《现代图书情报技术》版权所有,欢迎下载引用!
请注明引用地址:文本可视化及其主要技术方法研究*[J],现代图书情报技术,2008(8):24-30.

文本可视化及其主要技术方法研究*

赵　琦[1,2]　张智雄[1]　孙　坦[1]
[1](中国科学院国家科学图书馆　北京 100190)
[2](中国科学院研究生院 北京 100049)

【摘要】文本可视化是通过对文本资源的分析,发现特定信息,并利用计算机技术将其以图形化方式呈现来的一种方法。通过分析文本可视化典型系统,分析现今的文本可视化的特点。并从基于词汇、基于篇章、基于时间序列、基于主题领域4种不同的文本可视化方式入手来分析其的主要技术方法。最后探讨文本可视化如何在信息环境下发挥作用。

【关键词】文本可视化　知识表示　主题发现
【分类号】G250.76

A Research on the Methodological of Text Visualization

Zhao Qi[1,2]　Zhang Zhixiong[1]　Sun Tan[1]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](Graduate University of Chinese Academy of Sciences, Beijing 100049, China)

【Abstract】Text visualization is a method which uses computer technology to make a graphical show of the specific text resources. This paper analyzes the current text visualization characteristics through analysis of the typical text visualization system. There are four different classes of text visualization, including based on vocabulary, based on article, based on time series, based on topic which reflects the main text visualization techniques. The final part is about how text visualization used in the information environment now.

【Keywords】Text visualization　Knowledge expression　Topic discovery

1　概　述

随着海量信息的不断涌现,人们利用传统的阅读和检索方式理解大量、复杂信息的难度日益增大。为此,各种在海量的文字、数据信息中发现知识的方法纷纷涌现,而文本可视化就是这众多技术的重要组成部分。文本可视化是通过对文本资源的分析,发现特定信息,并计算机技术将其以图形化方式呈现来的一种方法,是信息可视化的重要分支之一。文本可视化的目的是以丰富的图形或图像揭示以文本为载体的信息内容。文本的可视化技术可以高度概括并且形象化表示文本信息中的核心内容,方便人们快速的理解和吸收文本中的核心思想。文本可视化技术亦可显示出文本中的隐含内容和隐含关系,为基于海量文本知识发现提供更好的支持。文本可视化涉及信息抽取、自然语言理解、人机交互、心理学等多方面内容,是一个综合性的研究课题。

文本可视化的本质在于针对海量的文本信息,最大程度的实现抽象和概括。它不仅仅是多样的图形、图表的组合,更大的作用在于发现一篇文档或者一系列文档集合中特定的、潜在的数学结构。Wise 认为[1],这种潜在的数学结构可以分为3类:

收稿日期:2008-06-16
*本文系国家社会科学基金项目"从数字信息资源中实现知识抽取的理论和方法研究"(项目编号:05BTQ006)的研究成果之一。

(1)基于词频统计的文本可视化。主要以词频统计方法为基础,以发现特定的词频模式(如高频词,异常词频)为目的来实现可视化。

(2)基于聚类算法的文本可视化。引入了聚类的思想,不单单考虑特定的词语或词频,而是利用文档中出现的词语比较文档的相似性,更进一步的产生聚类结果。基于聚类的文本可视化多应用在信息检索、网页的链接关系分析、主题探测、新兴趋势发现等领域。

(3)基于语义的文本可视化。不仅仅要求发现文本中特定的词,而且还要求通过标注、计算、统计、推断等各种技术手段,发现文章中的特定的隐含语义关系,并且更进一步的发现知识。

2　文本可视化的主要方法

目前,根据不同的应用目的,产生了众多的文本可视化研究和设计方案,各具特色,主要可以总结为以下几类:基于词汇的文本可视化、基于篇章内容的文本可视化、基于时间序列的文本可视化和基于主题领域的文本可视化。

2.1　基于词汇的文本可视化

词汇的可视化是文本可视化的一种基本的思路,其主要的可视化对象是文本中的词汇,希望通过对于词汇的不同呈现,来展现文本的特征。由以下几个典型系统,可以看到基于词汇来进行文本可视化的几种主要思路:

(1)基于命名实体的可视化——Contexter

Contexter[2] 由 Jozef Stefan 研究院知识技术部门开发。设计者认为,对于文本内容的分析往往是根据特定的需求。并不要求所有的关键词和关系都需要抽取和显示出来,而只有一些关键的人名、地名、术语及其关系是需要被发现的。

Contexter 系统首先利用信息抽取方法发现特定的命名实体,这些命名实体就是需要呈现的特定词汇。在此基础上,系统建立了命名实体之间的共现关系。

其可视化的内容主要包含以下两个方面:

①以网络图的形式表现实体之间的共现关系。

②反映实体之间的背景信息,包括:命名实体相关的关键词、相关的其他命名实体、以及根据选定命名实体和其相关命名实体共同计算出的出现频次较高的特定关键词。

Contexter 系统会首先将处理完成的语料,以命名实体、词袋(Bag - of - word)、原始文本等 3 种形式存储在数据库之中。词袋中包含原始文本的大多数词语,只是去掉了其中的停用词,并且做了共指消解的处理。利用词袋的目的,主要是为了全面的覆盖原始文本中的背景信息。利用一些特定的算法(典型的有 TF - IDF 算法),判断出文章中的关键词。从而建立命名实体与关键词的联系。

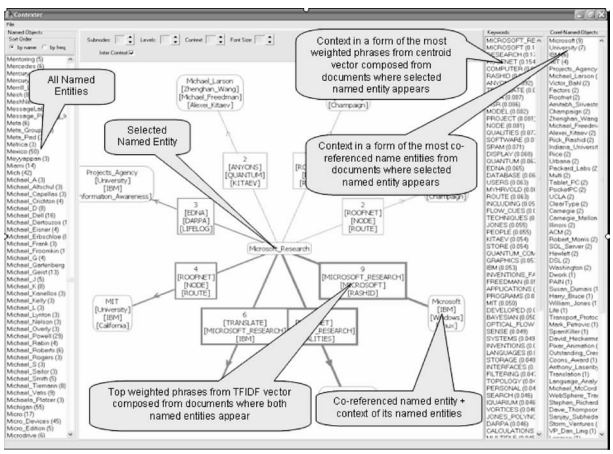

图 1　Contexter 系统可视化效果展示[2]

用户进入系统得到的界面如图 1 所示。点击左侧窗口选择命名实体,系统会返回特定命名实体背景信息。图中的中心结点就是选定命名实体,与其相连的外面一圈词语是从该命名实体出现的文本中,根据 TD - IDF 算法确定的最为相关的关键词。关键词相连接的外面的一层同样是一些命名实体,系统认为这些命名实体也与该关键词密切相关。这样,系统就建立了命名实体之间语义关系。同时,选择一个命名实体时,右侧两栏还会显示出与命名实体相关的关键词和其他命名实体。可以利用这些内容进行反查,发现更多的实体关系。

(2)词汇分布的可视化——Tilebars

Tilebars[3] 用于命中文献的显示,准确地说它提供的信息不是对文献的描述,而是表明查询条件与检索结果文献之间的相关性的中间的信息,通过这种方式显示使用户更清楚的了解文献内容,从而有针对性地选择文献原文。Tilebars 的基本原理是:将一篇文献从语义上划分为若干单元块(如章节、段落),假定用户有一组检索词,系统将显示出每个检索词在该文献中出现的分布情况。

Tilebars 提供了 3 方面的信息:文献的相对长度;查询条件集合在文献中的频率;查询条件集合关于文

献和它们彼此的分布状况。Tilebars 的行数等于检索条件组的数目,列数等于文献所划分的单元块数目。每个方块表示了在相应的文献片段中相应的条件组的命中数目。方块的颜色表示查询条件在那个原文片断中发生的次数,方块的颜色越深表明命中数目越大(白色表示 0,黑色表示 8 或者更多,一个条件组中所用条件的频率被加在一起)。

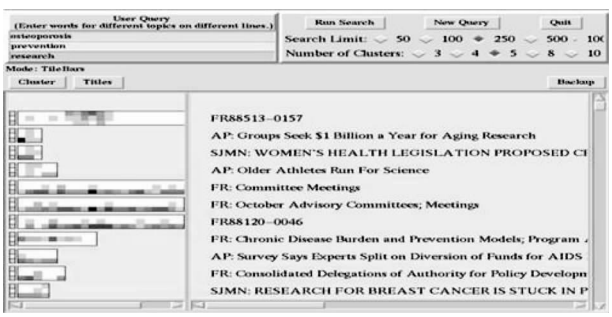

图 2　Tilebars 系统可视化效果展示[4]

如图 2 所示,Tilebars 显示的是关于骨质疏松症防治研究的一个查询结果。图形化的描述是这样形成的:每一个矩形表示一篇文献,矩形的每一个行表示与查询显示相关的条件组,例如:第一行与“骨质疏松症”相关,第二行是于“防治”相关,第三行是与“研究”相关。矩形还被细分为列,每一列表示一个原文片断。最左端的列表示文献的第一个片断或者段落,它右边的列表示文献的第二个片断,如此类推。可以看出,第一篇文献的图形显示出三个条件组只是交迭在相对靠前的位置,通过这个图形,用户能够设想这里只是在文献的前半部分稍微提及了医学应用软件;而第二篇文献中,论文篇幅较短,探讨骨质疏松症防治研究的内容较少。其他文献也可以以此类推,做同样的理解和解读。

用户能够迅速看到条件组的子集或全部在文献的相同片段上是否交迭在一起,以及在文献的何处发生了这个交迭。在 Tilebars 中文献可以首先根据交迭所有条件组的片段的数量排序,其次根据文献中查寻条件的全部频率排序。

(3)文本词汇可视化颜色标识理论

用颜色来区分和展示文本也是较为常见的一种可视化方式。在典型的信息抽取系统(如 GATE)中,会将不同的实体类型以不同的颜色标亮,加以区分。实际上,颜色除了区分的功能,还可以起到发现句子结构,区分文体的作用。Wibke Weber[5] 设计了一套基于颜色的文本标记体系。经过标注的文学小说文本显示出了比科学类文本更为明快的颜色,起到了显示文风的作用,如图 3 所示:

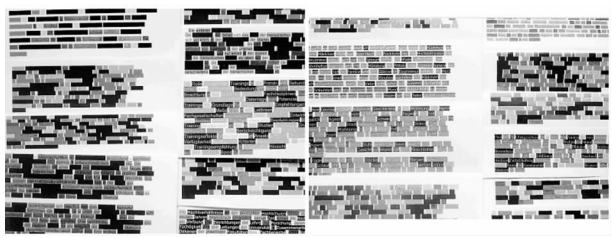

图 3　颜色系统对比 左 - 科学文献 右 - 小说[5]

2.2　基于篇章内容的文本可视化

篇章内容可视化不仅仅要求发现文本中特定的词,而且还要求通过标注、计算、统计、推断等各种技术手段,发现文章中的特定的隐含语义关系。这样读者可以快速地找到文章的主题和核心内容。以图形化的方式,更加有效阅读和理解文本内容。NLPWin[6]、TextArc[7] 具有一定的代表性。

(1)发现语义关系的篇章内容可视化——NLPWin

NLPWin 由微软研究院开发,系统通过抽取文本中命名实体和关系揭示语义关系,形成文本概览。系统研究者首先深入分析了文本的句法结构,对于每一个句子都抽取一个主谓宾的逻辑三元组关系。接下来利用跨句指代处理、共引处理、语义规范化等方法提炼精简生成的三元组关系,并映射到可视化图中。图 4 是 NLPWin 呈现的克林顿一次演讲的概览,反映出了的这次演讲中核心词及相应语义关系。

(2)发现阅读线索的篇章内容可视化——TextArc

TextArc 是一个在文本中发现模式和概念的工具。这个工具的最大特点在于,利用了人类可视化处理(Human Visual Processing)的方式与自然语言处理和计算语言学的技术相结合,提供一个文本文件的概览图形。接下来利用人的认知能力,与概览图形互动,最终可以从未阅读的文本中快速发现文中的主要任务、概念及核心思想。图 5 是经过 TextArc 处理过的 Lewis Carroll 作品《爱丽斯漫游奇境记》的可视化图形。TextArc 处理的这篇文章并没有使用任何一种算法抽取出特定的词,而是把文中所有的词都列在了图形之上,出现越频繁的词会以越明亮的颜色显示出来。TextArc

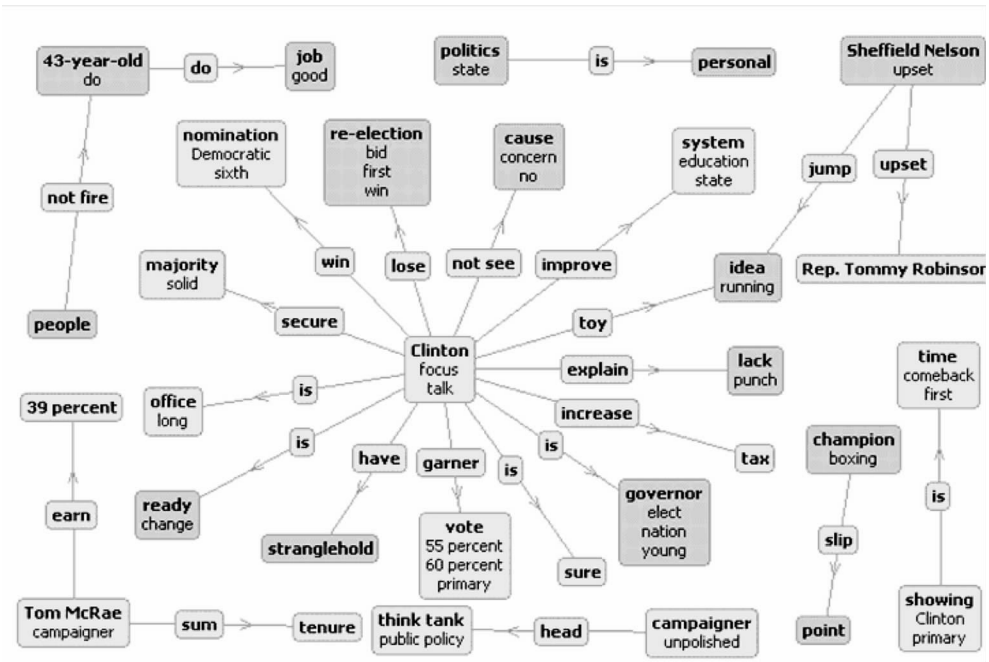

图 4　NLPWin 系统可视化演示[6]

图形的基本组成是两个同心的 360 度螺旋,第一层螺旋也就是外层从 12 点钟旋转一周,以非常小的、不可见的字体,按照在文本中的出现顺序,列出了所有的文本中的所有词语;第二层螺旋,依然是从 12 点钟的方向开始旋转,不同的是,这一次词汇要以可读的字体显示出来,而且出现多次的词只显示一次。词的位置由这个词到外层螺旋出现的不同位置的平均距离得出。出现频率越多,在文本中分布越均匀的词,就会出现在越接近圆心的位置。TextArc 同时还可以基于词在文本中出现的顺序,计算并绘制出阅读线索,达到快速了解文本内容的目的。

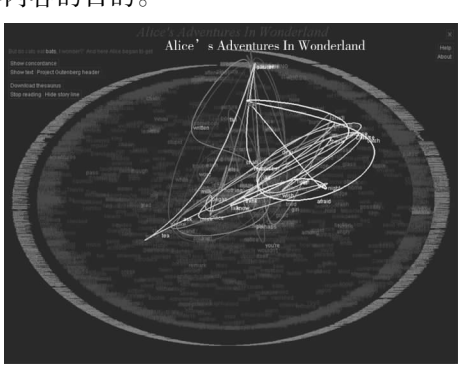

图 5　TextArc 可视化效果图[8]

基于篇章内容的文本可视化其难点也是重点就是发现文章中的语义关系。目前除上述两个典型系统外,利用 Ontology 呈现篇章语义关系也是研究的热点。Ontology 包含丰富的结构性信息,作为一种语义框架,是描述语义关系的良好载体。典型的系统有 SKET 项目的 Text Garden[9] 系统等等。

2.3　基于时间序列的文本可视化

时间是文本的一个重要属性,针对文本的时间关系进行可视化,同时在此基础进行一些特定分析任务,如发现研究趋势、技术进化发展规律等等是十分有意义的。基于时间序列的文本可视化正是为了满足这些需要而产生。代表性系统如 ThemeRiver[10]、LifeLine[11]、ThemeMountain[12]、History Flow[13] 等等。

(1) 主题变迁可视化——ThemeRiver

ThemeRiver 用一个河流的形式来表现特定主题领域随时间的变迁。图中的每一个彩色线条代表一个特定主题,线条的粗细表示主题的强度(可以是数量、频率、关注度等等),而总体线条的粗细代表一定时间段内特定主题的强度。如图 6 所示,描述的是某公司几年来专利集合的主题变化。图的底端是时间变化(年代),可以看出:专利申请增长和专利申请减少阶段交替出现。在第三个凸起部分,Histogram、Predictive、Inference 作为专利主题在 1994 年首次出现;Paradigm、在

1995 年是一个重点、大量研究的主题；Internet、Trends、Gesture、Dialogue 是 1996 年专利中的新主题。ThemeRiver 可以实现用户识别特定主题的变化模式，以验证或者推翻自己的假设。

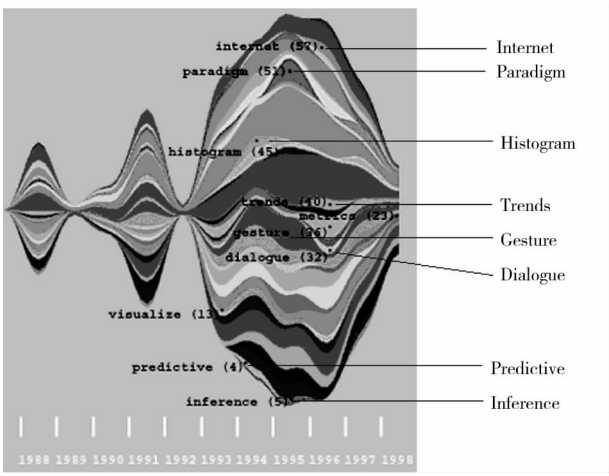

图 6　ThemeRiver 可视化效果图[10]

(2)内容变迁可视化——History Flow

History Flow 是由 IBM 开发的一个针对多作者的、动态改变的文本的可视化工具，能够随时间的变化展示文本所产生的内容是如何变迁的，以及多个作者以怎样的模式对同一个文本作出贡献。History Flow 将每一个版本的文本标识成一条线段，线段的位置表示该版本产生的时间，线段的不同颜色表示不同作者对该文本不同段落的贡献，而后将同一颜色的线段横向连接起来，便出现了曲折变化的效果，如图 7 所示：

图 7　History Flow 形成示意图[14]

History Flow 对于多作者、内容多变的文本的变迁模式有很好的可视化分析效果，最为典型的应用是对维基百科的可视化，如图 8 所示，是对维基百科中 Microsoft 这一词条内容变迁的可视化。左侧指示不同的颜色代表的不同作者，中间是 History Flow 可视化效果图，可以在图中用十字交叉线定位，在右侧查看不同的版本，下方是时间刻度，上方是 History Flow 提供的不同显示模式(包括多作者同时显示模式、突出单作者模式、突出文本改动模式、文本老化模式、链接模式等)。可以看出图 8 的词条随时间篇幅不断加长，并随着写作的作者越来越多，内容也不断发生变化。根据 History Flow 图还可以得出许多关于文本内容变迁以及作者合作模式的分析结果。如根据维基百科群体撰写行为的分析，通过 History Flow 图，研究者发现可以清晰反映出以下几种模式：文本内容的恶意破坏和修复；热烈讨论的话题；匿名和署名作者的不同贡献；文本内容的稳定性等。

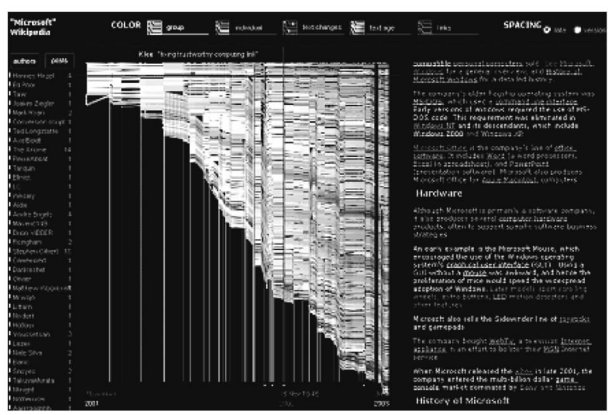

图 8　History Flow 系统演示图[13]

2.4　基于主题领域的文本可视化

基于主题的文本可视化，是针对大规模文本常见的一种模式，主要的目的是从大规模文本中发现特定的一个或者多个主题领域，并反映主题领域之间的关系。基于主题领域的文本可视化最常用于发现学科热点、演变和趋势。

IN－SPIRE™ 由 PNNL(Pacific Northwest National Laboratory)开发，是基于大规模的自由文本抽取和可视化典型系统[15]。IN－SPIRE 可视化系统包含两个主要的模块：ThemeView 和 Galaxy 两者的可视化效果不尽相同，如图 9 所示。ThemeView 主要是利用三维空间的山峰和山谷的形式表现主题和主题之间的关系。内容相近的文献在图中的距离也相近，最终形成山峰，图中不同山峰区域内表示某一特定技术主题中聚集的相应的主题。同一区域的文献数量与地图中山峰的高度相对应。文献内容越相似，文献点在图中的位置就越近；等高线表明了相关文献的密度：最高峰的高点区域包含的文献最多，低点区域包含的文献相对较少：峰间距离越近，表明所相应内容相似性越近；反之，山峰

越高表示的该类别的文档聚集越多,两山越近表示主题直接越密切相关。而 Galaxy 是在二维空间上,用星云团的来表现主题聚类,能呈现出特定文本集里文章根据主题聚集或离散的态势。

图 9　IN – SPIRE™可视化效果图 ThemeView 和 Galaxy[16]

IN – SPIRE™的两个子系统尽管表现主题聚类的形式不同,可是处理文档并最终形成可视化结果的基本流程与技术是一致的,可概括为以下几个步骤:

(1) 扫描特定的文档源,确定其中的记录、所在领域以及相应术语,同时进行标引;

(2) 将"领域—术语"的索引转化成"术语—领域"的索引,同时结合"领域—记录"索引建立"术语—记录"的索引,上面的转换过程利用了 FAST – INV 算法;

(3) 利用建立的索引发现相关联的术语集群,形成"主题"和"核心术语";

(4) 利于"主题"和"核心术语"计算得到关联矩阵;

(5) 利用关联矩阵,为每一个文档做"知识签名",把文档定位在一个 n 维的空间的特定正交维数上;

(6) 依据在 n 维空间测定的主题之间的距离形成聚类;

(7) 降维,最终实现可视化的结果。

基于主题的文本可视化典型的系统还有很多。美国 Thomson 集团公司推出的 Aureka IPAM 知识产权管理系统包含一个专利地图 Aureka ThemeScape[17] 文本分析工具,该工具以分析的专利样本为基础,对其中的相关词汇的词频应用聚类分析生成主题(词汇)地形图,以此来描述专利技术主题分布情况。该分析工具可以辨别和提出词汇系列中经常出现的关联词组,以及它们在文献中的相互关系。如同样是 PNNL 开发的 Topic Island[18],该系统以小波变换的理论为基础,显示某一主题在篇幅较长的文本分布情况,其特点在于为文档定位了不同的分辨率级别 MRLs (Multiresolution Levels),从而实现文本快速浏览和高精度主题定位。同时还有 HNC 的软件公司开发的 DEPICT[19],用于可视化大规模文本语料库。该系统是建立在两个单独的神经网络方法:上下文向量(Context Vectors,CVs)和自组织特征映射(Self – Organizing Map,SOM)。这些神经网络技术可以用于在缺乏相关背景数据和知识的情况下,自动识别语料中的主题,并以直观的形式呈现给用户。

3　结　语

文本可视化技术尚处在发展阶段,相应的技术和算法仍有待研究。作为信息分析处理的展示环节,其应用的前景十分广阔。文本可视化技术的作用不仅在于能更丰富和生动地表达结果,更在于能通过一系列的算法和设计,展示文本资源中的潜在语义联系,发现新颖的信息。利用文本可视化技术,生成丰富的图表和图像,可以充分概括文字和数据分析得到的结果,并以更加易于理解和接受的方式展现出来。从而使知识发现的分析结果为更多、更广泛的人群所理解,在情报研究、决策支持等相关领域发挥出巨大作用。

参考文献：

[1] Wise J A, Pennock K, Lantrip D, et al. Visualizing the Non – visual: Spatial Analysis and Interaction with Information from Text Documents [C]. *Proceedings on Information Visualization* 1995.

[2] Mladenic M G D. Visualization of News Articles [EB/OL]. [2008 – 06 – 12]. http://eprints. pascal – network. org/archive/00000742/01/GrobelnikMladenic – Contexter. pdf.

[3] Hearst M A. TileBars: Visualization of Term Distribution Information in Full Text Information Access [C]. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1995: 59 – 66.

[4] TileBars Examples [EB/OL]. [2008 – 06 – 12]. http://people.ischool. berkeley. edu/ ~ hearst/images/tb – example. html.

[5] Weber W. Text Visualization – What Colors Tell About a Text [C]. In: *Proceedings of the* 11*th International Conference Information Visualization*, 2007: 354 – 362.

[6] Leskovec J, Grobelnik M, Milic – Frayling N. Learning Sub – structures of Document Semantic Graphs for Document Summarization [C]. *LinkKDD*. 2004.

[7] Paley W B. TextArc: Showing Word Frequency and Distribution in Text [C]. *IEEE Symposium on Information Visualization*. 2002.

[8] TextArc – An Alternate Way to View a Text [EB/OL]. [2008 – 06 – 12]. http://textarc. org/.

[9] Grobelnik M, Mladenic D. Text – Garden——Text – Mining Software Tools [EB/OL]. [2008 – 06 – 12]. http://kt. ijs. si/Dunja/textgarden/.

[10] Havre S, Hetzler B, Nowell L. ThemeRiver™: In Search of Trends, Patterns, and Relationships [EB/OL]. [2008 – 06 – 12]. http://infoviz. pnl. gov/pdf/themeriver99. pdf.

[11] Plaisant C, Mushin R, Snyder A, et al. LifeLines: Using Visualization to Enhance Navigation and Analysis of Patient Records [J]. *The Craft of Information Visualization: Readings and Reflections*, 2003. 1(1): 1 – 5.

[12] Battiato S, Gianpiero Di Blasi, Gallo G, et al. Theme Mountain: a SVG – based Visual Data Mining Tool [EB/OL]. [2008 – 06 – 12]. http://www. svgopen. org/2005/papers/ThemeMountain/.

[13] History Flow – Visualizing the Editing History of Wikipedia Pages [EB/OL]. [2008 – 06 – 12]. http://www. research. ibm. com/visual/projects/history_flow/.

[14] Viégas F B, Wattenberg M, Dave K. Studying Cooperation and Conflict Between Authors with History Flow Visualizations [C]. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2004: 575 – 582.

[15] Krishnan M, Bohn S, Cowley W, et al. Scalable Visual Analytics of Massive Textual Datasets [C] In: *Parallel and Distributed Processing Symposium*, IPDPS. 2007.

[16] IN – SPIRE™ Visual Document Analysis [EB/OL]. [2008 – 06 – 12]. http://in – spire. pnl. gov/.

[17] Aureka ThemeScape 9. 2 User Guide [EB/OL]. [2008 – 06 – 12]. http://aureka. micropat. com/7w/html/customer _ support/documentation/user_guides/themepublisherug. pdf.

[18] Miller N E, Wong P C, Brewster M, et al. TOPIC ISLANDS™ – A Wavelet – Based Text Visualization System [C]. In: *Visualization'98. Proceedings*. 1998: 189 – 196.

[19] Rushall D A, Ilgen M R. DEPICT: Documents Evaluated as Pictures. Visualizing Information Using Context Vectors and Self – organizing Maps [C]. In: *Proceedings of the* 1996 *IEEE Symposium on Information Visualization* (*INFOVIS'96*). 1996, IEEE Computer Society.

（作者 E – mail: zhaoqi@ mail. las. ac. cn）