文章编号: 1001- 2486(2010) 01- 0116- 06

新闻视频数据库基于故事单元的“多线程”管理技术研究*

文 军[1], 吴玲达[2], 曾 璞[2], 谢毓湘[2]
(1. 国防科技大学 理学院, 湖南 长沙 410073; 2. 国防科技大学 信息系统与管理学院, 湖南 长沙 410073)

摘 要: 新闻视频关于新闻事件的报道是一种“多线程”的形式, 针对这种特性提出了一种基于有向图理论的新闻视频数据库管理方法。研究了故事单元相似关系与图论知识之间的联系, 结合最小部分树理论提出了一种将故事单元之间复杂相似关系图简化为“多线程”结构树的新闻视频数据库管理技术。实验显示, 这种管理方法对于视频数据库的浏览、检索、摘要等实际需求具有重要的理论意义和应用价值。

关键词: 新闻视频; 数据库; 多线程; 故事单元; 图论
中图分类号: TN941. 1 文献标识码: A

Organizing News Video Database Based on Stories with “MultiThreads”

WEN Jun[1], WU Ling-da[2], ZENG Pu[2], Xie Yu-xiang[2]
(1. College of Science, National Univ. of Defense Technology, Changsha 410073, China;
2. College of Information System and Management, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: In order to incarnate “multithreading” of news event reporting, the paper studies the validity of graph theory for managing news video database. Firstly, news stories similarity and graph theory are studied for digesting complicated relations of all stories reporting the same event. Then, a method based on minimal spanning tree is presented for predigesting the similarity graph into a simple tree with “multithreads”. Experiments show that the method is a research issue with great theoretical and pratical significance for browsing, searching and abstracting in news video database.

Key words: news video; database; multithreads; story; graph theory

新闻视频基于故事单元的管理技术研究通常直接为视频浏览和摘要等应用提供基础。新闻视频的故事单元是一个新闻事件的完整报道, 能够提供事件的各种语义信息, 因此在故事单元层次针对语义事件开展新闻视频的组织和管理, 能够更好地贴近用户实际需求。因此基于故事单元的新闻视频数据库管理技术的基本思路是通过一定方法获取新闻视频数据库中报道某一新闻事件的所有故事单元, 通过对故事单元之间的相似度的计算获得故事单元之间的依存关系, 构建新闻事件的线程结构, 这种线程通常以单一的线性结构或者相对复杂树形结构来表现。

虽然报道相同事件的故事单元具有相似性, 但是不同故事单元对事件的报道可能具有不同的侧重点。图 1 给出了新华网关于“汶川大地震”这一新闻事件的报道页面, 从中可以看出, 对于相同事件的新闻报道按照报道重点的不同可以分成多种分组: “救灾进展”、“灾区防疫”、“志愿者行动”、“赈灾需求”等。在新闻视频的故事单元中也存在相同现象。针对这种情况, 将所有的故事单元按照时间顺序简单线性排列的线程结构难以体现这种关系, 因此需要设计一种可以体现新闻事件报道“多线程”关系的新闻视频数据库管理技术。

本文研究针对这一问题, 提出了一种基于有向图理论的新闻视频故事单元结构化管理技术, 通过有向图的各种操作来满足用户的不同要求。与当前研究的各项线程化管理技术相比, 本文方法具有良好的理论基础, 能够更好地体现“多线程”的管理机制, 更加方便用户对于新闻事件的了解以及后期对于新闻视频信息的各种深入应用。

* 收稿日期: 2009- 06- 25
基金项目: 国家自然科学基金资助项目(60802080); 国家 863 计划资助项目(2009AA01Z335)
作者简介: 文军(1976—), 男, 讲师, 博士。

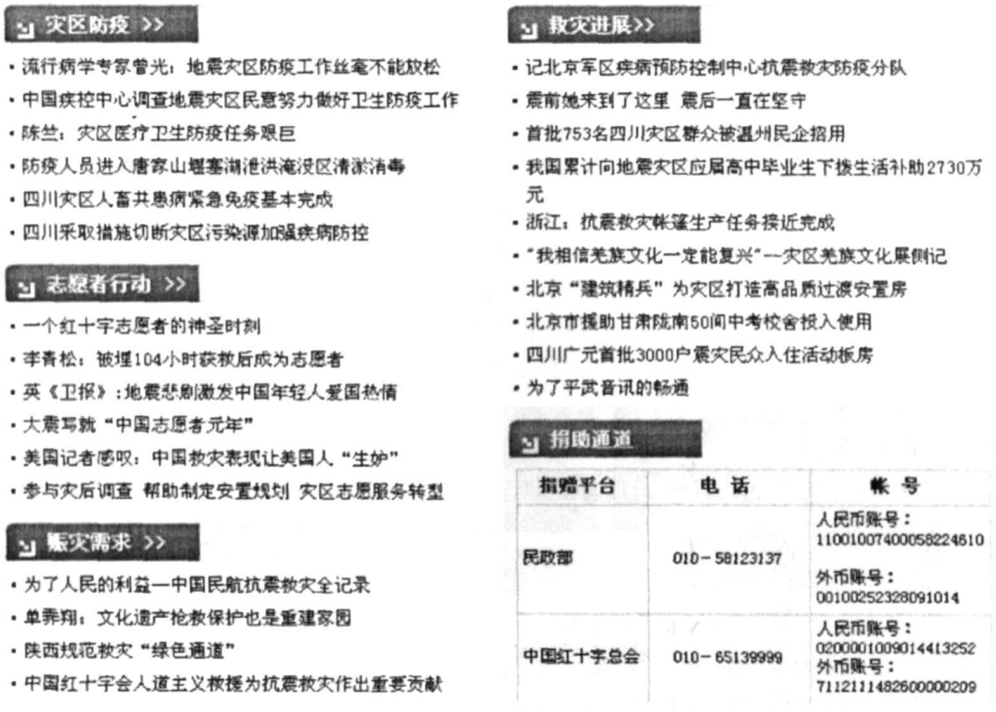

图 1　新闻事件报道的"多线程"结构示意图

Fig. 1　"Multithreads" of news reporting

1　相关研究

在早期的话题探测和跟踪(Topic detection and tracking, TDT)研究中, 往往只是将文档聚类到话题中, 没有充分研究话题内部文档之间的依存与相似关系。为给用户提供更加快速有效的信息, 部分研究提出了事件线程的概念[1], 并在文本 TDT 的研究中得到了关注, 但是对于新闻视频而言, 视频的高层语义获取比较困难, 因此新闻视频中事件探测与跟踪研究比较有限, 线程化组织技术研究也不多。

在文献[1]中, 结合 TDT 研究, 对文本媒体提出了故事、事件线程等定义, 通过事件模型获取事件丰富的结构信息和它们在话题中的依存关系。该研究提出了事件之间的依存关系对于用户完整地了解事件具有重要意义, 事件结构建模的方式对于获得故事的语义要比平面列表更加有效。在文献[2]中, 利用了 TDT 研究的成果, 对新闻视频的播报文本进行分析, 实现话题探测与跟踪等, 并将话题线程结构通过按时间排序的定向层次树来构建。

虽然在一些视频检索研究[3-5]中也提出了"线程"的概念, 但是基本都是比较简单的方法, 这种线程结构通常只是将视频的故事单元按照时间顺序或者相似度顺序简单地进行线性排列, 难以让用户对事件的发展和报道的不同重点有比较全面的了解。

日本国立情报研究所(National Institute of Informatics, NII)针对新闻视频的话题探测与跟踪、线程化组织等开展了一系列研究[2,6-10], 并通过这种结构化的结构来辅助用户的浏览和交互等。这些研究在完成话题探测与跟踪分析后, 获得所有故事单元之间的相似性和相互的依存关系, 构建每个故事单元相似关系的层次树, 并对树中的可能重复的子树进行删减、合并等一系列操作进行简化。然而, 这些研究初始的层次树结构复杂, 简化过程中对子树的各种操作处理也比较复杂。

Wu xiao 等在研究中[11-12]提出了一种融合文本和视觉语义概念的线程结构组织方法。该方法将故事单元之间的依存关系定义为: 新颖、重复、发展三种类型, 并通过一个二叉树来表示, 构建二叉树时考虑了两个方面的关系: 语义相似关系和时间相关性。二叉树的线程结构方便了用户的检索和视频摘要。然而研究中也面临一些问题: 完全利用二叉树来对视频数据库进行结构化的管理, 虽然结构简单, 能够方便检索等应用, 但正是因为二叉树结构简单, 在后续应用中对于充分体现故事单元之间的各种依存关

系, 满足用户部分特定的个性化需求存在不便。

2 新闻视频的故事单元相似关系与图论

本文研究在完成故事单元的关联分析和相似度计算[13]的基础上开展。故事单元的关联分析是指将新闻视频的故事单元按照所报道事件进行分组, 每个分组中的故事单元都报道相同的新闻事件; 故事单元相似度计算方法为数据库的线程化管理提供基础。

完成故事单元关联分析之后, 每个分组中的故事单元都报道相同的新闻事件。虽然在每个分组内部的故事单元都报道相同的新闻事件, 具有一定的相似性, 但是随着时间的变化和事件的发展, 以及故事单元报道的重点不同, 故事单元之间的相似程度会有比较明显的差异, 因此为体现这种差异性, 为"多线程"管理提供基础, 在开展线程化研究之前, 首先设置一个阈值, 相似度值大于该阈值的故事单元被看作是"相似的", 线程结构中主要体现这种相似关联, 而对于相似度值小于该阈值的故事单元之间的相似关系不在线程结构中进行体现。

例如: 报道一个新闻事件的故事单元集合中包含有五个故事单元$\{S_1, S_2, \cdots, S_5\}$, 两个故事单元之间的相似度具有对称性, 即 $Storysim(S_1, S_2) = Storysim(S_2, S_1)$, 因此上述集合中, 由任意两个故事单元组合所组成的"故事单元对"共有 10 种可能, 其中相似度大于设定阈值有 7 个, 分别为: $\{S_1, S_2\}$, $\{S_1, S_3\}$, $\{S_1, S_5\}$, $\{S_2, S_3\}$, $\{S_2, S_4\}$, $\{S_3, S_5\}$, $\{S_4, S_5\}$。而图是反映对象之间关系的一种工具, 由点及一些点之间关系的联线(称为边)所组成, 其中点表示研究对象, 边表示这些对象之间的联系, 可以用图来描述很多庞大复杂的系统, 解决研究和决策中的大量实际问题。因此相似关系也可以通过相似关系图表示, 如图 2 所示, 图中箭头表示时间先后顺序。

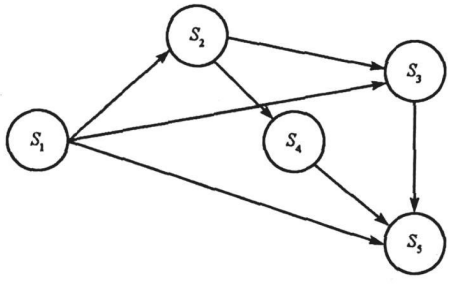

图 2 故事单元的相似关系图
Fig. 2 Similarity graph of news stories

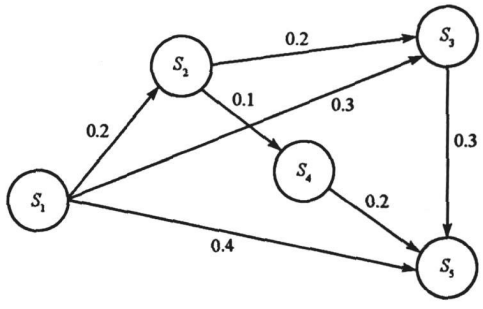

图 3 相似关系的有向赋权图表示
Fig. 3 Directed and weighted graph of similarity

显然, 这种相似关系图对于用户浏览和事件摘要分析而言过于复杂, 尤其是集合中包含的故事单元较多时; 因此这种直接生成的相似关系图不利于用户对于新闻事件信息的准确把握和分析。必须对上述相似关系图开展简化分析, 以获得简单有效的故事单元线程结构。

此外, 在相似关系图中, 对于顶点之间的弧还可以添加故事单元的相似度、时间距离等数量指标与这一相似关系相对应。因此, 可以将相似关系图进一步规范化为有向赋权图, 如图 3 所示, 将故事单元之间相似关系的距离信息作为相似关系图中边的权重, 即为 $D(S_i, S_j) = 1 - storysim(S_i, S_j)$。为方便表示, 图 3 中距离值通过四舍五入取小数点后一位。

图 3 中的相似关系图是一种有向赋权图。因此完全能够利用图论的相关知识和方法进行分析, 获得新闻视频故事单元简单有效的线程结构。树作为一种特殊的图, 其特殊属性对于复杂图的关系简化、图的最小路径计算等具有特殊的作用。其中与研究密切相关的概念是最小部分树(minimal spanning tree)。

在赋权图中, 其部分树各边权的总和称为部分树的权; 具有最小权的部分树, 称为最小部分树。

最小部分树与赋权图之间的联系对于研究具有重要的提示: 将复杂的相似关系图改进为相似关系最小部分树, 能够大大简化故事单元之间的相似关系, 得到比较简单的线程结构, 大大方便用户的浏览及摘要等应用。

3　新闻视频数据库故事单元基于有向图的“多线程”管理

如图 3 所示, 故事单元之间的相似关系图是一种有向赋权图。可以考虑作为权重的数据有时间跨度信息和相似度。计算所得的相似度信息主要用于判断新闻视频故事单元之间的相似程度, 通过相似度的数值来分析故事单元之间的相似关系是否在相似图中得以体现, 或者用于判断故事单元之间的重复程度, 为用户的摘要提供基础; 而时间信息主要反映故事单元报道内容的先后次序, 对于用户了解事件的发展信息同样有效, 因此对于故事单元线程化分析而言, 时间信息也具有重要的作用。

本文研究弧的权重以图 3 中相似度距离信息为例。首先研究考虑用户浏览的需求, 即生成的故事单元线程图中要包含集合或者某个时间段的子集中所有的故事单元。

对于上述相似关系赋权图, 如果不考虑边的方向, 则可以通过赋权图的最小部分树 Kruskal 顺序生枝法来求最小部分树。Kruskal 顺序生枝法就是首先将图中的所有边按权值从小到大进行排列, 在确保不出现回路的前提下, 将依次排列的边逐一绘出; 若在增加某条边时出现了回路, 则排除该边并继续寻找下一条边。

考虑到图中边的有向性, 对算法进行了一定的改进:

在进行排序时, 将图中的所有弧按权值从小到大进行排列, 权重相同的弧按照初始顶点的时间先后顺序进行排列; 在进行绘制时也是按照上述两种排序方式进行。

通过上述算法, 得到图 3 中相似关系图的最小部分树, 如图 4(a) 所示, 进一步将这种树形结构调整为图 4(b) 所示的布局。这是一种简单有效的“多线程”结构, 与相似关系图相比, 这种结构可以大大方便用户对于报道某一特定新闻事件全部故事单元的浏览和相似关系的了解, 方便用户清楚地了解故事单元之间的内在联系和发展历程。尤其在事件集合中包含的故事单元比较多时, 这种“多线程”管理方法的优势更加明显。

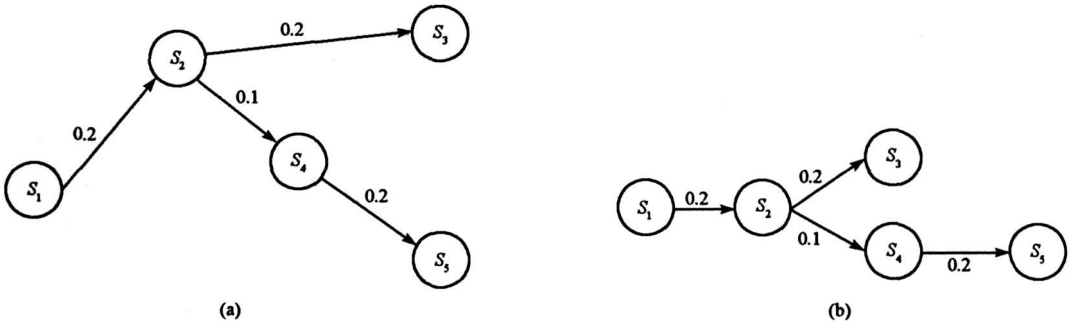

图 4　相似关系图的最小部分树

Fig. 4　Minimal spanning tree of similarity graph

通过本节基于有向图理论的“多线程”管理, 能够得到故事单元关于相似度、时间顺序等信息的线程结构。该结构作为一种有序的结构形式, 可以通过“多线程”体现每个事件集合内部故事单元之间的内在联系, 辅助用户对浏览、检索、摘要等的一系列应用。与当前研究中的其他组织形式相比, 这种组织形式可以提供更多的信息, 操作上更加方便, 可以更加广泛地满足用户的各种服务需求。

4　相关实验

本文实验从两个方面开展, 一方面在理论上比较本文方法与现有的线性结构管理方法、层次树管理方法之间的优缺点, 另一方面通过实验来对本文管理方法的用户满意度进行衡量。考虑不同用户对于故事单元之间的相似性、组织形式有不同的认识和要求, 具有较强的主观性, 难以通过一种标准的方法进行实验, 因此实验第二个方面主要以用户调查的方式来开展。

在实验第一个方面, 本文通过表 1 列出了各种不同管理方法之间的比较。

表1 各种管理方法之间的比较

Tab. 1 Comparing of different managing methods

	线性结构方法	层次树结构方法	本文方法
故事单元数量	n	n	n
初始结构的复杂度	简单	很复杂	较复杂
初始结构中相似关系数量	$n-1$	$\gg n$	$> n$
是否需要简化	否	是	是
简化算法理论基础		各种子树分析与处理	图论与最小部分树
简化算法难易		难	易
简化结构中相似关系数量		$n-1$	$n-1$
结构形式	单一线性线程	多线程	多线程
对事件发展的体现程度	一般	较好	较好

从表1中可以看出,线性结构在生成上最为简单,但是对于新闻事件发展变化的体现不如层次树结构和本文结构。虽然每种方法最后体现的相似关系数量都一样,但层次树结构由于在生成初始结构时需要为每个节点都生成完整的相似关系层次结构,因此具有大量的冗余信息,结构复杂,并且对于子树的分析和处理需要对大量冗余信息进行分析和删除、合并等各种处理,简化算法比较复杂。而本文方法初始结构复杂度虽然高于线性结构,但是较层次树生成方法的初始结构简单,并且简化过程具有图论知识的支持,简化算法简便。因此本文方法在"多线程"生成算法上具有更好的效率。

实验第二个方面的数据为动态采集的新闻视频数据库,调查重点关注3个事件:"伊朗核问题"(包含21个故事单元)、"联合国在黎巴嫩的维和行动"(包含15个故事单元)和"伊拉克爆炸事件"(包含9个故事单元),评估利用本文方式进行数据组织的用户主观满意度。实验选择9名没有经过任何训练的用户分成3组,分别对上述事件集合中的故事单元线程化组织形式进行主观评判。同时将本文的线程化组织方法与线性排列的组织方法进行了用户评判的对比:

让用户为两种不同的数据组织形式进行评估,用户对数据组织的满意度以10分制进行评估,分数越高表示用户满意度越高,每组的分数为该组成员的平均分数,评分只取整数,小数位四舍五入,实验结果如表2所示。

表2 新闻视频数据库管理方法的用户评价结果

Tab. 2 Experiment results of user evaluation

	第一组	第二组	第三组
线性结构	6	5	6
本文结构	9	8	8

由表2的结果可以看出,用户对于本文提供的数据组织形式要比时间线性结构的组织形式更加满意。这是因为本文提供的数据组织形式不仅体现了时间信息,也体现了故事单元之间各种内在联系等信息。而树形的线程结构也比单一维度的线性线程能够更好地体现故事单元之间的依存关系及发展变化的各个方面。与当前已存在的各种线程化方法相比,本文方法具有图论理论的支持,操作处理上更加方便有效。

5 总结

提出了一种基于有向图理论的新闻视频数据库基于故事单元的"多线程"管理方法。分析了故事单元之间相似关系与有向图之间的联系,根据这种联系,研究引入了赋权图的最小部分树理论,根据实际需求进行了一定的修改,提出了一种新的新闻视频故事单元线程化管理方法,方法利用图论理论的支持,对故事单元之间的相似关系图进行了分析,保留了故事单元之间主要的内在依存关系。实验显示,

这种基于有向图理论的“多线程”管理方法比目前广泛应用的线性结构更加能够体现故事单元之间的依存结构, 能够比较好地贴近用户服务需求。与其他线程化组织方法相比, 本文方法具有很好的理论基础支持, 处理上也比较简单, 能够保留和体现重要的相似关系。

参 考 文 献:

[1] Nallapati R, Feng A, Peng F, et al. Event Threading within News Topics[C]// Proceedings of the thirteenth ACM international conference on Information and Knowledge Management, Washington, USA, 2004: 446- 453.

[2] Ide I, Mo H, Katayama N, et al. Threading News Video Topics[C]// Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR2003), California, USA, 2003: 239- 246.

[3] Worring M, Snoek C, De Rooij O, et al. Mediamill: Advanced Browsing in News Video Archives[C]// International Conference on Image and Video Retrieval (CIVR 2006), Tempe, USA, 2006, 533- 536.

[4] De Rooij O, Snoek C, Worring M. Query On Demand Video Browsing[C]// Proceedings of the 15th International Conference on Multimedia (ACM MM2007), Augsburg, Germany, 2007: 811- 814.

[5] Tesic J, Natsev A, Seidl J, et al. IBM Multimodal Interactive Video Threading[C]// International Conference on Image and Video Retrieval (CIVR2007), Amsterdam, The Netherlands, 2007: 124- 126.

[6] Ide I, Mo H, Katayama N, et al. Topic Threading for Structuring a Large-scale News Video Archive[C]// Image and Video Retrieval: Third International Conference (CIVR2004), Dublin, Ireland, 2004: 123- 131.

[7] Katayama N, Mo H, Ide I, et al. Mining Large-scale Broadcast Video Archives Towards Inter-video Structuring[C]// Pacific Rim Conf. on Multimedia (PCM2004), Tokyo, Japan, 2004: 489- 496.

[8] Ide I, Mo H, Katayama N, et al. Exploiting Topic Thread Structures in a News Video Archive for the Semi-automatic Generation of Video Summaries[C]// 2006 IEEE International Conference on Multimedia and Expo (ICME2006), Toronto, Canada, 2006: 1473- 1476.

[9] Ide I, Noda K, Ogawa A, et al. Semantic Analysis of a Large-scale News Video Archive[C]// Proceedings of Asian-Pacific Workshop on Visual Information Processing (VIP 2006), Beijing, China, 2006: 166- 171.

[10] Ide I, Kinoshita T, Takahashi T. MediaWalker: A Video Archive Explorer Based on Time-series Semantic Structure[C]// Proceedings of the 15th International Conference on Multimedia (ACM MM2007), Augsburg, Germany, 2007: 162- 163.

[11] Wu Xiao. Threading Stories and Generating Topic Structures in News Videos Across Different Sources[C]// Proceedings of the 13th Annual ACM International Conference on Multimedia (ACM MM2005), Singapore, 2005: 1047- 1048.

[12] Wu Xiao, Chong-Wah N, Li Qing. Threading and Autodocumenting News Videos: A Promising Solution to Rapidly Browse News Topics[J]. IEEE Signal Processing Magazine, 2006, 23(2): 59- 68.

[13] 文军, 吴玲达, 曾璞, 等. 新闻视频中基于“场景词汇”的故事单元相似度分析[J]. 国防科技大学学报, 2009, 31(6): 121- 125.