

基于矩阵分解的伪相关反馈技术*

周 栋 刘建勋 张三蓉

(湖南科技大学 计算机科学与工程学院 湘潭 411201)

摘 要 伪相关反馈技术的性能很大程度上依赖 2 个参数的取值,在缺乏结果相关性评价的前提下,这些参数只能依靠经验设置.文中提出基于矩阵分解的伪相关反馈技术.该技术将多个伪相关反馈结果使用协同过滤的思想融合,自动选择最优化参数进行查询扩展.实验表明,与现有的伪相关反馈技术相比,无论使用哪种信息检索模型,文中方法的检索性能都能得到较好改善.

关键词 伪相关反馈, 矩阵分解, 查询扩展, 协同过滤

中图法分类号 TP 391

DOI 10.16451/j.cnki.issn1003-6059.201508009

Pseudo-Relevance Feedback Technique Based on Matrix Factorization

ZHOU Dong, LIU Jian-Xun, ZHANG San-Rong

(School of Computer Science and Engineering, Hunan University of Science and Technology,
Xiangtan 411201)

ABSTRACT

The performance of pseudo-relevance feedback technique is heavily dependent on two parameter values. Under the lack of relevance valuation results, these parameters can only rely on experience to set. In this paper, a pseudo-relevance feedback technique based on matrix factorization is proposed. This technique fuses multiple pseudo-relevance feedback results using the ideas of collaborative filtering together. And the optimal parameters are automatically selected for query expansion. Experimental results show that compared with the existing pseudo-relevance feedback techniques, the proposed method has a better retrieval performance, regardless of any underlying information retrieval model.

Key Words Pseudo-Relevance Feedback, Matrix Factorization, Query Expansion, Collaborative Filtering

* 国家自然科学基金项目(No. 61402167, 61300129, 61272063)、教育部留学回国人员科研启动基金项目(No. 教外司留[2013] 1792)、湖南科技大学博士启动项目(No. E51368)资助

收稿日期:2014-09-03;修回日期:2015-01-11

作者简介 周栋(通讯作者),男,1979年生,博士,副教授,主要研究方向为信息检索、自然语言处理. E-mail: donghou1979@hotmail.com. 刘建勋,男,1970年生,博士,教授,主要研究方向为服务计算与云计算、工作流. 张三蓉,女,1979年生,硕士,助教,主要研究方向为信息系统、经济学.

1 引言

在信息检索领域,特别是在搜索引擎等实际应用中,当查询词和相关文档中存在字面不同的词项时,词汇不匹配 (Vocabulary Mismatching) 问题就会出现. 而这些不同词项在语义上可能相近,仅依靠字面匹配可能会带来检索性能的下降. 伪相关反馈是一种常用的查询扩展方法^[1-3],可用来部分解决词汇不匹配问题. 它假设初始检索结果的前若干篇文档与查询词相关,然后利用一些技术从这些文档中抽取一些与查询词相关的词项,从而扩展用户提交的查询词^[4]. 理论和实际信息检索评测 (如 TREC (<http://trec.nist.gov/>)、CLEF (<http://www.clef-initiative.eu/>) 等) 都已证明这是一种简单有效的技术.

早期查询扩展方法是使用手动方法进行^[5]. 搜索引擎使用者被要求从排名前列的文档中提取扩展词项生成新的查询词. 由于需要用户手动干预和用户必须了解一定的信息检索系统知识,该方法并不受欢迎,并很快被自动查询扩展所代替. 在自动查询扩展中,附加词项 (通常从机器可读叙词表或文档语料库中提取) 添加到查询词这一过程无需用户干预^[6]. 伪相关反馈就是一种典型的自动查询扩展技术. 它从初始搜索结果中排名靠前的文档内提取用于扩展的词项^[7-9]. 研究者们也研究多种改进伪相关反馈性能的方法,但最佳参数设置仍是一个主要问题.

近期研究者在这方面的的工作主要集中于开发选择查询扩展 (Selective Query Expansion) 工具^[10]. 这些工具有着类似想法,即预估查询扩展的效果,在可能表现不佳的情况下禁用查询扩展. 其中最著名的预测函数被称为清晰度得分 (Clarity Score)^[11],通过计算查询词模型和文档模型之间的 Kullback-Leibler 距离得到. 然而,该预测方法的实际应用尚不明确.

相关工作还包括使用基于日志的方法进行选择查询扩展,如利用外部证据的方法,利用网络数据的方法等^[12]. 查询扩展仍是一个极其活跃的研究方向,He 等^[13] 甚至专门针对这一领域建立一个测试集.

在实际实施伪相关反馈技术时,检索性能很大程度上依赖 2 个重要参数的取值,包括前 α 个文档和从文档中提取的前 β 个关键词. 在缺乏结果相关性评价的前提下,这 2 个参数很难设置. 通常做法是借助标准化测试集进行参数调节,然后固定使用某

一数值在不同的数据集上进行查询扩展. 在很多情况下,这些参数只能依靠经验设置. 这一方法有 2 个问题:1) 在不同的数据集上使用固定的参数值往往会带来检索性能的下降^[14];2) 虽然在标准测试集上调试参数是可行的,但在实际应用 (如搜索引擎、数字图书馆等) 中往往没有评价集,不能进行参数调试.

协同过滤是一种常用的商业化推荐系统技术^[15]. 该技术对用户喜好某一物品做出可能性预测. 通常协同过滤算法的先决条件是需要一个物品评分矩阵. 这些评分由某一用户及所有其他用户给出. 假设该矩阵已构建好,是否向某一用户推荐某一物品的过程如下: 首先找到与当前用户评分相同的所有用户的集合,然后过滤该集合找到已评价某件物品的用户,最后计算该物品的平均用户评分,如果该平均评分为正数,向用户推荐该物品.

随着评分矩阵的扩大,计算成本迅速增高. 有许多研究致力于优化上述过程,提出一系列的基于内存和基于模型的算法^[16-17]. 但这 2 种算法在计算稀疏矩阵时都会遇到困难. 针对该问题,国内外提出一系列基于矩阵分解模型的协同过滤算法. 同时,如何准确推荐物品给新用户 (即“冷启动”或“Cold Start”问题) 也是推荐系统的另一个具有挑战性的问题^[18]. 目前尚缺乏使用协同过滤思想进行伪相关反馈的相关研究.

基于上述分析,本文提出基于矩阵分解的伪相关反馈技术 (Matrix Factorization Based Pseudo-Relevance Feedback, MPRF). 该技术将多个伪相关反馈结果使用协同过滤的思想相融合,自动选择最优参数进行查询扩展. 这样能避免在评价集缺乏的情况下不能进行参数调试的缺陷. 实验表明,与传统伪相关反馈技术相比,无论使用哪种信息检索模型,该方法的检索性能都能得到较大提升.

2 基于矩阵分解的伪相关反馈技术

本节解释基于矩阵分解的伪相关反馈技术 (MPRF) 如何工作. 假设信息检索系统接收某一查询词 q_a . 首先,使用某一信息检索排序模型得到一组结果集. 再使用参数 α (伪相关文档的数目) 和 β (扩展词项的数目) 的一系列取值结果调用常规的伪相关反馈算法. 该步骤生成一系列与 q_a 相关的扩展查询词集合 $Q = \{q_1, q_2, \dots, q_m\}$, $m = \alpha\beta$. 随后用清晰度

得分筛选这组扩展查询词,保留得分较低的查询词(即 Q_c),按照惯例,这些查询词很可能会产生更多元化的检索结果(见算法 1 第 3 行)。

随后,系统检索得到 Q_c 中每个查询词的靠前文档(见算法 1 中 4 ~ 7 行),与关于 q_a 的靠前文档一起生成用于协同过滤系统的评分矩阵 V (见算法 1 中 10 ~ 15 行). 在算法 1 中, x 表示扩展查询词检索得到的靠前文档数目, y 表示用于测试查询词检索得到的文档数目, $RATE()$ 函数返回信息检索排名分数. 本文将在第 3 节更详细讨论 x 和 y .

评分矩阵 V 建立好后,借用协同过滤的方法,可为每个文档 $d \in D$ 预测其相对于 q_a 的相关性评分 p_{ad} . 在算法 1 中, /* Generate predictions for each document */ 描述这一过程,其中 17 ~ 20 行计算除 q_a 检索到的文档外所有包含在矩阵中文档的协同过滤分数(即 CF-SCORE). 首先采用基于矩阵分解的协同过滤方法(见算法 1 第 18 行),用一个低秩矩阵 V' 逼近评分矩阵 V ,该操作对 D 中未包括 q_a 靠前文档的其他文档(即 $D \setminus \{d_1, \dots, d_y\}$) 也给予评分。

该步骤对于最终排序至关重要,因此有必要选择合适的协同过滤方法. 近年来,基于矩阵分解的协同过滤算法在推荐系统领域得到广泛应用^[19-20]. 与传统的协同过滤算法相比,基于矩阵分解的方法可较好解决稀疏矩阵和冷启动问题. 这里采用正则奇异值分解(Regularized Singular Value Decomposition)方法进行计算. 该方法希望找到一个低秩矩阵 V' 逼近一个维度为 $m \times n$ 的矩阵 V ,该 V 即为算法中构建的评分矩阵,其中

$$V' = UR^T, U \in \mathbf{R}^{m \times f}, R \in \mathbf{R}^{n \times f},$$

即分解为两个维度分别为 $m \times f$ 和 $n \times f$ 的矩阵. f 表示特征的个数,一般来说 $f \ll e$ (e 表示矩阵 V 的秩), $e \leq \min(m, n)$. 为找到这样一个低秩矩阵以最大程度逼近矩阵 V ,最小化如下损失函数:

$$L = \min_{U, R} \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^n I_{ij} (v_{ij} - \mathbf{u}_i^T \mathbf{r}_j)^2 + \frac{\lambda_1}{2} \|U\|_F^2 + \frac{\lambda_2}{2} \|R\|_F^2.$$

在该目标函数中, I_{ij} 为指示函数,当查询词 q_a 的搜索结果中包含 d_j 时其值为 1,否则其值为 0, λ_1 和 λ_2 均为正则参数,公式中矩阵用列向量表示. 使用简单的随机梯度下降方法,通过已得到的 v_{ij} 检索评分值,可得到如下更新规则:

$$\begin{aligned} \mathbf{u}_i &\leftarrow \mathbf{u}_i + \gamma_1 (\Delta_{ij} \mathbf{r}_j - \lambda_1 \mathbf{u}_i), \\ \mathbf{r}_i &\leftarrow \mathbf{r}_i + \gamma_2 (\Delta_{ij} \mathbf{u}_i - \lambda_2 \mathbf{r}_i), \end{aligned}$$

其中, $\Delta_{ij} = v_{ij} - \mathbf{u}_i^T \mathbf{r}_j$, γ_1 和 γ_2 为学习率.

使用基于矩阵分解的协同过滤算法后,得到推荐的文档排序列表. 最后阶段通过依据信息检索评分(IR-SCORE)和协同过滤评分(CF-SCORE)及两者线性结合的评分(COMB-SCORE)对检索文档进行重新插入及排序,产生最终的结果文档列表(算法 1 中 21 ~ 32 行). c 的取值通常大于 y 的取值.

结合 IR-SCORE 和 CF-SCORE 评分时,本文实验一系列的线性结合方法,如取最大值(CombMAX),取和(CombSUM),取百分比(规范化,CombRSV%,其中 RSV 指 Result Value,即结果值),集合检索-推理(Collection Retrieval Inference, CORI),Z 值(Z-SCORE)等^[21]. CombRSV% 效果最佳,通常计算如下(算法 1,第 25 行):

$$SUM(RSV_i / MAX_{RSV}),$$

其中 RSV 表示用于计算的 IR-SCORE 或 CF-SCORE. 一旦计算该结合评分,算法根据三类评分的大小进行排序,产生最后的文档列表.

下面给出基于矩阵分解的伪相关反馈(MPRF)的具体步骤.

算法 1 基于矩阵分解的伪相关反馈

Require: q_a A TEST QUERY

Require: C A DOCUMENT CORPUS

1 /* Populate the rating matrix V */

2 $Q \leftarrow QUERY-EXPANSION(q_a)$

3 $Q_c \leftarrow QUERY-CLARITY(Q)$

4 **for** $q_i \in Q_c$ **do**

5 $(d_1, d_2, \dots, d_x) \leftarrow IR-RETRIEVE(q_i, C)$

6 $D \leftarrow D \cup (d_1, d_2, \dots, d_x)$

7 **end for**

8 $(d_1, d_2, \dots, d_y) \leftarrow IR-RETRIEVE(q_a, C)$

9 $D \leftarrow D \cup (d_1, d_2, \dots, d_y)$

10 **for all** $q_i \in Q_c \cup q_a$ **do**

11 **for** $d_j \in D$ **do**

12 $v_{ij} \leftarrow RATE(q_i, d_j)$

13 $V \leftarrow v_{ij}$

14 **end for**

15 **end for**

16 /* Generate predictions for each document */

17 **for all** $d_i \in D \setminus (d_1, d_2, \dots, d_y)$ **do**

18 $p_{ai} \leftarrow CF-CALCULATION(q_a, d_i)$

19 $p_a \leftarrow p_{ai}$

20 **end for**

21 /* Integrate IR-generated and CF-generated

results to create the final ranking list */

```

22 ( $d_1, d_2, \dots, d_c$ ) ← IR-RETRIVER( $q_a, C$ )
23 for all  $d_i \in (d_1, d_2, \dots, d_c)$  do
24   if  $p_{ai}$  has corresponding score in  $p_a$ . then
25     COMB( $d_i$ ) ← COMB( $p_{ai}, IR-SCORE(d_i)$ )
26   end if
27 end for
28 for all  $d_i$  that  $p_{ad} > 0$  do
29    $D_{final} \leftarrow (d_1, d_2, \dots, d_c) \cup d_j$ 
30 end for
31 sort  $D_{final}$  by IR-SCORE, CF-SCORE,
   COMB-SCORE

```

Output: D_{final}

算法1的时间复杂度分析如下。设 n_f 表示特征的个数, n_q 表示查询词的数目, n_d 表示所有查询词靠前文档的数目。生成评分矩阵 V 的时间复杂度为 $O(n_q n_d)$, 进行基于矩阵分解的协同过滤算法时间复杂度为 $O(n_f^2 + n_f) n_i$, 其中 n_i 为迭代次数。而结合评分生成最后列表的时间复杂度为 $O(n_d)$ 。从上述分析来看, 该算法完全可行。

3 动态参数调节

算法1中有2个参数需手动设定: 扩展查询词所获得的前 x 个靠前文档数目, 测试查询词所获得的前 y 个靠前文档数目。现简要介绍一种动态调节方法, 自动设置这2个参数。

该动态调节方法通过计算查询词的效用 (Utility) 实现调节^[22]。这里效用是指在分离一组正、负的样例时获得的信息增益 (Information Gain)。给定某一查询词 q 和一组正、负样例 S, q 的信息增益通过划分 S 至子集 S_i 时信息熵 E 的变化进行计算, 其中子集 S_i 的划分根据信息检索评分确定, 定义如下:

$$IG(q, S) = E(S) - \sum_i E(S_i) \frac{|S_i|}{|S|},$$

其中, S 为查询词所获得的检索文档列表, 取前 i 个文档为正样例, 其余文档为负样例, i 的取值范围为1到结果文档列表的数目 d' (按照信息检索惯例, 这一数目通常设定为1000)。该方法循环测试 i 所有可能的取值结果, 当最佳信息增益出现时, 记录 i 的取值, 并把它作为 x 和 y 的取值。为简化操作, 设 $x = y$, 从而达到动态调节参数的目的。由于只需循环 i 次, 动态参数调节部分时间复杂度仅为 $O(nd')$ 。

4 实验及结果分析

本文实验所处理的语料是 CLEF 会议 (<http://www.clef-initiative.eu/>) 中 ad-hoc 任务提供的 CLEF-2006 测试集。其内容为 Glasgow Herald 1995 和 L. A. Times 1994 英文新闻数据集, 包含 169 477 份新闻文档。在查询词集方面, 亦使用 CLEF-2006 查询集, 包含 50 个查询主题。查询词由标题 (Title) Description 字段中内容构成, 因为该字段仅包含很少的查询词项, 与网络搜索的实际情况相符。在建立索引和搜索之前, 单词都进行停用词去除和词项归一化操作。所有相关性判定由 CLEF 会议提供。

实验采用的评价标准分为如下2类。

1) 平均准确率均值 (Mean Average Precision, MAP): 针对所有的查询主题求取平均准确率的均值, 平均准确率是指针对每个查询主题, 返回结果中每篇相关文档位置上的准确率的平均值。

2) 归一化折损累积增益 (Normalized Discounted Cumulative Gain, NDCG): 该评价方法不仅考虑二值评分 (即相关或不相关), 而且考虑针对查询结果的等级评分。

两种评价方法具体计算过程可参考文献[4]。给出的结果表明所有查询主题的平均表现。显著差异由配对样本 t 检验测定。

所有搜索实验由开源软件 Terrier (<http://terrier.org>) 完成。首先测试的信息检索模型为经典的 BM25 模型及 Terrier 提供的基于 BM25 的伪相关反馈方法 (缩写为 BM25-PRF)。 α 和 β 的取值范围分别设定为 [1, 10] 和 [1, 20]。在实验中亦采用语言模型作为信息检索模型, 以说明 MPRF 的稳定性。这里用 BM25-PRF- α - β 表示某一指定 α, β 取值下的一组伪相关反馈结果。

x 和 y 的取值使用第3节中的动态调节方法确定。实验结果表明, x 和 y 针对每个查询词的取值不同, 但一般不会大于200。这一结论与伪相关反馈相关研究结论相符, 排名靠后的文档相关度相对较低, 如果不加区分地与靠前文档一起使用, 会产生较大噪音, 降低搜索的准确率。

实验首先对比 MPRF 与传统的伪相关反馈方法的性能。实验结果见表1, 其中对比 MPRF 与最好的 BM25-PRF- α - β 结果, 表中与 BM25 有显著差异的实验结果用“*”号标识。

由表1可得出如下结论: 1) 在2种评价标准下, MPRF 的准确率要高于所有200组 BM25-PRF- α - β 的实验结果 (10×20, 且具有显著差异)。2) 与基线系

统 BM25 相比(无相关性反馈),在性能上 MPRF 实现统计学上显著意义的准确率提高,而在 BM25-PRF- α - β 的所有 200 组实验结果中,仅有 10% 的结果与基线系统 BM25 相比有显著提高。3) 从实验结果可发现,仅采用固定参数的 BM25-PRF- α - β 获得的实验结果非常不稳定,有一部分实验结果在准确率方面甚至低于 BM25 基线系统。

表 1 CLEF-2006 数据集检索结果

Table 1 Retrieval results on CLEF-2006 dataset

方法	MAP	方法	NDCG
MPRF	0.6359*	MPRF	0.3840*
BM25-PRF-2-11	0.6271*	BM25-PRF-6-20	0.3767*
BM25-PRF-2-9	0.6262*	BM25-PRF-6-16	0.3760*
BM25-PRF-2-10	0.6261*	BM25-PRF-6-17	0.3757*
BM25-PRF-6-16	0.6252*	BM25-PRF-6-19	0.3748*
BM25-PRF-2-8	0.6245*	BM25-PRF-6-18	0.3740*
BM25-PRF-6-17	0.6245*	BM25-PRF-6-15	0.3726*
BM25-PRF-6-20	0.6245*	BM25-PRF-2-11	0.3721*
BM25-PRF-6-19	0.6243*	BM25-PRF-2-9	0.3717*
BM25-PRF-2-7	0.6238*	BM25-PRF-6-14	0.3717*
BM25-PRF-6-18	0.6238*	BM25-PRF-2-20	0.3715*
BM25	0.5959	BM25	0.3392

为进一步说明 α 和 β 取值的合理性,验证固定参数的伪相关反馈方法的缺陷,图 1 展示部分 BM25-PRF- α - β 测试准确率方面的实验结果。

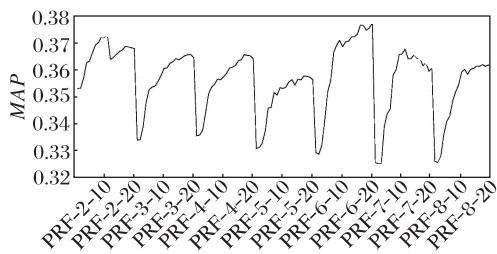


图 1 BM25-PRF- α - β 在 CLEF-2006 数据集上的部分实验结果

Fig. 1 Part of the experimental results of BM25-PRF- α - β on CLEF-2006 dataset

由图 1 可看出,传统的伪相关反馈方法,即 BM25-PRF 严重依赖 α 和 β 的取值。首先,当 α 和 β 的取值过大时,性能出现明显下降,如果不对其加以限制,准确率会随着取值的增大逐渐降低。其次,即使在存在训练集的情况下,如何准确取值也是一个大问题。因为从图中可看出曲线并非线性变化,亦不存在持续递增或递减的情况。更何况在实际网络搜索等环境中,训练集很难获得,给高效使用这一技术

带来很大困难。

为说明 MPRF 的稳定性,将信息检索模型替换为语言模型(简称为 LM)后,重复上述实验,得到一组实验结果如表 2 所示。MPRF 的准确率仍高于所有 200 组 LM-PRF- α - β 的实验结果及语言模型基线系统,证明该方法的稳定性。

表 2 语言模型在 CLEF-2006 数据集上检索结果

Table 2 Retrieval results of language models on CLEF-2006 dataset

方法	MAP	方法	NDCG
MPRF	0.3298*	MPRF	0.3800*
LM-PRF-9-1	0.3237*	LM-PRF-6-16	0.3746*
LM-PRF-9-5	0.3217*	LM-PRF-6-20	0.3745*
LM-PRF-10-6	0.3211*	LM-PRF-6-19	0.3744*
LM-PRF-9-2	0.3210*	LM-PRF-2-12	0.3735*
LM-PRF-8-1	0.3206*	LM-PRF-6-17	0.3734*
LM-PRF-9-3	0.3204*	LM-PRF-6-15	0.3731*
LM-PRF-10-4	0.3203*	LM-PRF-6-18	0.3728*
LM-PRF-5-16	0.3187*	LM-PRF-2-11	0.3724*
LM-PRF-9-6	0.3180*	LM-PRF-6-14	0.3718*
LM-PRF-5-15	0.3178*	LM-PRF-2-9	0.3713*
LM	0.3003	LM	0.3461

5 结束语

在传统的伪相关反馈技术的基础上,本文提出一种基于矩阵分解的伪相关反馈技术(MPRF)。该技术将多个伪相关反馈结果使用协同过滤的思想相融合。本文还介绍一种自动调节参数的方法以方便进行查询扩展。在 CLEF-2006 数据集上的实验表明,MPRF 明显优于传统的伪相关反馈技术,且该技术在不同的检索模型下取得稳定性能。在以后的工作中还将考虑其他协同过滤的算法,也可通过潜在语义分析方法过滤相似的扩展查询词,并用于伪相关反馈技术。

参 考 文 献

- [1] Bhatnagar P, Pareek N. Improving Pseudo-Relevance Feedback Based Query Expansion Using Genetic Fuzzy Approach and Semantic Similarity Notion. *Journal of Information Science*, 2014, 40(4): 523-537
- [2] Wu D, He D Q, Wang H L. A Relevance Feedback Based Query Translation Enhancement Technique in Cross Language Information Retrieval. *Journal of the China Society for Scientific and Technical Information*, 2012, 31(4): 398-406 (in Chinese)
(吴丹,何大庆,王惠临.一种基于相关反馈的跨语言信息检索

- 查询翻译优化技术研究. 情报学报, 2012, 31(4): 398-406)
- [3] Huang M X, Yan X W, Zhang S C. Query Expansion of Pseudo Relevance Feedback Based on Matrix-Weighted Association Rules Mining. *Journal of Software*, 2009, 20(7): 1854-1865 (in Chinese)
(黄名选, 严小卫, 张师超. 基于矩阵加权关联规则挖掘的伪相关反馈查询扩展. 软件学报, 2009, 20(7): 1854-1865)
- [4] Manning C D, Raghavan P, Schütze H. *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press, 2008
- [5] Harter S P. *Online Information Retrieval: Concepts, Principles, and Techniques*. Salt Lake City, USA: Academic Press, 1986
- [6] Bhogal J, Macfarlane A, Smith P. A Review of Ontology Based Query Expansion. *Information Processing & Management*, 2007, 43(4): 866-886
- [7] Lee C J, Croft W B. Cross-Language Pseudo-Relevance Feedback Techniques for Informal Text // Proc of the 36th European Conference on Information Retrieval. Amsterdam, The Netherlands, 2014: 260-272
- [8] Jiang L, Mitamura T, Yu S I, *et al.* Zero-Example Event Search Using Multi-modal Pseudo Relevance Feedback // Proc of the International Conference on Multimedia Retrieval. Glasgow, UK, 2014: 297-304
- [9] Wang X W, Wang X J, Sun Y P. Cross-Lingual Pseudo Relevance Feedback Based on Bilingual Topics. *Journal of Beijing University of Posts and Telecommunications*, 2013, 36(4): 81-84 (in Chinese)
(王序文, 王小捷, 孙月萍. 双语主题跨语言伪相关反馈. 北京邮电大学学报, 2013, 36(4): 81-84)
- [10] Cronen-Townsend S, Zhou Y, Croft W B. A Framework for Selective Query Expansion // Proc of the 13th ACM International Conference on Information and Knowledge Management. Washington, USA, 2004: 236-237
- [11] Cronen-Townsend S, Zhou Y, Croft W B. Predicting Query Performance // Proc of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Tampere, Finland, 2002: 299-306
- [12] Zhou D, Lawless S, Wade V. Improving Search via Personalized Query Expansion Using Social Media. *Information Retrieval*, 2012, 15(3/4): 218-242
- [13] He B, Ounis I. Studying Query Expansion Effectiveness // Proc of the 31st European Conference on Information Retrieval Research. Toulouse, France, 2009: 611-619
- [14] Billerbeck B, Zobel J. Document Expansion versus Query Expansion for Ad-hoc Retrieval // Proc of the 10th Australasian Document Computing Symposium. Sydney, Australia, 2005: 34-41
- [15] Zhang F Z, Liu S, Li Z H, *et al.* Collaborative Filtering Recommendation Algorithm Incorporating User's Reviews and Contextual Information. *Journal of Chinese Computer Systems*, 2014, 35(2): 228-232 (in Chinese)
(张付志, 刘赛, 李忠华, 等. 融合用户评论和环境信息的协同过滤推荐算法. 小型微型计算机系统, 2014, 35(2): 228-232)
- [16] Sarwar B, Karypis G, Konstan J, *et al.* Item-Based Collaborative Filtering Recommendation Algorithms // Proc of the 10th International Conference on World Wide Web. Hong Kong, China, 2001: 285-295
- [17] Cacheda F, Carneiro V, Fernández D L, *et al.* Comparison of Collaborative Filtering Algorithms: Limitations of Current Techniques and Proposals for Scalable, High-Performance Recommender Systems. *ACM Trans on the Web*, 2011, 5(1): 1-33
- [18] Sun D T, He T, Zhang F H. Survey of Cold-Start Problem in Collaborative Filtering Recommender System. *Computer and Modernization*, 2012, (5): 59-63 (in Chinese)
(孙冬婷, 何涛, 张福海. 推荐系统中的冷启动问题研究综述. 计算机与现代化, 2012, (5): 59-63)
- [19] Yang Y, Xiang Y, Xiong L. Collaborative Filtering and Recommendation Algorithm Based on Matrix Factorization and User Nearest Neighbor Model. *Journal of Computer Applications*, 2012, 32(2): 395-398 (in Chinese)
(杨阳, 向阳, 熊磊. 基于矩阵分解与用户近邻模型的协同过滤推荐算法. 计算机应用, 2012, 32(2): 395-398)
- [20] Ma H. An Experimental Study on Implicit Social Recommendation // Proc of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 2013: 73-82
- [21] Savoy J. Comparative Study of Monolingual and Multilingual Search Models for Use with Asian Languages. *ACM Trans on Asian Language Information Processing*, 2005, 4(2): 163-189
- [22] Egozi O, Markovitch S, Gabrilovich E. Concept-Based Information Retrieval Using Explicit Semantic Analysis. *ACM Trans on Information Systems*, 2011, 29(2): 1-38