文章编号: 1003-0077(2007)03-0069-07

基于小世界模型的中文文本主题分析

石 晶[1,2,3],胡 明[3],戴国忠[1]

(1. 中国科学院 软件研究所 人机交互技术与智能信息处理实验室,北京 100080;
2. 中国科学院 研究生院 北京 100049; 3. 长春工业大学,吉林 长春 130021)

摘 要: 本文旨在研究如何基于小世界模型进行文本分割,确定片段主题,进而总结全文的中心主题,使文本的主题脉络呈现出来。为此首先证明由文本形成的词汇共现图呈现短路径,高聚集度的特性,说明小世界结构存在于文本中;然后依据小世界结构将词汇共现图划分为"簇",通过计算"簇"在文本中所占的密度比重识别片段边界,使"簇"与片段对应起来;最后利用短路径,高聚集度的特性提取图"簇"的主题词,采取背景词汇聚类及主题词联想的方式将主题词扩充到待分析文本之外,尝试挖掘隐藏于字词表面之下的文本内涵。虽然国际上已有很多关于小世界结构及基于其上的应用研究,但利用小世界特性进行主题分析还是一个崭新的课题。实验表明,本文所给方法的结果明显好于其他方法,说明可以为下一步文本推理的工作提供有价值的预处理。

关键词: 计算机应用;中文信息处理;主题分析;小世界模型;文本分割;词汇聚类
中图分类号: TP391 **文献标识码**: A

Topic Analysis of Chinese Text Based on Small World Model

SHI Jing[1,2,3], HU Ming[3], DAI Guo-zhong[1]

(1. Computer Human Interaction and Intelligent Information Processing Laboratory Institute of Software, The Chinese Academy of Sciences, Beijing 100080, China;
2. Graduate University of Chinese Academy of Sciences, Beijing 100049, China;
3. Changchun University of Technology, Changchun, Jilin 130021, China)

Abstract: The paper aims to perform topic spotting of segments based on text segmentation using small world structure. Main topic of the whole text is generalized and the skeleton of text shows itself. It is explained that the term co-occurrence graph of text is highly clustered and has short path length, which proves that texts have small world structure. Clusters in the small world structure are detected. The density of each cluster is computed to discover the boundary of a segment. Topic words are extracted from clusters of the graph. Words which are not distinctly in the analyzed text can be included to express the topics with the help of word clustering of background and topic words association. The signification behind the words are attempted to dig out. Although much research on applications of small world structure, it is a new task to analyze texts with the characteristics of small world. The experiments tell that the result of tests is far better than that of other methods. Valuable pre-processing is provided for next work of text reasoning.

Key words: computer application; Chinese information processing; topic analysis; small world model (SWD); text segmentation; words clustering

文本的主题分析旨在确定一个文本的主题结构,即识别所讨论的主题,界定主题的外延,跟踪主题的转换,觉察主题间的关系等,分析结果对于信息提取、文摘自动生成、文本分类等领域都有极为重要

收稿日期: 2006-11-23 **定稿日期**: 2007-02-05
基金项目: 国家 973 重点基础研究发展规划资助项目(2002CB312103);国家自然科学基金资助项目(60503054);中国科学院软件研究所创新工程重大项目资助
作者简介: 石晶(1970—),女,博士,主要研究方向为自然语言处理。

的价值。主题分析的程度随着应用对象的不同有所区别,浅层次的分析仅仅确定主题边界(文本分割)[1,2],或者进而指明不同片段间的关系(是否讨论同一主题)[3];比较复杂的分析能够在识别边界的基础上讨论主题的内容[4]。作为文本推理的预处理,本文研究如何利用小世界模型特性识别文本的片段边界,并抽取片段主题及全文的中心主题。

常用的主题分析策略是选择合适的概率模型,利用统计的方法实现边界计算及主题提取。文献[4]以不附加任何统计假设的有限混合模型(Finite Mixture Model)代表文本中的词汇分布,直接利用期望极大算法对其进行训练。PLSA (Probabilistic Latent Semantic Analysis)[5]和LDA (Latent Dirichlet Allocation)[6,7]是另外两种可选的,也是目前较常用的主题模型。统计方法的最大弊端是需要大量的背景语料,对于应用来说,有时并不方便除了采用统计的方法,主题提取还可以基于其他策略,比如词汇链[8],但无法与文本分割集成在统一框架之下。

与上述策略完全不同,本文基于文本的小世界特性:①将文本表示为词汇共现图,通过聚类形成多个"簇";②利用密度公式计算不同的"簇"所对应的文本片段,识别片段边界;③提取片段主题词,并通过背景语料库的词汇聚类产生联想;④从联想后的片段主题词中提取全文中心主题词。实验表明以该方法分析文本的主题脉络,其结果基本符合人的直觉判断,且优于其他模型及方法。

本文的结构安排如下,第一节介绍小世界模型;第二节解释词汇聚类;第三节详述主题分析的方法;第四节给出测试手段及实验结果,并就实验结果进行讨论;最后总结全文。

1 小世界模型

小世界拓扑结构常见于生物、社会以及人造系统中。对于自然语言网络的小世界特性研究基本集中在印欧语系语言,比如英语[9,10],而汉语与这类语言有极大的差别。第四节的实验表明汉语文本同样存在小世界现象。

1.1 小世界现象

社会学家Milgram在1976年发现,任意一对美国人之间,大都可以找到不多于六个两两相识的人将他们联系起来,这就是著名的"六度分离(Six-Degree Separation)"。D.J. Watts[11]对这一网络特性进行了深入的研究,于1998年提出"小世界模型",该模型在某种程度上同时实现了短路径和高聚集度两种特性。

1.2 词汇共现图

通过文本建立词汇共现图的方法如下:

(1) 对文本进行预处理,包括分词,删除虚词及无意义的实词,忽略标题,图表及文章结构等。

(2) 选择出现频率 $f > f_{thr}$ (f_{thr}为一常数)的 n 个词汇作为节点。

(3) 针对每一对词汇 w_i, w_j 计算Jaccard系数 J_{w_i,w_j},如果 $J_{w_i,w_j} > J_{thr}$ (J_{thr}为一常数)则在 w_i, w_j 之间加边。Jaccard系数的计算公式为: $J_{w_i,w_j} = \frac{n_{w_i} + n_{w_j}}{n_{w_i,w_j}}$,其中 n_{w_i} 代表出现词汇 w_i 的句子数目,n_{w_j} 代表出现词汇 w_j 的句子数目,n_{w_i,w_j} 代表同时出现词汇 w_i, w_j 的句子数目。

1.3 小世界特性

为了将小世界特性形式化的表示出来,Watts[11]引入特征路径长度和聚集度两个变量。特征路径长度是指任意两个节点之间最短路径长度的平均值,聚集度是指一个随机抽取节点的两个邻接节点成为邻接节点的概率。具体计算方法如下:

假设 $\Omega_L = (W_L, E_L)$ 被定义为由文本抽取的词汇共现图,其中,$W_L = \{w_i\}, \{i = 1,2,\cdots,N_L\}$ 表示词汇集合,$E_L = \{\{w_i, w_j\}\}$ 表示词汇之间的边的集合。$\xi_{ij} = \{w_i, w_j\}, \{i,j = 1,2,..,N_L\}$ 表示词汇 w_i, w_j 之间是否存在边,若存在则 $\xi_{ij} = 1$,否则 $\xi_{ij} = 0$。以 $\Gamma_i = \{j \mid \xi_{ij} = 1\}$ 表示词汇 $w_i \in W_L$ 的邻接节点集合,该词的聚集度定义为 $C_v(i) = \frac{\phi_i}{\binom{|\Gamma_i|}{2}}$,其中 $\phi_i = \sum_{j=1}^{N_L} \xi_{ij} \left[\sum_{k \in \Gamma_i, j<k} \xi_{jk} \right]$,于是图 Ω_L 的聚集度为: $C = \frac{1}{N_L} \sum_{i=1}^{N_L} C_v(i)$。

给出两个词 $w_i, w_j \in W_L$,令 $d_{\min}(i,j)$ 表示两词间的最短路径长度,则一个词的平均路径长度定义为 $d_v(i) = \frac{1}{N_L} \sum_{j=1}^{N_L} d_{\min}(i,j)$,于是图 Ω_L 的特征路径长度为: $d = \frac{1}{N_L} \sum_{i=1}^{N_L} d_v(i)$。

令 $\bar{k}$ 表示一个词的平均连结数,则随机图的聚

集度 $C^{rand} \approx \frac{\bar{k}}{N_L}$,特征路径长度 $d^{rand} \approx \frac{\ln(N_L)}{\ln(\bar{k})}$,若 Ω_L 具有小世界结构,则 $C \gg C^{rand}$, $d \approx d^{rand}$。

为了将小世界特性以一个直观的量表示出来,Walsh 定义 $\mu = (C/d)/(C^{rand}/d^{rand})$。具有小世界结构的图满足 $\mu \gg 1$。

2　词汇聚类

仅仅依赖所在文本的内部信息确定主题词,错误较多,如果能够借助背景库使主题词产生联想,必然有助于准确率的提高,为此需要利用丰富的背景库知识聚类词汇。本文以 1998 年《人民日报》手工标注的语料为背景库,以《知网》词典中的每一个词作为种子词,选择与之最相关的 n 个词形成一个聚类。对于每一个词汇 w,按下式计算该词汇对于种子词 s 的 δSC 值,根据 MDL 原则[4],δSC 值越大,说明 w 与 s 的相关性越大。

$$\delta SC = H\left(\frac{m^+}{m}\right) - \frac{m_s}{m} H\left(\frac{m^+_s}{m_s}\right) - \frac{m_{\neg s}}{m} H\left(\frac{m^+_{\neg s}}{m_{\neg s}}\right) - \frac{1}{2m}\log\left(\frac{m_s m_{\neg s}\pi}{2m}\right)$$

其中,$H(z) = -z\log(z) - (1-z)\log(1-z)$,$0 < z < 1$,当 $z = 0$ 或 $z = 1$ 时,$H(z) = 0$;m^+ 表示出现 w 的文本数;m_s 表示出现 s 的文本数;m^+_s 表示 w,s 共现的文本数;$m_{\neg s}$ 表示不出现 s 的文本数;$m^+_{\neg s}$ 表示出现 w 但不出现 s 的文本数;m 表示总的文本数。

本文采用的聚类方法除了考虑种子词与其他词在不同文本的共现,还进一步考虑二者在同一文本的共现强弱,以 $rel(w,s) = freq(w,s)/k$ 计算,$freq(w,s)$ 表示在 s 出现的文本中,w 出现的频数,k 是语料库文本的平均词数。$rel(w,s)$ 直接体现出某一文本中 w 与 s 的相关程度。随着 δSC 的值逐渐减小,与种子词相关的词越来越难于选出,或者选择的词与种子词联系得越来越不紧密,此时以某一文本中与种子词频繁共现的词汇予以补充或替换,从而使聚类结果更令人满意。实验表明该方法的确比单独使用 MDL 的方法能够更好地吻合人的直觉,也有利于主题词的联想。

3　主题分析

3.1　文本分割

3.1.1　分割策略

因为小世界结构呈现短路径和高聚集度的特点,所以具有小世界结构的图本身隐藏着潜在的“簇”或者“块”,对于文本来说,这是作者围绕中心主题有次序、有条理地展开讨论各子主题的必然结果。利用小世界特性进行分割的方法是: ①将词汇共现图按照不同的主题划分成“簇”; ②识别片段边界,使“簇”与文本片段对应起来。

3.1.2　“簇”的形成

如果仅仅依照小世界特性对词汇共现图进行词簇聚类(Clustering)[12],则很难实现文本分割,因为对于文本,词汇位置决定其所属片段,或者说第一片段的词汇必然先于第二片段的词汇出现,但在图中这种位置信息被忽略掉了。为了进行弥补,词汇共现图的节点按词汇第一次出现的先后顺序被统一标号。假设一文本具有 N 个词汇并规定分割成 k 个片段,所形成的词汇共现图为 Ω_L,则图“簇”的具体形成算法如下:

(1) 分割: 若节点 n_i 与 n_j 连结(n_i 和 n_j 是节点的标号),且 $n_j - n_i > d$(d 为一常数),则将 n_i 与 n_j 的连线断开。遍历图中所有节点重复此步骤形成 k' 个“簇”,如果 $k' \leq k$,则算法结束,否则进入第(2)步。

(2) 归并: 如果 $k' > k$,说明图“簇”的个数多于规定的片段数目,应该按照如下 3 个原则进行归并。

a. 设某“簇”为 $n_1, n_2, \cdots, n_l$ ($n_1 < n_2 < \cdots < n_l$),其相邻“簇”为 $n_i, n_{i+1}, \cdots, n_{i+j}$ ($n_i < n_{i+1} < \cdots < n_{i+j}$),若 $n_1 < n_i$, $n_{i+j} < n_l$,则将其归并为“簇” $n_1, \cdots, n_i, \cdots, n_{i+j}, \cdots, n_l$ ($n_1 < \cdots < n_i < \cdots < n_{i+j} < \cdots < n_l$);若 $n_1 < n_i < n_l$, $n_{i+j} > n_l$,则将其归并为“簇” $n_1, \cdots, n_i, \cdots, n_l, \cdots, n_{i+j}$ ($n_1 < \cdots < n_i < \cdots < n_l < \cdots < n_{i+j}$)。遍历图中所有节点重复此步骤形成 k'' 个“簇”,如果 $k'' \leq k$,则算法结束,否则进入 b 步。

b. 设某“簇”为 P_1,其相邻的两个簇为 P_0, P_2,若在图 Ω_L 中,“簇” P_0P_1 和 P_1P_2 分别是相通的,且 $C_{P_0P_1} > C_{P_1P_2}$, $C_{P_0P_1}$ 是 P_0P_1 的聚集度, $C_{P_1P_2}$ 是 P_1P_2 的聚集度,则将 P_0, P_1 归并为一个簇。循环执行此步骤,直到 k 个“簇”出现。如果依然得不到 k 个“簇”,则进入 c 步。

c. 设某“簇”为 P_1,其词数为 L_1,其相邻的两个簇为 P_0, P_2,所包括的词数分别为 L_0, L_2 ($L_0 < L_2$),若 $L_1 < L_{thr}$,则将 P 归并入 P_0,其中 L_{thr} 为不断增长的变量。重复此步骤直到形成 k 个“簇”。

词汇共现图经过分割、归并后,会有 k 个“簇” $P_1, P_2, \cdots, P_k$ 出现,依照其先后顺序分别对应文本中的 k 个片段 $S_1, S_2, \cdots, S_k$。

3.1.3 边界识别

若图"簇"由词汇 $w_1, w_2, \cdots, w_n$ 组成,片段 $g_k g_j$(g_k, g_j 为片段首句、尾句)的"簇"密度计算公式为: $desity = \frac{\sum_{i}^{n} freq(w_i) \times freq(w_i)}{\sum_{i}^{n} freq(w_i)}$,其中 $freq(w_i)$ 是词汇 w_i 在片段 $g_k g_j$ 中的出现频率。假设已经得到的片段为 $g_1 g_2, \cdots, g_{n-1} g_n$,对于下一个"簇",以 g_n 为首句,计算 $g_n g_{n+1}$ 的"簇"密度,每次增加一个句子,直到"簇"密度不再变化,则 g_{n+1} 为所求的片段尾句。

3.2 片段主题提取

3.2.1 提取方法

定义一: 如果图 Ω_L 是非连通图,则将其路径长度扩展定义如下:

$$dis = \begin{cases} d_{\min}(i,j), & \text{如果 } i,j \text{ 间有路径} \\ num, & \text{如果 } i,j \text{ 间没有路径} \end{cases}$$

其中, $d_{\min}(i,j)$ 是连通图节点 i 与节点 j 之间的最短路径, num 是图中不相通部分的数目。

定义二: 扩展的图的特征路径长度 L 被定义为所有点对之间的扩展路径长度的平均值。

定义三: 删除节点 v,但保留节点 v 与其他节点的连结,所有点对之间的扩展特征路径长度被定义为 L_v';删除节点 v 及其所有连结,剩余点对之间的扩展特征路径长度被定义为 L_{G_v}'。

于是节点 v 的重要程度[13]通过公式 $CB_v = L_{G_v}' - L_v$ 加以计算。当一个节点拥有较大 CB_v 值,说明缺失该节点,将会使原图变得很松散,这样的节点既便其对应词汇在文本中出现次数很少,但由于对文本的上下文连接起着至关重要的作用,因而可以被看作是作者讨论的主题。

3.2.2 主题词联想

在词汇聚类表中选择种子词是主题词的聚类,并使主题词根据背景词汇聚类产生联想。联想包括归一,合并,替换三个过程。

归一: 令两个聚类分别为 $(s: w_1, w_2, \cdots, w_n)$, $(s': w_1', w_2', \cdots, w_m')$,其中 s, s 是种子词, $w_i (1 \le i \le n)$, $w_j' (1 \le j \le m)$ 是两个聚类中的非种子词,若 $s = w_j'$, $s' = w_i$, $(1 \le i \le n, 1 \le j \le m)$,且至少 s, s' 之一为主题词,则将主题词扩充为 $s - s', s, s$ 被称为主题词元素,处理后的主题词形成归一主题词表。

合并: 若两个主题词含有公共元素,则将这两个主题词合并为一个主题词。即对于主题词 $A - s - B$, $A' - s - B'$,合并二者为 $A - s - B - A' - B'$,其中 A, B, A', B' 可能由多个主题词元素构成。该过程遍历归一主题词表的所有主题词,直至没有重复的主题词元素,处理后的主题词表被称为合并主题词表。

替换: 若合并主题词表有主题词 $s - s'$,而原主题词表中有 s 或 s',将其替换为 $s - s'$,循环此过程直到合并主题词表中的所有主题词均得以替换,原主题词表中的其他主题词保持不变,形成替换主题词表。

最后删除替换主题词表中重复的主题词,形成新的主题词表,该主题词表即为经过背景词汇聚类联想后的主题词表。

3.3 中心主题获取

根据联想后的片段主题词计算全文的中心主题词。假设文本有 n 个片段,其中 s 作为不同片段的主题词元素出现 m 次,则 $P(s) = m/n$,取 $P(s) > \gamma$ 的主题词元素作为中心主题词, γ 为一小于 1 的常数。

4 实验设计及结果对比

本节首先证明小世界结构存在于文本中,然后阐述背景库词汇聚类的细节。对于主题结构,以文本分割和主题提取单独进行测试,以使结果客观。

4.1 小世界性证明

该实验选择 1998 年《人民日报》手工标注的 1 000 个文本进行验证,每个文本的平均词数大约为 250。实验中取 f_{thr} 为 2, J_{thr} 的值设置为保证每个词的平均连结数为 6。实验结果如表 1:

表 1 小世界特性

	C	C^{rand}	d	d^{rand}	μ
Max	0.697 058	0.053 087	2.475 824	2.307 162	32.023 328
Avg	0.475 489	0.035 672	2.222 582	2.567 897	15.400 429
Min	0.194 443	0.045 511	2.120 055	2.308 279	4.651 746

表中数据为 μ 的最大值、最小值和平均值,可见 $C \gg C^{rand}$, $d \approx d^{rand}$,同时 $\mu \gg 1$,说明汉语文本同样具有小世界模型的特性。

4.2 词汇聚类

本实验选择《知网》词典中的词汇作为种子词

s,以 1998 年《人民日报》手工标注的语料库及文本分类的部分语料库为背景知识(共 3 763 个文本),当 $P(w|s) > P(w)$ 时,取 7 个 $\delta SC > \gamma, \gamma = 0.005$ 的词汇(按 δSC 值从大到小的顺序)和 3 个 $rel(w,s) > \gamma', \gamma' = 0.002\ 5$ 的词汇(按 $rel(w,s)$ 值从大到小的顺序),构成同一个聚类。舍弃独词(只包括种子词)聚类,形成词汇聚类表。共有 7 128 个聚类出现。

4.3　文本分割的测试

4.3.1　测试集

本实验利用 1997 年 3 月《人民日报》手工标注的语料库构建 4 个测试集 $T_{3\text{-}11}$, $T_{3\text{-}5}$, $T_{6\text{-}8}$, $T_{9\text{-}11}$, $T_{x\text{-}y}$ 表示所含主题片段的句数在 x 和 y 之间。每一个测试集包括若干伪文本,即由不同类的文本连接而成的形式上的文本,要求相邻段落务必来自不同的类。其所含的主题数平均为 7,具体如表 2。

表 2　实验中的测试集

	$T_{3\text{-}11}$	$T_{3\text{-}5}$	$T_{6\text{-}8}$	$T_{9\text{-}11}$
主题片段的句数	3-11	3-5	6-8	9-11
伪文本数	109	127	115	98

4.3.2　度量标准

为了便于同类算法的对比,文本分割采用两种度量标准 P_k[14] 和 WindowDiff[15]。

$$P_k = P(seg)\,P(miss) + (1 - P(seg))\,P(false\ alarm)$$

$P(seg)$ 是算法分割一个片段的先验概率,本实验取 $P(seg) = 0.5$, $P(miss)$ 是算法分割结果缺少一个片段的概率, $P(false\ alarm)$ 是算法分割结果添加一个片段的概率。

$$WindowDiff(ref, hyp) = \frac{1}{N - k}\sum_{i=1}^{N-k}(|\,b(ref_i, ref_{i+k}) - b(hyp_i, hyp_{i+k})\,| > 0)$$

$b(i,j)$ 表示整句 s_i 和整句 s_j 间的边界数量, N 表示文本中的整句数量, k 取真实片段平均长度的一半, ref 代表真实分割, hyp 代表算法分割。

4.3.3　实验结果

作为与本文方法的对比,取 PLSA[16], LSA[17], Dynamic Programming[1] 三种算法在 $T_{3\text{-}11}$, $T_{3\text{-}5}$, $T_{6\text{-}8}$, $T_{9\text{-}11}$ 上进行测试,其中,本文算法取测试集测试结果的平均值。详见表 3(本文算法的结果亦给出 WindowDiff 的值)。

表 3　与 PLSA, LSA 以及动态规划的对比结果

	$T_{3\text{-}11}$ (%)	$T_{3\text{-}5}$ (%)	$T_{6\text{-}8}$ (%)	$T_{9\text{-}11}$ (%)
本文算法	6.47 (19.26)	7.64 (15.38)	9.13 (20.00)	6.38 (14.11)
PLSA 方法	16.79	13.81	13.26	11.94
LSA 方法	13.12	15.21	10.02	12.17
动态规划算法	17.38	14.09	19.40	11.31

注: 括号内为本文算法的 WindowDiff 值

可见,本文方法的错误率远远低于其他同类算法。不仅如此,利用该方法进行文本分割,可以得到边界识别完全正确的最佳结果,这是其他算法不易实现的。本文作者曾对基于 PLSA 模型的文本分割进行仔细研究[18],发现基于 PLSA 模型的分割,其结果的随机性较大,随迭代次数及主题数目的变化难于确定。

4.4　主题提取的测试

4.4.1　测试语料

本测试采用的是文本分类语料库,共包括环境、经济、艺术、教育、体育、计算机、医学、政治、交通、军事等十大类。测试语料库中的文本没有分词,所以首先利用中国科学院计算技术研究所的分词系统 ICTCLAS 对其进行处理,然后凭直觉给每个类以一定数目的标识词(不多于 5 个),如表 4:

表 4　类及其标识词

类	标　识　词
环境	环境 动物 土壤 植被
经济	经济 金融 财政 商品 贸易
军事	战 军 弹 航空
计算机	电脑 计算机 微机
交通	事故 车 路 站
教育	教育 思想 校 学
体育	赛 训练
艺术	文艺 拍摄 出版 剧院
医药	病 伤 药 饮食
政治	外交 会 访问 联合国 和平

4.4.2　度量标准

若从某类文本提取的主题词包含该类的标识词,即认为提取结果正确。准确率定义为: $precision = \frac{n_{correct}}{n_{total}}$,其中, $n_{correct}$ 指正确提取主题词的

文本数，n_{total}指测试文本的总数。

4.4.3 片段主题提取

以类为单位进行测试，每个类取大约100个主题片段，其测试集合如下表：

表5 测试集及所包含的片段数目

类	片段数目	类	片段数目
环境	101	教育	100
经济	113	体育	105
军事	97	艺术	114
计算机	95	医药	106
交通	102	政治	105

本文方法将与TF-IDF[19]及Z-SCORE[19]方法进行对比。TF-IDF的计算方法为：

$$weight_w(s) = \frac{tf_w(s) \times \log(\frac{N}{n_w})}{\sqrt{\sum_i^n (tf_w(s))^2 \times \log^2(\frac{N}{n_w})}}$$

其中，$tf_w(s)$表示词汇w在测试片段s中的出现频数，N为背景语料中所有的片段数目，n_w是背景语料含有w的片段数目。Z-SCORE的计算方法为：

$$weight_w(s) = \frac{tf_w(s) - \frac{\sum_i^N f_i(w)}{N}}{\sqrt{\sum_i^N \left(f_i(w) - \frac{\sum_i^N f_i(w)}{N}\right)^2 \times \frac{f_i(w)}{N}}}$$

其中，$f_i(w)$是词汇w在背景语料第i个片段中的出现频数。实验结果如表6。

表6 片段主题的提取结果

类	本文方法(%)	TF-IDF (%)	Z-score (%)
环境	96.86	64.81	31.48
经济	98.09	50.00	34.62
军事	87.09	63.54	29.17
计算机	100.00	68.00	47.91
交通	98.44	96.92	81.54
教育	100.00	88.24	56.86
体育	93.49	72.31	89.23
艺术	99.15	70.37	16.67
医药	98.36	78.69	55.74
政治	98.14	76.36	49.09

可见，本文方法的结果远远好于其他两种方法，主要原因在于充分利用背景语料库的知识，使主题词产生联想，以此挖掘出隐藏于文本之中的内涵。

4.4.4 中心主题获取

令$P(s)$为片段主题词元素的出现频率，取$P(s) > \gamma$的片段主题词元素为中心主题词，随着γ的提高，准确率不断降低，其结果如下表：

表7 γ的取值及相应的中心主题提取的准确率

γ	Precision (%)	γ	Precision (%)
0.5	99.12	0.6	96.35
0.7	70.49	0.8	54.54

5 结语

本文利用小世界特性描述文本，结合背景知识解析其主题结构——文本分割之上提取片段主题词并总结全文的中心主题词。为了提高主题词提取的准确性，本文以词汇聚类的方式使主题词产生联想，将主题词扩充到待分析文本之外，尝试挖掘隐藏于字词表面之后的文本内涵。虽然国际上已有很多关于小世界结构及基于其上的应用研究，但利用小世界特性进行主题分析还是一个崭新的课题。本文的实验结果表明，该方法有很好的分析表现，能够为下一步文本推理的研究提供坚实的基础。

参考文献：

[1] Ath. Kehagias, A. Nicolaou, P. Fragkou and V. Petridis. Text Segmentation by Product Partition Models and Dynamic Programming [J]. Mathematical and Computer Modeling, 2004. 39: 209-217.

[2] Gina-Anne Levow. Prosody-based topic segmentation for mandarin broadcast news [A]. In: Proceedings of HLT-NAACL [C]. 2004.

[3] Ferret, Olivier. Using collocations for topic segmentation and link detection [A]. In: Proceedings of COLING [C]. Taipei. 260-266.

[4] Hang Li and Kenji Yamanishi. Topic Analysis Using a Finite Mixture Model [J]. Information Processing & Management, 2003, 39(4): 521-541.

[5] Brants, T.; Chen, F. R.; Farahat, A. O. Arabic document topic analysis [A]. In: LREC-2002 Workshop on Arabic Language Resources and Evaluation [C]. Las Palmas; Spain, 2002.

[6] D.M. Blei, A. Y. Ng, and M. I. Jordan. Latent

Dirichlet allocation [J]. Journal of Machine Learning Research, 2003(3):993-1022.

[7] Steyvers, M. & Griffiths, T. Probabilistic topic models. In: T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), handbook of Latent Semantic Analysis [M]. Hillsdale, NJ: Erlbaum. 2007.

[8] 索红光, 刘玉树, 曹淑英. 一种基于词汇链的关键词抽取方法[J]. 中文信息学报, 2006, 20(6): 27-32.

[9] Ferrer-i-Cancho, R. and Sole, R. V. The small world of human language [A]. In: Proceedings of the Royal Society of London. Series B, Biological Sciences [C]. 2001. 268(1482): 2261-2265.

[10] Matsuo, Y.; Ohsawa, Y.; and Ishizuka, M. A document as a small world [A]. In: Proceedings the 5th World Multi-Conference on Systemics, Cybenetics and Infomatics [C]. 2001, 8: 410-414.

[11] D. Watts and S. Strogatz. Collective dynamics of small-world networks [J]. Nature, 1998, 393: 440-442.

[12] Yutaka Matsuo: Clustering using Small World Structure [A]. In: Proc. 6th Int'l Conf. on Knowledge-based Intelligent Information Engineering Systems & Applied Technologies (KES2002) [C]. IOS Press/ Ohmsha (ISSN: 0922-6389), Crema, Italy, 2002, 1252-1256.

[13] Yutaka Matsuo, Yukio Ohsawa and Mitsuru Ishizuka: KeyWorld: Extracting Keywords in a Document as a Small World [A]. In: DS-2001 [C]. 2001, 271-281.

[14] D. Beeferman, A. Berger, J. Lafferty. Statistical Models for Text Segmentation [J]. In: Machine Learning, 1999, 34, 1-34.

[15] L. Pevzner and M. Hearst. A critique and improvement of an evaluation metric for text segmentation [J]. Computational Linguistics. 2002, 28(1): 19-36.

[16] Thorsten Brants, Francine Chen, Ioannis Tsochantaridis. Topic-based document segmentation with probabilistic latent semantic analysis [A]. In: Proceedings of the eleventh international Conference on Information and knowledge management [C]. McLean, Virginia, USA. 2002. 211-218.

[17] F. Y. Y. Choi, P. Wiemer-Hastings, and J. Moore. Latent semantic analysis for text segmentation [A]. In: Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing [C]. 2001. 109～117.

[18] 石晶, 戴国忠. 基于 PLSA 模型的文本分割[J]. 计算机研究与发展, 2007, 44(2): 242-248.

[19] Liu Y, Ciliax BJ, Borges K, Dasigi V, Ram A, Navathe SB, Dingledine R. Comparison of two schemes for automatic keyword extraction from MEDLINE for functional gene clustering [A]. In: Procceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB'04) [C]. 2004. 394-404.