情报工程 ISSN 2095-915X
第2卷 第5期 50-57 2016 年 10 月

TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN 2095-915X Vol.2 No.5 50-57 Oct 2016

doi:10.3772/j.issn.2095-915x.2016.05.006

生物医学文献检索方法与问答系统

潘昊杰，周芳 ，张博文，张乐乐，方帆，殷绪成

（北京科技大学计算机科学与技术系 北京 100083）

摘要：如何有效的进行生物医学文献检索和信息挖掘，是计算机技术和生物信息技术研究领域中的一个经典课题。本文对生物医学文献中自然语言问题文档，片段，概念和 RDF 三元组，构建了高效的检索和问答系统。特别的，在文档检索中，我们搭建了基于顺序依赖模型，词向量，和伪相关反馈相结合的通用检索模型；同时，前 k 个文档被分离为句子和片段，并以此建立检索索引，并基于文档检索模型，完成片段检索；在概念挖掘中，提取生物医学的概念，列出相关的概念属于网络服务的五个数据库链接，通过得分排名得到最终的概念。在 CLEF BioASQ 几年的评测数据上，我们构造的检索系统都取得了不错的性能。

关键词：生物医学文献检索，序列依赖模型，词向量，伪相关反馈，排序学习

中图分类号：TP391

Query Processing in Biomedical Literature RetrievalandQuestion Answering System

Pan HaoJie, Zhou Fang,Zhang BoWen, Zhang LeLe,Fang Fan,Yin XuCheng

(School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China)

基金项目：本研究得到国家自然科学基金“结合前馈和反馈机制的自然场景文本识别技术”（编号：61473036）的资助，并在此基础上展开后续理论及应用研究。

作者简介：潘昊杰（1991–），硕士，研究方向：信息检索，数据挖掘，haojiepan@sina.com；周芳（1972–），博士，副教授，研究方向：机器学习、信息检索，zhoufang@ies.ustb.edu.cn；殷绪成（1977–），博士 / 教授 / 博导，北京科技大学计算机与通信工程学院计算机科学与技术系模式识别技术创新实验室主任，信息检索与推荐系统等应用技术国内知名青年专家；张博文（1992–），博士生，研究方向：机器学习、信息检索；张乐乐（1987–），硕士生，研究方向：信息检索；方帆（1992–），硕士生，研究方向：信息检索。

Abstract: How to effectively carry out the biomedical literature search and information mining is a classic topic in the field of computer technology and biological information technology research.This study constructed an efficient retrieval and question answering system refer to the related problem of natural language problems in biological medical literature documents,including snippets, concepts and RDF triplets.In particular, this research built a general search model based on Sequential Dependence Model, WordEmbedding and Pseudo Relevance Feedbackin the documents retrieval. Moreover, the former K documents were separated into sentences and snippets to establish the indexand complete the snippets search based on the documents retrieval model. In concepts mining, this study extracted biomedical concepts from the concepts, listed the related concepts belong to the web service of five URLs, and obtained the final concepts through the score rank. The results indicated that the retrieval systemof this study has achieved good performance based on the test data from CLEF BioASQ.

Keywords: Biomedical literature retrieval, sequential dependence model, word embedding, pseudo relevance feedback,learning-to-rank

1 引言

近几个世纪以来，随着互联网技术的发展，生物医学也得到了相应的发展。生物医学相关的文献不仅仅只有书籍的形式呈现，还可以以网络的形式呈现给用户。因此，如何使文献更快更准更好的呈现给用户已经成为信息检索领域的研究热点之一。文献逐渐以互联网内容的形式呈现，而生物医学检索的目的就是根据用户的查询，为其提供相应的文献。例如，在我们搭建的系统中用户以自然语言形式提出自己的需求，并得到其响应。即用户在搜索引擎中输入 The Anatomical Record，我们的系统就会返回给用户最相关的十篇文档。

当前国内外研究学者进行了一些卓有成效的研究，提出了一些具有启发性的方法与技术。总的来讲，我们将生物医学方法可分为三类，分别为基于顺序依赖模型的方法，基于词向量的方法和基于伪相关反馈的方法。首先，基于顺序依赖模型的方法主要关注于查询语句的结构，包括单个查询词特征、词组特征和乱序查询词组特征，三种特征分别进行搜索，然后再使用线性组合将结果融合在一起。其次，自然语言处理中最直观，也是到目前为止最常用的，向量的维度是词表大小[1]，其中绝大多数元素为 0，只有一个维度的值为 1，这个维度就代表了当前的词。最后，伪相关反馈，也称之为盲式相关反馈，把每个词表示为一个很长的向量，提供的是一种自动局部分析方法，它可以自动化相关反馈的手动操作部分，因此用户可不用参与额外的交互也可以获得更好的检索性能。

在本文中，我们提出了一个生物医学文献检索系统[2]。在文档检索中，我们搭建了基于顺序依赖模型，词向量，伪相关反馈相结合的通用检索模型。前 k 个文档被分离为句子和片段，以此来建检索索引。文档检索的相似模型被应用到片段检索。提取生物医学的概念，列出相关的概念属于网络服务的五个数据库链接，通过得分排名。

本文在第 2 节中详细描述了生物医学检索和相关工作；第 3 节中介绍了生物医学文献检索方法，包括：数据预处理，查询预处理，文档检索，片段检索和概念检索；在第 4 节中介绍了在测试

集上的实验结果和实验分析；最后进行总结。

2 生物医学检索和相关工作

生物医学的相关研究一直以来都是科学家的重要研究课题。现在，随着计算机和互联网技术的发展，越来越多的用户通过互联网查询自己想要了解的生物医学相关文献。生物医学文献检索的目的就是根据用户提出的查询，在指定的生物医学数据集中检索，为用户提供一个有序的生物医学文献推荐列表。

在我们的系统中，我们在索引和检索大型集群中部署 Galago，它是一种改进的 Indri 的 Java 版本的开源的搜索引擎。我们租赁了美国国家医学图书馆 2016 MEDLINE/PubMed Journal Citations，由 2200 万论文参考文献构成。

在以往的生物医学文献检索[3]的相关研究中，一些技术方法表现出了较好的检索效果。比如说顺序依赖模型，词向量和伪相关反馈，但是这些研究没有提出一种处理生物医学文献检索的通用方法[4]。我们在文档检索中将这三种方法融合在一起，形成一种通用的检索模型；并基于文档检索模型，完成片段检索；提取生物医学概念，计算得分进行排名。

3 文献检索方法

3.1 预处理

3.1.1 数据预处理

不管是本地的信息还是从网上下载的信息，都有可能来自不同种类的网页，有不同的标签、

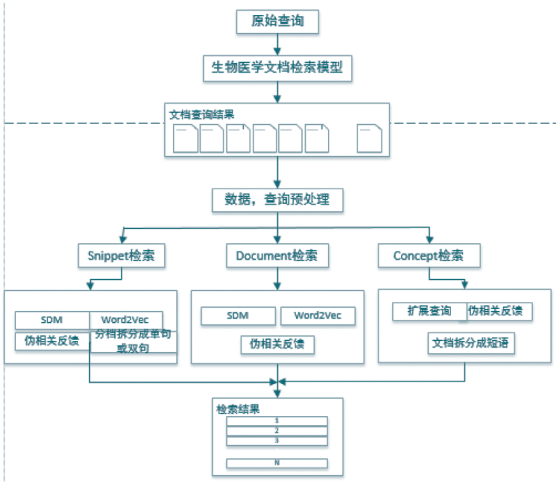

图 1 生物医学检索方法框架

表 1 生物医学检索 XML 文档示例

原始文档	处理文档
<MedlineCitationSet>	<MedlineCitation>
<MedlineCitation Owner="PIP" Status="MEDLINE">	<PMID>20981896</PMID>
<PMID Version="1">12336567</PMID>	<DateCreated>
<DateCreated>	<Year>1946</Year>
<Year>1981</Year>	<Month>12</Month>
<Month>03</Month>	<Day>01</Day>
<Day>10</Day>	</DateCreated>
</DateCreated>	<DateCompleted>
<DateCompleted>	<Year>2010</Year>
<Year>1981</Year>	<Month>10</Month>
<Month>03</Month>	<Day>28</Day>
<Day>10</Day>	</DateCompleted>
…	<DateRevised>
</MedlineCitation>	<Year>2014</Year>
<MedlineCitation Owner="PIP" Status="MEDLIN">	<Month>08</Month>
<PMID Version="1">12255534</PMID>	<Day>12</Day>
<DateCreated>	</DateRevised>
<Year>1980</Year>	<Article>
<Month>01</Month>	<Journal>
…	…
<MedlineCitationSet>	<MedlineCitation>

内容，如果要利用这些资源就必须进行格式的统一，转换成系统需要的格式[5]，比如原始文本、PTF、HTML、XML 等。除此之外，还需要统一编码格式。信息集合数据量比较大，存在多种字符，字符之间存在不同的编码问题。为了解决混乱的编码问题，人们制定了 Unicode 编码方案，解决了单个文件使用多种语言的问题[6]。一些网页中可能含有文本、链接、图片和标签，但这些内容和页面的主题并不是直接相关的。这些多余的内容绝大多数都是噪声，去除噪声也是数据预处理的一项重要的工作。有的噪声可以通过主观判断出来，对页面的排序起负面的作用。但是有些信息是模棱两可的，这就需要后期进行反复实验，通过检索结果的好坏来判断。这样经过数据的收集、加工处理，就得到了格式化的、可用的信息集合[7]。如图 1 所示，对检索文档进行了数据预处理。

3.1.2 查询预处理

词形整理包括一系列的步骤：去停用词、词干提取、大小写转换、词性标注等，采用与用户交互的方式实现的处理有：拼写纠错、查询扩展和相关反馈。可以看出，有些处理和文档处理是有相似之处的。一般情况下，如果存储空间允许，为了增加系统处理含有停用词的查询的灵活性，最好索引文档中所有的词。如果建立索引时对文档进行了词干提取，查询中的词最好也进行词干提取。这些处理都不是绝对的，在后期的实验中，可以根据系统的效果好坏而定。许多的处理技术都是根据后期的经验制定出来的规则，例如：常见的词干提取工具有 Porter、Krovetz，常用的英文停用词表有 INQuery 停用词表。斯坦福大学开发了词性标注器 Stanford POS Tagger 以及命名实体识别 Stanford Named Entity Recognizer（NER），前者能够标注出词的词性，比如名词，形容词等；后者能够识别出人名、地名、组织名以及一些专有名词等。其实这些也可以被用于自然语言处理的其他方面的研究[8]。如图 1 所示，对检索文档进行了查询预处理。

3.2 Documents 检索

如图 1 所示，文档检索应用到了顺序依赖模型，词向量和伪相关反馈，下面详细介绍这三项技术。

3.2.1 顺序依赖模型

马尔可夫随机域（Markov Random Field，MRF）模型是一种基于特征的线性模型，并且表示词项之间的依赖性。MRF 模型首先构建一个图，如图 3-1 所示，由一个文档节点 D 和查询中的每个词项对应的节点 q_1，q_2，q_3 组合而成，节点表示 MRF 中的随机变量。

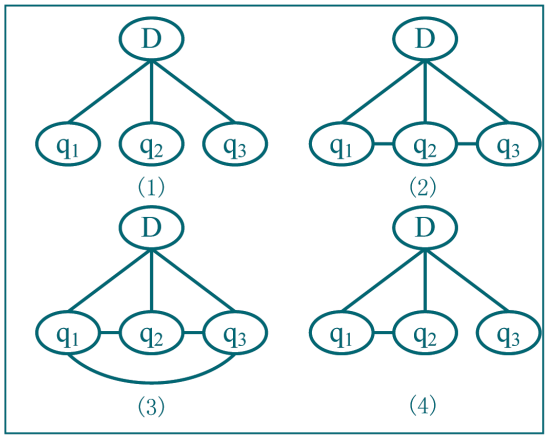

图 3-1 马尔科夫随机域模型假设

文档的相关性依赖于查询，所以图 3-1 中文档节点 D 连接着所有的查询词项 q 的节点。查询词项之间的联系是通过查询词项的节点之间的连线来实现的。图（1）是最简单的情况，全独立假设，在词项节点之间没有连线，说明词项之间是相互独立的，不存在任何依赖性，类似于一元模型；图（2）序列依赖假设，相邻的词项之间有连线，说明相邻的词项之间是相互依赖的，类似于二元

模型；图（3）是全依赖假设，每个词项与其他所有的词项都有连线，认为每个词项都依赖于其他所有的词项；图（4）是一般依赖，查询词项是根据某些有意义的方式进行连线。实践证明，序列依赖假设是最佳选择。

基于序列依赖假设的模型也称为顺序依赖模型（Sequential Dependence Model,SDM），它考虑了相邻词项之间的关系。本文中使用了这个模型，下面详细介绍一下这个模型。

顺序依赖模型[9]对查询语句进行处理，提取不同特征的短语来表征语句的信息，并赋予一定的权重来区分这些不同短语的重要性。

在顺序依赖模型中，从查询词中提取三个特征值对应的查询列表：简单词项列表 Q_T、有序短语列表 Q_O 以及无序短语列表 Q_U。处理之后的查询：$Q=q_1, q_2, q_3, \cdots q_i, \cdots q_{n-2}, q_{n-1}, q_n$，简单词项列表 Q_T：$q_1q_2q_3\cdots q_{n-2}q_{n-1}q_n$，有序短语列表 Q_O：$(q_1, q_2)(q_2, q_3), \cdots, q_i, \cdots, (q_{n-2}, q_{n-1})(q_{n-1}, q_n)$，无序短语列表 Q_U：$(q_1*q_2)(q_2*q_3), \cdots, q_i, \cdots, (q_{n-2}*q_{n-1})(q_{n-1}*q_n)$。其中*号表示零个或多个词。文档 D 的打分函数 $Score_{(Q,D)}$ 如公式 1 所示：

$$
\begin{aligned}
Score_{SDM}(Q,D) &= Score_{SDM}(Q_T, Q_O, Q_U, D) \\
&= \lambda_T \sum_{q \subset Q} f_T(q, D) \\
&+ \lambda_O \sum_{i=1}^{|Q|-1} f_O(q_i, q_{i+1}, D) \\
&+ \lambda_U \sum_{i=1}^{|Q|-1} f_U(q_i, q_{i+1}, D)
\end{aligned}
\qquad \text{（公式 1）}
$$

其中（1）$0 \le \lambda_T, \lambda_O, \lambda_U \le 1$ 且 $\lambda_T + \lambda_O + \lambda_U = 1$；（2）$\lambda_T \ge 0.60$；（3）$\lambda_O = 2 \times \lambda_U$。

式中 f_T——单个词项 q 与文档 D 之间置信概率的函数；f_O——无序短语 o 与文档 D 之间置信概率的函数；f_U——有序短语 u 与文档 D 之间置信概率的函数。

公式 1 中有三部分，每部分都对应模型中的一个特征类型，分别为计算匹配词项的贡献得分、计算查询中精确匹配有序二元短语的贡献得分以及计算相邻查询词在无序窗口中的贡献得分。而且每部分都是有加权的，λ_T、λ_o 和 λ_U 分别为三类特征设置的权重。λ_T 有更高的值，权重一般是根据经验或是实验调参得到的，一般取 $\lambda_T=0.85$，$\lambda_o=0.10$，$\lambda_U=0.05$。

3.2.2 Word2Vec

Word2Vec[10] 是 Google 公司在 2013 年开发的一款开源工具，能够从大规模的训练语料库中生成词向量，从而进行相似度计算找到近义词。

在顺序依赖模型的基础上，使用了词向量进行查询扩展。首先介绍一下词向量的构建。词向量是通过神经网络语言模型对大型的语料库建模训练得到的词的表示。本文使用 Word2Vec 工具进行词向量的训练，构建过程包括模型的选择、参数的设定以及模型的训练。参数设定主要包括上下文窗口的设置、词向量维度的选择、高频词采样等。

在顺序依赖模型中，有三个特征值对应的查询列表 Q_T、Q_O 和 Q_U，在这里，提出了一个新的特征概念 Q_W，即将原始查询进行扩展后的查询列表，用它来代替 Q_T。

一个经过预处理操作之后的查询词集合为：$Q=q_1, q_2, q_3, \cdots q_i, \cdots q_n$，通过 Word2Vec，使用神经网络模型训练词向量词表，通过词表，可以找出查询词的向量，利用余弦相似度计算词表中的所有词向量与查询词向量之间的相似度，得

到与查询词 q_i 相近的词 p_{ik} 作为其扩展词集合 p_i：$p_i=p_{i1}$，p_{i2}，…，p_{ik}，其中 p_i 集合中的每一个词都可以替换 q_i，通常取前 k 个词，作为 q_i 的扩展词并赋予一定的权重，重新组合得到了新的查询扩展语句：$Q_{new}=t_1$，t_2，…，t_i，….，t_n，其中 $t_i \subset T_i=q_i$，p_{i1}，p_{i2}，…，p_{ik}，T_i 集合表示原始查询词 q_i 和扩展词重新组合的所有情况。文档 D 的打分函数 $Score_{(Q,D)}$ 计算如公式 2 所示：

$$\begin{aligned} Score_{word2vec}(Q,D) &= Score_{word2vec}(Q_W,Q_O,Q_U,D) \\ &= \lambda_W \sum_{t \subset T} f_W(t,D) \\ &+ \lambda_O \sum_{i=1}^{|Q|-1} f_O(q_i,q_{i+1},D) \\ &+ \lambda_U \sum_{i=1}^{|Q|-1} f_U(q_i,q_{i+1},D) \end{aligned}$$

（公式 2）

式中 f_w 是扩展之后的查询项 t 与文档 D 之间置信概率的函数，在第一项中考虑了查询扩展，利用 Word2Vec 找到与查询词相近的词进行扩展[11]，这一项占有很大的比重。整个函数考虑到了词项之间的顺序依赖关系。

λ_W、λ_O 和 λ_U 分别为扩展词项、有序短语和无序短语三类特征设置的权重[12]。由于这个模型是在顺序依赖模型的基础上实现的，所以这三个特征值和顺序依赖模型的三个特征值的关系是一样的。

3.2.3 伪相关反馈

基于伪相关反馈的查询扩展，也称基于局部分析的查询扩展[13]，假设初检结果的前 N 篇文档都是与查询相关的，然后利用这前 N 篇文档的信息对原始查询进行扩展。基于伪反馈的方法解决了相关反馈中需要用户参与的问题，整个过程是自动的。但是缺点是过分依赖于初检结果的前 N 篇文档，如果返回的前 N 篇文档与查询相关度很小时，从这些文档提取出的关键词可能就与原始查询无关，这样重构出来的查询就会含有很多与原始查询无关的词，反而使检索性能下降。

3.3 片段检索

在片段检索任务中，前 k 个文档被分成句子，这些句子中可能含有答案，并将标题和摘要抽取出来。在测试集 3b 中，片段检索的答案可能是单个句子或两个句子。如图 1 所示，文档可能被分成单个句子或者混合的单句和双句，然后为片段检索搭建索引。文档检索的相似模型被应用到这部分。在片段检索中，每个文档可能返回多个结果，因此我们删除重复的句子使结果更好。

3.4 概念检索

对于每次查询，根据五个网络服务端 (MeSH, GO, SwissProt, Jochem 和 DO) 回馈相关的概念。我们用了三种方法检索，查询扩展，去停用词和相关反馈。五个网络服务端是用来搜索概念的请求和返回搜索结果。该请求包含两个基本元素：关键词，执行搜索查询。通常，这是一个简单的短语查询，用空格将其隔开。一句话可以包含字母数字和标点符号，每次搜索可能返回成千上万的概念。因此，使用分页机制。页面是一个数字代表页面来进行检索，并每个网页的概念是一个数字，表示每页的概念的数量；结果是一个 JSON 对象列表（Tsatsaronis et al.,2015）。

如图 1 所示，我们使用了概念检索，扩展查询和伪相关反馈。一个查询可能返回成千上万个概念，我们设置检索前十个概念。通过降序预测相关性排序返回的概念搜索查询，得分代表这一相关性。对于每一个查询，通过得分我们排序所有结果，排名前十的结果作为概念检索的答案。

4 实验结果

在包含 496 个查询的测试集 3b 上训练和验证

我们的方法。我们利用 trec_eval 来评估前十个排列搜索列表。MAP 作为我们的评估指标。我们从测试集 3b 选择最好的参数以及在测试集 4b 上使用这些参数。

总的来说，这些参数在测试集 3b 上优化运行的很好。显然，Word2Vec 比顺序依赖模型和伪相关反馈具有更好的性能。特别是，对 Word2Vec 平均得分高于其他两个。在 5 个 Batch 中我们使用 Word2Vec 模型并将文档的标题和摘要拆分成句子。

表 1 生物医学检索 XML 文档示例

Method	SDM	Word2Vec	PRF
Batch1	0.1893	0.1987	0.1997
Batch2	0.2384	0.2491	0.2410
Batch3	0.2724	0.2786	0.2754
Batch4	0.2507	0.2523	0.2514
Batch5	0.3243	0.3287	0.3278

检索到的文档的得分越高，所包含的句子的准确度越高。在这种方式中，我们选择实验里的前 50、前 20、前 10 个文件作为 snippets 检索数据。

表 2 top-50,top-20,top-10 用 MAP@10 评测结果

Method	Top-10	Top-20	Top-50
Batch1	0.1041	0.1013	0.0874
Batch2	0.1129	0.0932	0.0754
Batch3	0.1719	0.1382	0.1030
Batch4	0.1784	0.1355	0.0883
Batch5	0.2244	0.1635	0.1382

在任务 4b 的 phase A 中，由 BioASQ 官方提供测试集 4b 的结果。我们提交的最好结果，在 Batch1 和 Batch2 中取得了第一名，MAP 得分分别为 0.0981 和 0.1499。在表 3 中，除了 Batch3 我们的结果是比别人的更好，由于随机数据。因此，在生物医学检索中我们的检索系统是有效的。

表 3 BioASQ 官方文档检索提供的 MAP@10 评测结果

Method	SDM	Word2Vec	PRF
Batch1	0.0973	0.0981	0.0967
Batch2	0.1473	0.1474	0.1499
Batch3	0.1175	0.1192	0.1188

在片段检索中，用 Word2Vec 的文档结果被用作一个索引。我们选择了顺序依赖模型，词向量，和伪相关反馈为检索模型。表 4 显示了在片段检索中我们系统的结果。

表 4 BioASQ 官方片段检索提供的 MAP@10 评测结果

Method	SDM	Word2Vec	PRF
Batch1	0.0216	0.0212	0.0201
Batch2	0.0355	0.0355	0.0339
Batch3	0.0410	0.0411	0.0416

在 task4b 的 phase A 中，我们为片段检索选择前 50 个文档。结果表明，我们的系统不是很有效。表 4 显示了检索模型对片段的检索结果影响不大。表 2 显示了检索到的文档的数量对片段检索有重要的影响。

5 结论

随着互联网的不断发展，通过网络来查询自己想要的文献是将来的趋势，怎样使检索效率不断提高，并且检索出的文档具有较高的准确度，对我们来说，这将是我们不断探索，不断提高我们系统性能的动力。本文谈论到的是主要的技术方法，在提高系统性能的过程中，还有很多其他的技术本文并没有提及到。

我们对数据集进行了各种处理，搭建了各种不同的索引，并且测试了各种索引的效果，并且我们采用各种各样的检索模型以及调整了所有各种可能的参数来提高最终的性能。虽然片段检索对 Batch1 和 Batch3 的效果不是很好。对数据集的查询集进行了深入分析，该方法的召回率很

高，但是 MAP 比较低。我们发现，检索文件的数量对片段检索具有重要影响，然后在 Batch4 和 Batch5 中，我们采用前 10 篇文档，片段检索期望在竞争中表现良好。

在未来，我们将专注于生物医学文本查询扩展的策略，通过片段检索反馈结果提高文档检索的准确性概率特征，我们相信片段检索能提高系统的性能。此外，我们的研究将自然语言处理（NLP）进入我们的系统来改善性能。

参考文献

[1]Balikas G,Partalas I, Ngomo A C N, et al. Results of the BioASQ Track of the Question An-swering Lab at CLEF 2014. CLEF (Working Notes), 2014: 1181-1193

[2]Choi S, Choi S. Classification and Retrieval of Biomedical Literatures:SNUMedinfo at CLEF QA track BioASQ 2014. CLEF (Working Notes), 2014: 1283-1295

[3]Zhang B W, Yin X C, Cui X P, et al. Social Book Search Reranking with Generalized Content-Based Filtering. Submitted to CIKM14.

[4]Zhang B W, Yin X C, Cui X P, et al. USTB at INEX2014: Social Book Search Track. CLEF (WorkingNotes), 2014: 536-542

[5]Lingeman J, Dietz L. Umass atBioASQ 2014: Figure-inspired Text Retrieval. In CLEF(Working Notes), 2014:1296 - 1310.

[6]Neves M L. Hpi in-memory-basedDatabase System in Task 2b of BioASQ. In CLEF(Working Notes), 1337 - 1347.

[7]Peng S W, You R H, Xie Z K, et al. The Fudan Participation in the 2015 BioASQChallenge: Large-scale Biomedical Semantic Indexing and Question Answering.In CLEF(Working Notes), 2015.

[8]Balikas G, Malakasiotis P, Partalas I, et al. An Overview of the BioASQLarge Scale Biomedical Semantic Indexing and Question Answering Competition[J]. BMC bioinformatics, 2015, 16(1):1.

[9] Koolen M, Kazai G, Kamps J, et al. Overview of the INEX 2011 Books and Social Search Track[J]. Lecture Notes in Computer Science, 2012, 7424:1-29.

[10]Zhang B W, Yin X C, Cui X P, et al. Ustb at Index2014: Social Book Search Track.In CLEF (Working Notes), 2014:536 - 542.

[11]Zhang Z J, Liu T T, Zhang B W, et al. A Generic Retrieval System for Biomedical Literatures: Ustb at BioASQ 2015 Question Answering Task.In CLEF(Working Notes), 2015.

[12]Balikas G, Kosmopoulos A, Krithara A, et al. Results ofthe BioASQTasks ofthe Question Answering Lab at CLEF 2015.In CLEF(Working Notes), 2015.

[13]Zhang B W,Yin C X,Cui X P, et al.SocialBookSearchRerankingwithGeneralizedContent-basedFiltering.InProceedingsof the 23rd ACM International Conference on Information and Knowledge Management, 2014:361-370.