

个性化信息检索方法研究

王俊义 叶新铭

(内蒙古大学计算机学院 呼和浩特 010021)

摘 要 个性化信息检索是十分有用的检索方法,用户模型能够表示用户个人的爱好与兴趣,有许多研究工作以各种方式使用用户模型扩充问句。提出一种新的基于一元语言模型的方法。它通过对包含多个主题域的长期用户模型的学习得到相关的语义内容,对问句进行扩展后进行检索,得到更接近用户兴趣的结果,然后再与伪相关反馈模型相结合,进一步提高检索性能。通过实验证明,该方法取得了较好的效果。

关键词 信息检索,语言模型,个性化,问句扩展

中图法分类号 TP391.1 **文献标识码** A

Research of Personalized Methods of Information Retrieval

WANG Jun-yi YE Xin-ming

(College of Computer Science, Neimongol University, Hohhot 010021, China)

Abstract Personalization to information retrieval is very useful in information retrieval, the user profile can be used to represent the favorites or interests of an user. Many approaches to personalization have been studied in expanding query with user profile. We proposed a navel method which use the context of long-term user profile with multiple-domain to expand query model under the unigram language modeling framework, uses the new query model to retrieve and get more interesting results for users. Then combined with psudo-relevance feedback model, the method get better performance. Experimental results show that the proposed method is effective.

Keywords Information retrieval, Language model, Personalization, Query expansion

1 引言

信息检索是人们从海量信息中获取有用信息的必不可少的技术。随着信息的急剧增长,用户对信息检索技术的性能、多样性和个性化提出了更高的要求。信息检索有多种模型,如布尔模型、向量模型、概率模型和语言模型。自 1998 年以来,语言模型在信息检索中的应用取得了很好的效果^[1]。通常用户对信息需求的表达不够准确和明晰,往往只是几个单词,影响了检索性能。扩展问句和扩展文档是提高信息检索性能的有效方法。相关性反馈^[2,3]方法和在语言模型架下的各种扩展问句方法^[4-9]有效地提高了信息检索性能。个性化检索无疑是十分有用的,许多个性化方法都是以用户的短期模型或长期模型对问句进行扩展或对文档进行扩展,问句扩展与局部相关反馈技术可以在语义上支持个性化检索。

本文提出了一种新的基于长期用户模型的问句扩展方法和局部相关反馈相结合的方法,其考虑的应用情景是一个用户在一段时间以来浏览和点击过的文档组成的用户模型,它包含了用户所感兴趣的内容。当用户进行一个新的查询时,通常这个查询与其以前的兴趣是相同的或是有关联的,所以通过对用户模型的学习,会得到相关的语义内容,对问句扩展

后进行检索,可得到更接近用户兴趣的结果。我们在 TREC (Text Retrieval Conference)的几个文档集上,用 50 个问句进行了实验,结果表明我们的方法有较好的效果。

2 相关工作

本文工作所涉及到的相关主要模型和技术有语言模型、相关反馈和用户模型。

信息检索使用的语言模型是以概率与统计理论为基础的,主要思想是计算问句与文档的条件概率 $P(Q|D)$,文档依据它们的概率进行排名。一元语言模型主要有两种排名模型:query likelihood model^[1]和 KL-divergence^[10]。前者不易进行问句扩展,后者不仅给每个文档建模,也给问句建模,可以方便地进行问句扩展。

相关反馈是将初次检索到的部分文档交给用户进行相关性鉴别,以用户标记的相关文档扩充问句^[2]再进行检索,实际使用了一种短期的用户模型,能将用户所感兴趣的东西扩充进来,但在检索过程中需要和用户交互。伪相关反馈是认为初次检索到的前 k 个文档是相关的,并用它们扩充问句后进行检索^[12],该方法没有将用户的个人兴趣结合进来。语言模型能更自然地、更精确地对问句或文档进行推理式的扩充^[6]。

到稿日期:2009-07-08 返修日期:2009-09-24 本文受国家自然科学基金项目(60863015),内蒙古自然科学基金重点项目(20080404ZD20)资助。

王俊义(1958—),男,教授,硕士生导师,CCF 会员,主要研究方向为信息检索和计算机网络,E-mail: wangjunyister@gmail.com;叶新铭(1943—),男,教授,博士生导师,CCF 会员,主要研究方向为计算机网络和形式化理论。

文献[11]提出了语言模型下的几种相关反馈方法,根据用户标出的相关文档建立问句扩展语言模型,再用线性插入法与原问句合成新的问句模型;文献[4]提出了语言模型下的伪相关反馈方法;问句扩展可依据多方面的因素进行,如词的关系、用户兴趣、问句集语义等^[5,7-9]。

用户模型可以是用户使用过的问句、浏览过的页面、文档、邮件等历史记录^[13],常常用于对新的文档流过滤或对检索结果重新排名^[14]。文献[9]综合了多种方法扩展问句,其中之一是将用户模型按主题域分为不同的类,用最接近问句的类来扩展问句,将个性化成分加入到问句中,从而获得使用户更满意的检索结果。

上述方法有的是基于向量模型的,有的是基于语言模型的。多数使用用户模型做个性化检索或过滤的方法都是依据单一的主题域,虽然能够消除某些词的歧义性,但是会丢失一些用户感兴趣的东西。有很多情况下用户希望检索结果覆盖多个兴趣域。例如查询“游行”,希望查到诸如庆祝游行、示威游行和娱乐游行等,涉及到多个主题域。

3 用用户模型扩展问句

用户模型包含了用户感兴趣的内容,用其中的高频词扩展问句能更好地表示用户的信息需求。

文档与问句的模型都是一元语言模型,基本思想如图1所示。用问句 q 在用户模型上检索,取前 k 个文档作为反馈集,通过问句模型生成算法,依据原问句和反馈集生成扩展问句模型 Q ,再用 Q 在文档集上检索,得到检索结果。这里假设用户提出新问句多数属于其过去感兴趣的主题域,只有少数不属于其过去感兴趣的主题域,这也符合实际应用情况。

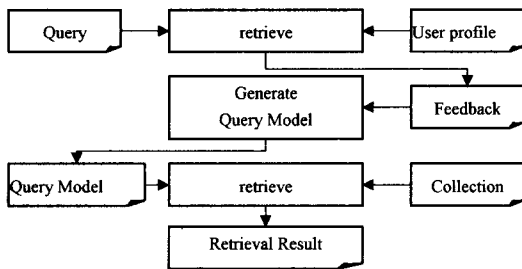


图1 问句扩展与检索过程

问句模型是由初始问句 q 的模型和反馈集模型线性结合而成:

$$p(w|Q) = (1-\alpha)p(w|Q_q) + \alpha p(w|Q_F) \quad (1)$$

式中, Q_q 是问句 q 的模型,用极大似然法估算。 Q_F 是反馈集的模型,用 Divergence minimization 方法建模^[4]可以降低对反馈集和非反馈集区分度不敏感的那些词的概率值,从而使得扩展问句的词能更准确地表达用户兴趣。

检索文档的排名函数使用 KL-divergence,其定义为:

$$Score(q, d) = \sum_{w \in q} p(w|Q) \log \frac{p(w|D)}{p(w|Q)} \propto \sum_{w \in q} p(w|Q) \log p(w|D) \quad (2)$$

式中, q 是问句, d 是文档, Q 和 D 分别是问句与文档的一元语言模型。 $p(w|D)$ 用极大似然法估算并做平滑处理。

我们的方法与传统方法和第2节中所述的方法有几点不同:

(1)相关反馈或伪相关反馈的初次检索和最终检索都是

在同一个文档集上进行的,我们的方法的初次检索是在用户模型上进行的,以获得用户所感兴趣的内容,检索范围小,效率高,用户感兴趣的文档更集中。

(2)相关反馈考虑了用户兴趣,但在检索中要和用户交互。伪相关反馈虽然不需要和用户交互,但没有考虑用户的个人兴趣。我们的方法既考虑了用户的兴趣,又不需要在检索过程中与用户交互。

(3)用户模型用于过滤时一般只包含一个主题域,检索时问句扩展只依据一个主题域。我们的方法中,用户模型包含了多个主题域,问句扩展可以依据多个主题域进行。

伪反馈方法是一种盲目反馈,对检索性能有较大的提高,但与用户个人兴趣毫无关系。多主题域用户模型扩展问句考虑了用户兴趣,将两者结合起来能取得更好的效果。扩展后的新问句模型由原问句模型、反馈集模型和盲目反馈集模型3部分线性组合而成:

$$p(w|Q) = \lambda p(w|Q_q) + \beta p(w|Q_F) + \gamma p(w|Q_P) \quad (3)$$

式中, $\lambda + \beta + \gamma = 1$, Q_P 是盲目反馈集的模型。

通过用户模型可以直接或间接地实现问句扩展的推理机制:

(1)发现主题域中和用户感兴趣内容中的高频共现词,例如用户对体育感兴趣,用户模型中会包含大量的“体育”、“足球”、“篮球”、“拳击”等词,查询“体育比赛”,从用户模型中得到的反馈集会包含大量这样的词,它们就有可能被扩展到问句中。

(2)推导出用户兴趣。例如,用户对计算机科学与技术感兴趣,“系统”一词具有很大的歧义性,如排水系统、银行系统、神经系统等,在用户模型中与“系统”关联的词是“计算机”、“软件”、“硬件”等词,所以当问句中含有“系统”一词时,它们被扩展到问句中的可能性就很大。

4 实验

实验中,使用 TREC 的 disk2 中与问句 51~100 相关的文档模拟用户模型,用 disk1 和问句 101~150 作为测试集。为了保证测试的公正性,用户模型的文档没有放入测试集,用于构造用户模型的问句在测试中也不使用。测试基线是一般的一元语言模型(记为 LM)。实验分为两个部分:第一部分是只针对以多主题域用户模型扩展问句的方法(记为 UF)进行实验,第二部分是 UF 和伪反馈(记为 PF)相结合的方法进行实验。问句 51~100 分属于 10 个主题域;问句 101~150 分属于 9 个主题域,有 8 个是重叠的。就主题域而言,这两组问句之间是有关联的。问句 51~150 在 13 个域中的分布如表 1 所列。从表中可看出,用户模型中没有 U. S. Politics 和 Medical and biological 这两个域的内容,测试用的 50 个问句中有 11 个是属于这两个域的,符合第 3 节开始的假设。

问句只抽取了 TREC 中 topics 的 title,文档和问句都使用了 Porter 词还原和去除了停用词,文档使用 Dirichlet prior 平滑,平滑系数是 1000。在第一部分实验中,式(1)中的 α 取 0.6。在第 2 部分实验中,式(3)的 λ , β 和 γ 分别取 0.2,0.3 和 0.5,这些参数的取值是经过优化的。对每个问句取它在用户模型中检索到的前 10 个文档为反馈集,一个问句最多扩展 20 个词,每个词的概率门限值为 0.001。对检索结果的评估指标是平均查准率 MAP、查全率 Recall 和前 20 个文档的查准率 $p@20$ 。实验使用 Lemmur 工具集。

表1 问句在主题域中的分布情况

No	Domain	51~100	101~150
1	International Economic	11	5
2	International Finance	1	0
3	U. S. Economic	2	4
4	U. S. Politics	0	3
5	Science and technology	11	5
6	Environment	4	0
7	International Relations	5	5
8	Law and government	4	6
9	Finance	2	2
10	international politics	2	9
11	Military	4	3
12	Political	2	0
13	Medical and biological	0	8

表2是多主题域用户模型扩展问句方法的实验结果。 $\alpha=0$ 是基线,没有扩展问句; $\alpha=1$ 是只包含反馈集的扩展模型,而不包含原问句模型。问句101~150在disk-1中共有7002个相关文档。 α 在0.1至0.9之间的结果均好于一般的语言模型检索结果, α 在0.8处查全率最高,在0.6处平均查准率最高,在0.4处,前20个文档的查准率最高。

表2 多主题域用户模型扩展问句实验结果

α	Recall/7002	MAP	p@20
0	4036	0.2133	0.3930
0.1	4083	0.2167	0.3970
0.2	4126	0.2202	0.3970
0.3	4153	0.2229	0.4050
0.4	4175	0.2245	0.4170
0.5	4198	0.2253	0.4050
0.6	4207	0.2261	0.4030
0.7	4223	0.2246	0.4070
0.8	4245	0.2210	0.4020
0.9	4225	0.2127	0.3850
1	3855	0.1771	0.3330

表3是 $\alpha=0.6$ 时UF与LM的比较,UF的各项指标均比LM有较好的改进。

表3 多主题域用户模型扩展问句模型与一般语言模型的比较

Model	Rec/7002	MAP	p@20
LM	4036	0.2133	0.3930
UF	4207	0.2261	0.4030
+	4.2	6.0	2.5

表4是UF与PF相结合的方法(记为UF+PF)的实验结果,此处 λ, β 和 γ 分别取0.2,0.3和0.5。与LM相比,各项指标均有较大的改进,与PF相比略有提高。

表4 UF与UF+PF的实验结果与比较

Model	Rec/7002	MAP	p@20
LM	4036	0.2133	0.3930
PF	4320	0.2388	0.4290
UF+PF	4364	0.2410	0.4390
+%to LM	8.1	13.0	11.7
+%to PF	1.0	0.9	2.3

结束语 个性化信息检索是复杂多样的,很难用一种通用的方法满足各种要求。本文提出的方法吸取了伪相关反馈方法和相关反馈方法的优点,考虑了用户希望检索结果涵盖

其多个感兴趣主题域的情况。通过实验证明了本方法的有效性。实验表明UF的性能优于LM,PF的性能优于UF,UF+PF的性能优于PF。用户模型具有很重要的作用,它的质量会直接影响到检索性能。许多细微的东西在我们的方法中还没有考虑,从选取好的用户模型内容、反馈集和扩展词等几方面入手,提高个性化检索性能还有一定的空间。

参考文献

- [1] Ponte J, Croft W B. A language modeling approach to information retrieval[C]//Proceedings of ACM SIGIR. 1998;275-281
- [2] Rocchio J. Relevance feedback in information retrieval[Z]. the SMART retrieval system. 1971;313-323
- [3] Salton G, Buckley C. Improving retrieval performance by relevance feedback[J]. American Society for Information Science, 1990,41;288-297
- [4] Zhai C X, Lafferty J. Model-based Feedback in the Language Modeling Approach to Information Retrieval[C]//Proceedings of ACM CIKM. 2001;403-410
- [5] Cao Guihong, Nie Jianyun. Integrating Word Relationships into Language Models[C]//Proceedings of ACM SIGIR. 2005;298-305
- [6] Nie Jian-yun, Cao Gui-hong, Bai Jing. Inferential Language Models for Information Retrieval[J]. Transactions on Asian Language Information Processing (TALIP), 2006,5(4):296-322
- [7] Bai Jing, Song Dawei, Bruza Peter, et al. Query Expansion Using Term Relationships in Language Models for Information Retrieval[C]//Proceedings of ACM CIKM. 2005;688-695
- [8] Bai Jing, Nie Jian-yun, Bouchard H, et al. Using Query Context in Information Retrieval[C]//Proceedings of ACM SIGIR. 2007;15-22
- [9] Bai Jing, Nie Jian-yun. Adapting information retrieval to query contexts[J]. Information Processing and Management, 2008,44;1901-1922
- [10] Lafferty J, Zhai Cheng-xiang. Document Language Models, Query Models, and Risk Minimization for information Retrieval[C]//Proceedings of ACM SIGIR. 2001;111-119
- [11] Shen Xue-hua, Zhai Cheng-xiang. Active Feedback in Ad Hoc Information Retrieval[C]//Proceedings of ACM SIGIR. 2005;59-66
- [12] Buckley C, Salton G, Allan J, et al. Automatic query expansion using SMART[C]//Proceedings of ACM SIGIR. 1988;321-331
- [13] Dumais S, Cutrell E, Caliz J, et al. Stuff I've seen: A system for personal information retrieval and re-use[C]//Proceedings of ACM SIGIR. 2003;72-79
- [14] Liu F, Yu C, Meng W. Personalized web search by mapping user queries to categories[C]//Proceedings of ACM CIKM. 2002;558-565
- [10] Hall P A V, Hitchcock P, Todd S J P. An Algebra of Relations for Machine Computation [C]//Conference Record of the Second ACM Symposium on Principles of Programming Languages. 1975
- [11] 王珊, 萨师煊. 数据库系统概论(第四版)[M]. 北京:高等教育出版社, 2006;52
- [12] 测绘科学数据共享服务网. 1比400万中国地图 [EB/OL]. http://sms.webmap.cn/, 2009-12-10

(上接第210页)

- [7] Jarke M, Koch J. Query Optimization in Database Systems [J]. ACM Computing Surveys, 1984,16;2
- [8] ESRI Inc. ArcGIS Server—企业级服务器 [EB/OL]. http://www.esrichina-bj.cn/templates/T_yestern_News/index.aspx?nodeid=155, 2009-12-10
- [9] 韩鹏, 王泉. 地理信息系统开发[M]. 武汉:武汉大学出版社, 2008;340