基于主题词对的文档重排方法

何婷婷[1,2]，许婷[2]，瞿国忠[2]，涂新辉[2]

（1.清华大学软件学院，北京 102201；2.华中师范大学计算机科学系，武汉 430079）

摘 要： 信息检索中相关文档的排序一直是一个至关重要的问题。本文提出一种基于主题词对的文档重排方法，使得检索结果在保持召回率的前提下提高精确率。主题词对意指能够共同表征同一主题的两个词语，其中一个来自于查询，另一个来自于文档，两者之间具有紧密的联系。本文中，主题词对的选择采用概率潜在语义索引的方法，并根据主题词对在文档中的分布状况对其进行重排。对 NTCIR-5 中文信息检索的文档集合进行测试，采用 trec 标准评估方法，结果表明采用该方法使得精确率在 rigid 和 relax 结果集上分别提高了 53.6% 和 55.8%。

关键词： 主题词对；概率潜在语义索引；文档重排；

Re-ranking Based on Topic Word Pairs

Tingting He[1,2], Ting Xu[2], Guozhong Qu[2], Xinhui Tu[2]

(1.Software College of Tsinghua University, Beijing 102201; 2.Department of Computer Science, Huazhong Normal University, Wuhan 430079)

Abstract: How to improve the rankings of the relevant documents plays a key role in information retrieval. In this paper, a re-ranking approach based on topic words pair is proposed to improve precision while recall is preserved. The topic word pairs contain two correlated words, one of which is the original query word and the other come from the documents. The selection is based on Probabilistic Latent Semantic Indexing (PLSI). Then, the distribution of the word pairs is used to re-rank documents. Results show a 53.6% and 55.8% improvement compare to the initial retrieval without any re-ranking or query expansion on NTCIR-5 document collection for SLIR.

Keywords: topic word pair; plsi; document re-ranking;

1 引言

如何有效、准确地从海量信息中向用户提供所需要的内容，是中文信息检索领域的基本目标。众多的检索模型、索引机制、查询扩展方法等各项研究的长足进步都为该领域的发展提供了极大的支持。而用户往往并不止步于得到了需要的内容，而是希望检索系统能够尽可能使相关文档排在最前，使需要的内容一目了然。因此结果文档集合的排序问题也在信息检索领域扮演了关键角色。文档重排是在首次检索的基础上对结果进行重新排序的一种方法，目的就在于保持召回率并提高精确率。一般地，有三类信息被用来进行文档重排：

1. 文档信息

Kyung-Soon Lee et al. 曾利用文档的聚类信息进行重排，他们综合考虑首次检索的结果和query-聚类的相似度，最终确定文档的排序。Jaroslaw Balinski和Czeslaow Danilowicz也曾经利用文档间的距离来进行重排。

2. 查询信息

基金资助：国家自然科学基金（60442005）；教育部科学技术研究重点项目（105117）；

作者简介：何婷婷(1964-)，女，湖北，教授，硕士生导师，博士，tthe@mail.ccnu.edu.cn.

Mitra et al. 在重排的过程中，考虑到了查询词之间的关系，利用这种关系对查询进行限制，并重新考虑文档的排名。查询词所在位置信息也被Qu Youli et al.利用于重排。

3． 额外信息

额外信息是指除文档本身信息以及查询信息之外的附加内容，例如词典以及本体。Dequan Zheng et al. 曾提到利用本体对文档进行重排序方法。

由于词语的多样性，单纯的利用查询信息对文档进行重排往往忽略了词的语义，使得那些与查询内容相同但是表述不同的文档排名靠后。而仅仅利用附加信息进行重排也会遇到诸多困难，首先词典或者本题的构造仍然是一项困难的任务。此外，将附加信息添加到查询中类似与查询扩展，有可能造成查询主题的扩散。

因此本文综合采用文档信息和查询信息，提出一种基于主题词对分布的重排方法。其主要过程如下：1、首次检索。2、选择主题词对，所谓主题词对是指能够共同表征某一主题的词对，其中有且仅有一个词来自于查询，另一个词来自于文档。两者之间具有很强的语义关联。利用主题词对为文档进行重新排序，是一个对查询主题的聚焦过程，而不是扩散过程。该方法使得查询意图更加明确，对相关文档能够产生很强的指示信息。因此，仅仅只有同时包含词对中两个词的文档将被影响。进而能够改善文档的排名状况，提高结果的精确率。主题词对的选择利用概率潜在语义索引，以期望得到语义上最相关的若干对词。3、利用主题词对的分布对文档进行重新打分并排序。

本文结构如下：下一节简要描述了混合索引以及检索模型。第三节介绍了主题词对及其选取方法。第四节给出利用主题词对的分布状况对文档进行重排的过程。第五节测试了我们的方法，评估结果并给出结果分析。最后，进行总结。

2 检索模型与索引机制

大量研究工作比较了不同的索引单元对检索效果的影响。单字、词或二元都可以用作索引的单位。基于单字的索引模型具有最好的召回率，而基于词和二元的索引则具有最好的准确率。与基于词的索引不同的是，二元索引不会受到未登录词的困扰，但是却需要消耗过多的存储空间。因此，取长补短我们采用词与二元的混合索引[1]机制。首先自动分词，当出现未登陆词时改用二元。最后，所有这些词与二元形成索引的基本单元。在我们的实验中，采用 tfidf 检索模型。

3 主题词对

主题词对是指能够共同表征某一主题的词对，其中有且仅有一个词来自于查询，另一个词来自于文档。两者之间具有很强的语义关联。设 $q = \{ q_1, q_2 \cdots q_k \}$是原始的查询，其中 q_i 为查询词，$D=\{d_1, d_2, \cdots d_n\}$是对 q 首次检索的结果文档集合。L 是词对的集合，每一对有且仅有一个词来自于查询[2]。L 可被表示为：

$$L = \{ (w_i, w_j, \text{association\_intensity}) \mid w_i \in q \text{ and } w_j \notin q \}$$

利用 D 中的前 1000 篇文档与 q 我们抽取出所有满足条件的词对，并根据概率潜在语义分析[3]计算每一对词的关联度。继而根据关联度选择最能够表征同一种主题的词对若干个，得到主题词对集合 L，用于重排。以下详细描述了我们的主题词对选择方法。

3.1 构造语义空间

3.1.1. 初始化

初始化 $p(z_1)= p(z_2)=\ldots= p(z_n)=\frac{1}{n}$， z 为某种潜在语义，n 为语义个数，其取值由经验决定，在本文实验中根据所采用的数据集大小发现 n 取 60 时效果较好。为 P (z | d)和 P (w | z)赋值，得到两个矩阵 $P(w_j \mid z_k)$ 和 $P(z_k \mid d_i)$，

使其满足$\sum_{j=1}^{M} P(w_j \mid z_k)=1$，$\sum_{k=1}^{K} P(z_k \mid d_i)=1$。

3.1.2. 采用 EM 算法求得结果

概率潜在语义分析使用最大期望(EM)算法对潜在语义模型进行拟合。在使用随机数初始化之后，交替实施 E 步骤和 M 步骤 进行迭代计算。

E-step:

$$P(z \mid w, d) = \frac{P(z)P(w \mid z)P(d \mid z)}{\sum_{z' \in Z} P(z')P(w \mid z')P(d \mid z')} \tag{1}$$

M-step:

$$P(w \mid z) = \frac{\sum_{d} m(d,w)P(z \mid d,w)}{\sum_{d,w'} m(d,w')P(z \mid d,w')} \tag{2}$$

$$P(d \mid z) = \frac{\sum_{w} m(d,w)P(z \mid d,w)}{\sum_{d',w} m(d',w)P(z \mid d',w)} \tag{3}$$

$$P(z) = \frac{\sum_{d,w} m(d,w)P(z \mid d,w)}{\sum_{d,w} m(d,w)} \tag{4}$$

根据最大似然估计原理，迭代(1)- (4)达到(5)的最大化，借此过程推导出模型的所有参数：

$$\lambda = \sum_{d \in D} \sum_{w \in W} m(d,w) \log P(d,w) \tag{5}$$

3.2 主题词对的选取

利用以上结果得到词 w 的向量 $P(w \mid z_i)$，因此词与词之间在潜在语义空间上的关系也可表示为两词向量的夹角。我们根据公式(6)计算词对的关联度：

$$P_{ij} = Cos(W_i, W_j) = \frac{\sum_{z \in Z} p(w_i \mid z)p(w_j \mid z)}{\sqrt{\sum_{z \in Z} p(w_i \mid z)^2}\sqrt{\sum_{z \in Z} p(w_j \mid z)^2}} \tag{6}$$

所有仅包含一个查询词的词对将按照此关联度进行排序，显然，位于前方的词对具有很强的语义关联，P_{ij} 越大，两个词能够共同表征同一主题的能力越强。因此选择前 N 个形成主题词对，在后续章节将利用这些主题词对的分布状况进行重排。

4 文档重排

文档的重排指在首词检索的基础上对其进行重新排序，而不需要二次检索的一种方法。它期望与查询越相关的文档，排名越靠前。本节中详细描述了根据主题词对的分布状况对文档进行重排的过程，我们融合了主题词对在文档中的跨度、主题词对的相对文档频率及其相关度等信息：

1. 主题词对的跨度 span(a)：主题词对 a 中两词在文档中的距离$|\ position(w_i)-position(w_j)\ |$。其中 position(w)表示词 w 在文档中的位置序号[4]。跨度越小，主题词对的凝聚力越大，文档的权重就越大。
2. 主题词对的相关度 P_a：根据公式(6)计算得出。相关度越大，主题词对表征同一语义的能力就越强。包含该主题词对的文档主题就越明确，与查询的相似度越大。
3. 相对文档频率：主题词对在首次检索结果中的前 1000 篇文档中的文档频率与在所有文档中的文档频率的比

例。很明显，比例越大表示该主题词对越重要[5]。

4. 查询与文档间的原始相似度 S：首次检索时的相似度。

综合考虑以上因素后，文档被重新算分：

$$\text{DocumentScore(d)} = (\sum_{a \in L} \frac{P_a \times df(a,D)/1000}{span(a) \times Df(a,C)/|C|} + 1) \times S \tag{7}$$

其中 d 为文档，a 为某一主题词对，L 为主题词对集合， df(*a* ,D)和 Df(*a*, C)表示 a 在集合 D 以及 C 中的文档频率。最后根据得出的分数，对文档进行重排序。

5 实验与结果评估

5.1 测试语料

我们采用 NTCIR-5 中文信息检索的测试集合，选取其中 20 个查询，并利用其 DESC 部分作为查询进行检索和重排。实验结果利用标准 TREC 评估工具计算平均精确率（MAP）以及检索出 R 篇文档时的精确率。

5.2 实验与结果

本节给出利用主题词对进行文档重排对检索结果的影响。作为比较的基线， Lemur4.1 工具包被用来进行首次检索。

在第一组实验中，我们比较了使用不同个数的主题词对进行文档重排的效果。由于太多的词对会给原始的查询主题带来噪音，而太少词对则不能携带足够的信息，不能很好的概括查询主题，因此我们提出一个假设：重排所使用的词对个数取决于原始查询的长度。假设原始查询的长度为 m（排除停用词后），根据经验，自动提取 2m-1 个词对用语重排。表 1 列出了测试结果。分别列出了在 rigid 和 relax 测试条件的平均精确率。列[normal]给出首次检索后的结果，列[enhanced]给出使用重排后的平均精确率。行[m=3]给出当原始查询长度为 3 时的实验结果，依次类推。

表 1 不同长度查询的平均精确率(MAP)

Length of query	Rigid		Relax	
m	normal	enhanced	normal	enhanced
m=3	0.2210	0.3112	0.2415	0.3781
m=4	0.2018	0.3247	0.2387	0.3725
m=5	0.2154	0.3031	0.2328	0.3547

从表中结果可以看出，采用主题词对进行重排能够明显改善检索的结果。

表 2 给出进行文档重排后检索出 R 篇文档时的平均精确率，经过重新排序后，排名靠前的文档中相关文档的数量得到明显增加。其中在 relax 集合下，PreAt5 由 0.4960 上升为 0.5720,而在 rigid 集合下，PreAt5 由 0.3520 上升为 0.3980。分别提高了 15%和 13%。使得相关文档能够更加集中的出现在靠前的位置。

表 2 R 篇文档时的平均精确率（MAP）

Precision at	relax		rigid	
R documents	Original	After re-ranking	Original	After re-ranking
At 5 docs	0.4960	0.5720	0.3520	0.3980
At 10 docs	0.4660	0.5200	0.3420	0.3680
At 15 docs	0.4587	0.4953	0.3293	0.3487
At 20 docs	0.4360	0.4660	0.3150	0.3370

第二组实验中，我们比较了主题词对的选择方法对于结果的影响。分别采用本文所述的 PLSI 方法以及互信息作为词对的选择依据。表 3 给出了其结果。列[Rigid]给出在 Rigid 评估数据集下，实验中 20 个查询的平均精确率；同样，列[Relax]给出 Relax 评估数据集下，20 个查询的平均精确率。实验发现，两种方法选择词对能够达到相似的效果。因此主题词对的选择并不依赖于特定的算法，这一结论对于探讨主题词对改进检索系统的结果排序有一般性意义。

表 3 两种主题词对选择方法的平均精确率（MAP）

different topic word pair selection	Rigid		Relax	
	MAP	%Change	MAP	%Change
Baseline	0.2024	-	0.2407	-
PLSI	0.3110	53.6%	0.3750	55.8%
MI	0.2980	47.2%	0.3772	56.7%

5.3 结果分析

实验数据显示总体上该方法使得检索结果得到了较大提升，但是对于某些查询，效果也不尽如人意。例如查询 001 查询时代华纳与美国线上合并案的後续影响：分析发现，对该查询的首次检索结果进行重排无效的原因在于错误的分词。例如“时代华纳”将被划分为“时代/华纳”，继而“时代”将导致错误的意思“时间”，而不是作为“时代华纳”公司名称的一部分。因此利用包含“时代”的主题词对进行文档重排将会使谈论时间的文档排名提前，但这并不符合我们的初衷。该类问题也许可由改变索引单位解决，如果采用术语自动抽取，并利用术语作为索引单元，诸多类似词语将被作为整体被划分，从而改善重排的效果。

6 结语

本文介绍了一种基于主题词对的文档重排方法，该方法是一个对查询主题的聚焦过程，而不是扩散过程。它使得查询意图更加明确，对相关文档能够产生很强的指示信息。因此在保持结果文档召回率的前提下能够明显提高结果的平均精确率。

实验中，我们对如何选取主题词对的个数以及采用何种方法选择词对做了分析，发现主题词对的选择并不依赖于特定算法，这一结论对于探讨主题词对改进检索系统的结果排序有一般性意义。同时在实验中我们也发现该方法在分词错误的情况下会失效，因此在下一步的研究中，我们试图采用术语作为索引的基本单元，从而减小分词错误给重排结果带来的影响。同时，提高主题词对的质量，扩大实验规模，对主题词对的选择方法进行更深入的研究。

参考文献：

[1] Tsang, T.F., R.W.P. Luk and K.F. Wong, Hybrid term indexing using words and bigrams, Proceedings of IRAL 1999, Academia Sinica, Taiwan, 112-117, 1999.

[2] David Carmel, Eitan Farchi, Yael Petruschka, Aya Soffer, Automatic Query Refinement using Lexical Affinities with Maximal information Gain. In Proceedings of the ACM SIGIR'02 Conference, Tampere, Finland, August 11-15, 2002.

[3] Hofmann, T. Probabilistic latent semantic analysis. In Proceedings of the 15th Conference on Uncertainty in AI (1999).

[4] Olga Vechtomova, Murat Karamuftuoglu. Approaches to High Accuracy Retrieval: Phrase-Based Search Experiments in the HARD track.

[5] L.P. Yang, D.H. Ji. I2R at NTCIR5. Proceedings of NTCIR-5 Workshop Meeting, December 6-9, 2005, Tokyo, Japan.