DOI:10.13530/j.cnki.jlis.156007

基于实时新闻分析的馆藏资源推荐方法研究*

陈俊鹏　虞　为

摘　要　如何在信息时代增加馆藏资源的可见度,提高馆藏资源的利用率,是一个急需研究和解决的问题。实时新闻和图书馆馆藏资源间的连接可以提高图书馆馆藏资源的可见度,增加图书馆馆藏资源的利用率,为用户提供丰富、全面的阅读资料和专业知识,帮助用户形成全面、深入阅读和思考的良好习惯。基于快数据处理技术的实时新闻分析和馆藏资源推荐框架,通过分析网络实时新闻获取用户感兴趣的话题,应用快数据处理技术、潜在语义分析、非负矩阵分解、权重矩阵分解等方法对数据进行语义分析和处理,对图书馆馆藏资源进行相关话题的分类和推荐。对 OCLC 的百万数据集和雅虎新闻的分析和实验表明,这种资源推荐框架和方法有较好的应用效果。图 2。表 1。参考文献 18。

关键词　馆藏资源　资源推荐　实时新闻　快数据处理　矩阵分解

分类号　G250

Library Resource Recommendation Based on Analysis on Newswires

CHEN Junpeng & YU Wei

ABSTRACT

With the development of the web, reading is more regarded as a kind of entertainment such as reading twitter or blog than the study with in-depth thoughts. The real-time news, for example, is a kind of popular web-published information which can help people catch the update news from the web. At the same time, the works in library which contain in-depth thoughts and domain knowledge are often overlooked in daily life.

There is a lack of research for providing professional domain knowledge and extending reading list to users who are interested in the special topics mentioned in the real-time newswires. Meanwhile, there is a large scale of domain knowledge and application examples in library collections which can help users have a good understanding for those special topics. Hence, in this paper, we provide a novel method to link the corresponding real-time news and records in the library. The extending reading list from the library can be

* 本文系国家社会科学基金青年项目“基于关联数据的图书馆语义云服务研究”(编号:12CTQ009)和江苏省社会科学基金青年项目“基于语义云服务的数字阅读推广研究”(编号:14TQC003)的研究成果之一。(This article is an outcome of the project “The research of semantic cloud service for library based on linked data” (No. 12CTQ009) supported by Youth Program of National Social Science Foundation of China and the project “The research for digital reading promotion based on semantic cloud service” (No. 14TQC003) supported by Youth Program of Social Science Foundation of Jiangsu Province.)

通信作者:虞为,Email:yuw. nju@ gmail. com,ORCID:0000-0003-1933-5380(Correspondence should be addressed to YU Wei, Email:yuw. nju@ gmail. com, ORCID:0000-0003-1933-5380)

recommended with the technology of natural language processing and semantic analysis.

We recommend the related library records to the users who are interested in the target news. We adopt natural language processing technology and LSA, NMF, and WMF methods to carry out our experiments.

We use the catalogue records corpora: WorldCat-million dataset released by the OCLC in 2012. The dataset contains metadata records of nearly 1.2 million materials most widely held in libraries. The metadata contains approximately 80 million linked data triples, which can help users find the linked resources easily on the web. For the corpus of news articles, we collect the news articles of Yahoo! news from RSS feeds, dated from the 5th of April to 7th of July, 2014, totally 95 days. In order to get an objective observation of the performance, we randomly selected 500 news articles (about 10% of the news articles set) for evaluation. The results are evaluated with TOP10 recall hit rate, from which we can see WMF has better performance than LSA and NMF.

This newswire-library linking offers a number of unique advantages to both libraries and information seekers: the up-to-dateness, the extensive coverage and comprehensiveness, the rich description. Using newswires as a complementary information resource in library catalogues addresses users' information need by offering a vast pool of everyday life subject headings to complement the traditional library vocabularies constructed mainly by experts knowledge.

For future work, we will involve library users in the evaluation of the system and make necessary improvements. 2 figs. 1 tab. 18 refs.

KEY WORDS

Library resources.　Resource recommendation.　Real-time news.　Fast data processing.　Matrix factorization.

0　引言

随着互联网和电子信息的飞速发展，浅阅读、大阅读等快餐式、跳跃式、碎片式的阅读方式开始出现[1]。一部分人将阅读单纯视作一种休闲与娱乐，或者某种纯功利性的行为，往往浅尝辄止，流于表面，缺乏深层次的探索和思考。而另一方面，包含了大量深刻思考和感悟的书籍却在图书馆中被人们所疏远。2010 年 OCLC《图书馆的认知度》[2]研究报告描述了这样一个现实：数字化网络环境中，搜索引擎占据了信息消费者的检索起点。Jaorabchi 和 Mahdi[3]通过研究发现，实时更新的开放型网络资源，如 WikiPedia 等正逐步成为网络用户满足信息需求的新途径。在这种情况下，如何把图书馆馆藏资源有效地推荐给普通读者，使用户在网络浅阅读的潮流中获得深刻的思考和领悟，是图书馆研究和阅读推广工作的当务之急。

为了使图书馆获得更多的用户关注，目前有许多研究致力于将图书馆馆藏资源和外部的网络资源互联。Jaorabchi 和 Mahdi[3]通过寻找共同的主题词将图书馆馆藏信息和 Wikipedia 中的概念词条信息相联系，利用 wikipedia 的用户流量来增加图书馆馆藏资源的可见度和利用率。Golub[4]和 Yi[5]都以受控词表为基础，将图书馆资源和网络分类系统相联系，增加图书馆资源的利用率。夏明春[6]等对国内图书馆资源的共享状况做了调查和研究，并提出建议，图书馆应该根据不同资源的特点，加强资源间的内部联系。这些工作都在一定程度上增强了图书馆馆藏资源的能见度，促进了图书馆馆藏资源的利用和发展。但是，这些研究都是构建网络分类系统、知识库等专家系统和图书馆之间的连接，有很强的针对性，同时也需要用户对这些网络分类系统和知识库有较深的理解和较高的网络搜索技能。

因此,这些将网络资源和图书馆馆藏资源相联的研究存在很强的专业性,不能面向普通的网络搜索用户,在扩展图书馆服务及扩大馆藏资源利用率方面受到很大限制。

实时新闻是一种通过网络信息技术将世界各地及用户身边发生的事件即时发布,供用户浏览、查询、评论的一种新闻播报方式。在网络信息时代,实时新闻以其时效性、丰富性、交互性等特点受到广大网络用户的关注。实时新闻信息量大,产生速度快,在为用户提供大量实时信息的同时也致使用户疲于被动接受,无法进行深入思考和探索。为了帮助用户更有效地理解新闻,Alfonseca[7]提出了一种通过 NOISY-OR 模型提取事件模式的方法,能够根据新闻中的描述信息自动提取概要,帮助人们快速了解新闻内容。Wei[8]等通过对和新闻相关的微博信息进行过滤,自动提取相关新闻中的关键语句,帮助用户快速了解新闻重点。刘晓娟等[9]通过对网络新闻进行分析和处理,对研究热点和发展方向进行分析。这些工作都在一定程度上帮助用户更加快捷有效地理解实时新闻的要点,但没有为用户提供更深入的理解和思考的途径和方法。用户面对海量高速更新的实时新闻,容易陷入盲目、混乱甚至自相矛盾的境地,不利于对实时新闻的深入理解和知识发现。

在目前的实时新闻研究中,缺少针对实时新闻为用户提供主题相关的专业知识和扩展阅读资源的方法的研究。图书馆馆藏资源包含了大量的专业知识和实例分析,能为用户提供充分的分析和专业的知识理解,促使读者进行更广泛、更深入的阅读和思考。因此,本文提出了一种基于实时新闻分析的图书馆馆藏资源推荐方法,通过对实时新闻进行自然语言处理和语义分析,提取新闻的相关主题,再对相似主题的图书馆馆藏资源进行分类和推荐。

1 从实时新闻到图书馆馆藏资源

通过对实时新闻进行语义分析并找出与之相关的馆藏资源进行阅读推荐,可以为用户提供一个由图书馆馆藏资源组成的扩展阅读清单。用户如果对实时新闻的某个主题感兴趣,可以进一步阅读与之相关的图书馆资源,并从图书馆网站下载或到馆借阅这些阅读资料。实时新闻和图书馆间的连接可以极大地提高图书馆馆藏资源的可见度,增加馆藏资源的利用率,同时为用户提供更丰富、全面的阅读资料和专业知识,一定程度上有助于用户逐渐形成全面、深入阅读和思考的良好习惯。

实时新闻和图书馆馆藏资源间的连接有如下几个优点。

第一,提高馆藏资源的信息覆盖率。对于图书馆馆藏资源中所包含的大量理论知识,实时新闻可以为其提供丰富的实例和内容的拓展,能够有效帮助用户理解书本中抽象的理论知识,并提供各种相同的实例类比,方便用户进行比较和分析。

第二,加强馆藏资源的时效性。由于实时新闻的特性,所包含的主题往往带有很强的实时性,通过连接实时新闻,可有效提高图书馆主题信息的时效性,使图书馆资源建设与时俱进。

第三,更丰富的描述信息。实时新闻由于事件发生的即时性,不可能对所发生的事件有非常全面的描述,同时网络技术也对实时新闻的字数长短有较多的限制。因此,读者在阅读实时新闻时可能会因为缺少背景知识或更全面的信息而产生误解,从而做出不正确的判断。图书馆馆藏资源可以提供更丰富的描述信息,帮助用户扩展阅读,较全面地了解事情的来龙去脉。

第四,更丰富的语义信息。实时新闻往往只包括标题、时间、内容等较少的元数据信息,不利于与其他网络资源互联和进行语义分析。图书馆馆藏记录往往包含大量的元数据信息,可以为实时新闻数据提供扩展和补充,方便进行应用程序的开发和多种服务功能的建立。

第五,多语化。针对单一语种的实时新闻,图书馆馆藏记录中可能包括多种语言的馆藏记录,从而实现实时新闻与多种语言的连接,在为

用户提供丰富知识的同时，满足不同语种用户的信息需求。

根据 OCLC 在 2009 年做的一项用户调查[10]，大部分被调查的在线图书馆终端用户希望在图书的元数据信息中添加更多的主题信息，以增加图书馆服务的效果。因此，实时新闻作为对传统的图书馆资源和分类主题的补充，能够增强图书馆主题的时效性，引起读者的阅读兴趣，吸引更多的图书馆用户，有效提高图书馆的服务质量。

虽然实时新闻和图书馆馆藏资源之间的连接可以很大程度上促进图书馆和网络实时资源的整合，但如果仅通过图书馆的人力资源实施手动连接会耗费大量的人力物力。因此，考虑到图书馆的应用需求和实际的成本费用，本文致力于构建一个自动的、基于实时新闻分析的图书馆馆藏资源的推荐方法。

2　实现思路

2.1　应用框架

对实时新闻进行分析，需要根据实时新闻产生速度快、数据处理复杂等特点，设计与之相对应的阅读推荐框架。

基于快数据处理的实时新闻分析和馆藏资源推荐框架可以更好地结合“大”和“快”的特点，提供图书馆阅读服务，增加用户的阅读兴趣，达到阅读推广和知识发现的目的。因此，本文提出了一个基于快数据处理技术的实时新闻分析和馆藏资源推荐框架，如图 1 所示。

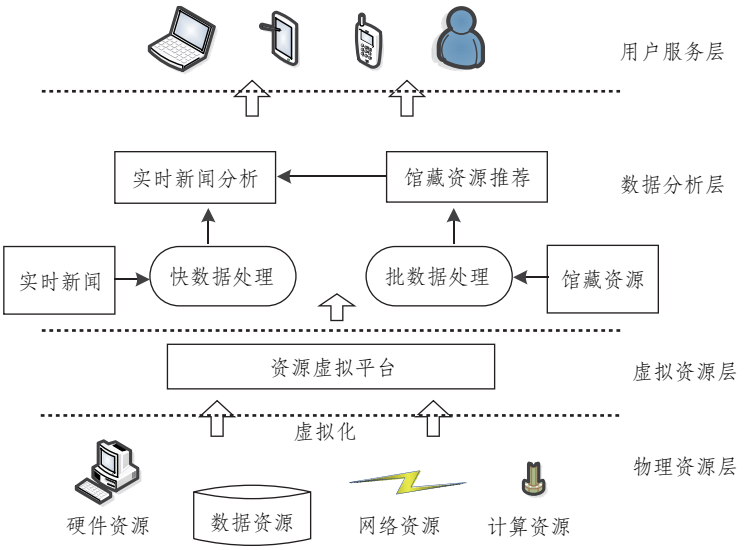

图 1　快数据处理框架图

快数据处理技术是继云计算技术和流计算技术[11]之后，融合对大数据的稳定、精准的延时处理和对实时数据快速可靠的实时处理技术开发的新型的大数据处理框架，目前已被谷歌等主流的网络应用公司所采用，并逐步取代传统的云计算处理平台，成为大数据处理技术的新趋势[12]。新型的大数据处理平台 Spark 是由加州大学伯克利分校开发的一种可伸缩(scalable)的基于内存计算(In-Memory Computing)的数据分析平台，用于构建大规模、低延时的数据分析应用。Spark 比基于云计算的 Hadoop 集群存储方法更有性能优势。Spark 采用 Scala 语言实现，提供单一的数据处理环境。Spark 采用基于内存的分布式数据集，优化了迭代式的工作负

载以及交互式查询。

实时新闻分析和馆藏资源推荐框架主要分为四个部分:物理资源层、虚拟资源层、数据分析层、用户服务层。其中,物理资源层和虚拟资源层是数据分析层、用户服务层的基础,提供所需的资源和计算、存储能力;数据分析层是框架的核心部分,为用户提供基于实时新闻分析的图书馆馆藏资源推荐服务;用户服务层可将不同用户进行细分,满足移动用户、传统 PC 用户、便捷设备用户等不同用户对阅读方式的不同需求。这几个层次层层递进,互为补充,为用户提供全面、高效、快速的阅读推荐服务。

2.2 应用流程

为了实现实时新闻的分析和图书馆馆藏资源的推荐,我们设置了应用任务和具体实施流程,如图 2 所示。应用流程主要分为三个部分:数据的采集和存储、数据分析、数据应用。

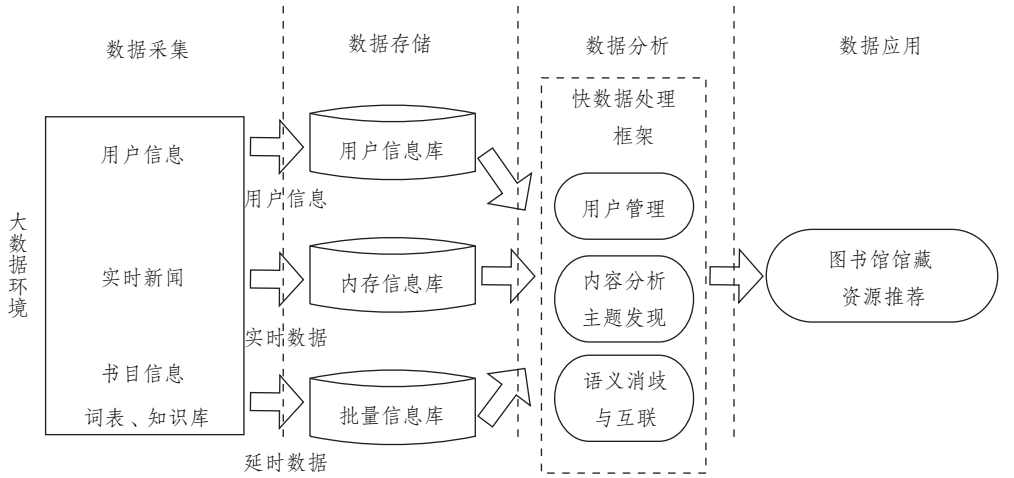

图 2 馆藏资源推荐应用流程

在数据采集阶段,可以从大数据环境中获取实时新闻、用户信息和图书馆馆藏数据(包括书目信息和词表、知识库等)。其中,用户信息可以存入用户信息库,方便对用户进行安全、可靠的管理;实时新闻数据可以存入内存数据库,方便进行快速处理;而图书馆馆藏资源可以存入批量信息数据库,有较好的存储效率和稳定性。

在数据分析阶段,需要对采集到的不同数据进行不同的分析。对于实时新闻,可以通过语义分析和自然语言处理对新闻的内容进行分析,并进行新闻的主题发现和分类工作。图书馆馆藏信息含有丰富的语义信息和链接,因此可以对其进行语义消歧和语义互联,使之能够更好地提供分类和推荐服务。有了数据分析的基础,可以应用面向对象的服务框架(Service Oriented Architecture,SOA)对其网络服务的功能进行构建,为用户提供安全可靠、语义互联的基于实时新闻分析的图书馆馆藏资源推荐服务。

3 技术与应用

3.1 关键技术

快数据处理技术是基础架构,主要是数据的采集和存储,而所用的潜在语义分析、非负矩阵分解、权重矩阵分解等三种机器学习方法,是在做具体的馆藏资源推荐时使用,主要用于提高系统的性能。

(1)快数据处理技术

通过采用 Spark 的快数据处理架构,可以利

用 Spark 框架中的类查询语言(Structured Query Language,SQL)操作指令集对存储数据进行准确高效的调度和操作。应用改进的 MapReduce[13] 并行处理框架对需要全面分析和深入理解的海量延时数据进行处理,通过机器学习方法对存储实时数据的存储块 RDD(Resilient Distributed Dataset)进行调度和运算。从大数据环境中获取数字阅读资料、实时数据信息、用户信息等。根据不同的数据格式和应用目的对获取数据进行存储和预处理。一方面,海量的延时数据可以存放到高容量、可扩展的 NoSQL 数据库[14] 中;另一方面,实时的高速产生的数据以 RDD 的形式存储在内存数据库中。

(2)潜在语义分析

潜在语义分析(Latent Semantic Analysis,LSA)[15] 通过构建向量语义空间来提取文档与词中的概念,从而分析文档与词之间的关系。LSA 的基本假设是,如果两个词在同一文档中多次出现,则这两个词在语义上具有相似性。LSA 使用大量的文本构建一个矩阵,这个矩阵的一行代表一个词,一列代表一个文档,矩阵元素代表该词在该文档中出现的次数,然后在此矩阵上使用奇异值分解(Singular Value Decomposition,SVD)减少矩阵行数。每两个词语的相似性则可以通过行向量点积来进行表示,值越接近于 1 则说明两个词语越相似。

具体实现过程是首先将文档集构造成文本—词语矩阵 M,矩阵中每个位置的值可以是该行代表的词在该列代表的文档中的词频、TF/IDF 值等。然后,需要对文本—词语矩阵进行奇异值分解,此时 M = U * S * V 的转置。再对 SVD 分解后的矩阵进行降维,只保留矩阵 S 前 K 个最大的奇异值得到 S'。相应的 U、V 分别为 U'、V'。V'中的每行即为每个文档在潜在语义空间上的 K 维表示。使用降维后的矩阵重建文本—词语矩阵 M' = U' * S' * V'的转置。对于一个列向量表示的新文档 Q,其在潜在语义空间上的 K 维表示为 $Q' = Q^T * U' * S'$ 的逆。将新文档 Q 和文档集中的每个文档在潜在语义空间进行相似度计算,得到与 Q 最相似的文档。

(3)非负矩阵分解

非负矩阵分解(Non-negative Matrix Factorization,NMF)[16] 也通过矩阵分解的方法来达到相似度计算和文本分类的目的,不同的是,NMF 分解矩阵后的结果不能为负值。NMF 的思想是,虽然从数学上看,分解结果中存在负值是正确的,但负值元素在实际问题中往往是没有意义的。NMF 在图像处理、单词统计、股价估算中都有很好的应用效果。

NMF 算法流程是,对于任意给定的一个非负矩阵 A,NMF 算法能够寻找到一个非负矩阵 U 和一个非负矩阵 V,使得满足 A=U * V。从而将一个非负的矩阵分解为左右两个非负矩阵的乘积。由于分解前后的矩阵中仅包含非负的元素,因此,原矩阵 A 中的一列向量可以解释为对左矩阵 U 中所有列向量(称为基向量)的加权和,而权重系数为右矩阵 V 中对应列向量中的元素。非负矩阵分解是个 NP(Non-deterministic Polynomial)问题,可以划为优化问题用迭代方法交替求解 U 和 V。NMF 算法提供了基于简单迭代的求解 U 和 V 的方法,求解方法具有收敛速度快、左右非负矩阵存储空间小的特点,它能将高维的数据矩阵降维处理。

(4)权重矩阵分解

权重矩阵分解(Weighted Matrix Factorization,WMF)[17] 通过对矩阵中不为零的特征向量加权来强化非零值在矩阵分解中的重要性。对于一些短文本信息,所构成的文本—词语矩阵往往非常稀疏,对矩阵分解的精确度造成极大的影响。通过加权对非零的矩阵单位进行强化,不影响整个稀疏矩阵特征值为零的特点,但能够更好地对矩阵进行分解,找到不同文本和词语间的潜在语义关系。

WMF 算法的流程是任意给定的一个矩阵 X,$X=P^TQ$,其中 P 是一个 D×M 的矩阵,而 Q 是一个 D×N 的矩阵。WMF 模型的参数通过目标函数进行优化,其中 λ 是一个自由参数,而 W 定

义了 X 中每个单元的权重，$\sum_i \sum_j W_{ij}(P_{.,i} \cdot Q_{.,j} - X_{ij})^2 + \lambda \| P \|_2^2 + \lambda \| Q \|_2^2$。P 和 Q 中的初值可以随意选定，然后通过优化公式进行迭代，使得目标函数取值最小，并在此条件下得到分解的 P 和 Q 矩阵。其中，W 可以使矩阵中的零值和非零值差异化，可根据应用自行确定取值。

3.2 应用方法

(1)主题发现和关联

通过应用 3.1 章中的文本分析和矩阵分解方法，可以把书目数据和新闻数据表示成 D 维文本向量，每个向量之间可以计算相似度。算出每条新闻数据和所有书目数据的相似度，通过相似度对书目数据进行排序，取和此条新闻最相似的前 T 个书目，每一个书目数据都有受控词表定义的关键词或主题词。通过统计前 T 个书目中出现频率最高的前 N 个主题词作为该新闻的相关主题，通过受控词表对实时新闻中的主题给出具有权威性的主题发现和检测。同时，每一个相关主题可以通过这个主题和其他书目数据间的关联关系找到与该相关主题关联的所有书目，满足用户的扩展阅读需求。在 OCLC 中，通过关联数据定义的每个主题都可以关联到与之相关的书目关联数据中，用户可以通过感兴趣的主题词找到相关的书目。

(2)语义消歧和互联

通过主题发现和关联关系，可以对实时新闻中不确定的主题和语义关系进行消歧和互联。比如说，在实验中我们遇到这样一个案例，在一段实时新闻中出现了“spelling bee”这种非英语母语用户所不熟悉的说法。一般情况下，阅读者倾向于认为这个词表示与“蜜蜂”相关的含义。但是在主题发现的过程中，我们发现和此条新闻相关的主题是“spelling ability”，因此，这个词的意思应该和“拼读能力”相关，而不是直译为“蜜蜂”。通过进一步的检索，我们发现“spelling bee”指的是一种英语拼读比赛。因此，通过和相似的主题相关联，可以起到一定的语义消歧和语义互联的作用，帮助用户更好地理解新闻数据。语义消歧和互联的应用原理是，由于书目数据中的主题是由专家确定的具有权威性的受控词表，因此能够对实时新闻中形式自由的主题和词语起到一定的归类和权威化的作用，并把这些相关的主题和词语同受控词表以及其他的图书馆资源联系起来。

(3)分类和推荐

通过以上的主题发现和互联，以及语义消歧，我们可以针对每条新闻得到与之相关的主题和相关书目，从而进行书目数据和主题信息的分类。通过对不同的主题进行分类，可以有针对性地对用户进行书目数据及相关图书馆资源的推荐，引起用户的阅读兴趣，达到阅读推广的目的。

4 实验与讨论

4.1 实验评估

为了验证基于实时新闻分析的图书馆馆藏资源推荐的可行性，实现语义分析和文本分类的技术和方法，本节针对实时新闻和馆藏资源的特征进行分析和实验。

选取 OCLC 在 2012 年发布的 WorldCat 百万数据集[18]作为实验数据。这个数据集以关联数据的形式对 WorldCat 的书目数据进行公开发布，使它们转变成关联开放数据的一部分。只要有其他的应用或者数据集和这些关联开放数据相联，就可以通过数据中所包含的关联关系和其他的数据集、知识库等网络资源互联，方便用户更高效、准确地获取信息，从而促进知识发现。

WorldCat 百万数据集共包括 120 万条图书馆书目数据，8 000 万个 RDF 三元组。每一条书目数据包含大量的属性信息，包括书的作者、书名、出版社、体裁、书目所属数据库、该书的内容描述、主题、该书存储的物理媒介、文本格式、ISBN 号、OCLC 编号等。这些属性信息为应用开发提供了丰富的分析资源。其中，“该书的内

容描述”属性包括书的摘要、总结、内容描述信息等内容，是读者了解该书内容，判断是否深入阅读的重要途径。OCLC 采用多种分类方法对书目进行标引，包括杜威十进制分类法（Dewey）、美国国会图书馆标题表（LCSH）、主题分类词表（FAST）等，帮助用户和图书馆员更准确、全面地对书目数据进行分类和编目。

WorldCat 百万数据集中的每一条书目数据，根据同一个 ID 号“http://www.worldcat.org/oclc/99998687”可以搜索到相关的书目属性，而这些数据都对应着 WorldCat 中的同一条书目数据。本文选取 WorldCat 百万数据集中的“标题”“该书的内容描述”，以及主题分类信息，作为实验中训练数据模型的训练集，因此去掉数据集中不包含这些属性的条目，一共得到 217 147 条书目记录。其中，1 000 条作为测试样本，余下 216 147条书目记录作为训练集。为了更好地进行语义分析，我们把书目记录中的“该书的内容描述”属性和“标题”属性合并，形成一段短文本。通过对所有训练集中的短文本进行预处理、词干化、去噪处理，以及词性标注，一共获得 67 932个单词。我们使用 TF-IDF 量度作为文本相似度的测量方法。

为了进行实时新闻分析，我们通过雅虎新闻的 RSS 种子进行实时新闻的采集，时间从 2014 年 4 月到 7 月。通过对采集的新闻进行去重，一共得到 4 702 条新闻数据集。我们使用训练数据样本集对在章节 2 和章节 3 中提出并搭建的系统进行参数训练后，利用这4 702条新闻模拟实时的新闻分析环境，向训练后的系统提交新闻数据，通过系统自动分析，进行馆藏资源的推荐。新闻数据集中的每条数据都包含“标题”“内容”“时间戳”“网址”四条属性。在实验中我们主要用到新闻的“标题”和“内容”属性，并把这两个属性合并成一段短文本提交给系统进行分析。

为了评估实验系统的性能，我们从 4 702 条新闻数据及其系统返回结果中随机选取 500 条数据和相关结果，请图书馆的三位专业人员对数据进行评估。采用计算主题词前 10 个搜索结果的命中率的方法（TOP10 Recall Hit Rate）。对于一个提交给系统的目标新闻，系统为其推荐相关的书目数据，取排在前面的最相似的 50 条书目记录，并把这些记录的主题词集合起来。对这 50 条书目记录中所有主题词出现的频率进行排序，取排在前 10 位的主题词，并对这 10 个主题词进行评估，如果包含了和提交的目标新闻相关的主题，就认为系统在用户可以读到的范围之内找到了和系统提交的目标新闻相关的书目信息。这种对书目数据的评估方法可以把对大段的书目描述信息的评估转化为对书目的主题词的分类和评估，从而避免了阅读大量非结构化文本所带来的歧义信息和评估者之间产生的主观分歧。由于主题词是由图书馆专家制定的受控词表，并有各级图书管理机构和出版单位分配给相关书目作为分类和标识的有效信息，故具有一定的权威性，并较少含有歧义，从而能够客观有效地对书目数据推荐的准确性进行评估。

我们对三种书目推荐方法进行评估，结果显示：WMF 的搜索结果命中率最好，达到了 70.4%；其次是 LSA，为 45.4%；NMF 的评估性能较差，为 31.4%。以上评估效果显示，WMF 更适用于新闻文本和馆藏书目记录这种短文本的分类环境。

4.2　分析和讨论

基于图书馆馆藏资源推荐结果，我们对系统进行分析，部分数据如表 1 所示。表 1 中随机选取了实验结果中的三个实例，每个实例分为三个部分，分别是：实时新闻，包括标题和内容；关键词，取前五个作为样例；相关书目，取前五个作为样例。

表 1 实验结果分析(部分)

新闻数据	关键词	相关馆藏书目数据
Title:Employees suspended following Mo. school rape allegations Contents:Months after vowing to boost security at a Kansas City school where a student says she was dragged to a room and raped,district officials have suspended three employees amid new allegations from a 14-year-old girl who alleges a boy repeatedly raped her at school.	①Rape—United States—Trial practice ②Rape victims—Legal status, laws,etc. —United States ③Date rape—United States ④Victims of crimes—United States ⑤Acquaintance rape—United States	①The criminal justice and community response to rape ②Punishing violence ③Crimes of violence ④Against our will:men,women and rape ⑤Confronting rape: the feminist anti-rape movement and the state
Title: States rebel against powerful new painkiller Contents:Officials fear pill that hit market last month will deepen prescription drug abuse crisis.	①Drug abuse ②Substance—Related Disorders ③Drugs of abuse—Law and legislation—United States—Criminal provisions ④Psychotropic drugs ⑤Drug addiction—Treatment	①Drugs and drug abuse ②From chocolate to morphine:everything you need to know about mind-altering drugs ③Drugs:a very short introduction ④The consumer's guide to drug interactions ⑤Prescription drugs
Title:So what? U. S. regains jobs lost in the recession Contents:Analysts downplay milestone as economy is still millions of jobs short of goal.	①Job hunting ②Unemployment—United States ③Employment interviewing ④Vocational guidance ⑤Labor market—United States	①The job description handbook ②The job search solution:the ultimate system for finding a great job now! ③The job-hunter's survival guide: how to find hope and rewarding work even when there are no jobs ④Everything you need to know about getting a job ⑤The coming jobs war: what every leader must know about the future of job creation

第一个实例的新闻是关于一起美国的校园强奸案,关键词包括强奸受害者、约会强奸、暴力受害者等。与新闻相关的书目包括《罪案中的正义和社区对强奸的态度》《对暴力的惩戒》《违背我们的意愿:男人、女人和强奸》《面对强奸:女性反强奸活动和现状》等。通过这些书目,我们可以对这起罪案发生的社会背景、法律救援措施、受害者的应对措施等有一个更加深入的思索和探讨,也可以告诉读者一些防卫措施和法律条例。

第二个实例的新闻是美国政府禁止销售新的止痛药,怕引起药物滥用。关键词包括药物滥用、药物引起的不良反应、药物滥用相关的法律法规、药物依赖的治疗等。相关的书目数据包括《药物和药物滥用》《从巧克力到吗啡:你所需要知道的对神经产生改变的药物》《药物的简短介绍》《药物互作用的消费者指引》《处方药物》。这个扩展阅读书单主要从药物的机理、法律法规、消费者指引方面为读者提供了一系列专业的和深入的阅读书籍,使读者可以获得与新闻相关的更多专业方面的知识。

第三个实例的新闻是美国经济尚有百万工作职位的缺口。关键词包括美国的劳务市场、美国的失业、找工作、工作面试、职业指南等。相关的书目有《职业描述手册》《找到一个好工作的终

极方法》《找工作者的生存指导:如何找一份有希望和回报的工作,即使现在没有工作》《关于找工作所需要知道的每件事》《即将到来的工作之战:每位领导人都应该知道的就业前景》。读者通过这份馆藏资源推荐单,可以方便地获取高失业率情况下寻找工作的方法,或者是为了增加就业机会、提高就业率政府应该采取的有效措施。

通过样例分析我们可以发现,本文构建的馆藏资源推荐系统可以通过对简短的实时新闻进行分析,有效扩充用户感兴趣的主题和内容。通过推荐馆藏资源从各个不同的角度对与新闻主题相关的主题和内容进行深度的阐述,可以扩展用户的视野,深化用户对相关主题的理解,甚至为用户提供解决问题的方法和途径。

5　结语

基于实时新闻的兴趣主题分析是近几年网络分析应用研究的一个重要领域。目前,相关研究已经在舆情分析、股票价格预测、商品推荐、用户兴趣分析等多个领域进行有益的尝试和探索,但是在图书馆馆藏资源推荐方面的应用还较少。本文认为,基于实时新闻的图书馆馆藏资源推荐可以提高图书馆资源的能见度,对实时新闻分析和图书馆馆藏资源推荐都是很好的尝试。本文通过对 OCLC 的百万数据集和雅虎的实时新闻两者之间相关的话题进行分析和匹配,验证了实时新闻和图书馆馆藏资源之间可以互相补充,互为扩展。如何将实时新闻和图书馆馆藏资源有机融合,实现一个语义互联的整体,并实现多语言的实时新闻分析和图书馆馆藏资源推荐,是今后的研究方向。在后续研究中,我们还将针对图书馆用户进行资源推荐服务的应用调查和评估,以便更好地对系统进行评估和改进。

参考文献

[1] 任福兵. 微时代浅阅读对网络信息危机生成的影响机制[J]. 情报理论与实践,2013,36(4):53-58. (Ren Fubing. The influence on the generation of network information crisis caused by the masses' reading in the micro-era[J]. Information Studies:Theory and Application,2013,36(4):53-58.)

[2] Perceptions of libraries, 2010: context and community [EB/OL]. [2015-10-20]. http://www.oclc.org/zhcn-asiapacific/reports/2010perceptions.

[3] Joorabchi A, Mahdi A E. Towards linking libraires and Wikipedia: automatic subject indexing of library records with Wikipedia concepts[J]. Journal of Information Science,2014,40(2):211-221.

[4] Golub K. Automated subject classification of textual web pages, based on a controlled vocabulary: challenges and recommendations[J]. New Review of Hypermedia and Multimedia 2006,12(1):11-27.

[5] Yi K. Automated text classification using library classification schemes: trends, issues, and challenges[J]. International Cataloguingand Bibliographic Control,2007,36(4):78-82.

[6] 夏明春,强切云. 我国高校图书馆资源整合的现状:调查与建议[J]. 大学图书馆学报,2008 (1):39-44. (Xia Mingchun, Qiang Qieyun. Survey and proposal on academic library resources integration in China[J]. Journal of Academic Libraries,2008(1):39-44.)

[7] Alfonseca E, Pighin D, Garrido G. HEADY: news headline abstraction through event pattern clustering[C/OL]// Proceedings of ACL 2013, the 51st of Annual Meeting of Association of Computational linguistics,2013[2015-07-15]. http://www. aclweb. org/anthology/P13-1122.

[8] Wei Z Y, Gao W. Utilizing microblogs for automatic news highlights extraction[EB/OL]. [2015-09-14]. http://www. hlt. utdallas. edu/~zywei/paper/676-wei-coling2014-slides. pdf.

[9] 刘晓娟,王凌云. 面向社科领域的网络新闻分析与监测[J]. 情报科学,2011,29(10):1569-1574. (Liu Xiaojuan, Wang Lingyun. Analysis and monitoring of web news for the social sciences[J]. Information Science, 2011,29(10):1569-1574.)

[10] Calhoun K, Cantrell J, Gallagher M, et al. Online catalogs: what users and librarians want: an OCLC report[R]. Dublin: OCLC, 2009.

[11] 孙大为,张广艳,郑纬民. 大数据流式计算:关键技术及系统实例[J]. 软件学报,2014 (4):839-862. (Sun Dawei, Zhang Guangyan, Zheng Weimin. Big data stream computing: technologies and instances[J]. Journal of Software, 2014 (4):839-862.)

[12] Zaharia M, Chowdhury M, Das T, et al. Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing, NSDI 2012[EB/OL]. [2015-06-30]. http://www.eecs.berkeley.edu/Pubs/TechRpts/2011/EECS-2011-82.pdf.

[13] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters[C/OL]//Proceedings of the 6th OSDI, 2004[2015-06-30]. http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en/us/archive/mapreduce-osdi04.pdf.

[14] Chang F, Dean J, Ghemawat S, et al. Big table: a distributed storage system for structured data[C/OL]//Proceedings of the 6th OSDI, 2004[2015-06-30]. http://courses.cse.tamu.edu/caverlee/csce438/readings/bigtable.pdf.

[15] Mihalcea R, Corley C, Strapparava C. Corpus-based and knowledge-based measures of text semantic similarity[C/OL]//The 21st National Conference on Articial Intelligence, 2006[2015-06-30]. http://www.aaai.org/Papers/AAAI/2006/AAAI06-123.pdf.

[16] Chen J P, Feng S, Liu J. Topic sense induction from social tags based on non-negative matrix factorization[J]. Information Sciences, 2014(280):16-25.

[17] Stajner T, Thomee B, Popescu A, et al. Automatic selection of social media responses to news[C/OL]//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013[2015-06-30]. http://www.liacs.nl/~bthomee/assets/13social_p.pdf.

[18] Murphy B. OCLC provides downloadable linked data file for the 1 million most widely held works in WorldCat[EB/OL]. [2014-06-30]. http://www.oclc.org/news/releases/2012/201252.en.html.

陈俊鹏 南京财经大学信息工程学院讲师。江苏 南京 210046。

虞 为 南京大学信息管理学院副教授。江苏 南京 210046。

(收稿日期:2015-06-03;修回日期:2015-08-02)