Keyword Cloud在文献检索中的应用研究

廖　凤[1,2]　张建勇[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　100190)

文　摘　理论部分对 Keyword Cloud的来源、概念、功能以及在图书馆服务中的需求分析进行介绍,为实际应用奠定理论基础。实践部分将 Keyword Cloud应用于文献数据库检索服务中,用于汇总检索结果和辅助二次检索;同时引入 Tag Line技术为传统 Keyword Cloud增加时间框架,便于用户观察热点趋势变化;通过用户调查对 Keyword Cloud的实用性和适用性进行评价;总结下一步工作的重点在于关键词语义关系的构建。

关键词　关键词云图　标签云图　标签线图　关键词检索　二次检索

Keyword Cloud and its Application in Document Retrieval

Liao Feng[1,2]　Zhang Jianyong[1]

(1National Science Library, Chinese Academy of Sciences, Beijing 100190)
(2Graduate University of Chinese Academy of Sciences, Beijing 100190)

Abstract: The theoretical part is an overview of the keyword cloud, including origin, concept, function and its demand analysis in library services, which lay the foundation for the application part. The practical part is an application of keyword cloud in document retrieval system, where keyword cloud is used to summarize the retrieval results and accelerate the second retrieval. In addition, we add a time frame named tag line for the traditional keyword cloud, through which users can observe the trends of the hot topic. Then we carry out a user investigation in order to evaluate the suitability and practicability of this application. At last, we make a conclusion that the future research will focus on the semantic relation construction of keywords.

Key words: Keyword cloud; Tag cloud; Tag line; Keyword search; Refine search

在传统文献检索中,关键词既有描述和揭示文章主题的作用,也能够提供检索点,成为用户常用的检索入口之一。一般情况下,检索系统只在特定文章层次为用户提供关键词浏览,却很少有关注文章集合层次关键词的特点和功能。本文将大众标注系统(Folksonomy)流行的 Tag Cloud呈现方法引入文献数据库检索中,利用检索结果的关键词集合生成 Keyword Cloud,以期为用户提供可视化的浏览、检索和主题分析功能。

1　Keyword Cloud概述

1.1　源起 Tag Cloud

Web2.0环境下,大众标注的理念得以迅速推广,为了提高网络资源的发现和共享效率,需要将大众分类法的标签以某种方式展示,供用户浏览。Tag Cloud,可译为标签云图,是目前普遍使用的对标签的可视化组织和表现方式。标签云图中的标签通常是单词,一般按字母顺序排列,标签的重要性(权重)通过字体大小或者颜色来标示,这就使得通过字母顺序和重要程度查找标签成为可能。Tag Cloud中的标签一般都具备超链接,关联到被该标签所标注的一组对象。一个 Tag Cloud一般拥有 30到 150个标签[1]。Tag Cloud的实现主要依靠内嵌 HTML元素。

Tag Cloud不仅广泛应用于大众标注网站,其理

念和技术还被推广应用于展示非标签类型数据(Non-Tag Data),由此产生了其他类型云图[2]。它们显示原理与Tag Cloud相似,只是将标签集合替换为其他类型的数据单元集合。常见的有数据云图(Data Cloud)、文本云图(Text Cloud/Word Cloud)、搭配云图(Collocate Cloud)。

Keyword Cloud是文本云图的一种,是关键词集合以标签云图的呈现方式。之所以选择这种呈现方式,是因为关键词与标签之间的一些共同特征:①都属于自然语言范畴,是未经加工、规范的语词,源于用户或者作者自由标注,使用起来比较自由。②标签是用户对资源属性、特征或功能描述的元数据,关键词是篇名、文摘、正文中对揭示和描述文献主题内容具有实质意义的语词,因此它们都能够起到描述和揭示资源对象内容的作用。③不论是标签还是关键词,都能为用户查找资源提供检索入口,关联到包含该标签或关键词的一组资源。

1.2 Keyword Cloud功能

Keyword Cloud的形成需要根据权重算法计算各个关键词的权重,然后设计显示方式和排序方式,将关键词集合呈现出来供用户浏览。尽管形式简单,但笔者认为可以用"具备超链接的词汇摘要"来概括关键词云图的功能。具体来说,分以下几个方面:

①内容概览。关键词是对文章内容的深度揭示,因此关键词云图是一组文章集合内容的浓缩。通过浏览云图,用户可以获得对文献资源主题内容的大致了解,这是一种快捷而高效的词汇摘要。

②资源定位。由于关键词云图中的关键词是具备超链接的,点击其中任何一个就可以跳转到包含该关键词的一组文献资源,为用户提供了准确的内容定位。

③专题导航。关键词云图按关键词的重要性设置不同的显示特征,权重较大的关键词要么字体较大,要么颜色突出,在视觉效果上能够首先吸引用户的注意。通过浏览关键词云图,用户能够很快捕捉到热门关键词和重点关键词,便于进行话题导航。

④挖掘潜在需求。通过检索某一主题得到的关键词云图,除了涵盖用户已知的关键词外,同时包含了同一主题下用户不知道的其他关键词,用户可以利用这些关键词扩展查询。因此,关键词云图有帮助用户挖掘潜在需求的功能。

综上所述,关键词云图既是个性化的索引,因为它能够为用户指引同类信息的所在;也承担了文摘的功能,因为它从词汇角度揭示原文内容,是原文信息的浓缩。

1.3 Keyword Cloud在图书馆服务中的需求分析

传统的检索中,用户根据自己的信息需求,利用系统提供的关键词检索入口,输入自己选定的关键词,系统按照用户的查询指令查找符合条件的对应内容,并把检索结果组织起来提供给用户。相比于传统的关键词检索和结果展现方式,关键词云图体现了一种新的服务理念和服务方式,有着重要的应用价值:

①可视化服务:关键词云图的特点在于直观,用户可以根据字体大小或者颜色深浅很快地发现重点和热点。直观便捷,这符合用户使用服务的最省力原则,易为用户接受。

②个性化服务:任何形式的文献集合,都可以产生相应的关键词云图。以用户收藏的文献为例,不同用户有着不同的关键词云图。该云图不仅汇总了用户的研究主题和关注重点,并且可以帮助用户进行文献管理和内容查找,是一种个性化的服务工具。

③深层次服务:关键词云图体现了一种更深入的服务模式:在内容维度上,可以帮助用户全面分析特定主题领域的文献信息,概览体现的是广度,关键词细化体现的是深度;在时间维度上,可以帮助用户了解研究重点随时间的变化趋势。

上文对Keyword Cloud概念、功能和应用需求进行了简要分析,下面将在实际系统环境中将关键词云图付诸应用,验证并分析其应用效果。

2 Keyword Cloud在文献检索中的应用

2.1 应用背景

Keyword Cloud对于非结构化数据具有良好的导航和汇总功能,由于其突出强调了重要概念,使得人们可以很快通过浏览获得概要信息。因此,本研究将Keyword Cloud应用于国际西文引文数据库的检索服务中:从用户的检索结果中抽取权重符合一定标准的关键词制作关键词云图,用以汇总此次的检索结果;通过云图中带链接的关键词,用户可以跳转到相关主题实现二次检索。考虑到传统Tag Cloud缺少时间框架,在Keyword Cloud中加入了时间控件,可以按年显示关键词云图。

利用云图汇总检索结果的相关研究有:PubCloud[4]使用Tag Cloud汇总从PubMed数据库中检索出的生物医学文献结果,其标签集合是从查询结果记录的文摘中提取而成的;Tag Cloud展示汇总关键词的功能也被应用到Email中[5];CourseCloud[6]

通过标签云图汇总检索结果, 方便用户重定义检索关键词, 获得更深入更多样化的结果。

2.2 系统流程结构

Keyword Cloud系统结构分三层, 如下图所示。

用户界面层: 负责与用户的交互。接收用户的查询请求, 将查询结果和关键词云图以特定的格式呈现给用户, 供用户浏览和检索。

逻辑处理层: 负责逻辑功能实现。接收用户查询参数, 构造为数据库可以执行的 SQL 语句, 发送至数据库服务器查询; 接收数据库返回的查询结果集, 将其按一定格式显示到用户界面; 同时从查询结果中抽取关键词, 分年份计算权重, 挑选在关键词云图中显示的关键词, 再根据标签云的显示技术将其呈现到用户界面。

数据层: 负责数据存储以及底层数据查询。需要响应逻辑处理层的查询请求, 返回查询结果。数据库中数据按关系模式存储。

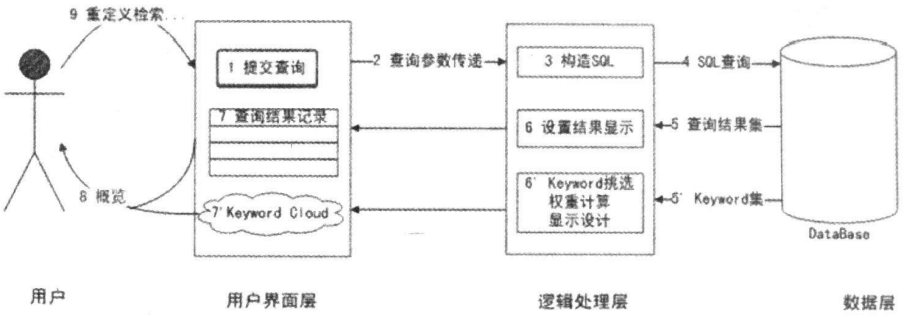

图 1 **Keyword Cloud**系统结构流程图

2.3 时间框架

传统云图能够为用户提供信息概览, 但却缺乏时间框架。时间框架之所以重要, 因为 Keyword Cloud一般根据频次来选择显示的关键词。由于出版时间不一致, 这种选择方法让旧关键词可以通过时间累积频次, 而部分新关键词固然重要但由于使用频次低而无法显示。另一方面, 这种 Keyword Cloud也不便于观察关键词随时间的变化趋势。

为了给 Keyword Cloud增加时间框架, 我们引入了 Tag Line[7]。Tag Line是目前最典型的带时间维度的标签云, 它允许用户选择查看特定时间段内的热门标签集合, 直观呈现了热点变化趋势。Tag Line是 2006年由 Dubinko等提出的概念[8]。他们的项目目标在于观察 Flickr网站上流行标签的变化历程。在他们的 Tag Line中, 用户可以观察到 2004年6月到 2005年 9月这个时间段内的标签云图, 拖动时间轴滑块可以查看任一个时间点的图片以及相应的热门标签集。

本研究将 Tag Line技术应用 Keyword Cloud中, 将时间因素纳入权重计算方法中。

2.4 关键词权重计算

一般情况下直接用频数 TF表示权重, 但是存在几个问题: 当关键词 TF相同的时候如何进行权重区分? 如何消除时间累积效应, 将最新且有代表性的关键词展示出来? 如何准确表示关键词列表长度与关键词权重的关系?

综合上述问题, 设关键词 Ki(i= 1, 2, …, n, n为关键词总数), 则 Ki的权重计算公式为:

$$W = TF^{*} \sum_{1}^{m} 1/aj$$

如果 Ki出现在 title中, 则 TF= TF+ k, k表示在 title中出现过的次数

其中, TF代表关键词 Ki出现的频数, TF越大说明该关键词被使用的越频繁, 越能反映该检索主题的核心内容。m代表包含 Ki的结果记录数目, aj表示各记录拥有的关键词数目, j是记录编号。我们认为记录包含的关键词越多, 那么 Ki在描述对应资源时的作用就越小, 或者说有更多的关键词协助揭示主题内容, Ki不再是独当一面, 因此权重均衡下调。这是假设各个关键词地位等同, 但事实上有的关键词确实是举足轻重的, 不论这篇文章有多少个关键词, 都不会影响其重要性。经验表明这类关键词一般会出现在题名中, 因此对于这种关键词频数会相应累加。为了消除时间累积效应的影响, 在不同的时间段内分别计算权重。即将所有的关键词按年分组, 在每年的关键词集合中分别计算权重、筛选和显示, 再通过时间轴控件将各年的云图联系起来, 形成一个完整的关键词云图。

对于关键词的筛选标准, 我们将阈值设置为权重最大值的 10%, 在这个范围内的关键词可以在 Keyword Cloud中显示。

2.5 结果展示

下图展示了用户输入某个检索词之后的检索结果界面,左边是常规的检索结果列表,右边是关键词云图。关键词按字母顺序排列;权重以字体大小区分;为了让字体大小区分更明显,设计了不同的颜色层次;拖动时间轴,可以查看各年的关键词云图;每个关键词都是可链接的,点击进入相当于在当前结果范围内输入该关键词进行二次检索。此外,关键词云图的数据源是左边的检索结果关键词集合,因此与左边的检索列表是保持同步动态更新的。

图2 检索结果及相应的Keyword Cloud展示

2.6 意义及评价

任何一种新技术或者新思想,只有用户认可,方能成就其价值。为了评价Keyword Cloud的可用性,设计了一个简单的性能评价实验。

方法过程:首先采用嵌入式网络问卷调查,再结合统计结果进行用户访谈。设定检索主题为“lymphoma/淋巴癌”和“inflammation caused by Helicobacter/螺旋杆菌引发的炎症”,问卷包含5个选择题和1个填空题,调查内容:相比于传统检索,Keyword Cloud是否能够帮助确定更准确的检索词,是否能够帮助全面了解检索主题,是否能够帮助了解热门主题的变化趋势,是否提高了检索效率,是否喜欢此类可视化工具,从输入检索式到获得满意结果所花费的时间。

调查对象:选择50名用户进行调查。选择标准:使用过Keyword Cloud和Keyword Search两个界面进行检索;对检索主题相关领域熟悉程度一般且一致。

结果分析:回收有效问卷46份。用户反馈分析如下:

解决问题的准确度和时间耗费

对于比较简单的问题,Keyword Cloud的答案比Keyword Search要准确;但是当面对概念需要组合的问题时,Keyword Cloud就显得无能为力了,因为它不能引导用户走得更广。同时,时间统计表明,使用Keyword Cloud的时间耗费要高于Keyword Search,因为云图的浏览和关键词选择比较耗时。

汇总和辅助检索功能

多数用户认为Keyword Cloud的汇总功能帮助用户获得了对检索课题的全面理解,挖掘出了相关主题下用户不知道的其他关键词。这样首先可以通过关键词链接将之前淹没在众多结果记录中对用户有用的记录发掘出来;其次可以帮助用户调优检索式,比如通过浏览可以选择更精准的检索词;再者也有可能激发用户的潜在需求,即用户没有意识到或者没有表达出来的需求。所以,69%的用户认为Keyword Cloud从整体上提高了检索效率,在辅助检索方面是比较有价值的。

帮助分析热点变化的功能

由于Keyword Cloud中加入了时间框架,用户可以拖动时间轴观察特定检索主题下各年的主要关键词,从而可以对该领域的发展情况和变化趋势进行比较分析,既可以掌握较全面的内容亦可以捕捉到前沿信息,这也是让用户比较满意的。

3 下一步工作及总结

3.1 进一步工作:关键词语义关系构建

上述实验展现了如何用关键词云图汇总检索结果。但是由于关键词是自然语言范畴,不可避免地存在同义、近义、多义的问题,很大程度上影响着用户的检全率和检准率。如果关键词能够和主题词一样拥有规范的语义关系结构(上位词、下位词等),那么上述问题就可以很好地解决。因此,关键在于如何发掘关键词之间的基本语义关系。

基于共现的聚类是解决 Keyword Cloud中语义关联缺失的常用方法。这需要计算关键词相似度,衡量关键词相似度的基础是关键词共现次数。关键词共现指两个关键词被赋予同一篇文献的次数,共现次数越高,说明这两个关键词之间的相关性越高。关键词的共现相关系数 RC定义如下:[9]

$$RC(A, B) = \frac{|A \cap B|}{|A \cup B|}$$ [9]

其中 A和 B是两个关键词所描述的文献资源集合; $|A \cap B|$表示两个关键词共同描述的文献数目,即两个关键词的共现次数; $|A \cup B|$表示两个关键词标引过的资源总数,即两个关键词出现的总次数;二者之商即为共现相关系数。一般只采用 $|A \cap B|$来衡量词汇相似度,却忽略了规模效应的影响。$|A \cup B|$正是为了消除规模效应的影响,使得各类关键词能够平等地计算共现系数。

因此,下一步工作将从语义角度对关键词关系进行构建,以期改进和完善关键词云图的应用价值。同时,关于 Keyword Cloud的使用反馈,应该有一个更科学可行的评价方案,用以评估 Keyword Cloud的引入是否切实改进了用户的资源访问效率。

3.2 总结

本文介绍的 Keyword Cloud是 Tag Cloud应用的扩展,是文献关键词的云图展示方式。实践部分在传统检索系统中引入关键词云图来汇总检索结果,并可以辅助二次检索。该应用结合了关键词搜索、云图展现以及 Tag Line技术,用户可以概览检索主题下的热门关键词,也可以通过关键词链接进行二次检索,缩小检索范围,精确检索结果。实践表明,图书馆传统服务在吸收和引入一些新的应用理念的基础上,可以使其服务增值。Web 2.0信息环境下,各领域的用户服务必将沿着个性化、知识化的方向发展。关键词云图虽然简单,但却充分体现了这种思想,起到了很好的抛砖引玉的作用,期待以后能有更多更好的服务模式,在帮助用户组织和发现资源上起到更好的作用。

参考文献

1 Horse Luke. 概念验证: Tag cloud生成工具制作过程. [2009-08-04]. http://blog.sina.com.cn/s/blog_56b7981801009mb.html

2 Tag Cloud [2009-08-04]. http://en.wikipedia.org/wiki/Tag_cloud

3 Mogens Nielsen. Functionality in a second generation tag cloud[D]. Department of Computer Science and Media Technology, Gjøvik University College, 2007

4 Byron Y-L. Kuo, Thomas Hentrich, Benjamin M. Good, and Mark D. Wilkinson. Tag Clouds for Summarizing Web Search Results. WWW, 2007, 1203-1204

5 M. Dredze, H. Wallach, D. Puller, and F. Pereira. Generating summary keywords for emails using topics. IUI, 2008, 199-206

6 Georgia Koutrika, Zahra Mohammadi Zadeh, and Hector Garcia-Molina. Data Clouds: Summarizing Keyword Search Results over Structured Data. EDBT, 2009, 391-402

7 Taglines [2009-08-10]. http://research.yahoo.com/taglines/

8 Chirag Mehta. Timeline-based Tag Clouds [2009-08-10]. http://chir.ag/projects/tagline/

9 Yusef Hassan-Montero, Víctor Herrero-Solana. Improving Tag-Clouds as Visual Information Retrieval Interfaces. International Conference on Multidisciplinary Information Sciences and Technologies, Spain, October 25-28, 2006

廖 凤 女,硕士研究生。

张建勇 研究馆员。

(收稿日期: 2010-04-19 编发: 刘炜 赵亮)