

文章编号:1001-9081(2016)08-2099-04

doi:10.11772/j.issn.1001-9081.2016.08.2099

## 基于检索结果排序的伪相关反馈

闫 蓉\*, 高光来

(内蒙古大学 计算机学院, 呼和浩特 010021)

(\*通信作者电子邮箱 csyanr@imu.edu.cn)

**摘要:**针对传统伪相关反馈(PRF)算法扩展源质量不高使得检索效果不佳的问题,提出一种基于检索结果的排序模型(REM)。首先,该模型从初检结果中选择排名靠前的文档作为伪相关文档集;然后,以用户查询意图与伪相关文档集中各文档的相关度最大化、并且各文档之间相似性最小化作为排序原则,将伪相关文档集中各文档进行重排序;最后,将排序后排名靠前的文档作为扩展源进行二次反馈。实验结果表明,与两种传统伪反馈方法相比,该排序模型能获得与用户查询意图相关的反馈文档,可有效地提高检索效果。

**关键词:**伪相关反馈;潜在狄里克雷分配;主题模型;查询扩展

**中图分类号:** TP391.3    **文献标志码:**A

### Pseudo relevance feedback based on sorted retrieval result

YAN Rong\*, GAO Guanglai

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

**Abstract:** Focusing on the low quality of expansion source of traditional Pseudo Relevance Feedback (PRF) algorithms, which lead to low retrieval performance, a retrieval result based sorting model, namely REM, was proposed. Firstly, the first-pass retrieval result was considered as a pseudo relevant set. Secondly, documents in the pseudo relevant set were re-ranked based on rules of maximizing the relevance between the user query intention and the documents of pseudo relevant set and minimizing the similarity between documents. Finally, the top ranked documents of the re-ranking were regarded as the expansion source to the second-retrieval. The experimental results show that, compared with two classical PRF methods, the proposed model can improve the performance of retrieval and obtain more relevant feedback document to the user query intention.

**Key words:** Pseudo Relevance Feedback (PRF); Latent Dirichlet Allocation (LDA); topic model; query expansion

### 0 引言

随着 Web 的普及,越来越多的用户希望从互联网上获取信息。对于目前主流的基于关键词的搜索方式,用户必须通过构造有限的查询词来表达信息需求 (information need)。Carpinetto 等<sup>[1]</sup>在查询扩展综述中明确指出,大多数用户喜欢构造短查询交给搜索引擎,且构造的查询词多以 1~3 个词居多;并且用户的查询构造本身就是一个抽象的过程,查询构造结果具有模糊性、不确定性和描述的多样性。在这种情况下,由于缺乏上下文语境,搜索引擎很难完全理解用户的查询意图,返回的结果中经常会包含大量无关或相似的文档。特别是当查询词出现歧义时,返回的文档集会偏向于某一个主题,而该主题往往并不是用户潜在查询意图<sup>[2]</sup>。如果搜索引擎能够将与用户初始查询构造相关的信息全部返回给用户,那么,用户就可以在多个不同查询结果中找到自己最想要的结果。文献[3]的研究表明,提高用户体验较好的办法就是给用户提供尽可能多的不同信息,而这些信息中至少会有一个是与用户需求相关的。

查询扩展可以有效地解决用户表达问题。其基本思想是

利用与关键词相关的词语对用户原始查询进行修正,弥补用户初始查询信息的不足,提高查全率。伪相关反馈 (Pseudo Relevance Feedback, PRF) 作为一种有效的自动查询扩展方法<sup>[4-6]</sup>,其假设初检查询结果集中排名靠前的  $k$  个文档是与用户查询相关的,记为伪相关文档集,并从中抽取扩展词进行查询扩展。该方法的查询效果主要受制于选取的前  $k$  个文档的数目及质量<sup>[7-8]</sup>,在其质量偏低的情况下,容易产生“查询主题偏移”现象。提升前  $k$  个相关文档的质量可以有效避免这种现象,形成真正与用户查询需求相关的伪相关文档集合。通常,改善伪反馈文档质量包括调整<sup>[9-11]</sup>和聚类<sup>[12]</sup>两种方法。其中,调整的方法包括对查询结果重排序和过滤两种方式:重排序的方法通过给查询结果集中各文档赋予不同的值来进行排序,通过构造算子<sup>[9]</sup>或是加权<sup>[10]</sup>完成;过滤的方法<sup>[11]</sup>主要通过给查询结果集中各文档添加若干特征,突显相关文档,提高相关文档的排名,从而达到过滤的目的。

以上这些伪相关反馈方法关注的重点仍是用户查询词的表象形式,而不是用户的内在实际信息需求,得到的伪相关文档中往往有很多是非常相似的,造成查询结果冗余的增加,不能很好地体现用户不同层面的查询需求<sup>[8]</sup>。本文研究认为,

收稿日期:2016-03-01;修回日期:2016-05-03。

基金项目:国家自然科学基金资助项目(61263037);内蒙古自然科学基金资助项目(2014BS0604, 2014MS0603)。

作者简介:闫蓉(1979—),女,内蒙古鄂尔多斯人,讲师,博士研究生,CCF 会员,主要研究方向:信息检索、自然语言处理; 高光来(1964—),男,内蒙古扎赉特人,教授,硕士,CCF 会员,主要研究方向:智能信息处理。

用户的查询需求并不是单一的,而是多层面和多角度的,要实现自动的查询扩展,就要求伪相关文档中的各文档内容既保证与用户原查询相关,又要保证其与用户多层次需求的一一映射关系,从而降低查询主题偏移的风险,进而获取与用户查询尽可能相关的信息来进行伪反馈。有鉴于此,本文提出一种提高伪反馈文档质量的排序模型 REM( REorder Model)。该模型从文档隐含语义角度出发,通过对初检查询结果集中各文档进行重调序的方式,提高与用户查询主题相关文档的位序,确保二次反馈扩展源的质量,进而提高检索效果。

## 1 基于检索结果排序的 PRF 模型

伪反馈文档质量不高实质上是由于搜索引擎对于用户查询理解不充分造成的,而要让搜索引擎完成这种充分理解是不大可能的。那么,如果能够将用户查询本身所有相关内容都尽可能地覆盖到,这样就可以在伪相关文档中减少不相关文档的数量,从而提高查询准确率。为了确保伪相关文档中各文档满足用户查询覆盖度的要求,本文提出一个排序模型 REM。该模型将初检查询结果文档集中的各文档依据满足用户查询意图相关度程度进行重新排序,选择排名靠前的 top- $k$  个文档来构造二次反馈的扩展源集合。

### 1.1 排序原则

Carbonell 等<sup>[13]</sup> 提出的最大边缘相关算法 (Maximal Marginal Relevance, MMR) 是用来解决查询结果多样化问题的一种方法。该算法分别对各文档与用户查询间的相关度和文本之间的相关度进行度量,所谓的边缘相关即为二者的线性组合。按照各文档的边缘相关最大化作为排序依据,提升在已有查询结果中与查询相关性尽量大、且与先前被选择的文档间相似性尽量小的文档的排名次序,完成对各文档的重定序。

本文的排序策略与 MMR 很类似,区别在于:本文认为初检查询结果集中的各文档还应当依据其与用户查询意图相关度高低来进行排序,并从排序结果中构造伪相关文档集。这就要求构造的 REM 排序模型,一方面要保证伪相关文档集中各文档与用户各层次查询需求的一一映射关系,另一方面要保证其中的文档间的相似度最小。本文假定初检查询结果各文档相关主题的语义集合涵盖了用户的查询需求。由此,构造 REM 模型的排序准则如下:排序结果集中的各文档要满足用户各层次查询需求,即需求覆盖度的最大化;同时还应保证各文档之间尽可能的不相似,即冗余度的最小化。

### 1.2 基于检索结果排序模型构造

按照上述排序准则,本文构造的排序模型 REM 描述如下。

假设: $C$  表示一个文档集合, $d_i$  是  $C$  中的文档, $d_i \in C$ ;  $Q$  表示一个用户查询, $Q$  的真相关 Judgments 集,记为  $RR$ ; 从  $C$  集中得到的以相关度为基础的初检结果文档集合,取其前  $K$  个文档构成初始伪相关文档集,记为  $D_r$ ;  $S$  表示  $D_r$  中已经被选取的文档集合, $d_j \in S$ ;  $Y = D_r - S$  表示未被选取的文档集合, $d_i \in Y$ ;  $\arg \max^K[*]$  表示给出集合  $K$  个最大元素的索引。MMR 算法计算如式(1) 所示:

$$MMR(d_i) = \arg \max_{d_i \in Y}^K [\lambda Sim_1(d_i, Q) -$$

$$(1 - \lambda) \max_{d_j \in S} (Sim_2(d_i, d_j)) ] \quad (1)$$

其中: $Sim_1(d_i, Q)$  表示查询  $Q$  与文档  $d_i$  之间的相似度, $Sim_2(d_i, d_j)$  表示两个文档  $d_i$  和  $d_j$  间的相似度, $\lambda \in [0, 1]$  是调节系数。本文构造的伪相关文档排序模型 REM 如式(2) 所示:

$$Rele\_score(d_i) = \arg \max_{d_i \in Y}^K [\lambda Rele\_Value(d_i, RR) * \\ Sim_1(d_i, Q) - (1 - \lambda) \max_{d_j \in S} (Sim_2(d_i, d_j))] \quad (2)$$

其中: $Rele\_value(d_i, RR)$  是文档  $d_i$  的排名调节函数,如式(3) 所示。

$$Rele\_Value(d_i, RR) = \frac{Count^-(S)}{Count^+(S)} \quad (3)$$

其中: $Count^-(S)$  和  $Count^+(S)$  分别表示  $S$  集中各文档不属于  $RR$  的文档个数和属于  $RR$  的文档个数。设置  $Rele\_value(d_i, RR)$  主要是针对查询  $Q$  的真相关文档而言的。若文档  $d_i \in RR$ , 应当在排序模型适当调整其位序,尽量让其排名靠前。具体方法:当文档  $d_i$  是查询  $Q$  的真相关文档,且与其他文档  $d_j$  的相似度低,则文档  $d_i$  的得分应该相应地提高;反之,则降低。由式(2)可以看出,文档  $d_i$  的排序结果是由它本身、与其相似文档和其是否为查询的真相关文档三者共同决定。图 1 所示为基于检索结果排序的伪相关反馈的实现过程。

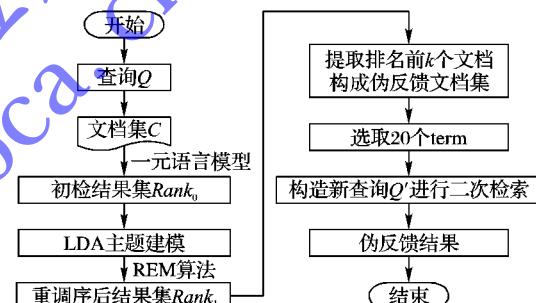


图 1 基于检索结果排序的伪相关反馈实现流程

## 2 文本相似度计算

式(2)中列出了两个相似度计算,它们是构造本文排序模型的关键。对于文本间相似度的计算方法大部分以基于向量空间模型<sup>[14]</sup>为主。该方法通过构造词典空间,将文本在词典空间表示为词向量的方式进行建模。但在真实数据集中构造的词典空间存在维度过高和数据稀疏的问题,而且在建模过程中未考虑文本中各词项的语义特征。在本文的排序模型中,目的是让伪相关反馈集中的各文档尽量满足用户各层面上的信息需求,那么,在相似度计算中,应该选取一种更合适的,能考虑文本中各词项语义特征的文本表示方法。近年来,主题模型——潜在狄利克雷分配 (Latent Dirichlet Allocation, LDA)<sup>[15]</sup> 被研究应用在文本相似度计算<sup>[16]</sup>中。LDA 通过引入隐含主题(latent topic)概念,在主题空间(topic space)中用有限主题数目将文档表示成低维的文档-主题向量,并且考虑了文本的语义特征,通过构造“词汇-主题-文档”模式来提取大规模数据集中潜在的主题(语义)信息。基于此,本文选用 LDA 主题模型抽象表示文本,用于计算文本间语义相似度。

信息检索本身对于词汇的精确度要求高。但是,LDA 在建模过程中抽象的主要对象是整个数据集,对应用 LDA 模型生成的文本来说,文本被表示成所有主题的特定比例的混合。

如果依此方式对用短文本构造的用户查询直接进行 LDA 建模,会由于数据稀疏的原因,使得这种文本表示结果不合适<sup>[17]</sup>,势必会造成检索性能较差。所以本文在实验过程中,仅对进行  $Sim_2(d_i, d_j)$  计算的两个文本进行 LDA 建模,而对  $Sim_1(d_i, Q)$  相似度计算,本文将直接利用经典的 BM25<sup>[18]</sup> 检索结果。对于 LDA 建模后的两个文本,本文使用 JS(Jensen-Shannon) 距离<sup>[19]</sup> 计算文本相似度,如式(4)所示:

$$D_{JS}(p, q) = \frac{1}{2} \left( D_{KL}\left(p, \frac{p+q}{2}\right) + D_{KL}\left(q, \frac{p+q}{2}\right) \right) \quad (4)$$

其中: $p$  和  $q$  分别表示两个随机变量的概率分布; $D_{KL}(p, q)$  表示随机变量  $p$  和  $q$  的 KL(Kullback-Leibler) 距离,如式(5)所示。

$$D_{KL}(p, q) = \sum_{j=1}^T p_j \ln(p_j/q_j) \quad (5)$$

JS 距离是介于 0 和 1 的实数,且 JS 距离与文本间相似度成反比,即 JS 距离越小,文本间相似度就越大。

### 3 实验设置及评价

#### 3.1 实验设置

##### 3.1.1 索引建立

本文使用 lemur(<http://www.lemurproject.org>) 工具建立文档索引和查询。实验数据集包括文档集和查询集,其中:文档集包括简体中文 Xinhua(2002—2005)四年的新闻文档,共 308 845 个文档;查询集包括简体中文 ACLIA2-CS(0001 ~ 0100),共 100 个查询。由于数据集为中文数据,所以在进行检索和查询前,首先对文档集和查询集都进行了预处理,包括分词(采用的是中国科学院计算技术研究所的 ICTCLAS)和去除停用词。

##### 3.1.2 LDA 建模

在进行 LDA 建模前,为了降低少数低频词对文本建模结果的影响,对实验文档集作了进一步的预处理,去除部分虚词、形容词、副词等意义不大的词;删除文档集中出现频度小于 5 的词汇。最后对剩余的 65 082 429 个词项进行 LDA 主题建模。LDA 建模的参数估计利用 MCMC(Markov Chain Monte Carlo)方法中的 Gibbs 抽样<sup>[20]</sup> 算法。初始设置主题个数  $M = 10$ ,  $\alpha = 50/M$ ,  $\beta = 0.01$ , Gibbs 抽样的迭代次数为 100。

LDA 建模过程中主题数目  $M$  的设置非常关键,主要是因为主题数目与数据集密切相关。用 LDA 对数据建模后,数据会通过主题进行高度抽象和压缩,主题数目的设置应当以数据为根本,因为不同的主题数目会导致每个主题-词项分布结果的不一样,直接影响文本的语义表达。所以对于不同的数据集,主题数目  $M$  的取值是不固定的。困感度(Perplexity)<sup>[21]</sup> 可以用来评价主题模型的生成性能,本文采用该方法作为评价指标来确定最佳主题数目  $M$ 。一般地,困感度取值越低,就表示模型更能发现数据中深层次的语义结构,模型的推广性就越好。困感度的计算如式(6)所示:

$$Perplexity(R) = \exp\left(\frac{\sum_{n=1}^N \lg(P(r_n))}{\sum_{n=1}^N L_n}\right) \quad (6)$$

其中: $R$  表示具有  $N$  个文档的测试集,  $L_n$  为第  $n$  篇文本  $r_n$  的长度,  $P(r_n)$  表示模型产生文本  $r_n$  的概率。

本文实验中,依次取主题个数  $M = 10, 20, \dots, 100$ , 分别对 LDA 建模,分析困感度的变化。实验结果如图 2 所示。从图 2 可以看出,当  $M = 60$ , 模型困感度达到最小峰值,此时模型的生成性能最佳。因此,实验中选取主题数目  $M = 60$ 。

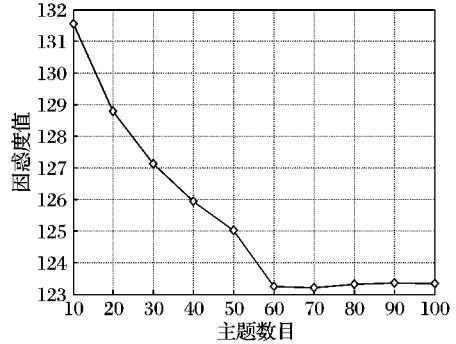


图 2 Perplexity 值随主题个数变化曲线

#### 3.2 实验评测标准和结果分析

初检的相关度排序方法选用的是典型的一元语言模型(Language Model, LM)方法,采用 Dirichlet 平滑方法,设置值为 1000。LM 是基于词项空间的统计结果来对用户查询和文档的相关度进行计算的,并没有考虑词语所表达的语义信息。选取其结果作为初检结果的目的,是为了验证引入表达语义信息的文本表示方法后,从浅层语义的角度是否可以通过文档顺序的调整来达到提升扩展源质量的目的。因为大多数用户在检索过程中主要关注排名靠前的结果,实验结果应该考察其是否符合大多数检索用户的习惯,所以实验中主要从查询准确率方面进行评价,分别采用前  $n$  个结果的查准率  $Precision@n$ (简记为  $P@n$ ) 和平均查准率(Mean Average Precision, MAP) 来衡量。

实验中初检查询结果文档个数设定为  $K = 50$ ,并设置统一从排名前 10 个文档(即  $k = 10$ ) 中抽取扩展词。文献[22]研究表明,扩展词个数的数目设定为 10 ~ 20 时,检索效果最好,所以实验中设置  $feedbackTermCount = 20$  进行伪反馈。Baseline 选取标准的 BM25<sup>[18]</sup> 伪反馈。REM 算法中关于参数  $\lambda$  取值,本文采用贪心策略,当  $\lambda$  取 0.7 时,检索效果最好。为了有效验证 REM 方法,本文还和 TF-IDF(Term Frequency-Inverse Document Frequency) 伪反馈方法进行了比较。实验结果如表 1 所示。

表 1 本文方法和两种传统 PRF 方法检索结果对比

方法	MAP	P@5	P@10
BM25	0.2690	0.5239	0.5056
TF-IDF	0.2815	0.5296	0.4817
REM	0.3098	0.5355	0.5000

从表 1 的结果可以看到,REM 方法比 Baseline(BM25) 和 TF-IDF 伪反馈方法在 MAP 和 P@5 指标上有了明显的提高,说明 REM 方法对于提高检索效果是有效的;但在指标 P@10 上的结果略有下降,该结果其实正体现了本文的核心思想,即实际应用中对于搜索引擎提供的查询结果,应该做到查询的多样性与查询内容的相关性及有用性的折中。

为了进一步验证本文提出方法的有效性,将 REM 方法结果与直接对初检结果利用 MMR 算法进行调序的结果(VSM\_PRF)进行了比较。在 VSM\_PRF 方法中,文档采用向量空间

模型文本表示方法，并基于 Cosine 系数计算文档间相似度。这样做，还可以比较两种不同文本表示方法对于检索效果的影响。另外，本文同时还与初始结果直接进行伪反馈的结果 (LM\_PRF) 进行了比较。结果如表 2 所示。

从表 2 结果可以看出，在各项评测指标上，VSM\_PRF 和 REM 均明显高于 LM\_PRF 检索结果，说明从文本语义角度出发对初检结果进行重排序的方法是切实可行的。另外，对 MMR 结果进一步改进，可以达到更好的检索效果，REM 方法在 MAP 指标上比 VSM\_PRF 高出 6.4%，表明引入主题空间的统计信息，可以更有效地改善词项空间的统计结果；但二者在 P@5 和 P@10 指标上相差无几，主要是由于本文提出的算法对于文本的主题建模精度要求高所造成的。

表 2 三种方法的检索评价指标对比

方法	MAP	P@5	P@10
LM_PRF	0.2603	0.4767	0.4575
VSM_PRF	0.2912	0.5389	0.4917
REM	0.3098	0.5355	0.5000

## 4 结语

主流的关键词查询表达多样性使得传统的查询扩展会发生“查询主题偏移”问题，为此提出一种新的伪相关反馈方法，通过引入排序模型 REM 对初检结果文档集中各文档进行重排序，从而获取高质量伪相关文档，减小查询主题偏移的风险。实验结果验证了本文提出方法的有效性。与传统的伪反馈方法比较而言，本文提出的 REM 模型更有助于提高查询效果；而且实验结果还表明，在重排序过程中，与基于词汇级别的文本建模方式相比，基于主题级别的文本建模方式能够获取更多的语义信息，有助于提升伪相关文档的质量，改善检索效果。

本文将浅层语义应用于文本相似度计算中，并将其用于解决实际的检索问题进行了初步尝试。但在实际的检索实现中，需要用户的参与或分析和挖掘用户检索行为来获取与用户查询真相关的 RR 集合，这是一件很困难的事情。所以进一步的工作重点在于，在对大规模文本数据集进行主题建模基础上，有效利用隐藏在伪反馈文档中的主题信息，进而提取与用户查询相关的语义信息，以达到用浅层语义指导检索过程的目的。

## 参考文献：

- [1] CARPINETO C, ROMANO G. A survey of automatic query expansion in information retrieval [J]. ACM Computing Surveys, 2012, 44(1): Article No. 1.
- [2] VARGAS S, SANTOS R L T, MACDONALD C, et al. Selecting effective expansion terms for diversity [C]// OAIR2013: Proceedings of the 10th Conference on Open Research Areas in Information Retrieval. Paris: Le Centre de Hautes Etudes Internationales D'Informatique Documentaire, 2013: 69–76.
- [3] TEEVAN J, DUMAIS S T, HORVITZ E. Characterizing the value of personalizing search [C]// SIGIR2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2007: 757–758.
- [4] COLLINS-THOMPSOM K. Reducing the risk of query expansion via robust constrained optimization [C]// CIKM2009: Proceedings of the 18th ACM Conference on Information and Knowledge Management. New York: ACM, 2009: 837–846.
- [5] RAMAN K, UDUPA R, BHATTACHARYA P, et al. On improving pseudo-relevance feedback using pseudo-irrelevant documents [C]// ECIR 2010: Proceedings of the 32nd European Conference on IR Research, LNCS 5993. Berlin: Springer-Verlag, 2010: 573–576.
- [6] ZHAI C, LAFFERTY J. Model-based feedback in the language modeling approach to information retrieval [C]// CIKM2001: Proceedings of the 10th International Conference on Information and Knowledge Management. New York: ACM, 2001: 403–410.
- [7] HUANG Q, SONG D, RÜGER S. Robust query-specific pseudo feedback document selection for query expansion [C]// ECIR 2008: Proceedings of the 30th European Conference on IR Research, LNCS 4956. Berlin: Springer-Verlag, 2008: 547–554.
- [8] HE B, OUNIS I. Studying query expansion effectiveness [C]// ECIR 2009: Proceedings of the 31th European Conference on IR Research, LNCS 5478. Berlin: Springer-Verlag, 2009: 611–619.
- [9] MITRA M, SINGHAL A, BUCKLEY C. Improving automatic query expansion [C]// SIGIR1998: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 1998: 206–214.
- [10] AMO P, FERRERAS F L, CRUZ F, et al. Smoothing functions for automatic relevance feedback in information retrieval [C]// DEXA 2000: Proceedings of the 11th International Workshop on Database and Expert Systems Applications. Washington, DC: IEEE Computer Society, 2000: 115–119.
- [11] 叶正. 基于网络挖掘与机器学习技术的相关反馈研究[D]. 大连: 大连理工大学, 2011: 51–55. (YE Z. The research of machine learning techniques and external Web resources for relevance feedback [D]. Dalian: Dalian University of Technology, 2011: 51–55.)
- [12] PU Q, HE D. Pseudo relevance feedback using semantic clustering in retrieval language model [C]// CIKM2009: Proceedings of the 18th ACM Conference on Information and Knowledge Management. New York: ACM, 2009: 1931–1934.
- [13] CARBONELL J, GOLDSTEIN J. The use of MMR, diversity-based reranking for reordering documents and producing summaries [C]// SIGIR 1998: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 1998: 335–336.
- [14] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing [J]. Communications of the ACM, 1975, 18(11): 613–620.
- [15] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993–1022.
- [16] ZHOU D, DING Y, YOU Q, et al. Learning to rank documents using similarity information between objects [C]// ICONIP 2011: Proceedings of the 18th International Conference on Neural Information Processing, LNCS 7063. Berlin: Springer-Verlag, 2011: 374–381.
- [17] HONG L, DAVISON B D. Empirical study of topic modeling in twitter [C]// SOMA '10: Proceedings of the First Workshop on Social Media Analytics. New York: ACM, 2010: 80–88.

(下转第 2143 页)

别器 MI 采用了高效的删减策略与匹配方法,有效地缩减了匹配空间,大大减少执行代价相对较高的属性匹配器 AM 需要处理的数据量,从而获得了较好的运行效率。

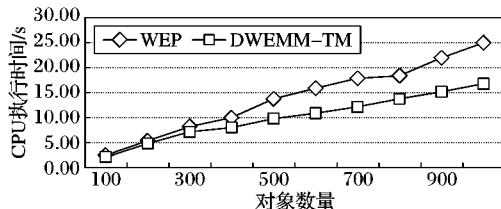


图 3 DWEMM-TM 与 WEP 运行效率比较

## 6 结语

Deep Web 数据的迅速增长对实体匹配的效率和性能提出了更高要求。本文借鉴聚类思想,将实体匹配看作类簇归并过程,提出基于二次归并的 Deep Web 实体匹配方法 DWEMM-TM,将不同的匹配关系分阶段交予不同的处理模块,使匹配逐渐精化,并通过引入自匹配,实现两次归并间的自动转接和不同匹配策略的自动选择。实验结果表明,DWEMM-TM 方法在缩减匹配空间、降低复杂数据处理量和提高匹配精度方面有不错表现,达到了性能与效率的双重提高。

### 参考文献:

- [1] 陈丽君,林怀忠.一种用于深层网接口集成的模式匹配方法[J].计算机工程,2012,38(12):42–44. (CHEN L J, LIN H Z. Pattern matching method for Deep Web interface integration [J]. Computer Engineering, 2012, 38(12): 42 – 44.)
- [2] KÖPCKE H, RAHM E. Frameworks for entity matching: a comparison [J]. Data & Knowledge Engineering, 2010, 69(2): 197 – 210.
- [3] HAN X, SUN L, ZHAO J. Collective entity linking in Web text: a graph-based method [C]// SIGIR '11: Proceedings of the 34th Annual ACM SIGIR Conference on Research and development in Information Retrieval. New York: ACM, 2011: 765 – 774.
- [4] RASTOGI V, DALVI N, GAROFALAKIS M. Large-scale collective entity matching [J]. Proceedings of the VLDB Endowment, 2011, 4 (4): 208 – 218.
- [5] WANG Z, LI J, WANG Z, et al. Cross-lingual knowledge linking across Wiki knowledge bases [C]// WWW '12: Proceedings of the 21st International Conference on Word Wide Web. New York: ACM, 2012: 459 – 468.
- [6] FAN J, LU M, OOI B C, et al. A hybrid machine-crowdsourcing system for matching Web tables [C]// Proceedings of the 2014 IEEE 30th International Conference on Data engineering. Washington, DC: IEEE Computer Society, 2014: 976 – 987.
- [7] 崔晓军,肖红宇,丁立新.基于距离的自适应 Web 数据库记录匹配方法[J].武汉大学学报(理学版),2012,58(1):89 – 94. (CUI X J, XIAO H Y, DING L X. Distance-based adaptive record matching for Web database [J]. Journal of Wuhan University ( Science Edition), 2012, 58(1): 89 – 94.)
- [8] LIU W, MENG X. A holistic solution for duplicate entity identification in deep Web data integration [C]// SKG '10: Proceedings of the 2010 Sixth International Conference on Semantics, Knowledge and Grids. Washington, DC: IEEE Computer Society, 2010: 267 – 274.
- [9] 徐红艳,党晓婉,冯勇,等.基于 BP 神经网络的 Deep Web 实体识别方法[J].计算机应用,2013,33(3):776 – 779. (XU H Y, DANG X W, FENG Y, et al. Method of Deep Web entities identification based on BP network [J]. Journal of Computer Applications, 2013, 33(3): 776 – 779.)
- [10] LIU W, MENG X, YANG J, et al. Duplicate identification in Deep Web data integration [C]// WAIM '10: Proceedings of the 11th International Conference on Web-age Information Management, LNCS 6184. Berlin: Springer-Verlag, 2010: 5 – 17.
- [11] 李亚坤,王宏志,高宏,等.基于实体描述属性技术的 XML 重复对象检测方法[J].计算机学报,2011,34(11):2131 – 2141. (LI Y K, WANG H Z, GAO H, et al. Efficient entity resolution on XML data based on entity-describe-attribute [J]. Chinese Journal of Computers, 2011, 34(11): 2131 – 2141.)
- [12] EFTHYMIOU V, PAPADAKIS G A, PAPASTEFANATOS G, et al. Parallel meta-blocking: realizing scalable entity resolution over large, heterogeneous data [C]// Proceedings of the IEEE 2015 4th International Conference on Big Data. Piscataway, NJ: IEEE, 2015: 411 – 420.
- [13] 寇月,申德荣,李冬,等.一种基于语义及统计分析的 Deep Web 实体识别机制[J].软件学报,2008,19(2):194 – 208. (KOU Y, SHEN D R, LI D, et al. A Deep Web entity identification mechanism based on semantics and statistical analysis [J]. Journal of Software, 2008, 19(2): 194 – 208.)
- [14] MCCALLUM A. Cora citation matching [EB/OL]. (2004-02-09)[2015-08-22]. <http://www.cs.umass.edu/~mccallum/data/cora-refs.tar.gz>.

### Background

This work is partially supported by the National Research Fund for Educational Information Technology (136241401), the Research Foundation of Zhejiang Yuexiu University of Foreign Languages (N201375).

**CHEN Lijun**, born in 1979, M. S., lecturer. Her research interests include Deep Web data mining, intelligent information processing, information technology in education.

(上接第 2102 页)

- [18] 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.
- [19] LIN J. Divergence measures based on Shannon entropy [J]. IEEE Transactions on Information Theory, 1991, 37(14):145 – 151.
- [20] GRIFFITHS T L, STEYVERS M. Finding scientific topics [J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(Supp 1): 5228 – 5235.
- [21] BLEI D B, LAFFERTY J D. Correlated topic models [C]// NIPS 2005: Advances in Neural Information Processing Systems 18.

Cambridge, MA: MIT Press, 2005, 18: 147 – 155.

- [22] OGILVIE P, VOORHEES E, CALLAN J. On the number of terms used in automatic query expansion [J]. Information Retrieval, 2009, 12(6): 666 – 679.

### Background

This work is partially supported by the National Natural Science Foundation of China (61263037), the Natural Science Foundation of Inner Mongolia Autonomous Region (2014BS0604, 2014MS0603).

**YAN Rong**, born in 1979, Ph. D. candidate, lecturer. Her research interests include information retrieval, natural language processing.

**GAO Guanglai**, born in 1964, M. S., professor. His research interests include intelligent information processing.