基于迭代方法的多层 Markov 网络信息检索模型*

洪欢，王明文，万剑怡，廖亚男

（江西师范大学 计算机信息工程学院，南昌 330022）

摘要：查询扩展是提高检索效果的有效方法，传统的查询扩展方法大都以单个查询词的相关性来扩展查询词，没有充分考虑词项之间、文档之间以及查询之间的相关性，使得扩展效果不佳。针对此问题，本文首先通过分别构造词项子空间和文档子空间的 Markov 网络，用于提取出最大词团和最大文档团，然后根据词团与文档团的映射关系将词团分为文档依赖和非文档依赖词团，并构建基于文档团依赖的 Markov 网络检索模型做初次检索，从返回的检索结果集合中构造出查询子空间的 Markov 网络，用于提取出最大查询团，最后，采用迭代的方法计算文档与查询的相关概率，并构建出最终的基于迭代方法的多层 Markov 网络信息检索模型。实验结果表明：本文的模型能较好地提高检索效果。

关键词：Markov 网络；查询扩展；文档依赖；团；信息检索

中图分类号：TP391　　**文献标识码：**A

A Multi-layer Markov Network Information Retrieval Model Based on Iteration

Hong Huan, Wang Mingwen, Wan Jianyi, Liao Yanan

(School of Computer Information Engineering, Jiangxi Normal University, Nanchang 330022, China)

Abstract: Query expansion is an effective way to improve the retrieval effectiveness, traditional query expansion methods mostly extend the query words only considered the relevance of a single query word, without fully considering the relevance between terms, documents, as well as between queries, so this makes the expansion effect poorly. To solve this problem, first, we construct the Markov network of terms' and documents' subspace for extracting the maximum term cliques and document cliques, then, we divide the maximum word cliques into documents dependent word cliques and non-documents dependent word cliques through the mapping relation between term and document cliques, and build the Markov network retrieval model based on document cliques dependency to do the initial search, then we construct the Markov network of queries' subspace from the search results, which are used for extracting the maximum query cliques, finally, we calculate the probability between document and query in an iterative method, and build the final multi-layer Markov network information retrieval model based on iteration. Experimental results show that our model can improve the retrieval results.

Key words: Markov network; query expansion; document reliance; clique; information retrieval

1. 引言

对于大部分用户来说，在没有充分了解文档集的情况下很难给出专为检索而设计的查询，搜索引擎用户经常需要重构他们的初始查询表式以获得他们所感兴趣的更好结果[1]。随着信息检索技术的不断发展，出现了许多通过使用和查询意图相关的信息来提升初始查询表式的方法。查询扩展就是一种使用和查询意图相关的附加信息来重构查询的反馈方法，它是

* **基金项目：**国家自然科学基金资助项目(61272212，61163006，61203313)

作者简介：洪欢（1991—），男，硕士研究生，主要研究方向为信息检索、数据挖掘；王明文（1964—），男，博士，教授，主要研究方向为信息检索、数据挖掘、机器学习；万剑怡（1974—），女，博士，教授，主要研究方向为并行分布式算法、大规模计算研究。

指利用计算机语言学、信息学等多种技术，在初始查询表式的基础上通过一定的方法和策略把与原查询词相关的词、词组添加到原查询中，组成新的、更能准确表达用户查询意图的查询词序列，然后用新查询对文档重新检索，从而改善信息检索中的查全率和查准率低下的问题，弥补用户查询信息不足的缺陷。

在最近的一些研究当中，Zhai 通过 Boosting 算法将相关模型[2][3]和混合模型[4]合并来提取查询扩展词，这种方法将两个弱模型合并成一个强模型[5][6]，但是没有考虑文档与词项之间的关联性。Lee 通过聚类的方式将初始检索出的文档聚成多个簇，然后基于文档簇利用相关模型选取查询扩展词[7]，虽然考虑了文档与词项之间的关系，但是在文档表示方面存在缺陷，他仅仅将文档簇看成一个大的文档，忽略了每个文档之间的相关性。上述模型都是基于词独立性的假设，但实际上词之间的关联信息对检索效果有很大的影响。甘丽新等提出一种基于 Markov 概念的信息检索模型[8]，对于每个查询，使用团的提取算法在词项子空间的 Markov 网络中提取扩展词，该模型取得了良好的检索效果，但是该工作没有考虑文档之间和查询之间的相关性信息。付剑波等提出基于团模型的文档重排算法研究[9]，该模型通过对文档集的学习，构造文档子空间的 Markov 网络，提取出文档团，使用文档团信息进行文档重排，但是该工作没有将文档团信息用于查询扩展中。汤皖宁等提出基于文档团依赖的 Markov 网络检索模型[10]，该模型首先使用团的提取算法提取出文档团和词团，然后根据词项子空间和文档子空间的映射关系将词团分为文档依赖词团和非文档依赖词团，并用于查询扩展中。虽然扩展中充分利用了索引词项、文档子空间的 Markov 网络信息，但是没有考虑到查询子空间的 Markov 网络信息。

本文将汤皖宁等提出的基于文档团依赖的 Markov 网络检索模型[10]作为基础模型，对文档集做初次检索；接着，将初次检索得到的结果用于提取查询子空间的 Markov 网络，并提取出最大查询团；最后，结合前两个步骤构建基于迭代方法的多层 Markov 网络信息检索模型，即结合词项、文档和查询子空间的 Markov 网络信息用于查询扩展以及采用迭代方法计算文档与查询之间的相关概率。文中首先介绍了 Markov 网络检索模型、Markov 网络的构造方法、最大团的提取等工作；接着，重点介绍了初始的基于文档团依赖的 Markov 网络检索模型以及基于迭代方法的多层 Markov 网络查询扩展模型。最后，对本文提出的模型与一些经典的模型作对比实验，并对实验结果进行分析以及提出下一步的工作展望。

2. Markov 网络检索模型构造方法

2.1 Markov 网络检索模型

Markov 网络是一种不确定性推理的有利图形工具[11]，它可以较好地表示知识关联，很容易从实例数据中训练获得，具有强大的学习功能和推导能力。通过它的无向边来解释信息检索中的语义关系更直接、恰当。Markov 网络可以表示为一个二元组(V,E)，V 为所有节点的集合，E 为一组无向边的集合，$E=\{(x_i, x_j)|\ x_i \neq x_j \wedge x_i, x_j \in V\}$，$E$ 中的边表示节点之间的依赖或相关关系。

基于 Markov 网络的检索模型分为三层:查询子空间，索引词项子空间，文档子空间。如图 1 所示，三层子空间构成了一个推理网络，根节点为词项子空间的词节点，我们利用词与词之间的相关性、文档与文档之间的相关性、查询与查询之间的相关性分别构造词项、文档、查询子空间的 Markov 网络。通过设定阈值，对三层子空间的 Markov 网络信息分别进行最大团的提取，得到最大词团、文档团、查询团。

在下图 1 中，q 代表用户给定的初始查询，q_i 代表查询子空间中的查询，t_j 代表索引词项子空间中的词项，d_k 代表文档子空间里的文档。图中从词项 t_j 存在指向查询 q_i 和文档 d_k 的有向边，分别代表查询 q_i 和文档 d_k 由这些词项 t_j 构成；在查询、词项和文档三层子空间中存在的多条无向边，分别表示查询间、词项间及文档间存在关联，并且边的权重取决于关

联的程度：

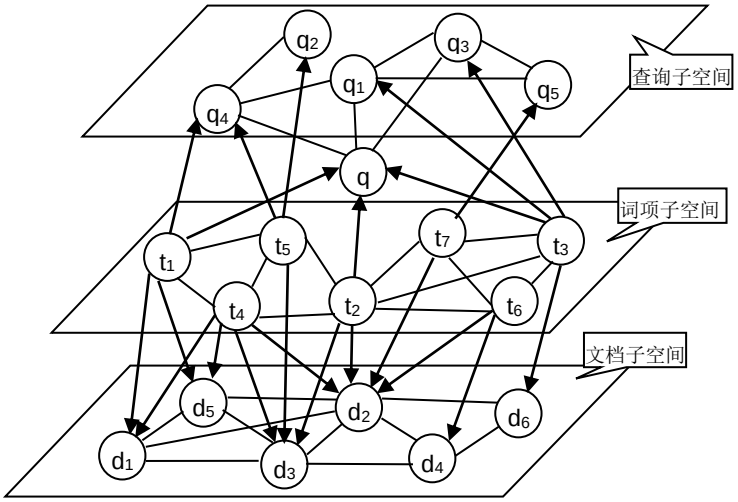

图 1 多层 Markov 网络结构图

2.1.1 词项子空间的 Markov 网络

在图 1 所示结构图的词项子空间中，节点表示词项，词项之间有边相连构成了词项子空间的 Markov 网络，边的权重表示词间的相关性程度。通过利用词的共现性来提取词与词之间的关系已经运用到许多研究中，在计算词共现的词频中，一般可以以整个文档、段落或是一个固定长度为窗口[12]。出于考虑效率方面的因素，本文选择文档作为窗口单位。实验中采用词的共现性来提取词项与词项之间的关系，鉴于 Markov 网络的无向性，因此在构造词项子空间的 Markov 网络时，采用两个词的综合共现性来计算，公式如下：

$$Sim(t_i,t_j)=\frac{P_{co}(t_i|t_j)+P_{co}(t_j|t_i)}{2} \quad (1)$$

$$P_{co}(t_i|t_j)=\frac{C(t_i,t_j)}{C(t_j)} \quad (2)$$

其中 t_i 和 t_j 指两个词项，$C(t_i,t_j)$指在训练文档集中 t_i 和 t_j 在同一篇文档中同时出现的频率，$C(t_i)$和 $C(t_j)$分别表示在训练文档集中 t_i 和 t_j 出现的频率，$Sim(t_i,t_j)$表示 t_i 和 t_j 之间的相关性，*Sim* 值越大，两个词的相关性就越高。当 *Sim* 值大于给定的阈值时，则词项 t_i 和 t_j 相互依赖，即在词项子空间的 Markov 网络中有边相连。

2.1.2 文档子空间的 Markov 网络

同样，在图 1 所示的文档子空间中，节点表示文档，文档之间有边相连构成文档子空间的 Markov 网络，边的权重表示文档间的相关性程度。本文中的文档均表示为向量，采用文档向量之间的夹角余弦公式来度量文档之间的相关性，文档与文档之间的相关性记为 $Sim(d_i,d_j)$ ，计算公式如下：

$$Sim(d_i,d_j)=\frac{\vec{d_i}\times\vec{d_j}}{\left|\vec{d_i}\right|\times\left|\vec{d_j}\right|} \quad (3)$$

当 $Sim(d_i,d_j)$的值大于给定的阈值时，则文档 d_i 和 d_j 相互依赖，即在文档子空间的 Markov 网络中有边相连。

2.2 团的提取

通过三层子空间的 Markov 网络结构分析可知，它实际构成一个相容关系图。在相容关系图中，我们发现许多完全多边形，就是每个节点都与其他节点相连的多边形，即构成了团。团的提取分为索引项词团的提取、文档团的提取和查询团的提取。

对于词项和文档子空间的 Markov 网络，团内的词和文档彼此相互依赖，即存在某种语义关联，可以认为它们集中表达同一个概念[13]（或主题）。本文按照词的最大团以及文档团与词团之间的映射关系来选择扩展查询词，以最大团为单位进行扩展，具体最大团的提取方法可以参照文献[8]。如果一个词团中的某个词项同时映射到一个文档团的多篇文章，则认为这个词比其他的词对主题更具有代表性，在后续的检索阶段中提高包含此类词项词团的权重，我们将这种词项称为文档依赖词。

在词项和文档子空间的 Markov 网络中，关键的步骤是提取出最大词团、文档团以及将最大词团映射到文档团上。按照最大团的相关性权重来选择用于扩展的最大词团，以最大团为单位作为一个概念整体扩展进来，这样有利于把语义比较集中的词扩展进来，同时也提高了那些与某个查询词的关系不是很强，但与查询主题很相关的词加入查询扩展的可能性，从而提高检索效果。在图 1 的词项子空间中，词项 t_2 的最大团有{t_2、t_4、t_5}，{t_2、t_3、t_6、t_7}，且每个最大词团会有相应的权重，取决于团内各词项间的依赖程度。

2.3 词团与文档团的映射

词团与文档团之间的映射关系可以用来强化索引词项与用户给定的初始查询之间的关联性。如果最大词团中的一个索引词项出现在一个文档团的多篇文档中，则认为包含该词项的词团与该文档团存在语义上的关联性，同时也认为这个词团与查询的主题更加相关。因此，本文通过这种映射关系将最大词团分为文档依赖词团和非文档依赖词团，并将这两种词团用于查询扩展。由于文档依赖词团可能对文档团的主题更具代表性，因此，在检索阶段会给文档依赖词团赋予更大的权重。如图 1 所示，词项子空间内有三个最大词团 T_1={t_1、t_4、t_5}、T_2={t_2、t_4、t_5}和 T_3={t_2、t_3、t_6、t_7}，文档空间内有三个最大文档团 D_1={d_1、d_2、d_3、d_5}、D_2={ d_2、d_3、d_4}、D_3={ d_2、d_4、d_6}。词项 t_4 在文档团 D_1 中的所有文档中都出现，所以词团 T_1 和 T_2 都是文档依赖词团，而词团 T_3 中的任何词项都没有在同一个文档团中的所有文档中出现，所以词团 T_3 是非文档依赖词团。

2.4 基于文档团依赖的 Markov 网络检索模型

对于用户给定的初始查询 q，通过利用词项、文档子空间的 Markov 网络信息，计算文档集 D 中任意文档 $d_j \in D$ 和查询 q 的相关概率 $p(d_j|q)$。然后按照 $p(d_j|q)$的大小排列文档集中的文档，从而得出我们需要的文档。由左家莉等提出的基于 Markov 网络的信息检索扩展模型[14]可得：

$$p(d_j|q) \approx \sum p(t_i|q)\, p(t_i|d_j) \tag{4}$$

其中，$p(t_i|q) = \begin{cases} w_{i,q} & t_i \in q \\ 0 & t_i \notin q \end{cases}$，$p(t_i|d_j) = \begin{cases} w_{i,j} & t_i \in d_j \\ 0 & t_i \notin d_j \end{cases}$。

本文采用 BM25 类似权重方式来计算 $w_{i,q}$ 和 $w_{i,j}$，具体的计算公式参照文献[14]。

在查询扩展阶段，扩展词的选择方案采用基于最大团的方式，以最大词团为单位，作为一个概念整体对初始查询表式中对应的词项进行扩展。在选取最大词团阶段我们将词团分为两类，分别是：文档依赖词团和非文档依赖词团。利用词团与文档团之间的映射关系来提取文档依赖词团，并在检索阶段加大这类词团的权重。对于用户给定的初始查询 q，文档和查询的相关概率计算公式由式(4)修正为：

$$P(d_j|q)=\sum_{t\in q}(1-\alpha-\beta)P(t|q)P(t|d_j)+\alpha\sum_{t_k\neq t\cap t_k\in C_{\max}(t)}P(t_k|q)P(t_k|d_j)$$
$$+\beta\sum_{t_l\neq t,t_k\cap t_l\in C_{\max}(t)}P(t_k|q)P(t_l|d_j) \quad (5)$$

其中，α：非文档依赖词团平滑参数（$0\leq\alpha\leq1$），β：文档依赖词团平滑参数（$0\leq\beta\leq1$），$\alpha+\beta<1$，$\begin{cases}P(t_k|q)=sim(t_k,t)\cdot P(t|q)\\P(t_l|q)=sim(t_l,t)\cdot P(t|q)\end{cases}$，***sim*** 指用公式(1)计算词项之间相关性的值。

公式(5)中的 $C_{max}(t)$代表词项 t 的最大词团，在提取最大词团时，会给每个词项的最大词团赋予不同的权重。本文提出这样的假设：最大词团的权重越大，则它所表达的概念与查询主题越相关，在查询扩展的时候优先被考虑进来。在实验中，利用公式(5)对文档集做初次检索，通过调整最大词团的个数以及式(5)中的平滑参数，使该模型的检索效果达到稳健。

3. 基于迭代方法的多层 Markov 网络信息检索模型

3.1 查询子空间的 Markov 网络

在图 1 中，查询子空间中的节点（查询）之间也有边相连，构成查询子空间的 Markov 网络。本文中查询之间的相关性由查询返回的结果集合中排名靠前的文档所决定，也就是利用相关反馈方法中的隐式反馈信息[1]。通过使用 2.4 节介绍的基于文档团依赖的 Markov 网络检索模型[10]对文档集做初次检索，得出用户给定查询返回的结果集合，取结果集合中查询 q_i 返回的前 2 篇文档，将这 2 篇相关文档合并成一个新的“文档”D_i，采用类似 2.1.2 节中文档之间相关性的度量方法来确定查询间的相关性，记为 $Sim(q_i,q_j)$，计算公式如下：

$$Sim(q_i,q_j)=\frac{\overrightarrow{D_i}\times\overrightarrow{D_j}}{\left|\overrightarrow{D_i}\right|\times\left|\overrightarrow{D_j}\right|} \quad (6)$$

当 $Sim(q_i,q_j)$的值大于给定的阈值时，则查询 q_i 和 q_j 相互依赖，即在查询子空间的 Markov 网络中有边相连。

3.2 查询团的提取

使用公式(6)计算出查询间的相关性，从而构建查询子空间的 Markov 网络。将所得到的的 Markov 网络信息用于最大查询团的提取，具体的提取方法与 2.2 节中的方法一致，所得到的查询子空间的 Markov 网络信息也可以用在查询扩展中，从而提高检索效果。

3.3 基于迭代方法的多层 Markov 网络查询扩展模型

结合 2.4 节基于文档团依赖的 Markov 网络检索模型[10]以及最大查询团来构建本文的基于迭代方法的多层 Markov 查询扩展模型。这样，就将词项、文档和查询三层子空间的 Markov 网络信息用于查询扩展中。加入了查询子空间的 Markov 网络信息后，将文档和查询的相关概率计算公式由式(5)修正为：

$$P(d_j|q)=(1-\theta)P'(d_j|q)+\theta\sum_{q_i\neq q\cap q_i\in C_{\max}(q)}sim(q,q_i)\cdot P'(d_j|q_i) \quad (7)$$

上式(7)中的 $P'(d_j|q_i)$ 是在模型中加入了查询子空间的 Markov 网络信息，计算文档 d_j 与初始查询 q 的最大查询团 $C_{max}(q)$中的其他查询 q_i 的相关概率，计算公式类似于公式(5)，如下所示：

$$P'(d_j|q_i)=\sum_{t\in q}(1-\alpha-\beta)P(t|q_i)P(t|d_j)+\alpha\sum_{t_k\neq t\cap t_k\in C_{\max}(t)}P(t_k|q_i)P(t_k|d_j)$$
$$+\beta\sum_{t_l\neq t,t_k\cap t_l\in C_{\max}(t)}P(t_k|q_i)P(t_l|d_j) \quad (8)$$

其中，α：非文档依赖词团平滑参数（$0\leq\alpha\leq1$），β：文档依赖词团平滑参数（$0\leq\beta\leq1$），$\alpha+\beta<1$，$\begin{cases}P(t_k|q_i)=sim(t_k,t)\cdot P(t|q_i)\\P(t_l|q_i)=sim(t_l,t)\cdot P(t|q_i)\end{cases}$。

构造本文基于迭代方法的多层 Markov 网络信息检索模型时，迭代的具体步骤如下：

1) 第一次迭代过程：
   a) 利用公式(5)中得到的检索结果提取出最大查询团信息；
   b) 取出检索结果中文档与查询的相关概率值，并代入公式(7)中计算新的相关概率，即将式(5)中的 $P(d_j|q)\rightarrow P'(d_j|q)$；
   c) 用公式(7)对文档集进行重新检索；
2) 第 n(n>1)次迭代过程：
   a) 从前一次(n-1)迭代过程中得到的检索结果中提取出最大查询团信息；
   b) 将检索结果中文档与查询的相关概率值代入公式(7)中重新计算文档得分；
   c) 利用公式(7)对文档集再次检索。

重复步骤 2)进行多次迭代，直到实验的检索效果收敛（即检索效果已经不再提高或者提高的比例不大）为止。实验过程中，通过调整最大词团的个数以及式(7)、(8)中的平滑参数θ、α、β，使该模型的检索效果达到稳健。

4. 实验设计和结果

4.1 对比实验

本文采用的是一种常用的测试集，该测试集由 adi、med、cran、cisi 及 cacm 五个标准测试文档集组成，常用于评价检索系统的性能。五个测试文档集的文档较小，且评价效果良好，下载地址：ftp://ftp.cs.cornell.edu/pub/smart. 测试数据集的具体情况如下表 1 所示：

表 1 实验中的数据集

数据集	类型	文档总数	查询数	词项数
adi	信息科学	82	35	893
med	医学	1033	30	8702
cran	航空	1398	225	4110
cisi	图书馆科学	1460	76	5494
cacm	计算机科学	3204	64	5041

为了验证基于迭代方法的多层 Markov 网络信息检索模型(IMMR)的检索效果，本文选取使用普通 BM25 模型和基于 Markov 概念的信息检索模型[8](MRC)来进行对比实验。并把基于文档团依赖的 Markov 网络检索模型[10]作为 Baseline，与本文提出的模型的主要区别在于查询扩展中没有利用查询子空间的 Markov 网络信息，以及没有采用基于迭代的方法来计算文档与查询的相关概率。在表中可以看出本文提出的检索模型实验结果比起基准方法有比

较大的提高，我们采用的评价指标是 3-avg 和 11-avg。3-avg 和 11-avg 分别代表一组测试查询在 3 个召回率点（0.2，0.5，0.8）和 11 个召回率点（0，0.1，0.2，……，1.0）上精确率的平均值，这种平均精度-召回率数值如今已成为信息检索系统的标准评价指标，在信息检索文献中被广泛采用[1]。本文实验的具体结果见表 2 和表 3。

从实验结果中，我们可以发现本文提出的模型在数据集 cacm 和 adi 上相对于 Baseline 模型有最大幅度的提高。对于数据集 cacm 主要是因为其含有的文档数最多，文档团与词团的映射关系更为明显，可以有效地将最大词团划分成文档依赖词团和非文档依赖词团，并提高从较多的文档里找到与用户需求真正相关文档的概率；对于 adi 数据集，虽然数据集最小，但是由于其自身的特殊性，可以加入的修正信息最多，也能取得较好的检索效果。而数据集 med 虽然包含的词项数多，但文档总数及查询数较少，构造出的文档、查询子空间的 Markov 网络信息较少，使得查询扩展时加入的修正信息少，导致最后的检索效果提升不明显。

表 2 3-avg 实验结果

	adi	**med**	**cran**	**cisi**	**cacm**
BM25	42.97%	29.56%	42.72%	21.43%	31.23%
MRC	43.93%	54.19%	45.06%	22.87%	31.72%
Baseline	48.56%	55.97%	45.22%	23.22%	32.88%
IMMR	**51.03%** **(+5.09%)**	**56.36%** **(+0.70%)**	**47.26%** **(+4.51%)**	**24.74%** **（+6.55%）**	**35.63%** **（+8.36%）**

表 3 11-avg 实验结果

	adi	**med**	**cran**	**cisi**	**cacm**
BM25	41.12%	30.96%	39.36%	23.67%	32.48%
MRC	42.31%	53.31%	45.16%	25.08%	33.92%
Baseline	46.36%	53.99%	45.27%	25.23%	34.50%
IMMR	**49.04%** **(+5.78%)**	**54.47%** **(+0.89%)**	**46.55%** **（+2.83%）**	**26.50%** **（+5.03%）**	**36.23%** **（+5.01%）**

本文提出的基于迭代方法的多层 Markov 网络信息检索模型中，采用迭代的方法计算文档与查询的相关概率，以及从检索结果集合中构造查询子空间的 Markov 网络。理论上，在检索结果收敛之前，每一次迭代过程都能将与用户查询更加相关的文档排在前面，从而构造出更加符合用户查询意图的查询子空间的 Markov 网络，并最终提高检索效果。实验中我们发现，迭代 2~3 次检索结果就基本达到收敛，且第 1 次迭代过程检索效果提升的幅度较大。主要的原因有：1)第 1 次迭代过程中，在基于文档团依赖的 Markov 网络检索模型[10]的基础上初次加入了查询子空间的 Markov 网络信息，使得检索结果有明显提高；2)本文用于构造查询子空间的 Markov 网络采用的方法是从检索结果集合中取出查询 q_i 返回的前 2 篇文档，来计算查询间的相似性，加上实验中采用的数据集文档数相对较少，返回的前 2 篇文档不会有太大改变，使得后续构造的查询子空间的 Markov 网络基本不变。未来的工作中将在更大的数据集上做测试，同时改进构造查询子空间的 Markov 网络的方法（比如采用查询日志信息），相信通过迭代方法能够取得较好的检索效果。

4.2 相关性阈值的选取

在实验中，发现调整词项间、文档间以及查询间相关性的阈值会使得提取最大团的计算量变化很大，最大团的个数以及最大团中包含词、文档和查询的数量也会有较大的影响，最终对检索效果也会产生比较明显的影响。对于词项间相关性阈值的设定，数据集 adi 中取 0.7，其余四个数据集均取 0.8，这是因为 adi 数据集词项数最少，需要加入用于构造词项子空间的 Markov 网络的词项数多些。词项间的相关性阈值越大，得到的最大团的个数 s 越少，相反则最大团的个数 s 越多，用于提取最大团的计算时间也越长。当词项网络固定时，用于加入查询扩展中词项的最大团个数 s 会对实验结果产生较大影响：一开始随着 s 的增加，检索效果会随之提高；当 s 增加到一定的数值时，检索效果达到最优，若再增大则结果会逐渐降低。在文档间相关性阈值的设定上，根据 5 个数据集中文档总数的不同取不同的值，实验中数据集 adi 的文档数最少，取阈值 0.1，med 取 0.2，cran、cisi、cacm 取 0.3。同样地，对于查询间相关性阈值的取值也与数据集中包含的查询数和需要加入的修正信息量有关，实验中数据集 adi 的查询数少且可加入修正的信息量大，取阈值 0.2，med、cisi、cacm 取 0.3，cran 数据集含有的查询数最多阈值取 0.5。

4.3 平滑参数的调整对实验结果的影响

本文提出的模型中包含 θ、α、β 这 3 个平滑参数，分别对应加入查询团信息的权重、非文档依赖词团权重以及文档依赖词团权重。通过调整平滑参数的取值，使检索效果达到最优，由于文档依赖词团可能对文档团的主题更具代表性，需要赋予更高的权重。因此，实验中平滑参数 α 和 β 初始值为 0，令参数 α 以 0.01，β 以 0.03 的增量来循环计算文档与查询的相关概率，并记录在每次循环中各参数取值确定时的实验结果。实验中通过调整 3 个平滑参数会使得检索结果有较大影响，以数据集 adi 和 cacm 为例，下图 2 和图 3 表示在第一次迭代过程中，调整平滑参数对实验结果的影响(实验中，取 $\theta=\alpha$，$\beta=3\alpha$)：

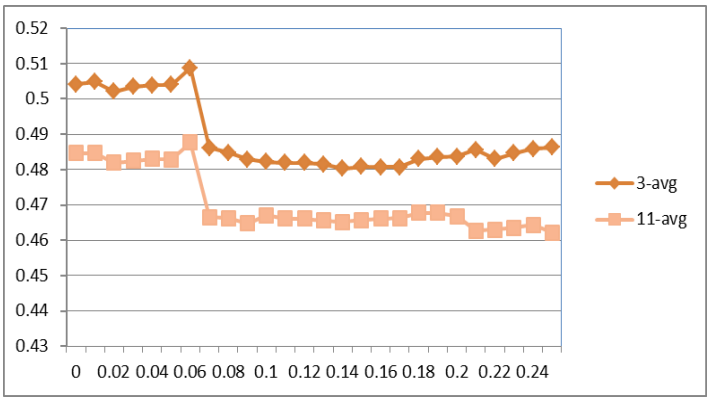

图 2 adi 数据集的实验结果

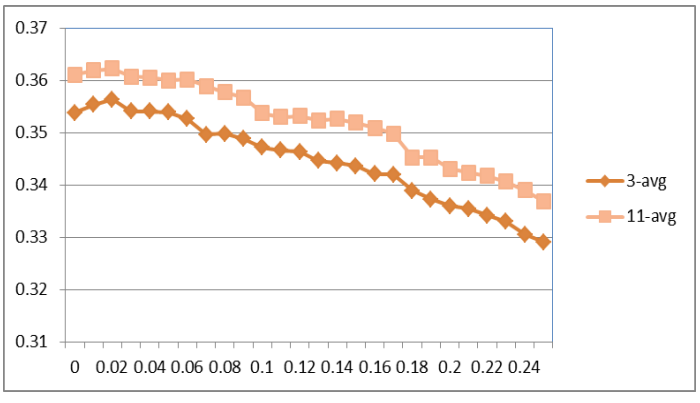

图 3 cacm 数据集的实验结果

在上图中，横坐标表示α的取值，纵坐标表示平均准确率。实验表明，α和β的取值对结果的影响遵循一定的规律，在一定的范围内平滑参数α越大，检索效果越好，但达到一定数值以后结果开始变差。另外，它们的取值与文档集规模也有一定的关系。在实验当中，当文档集规模较小时，α和β取较大的数值实验结果才能达到相对较好水平，如上图 2, 3 所示，在 adi 数据集中α=0. 06，β=0. 18 时，实验结果达到稳健，然而在 cacm 数据集中α=0. 02，β=0. 06 时，实验结果才能达到稳健，这是因为 adi 数据集中文档数目要远远少于 cacm 数据集中的文档数目，用于修正查询的信息量少，需要通过提高修正信息的权重来取得好的检索效果。

5. 总结与展望

本文通过对训练文档集和检索结果集合的学习，构造出词项、文档和查询三层子空间的 Markov 网络，将得到的三层 Markov 网络信息用于最大词团、文档团和查询团的提取，利用文档团与词团之间的映射关系将最大词团分为文档依赖词团和非文档依赖词团，在查询扩展中加大文档依赖词团的权重。模型中通过迭代的方法从已有的检索结果集合中提取出最大查询团信息，并迭代地计算文档与查询的相关概率，直到检索效果收敛。下一步的工作有：(1)在更大的数据集上（比如 TREC 数据集）进行实验检测该模型的通用性；(2)本文是对已有的检索结果集合采用隐式的相关反馈方法，收集检索结果返回的前 2 篇相关文档信息来计算查询间的相关性，未来我们将利用用户查询日志来构造查询子空间的 Markov 网络；(3)随着数据集规模的增大，构造 Markov 网络计算量会变得非常大。因此，在面对海量数据检索时，一方面需优化 Markov 网络构造方法；另一方面可将词项、文档和查询间的相关性计算算法采用 MapReduce 编程模型并行化，加快执行效率。

参 考 文 献

[1] 黄萱菁, 张奇，邱锡鹏 译. 现代信息检索[M]. 第一版. 机械工业出版社, 2012 年 10 月.

[2] Zhai C, Lafferty J. Model-based feedback in the language modeling approach to information retrieval[C]//Proceedings of the tenth international conference on Information and knowledge management. ACM, 2001: 403-410.

[3] Tao T, Zhai C X. A mixture clustering model for pseudo feedback in information retrieval[M]//Classification, Clustering, and Data Mining Applications. Springer Berlin Heidelberg, 2004: 541-551.

[4] Soskin N, Kurland O, Domshlak C. Navigating in the dark: Modeling uncertainty in ad hoc retrieval using multiple relevance models[M]//Advances in Information Retrieval Theory. Springer Berlin Heidelberg, 2009: 79-91.

[5] Tao T, Zhai C X. Regularized estimation of mixture models for robust pseudo-relevance feedback[C]//Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006: 162-169.

[6] Lv Y, Zhai C X, Chen W. A boosting approach to improving pseudo-relevance feedback[C]//Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM, 2011: 165-174.

[7] 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. ACM, 2008: 235-242.

[8] 甘丽新. 基于 Markov 概念的信息检索模型[D]. 江西师范大学, 2007.

[9] 付剑波, 王明文, 罗远胜, 等. 基于团模型的文档重排算法研究[J]. 中文信息学报, 2009, 23(1): 71-78.

[10] 汤皖宁. 基于文档团的 Markov 网络检索模型[D]. 江西师范大学, 2013.

[11] 何盈捷, 刘惟一. 由 Markov 网到 Bayesian 网[J]. 计算机研究与发展, 2002, 39(1): 87-99.

[12] 王斌 译. 信息检索导论[M]. 第一版. 人民邮电出版社, 2010 年 9 月.

[13] 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. ACM, 2007: 311-318.

[14] 左家莉, 王明文, 王希. 基于 Markov 网络的信息检索扩展模型[J]. 清华大学学报: 自然科学版, 2005, 45(1): 1847-1852.

作者联系方式：

姓名：洪欢

地址：江西师范大学瑶湖校区紫阳大道 99 号

邮编：330022

手机号码：15170474916

电子邮箱：honghuan252008@126.com