一种面向研讨环境的摘要生成方法

王 艾 李耀东
(中国科学院自动化研究所复杂系统与智能科学实验室 北京 100190)

摘 要 针对当前研讨厅中对专家大段发言缺乏归纳、概括方法的问题,提出了一种面向研讨环境的摘要生成方法,该方法采用概率混合模型抽取专家发言的话题集,对相邻话题的变化情况进行判断,不仅可以发现话题(Topic)的转变趋势,还可根据话题的演化生成相应的摘要,提供给专家。这些自动生成的摘要既有助于增强专家之间的良性互动、激发专家思维,也可用于决策方案和会议总结的辅助生成。实验结果表明,提出的面向研讨环境的摘要生成方法合理、有效。

关键词 综合集成研讨厅,概率混合模型,群体交互,摘要生成
中图法分类号 TP18 **文献标识码** A

Probabilistic Mixture Model Based Summarization Approach for CWME Discussions

WANG Ai LI Yao-dong
(Laboratory of Complex Systems and Intelligence Science, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

Abstract To solve the problem of the lack of summarization method of experts' speeches in CWME, a new method was proposed. This method first adopts a probabilistic mixture model to discover latent themes from experts' speeches. Then, it judges whether themes have changed. Once they are changed, this method will generate summary and provide it for experts. The summaries can help experts to inspire each other in CWME. Experimental result shows that the proposed method is feasible and effective.

Keywords Cyberspace for workshop of metasynthetic engineering(CWME), Probabilistic mixture model, Collective interaction, Text summarization

1 引言

1990年钱学森等人正式提出开放的复杂巨系统及其方法论,用来描述、处理自然界和人类社会中复杂的巨系统[1]。1992年钱学森与其合作者进一步提出了人机结合、从定性到定量的综合集成研讨厅(HWME)体系[2]。此后,随着Internet和网络的迅速普及,"Cyberspace(电子空间和数字空间)"成为一个重要的概念,它使参与者跨越时间和地域的限制,随时随地就所关心的问题进行研究、交流和探讨,并可随时利用网络上的大量资源。在Cyberspace中构建综合集成研讨厅,即基于Cyberspace的综合集成研讨厅(Cyberspace for Workshop of Metasynthetic Engineering,简记为CWME)[3]成为一个新的发展方向。通过多年的探索与实践,戴汝为、李耀东等人已经成功建立了几个典型的CWME系统[4,5]。

综合集成研讨厅的构思是把人集成于系统之中,采用人-机结合、以人为主的技术路线,充分发挥人的作用,使研讨的集体在讨论问题时互相启发,互相激活,使集体创见远远胜过一个人的智慧[7]。其中在线组织研讨过程是问题分析、求解的核心,通过在线的组织化研讨,专家之间交换求解问题所需的专业知识,阐述自己的观点以及意见。由此呈现出大量的专家定性知识,如何在研讨厅中对其进行分析和归纳,从而清晰地展现研讨话题的演变脉络,进而为参与研讨的专家和决策者提供相应的信息支持,对解决重大决策问题有着积极的促进意义。

对于这一问题,综合来看,目前的工作主要集中在对专家个体意见与群体意见差异性的评估与比较上。比如:文献[9]中提出在综合集成研讨环境中将专家对多个方案评价意见的多维数据组进行降维,并将降维后的结果在低维数据空间中进行可视化表示,通过这种对专家评价意见在低维空间内的可视化表示,可直观地观测到专家群体意见的聚类情况;文献[10,11]提出群体一致性算法,当专家群体通过研讨形成了一些定性的方案之后,用该方法使专家的思维逐渐趋于收敛,最终达到群体意见一致。这些方法对于汇总专家意见,促进群体智慧都发挥了有益的作用,但是尚未对研讨中专家的发言文本进行分析归纳,从而使得大量的专家定性知识无法得到充分利用。因此,本文提出一种面向研讨环境的摘要生成方法。该方法对专家发言文本进行分析和归纳,首先采用概率混合模型抽取专家发言的话题集;然后判断专家发言话题是否发生转变,当专家发言的话题转变时,就生成相应的摘要,提供给专家。专家可由此了解话题演变的清晰脉络,对话题的走向予以更有效的把握,有望促进研讨质量和问题求解效率的提高。

到稿日期:2010-03-17 返修日期:2010-06-23 本文受973国家重点基础研究发展计划(2007CB311007)资助。

王 艾(1980—),女,博士生,主要研究方向为模式识别、系统复杂性,E-mail:ai.wang@ia.ac.cn;李耀东(1977—),男,博士,副研究员,主要研究方向为系统复杂性、智能信息处理。

2 面向研讨环境的摘要生成方法

本方法的主要目的是根据研讨文本发现研讨厅中话题的演变趋势，并将其以摘要的形式呈现给各位专家。这一方法基于语言模型，主要由3个部分组成，分别是话题提取、话题判别以及摘要生成，如图1所示。

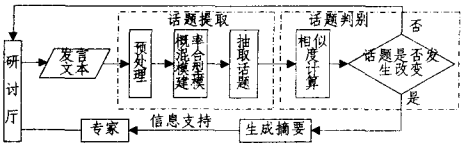

图1 面向研讨环境的摘要生成方法的框图

2.1 基于概率混合模型的话题提取

根据专家在研讨中的发言顺序，获得了一个有时间索引的文本集合 $C=\{d_1,d_2,\cdots,d_T\}$，话题提取的主要目的是从文本集合 C 中发现专家发言的话题演变趋势，包括下面两个步骤：

(1)将文本集合 C 分成 n 个子集合，按照固定的时间间隔分割为 $C=C_1\cup C_2\cup\cdots\cup C_n$ 且 $C_i\cap C_j=\varnothing, i,j=1,2,\cdots,n, i\neq j$，其中子集合 $C_i=\{d_{t_i},\cdots,d_{t_i+l_i-1}\}$ 包含 l_i 个文本。

(2)本文对已有的简单概率混合模型[16,17]进行改进，利用新的概率混合模型，从每个子集合中抽取出显著的话题集。这里的话题指的是词条的概率分布，它能描述语义的一致性(即话题)。通常每个话题都是通过一元语言模型(特定于某个话题的)表示的，即词条的概率分布 $\{p(w|z)\}_{w\in V}$ 同时满足 $\sum_{w\in V}p(w|z)=1$，下面详细说明这个概率混合模型。

简单概率混合模型是由 k 个一元话题语言模型($z_1,z_2,\cdots,z_k$)(针对子集合 C_i)以及一个背景模型 B(针对整个集合 C)构成，文本 d 中的每个词条都是由简单混合模型生成的。

$$p_d(w)=\lambda_B p(w|B)+(1-\lambda_B)\sum_z[p(z|d)p(w|z)] \quad (1)$$

式中，λ_B 为 B 的混合权重。由于 B 会以较高的概率生成一些非功能词条(没有任何语义信息的词条，例如："是"，"而且"等)，从而话题模型 z_j 就会以较高的概率生成具有丰富语义信息的词条，因此背景模型 B 的使用可以使得话题模型更加有判别力。

但是通过式(1)所估计的参数，在接下来的自动摘要生成中不能得到利用。因此，需要一个模型，既能从每个子集合 C_i 中抽取出话题来，还能将所估计的参数适用于摘要生成。采用贝叶斯规则对式(1)重写后，文本 d 中的每个词条通过下述改进概率混合模型生成：

$$p(w,d)=\lambda_B p(d|B)p(w|B)+(1-\lambda_B)\sum_z[p(z)p(d|z)p(w|z)] \quad (2)$$

在接下来的摘要生成所采用的句子排序打分机制中，$p(z)$ 和 $p(d|z)$ 将会被用到。

采用整个集合 C 来估计背景模型 B：

$$p(w|B)=\frac{\sum_{i=1}^{T}c(w,d_i)}{\sum_{w\in V}\sum_{i=1}^{T}c(w,d_i)} \quad (3)$$

$$p(d|B)=\frac{\sum_{w\in V}c(w,d)}{\sum_{i=1}^{T}\sum_{w\in V}c(w,d_i)} \quad (4)$$

式中，$c(w,d_i)$ 是文本 d_i 中词条 w 的计数。另外还需要估计的参数为 $\Lambda=\{p(z),p(d|z),p(w|z)\,|\,d\in C_i,1\leqslant z\leqslant k\}$。集合 C_i 的对数似然函数为：

$$\log p(C_i|\Lambda)=\sum_{d\in C_i}\sum_{w\in V}[c(w,d)\log(\lambda_B p(d|B)p(w|B)+(1-\lambda_B)\sum_z p(z)p(d|z)p(w|z))] \quad (5)$$

式(5)中，采用 EM(Expectation Maximization)算法来估计这些参数 Λ，使集合 C_i 的对数似然函数的期望 $E(\log p(C_i|\Lambda))$ 最大化：

$$E(\log p(C_i|\Lambda))=\sum_{d\in C_i}\sum_{w\in V}c(w,d)[p(B|d,w)\log(\lambda_B p(d|B)p(w|B))+\sum_z(1-p(B|d,w))p(z|d,w)\log((1-\lambda_B)p(z)p(d|z)p(w|z))] \quad (6)$$

式中，$p(z|d,w)$ 表示在背景模型 B 不生成词条 w 的假设下，话题 z 生成文本 d 中的词条 w 的概率。

通过贝叶斯理论推算 $p(z|d,w)$ 的过程为 E 步骤。E 步骤推导公式如下：

$$p(B|d,w)=\frac{\lambda_B p(d|B)p(w|B)}{\lambda_B p(d|B)p(w|B)+(1-\lambda_B)\sum_z p^{(n)}(z)p^{(n)}(d|z)p^{(n)}(w|z)} \quad (7)$$

$$p(z|d,w)=\frac{p^{(n)}(z)p^{(n)}(d|z)p^{(n)}(w|z)}{\sum_z p^{(n)}(z)p^{(n)}(d|z)p^{(n)}(w|z)} \quad (8)$$

根据 E 步骤的计算结果来估计参数为 M 步骤。在 M 步骤中，$E(\log p(C_i|\Lambda))$ 分别对 $p(z)$，$p(d|z)$ 以及 $p(w|z)$ 求偏导，使其为0，并满足 $\sum_z p(z)=1$，$\sum_d p(d|z)=1$ 和 $\sum_w p(w|z)=1$，通过拉格朗日乘法可得：

$$p^{(n+1)}(z)=\frac{\sum_{d,w}c(w,d)(1-p(B|d,w))p(z|d,w)}{\sum_{d,w}c(w,d)(1-p(B|d,w))} \quad (9)$$

$$p^{(n+1)}(d|z)=\frac{\sum_{w}c(w,d)(1-p(B|d,w))p(z|d,w)}{\sum_{d',w}c(w,d')(1-p(B|d',w))p(z|d',w)} \quad (10)$$

$$p^{(n+1)}(w|z)=\frac{\sum_{d}c(w,d)(1-p(B|d,w)p(z|d,w)}{\sum_{d,w'}c(w',d)(1-p(B|d,w')p(z|d,w')} \quad (11)$$

基于概率混合模型的话题提取算法的实现过程如图2所示。

1. 初始化：给参数 $\Lambda=\{p(z),p(d|z),p(w|z)\,|\,d\in C_i,1\leqslant z\leqslant k\}$ 赋初值。

(1)$p(z)=1/k$，其中 $1\leqslant z\leqslant k$。

(2)$p(d|z)=random$，其中 $d\in C_i,1\leqslant z\leqslant k$，并对其归一化满足 $\sum_d p(d|z)=1$。

(3)$p(w|z)=random$，其中 $w\in V,1\leqslant z\leqslant k$，并对其归一化满足 $\sum_w p(w|\theta_j)=1$。

2. Estep

根据式(7)计算 $p(B|d,w)$。

根据式(8)计算 $p(z|d,w)$。

3. Mstep

根据式(9)计算 $p(z)$。

根据式(10)计算 $p(d|z)$。

根据式(11)计算 $p(w|z)$。

4. 判断集合 C_i 的对数似然函数是否收敛，如果收敛条件不满足，则回到步骤2，继续迭代，否则输出估计的参数集 Λ。

图2 基于概率混合模型的话题提取算法实现过程

2.2 话题判别

从每个子集合抽取话题集以后，就需要判断相邻的子集

合之间的话题是否发生了变化。本文采用相对熵(Kullback-Leibler divergence)[18]来度量话题的演化距离(即判断相邻的子集合的话题是否发生了转变)。假设 z_i 为子集合 C_i 的某个话题,z_{i+1}为子集合 C_{i+1}的某个话题,如果这两个一元话题语言模型 z_i 和 z_{i+1}十分接近,就认为这两个话题之间的演变距离小。由于相对熵 $D(z_{i+1}|z_i)$可以对 z_{i+1}相对于 z_i 的一些额外信息建模,本文认为选择相对熵作为话题之间演变距离的度量标准是可行的。话题之间演变距离计算公式如下:

$$D(z_{i+1}|z_i)=\sum_{n=1}^{|V|}p(w_n|z_{i+1})\log\frac{p(w_n|z_{i+1})}{p(w_n|z_i)} \tag{12}$$

由于相对熵是非对称的,使用 $D(z_{i+1}|z_i)$比使用 $D(z_i|z_{i+1})$更有意义。如果 $D(z_{i+1}|z_i)$大于某个阈值 ξ,就判断两个相邻的子集合的话题发生了演化。

2.3 自动摘要生成

在研讨环境中,一旦相邻的发言子集合之间的话题发生了转移(相对熵超过了阈值),就需要对后一个子集合抽取摘要,提供给参与研讨的各个专家新的信息支持。第 2.1 节中介绍的概率混合模型可对文本中含有的丰富语义信息建模,在采用该模型抽取演变的话题集后,为了充分发挥该模型的这一优势,我们采用一种基于概率混合模型的语句打分机制,以此为基础计算句子的权重,从发言文本中抽取重要的句子。

一般情况下,文本都包含几个话题。因此人工书写的文摘应涵盖全部话题,以便读者对原文内容的把握。理想情况下,自动文摘技术生成的摘要应能涵盖更广的话题范围。在此基础上,我们采用概率混合模型对需生成摘要的文本建模。理论上,该方法可以有效地抽取话题集,还可根据语句所属的话题来对其分类以及排序,从不同的话题中挑选语句,使生成的摘要覆盖多个话题。

PLSI(Probabilistic Latent Semantic Indexing)是自动文本标引的新方法,由于它是基于最大似然原则的生成模型,因此它有很坚实的统计基础,生成的摘要属于通用型,可以涵盖多个话题[12,13],对数似然函数是:

$$L=\sum_d\sum_w n(d,w)\log\sum_z p(z)p(d|z)p(w|z) \tag{13}$$

在文献[12]中,对句子排序利用了多个话题的语义信息融合,给句子打分的函数是:

$$R=\sum_z p(z)p(d|z) \tag{14}$$

式中,$p(d|z)$表示为话题 z 中语句 d 的得分。

文献[12]的结果表明,这种句子打分的机制,可以将影响范围跨越几个话题的语句挑选出来,生成的摘要通用性更好;同时,文献[12]的实验结果表明该方法是有效的,性能优于 LSA[14]和 LexPageRank[15]。

基于此,为了将影响范围跨越几个话题的语句挑选出来,生成通用性更好的摘要,我们也采用式(14)作为打分方法。本文与 PLSI 的打分方法的区别是:PLSI 中参数 $p(z)$和 $p(d|z)$是通过最大化对数似然函数,即式(12)来确定的;本文的打分方法中的参数 $p(z)$和 $p(d|z)$是通过最大化集合 C_i 的对数似然函数,即式(6)来确定的。在式(6)中,由于背景模型 B 会滤去一些非功能词条,从而使得话题模型 z 会以较高的概率生成具有丰富语义信息的词条,去除更多干扰信息。因此话题模型 z 含有的语义信息相较 PLSI 的更加丰富,即 $p(z)$和 $p(d|z)$可以表示蕴含更多的话题信息。由此可知,尽管采用同一公式(14),本文方法所挑选的摘要语句的质量还是好于 PLSI 方法。

3 实验结果及其分析

下面以一个专家研讨过程为例来说明本文提出的方法。针对当前如何顺利度过金融风暴使中国经济复苏这一议题,参与研讨的专家从财政政策、货币政策、基础设施建设以及产业结构调整这 4 个方面发表自己的观点。其中,发言专家以博士研究生与硕士研究生为主,但发言内容是经济专业人士提供。

首先将专家的发言分成 4 个子集合,采用第 2.1 节中的概率混合模型来抽取每个子集合显著的话题。这里设置 $\lambda_B=0.7$,每个子集合的话题数 $k=4$,表 1 显示了从每个子集合的隐话题(hidden topic)下抽取的前 5 个高概率词条。分析表 1 可知,子集合 1 的话题 2 主要是描述出口退税,子集合 3 的话题 3 主要是描述医疗以及户籍制度的建立。

表 1 从专家发言中抽取出的话题集

子集合	话题 1	话题 2	话题 3	话题 4
C_1	改革 0.10877	调整 0.06506	出口 0.07371	经济 0.04205
	财政 0.04145	问题 0.06263	企业 0.05127	家庭 0.03772
	增值税 0.03395	出口 0.05767	产品 0.03469	差异 0.03759
	减税 0.03379	税 0.03604	地方 0.03273	所得税 0.02842
	政策 0.02848	退税率 0.01891	目的 0.02317	收入 0.02730
C_2	政策 0.09554	贷款 0.12600	经济 0.06347	金融 0.08131
	信贷 0.08405	货币 0.05740	存款 0.05505	降息 0.04745
	货币 0.06821	规模 0.04761	准备金 0.05505	改变 0.03583
	投放 0.02311	扩大 0.04728	商业 0.03690	预期 0.03583
	成本 0.02185	增长 0.04676	银行 0.03419	公司 0.03559
C_3	社会 0.04758	农村 0.04923	服务 0.07321	就业 0.13400
	需要 0.03970	设施 0.04095	医疗 0.04859	压力 0.02815
	发展 0.03969	铁路 0.04000	建立 0.03282	大学生 0.02711
	企业 0.03750	保障 0.032598	制度 0.02690	农民工 0.02637
	措施 0.03070	运力 0.024821	户籍 0.01651	问题 0.02503
C_4	发展 0.10508	结构 0.12095	国家 0.02792	农业 0.03888
	服务业 0.08954	产业 0.07379	第三产业 0.02247	劳动力 0.03634
	劳动力 0.07354	调整 0.05018	学者 0.02237	中国 0.03290
	就业 0.06989	失业 0.02949	规律 0.02222	增加值 0.02742
	中小企业 0.06625	技术 0.02697	失业率 0.01749	发展 0.02604

从每个子集合抽取出相应的话题集以后,首先用 $p(z)$来度量话题在各子集合 $C_i(i=1,2,3,4)$中的显著性,找到每个子集合最显著的话题后,再采用相对熵来判断相邻的子集合之间最显著的话题是否发生了变化。计算结果如表 2 所列,这里设置话题演变距离的阈值 $\xi=12$。从表 2 可清楚地发现相邻的子集合之间的话题发生了演变。

表 2 各子集合间最显著话题的相对熵

子集合	子集合	KL 熵
C_1 话题 1	C_2 话题 1	12.55
C_2 话题 1	C_3 话题 3	13.63
C_3 话题 3	C_4 话题 2	13.43

3.1 自动文摘性能评估

摘要生成的流程如下:一旦相邻的子集合之间的话题发生了转移,即相对熵超过了阈值,就需要对后一个子集合抽取摘要,并提供给参与研讨的各个专家对刚才讨论内容的一个回顾或信息支持。在已对发言内容建模的基础上,采用基于概率混合模型的打分机制对候选语句打分,生成相应的摘要。

为了验证本文提出的自动文摘方法的有效性,采用 PLSI[12]以及 LSA(Latent semantic analysis)[14]方法作为基准方法进行对比实验。文摘的评估手段可以归纳为两种:第一,内

部评测，就是测试文摘本身是否与文章要点一致，以及是否包含文章的基本要点；第二，外部评测，就是通过文摘方法对其他相关的任务或应用所产生的影响做出评价的方法。本文的评测方法属于前者。

本文对文摘结果使用召回率、准确率及 F-measure 3 个参数进行评估。召回率指系统对标准文摘的覆盖率，准确率指系统准确识别的比率，F-measure 是召回率和准确率的综合。设 Nh 为人工标注的标准文摘抽取的句子数目，Nm 为自动文摘系统抽取的句子数目，Nhm 为同时被自动文摘系统和标准文摘抽取的句子数目，具体公式为：召回率 $R=Nhm/Nh$，准确率 $P=Nhm/Nm$，$F\text{-}measure=2\times P\times R/(P+R)$。

本文从研讨语料中摘录了 10 段专家研讨发言，请 3 个相关专业的研究生对每段发言进行手工编制摘要，将其综合后得到标准摘要，摘要长度比例分别为每段发言的 10% 和 20%。然后我们计算出各自的平均召回率、平均准确率、平均 F-measure，如表 3 所列。

表 3　准确率、召回率、F-measure 的评测情况

文摘比例	文摘方法	P	R	F-meaasure
10%	LSA	0.583	0.567	0.573
	PLSI	0.583	0.567	0.573
	PMM	0.767	0.750	0.757
20%	LSA	0.521	0.494	0.504
	PLSI	0.568	0.542	0.552
	PMM	0.673	0.653	0.660

表 3 中，PMM 表示的是基于概率混合模型的自动摘要方法，由表 3 可见，以本文的方法在 10%，20% 的摘要比例下的召回率、准确率和 F-measure 值均高于 PLSI 和 LSA，说明本方法生成的文摘较好地兼顾了覆盖度和准确率，具有较高的质量。

3.2　自动文摘应用举例

下面举一个应用实例说明本文所提方法的有效性。首先摘录一部分专家参与研讨的发言，其中，T，C，G，L 代表不同的专家：

T：经过向全社会征求意见，医改的实施方案下一步将围绕基本医疗保障体系建设、健全基层医疗卫生服务、促进基本公共卫生服务逐步均等化、建立国家基本药物制度、推进公立医院改革试点等 5 个重点推进。其中，多方关注的两个重要难点需要有力的解决方案：一是公立医院保证公益性的成本谁来买单；二是尽快完善基本药物制度。

C：此外，需要进一步明确和细化的内容，还包括完善新型农村合作医疗制度，建立稳定可靠、合理增长的筹资机制；加快县乡村三级医疗卫生服务机构和城市社区卫生服务机构建设，实现基层医疗卫生服务网络的全面覆盖；制定国家基本公共卫生服务项目等。

G：按照中央要求，公安机关将积极推进户籍制度改革，推动建立城乡统一的户籍登记管理制度，创新流动人口服务和管理体制。

L：当前经济危机条件下，部分产业应从东部向中西部转移，形成与返乡农民相一致的转移路径，这样既能解决失业农民工再就业难题，也有助于中西部经济的崛起，和沿海经济的进一步升级。

T：改革开放以来，中国东部地区发展远远快于中西部地区，经济发展地区差异日益拉大。产业结构的调整有助于地区之间的平衡发展，非常符合经济发展趋势，应该去做，也必须去做。这个调整是有惰性的，在经济发展好的情况下，该调的时候未必及时调整。而经济危机有一个重要的作用就是对落后的、应该被淘汰的产业结构进行调整。一个地区能不能形成更高级的产业结构、更先进的经济发展模式，取决于它能不能将产业结构调整出去。

实验时，分别采用概率混合模型、PLSI 以及 LSA 对其建模，按照压缩率 10% 生成摘要。这里采用的 3 种自动文摘方法都是基于统计的摘录式摘要，对文本的语句分割不仅考虑了常规的语句标点符号（句号、问号、分号以及冒号等），还考虑了更小的分割符号——逗号。因此下述的发言文本共包含 34 个语句，每种方法生成包含 3 个语句的摘要。

采用 LSA 方法生成的摘要为：

① 医改的实施方案下一步将围绕基本医疗保障体系建设、健全基层医疗卫生服务、促进基本公共卫生服务逐步均等化、建立国家基本药物制度、推进公立医院改革试点等 5 个重点推进。

② 加快县乡村三级医疗卫生服务机构和城市社区卫生服务机构建设。

③ 一个地区能不能形成更高级的产业结构、更先进的经济发展模式。

采用 PLSI 方法生成的摘要为：

① 医改的实施方案下一步将围绕基本医疗保障体系建设、健全基层医疗卫生服务、促进基本公共卫生服务逐步均等化、建立国家基本药物制度、推进公立医院改革试点等 5 个重点推进。

② 加快县乡村三级医疗卫生服务机构和城市社区卫生服务机构建设。

③ 而经济危机有一个重要的作用就是对落后的、应该被淘汰的产业结构进行调整。

采用概率混合模型生成的摘要为：

① 加快县乡村三级医疗卫生服务机构和城市社区卫生服务机构建设。

② 而经济危机有一个重要的作用就是对落后的、应该被淘汰的产业结构进行调整。

③ 推动建立城乡统一的户籍登记管理制度。

比较这 3 种方法生成的摘要，可以发现基于概率混合模型生成的摘要质量要比基于 LSA 以及 PLSI 生成的摘要好。LSA 生成的前 2 个摘要语句和 PLSI 生成的前 2 个摘要语句是一模一样的。但是 LSA 生成的第 3 个摘要语句明显没有什么语义内容，相比之下 PLSI 生成的第 3 个摘要语句对专家而言就有更好的提示作用。概率混合模型生成的摘要有 2 个摘要语句是和 PLSI 生成的摘要语句相一致的，即语句①和语句②。但是 PLSI 生成的摘要语句①和②内容有些冗余，都是讲述医疗卫生改革的，而概率混合模型生成的摘要语句①谈论的是医疗卫生改革；摘要语句②谈论的是产业结构调整；摘要语句③是谈论户籍制度的，这是 LSA 以及 PLSI 生成的摘要都没有涉及到的内容。可以发现概率混合模型生成的摘要更加综合，涵盖了发言文本各话题下的内容，生成的摘要更具有通用性。

结束语　针对当前研讨厅中缺乏对专家发言文本进行归

（下转第 209 页）

关系图、层次交互图3种方式分别展示了论坛中的不同属性数据，并对结果进行了总结和分析，结果表明，其能够方便地进行海量论坛数据的层次多维可视化。

网络论坛数据的可视化为管理人们在论坛中的行为模式提供了一种直观的方法，对论坛的管理人员和其他分析人员有一定的指导意义。由于网络论坛数据的复杂性和大规模性，海量数据的可视化分析速度相对较慢，如何加快处理的速度是我们下一步的工作，以及进一步在网络论坛数据中挖掘出更多的聚类信息，如社会群体事件的信息。

参 考 文 献

[1] Smith M A, Fiore A T. Visualization Components for Persistent Conversations[C]// Proceedings of the SIGCHI conference on Human factors in computing systems. New York, NY, USA: ACM, 2001: 136-143

[2] Viégas F B, Smith M. Newsgroup, Crowds and AuthorLines: Visualizing the Activity of Individuals in Conversational Cyberspaces[C]// Proceedings of the 37th Hawaii International Conference. Waikoloa, Big Island, Hawaii, 2004

[3] Welser H, Gleave E, Fisher D, et al. Visualizing the Signatures of Social Roles in Online Discussion Groups [J]. Journal of Social Structure, 2007, 8(2): 1-13

[4] Fisher D, Smith M, Welser H T. You are who you talk to: Detecting roles in Usenet newsgroups [C]// Proceedings of the 39th Hawaii International Conference. Waikoloa, Big Island, Hawaii, 2006

[5] Paolillo J C. Visualizing Usenet: A Factor-Analytic Approach [C]// Proceedings of the 33rd Hawaii International Conference. Waikoloa, Big Island, Hawaii, 2000

[6] Heer J, Boyd D. Vizster: Visualizing Online Social Networks[C]// Proceedings of the 2005 IEEE Symposium on Information Visualization. Washington, DC, USA: IEEE Computer Society, 2005: 33-40

[7] Viegas F, Golder S, Donath J. Visualizing email content: Portraying relationships from conversational histories[C]// Proceedings of the SIGCHI conference on Human Factors in computing systems. New York, USA: ACM, 2006: 978-988

[8] Kerr, Bernard. Thread Arcs: An Email Thread Visualization[C]// Proceedings of the 2003 IEEE Symposium on Information Visualization. Washington, DC, USA: IEEE Computer Society, 2003: 211-218

[9] Rohrer R, Swing E. Web-based information visualization [J]. IEEE Computer Graphics and Applications, 1997, 7(8): 52-59

[10] Rohall S L, Gruen D, Moody P, et al. ReMail: A Reinvented Email Prototype[C]// Proceedings of ACM Human Factors in Computing Systems. 2004: 791-792

[11] Venolia G, Neustaedter C. Understanding Sequence and Reply Relationships within Email Conversations: A Mixed-Model Visualization[C]// Proceedings of ACM Human Factors in Computing Systems. 2003: 361-368

[12] 裴新. 网络中极大 k-plex 发现算法和网络社群简历挖掘研究[D]. 北京：北京邮电大学，2008

[13] 黄雄伟，陈定方，祖巧红. Web 数据挖掘可视化研究与应用[J]. 湖北工业大学学报，2009，24(4)：54-56

[14] 王柏，吴巍，徐超群，等. 复杂网络可视化研究综述[J]. 计算机科学，2007，34(4)：17-23

[15] 蔡磊，龚健雅. 分布式海量多源空间数据的组织与网络可视化[J]. 测绘信息与工程，2009，34(6)：28-30

[16] 杨育彬，李宁，张瑶. 基于社会网络可视化分析的数据挖掘(英文)[J]. 软件学报，2008，19(8)：1980-1994

(上接第 194 页)

纳与总结方法的问题，本文提出了一种面向研讨环境的摘要生成方法。该方法首先采用概率混合模型对专家发言进行分析，抽取专家发言的话题集，再判断专家发言话题是否发生转变，一旦专家发言的话题转变，就生成相应的摘要，提供给专家，专家可对其进行提炼以及升华，为最终形成有效的决策方案提供信息支持。实验结果表明，本文提出的面向研讨环境的摘要生成方法合理、有效。

参 考 文 献

[1] 钱学森，于景元，戴汝为. 一个科学新领域—开放的复杂巨系统及其方法论[J]. 自然杂志，1990，13(1)：3-10

[2] 王寿云，于景元，戴汝为，等. 开放的复杂巨系统[M]. 浙江：科学技术出版社，1996：278-282

[3] 戴汝为. 系统学与中医药创新发展[M]. 北京：科学出版社，2008

[4] 李耀东，崔霞，戴汝为. 综合集成研讨厅的理论框架、设计与实现[J]. 复杂系统与复杂性科学，2004，1(1)：27-32

[5] 赵明昌，李耀东. 一个新的综合集成研讨厅软件框架[J]. 计算机工程与应用，2008，44(11)：1-4

[6] 戴汝为. 人-机结合的智能工程系统—— 处理开放的复杂巨系统的可操作平台[J]. 模式识别与人工智能，2004，17(3)：257-261

[7] 戴汝为，李耀东. 基于综合集成的研讨厅体系与系统复杂性[J]. 复杂系统与复杂性科学，2004，1(4)：1-24

[8] 操龙兵，戴汝为. 综合集成与决策[J]. 计算机研究与发展，2003，40(4)：531-537

[9] 刘春梅，戴汝为. 综合集成研讨厅专家群体评估结果的可视化[J]. 模式识别与人工智能，2005，18(1)：6-11

[10] 王丹力，戴汝为. 群体一致性及其在研讨厅的应用[J]. 系统工程与电子技术，2001，23(7)：33-37

[11] 王丹力，戴汝为. 专家群体思维收敛的研究[J]. 管理科学学报，2002，5(2)：1-5

[12] Bhandari H, Ito T, Shimbo M, et al. Generic Text Summarization Using Probabilistic Latent Semantic Indexing [C]// Proceedings of IJCNLP. 2008

[13] Hofmann T. Probabilistic Latent Semantic Indexing [C]// Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval. 1999: 50-57

[14] Gong Y, Liu X. Generic text Summarization using relevance measure and latent semantic analysis [C]// Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. 2001: 19-25

[15] Erkan G, Radev D R. LexPageRank: Prestige in Multi-Document Text Summarization [C]// Proceedings of EMNLP. 2004

[16] Mei Qiao-zhu, Zhai Cheng-xiang. Discovering Evolutionary Theme Patterns from Text-An Exploration of Temporal Text Mining [C]// Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining

[17] Zhai Cheng-xiang, Velivelli A, Yu Bei. A Cross-Collection Mixture Model for Comparative Text Mining[C]// Proceedings of KDD '04

[18] Cover T M, Thomas J A. Elements of Information theory [M]. Wiley, 1991