文章编号：1003-0077（2011）00-0000-00

基于伪文档的伪相关反馈方法*

闫蓉，高光来

（内蒙古大学 计算机学院，内蒙古 呼和浩特 010021）

摘要：传统的伪相关反馈(Pseudo Relevance Feedback, PRF)方法通常是以文档作为扩展源单元提取扩展词，提取粒度过大造成扩展源质量下降，使得检索结果鲁棒性差。该文研究利用主题分析技术，尝试将文本语义内容作为扩展源单元，缓解扩展源质量不高的问题。提出并实现了对文本集中各文档内容的伪文档描述，通过对其进行隐式多样化处理，实现了从更细微的文本内容角度出发提取扩展词。通过在真实 NTCIR8 中文语料的检索结果表明，该方法可以有效的提升伪相关反馈的检索性能。

关键词：伪相关反馈；伪文档；主题分析；隐含主题

中图分类号：TP391　　**文献标识码：**A

A New Pseudo Relevance Feedback Approach Based on Pseudo Document

YAN Rong , GAO Guang-lai

(College of Computer Science, Inner Mongolia University, Hohhot, Inner Mongolia 010021, China)

Abstract: Traditional Pseudo Relevance Feedback (PRF) algorithms usually regarded the document as an expansion unit, which would decrease the quality of expansion source due to the larger extraction unit, and make the robustness of retrieval performance is poor. By virtue of the topic analysis techniques, this paper attempted to use the semantic content of text as the expansion unit so as to relieve the low quality of expansion source. Proposed and realized the pseudo document description of the content of each document in collection, on which it achieved extracting the expansion terms by using implicit diversification on the more subtle document content level. The experimental results on real NTCIR8 dataset show an important improvement in terms of PRF performance.

Key words: pseudo relevance feedback (PRF); pseudo document; topic analysis; latent topic

1 引言

对于基于关键词的检索方式，通常用户是通过构造短查询来表达其查询需求的。这样的结果是，检索效果会因为用户查询需求表达不全而表现不佳。为了弥补用户查询表达不全的问题，查询扩展(Query Expansion)技术[1]通过将与用户查询意思相近的词语引入用户初始查询，达到提高检索性能的目的。伪相关反馈(Pseudo Relevance Feedback, PRF)是一种自动局部查询扩展技术，假设用户查询的初检结果中排名靠前的文档，是与用户查询相关的，记为伪相关文档集(pseudo relevant set)，并从中抽取扩展词来对用户查询进行重构。该项技术在信息检索中被广泛应用[2-6]。然而，PRF 的检索性能严重受限于伪相关文档集的质量和数量，也是影响 PRF 查询精度的主要原因，具体表现在两个方面。一方面，当伪相关文档集中大多数的文档与用户查询是不相关的，将直接导致查询结果偏离用户初始查询意图，用户会发现在查询结果中有很多都不是自己想要的查询信息，即主题偏移(topic drift)现象的产生。特别的，当用户查询有歧义时，这种现象就会更加严重[1,7]，用户查询满意度差。另一方面，

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金项目（61263037）；内蒙古自然科学基金项目（2014BS0604）

作者简介：闫蓉（1979—），女，讲师，博士生，主要研究领域为自然语言处理和信息检索；高光来（1964—），男，教授，博士生导师，主要研究领域为智能信息处理。

随着选取的伪相关文档数目的增加，其中内容相似的文档数量就会增多，将导致查询结果会集中偏向于用户初始查询意图中的某一个主题，而且这些相似的文档往往会集中分布在查询结果靠前的位置。在这种情况下，提取的反馈文档信息效用会明显降低。传统的 PRF 方法一般会在伪相关文档集中选取频度较大的词项作为候选扩展词，这样做的结果是，那些排名靠后并且涉及其它主题的文档，对于满足用户查询满意度的效用就会进一步减少，甚至会达到对用户无用的程度[8]。另外，对于传统的信息检索方法，大多是基于用户的信息需求(information need)是语义单一的假设前提下来寻找相关文档的。而事实上，用户所提交查询的实际或潜在信息需求并不具有单一性，而是具有模糊性、不确定性和未知性的。

其实，以上这些问题都可以通过将查询结果进行多样化处理的方式来解决。查询结果的多样化强调将用户查询的潜在信息需求，通过在查询结果中进行多样化表达来完成，目的是更好的保证用户查询满足度。多样化研究表明[7,9-10]，相比于传统的搜索引擎反馈结果，提高用户查询满意度较好的办法就是给用户提供尽可能多不同的信息，而这些信息中至少会有一个是与用户需求是相关的。依据多样化问题是否直接对用户潜在意图进行估计，多样化研究方法大体上包括显式方法和隐式方法两种[7]。本文延用隐式多样化思想，即不直接对用户潜在意图进行估计，基于 PRF 方法假设，寻找文档间的差异性来实现文档多样化，提出一种新的提高反馈源质量的 PRF 方法。解决的核心思想就是通过伪相关文档集抽象内容来表达用户查询意图的多样化。查询结果多样化研究[7,9-10]和 PRF 方法的研究中[2-6]，其研究对象和反馈源都是基于文本为处理单元进行的，使得 PRF 检索性能受制于其质量。本文打破这种选取模式，选用的是更小粒度的文档内容作为反馈处理单元，以期获得更加细化、更加相关的反馈源。主题模型[11]认为每个文档是与多个隐含主题(latent topic)相关的，并且可以用有限隐含主题描述文档，这是与其它限制每个文档只与一个主题相关的文本建模模型本质上的区别。本文利用主题模型中的隐含主题信息，从浅层语义角度出发，通过提取表示文本语义的特征信息来表征文本，提出并实现了文本数据集中各文档的伪描述处理，构成各文档的伪文档(pseudo document)表达，将用户尽可能多的查询意图在由伪文档所构成的伪相关文档集中表达出来，提出了一种基于伪文档的 PRF 方法。

2 方法学

传统的基于关键词的查询方式，是基于假设用户的查询需求单一的前提下进行的，然后按照用户查询与各文档间相关度大小，将排序后的文档呈现给用户。在这种假设前提下，查询结果中势必会存在表示同一相似内容的、冗余的文档。然而，在实际用户检索行为中，用户往往只关注前 20-30 个搜索结果。最坏的情况是，如果用户在这其中未找到其所关注的内容，用户将会重新构造查询，用户体验得不到满足。本文认为，用户真实的查询需求应该是多层面的，但往往由于用户短查询的不完全需求表达构造，用户真实的查询需求被描述成为一种带有多变性或多样性的表达结果，使得搜索引擎不可能完全捕获用户的真实需求。与此同时，用户要查询的对象，即文本数据，其所表达的信息却是相对稳定的，并且这些信息是可以通过合适的方式获取的。这样，用户的检索行为就可以描述为：用不确定的查询需求表达查找确定内容文本的过程。为了能够让用户快速的找到其所关注的内容，迎合用户查询需求的多变性或未知性，本文认为可以在相对稳定的文本数据中将其所表达的语义抽取出来，然后寻找与用户查询需求相匹配或是符合的语义层面的体现，最终按照匹配程度的不同，抽取扩展词，用于反馈行为。

综上所述，针对 PRF 对伪相关文档受限问题，本文所需要解决的问题包括两个方面。一个是文本数据中所包含的语义如何表达的问题；另一个是如何在伪相关文档集中，将用户的多样化查询意图表达出来的问题。

第一个问题的解决，可以通过对文本数据集进行主题分析(topic analysis)来完成，通过挖掘文本数据集中的隐含主题，实现文档与词项的关联，实现用隐含主题信息描述文档的浅

层语义。通常，对文本数据集进行主题分析的方式有两种，包括有监督的主题标注和无监督的主题建模。其中无监督的文本建模方法有文本聚类和潜在主题建模两种[12]。因为本文处理的数据集是无任何标识的离散文本数据，所以选择无监督的主题建模方式来对文本数据进行主题分析。本文选用的是标准的主题建模方式 LDA[13](Latent Dirichlet Allocation)方法。

第二个问题的解决，可以利用对文本数据集的 LDA 建模结果，按照与用户查询意图相关程度的不同，抽取与用户查询意图相关的多样化表达信息。抽取对象的获取可以通过类似自动文摘(text summarization)的思想完成。自动文摘的目的，是生成能够包含并概括整个文本主要信息的摘要，实现文本的简短描述。本文希望通过挖掘文本数据集的浅层语义描述，实现对文档在主题内容粒度下的文档信息的短描述，即文档的伪文档描述。然后，从伪相关文档集中各文档的伪文档描述中，抽取与用户查询意图多样化表达相关的信息，达到改善 PRF 反馈源质量的目的。具体实现可以描述为：首先，将 LDA 建模后得到的各隐含主题，认为是描述文本数据的“特征”，但是这些特征并没有达到标识文本的能力，使得文本之间具有可区别性，所以本文要对这些“特征”做进一步的分析处理。在 LDA 中，隐含主题可以看作是所有词项按一定比例的混合。同时，文档可以看作是所有隐含主题按一定比例的混合。依此，对于初检结果中的各文档 d，也可以看作是各个词项的混合文本 d'，获得混合文本 d' 的过程，称为文档 d 的伪表达抽象描述过程。其中，文本 d' 记为文档 d 的伪文档。经过伪表达处理之后的文档集记为伪文档集。接着，采用最大边缘相关 MMR(Maximal Marginal Relevance,MMR)[14]方法，将在伪文档集中与用户查询意图相关内容，进行多样化处理后重定序。MMR 方法是信息检索研究中，实现对检索结果文档进行多样化处理的经典方法。其实现思想是，将那些与用户查询相关，且与先前被选择的文档相似性小的文档认为具有较高的边缘相关，从而实现多样化。本文的重定序对象并不是文档本身，而是针对文档的伪文档，即文档内容进行的。按照与用户初始查询语义相近且与之前所选伪文档内容不同的方式，对伪文档集中的伪文档进行重定序。最后，从排名靠前的伪文档集中，抽取扩展词，实施二次反馈。

3 基于伪文档的伪相关反馈

3.1 LDA 文本建模

LDA[13]最早在 2003 年由 Blei 提出，是一种贝叶斯层次模型(bayes hierarchy model)，从浅层语义角度实现文本抽象建模。LDA 模型通过构建主题空间(topic space)，并在其上抽取文本数据的隐含主题信息，实现用隐含主题抽象表示文档。同时，每个隐含主题又被抽象表示为各词项的概率分布。图 1 所示为 LDA 的图模型表示。

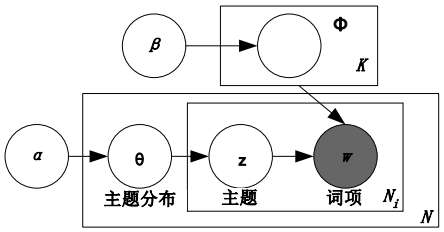

图 1 LDA 的图模型表示

通过 LDA 对文本进行抽象建模，不仅可以实现文档的相对短小描述，也可以挖掘文本中潜在的统计语义信息。LDA 中假设文本数据集中组成文档的各词项，是被隐含主题通过 Dirichlet 先验主题分布进行反复抽样产生的。假设文本数据集 C 中有 N 个文档，共包含 V 个不同的词项，$C=\{d_1,d_2,...,d_N\}$，每个文档 d_i 由 N_i 个不同的词项构成。假定 K 为 C 上的主题数目，向量 z 定义为每个词项对隐主题的分配。α 和 β 分别对应文档的主题 Dirichlet 先验分布 θ 和每个主题的词项 Dirichlet 先验分布 Φ 的超参数。

由图 1 可知，整个产生式模型可以用公式(1)来描述：

$$P(d_1,...,d_N,\phi,\theta|\alpha,\beta) \propto \left(\prod_k^K P(\phi_k|\beta)\right)\cdot\left(\prod_j^N P(\theta_j|\alpha)\right)\cdot\left(\prod_i^V \phi_{z_i}(w_i)\theta_{d_i}(z_i)\right) \quad (1)$$

其中， $\phi_{zi}(w_i)$和 $\theta_{di}(z_i)$分别表示向量 ϕ_{z_i} 和向量 θ_{di} 中的第 i 个元素。LDA 建模的目的就是推导潜在变量(z,ϕ,θ)。由于复杂的后验分布，LDA 模型无法实现精确的参数推导，本文采用 Gibbs 抽样[15]来对参数进行估计推断。在对实际文本数据集的 LDA 建模过程中，由于主题数目与数据集是严重相关的，不同的主题数目会导致主题建模结果的不同，所以有必要确定文本数据集的最佳主题数目。目前对于主题模型的评价尚无标准的方法，本文采用如下方式来获取最佳主题数目。首先，本文采用常用的语言模型中的困惑度[13](Perplexity)方法来获取最佳主题数。其中，困惑度值越低意味着模型产生文本的能力越高，模型效果越好，模型具有更好的推广性。接着，将 LDA 建模结果用于解决有明确评价指标的实际应用问题，进一步评价建立的主题模型的有效性。

3.2 文档的伪表达获取

3.2.1 获取原则

对文本数据进行 LDA 建模后，其结果可以表示为如下形式：文本-主题分布记为 $\theta \in R^{N\times K}$，主题-词项分布记为 $\Phi \in R^{K\times V}$，每一个 $\phi_{i,m}$ 表示在主题 i 中生成第 m 个词项的概率值，每一个 $\theta_{j,i}$ 表示第 j 个文本中主题 i 所占比重。其中，不同文本的差别仅限于 K 个隐含主题所占比重的不同。同时，隐含主题间的差别仅限于 V 个词项所占比重的不同。文本间的差异表现并不明显且描述抽象，很难将 LDA 建模结果直接应用于实际的检索行为中。另外，隐含主题之间没有明显的可供区别的特征。本文认为各词项所表达隐含主题的含义和各隐含主题在表达文本隐含语义时，作用或表达能力是有差异的。若将隐含主题作为特征来描述文本数据的话，必须将这种抽象的差异转换成是人们可以理解的形式。具体的讲，就是应该将表达所在文本语义能力尽可能高的隐含主题突显出来，表达隐含主题能力尽可能高的词项突显出来，即将文本间可以区别的特征内容突显出来，真正实现文档在主题粒度下的短文本描述。这样，文本的 LDA 建模结果就可以直接应用于实际的 PRF 中。首先，本文认为对于表示各文本浅层语义的 K 个隐含主题的贡献程度或语义表达程度是不同的。若某个隐含主题对于描述其所在文本的重要程度越大，同时，该隐含主题对于描述文本数据集的重要程度越大，那么，该隐含主题表征其所在文本的浅层语义的描述能力就越大。其次，组成各隐含主题的各词项，对于描述其所在隐含主题语义的贡献程度也是不一致的。若某个词项在其所在隐含主题中概率值越大，同时，该词项在其它隐含主题中出现概率值越小。那么，该词项对描述其所在隐含主题语义重要程度就越大。依此方式，就可以突显每个文档中描述其浅层语义的特征信息。

3.2.2 伪文档获取方法

文献[12]为了帮助用户更直观的理解文档的主题描述，对主题进行了排序。本文借鉴该方法，对 LDA 建模后的结果进行排序，包括对文本-主题分布 θ 和主题-词项分布 Φ 排序。排序依据分别是隐含主题在文本数据集中的重要程度和词项在隐含主题中的重要程度。其中，隐含主题在文本数据集中的重要程度衡量，是通过计算隐含主题在数据集中的内容覆盖度和方差的组合得到。词项在隐含主题中的重要程度衡量，是通过计算词项在其所在隐含主题的 TF(Term Frequency)和在所有隐含主题中的 IDF(Inverse Documentation Frequency)权重方式得到。目的是通过两个排序行为，实现在主题粒度下对于表示特定文本的隐含主题描述进行排序，抽取能够表征所在文本浅层语义的各主题的关键词，获取文档的伪文档表达。文档经过伪表达处理后，将会用从主题-词项分布中提取的有限关键字(*Keywords*)来表达。算法 1 描述的是文档的伪文档获取过程。

算法 1：*Pseudodoc_get*(*C*)

输入：文本数据集 C。

输出：C 中各文档的伪表达 C'。

Step1 文本数据集的浅层语义建模。设置初始主题数目 L，对 C 进行 LDA 建模。

Step2 计算困惑度，如公式(2)所示。其中，对于一个具有 M 个文档的测试集 R，N_m 为第 m 篇文本 d_m 的长度；$P(d_m)$表示模型产生文本 d_m 的概率。

$$perplexity(R)=\exp\left\{\frac{-\sum_{m=1}^{M}\log(P(d_m))}{\sum_{m=1}^{M}N_m}\right\} \tag{2}$$

根据困惑度计算结果，为 C 选取适合的主题数目 K。

Step3 计算隐含主题分布中各词项权重。利用公式(3)，重新计算主题-词项分布 Φ 中的各词项权重值，将隐含主题内容进行压缩和抽象，过滤对隐含主题间区分能力描述弱的词项，使隐含主题之间可以有效区分，并提取描述隐含主题的关键字序列。其中，w_m 表示隐含主题 $i(i\in[1,K])$中第 m 个词项。

$$WC_{weight}(w_m)=\phi_{i,m}\cdot\log\frac{\phi_{i,m}}{\sqrt[K]{\prod_{j=1}^{K}\phi_{j,m}}} \tag{3}$$

Step4 计算每个隐含主题的重要度 r。对每个隐含主题 $i\in[1,K]$，利用公式(4)、(5)和(6)，分别计算其对于描述文本数据集的覆盖度、方差及重要度 r_i 值，并按 r_i 值，实现对各隐含主题的排序。其中，μ_i 和 σ_i 分别表示隐含主题 i 在数据集中的覆盖程度和方差。

$$\mu_i=\sum_{j=1}^{N}N_j\cdot\theta_{j,i}\Big/\sum_{j=1}^{N}N_j \tag{4}$$

$$\sigma_i=\sqrt{\sum_{j=1}^{N}N_j\cdot\left(\theta_{j,i}-\mu_i\right)^2\Big/\sum_{j=1}^{N}N_j} \tag{5}$$

$$r_i=\mu_i\cdot\sigma_i \tag{6}$$

Step5 伪文档获取。对 C 中的每一个文档 d_j，利用公式(7)，计算各隐含主题 i 描述其所在文本 d_j 的描述能力 $ability_i$，并按此将描述该文档的隐含主题序列进行排序。通过设定阈值的方法，将序列中 $ability_i$ 值超过阈值 η 的隐含主题的 *Keywords* 集进行合并，作为文档 d_j 的伪文档 d_j'表达，获取 $C'=\{d_1',\ldots,d_N'\}$。

$$ability_i=r_i\cdot\theta_{j,i} \tag{7}$$

3.3 基于伪文档的伪相关反馈

算法 2 是基于伪文档的伪相关反馈方法描述。

算法 2：*Pseudodoc_PRF*(*Q*,*C*)

输入：用户查询 Q 和文本数据集 C。

输出：伪相关反馈结果。

Step1 对 Q 进行预处理，包括分词和去除停用词处理，记为 $Q=\{q_1,...,q_n\}$。

Step2 用 Q 去检索 C，在初检结果中取前 P 个文档，记为 $D_r=\{d_1,\ldots,d_P\}$。

Step3 对 D_r 中各文档按照算法 1 处理结果，获取各文档的伪文档表达，构成伪文档集，记为 $D_r'=\{d_1',\ldots,d_P'\}$。

Step4 按照公式(8)计算 D_r'集中每个伪文档 d_i'的 MMR 值，并按 MMR 值，将各伪文档 d_i'进行降序排列，排序结果集记为 *Rank*。

$$MMR(d_i') = \arg\max_{d_i' \in Y}(\lambda Sim_1(d_i', Q) - (1-\lambda) \max_{d_j' \in S}^{k} Sim2(d_i', d_j')) \quad (8)$$

其中，S 表示 D_r'中已经被选取的伪文档集合，$d_j' \in S$。$Y=D_r'-S$ 表示未被选取的伪文档集合，$d_i' \in Y$，λ 是调节参数。$Sim_1(d_i',Q)$表示查询 Q 与伪文档 d_i'之间的相似度，$Sim_2(d_i',d_j')$表示两个伪文档 d_i'和 d_j'之间的相似度。

经过算法 1，D_r 中各文档都已经被抽象描述为主题空间下的伪文档形式。由于伪文档是由若干 *Keywords* 构成，那么，在公式(8)中进行相似度计算的对象就可以归结为计算两个词向量之间的相似程度。计算两个词向量相似度的方法有很多，本文采用经典的词向量相似度计算 Cosine 方法来计算 $Sim_1(d_i',Q)$和 $Sim_2(d_i',d_j')$。

Step5 从 *Rank* 集取前 k 个文档作为伪相关文档，从中抽取扩展词，实施二次反馈。

4 实验

4.1 实验设置

4.1.1 实验数据

本文实验是对简体中文数据进行处理的，数据集包括文档集和查询集两部分。文档集包括 xinhua2002-2005，涉及各种主题的四年新闻语料，共有 308,845 个文件，其中含有文本内容的有 308,827 个文档；共有词数 65,079,191 个，其中，不同词数 223,632 个。查询集包括 ACLIA2-CS-0001～ACLIA2-CS-0100，共 100 个查询。表 1 列出了数据集中文档集和查询集的统计结果描述。

表 1 数据集中文档集和查询集的统计描述

文档集名称	文档数	词项数	最长词项数	最短词项数	查询词项数	个数
xinhua2002	64251	12,721,776	1875	4	1-3	36
xinhua2003	73431	15,444,634	1874	1	4-6	48
xinhua2004	84287	18,268,710	1716	1	7-10	16
xinhua2005	86858	18,647,309	1805	1		

本文使用 Lemur[1] 工具建立索引和查询。建索引前对文档集和查询集均进行了预处理，包括分词(中科院计算所 ICTCLAS)和去除停用词操作。对文档集进行 LDA 建模时，还去除了文档集中的部分虚词、形容词、副词和频度小于 5 的词项(共 130,363 个不同词)。

4.1.2 参数设置

初始 LDA 建模初始设置主题数目为 10，α=50/K，β=0.01，Gibbs 抽样的迭代次数设定为 100 次，设置返回描述隐含主题的词项个数为 30。本文初检结果的相关度排序方法选用的是典型的一元语言模型 LM(Language Model)方法，目的是考察能否通过主题空间的统计结果来改善词项空间的统计结果，进而改善 PRF 反馈源质量。实验中统一采用 Dirichlet 平滑方法，并在实验过程中统一设置平滑参数 μ 为固定值 1000。按照文献[16]研究表明，扩展词个数的数目设定为 10-20 时，效果最佳。本文设定选取的扩展词个数为 20。

4.2 实验结果与分析

实验 1，确定文档集的最佳主题数目。依次将主题数目分别取 K=10、20，直至 100，分别对文本数据集进行 LDA 建模，分析 Perplexity 值的变化，Perplexity 值随主题数目变化曲线如图 2 所示。

从图 2 可知，LDA 在主题数目为 60-70 时性能最好，实验选取最佳主题数 K=60。实验 2，获取文档集中各文档的伪文档。本文的目的是通过对文档的伪表达抽象，提高 PRF 反馈源质量，改善检索性能。与设置伪反馈的扩展词数一致，在算法 1 实现过程中，对公式(3)计算结果中取前 20 个词项，作为描述主题的关键字序列。实验中，将文档数据集 2002-2003 两年共 137,682 个文档作为训练集，剩余 2004-2005 两年文档 171,145 个文档作为测试集。在训练集中进行实验，隐含主题排序序列中设置阈值 η=0.12 时，效果最好。

[1] http://www.lemurproject.org

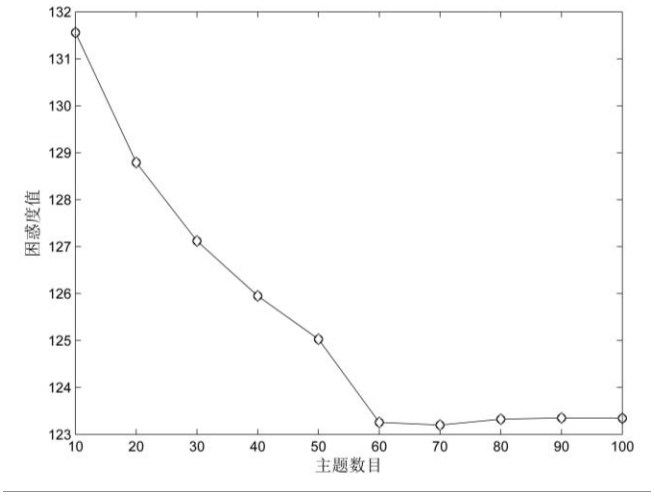

图 2 Perplexity 值变化曲线

表 2 所示为文档集中部分文档的原始文档的词项数、伪表达后的文档词项数和文档的主题表示列表 *Topic_list*(*Topic_list* 是按照各主题的 *ability* 降序排列且超过阈值 η 后的主题编号列表)。

表 2 部分文档的伪表达前后对比

文档编号	文档原始词项数	伪文档词项数	*Topic_list*
20020101.0069	223	116	29,35,17,7,18
20030314.0158	114	70	43,58,9
20040202.0155	1308	152	52,18,51,11,3,21,13,7,32
20050611.0175	221	102	5,11,13,29,55,40

图 3 所示为文档集中所有文档的原始词项数和相应伪文档表达后词项数的对比结果。

从表 2 和图 3 的结果可以看出，通过对各文档进行伪表达处理之后，各文档的伪文档的词项个数明显降低，达到了文本简短描述目的。另外，从图 3 可以看出各文档的伪文档词项数目分布均匀，降维效果明显。

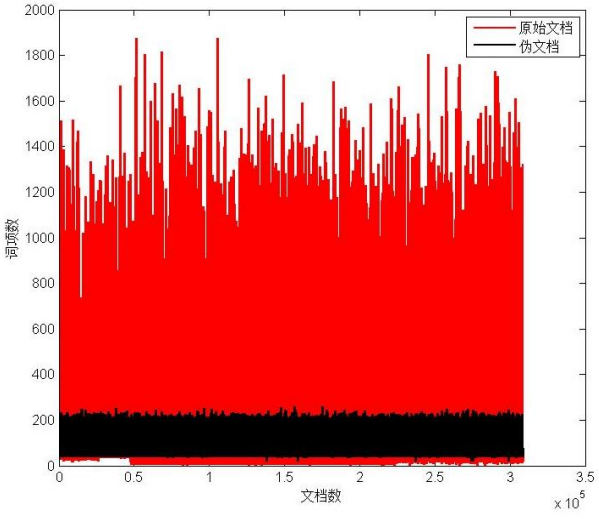

图 3 文档词项数和对应伪文档词项数对比

实验 3，用三个信息检索经典评价指标进行检索性能的评测，分别是平均查准率 MAP(Mean Average Precision)、前 n 个结果的查准率 Precision@n 和衡量排序质量的指标 NDCG(Normalized Discounted Cumulative Gain)。Baseline 选取标准的 BM25[17]伪反馈，实验中还与经典的 TF-IDF 伪反馈方法进行了比较，并统一设置 *feedbackDocCount*=10。进一步的为了验证提出方法的有效性，本文还与未进行伪文档表达的初检结果，同样也应用 MMR

算法的结果进行了比较，结果表示为 No_pseudo。本文提出的方法表示为 Have_pseudo。关于调节参数 λ 的选取在实验中采用贪心策略，当 λ 值取 0.7 时检索效果最好。表 3 列出了四种方法在各项指标上的结果对比。

表 3 四种方法结果对比

Metric	MAP	NDCG	P@5	P@10
Baseline	0.2690	0.3997	0.5239	0.5056
TF-IDF	0.2815	0.4128	0.5296	0.4817
No_pseudo	0.2912	0.4224	0.5389	0.4917
Have_pseudo	**0.3010**	**0.4328**	**0.5452**	0.4986

实验结果表明该方法的有效性，达到了改善反馈质量源的目的。从表 3 可以看出，本文提出方法，分别在 MAP、NDCG 和 P@5 三个指标上，高于 Baseline：11.90%、8.28%和 4.07%；高于 TF-IDF：6.92%、4.84%和 2.95%；高于 No_pseudo：3.4%、2.5%和 1.2%，达到了改善 PRF 扩展质量源的目的。尽管在指标 P@10 上，结果逊于 Baseline 方法，这其实和实验预期是一致的，因为本文的 PRF 方法，目的是将伪相关文档集中，迎合用户查询意图多样化的内容作为查询结果。因为在文本检索中，对于用户的检索行为结果，要尽量做到查询结果的相关性和查询结果类别多样性的折中[18]。虽然本文提出方法在各项指标上均高于 No_pseudo 方法，但提高并不是很明显，说明本文方法对主题分析精度要求较高，理想的主题分析结果可以进一步提高 Have_pseudo 方法的伪反馈检索效果。

实验 4，通过不断扩大选取伪反馈文档数目 k，观察 MAP 的变化，考察本文提出方法是否受限于伪相关文档集数目。图 4 所示为三种方法不同 k 值的 MAP 对比曲线。

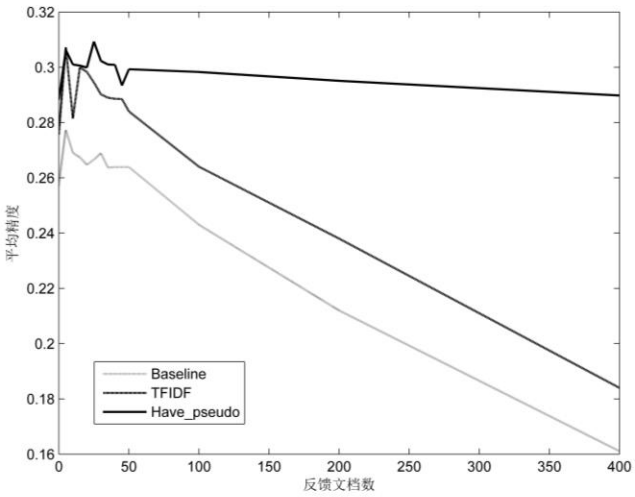

图 4 三种方法 MAP 对比曲线

从图 4 可以看出，随着选取的伪反馈文档数目 k 的增加，BM25 和 TF-IDF 两个方法 MAP 变化明显，说明 BM25 和 TF-IDF 这两种方法对 k 值的变化非常敏感，而且随着 k 值的增加检索性能下降非常明显。相对地，Have_pseudo 方法 MAP 变化平缓，检索性能起伏变化平缓，说明 k 的选取对 Have_pseudo 方法的检索性能影响并不明显，并且在选取 k 值并不是很大的情况下，检索效果就很明显。

综上的实验结果表明，这种将文档内容作为扩展源来抽取扩展词对象的方法是切实可行的。

5 结论

本文通过对文本数据集进行浅层语义分析，尝试将反馈源对象，从传统的文档粒度，换之以更能体现语义的文档内容粒度进行。实验结果验证了提出方法的科学性。通过明确的信息检索评价指标结果，也进一步证实了主题建模结果的有效性。从实验结果可以看出，通过不断的提升对文本数据的主题分析精度，本文的方法有一定的改善和提升空间。

参考文献

[1] Carpineto C, Romano G. A survey of automatic query expansion in information retrieval[J]. ACM Computing Surveys (CSUR), 2012, 44(1): 1-56.

[2] Metzler D, Croft W B. Latent concept expansion using markov random fields[C]// Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. Amsterdam, the Netherlands. 2007: 311-318.

[3] Lee K S, Croft W B, Allan J. A cluster-based resampling method for pseudo-relevance feedback[C]// Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. Singapore. 2008: 235-242.

[4] Mei Q, Zhang D, Zhai C. A general optimization framework for smoothing language models on graph structures[C]// Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. Singapore. 2008:611–618.

[5] Huang Y, Sun L, Nie J Y. Query model refinement using word graphs[C]// Proceedings of the 19th ACM conference on Information and knowledge management. Toronto, Ontario, Canada. 2010:1453–1456.

[6] Parapar J, Presedo-Quindimil M A, Álvaro Barreiro. Score distributions for Pseudo Relevance Feedback[J]. Information Sciences, 2014, 273(8): 171–181.

[7] Vargas S, Santos R L T, Macdonald C, et al. Selecting effective expansion terms for diversity[C]// Proceedings of the 10th conference on Open research areas in information retrieval. Lisbon, Portugal. 2013: 69-76.

[8] Clough P, Sanderson M, Abouammoh M, et al. Multiple approaches to analysing query diversity[C]// Proceedings of the 32nd annual international ACM SIGIR conference on Research and development in information retrieval. Boston, USA. 2009: 734-735.

[9] Clarke C L A, Kolla M, Cormack G V, et al. Novelty and diversity in information retrieval evaluation[C]// Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. Singapore. 2008: 659-666.

[10] Teevan J, Dumais S T, Horvitz E. Characterizing the value of personalizing search[C]// Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. Amsterdam, the Netherlands. 2007: 757-758.

[11] Blei D. M., Lafferty J. Text Mining: Theory and Applications[M]. Chapter Topic Models, Taylor and Francis, London, 2009.

[12] Wei F, Liu S, Song Y, et al. Tiara: a visual exploratory text analytic system[C]// Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. Washington. 2010: 153-162.

[13] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. Journal of machine learning research, 2003, 3: 993-1022.

[14] Carbonell J, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summarics[C]// Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. Melbourne. 1998: 335-336.

[15] Griffiths T L, Steyvers M. Finding scientific topics[J]. Proceedings of the National Academy of Sciences of the United States of America (PNAS), 2004, 101(z1): 5228–5235.

[16] Ogilvie P, Voorhees E, Callan J. On the number of terms used in automatic query expansion[J]. Information Retrieval, 2009, 12(6): 666-679.

[17] Jones K S, Walker S, Robertson S E. A probabilistic model of information retrieval: development and comparative experiments: Part 1[J]. Information Processing & Management, 2000, 36(6): 779-808.

[18] Karimzadehgan M, Zhai C. A learning approach to optimizing exploration–exploitation tradeoff in relevance feedback[J]. Information Retrieval, 2013, 16(3): 307-330.

闫蓉（1979—），女，讲师，博士生，主要研究领域为自然语言处理和信息检索。Email: csyanr@imu.edu.cn；

高光来（1964—），男，教授，博士生导师，主要研究领域为智能信息处理。 Email:csggl@imu.edu.cn。

作者联系方式：

姓名：闫蓉
地址：内蒙古呼和浩特市赛罕区大学西路 235 号内蒙古大学计算机学院
邮编：010021
电话：13674832054
电子邮箱：csyanr@imu.edu.cn