）的 2 个样本 x_{i_1}，x_{i_2} 作为前 2 个聚类中心（这里的样本就是文本单元，可以是 1 篇文档，1 个段落或是 1 个句子，称为文摘句）；其余的聚类中心的选取可以用递推式表达，也就是若已经选取了 m 个聚类中心（$m<K$），则第 m+1 个聚类中心的选取原则为

$$\min\left\{d\left(x_{i_{m+1}},x_{i_r}\right),r=1,2,\cdots,m\right\}=$$
$$\max\left\{\min\left[d\left(x_j,x_{i_r}\right),r=1,2,\cdots,m\right],j\neq i,\cdots,i_r\right\},\quad(1)$$

上式右端表示：先选取每个样本（除聚类中心点外）与前 m 个聚类中心的距离的最小值，然后在这些最小值中选取距离最大的那个值，最后把这个值对应的样本作为第 m+1 个聚类中心。

聚类中心的选择如图 2 所示。假设前 2 个聚类中心 x_1、x_2 已经确定，则第 3 个聚类中心应该根据如下规则确定：第 3 个聚类中心与前 2 个聚类中心的距离最小者（相似度最大）等于所有其它的点中与前 2 个点 x_1、x_2 的较小距离（相似度最大）中距离最大（相似度最小）的。所以在确定 x_1、x_2 后，下一个取出点是 x_3 而不可能是 x_4。

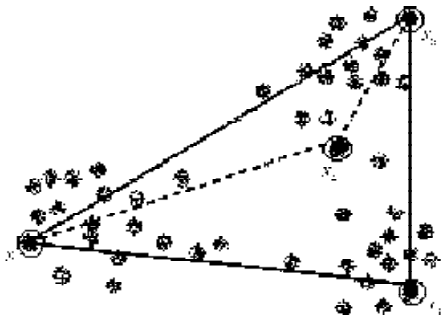

图 2　聚类中心的选择方法

Fig. 2　Selection method for clustering center

传统上该方法使用的前提条件是假设聚类数目 K 已知。但聚类本身是一个无指导的算法，在很多应用中，往往需要聚类算法发现聚类数目 K。因此，聚类算法中良好的收敛策略成为影响聚类结果的重要因素。K-means 算法通过设置 1 个最大迭代次数，或设置 1 个固定的阈值来达到收敛。但迭代过程中，各个类别的类间距离依据类别特征和数据分布而定，不同的类别其类内元素之间的距离不尽相同，而更新后的类中心值由划分后的类别元素值决定，所以固定阈值的收敛判断法容易引起类别间的不公平，影响迭代速度，严重的将造成错分、误分的情况。而使用适合各类别内数据分布的阈值从理论上能够减少这种不公平性。

2.2 改进的最小最大聚类算法（Improved Min -Max Clustering，IMMC）

IMMC引入平均自相似度概念来控制算法收敛，自动获得相似度值。

任意2个文本单元间的距离用“向量空间模型+余弦相似度”来计算。比如文本单元 i 的特征向量为 $\boldsymbol{x}_i=(a_1,a_2,\cdots,a_n)$，$j$ 的特征向量为 $\boldsymbol{x}_j=(b_1,b_2,\cdots,b_n)$，则2个文本单元的相似度为

$$sim(\boldsymbol{x}_i,\boldsymbol{x}_j)=\frac{\sum_{k=1}^{n}a_k b_k}{sqrt\left(\sum_{k=1}^{n}a_k^2\right)\times sqrt\left(\sum_{k=1}^{n}b_k^2\right)}, \quad (2)$$

式中：a_m、$b_m(1\le m\le n)$为文本单元对应的第 m 个特征对应的权值；

n 为2个文本单元的特征并集总数，特征为文本的关键词。

设聚类样本集合也就是文本单元集合：$X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$，$N$ 为样本个数。计算出所有样本间的相似度 $sim_{ij}=sim(\boldsymbol{x}_i,\boldsymbol{x}_j)$，当 $i=j$ 时 $sim_{ij}=1$。

定义全局平均相似度：

$$\bar{d}=\frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}sim_{ij}}{N(N-1)/2}, \quad (3)$$

定义最小最大平均相似度：

$$\bar{d}_{min\text{-}max}=\frac{\max_{i\neq j}\{sim_{ij}\}+\min_{i\neq j}\{sim_{ij}\}}{2}, \quad (4)$$

定义类别 ω_k 的平均自相似度：

$$\bar{d}_{kk}=\frac{d_{kk}}{|\omega_k|(|\omega_k|-1)/2}, \quad (5)$$

式中：$d_{kk}=\sum_{i\in\omega_k,j\in\omega_k,j\neq i}sim_{ij}$；$i$、$j$ 同为类 ω_k 中的样本，因此称 d_{kk} 为类 ω_k 自相似度。如果 $i\in\omega_{k_1}$，$j\in\omega_{k_2}$，则 $2d_{k_1k_2}=\sum_{i\in\omega_{k_1}}\sum_{j\in\omega_{k_2}}sim_{ij}$ 称为类 ω_{k_1}、ω_{k_2} 的互相似度。聚类的目的是尽量使类内的平均自相似度值大，使各类之间的互相似度减小。

定义全局平均自相似度门限：

$$\bar{d}_{Gsth}=\max\left(\bar{d},\bar{d}_{min\text{-}max}\right)。 \quad (6)$$

IMMC算法的收敛策略如下：

当选取了 K 个参照点时，各类的平均自相似度分别为：$\bar{d}_{11}$，$\bar{d}_{22}$，$\cdots$，$\bar{d}_{kk}$，增加1个参照点后，各类的平均自相似度为：$\bar{d}_{1}'$，$\bar{d}_{22}'$，$\cdots$，$\bar{d}_{kk}'$，$\bar{d}_{(k+1)(k+1)}'$。此时局部自适应的平均自相似度门限为：

$$\bar{d}_{gsth}=\frac{\bar{d}_{1}'+\bar{d}_{22}'+\cdots+\bar{d}_{kk}'+\bar{d}_{1}'+\bar{d}_{22}'-\cdots+\bar{d}_{kk}'}{2k}, \quad (7)$$

该门限值随每次聚类动态变化。

如果满足

$$\frac{\bar{d}_{11}'+\bar{d}_{22}'-\cdots+\bar{d}_{kk}'}{\left|\bar{d}_{1}'-\bar{d}_{gsth}\right|+\left|\bar{d}_{22}'-\bar{d}_{gsth}\right|+\cdots+\left|\bar{d}_{kk}'-\bar{d}_{gsth}\right|}\ge\frac{\bar{d}_{1}+\bar{d}_{22}+\cdots+\bar{d}_{kk}}{\left|\bar{d}_{1}-\bar{d}_{gsth}\right|+\left|\bar{d}_{22}-\bar{d}_{gsth}\right|+\cdots+\left|\bar{d}_{kk}-\bar{d}_{gsth}\right|}, \quad (8)$$

同时，$d_{(k+1)(k+1)}>d_{Gsth}$， (9)

则继续选取下一个参照点聚类，直到不满足上述条件为止。

IMMC的具体过程如下：

1）抽取相似度最小的2个样本 d_1、d_2 作为初始参照点，其它样本划分到与 d_1 和 d_2 相似度较大的类中，然后分别计 $\bar{d}_{11}$ 和 $\bar{d}_{22}$，此时聚类类别数 $K=2$。如果出现具有同样最小相似度的其它样本对，则同样重新计算出以此对样本为参照点后各类的平均自相似度 $\bar{d}_{11}'$ 和 $\bar{d}_{22}'$。（具体选取哪一对样本作为初始参照点，方法介绍如下：不计算互相似度，仅利用平均相似度进行初始中心的选取。比较 $\frac{\bar{d}_{11}+\bar{d}_{22}}{\left|\bar{d}_{1}-\bar{d}_{Gsth}\right|+\left|\bar{d}_{22}-\bar{d}_{Gsth}\right|}$ 与 $\frac{\bar{d}_{11}'+\bar{d}_{22}'}{\left|\bar{d}_{11}'-\bar{d}_{Gsth}\right|+\left|\bar{d}_{22}'-\bar{d}_{Gsth}\right|}$，选取比值较大的那对样本作为初始参照点；如果比值相等，则选择 $\bar{d}_{11}-\bar{d}_{22}$ 与 $\left|\bar{d}_{11}'-\bar{d}_{22}'\right|$ 较小的那对样本。此原则选取的样本对，一方面尽量保证聚类后各类的平均自相似度不能太小，另一方面也避免了选取的样本聚类后2个类的平均自相似度相差太大的情况。式中分母是平均自相似度与全局门限的距离和，值越小则表示越靠近全局门限 $\bar{d}_{Gsth}$。）

2）根据最小最大原则选取下一个参照点，聚类后再计算每个类别的平均自相似度，直到算法收敛，确定类别数 K。

3 基于NS-IMMC的主题来源生成方法

根据新闻文档的特殊结构（New Structure，NS）和内容特点，利用改进的最小最大聚类算法对新闻文档进行聚类，提出了一种基于NS-IMMC的主题来龙去脉生成方法。其流程如图3所示。

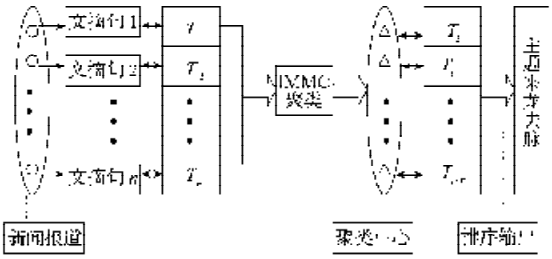

图 3　主题来龙去脉生成流程

Fig. 3　Flow of generating the cause-and-effect of news topic

NS-IMMC 算法的步骤如下：

1）新闻时间信息的获取。抽取每篇新闻报道的时间，如图 3 中的 $T_1,T_2,\cdots,T_n$，并保存到链表中。

聚类完毕后形成的 K 个聚类中心就是形成多文档摘要所需的文摘句，如把这些句子简单组合则会造成形成的文摘逻辑性不强，影响阅读。当今新闻要求及时性强，报道越及时，信息量越大。所以这些时间与新闻事件发生的时间非常相近，时间的变化随着事件的发展动态，准确地挖掘出新闻的时间信息，对于正确描述事件的性质有重要意义。抽取的时间信息如图 4 所示。

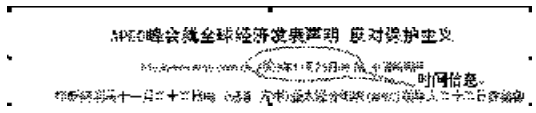

图 4　新闻中的时间信息

Fig. 4　Information of time in news

由于文摘句都是从新闻报道中抽取的，因此可以将新闻报道的时间与文摘句对应。文本摘要输出可根据其对应时间的先后顺序输出，这样可使文摘可读性强，并能准确反映主题的发展。为了比较时间的先后顺序，把时间转换成整数来比较大小。以“2006 年 04 月 20 日 10：25”为例，本章提取的时间信息精确到时，算法如下：

i) 抽取时间信息字符串，如“2006 年 04 月 20 日 10:25”所示；

ii) 查找字符“年”，抽取该字符左边部分字符串“2006”，将其转化为整型，用 $year$ 表示；

iii) 查找字符“月”，抽取该字符左边部分字符串“04”，将其转化为整型，用 $month$ 表示；

iv) 查找字符“日”，抽取该字符左边部分字符串“20”，将其转化为整型，用 day 表示；

v) 查找字符‘：’，抽取该字符左边部分字符串“10”，将其转化为整型，用 $hour$ 表示；

vi) 令 $time=year\times 365\times 24+month\times 30\times 24+day\times 24+hour$。

由算法可以看出，时间越短，报道越早；时间越长，报道越晚。

2）新闻预处理。

i) 扫描该主题下的所有新闻文档，先对其进行分词并去停用词；再根据信息增益（IG）法提取特征词，将其保存到 1 个词表中，称为特征词表；

ii) 提取命名实体，配以加权系数，即 $\alpha:\beta=3.5:1$；

iii)计算特征词的权值。算法如下：

$$P\_c=1.0*(N/N\_ft); \quad (10)$$

$$P\_ft=\log(P\_c); \quad (11)$$

$$P\_n\_c=1.0*N\_W\_ft; \quad (12)$$

$$P\_n\_ft=P\_n\_c*P\_ft; \quad (13)$$

$$weight=P\_n\_ft*value; \quad (14)$$

其中：N 为总的新闻文档个数；

ft 为特征项；

N_ft 为整个主题中含有 ft 的新闻文档个数；

N_W_ft 为 ft 在整个主题中出现的总次数；

$weight$ 为特征项 ft 的权值；

$value$ 是加权系数，当 ft 是命名实体时，$value$ 等于 3.5，否则 $value$ 等于 1。

3）文摘句的生成。将新闻报导的首段首句抽取出来；按句号拆分末段，将所有句子计算其与新闻标题的相似度；选择相似度最大的句子与首段首句结合作为该篇新闻的文摘句。相似度算法如下：

i) 将标题与特征词表匹配生成标题向量，记为 $str\boldsymbol{T}$；将末段的每个句子分别与特征词表匹配生成句子向量，记为 $str\boldsymbol{S}$。

ii) 将 $str\boldsymbol{T}$ 与 $str\boldsymbol{S}$ 的特征相对比，相同的特征权值做内积，再比较 $str\boldsymbol{T}$ 与 $str\boldsymbol{S}$ 所有特征的权值模。例如：

$str\boldsymbol{T}\langle(t_1,w_1),(t_2,w_2),\cdots,(t_n,w_n)\rangle$ 有 n 个词 t，每个的权值为 w；

$str\boldsymbol{S}\langle(s_1,u_1),(s_2,u_2),\cdots,(s_k,u_k)\rangle$ 有 k 个词 s，每个的权值为 u；

$str\boldsymbol{S}\langle(s_1,u_1),(s_2,u_2),\cdots,(s_k,u_k)\rangle k$，相同的词有 $t_1=s_2$，$t_5=s_1,\cdots$，其实共同的词的权重是相同的，也就是 $w_1=u_2$，$w_5=u_1$，则：

$$sim=(w_1\cdot u_2+w_5\cdot u_1+\cdots+w_n\cdot u_k)/(sqrt(w_1\cdot w_1+\cdots+w_n\cdot w_n)\cdot sqrt(u_1\cdot u_1+\cdots+u_k\cdot u_k)). \quad (15)$$

iii) 统计末段中所有句子与标题的相似度，取相似度最大的那句，令为 $str\boldsymbol{S}_{\max}$。

iv) 将首段首句和 $str\boldsymbol{S}_{\max}$ 相加作为该篇新闻报导的文摘句。

4）将所有新闻文摘句用 IMMC 进行聚类，迭代收敛后保存在聚类中心，如图 3 中 $T_i,T_{i+1},\cdots,T_{i+k}$ 所示。

5）将成为聚类中心的文摘句按其对应的时间从小到大输出形成多文档摘要，作为主题的来龙去脉。

4 实验结果与性能分析

验证实验是将本文方法与2种通用方法进行比较。方法1，提取每个新闻文档的首句来组成多文档摘要的方法[7]（简称Baseline）；方法2，基于*K*-means的多文档摘要生成方法[8]（简称*K*-means）；方法3，本文提出的方法（简称NS-IMMC）。

实验1：将3种方法对同一主题的新闻文档生成的摘要作对比，从直观上说明摘要生成的好坏；主题名称是“伊朗核问题”，抽取其中的30篇新闻报道来生成摘要。从生成的摘要来看，Baseline生成的摘要只是将各篇新闻的首段首句简单拼接，顺序颠倒，缺乏逻辑性；*K*-means方法生成的摘要，文本单元是新闻首段，冗余太多，主题不突出；本文生成的摘要内容简洁，逻辑性强且主题突出。

实验2：将3种方法对3个主题的新闻文档生成的摘要分别在主题覆盖性、主题表达充分性和生成时间上作比较。（主题覆盖性是指文摘内容不集中偏颇于某个主题，而是能够覆盖各个局部子主题。主题表达充分性是指提取的主题代表句在对该主题的内容表达上是否充实、完整、鲜明。）在测试语料中选取的3个主题名称分别是“以色列大选”、“斐济发生政变”、“格鲁吉亚间谍风波”。由于多文档摘要没有参考语料，实验2由8名具备该方面专业知识的评判员对这3种方法生成的摘要予与打分，分值范围为（0.1~1），结果详见表1。

表1 对多文档打分的平均分值

Tab. 1 Average evaluating value of multi-document

算法	主题覆盖性	主题表达充分性	摘要平均耗时/s
Baseline	0.476	0.552	13.08
K-means	0.560	0.433	96.78
NS-IMMC	0.685	0.724	114.31

表1表明：NS-IMMC在主题覆盖性上比Baseline高20.9 %，比*K*-means高12.5 %；在主题表达充分性上比Baseline高17.2 %，比*K*-means高29.1 %。这是由于NS-IMMC分析了新闻文档的结构特点，并用IMMC算法解决了初始聚类中心的选择问题和最终聚类中心数目的确定问题。在生成时间上，NS-IMMC耗时稍多，这是本文方法利用最小最大聚类的迭代次数多和需要计算与新闻标题的相似度所造成的。但基于生成摘要的质量考虑，为了使文摘句更具代表性，阐述更充分，牺牲一点时间也是允许的。

5 结语

针对新闻主题包含文档数目多、冗余信息大给人们造成的阅读困难，根据新闻文档的结构和内容特点，提出了一种基于NS-IMMC的主题来龙去脉生成方法。该方法抽取每篇新闻的首段首句和末段中与标题最相似的一句作为文摘句，有效地概括了该新闻的主题内容；然后利用改进的最小最大聚类原则对所有文摘句进行聚类，解决了聚类初始值的选定和值的确定问题，有效地去除了大量重复新闻的冗余信息；最后将文摘句按照时间的先后顺序输出，使形成的多文档摘要具有良好的逻辑性。实验结果表明：与2种通用的方法相比，新方法生成的摘要可理解性更强，内容更全面，更好地反映了主题的演变。

参考文献：

[1] Wan Xiaojun，Yang Jianwu. Design and Application of an On-line News Topic Detection System[J]. Journal of South China University of Technology：Natural Science Edition，2004(32)：42-46.

[2] Xu Yongdong，Xu Zhiming，Wang Xiaolong. Multi-Document Automatic Summarization Technique Based on Information Fusio[J]. Chinese Journal of Computers，2007，30(11)，2048-2054.

[3] Chin YewLin，Eduard Hovy. From Single to Multi-Document Summarization： A Prototype System and its Evaluation[C]// In Proceeding of the 40th Anniversary Meeting of the Association for Computational Linguistics(ACL-02). Philadelphia：[s.n.]，2002：25-34.

[4] Sun Jigui，Liu Jie，Zhao Lianyu. Clustering Algorithms Research[J]. Journal of Software，2008，19(1)：48-61.

[5] Hu Bai. Research and Development of Phrase-Representation Summarization Method for Chinese[D]. Shanghai：Shanghai Jiao Tong University，2007.

[6] Qian Aibing. Design and Implementation of Focused Web News Aggregator Based on RSS[J]. New Technology of Library and Information Service，2007(4)：56-61.

[7] Goldstein，MittalV，Carbonell，et al. Multi-Document Summarization by Sentence Extraction[C]//In proceedings of ANLP/NAACL 2000 Workshop on Automatic Summarization. Morristown：[s.n.]，2000：40-48.

[8] Hu Po，He Tingting，Ji Donghong. A Study of Chinese Text Summarization Based on Thematic Area Discovery [J]. Computer Science，2005，32(1)：177-181.

（责任编辑：罗立宇）