软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(5):1163−1173 [doi: 10.13328/j.cnki.jos.004962]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于文本摘要及引用关系的可视辅助文献阅读*

张加万[1], 杨思琪[1], 李泽宇[1], 杨伟强[1], 王锦东[1], 贺瑞芳[2], 黄茂林[1,3]

[1](天津大学 软件学院,天津 300350)

[2](天津大学 计算机科学与技术学院,天津 300350)

[3](Faculty of Engineering and Information Technologies, School of Software, University of Technology Sydney, Australia)

通讯作者: 张加万, E-mail: jwzhang@tju.edu.cn

摘 要: 近年来,科技论文发表数量与日俱增,科研人员需要阅读文献的数量也随之迅速增长.如何快速而有效地阅读一篇科技论文,逐渐成为一个重要的研究课题.另一方面,在阅读科技论文时,理解与其相关的重要参考文献可帮助读者更好地理解文章的内容.然而,如何从众多的参考文献中快速找到最重要、最相关的几篇,如何避免在阅读过程中迷失在文档的多维空间,仍是值得研究的问题.为了解决上述问题,提出了一个基于文本摘要和引用关系的可视辅助文献阅读系统.该系统利用一种基于阅读目的的文本摘要技术提取出论文中重要的句子,并采用多尺度的可视化方式进行展示;使用 LDA(latent dirichlet allocation)话题模型抽取参考文献的核心话题;记录用户的阅读行为,用于提示其阅读上下文,以保证用户关注点不发生迷失.同时,在一个具体的案例场景中详细介绍了系统的使用方法,并进行了用户研究以验证系统的可用性.

关键词: 文档可视化;文本摘要;引用网络;阅读行为分析;文本可视分析

中图法分类号: TP391

中文引用格式: 张加万,杨思琪,李泽宇,杨伟强,王锦东,贺瑞芳,黄茂林.基于文本摘要及引用关系的可视辅助文献阅读.软件学报,2016,27(5):1163−1173. http://www.jos.org.cn/1000-9825/4962.htm

英文引用格式: Zhang JW, Yang SQ, Li ZY, Yang WQ, Wang JD, He RF, Huang ML. Visualization guided document reading by citation and text summarization. Ruan Jian Xue Bao/Journal of Software, 2016,27(5):1163−1173 (in Chinese). http://www.jos.org.cn/1000-9825/4962.htm

Visualization Guided Document Reading by Citation and Text Summarization

ZHANG Jia-Wan[1], YANG Si-Qi[1], LI Ze-Yu[1], YANG Wei-Qiang[1], WANG Jin-Dong[1], HE Rui-Fang[2], HUANG Mao-Lin[1,3]

[1](School of Computer Software, Tianjin University, Tianjin 300350, China)

[2](School of Computer Science and Technology, Tianjin University, Tianjin 300350, China)

[3](Faculty of Engineering and Information Technologies, School of Software, University of Technology Sydney, Australia)

Abstract: With growing volume of publications in recent years, researchers have to read much more literatures. Therefore, how to read a scientific article in an efficient way becomes an importance issue. When reading an article, it's necessary to read its references in order to get a better understanding. However, how to differentiate between the relevant and non-relevant references, and how to stay in topic in a large document collection are still challenging tasks. This paper presents GUDOR (GUidedDOcument Reader), a visualization guided reader based on citation and summarization. It (1) extracts the important sentences from a scientific article with an objective-based summarization technique, and visualizes the extraction results by a multi-resolution method; (2) identifies the main topics of the

* 基金项目: 国家社会科学基金(12&ZD213); 国家科技支撑计划(2013BAK01B05, 2014BAK09B04)

Foundation item: National Social Science Foundation of China (12&ZD213); National Key Technology R&D Program of China (2013BAK01B05, 2014BAK09B04)

收稿时间: 2015-07-30; 修改时间: 2015-09-19, 2015-11-09; 采用时间: 2015-12-05

references with a LDA (Latent Dirichlet Allocation) model; (3) tracks user's reading behavior to keep him or her focusing on the reading objective. In addition, the paper describes the functions and operations of the system in a usage scenario and validates its applicability by a user study.

Key words: document visualization; text summarization; citation network; reading behavior analysis, visual text analysis

近年来,科技论文的发表数量持续增长,这给科学研究人员造成了一定的压力.为了掌握研究领域内同等比例的最新研究,科研人员需阅读和原来相比几倍数量的文献.已有研究[1]表明:每个科研工作者每年所阅读的平均文献数持续增长,然而他们花在阅读同一篇文献上的时间却呈减少趋势.对于读者来说,阅读大量的文献是一项很耗费时间和脑力的工作.如何在有限的时间内有效地阅读一篇科技论文,成为一个亟需解决的问题.

在文档集合的辅助阅读这一研究问题上已有一些相关的研究[2,3],然而,仍有一些问题在这些研究中并未涉及.一篇科技论文的参考文献不仅可以用于阐明和佐证文章提出的方法和观点,同时也可以体现出作者对文章所涉及研究背景的理解[4].在阅读一篇科技论文时,理解与文章紧密相关的一些参考文献对于读者来说是十分必要的.然而,仅依赖文章中对参考文献的简短描述,读者无法对一篇参考文献的内容形成整体认知.而通篇阅读一篇参考文献也并不是一个有效率的方法.另一方面,在读者阅读一篇文章和相关的参考文献时,难免会在当前文章与其参考文献间频繁地切换,而这种频繁的切换会使读者丢失其阅读的上下文,这种现象被称为文档多维空间中的迷失(lost in hyperspace)[5−8].

为了解决上述问题,本文提出了一个可视辅助文献阅读系统.它是一个帮助读者阅读科技论文及其相关参考文献的可视化阅读工具,它利用可视文本摘要技术辅助用户进行文献阅读,并在用户阅读过程中记录其阅读行为,提示用户其阅读上下文,以保证其关注点不会迷失.整个系统由 3 个协同视图组成:文本摘要视图、话题概览视图和阅读行为记录器.文本摘要视图中,我们利用一种基于阅读目的的文本摘要方法抽取文章的主要内容,并以便利贴为可视隐喻,通过一种多尺度的可视化方法进行展现.话题概览视图中,我们使用 LDA 话题模型进行分析,得出一篇参考文献的核心话题,由此使用户通过文献所涉及的话题了解其概要内容,从而分辨出相关的参考文献.阅读行为记录器负责记录用户在文档集合中跳转、浏览文献等行为,在用户阅读文献的过程中,可以利用其记录的阅读行为数据提示用户某一时刻的阅读上下文,避免用户产生迷失.同时,通过分析这些阅读行为的记录,本文发现了一些在阅读科技论文时的模式.

综上,本文的贡献主要包括以下几点:

1) 一种基于可视文本摘要技术的多尺度论文阅读方式;

2) 一种帮助用户在阅读文档集合过程中不丢失阅读上下文,防止迷失在文档多维空间的可视化方法;

3) 一个可视辅助的文献阅读系统,帮助用户快速、有效地理解论文及其相关参考文献.

本文第 1 节讨论已有的与文档阅读相关的工作.第 2 节提出可视辅助文献阅读系统,并详细阐述系统使用的文本挖掘分析方法及可视化设计.第 3 节描述系统的使用场景,并针对用户的使用反馈做简要分析.第 4 节对本文做简要的总结,同时提出今后工作的方向.

1 相关工作

1.1 文档及文档集合的可视分析

许多研究关注于文档内容的可视化.Tag Clouds[9]和 Wordle[10]是最具代表性的两种词频可视化的方法.TextArc[11]不仅是一种词频可视化方法,还可以展示词的分布情况.DocuBurst[12]以放射状层次圆环的形式展示文本结构及词频.这几种可视化方法虽简单直观,但无法涵盖文档中大部分的语义信息.Stoffel 等人[13]利用高亮文档缩略图中一些词语及其上下文的方法,增强了文档查询时文档缩略图的可读性.而本文的关注点不在于文档查询操作时,而是在文档阅读时的高效性.Document Cards[2]利用图片和基于 TF/IDF(term frequency inverse document frequency)的关键词来代表一篇文章的内容.VarifocalReader[14]融合了 focus+context 和 overview+detail 两种可视化技术的特点,使读者在阅读文档可视摘要的同时,可以获得相应的文章细节.然而,上述两篇文

章并没有考虑文章与参考文献之间的引用关系.

一些研究关注于引用网络中的文档集合.CiteSpace II[15]将协同引用网络进行可视化,可以方便地对某个特定科学领域的发展趋势进行分析和研究.CircleView[5]利用一个基于网络的图形界面,实现对引用网络的可视化和导航.Dunne 等人[3]提出了一个可视化工具,用于浏览某一科学领域内的文章,发现领域内重要的文章等.PaperVis[6]利用一个信息聚类算法将科学文献分类,同时,可视化地展现科学文章之间的相关关系.Schafer 等人[16]提出了一个帮助科研人员浏览大量文档集合的可视化工具,方便用户快速了解一个科研领域.上述几篇文章研究引用网络中论文之间的关系,但并没有在科技论文的内容阅读方面做过多研究.

综上,基于更好地理解文章内容的要求,本文不仅需要考虑到针对一篇科技论文的内容可视化,其引用网络上参考文献的可视化也同样重要.还没有一篇文章可以综合考虑这两点内容,所以本文有必要提出一种新的文档集合可视化方法.

1.2 文档摘要技术

在文本分析领域,已经有了许多文本自动摘要的方法.这些方法主要可以分为两类:基于句子的文本摘要和基于关键词组的文本摘要.

基于句子的文本摘要方法通过提取出文档中最重要的句子来完成摘要.Teufel[17]提出了划分不同论证区域的方法(argumentative zoning),通过句子的修辞状态(rhetorical status)对句子分类.这种方法在文献[18,19]中用作抽取科学论文摘要的一种策略.Mei 等人[20]认为,其他文章对一篇文章的引用信息,是衡量一篇文章内容的重要性的一种方式.文中在构建文章摘要时,充分考虑了其他文章对这篇文章的引用信息.而 Dunne[3]则直接从其他文章对当前文章的引用句中抽取重要句子,作为一篇文章的摘要.

基于关键词组的文档摘要方法,如文献[21−23],利用话题模型来完成关键内容的抽取.一篇文档被拆分成多个话题,一个话题由一组关键词组来代表.然而,由于这些方法得出的结果是词组,其缺陷在于,通过词组来对一篇文档进行展现会丧失大部分的语义信息,适合对文章进行概要展示.

本文进行文档摘要的目的是使读者更加方便地了解文章的主要内容,而并非仅仅是一篇文章所涉及的话题,所以本文的文本摘要方法是在基于句子的文档摘要方法上改进的,同时考虑到了阅读上下文的信息.而在做文章的概览展示时,更关注于一篇文章所涉及的话题,所以这部分使用了基于关键词组的话题抽取技术.

1.3 文档阅读中的用户行为分析

本文旨在能够在记录、分析用户阅读行为的基础上,帮助用户解决在阅读文档的过程中遇到的一些问题.在分析用户阅读行为这一问题上已经有了许多研究,这里只回顾在阅读文档方面的几项相关工作.

文献[24]将一篇文档进行碎片化处理,并提出了一种对碎片化文档进行平行化阅读的工具.Paris 等人[8]帮助用户在一个高度相通的信息空间中阅读,并同时保证用户关注点不会迷失.文中提出了多维空间迷失(lost in hyperspace)这一问题,这也是本文需要解决的问题之一.然而,上述两篇文章并没有关注具有引用关系的文档集合.文献[25]中指出,在阅读中,不同读者会表现出不同的信息检索行为,这些行为表现和他们所要完成的任务有关.Tenopir 等人[1]在研究中发现,每位科研人员每年所阅读的文章数量一直在增加,而他们花费在阅读同一篇文章上的时间却在减少.这表明,科研人员急需一种有效地进行论文阅读的工具.这两篇文章虽然详细分析了用户的阅读行为,但却并没有针对在阅读过程中所遇到的问题提出一个有效的解决办法.

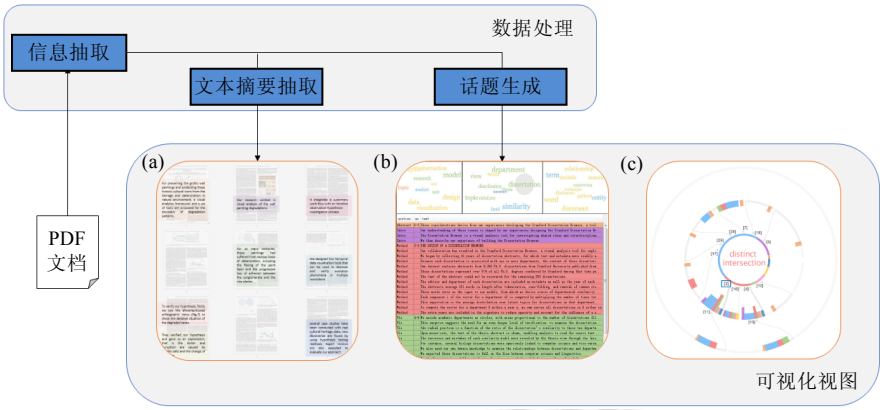
2 可视辅助文献阅读系统

2.1 系统概览

系统在 ICEpdf 类库[26]的基础上开发,主要包括两个部分:数据处理模块和可视分析模块,如图 1 所示.

数据处理模块主要完成文献中相应的文本内容和结构信息的抽取,抽取的结果用于下一步的文本摘要和话题分析.可视分析模块包括 3 个协同视图:文本摘要视图、话题概览视图以及阅读行为记录器,如图 1(a)~图 1(c)所示.文本摘要视图通过多尺度的可视文本摘要辅助用户进行文献阅读;话题概览视图可视化地展示出一

篇参考文献的核心话题,从而使用户决定是否需要对这篇参考文献进行深入阅读;阅读行为记录器用于记录用户在整个阅读过程中的行为,使其关注在自己的阅读目的,不发生迷失.



(a) 文本摘要视图; (b) 话题概览视图; (c) 阅读行为记录器

Fig.1 System overview

图 1 系统概览图

2.2 数据预处理

文献数据通常为 PDF 格式的文件,我们首先利用 Xpdf 工具包[27]对 PDF 格式的文件进行处理.对于每篇文章,我们抽取其文本内容信息及其 PDF 文档的格式信息.得到两种数据结果后,对文本内容数据进行清洗,消除乱码等问题,用于文本摘要和话题的抽取.文档格式信息主要包括原 PDF 文件的页面大小、页数,每个句子在 PDF 文件中的位置信息等,主要用于可视化的布局.

2.3 文本摘要视图

不同的研究人员有不同的科研背景,在阅读文献时也会有不同的阅读目的,因此,他们会有不同的侧重点和信息需求[3].另一方面,用户在阅读一篇文章时与在阅读其参考文献时,阅读内容的详略程度也会有所不同.总之,用户在阅读文章时,会根据阅读目的、文章类型等不同对文章有不同的内容侧重和粒度要求.

2.3.1 基于阅读目的的多尺度文本摘要

由于基于关键词汇的文本摘要会丧失部分语义信息,我们选择使用 MEAD[28]自动文本摘要工具来抽取文章的关键句子.使用 MEAD 时,先确定抽取句子需要考虑到哪些句子特征.在 MEAD 中,默认特征包括表示句子长度的特征 *Length*、句子在文中位置的特征 *Position*、衡量句子与文中话题相似度的特征 *Centroid*.确定特征后,利用默认分类器(default-classifer),根据每个特征对每个句子进行打分,然后,通过默认重排名器(default-reranker)对句子重新进行排序,以降低抽取出的句子在语义上的重复程度.

考虑到用户在阅读时有不同的阅读目的,在使用 MEAD 进行关键句子抽取时,通过增加一个新的特征 *objective* 重新定义了句子的重要程度.*objective* 变量用来表示用户的阅读目的,它由一组词来表示:*objective*={$term_1$,$term_2$,…,$term_n$},每个 $term_i$ 都为文中出现的词.当一个句子中出现 *objective* 中包含的某个 $term_i$ 时,会提高句子的重要程度.用户可以在阅读一篇参考文献的话题概览视图时,通过点击某个词将其添加到 *objective* 中,或从中删除.

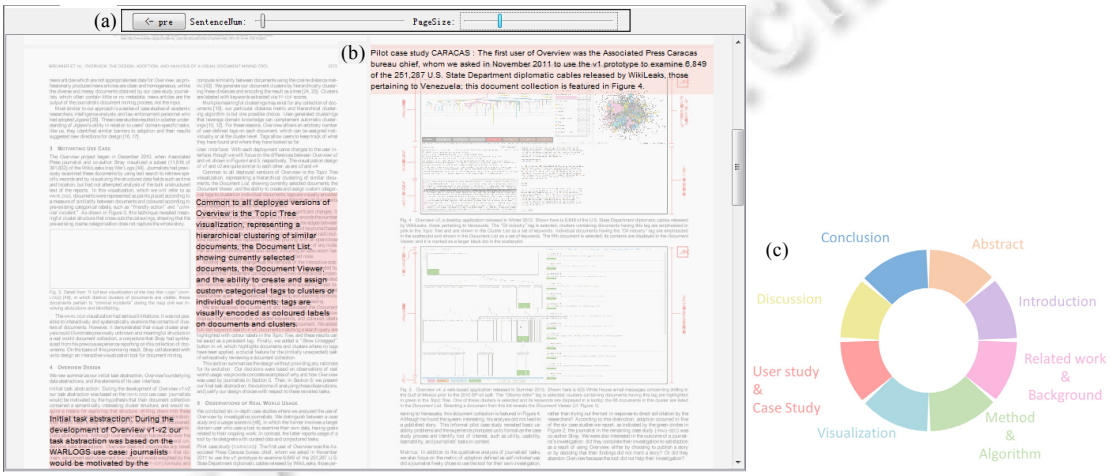
由于不同的用户对摘要的粒度会有不同的要求,本文以一种多尺度的方式来进行文本摘要.一篇文章被划分成不同的粒度,最粗的粒度为整篇文章,依次为章节、子章节,最细的粒度为段落.这种方式允许用户以多尺度的方式在任何粒度上阅读文本摘要.在用户切换阅读粒度时,会重新生成当前粒度范围内的文本摘要(阅读粒度的切换会在下一节及第 3.1 节中详细介绍).同时,用户还可以通过控制抽取句子的数量来增减文本摘要的信息

丰富程度.

2.3.2 可视化设计

本文提出了一种以便利贴为隐喻来展现一篇文章摘要的可视化方法,如图 2 所示.可视化设计共有 3 层,从上到下依次是摘要便签层、缩略图层、原文层.通常情况下,原文层不显示,只有在摘要便签层和缩略图层隐藏的情况下,原文层才可见(图 2 只显示了摘要便签层和缩略图层).在摘要便签层,每一个便签代表摘要中的一个句子.便签颜色代表句子摘取自不同的章节,如图 2(c)所示(本文其他可视化部分也将用此作为统一的颜色编码).缩略图层由多个文章页面的缩略图组成(图 2 显示了两页缩略图).

通常情况,每个便签都贴在句子在缩略图中的对应位置,但当一个句子被分开到两个页面时,为了不影响阅读的连续性,便签会被贴到后一页的顶部,如图 2(b)所示.



(a) 工具栏; (b) 被两页分隔的句子放在后一页的顶端; (c) 颜色图注

Fig.2 Text summarization view

图 2 文本摘要视图

缩略图层只显示当前用户所处粒度所包含的页面,例如,若当前粒度为整篇文章,则缩略图层显示所有页面的缩略图;若当前粒度为段落,则显示当前段落所在的页面.通过缩略图层,用户可以对当前粒度层级的文章结构形成直观的了解.同时,为了避免缩略图上的文字分散用户的注意力,缩略图层上覆盖了一层半透明的层.这样不仅保留了缩略图层所显示出的文章结构信息,又可以使用户的关注点保留在摘要便签层的句子上.

2.4 话题概览视图

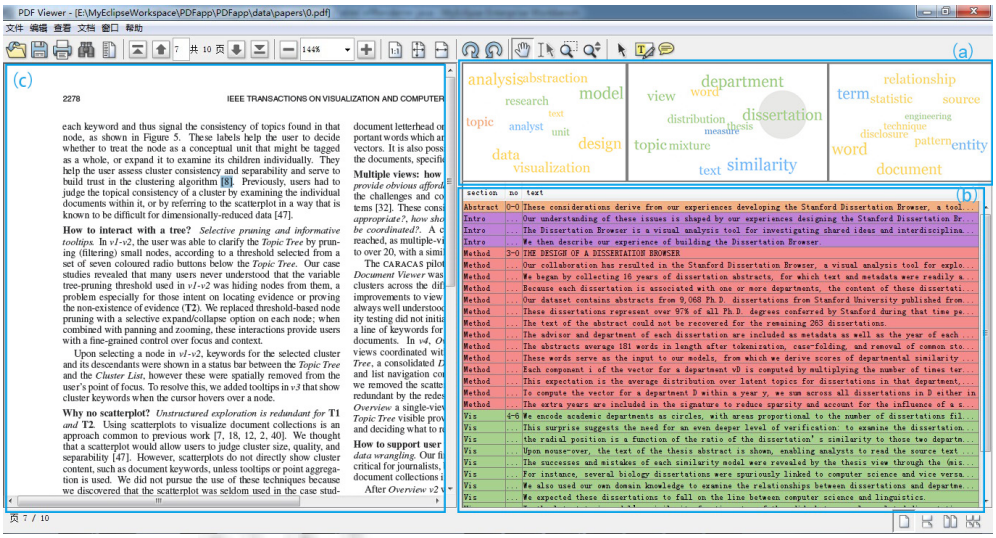
在阅读一篇文章时,根据阅读任务的不同,读者通常会有明确且不同的阅读目的[24],所以在阅读其参考文献时,只需阅读与阅读目的相关的那些参考文献.然而,一篇文章中对某一参考文献的描述都是非常简短且片面的,仅通过阅读这部分简短描述,读者无法全面地了解一篇参考文献.而若打开某篇参考文献进行详细阅读,则是非常费时且无益的,因为通过阅读会发现,许多文献与用户的阅读目的并不相关.

因此,本文提出了一种基于话题模型的可视概览技术,用来概括一篇参考文献的主旨.通过阅读一篇参考文献的话题概览,读者就可以判断出此篇文献是否与其阅读目的相关,是否值得继续深入阅读.本文引入了 LDA 话题模型[29],该模型是一种包含词、主题和文档这 3 层结构的贝叶斯概率模型.它基于一个多文档集合,将文本看成由多个主题混合组成,这些主题被集合中的所有文档所共享,每个主题对应所有词汇上的一个多项式分布,每个文档有一个特定的主题分布,多项式分布的先验分布取其共轭先验 Dirichlet 分布.假设主题的个数为 T,那么在文档 d 中的词汇 w_i 的概率表示为

$$P(w_i \mid d) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j \mid d) \quad (1)$$

其中,$P(w_i|z_i=j)$表示主题 j 中词汇 w_i 的概率,$P(z_i=j|d)$表示文档 d 属于主题 j 的概率.

本文将一篇文章中的每个句子 s_i 当做是一个文档 d_i,并在句子级别上进行主题建模,由此得出一篇文章所包含的所有话题.在得到的话题中选出最重要的 3 个话题,并将每个话题作为一组,生成一个 WordCloud 可视化图.同时,列出话题中包含的词在文中出现的句子,也可通过交互来选择某个话题中特定单词出现在文中的所有句子,如图 3(a)、图 3(b)所示.



(a) 抽取出的话题; (b) 文献[8]中包含“dissertation”的句子; (c) 原文层

Fig.3 System interface

图 3 系统界面

2.5 阅读行为记录器

在阅读文章的过程中,当读者对某篇参考文献感兴趣,想要进行深入阅读时,通常会有两种行为表现[8]:一种是将参考文献作为一个新的页面打开,待当前文章阅读结束后再来阅读这篇参考文献;另一种是立即阅读这篇参考文献.当选择第 1 种处理方式时,随着阅读的进行,会打开许多将要阅读的参考文献页面.当读者把当前文章阅读完,想要找到一篇参考文献继续阅读时,往往会忘记这篇参考文献的阅读目的.如果选择第 2 种处理方式,当完成参考文献的阅读后,读者会忘记原文阅读到了哪部分.无论如何,频繁地在一篇文章与其参考文献之间跳转,都会让读者丢失他们的阅读上下文,这种现象被称为多维空间迷失(lost in hyperspace).

为了避免发生上述情况,本文提出,通过记录读者在整个阅读过程中的行为,使其专注于他们的阅读目的,从而防止关注点发生迷失.本节将详细介绍阅读行为记录器的可视化设计及布局算法.

2.5.1 可视编码

在阅读一篇文章时,尽管当前正在阅读的文章与其他参考文献在本质上都是一篇独立的文章,但由于本文的研究内容以理解某一篇文章为核心,所以在可视编码上,将当前文章与其他参考文献做了区分,如图 4 所示.本节将详细介绍可视化设计中各部分的可视编码.

同心环表示参考文献层级,不同的同心环代表不同的参考文献层级.中心环的深度为 0,表示的是当前正在阅读的文章,环的中心显示参考文献的阅读上下文.外侧的环是其内侧一层环上文章的参考文献.用深度 0,深度 1,...,深度 n 来表示参考文献的层级深度,同时也表示阅读的深度.

弧表示章节,一条弧表示一篇文章中的某一章节,而弧的颜色表示章节的分类.弧的长短与章节中的引用数

量成正比.对于参考文献来说,弧的宽度表示章节的阅读次数.但中心文章是需要用户去仔细阅读的,无需考虑它被阅读的次数,所以中心弧的宽度是不变的.一篇文章可能在不同章节多次引用同一篇参考文献,但每次引用所涉及的内容可能是不同的.当从不同的引用句跳转,去阅读同一篇参考文献时,阅读上下文也随之不同.所以在不同的阅读上下文中,引用句所涉及的是这篇参考文献中不同的内容,这种来自不同阅读上下文跳转而进行的阅读用不同的弧表示.如图 4 中标记为[1]的参考文献,它们所连接的两组彩色弧分别代表了在不同阅读上下文中同一篇参考文献被阅读的部分.

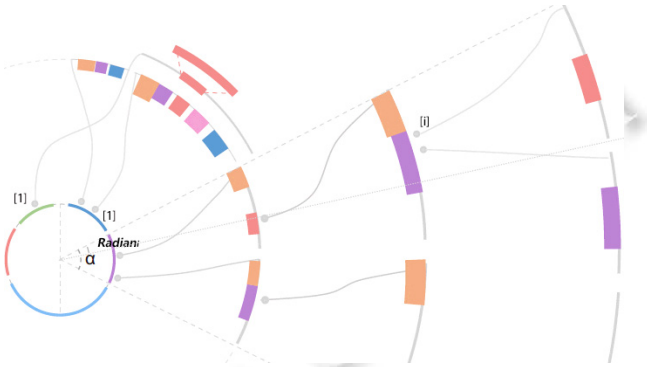


Fig.4 Visual encoding of reading behavior recorder

图 4 阅读行为记录器的可视编码

圆点表示一处引用,短弧上的圆点代表在某一章节中出现的引用.圆点的顺序与引用在文中出现的顺序相同.圆点旁的标注表示了引用在文章中的顺序号.

曲线表示阅读路径,一条连接圆点和弧的曲线代表用户在两篇文章之间的一个跳转.跳转的方向总是从圆点到弧的方向,曲线起点的圆点代表用户跳转之前阅读到的引用,终点的弧代表用户跳转之后将要阅读的参考文献,曲线的宽度代表了用户跳转的次数.

当在整个引用网络中,某一篇参考文献被不同的文章引用时,它可能会出现在不同的层级上.为了保证可视化设计中每篇文章的唯一性,这种情况被归类为上述不同阅读上下文的跳转.发生上述跳转时,用一条曲线连接引用标记与已经出现在图中的参考文献,并将此次阅读的部分用一组新的弧表示,以体现其来自不同的阅读上下文.

2.5.2 布局算法

本节将详细介绍上面提到的可视化元素的布局算法.

曲线曲率.曲线曲率代表阅读路径的曲线是一种贝塞尔曲线,曲线的控制点可以通过用户交互来改变.

同心环的半径.随着参考文献层级变深,参考文献的数量呈指数级增长.如果将参考文献全部显示,为了让环上的信息不发生覆盖,每增加一层深度,弧长也需要呈指数级增长.然而,并不是所有的参考文献都需要被阅读,读者只会根据他们的需求选择性地阅读.综合考虑这两点,在本文的方法中,弧长随层级深度增加呈线性增长,后面的结果显示出这种方法是具有可扩展性的.

扇区划分.中心圆环代表当前正在阅读的文章,而外侧的圆环中,参考文献是随着阅读逐渐产生的,数量不确定,导致每篇文章所占的扇形区域是不确定的.为了让每篇文章上的引用点不发生覆盖,每一个引用点的初始弧长都为 l,如果不满足,就按照比例缩减.我们定义了一个词——中心引用章节(center cited section,简称 CCS).当一篇文章 i 被中心文章的某一章节直接或间接引用(被其中某个参考文献引用)时,此章节就叫做文章 i 的 CCS.文章 i 所占扇形的弧度 $Radian_i$ 的计算公式为

$$\alpha_0=l/r_{level} \tag{2}$$

$$Radian_i=\begin{cases} numOfRef_i\cdot\alpha_0, & \sum_{j=1}^{n}numOfRef_i\leqslant\alpha/\alpha_0 \\ numOfRef_i\Big/\left(\sum_{j=1}^{n}numOfRef_i\cdot\alpha\right), & \sum_{j=1}^{n}numOfRef_i>\alpha/\alpha_0 \end{cases} \tag{3}$$

其中,l/r_{level}为文章 i 所处层级的半径;$numOfRef_i$为文章 i 的参考文献数,$\sum_{j=1}^{n}numOfRef_i$ 是与文章 i 在同一引用层级且有相同 CCS 的所有 n 个文章(包括 i)参考文献数的总和,α为文章 i 的 CCS 所占的弧度,如图 4 所示.

弧的长度.短弧的长度用来区分章节之间参考文献个数的不同.通常,一些章节是没有参考文献的,如摘要、结论等.然而这些章节被阅读的情况需要被可视地展示出来,所以本文使用了一种算法来重新计算一篇文章 i 的各章节所占的弧度.算法的伪代码如算法 1 所示.*ReferenceSet* 是一个集合,其中元素为文章 i 每一个章节中参考文献的个数.*Angle* 和 *TotalLength* 分别是文章 i 所占的弧度和弧长.算法输出的结果集合为 *RadioSet* 和 *LengthSet*,其元素分别为每个章节短弧的弧度和弧长.

算法 1. 计算弧度和弧长.

Input: ReferenceSet, TotalLength, Angle.

Output: RadioSet, LengthSet.

```
1: function ARCLENGTH(ReferenceSet,TotalLength,Angle)
2:     for i=0→ReferenceSet.length do
3:         RatioSet[i]←ReferenceSet[i]/SUM(ReferenceSet)*Angle
4:     end for
5:     while MIN(RatioSet)⇐1 do
6:         for i=0→ReferenceSet.length do
7:             ReferenceSet[i]←ReferenceSet[i]+1
8:         end for
9:         for i=0→ReferenceSet.length do
10:            RatioSet[i]←ReferenceSet[i]/SUM(ReferenceSet)*Angle
11:            LengthSet[i]←RatioSet[i]*TotalLength
12:        end for
13:    end while
14:    return RadioSet, LengthSet
15: end function
```

3 使用场景及案例研究

本节将通过一个具体的案例,详细描述可视辅助文献阅读系统.案例选取了 InfoVIS 会议 2014 年的一篇文章 Overview: The Design, Adoption, and Analysis of a Visual Document Mining Tool For Investigative Journalists (以下简称 Overview),并找到了 9 位对可视化领域有所了解的志愿者参与案例的研究,他们中有 2 位教授、4 位研究生和 3 位本科生.我们首先向每一位参与者介绍了系统的主要功能和操作,然后,要求他们使用系统阅读 Overview 这篇文章,并根据需要阅读相关参考文献,在完成后对系统的使用做简要的评价.通过分析参与者在阅读过程中的行为数据,我们发现了一些规律.

3.1 阅读InfoVis 2014年的可视化文章

首先,遍历以这篇文章为根节点的 4 层参考文献网络,由于一些文章无法从网络上获取到,导致部分子网络上的文章缺失,共得到 1 927 篇文章,其中包括重复文章 362 篇,实际得到 1 565 篇文章.文章都是 PDF 格式的文档,通过数据清洗和处理,得到了系统输入的文档集合.

在系统界面打开该篇文章,左侧是文章的摘要视图,如图 2 所示.通过视图上半部分的工具栏,用户可以调节

视图上显示的摘要句子的数量以及每个页面的大小.点击某个彩色的便签,可以缩放到更细粒度的摘要视图,通过工具栏的 pre 按钮,可返回到更粗粒度.点击没有便签的部分,摘要视图将切换到原文层,此时,摘要便签层和缩略图层隐藏,如图 3(c)所示.当用户在阅读到某一包含引用的句子时,通过点击引用标号,右侧的视图会显示该参考文献的话题概览视图,如图 4 所示.话题概览视图上部显示了这篇参考文献最重要的 3 个话题,通过单击话题上的某个词,下部的句子列表会显示文中包含该词的所有句子.双击话题上的词,将其作为 *objective* 变量的一个特征值,同时也被记录为这篇参考文献的阅读上下文.用户可以通过双击句子列表中的任意句子,打开该参考文献进行深入阅读.在将该参考文献作为新窗口打开时,原文章的阅读窗口隐藏.阅读后关闭当前窗口,原始窗口会再次可见.整个过程中,阅读行为记录器会随着用户逐步阅读,记录行为数据并更新视图.当关闭一篇参考文献的窗口时,用户可以通过阅读行为记录器上的阅读路径,找到前一篇文章正在阅读的章节,文档上也会显示出之前阅读的位置.

3.2 用户反馈及阅读行为分析

参与者的反馈主要集中在以下几个方面:

(1) 文本摘要抽取的句子是否能够反映文章的主旨,从而起到辅助阅读的作用?

(2) 话题概览视图能否有效地反映出一篇参考文献的主题?

(3) 记录阅读行为数据及阅读上下文能否对用户起到提示作用?

(4) 整个可视化设计和系统的使用能否提供一个自然的阅读体验?

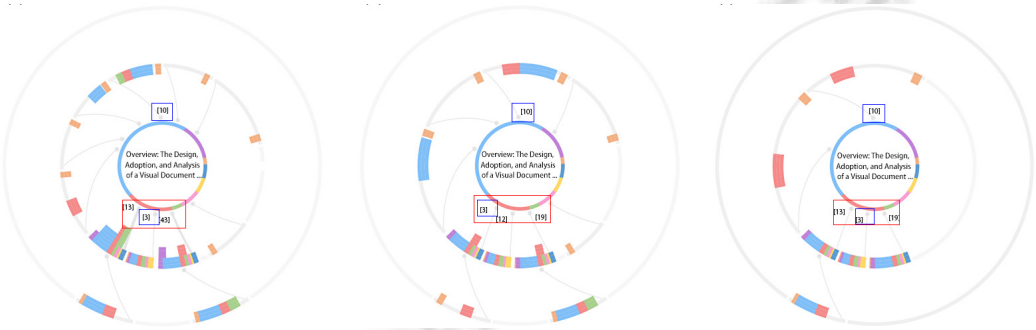
大部分参与者对于系统给予了正面的评价,如,“阅读完一篇参考文献并关闭页面后,会提示之前的阅读进度,让人感觉很方便”,“有时不确定一篇参考文献的内容是否相关,通过浏览相关的话题也能大概了解其内容”等.从中可以看出,记录阅读行为防止关注点迷失这一特性及话题概览视图都得到了较好的评价.

通过分析参与者的阅读记录,我们发现了几条规律:

(1) 不同经验和知识水平的人所阅读文献的数量有所差别:本科生所阅读的参考文献数明显高于硕士生和教授,因他们对特定领域的知识水平不高,需要阅读更多的参考文献,如图 5(a)~图 5(c)所示.

(2) 大部分人都阅读了几篇相同参考文献,图 5 中蓝色框标出的 2 篇文章都被详细阅读了.通过分析发现,这几篇参考文献的话题与中心文章相似度很高,说明阅读这几篇文章对于理解中心的文章十分重要.

(3) 阐述论文方法的章节中的参考文献通常会被详细地阅读,如图 5 中红色框标注的部分.我们尝试揣测其中的原因:通常情况下,阐述方法的章节是一篇科技论文最为核心的部分,这部分引用的文献也通常是和文章联系最紧密的,阅读这部分的文献对理解这篇文章起到关键的作用.



(a) 本科生 (b) 研究生 (c) 教授

Fig.5 Reading trajectory generated by participants read the Overview

图 5 参与者阅读 Overview 这篇文章生成的阅读轨迹图

4 结论和展望

本文提出了一个可视辅助文献阅读系统,帮助科研人员以一种更有效的方式进行科技文献的阅读.文中提出一种基于阅读目的的摘要方法抽取文献的关键句子,并利用 LDA 话题模型得到文献的话题概览,通过记录并分析用户的阅读行为,解决了科技文献阅读中多维空间迷失的问题.

为了评估系统的可用性,我们邀请不同经验和知识背景的科研工作者进行了用户研究.通过研究结果发现:可视辅助文献阅读系统可以对科技论文的阅读起到有效的辅助作用,其可视化方法也提供了自然的阅读体验.通过分析用户阅读行为,我们还发现了一些科技论文的阅读模式.

今后我们将会进行更多的用户研究工作,以改善可视辅助文献阅读系统的用户体验.同时,我们希望收集更多用户的阅读行为,深入挖掘用户在阅读科技论文时的行为模式,并以此为理论基础尝试进行阅读推荐.

References:

[1] Tenopir C, King DW, Edwards S, Wu L. Electronic journals and changes in scholarly article seeking and reading patterns. Aslib Proc.: New Information Perspectives, 2009,61(1):5−32. [doi: 10.1108/00012530910932267]

[2] Strobelt H, Oelke D, Rohrdantz C, Stoffel A, Keim DA, Deussen O. Document cards: A top trumps visualization for documents. IEEE Trans. on Visualization & Computer Graphics, 2009,15(6):1145−1152. [doi: 10.1109/TVCG.2009.139]

[3] Dunne C, Shneiderman B, Gove R, Klavans J, Dorr B. Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. Journal of the American Society for Information Science & Technology, 2012,63(12):2351−2369. [doi: 10.1002/asi.22652]

[4] Hoang VCD, Kan MY. Towards automated related work summarization. In: Aravind KJ, ed. Proc. of the 23rd Int'l Conf. on Computational Linguistics: Posters. Stroudsburg: Association for Computational Linguistics, 2010. 427−435.

[5] Bergstrom P, Whitehead EJ. CircleView: Scalable visualization and navigation of citation networks. In: Proc. of the 2006 Symp. on Interactive Visual Information Collections and Activity IVICA. Texas: Citeseer, 2006. e4806.

[6] Chou JK, Yang CK. PaperVis: Literature review made easy. Computer Graphics Forum, 2011,30(3):721−730. [doi: 10.1111/j.1467-8659.2011.01921.x]

[7] Lehmann S, Schwanecke U, Ralf D. Interactive visualization for opportunistic exploration of large document collections. Information Systems, 2010,35(2):260−269. [doi: 10.1016/j.is.2009.10.004]

[8] Paris C, Wan S. Capturing the user's reading context for tailoring summaries. In: Houben GJ, McCalla G, eds. Proc. of the 17th Int'l Conf. on User Modeling, Adaptation, and Personalization: Formerly UM and AH. Heidelberg: Springer-Verlag, 2009. 337−342. [doi: 10.1007/978-3-642-02247-0_33]

[9] Viegas FB, Wattenberg M. TIMELINES: Tag clouds and the case for vernacular visualization. Interactions, 2008,15(4):49−52. [doi: 10.1145/1374489.1374501]

[10] Viegas FB, Wattenberg M, Feinberg J. Participatory visualization with wordle. IEEE Trans. on Visualization & Computer Graphics, 2009,15(6):1137−1144. [doi: 10.1109/TVCG.2009.171]

[11] Paley WB. TextArc: Showing word frequency and distribution in text. In: Proc. of the Poster at IEEE Symp. on Information Visualization. 2002.

[12] Collins C, Carpendale S, Penn G. DocuBurst: Visualizing document content using language structure. Computer Graphics Forum, 2009,28(3):1039−1046. [doi: 10.1111/j.1467-8659.2009.01439.x]

[13] Stoffel A, Strobelt H, Deussen O, Keim DA. Document thumbnails with variable text scaling. Computer Graphics Forum, 2012, 31(3pt3):1165−1173. [doi: 10.1111/j.1467-8659.2012.03109.x]

[14] Koch S, John M, Worner M, Muller A, Ertl T. VarifocalReader: In-depth visual analysis of large text documents. IEEE Trans. on Visualization and Computer Graphics, 2014,20(12):1723−1732. [doi: 10.1109/TVCG.2014.2346677]

[15] Chen C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for information Science and Technology, 2006,57(3):359−377. [doi: 10.1002/asi.20317]

[16] Schafer U, Kasterka U. Scientific authoring support: A tool to navigate in typed citation graphs. In: Piotrowski M, Mahlow C, eds. Proc. of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids (CL&W 2010). Stroudsburg: Association for Computational Linguistics, 2010. 7−14.

[17] Teufel S. Argumentative zoning: Information extraction from scientific text university of Edinburgh [Ph.D. Thesis]. University of Edinburgh, 1999.

[18] Teufel S, Kan MY. Robust argumentative zoning for sensemaking in scholarly documents. In: Bernardi R, Chambers S, eds. Proc. of the Advanced Language Technologies for Digital Libraries. Heidelberg: Springer-Verlag, 2011. 154−170. [doi: 10.1007/978-3-642-23160-5_10]

[19] Teufel S, Moens M. Summarizing scientific articles: Experiments with relevance and rhetorical status. Computational Linguistics, 2002,28(4):409−445. [doi: 10.1162/089120102762671936]

[20] Mei Q, Zhai C. Generating impact-based summaries for scientific literature. In: Kathleen M, ed. Proc. of the ACL 2008: HLT. Columbus: Association for Computational Linguistics, 2008. 816−824.

[21] Bhaskar P, Nongmeikapam K, Bandyopadhyay S. Keyphrase extraction in scientific articles: A supervised approach. In: Martin K, Christian B, eds. Proc. of the COLING 2012: Demonstration Papers. Mumbai: The COLING 2012 Organizing Committee, 2012. 17−24.

[22] Nguyen TD, Kan MY. Keyphrase extraction in scientific publications. In: Goh DH, Cao TH, eds. Proc. of the 10th Int'l Conf. on Asian Digital Libraries: Looking Back 10 Years and Forging New Frontiers (ICADL 2007). Heidelberg: Springer-Verlag, 2007. 317−326. [doi: 10.1007/978-3-540-77094-7_41]

[23] Nguyen TD, Luong MT. Wingnus: Keyphrase extraction utilizing document logical structure. In: Katrin E, Carlo S, eds. Proc. of the 5th Int'l Workshop on Semantic Evaluation (SemEval 2010). Stroudsburg: Association for Computational Linguistics, 2010. 166−169.

[24] Ribaupierre HD, Falquet G. New trends for reading scientific documents. In: Kazai G, Eickhoff C, eds. Proc. of the 4th ACM Workshop on Online Books, Complementary Social Media and Crowdsourcing. New York: ACM Press, 2011. 19−24. [doi: 10.1145/2064058.2064064]

[25] Nicholas D, Huntington P, Jamali HR, Dobrowolski T. Characterising and evaluating information seeking behaviour in a digital environment: Spotlight on the bouncer. Information Processing & Management, 2007,43(4):1085−1102. [doi: 10.1016/j.ipm.2006.08.007]

[26] ICEpdf. http://www.icesoft.org/java/home.jsf

[27] Noonburg D. Xpdf. 2002. http://www.foolabs.com/xpdf

[28] MEAD. 2006. http://www.summarization.com/mead/

[29] Blei D, Ng A, Jordan M. Latent dirichlet allocation. Journal of Machine Learning Research, 2003,3:993−1022.

张加万(1975－),男,山东蒙阴人,博士,教授,博士生导师,CCF 高级会员,主要研究领域为文化遗产保护与传承信息技术,信息可视化与可视分析,图像与视频处理,图形学,真实感绘制.

杨思琪(1991－),女,硕士,主要研究领域为文本可视分析.

李泽宇(1994－),男,博士生,主要研究领域为可视分析,可视化.

杨伟强(1992－),男,硕士生,主要研究领域为可视化,可视分析.

王锦东(1991－),男,硕士生,主要研究领域为文本数据的可视分析,社交媒体数据可视分析.

贺瑞芳(1979－),女,博士,副教授,博士生导师,CCF 专业会员,主要研究领域为自然语言处理,社会媒体挖掘,机器学习.

黄茂林(1957－),男,博士,教授,博士生导师,主要研究领域为数据可视分析,信息可视化,网形可视化,软件形象化,图表用户界面.