计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2014.20130343
51(10):2248-2254,2014

# 基于文档团的 Markov 网络检索模型

汤皖宁　王明文　万剑怡
(江西师范大学计算机信息工程学院　南昌　330022)
(tangwanning23@126.com)

# Markov Network Retrieval Model Based on Document Cliques

Tang Wanning, Wang Mingwen, and Wan Jianyi
(*School of Computer and Information Engineering, Jiangxi Normal University, Nanchang* 330022)

**Abstract**　The query expansion is an effective way to improve the efficiency of information retrieval. But many of the query expansion methods to select the expansion terms did not take fully account of the correlation between the terms as well as terms and documents, which may reduce retrieval performance. Due to the information of the correlation between terms and documents is able to improve the efficiency of retrieval, this paper calculates the correlation between documents and terms, and mapping terms to documents to build a Markov network retrieval model; and then extracts term clique according to the mapping information. The mappting information is used to divide the term cliques into two categories. One is based on document and another is not based on document. The terms cliques based on document are more relevant with the query topic, so to the terms cliques are given greater weight based on document and the information of the two kinds of terms cliques is used to assist retrieval. Therefore, the method we propose in this paper can make the extension content more relevant to query. Experimental results show the proposed model can improve the retrieval efficiency.

**Key words**　information retrieval; query expansion; Markov network; retrieval model; clique

**摘　要**　查询扩展是提高检索效率的有效方法.但是许多查询扩展方法中扩展词的选择没有充分考虑词项之间以及词项与文档之间的相关性,这样可能在查询扩展时加入太多不相关信息降低检索的性能.通过对文档间相关性和词间相关性的计算,把文档和词关联起来构建 Markov 网络检索模型,然后根据词项子空间和文档子空间的映射关系提取词团,将提取的词团信息用于查询扩展,使得查询扩展的内容更为相关.实验表明:基于文档团依赖的 Markov 检索模型能有效地提高检索效果.

**关键词**　信息检索;查询扩展;Markov 网络;检索模型;团

**中图法分类号**　TP391

随着计算机网络的出现和迅速发展,人们能接触到的信息越来越多.一方面用户可以迅速、方便地接触到丰富的信息,另一方面,如何在如此繁杂庞大的信息中找到自身真正需要的信息却又是异常困难.如今,人们越来越关注如何从浩如烟海的信息中迅速而准确地查找到所需要的资料.因而,信息检索的地位也就显得日益重要.典型的查询扩展模型有相关模型[1-3]、混合模型[4]等.这些模型都是基于词

**收稿日期**:2013-03-27;**修回日期**:2013-06-28
**基金项目**:国家自然科学基金项目(61272212,61163006)
**通信作者**:王明文(mwwang@jxnu.edu.cn)

独立性的假设,但实际上词之间的关联信息对信息检索最后的性能有很大的影响. 在近年的一些研究中,Zhai 等人通过 Boosting 算法将相关模型和混合模型合并来提取查询扩展词,该方法将 2 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 等人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. Xu 等人将用户提出的查询分为 3 类,对不同的类别采用不同的查询扩展方式,并且通过维基百科来辅助查询扩展词的选择[8],该方法不仅加入初始检索出的文档而且还利用维基百科的内容来辅助查询扩展,但是由于辅助内容并没用考虑与初始检索文档之间的相关性,这样可能造成主题漂移,从而影响检索效率. 在信息检索中查询扩展已被证明能有效地提高检索的性能. 虽然很多查询扩展能够对大多数查询提供帮助,但是这种技术可能造成主题漂移而损害另外一些查询的性能.

因此,本文利用 Markov 网络模型将词之间、文档之间的关联信息以及词与文档之间的映射信息加入到查询扩展技术中. 本文的方法首先通过计算词之间的相关性、文档之间的相关性构造索引词空间和文档空间,然后提出最大团概念. 在这 2 个空间中提取出最大词团和最大文档团,将最大词团与最大文档团进行映射. 最后查询扩展阶段将最大词团分为 2 类:一类是文档依赖最大词团,另一类是非文档依赖最大词团. 因为文档依赖的词团更可能表达的是同一个主题,防止主题漂移. 本文的方法考虑了词之间的相关性和文档之间的相关性,并且将词和文档之间的映射信息加入到查询扩展中. 将我们的模型与一些经典的模型进行了比较,实验表明本文模型的检索性能更优.

## 1 Markov 网络检索模型

基于 Markov 网络的检索模型分为 3 层:查询空间、索引词项空间、文档空间. 如图 1 所示,所有的层构成了一个推理网络,根节点为词子空间的词节点,我们利用词与词之间的相关性构造词项空间. 如果索引项空间的一个词在同一文档团的多篇文档出现,我们利用这个信息来构造索引项空间与文档子空间之间的映射关系.

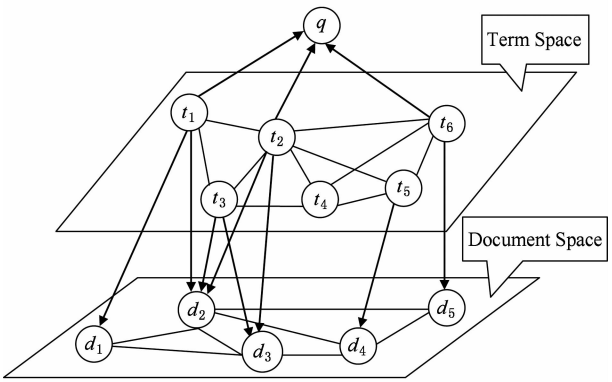


Fig. 1 Markov network retrieval model based on document cliques.

图 1 基于文档团的 Markov 网络模型

从模型的网络结构直观地来看,对于一个给定的文档集,索引项空间只要建立就不会改变,其结构仅仅受到索引项之间的相关性的影响. 查询空间针对特定用户的信息需求,在查询处理过程中,根据索引词之间的相关性以及索引词与文档的相关性进行的扩展,那么查询空间的查询词也会有一定的调整. 这样调整后的查询词能更好体现用户信息需求,从而也使检索结果更符合用户信息需求.

在 Markov 网络信息检索模型中,给定一个用户的查询,我们要计算文档集中的文档和查询的相关概率即条件概率. 通过 Markov 网络信息检索模型的学习,构造索引项空间,这样使得提取了索引项之间的相关信息,并且利用索引项与文档之间的映射关系优化查询与索引项之间的相关性. 利用从文档中训练得到的这些相关信息,Markov 网络模型可对原始查询进行扩展,从而达到提高检索效果的目的.

## 2 Markov 网络检索模型构造方法

Markov 网络检索模型分为词项空间构造、文档空间构造以及这 2 个子空间之间映射 3 个方面. 通过下面给出的词项以及文档项相关性的度量方式构造词项及文档项空间,然后通过词项与文档项之间的映射建立 2 个子空间之间的关联. 最后通过这 2 个子空间的构造和映射来构造 Markov 网络检索模型.

### 2.1 相关性度量

1) 词项相关性的度量

利用词的共现性获取词之间的关系已经运用到很多研究中. 通常计算词共现的词频时以整个文档、段落或是一个固定长度为窗口. 因此词之间的关系强度计算如下:

$$P_{co}(t_i|t_j)=\frac{C(t_i,t_j)}{C(t_j)}. \quad (1)$$

本文实验均采用词的共现性来提取词之间的关系,鉴于 Markov 网络的无向性考虑,在构造 Markov 网络时,采用综合词的共现性来计算如下:

$$sim(t_i,t_j)=\frac{p_{co}(t_i|t_j)+p_{co}(t_j|t_i)}{2}. \quad (2)$$

2) 文档项相关性的度量

在文档空间内,文档与文档之间的关系,记为 $sim(d_i,d_j)$,本文采用文档之间的夹角来度量文档之间相关性,即

$$sim(d_i,d_j)=\frac{d_i\times d_j}{|d_i|\times|d_j|}. \quad (3)$$

### 2.2 文档与索引词项之间的映射

本文利用文档团与词团之间的映射信息强化查询与索引词项之间的关联性. 如果词团的一个索引词项出现在一个文档团多篇文档中,则认为这个词团与查询的主题更加相关,因此在词项扩展阶段加大该词团的权重. 如图 1 所示,词项空间内有 2 个最大词团 $T_1=\{t_1,t_2,t_3\}$ 和 $T_2=\{t_2,t_4,t_5,t_6\}$,文档空间内有 3 个最大文档团 $D_1=\{d_1,d_2,d_3\}$,$D_2=\{d_2,d_3,d_4\}$,$D_3=\{d_3,d_4,d_5\}$. 词团 $T_1$ 内所有词项都在文档 $d_2$ 内出现,然而词团 $T_2$ 内词项没有同时出现在一个文档内,由于最大文档团内部具有较强的语言关联性,因而,词团 $T_1$ 内的词项同时出现在一个文档团内的文档中,所以与词团 $T_1$ 相比词团 $T_2$ 更能够表达查询的意图.

### 2.3 Markov 网络检索模型构造方法

构造 Markov 网络需要构造索引词项空间、文档词项空间. 本文是选用 2.1 节中所述词的共现性来确定词与词之间的关系. 本文选择文档作为窗口单位,这是因为考虑效率方面因素. 对文档集的倒排文件进行统计,利用式(1)(2)就可以得到词与词之间的相关性. 文档与文档之间的关系采用 2.2 节中文档之间夹角来度量. 利用式(3)就可以得到文档与文档之间的相关性.

根据式(1)~(3)可以得到词与词、文档与文档组成的 Markov 网络,网络中的边表示词之间、文档之间的依赖关系.

## 3 团的提取和词团与文档团的映射

### 3.1 团的提取

团的提取分为索引项词团的提取和文档团的提取.

通过词以及文档的 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.

词和文档空间构成的 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(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. 如果一个词团中的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

在词和文档的 Markov 网络中提取最大团以及将词团映射到文档团上是比较关键的步骤. 根据离散数学的定理:$C$ 是一个团,那么必存在一个最大团 $C_{\max}$,使得 $C\subset C_{\max}$,在文献[9]中该定理的证明如下:设 $T=\{t_1,t_2,t_3,\cdots,t_k\}$,构造团序列 $C_0\subset C_1\subset C_2\subset\cdots\subset C_{\max}$. 其中 $C_0=C$ 且 $C_{i+1}=C_i\cup\{t_i\}$,其中 $j$ 是满足 $t_j\notin C_i$ 而 $t_j$ 与 $C_i$ 中各节点都有相容关系的下标. 由于 $T$ 的词节点个数 $|T|=n$,所以至多经过 $n-|C|$ 步,就使得这个过程终止,而此序列的最后一个团就是要找的最大团.

从这个定理的证明过程可以看出:1)在 $C_i$ 基础上每一步增加一个与 $C_i$ 中的每个节点都有边相连的节点 $t_j$,得到 $C_{i+1}$,通过步步迭代,最终得到 $C_{\max}$;2)属于 $C_{k+1}(t_i)$ 的任何 2 个团,若它们真包含在 $C_{k+1}(t_i)$ 中,则这 2 个团只存在 2 个不同的词节点,且这 2 个词节点之间是有边相连的,即相关. 因此,本文在词的 Markov 网络中提取词的团采用了上述思想,即在 $C_k(t_i)$ 基础上来获取 $C_{k+1}(t_i)$. 该算法的时间复杂度为 $O(n^2)$.

### 3.2 文档团与词团的映射

提取了文档团和词团后,团内部的文档和词项相互之间存在较强的语义关联性,在文档团与词团之间也存在着映射关系,如果一个词团中索引词项同时映射到一个文档团中的多篇文档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. 因此,本文通过这种映射信息将词团分为文档依赖词团和非文档依赖词团. 由于文档依赖词团可能对文档团的主题更具代表性,因此,在检索阶段给赋予文档依赖词团更大的权重.

## 4 模型的构造

基于文档团的 Markov 网络检索模型给定查询 $q$,利用 Markov 网络可以计算文档集 $D$ 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. 因此需要 $p(d_j|q)$. 在构造的 Markov 网络中,$M_T$ 是词子空间的 Markov 网络,由条件概率定义可得:

$$p(d_j|q)=\frac{p(d_j,q)}{p(q)}\propto p(d_j,q)=$$
$$p(d_j|T,D,M_T)p(q|T,M_T)p(T|M_T)p(M_T). \tag{4}$$

若索引项和文档的网络固定,$p(T|M_T)$ 和 $p(M_T)$对任一 $d_j$ 均相同,则由式(4)可知:

$$p(d_j|q)\propto p(d_j|T,D,M_T)p(q|T,M_T)). \tag{5}$$

式(5)为模型检索算法的一般形式. 从网络结构可以推出:

$$p(q|T,M_T)\propto p(T,M_T|q)$$
$$\frac{P(q)}{p(T,M_T)}\propto \sum_{t_i\in q}P(t_i|q), \tag{6}$$
$$p(d_j|T,D,M_T)=p(T,D,M_T|d_j)$$
$$\frac{P(d_j)}{p(T,D,M_T)}\propto \sum_{t_i\in T}P(t_i|d_j). \tag{7}$$

那么将式(6)(7)代入式(5)中,式(4)最终可以变形为

$$p(d_j|q)\propto p(t_j|d_j)p(t_i|q)). \tag{8}$$

本文实验采用 BM25 类似权重方式:

$$w_{i,q}=\frac{\ln(tf_i)^2}{\Phi\sum_{t_i\in q}\ln(tf_i)^2}, \tag{9}$$

其中,$\Phi$ 为归一化因子.

$$w_{i,j}=\ln\left(\frac{N-n_i+0.5}{n_i+0.5}\right)\times$$
$$\frac{(k+1)\times f_{j,i}}{k\times\left\{[(1-b)]+b\times\dfrac{dlf_j}{avedlf}\right\}+f_{j,i}}. \tag{10}$$

信息检索领域通常忽略词之间的相关性,即假设词与词之间是独立不相关的,实际上这种假设并不成立. 本文假设如果词团的词项映射到同一文档团中的多篇文章,则认为包含这个词的词团更能代表对应的文档团的主题,对最后的检索性能影响更大. 因此提出了基于文档依赖最大词团的 Markov 网络检索模型. 在查询扩展阶段,本文采用的是结构化查询扩展方法,检索词的选择方案采用基于最大团的方式,以最大团为单位,作为一个概念整体与原始查询词重新组成一个新查询. 在选取最大团阶段我们将团分为 2 类:文档依赖最大团和非文档依赖最大团. 利用词团与文档团之间的映射来提取文档依赖最大团. 查询扩展阶段加大这类词团的权重.

该模型实质上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. 因此,在检索阶段,本文以最大团为单位,作为一个概念整体与原始查询词组成一个新查询,通过修正词的权重,重新构造文档和查询之间的相关性. 因此给定查询 $q$、文档 $d$ 和 $q$ 的相关概率,式(8)修正为

$$p(d_j|q)\propto \sum_{t_i\in q}(1-\alpha-\beta)p(t_i|q)p(t_i|d_j)+$$
$$\alpha\sum_{t_k\neq t_i\wedge t_k\in C_{\max}(t_i)}p(t_i|q)p(t_i|d_j)+$$
$$\beta\sum_{t_l\neq t_i\wedge t_k\wedge t_l\in C_{\max}(t_i)}p(t_l|q)p(t_l|d_j), \tag{11}$$

其中,$t_k$,$t_l$ 表示一个词;$\alpha$ 为非文档关联的词团权重参数($0\leqslant\alpha\leqslant1$);$\beta$ 为文档关联的词团权重参数($0\leqslant\beta\leqslant1$);$P(t_k|q)=S_{i,k}P(t_i|q)S_{i,k}\propto sim(t_i,t_j)$且 $0\leqslant S_{i,k}\leqslant1$. 通过系数 $S_{i,k}$可以加入相关性程度信息,即 $sim(t_i,t_j)$越大,$S_{i,k}$越大.

本文利用词的共现性构造词和文档的 Markov 网络,从词的 Markov 网络提取出词的最大团. 本文提出了一个这样的假设:最大团的权重越大,认为它所表达的概念与查询主题越相关,对查询越有利,在查询扩展时被考虑优先扩展进来. 在式(11)中,通过调整权重参数,使得本文的模型达到检索性能的最优效果.

## 5 实验设计和结果

### 5.1 对比实验

为了验证本文提出方法的效果,本文选取了 5 个检索算法[10]和基于团的 Markov 网络模型[11]以及经典的相关反馈算法 Rocchio 来进行对比实验. 我们从中选取了以下 5 种检索模型进行比较:hits 模型、tf 模型、idf 模型、tf * idf 模型以及 BM25 模型.

在相同的数据预处理的方式下,我们比较不同模型的性能差异. 其中,对于 5 个标准测试集如表 1 所示,我们利用这 5 个数据集分别进行实验,得到实验结果.

**Table 1 Data Set of Experimental**

**表 1 实验的数据集**

| Data Set | Topic | # Document | # Query | # Term |
|---|---|---|---|---|
| ADI | Information | 82 | 35 | 893 |
| MED | Medical | 1 033 | 30 | 8 702 |
| CRAN | Aviation | 1 400 | 225 | 4 110 |
| CISI | Library | 1 460 | 76 | 5 494 |
| CACM | Computer | 3 024 | 64 | 5 041 |

以文献[10]中提到的模型实验结果为基准,文献[10-11]中的模型与本文的模型区别在于没有将索引空间的词项与文档空间的文档进行映射,从表 1 可以看出我们的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 11-*avg* 和 3-*avg*. 实验结果如表 2、表 3 所示,本文提出的基于文档团的 Markov 网络信息检索(DCMR)是取最优检索结果.

**Table 2 Experimental Result of 11-*avg***

**表 2 召回率不同时准确率的平均值 11-*avg***

| Model | ADI | MED | CRAN | CISI | CACM |
|---|---|---|---|---|---|
| hits | 0.279 9 | 0.404 1 | 0.276 4 | 0.140 5 | 0.186 7 |
| tf | 0.018 | 0.056 3 | 0.148 | −0.063 5 | 0.110 8 |
| idf | 0.299 7 | 16.37 | 0.245 7 | 0.189 5 | 0.276 9 |
| tf * idf | 0.364 3 | 16.10 | 0.326 9 | 0.211 5 | 30.74 |
| BM25 | 0.411 2 | 0.309 6 | 0.393 6 | 0.236 7 | 0.324 8 |
| Rocchio | 0.422 | 0.401 3 | 0.413 9 | 0.241 | 0.331 5 |
| Baseline | 0.431 5 | 0.567 4 | 0.453 1 | 0.251 | 0.341 7 |
| DCMR | 0.463 6 | 0.594 1 | 0.482 99 | 0.281 | 0.363 6 |

**Table 3 Experimental Result of 3-*avg***

**表 3 召回率不同时准确率的平均值 3-*avg***

| Model | ADI | MED | CRAN | CISI | CACM |
|---|---|---|---|---|---|
| hits | 0.282 1 | 0.394 1 | 0.261 5 | 0.112 1 | 0.160 1 |
| tf | 0.026 | 0.090 7 | 0.196 5 | 0.084 5 | 0.151 4 |
| idf | 0.263 2 | 0.209 8 | 0.283 2 | 0.164 9 | 0.198 2 |
| tf * idf | 0.364 | 0.216 | 0.390 1 | 0.208 5 | 0.290 6 |
| BM25 | 0.429 7 | 0.295 6 | 0.627 2 | 0.214 3 | 0.312 3 |
| Rocchio | 0.445 7 | 0.384 5 | 0.464 1 | 0.231 4 | 0.315 6 |
| Baseline | 0.447 2 | 0.553 6 | 0.452 5 | 0.229 6 | 0.323 5 |
| DCMR | 0.485 6 | 0.570 5 | 0.476 6 | 0.240 9 | 0.360 4 |

在实验中,发现调整词之间相似度的阈值会使得提取最大团的计算量变化很大,最大团包含词的数量有较大的影响,网络中的词越多最大团中包含的词可能也就越多. ADI 数据集中词之间相关性阈值取 0.4,其余 4 个数据集均取 0.5. ADI 中阈值最小,因为该数据集词空间较小,需要加入网络中的词多些. 当词网络固定时,用于修正词权重的最大团的个数 $s$ 的变动会使实验结果剧烈地波动. 随着 $s$ 的增加,检索效果会随之提高. 当 $s$ 增加到一定的值时效果达到最优,若再增大则结果会逐渐降低. 本实验中每个文档集 $s$ 取值大概在(10,15)之间. 由于许多最大团中包含较多相同的词,因此真正用于修正的词一般在 30 个左右. ADI 中 $s$ 的值最大,这是由于 ADI 的文档数较少,词空间较小,需要加入更多的修正信息. 由此可见,$s$ 的取值对结果的影响遵循一定的规律,它与文档集的规模密切相关. 在实验的调参过程发现,ADI 数据集中用于反馈的词团数目 $s=15$ 时,检索的结果达到最优,在 MED,CRAN,CISI,CACM 数据集中 $s$ 分别为 10,14,12,12 时检索结果达到最优. 因此在实验中 $s$ 取以上几个数值进行实验.

本文最大词团代表的含义就是一个主题,实验数据集词团中包含词的数目越少表达的主题越宽泛,例如最大词团 $A=${排序,树,链表,图},这个词团可能表达的主题是数据结构,最大词团 $B=${插入排序,Hash 排序,冒泡排序,基数排序},这个词团的主题可能表达的是排序,词团 $B$ 比词团 $A$ 表达的主题更加细化,产生这种原因是数据集中词项越多提取的词团表达的主题越细化,数据集词项越少词团表达的主题越宽泛.

### 5.2 文档规模对词团权重参数的影响

当 $s$ 固定时,调整 $\alpha$ 和 $\beta$ 的取值,进一步提高检索精度. 对于不同的 $s$,$\alpha$ 的变化范围不大,而且对于检索效果的影响也较缓慢. 在最优结果下 $\alpha$ 的取值都较小,而 $\beta$ 的变化对检索结果影响较大,由于文档依赖词团对主题更具代表性,因此 $\beta$ 在一定范围内 $\beta$ 越大检索结果越好,但达到一定值后反而不利于检索结果. 同时 $\alpha$ 和 $\beta$ 的取值对结果的影响遵循一定的规律. 它们的取值与文档集规模有一定的关系. 在实验中,当文档集规模较小时,$\beta$ 取较大的数值实验结果才能达到最优,如图 2 和图 3 所示,在 ADI 数据集中 $\beta=0.18$,$\alpha=0.06$ 时,实验结果达到最优,然而在 MED 数据集中,$\beta=0.06$,$\alpha=0.02$ 时,实验结果才能达到最优,这是由于 ADI 数据集中文档数目要远远少于 MED 数据集中的文档数目,因此只有提高 $\alpha$ 和 $\beta$ 的数值才能使得检索结果达到最优.

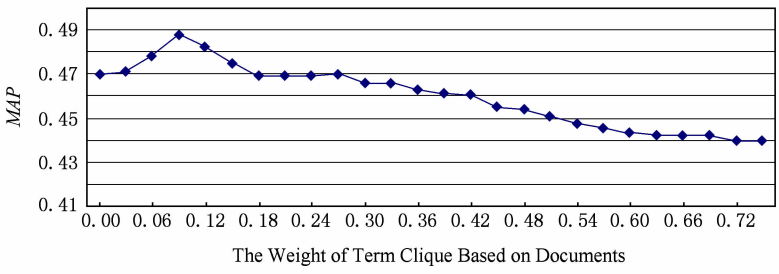


Fig. 2 ADI data set 3-*avg* result.

图 2 ADI 数据集 3-*avg* 结果

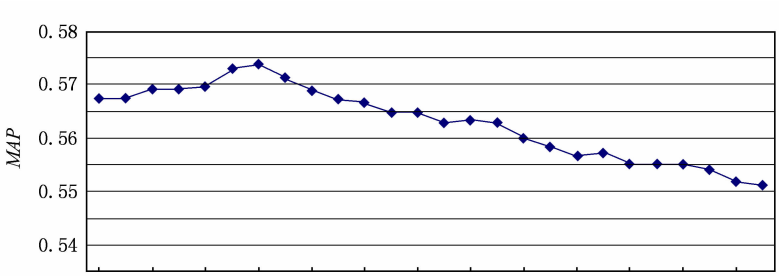


Fig. 3 MED data set 11-*avg* result.

图 3 MED 数据集 11-*avg* 结果

## 6 总结与展望

由于之前的一些信息检索模型都作了独立性假设,即索引词之间是独立的,这样的检索效果均不佳.信息检索可以看成是一种不确定的推理过程,Markov 网络检索模型恰好能有效地表示和推理不确定知识.然而已有对 Markov 网络模型的研究都存在一个这样简单的假设:即一个词和查询中某个查询词的相似性越高,则认为该词语查询越相关,因此被扩展进来的概率越大,但是存在一个这样的问题:很少将文档信息考虑进来.上述假设很容易将不合适的查询词考虑进来,使得检索精度降低,甚至造成主题漂移.

通过对 Markov 网络结构的分析可知,在由词空间和文档空间构成的 Markov 网络中形成了许多完全连通图,即团.构成团的文档对和词项对都相互依赖,即存在语义上较强的关联性,可以认为它们很集中表达同一个概念.因此,本文提出基于文档团的 Markov 网络查询扩展模型,如果一个词团中的词项在同一个文档团中多篇文档出现,通过这种方式将文档团中的信息与词团中的信息进行映射,利用这些信息来将词团分为文档依赖词团和非文档依赖词团,在查询扩展中增加文档依赖的词团的权重.通过实验表明:基于文档团依赖的 Markov 网络信息检索扩展模型的检索效果优于基于团的 Markov 网络信息检索模型的检索效果.

实验结果表明,本文提出的基于文档团的 Markov 网络检索模型的检索性能是很优秀的.但是,本文工作也存在不足之处:1)本文提出的模型只是在小数据集上进行实验,在下一步工作中希望在大型的数据集上进行实验检测该模型的通用性;2)由于查询词之间也存在着依赖关系,这对检索性能进一步提高有所帮助,因此,在以后研究中将查询词空间的词项的依赖关系加入到模型当中,这样可能更有利于查询扩展;3)Markov 网络最大的问题是随着规模的增大,计算量会变得非常大,尤其在搜索引擎的海量文本库中,因此面对海量数据检索时,一方面优化 Markov 网络构造方法,另一方面将文档和词项的相关性计算算法并行化,可以加快计算速度,也是未来发展的趋势.

## 参 考 文 献

[1] Zhai Chengxiang, Lafferty J. Model-based feedback in the language modeling approach to information retrieval [C] // Proc of the 10th Int Conf on Information and Knowledge Management (CIKM01). New York: ACM, 2001: 403-410

[2] Tao Tao, Zhai Chengxiang. Mixture clustering model for pseudo-feedback in information retrieval [C] //Proc of Int Federation of Classification Societies 2004 (IFCS2004). Berlin: Springer, 2004: 541-552

[3] Bai Jing, Song Dawei, Bruza Peter, et al. Query expansion using term relationships in language model for information retrieval [C] //Proc of the 14th ACM Int Conf on Information and Knowledge Management (CIKM 2005). New York: ACM, 2005: 688-695

[4] Soskin N, Kurland N, Domshlak N. Navigating in the dark: Modeling uncertainty in ad-hoc retrieval using multiple relevance models [C] //Pro of the 2nd Int Conf on the Theory of Information Retrieval (ICTIR2009). Berlin: Springer, 2009: 79-91

[5] Tao Tao, Zhai Chengxiang. Regularized estimation of mixture models for robust pseudo-relevance feedback [C] // Proc of the 29th Int ACM SIGIR Conf on Research and Development in Information Retrieval (SIGIR 2006). New York: ACM, 2006: 162-169

[6] Lv Yuanhua, Zhai Chengxiang, Chen Wan. A boosting approach to improving pseudo-relevance feedback [C] //Proc of the 34th Int ACM SIGIR Conf on Research and Development in Information Retrieval (SIGIR 2011). New York, ACM: 2011: 165-175

[7] Lee K, Croft W. A cluster-based resampling method for pseudo-relevance feedback [C] //Proc of the 31st Int ACM SIGIR Conf on Research and Development in Information Retrieval (SIGIR 08). New York: ACM, 2008: 426-436

[8] Xu Yang, Jones G, Wang Bin. Query dependent pseudo-Relevance feedback based on wikipedia [C] //Proc of the 32nd Int ACM SIGIR Conf on Research and Development in Information Retrieval (SIGIR2009). New York: ACM, 2009: 614-624

[9] Zuo Xiaoling, Li Weijian, Liu Yongcai. Discrete Mathematics [M]. Shanghai: Shanghai Science and Technology Publishers House, 1982: 44-45 (in Chinese)
(左孝凌, 李为鑑, 刘永才. 离散数学[M]. 上海: 上海科学文献出版社, 1982: 44-45)

[10] Zuo Jiali, Wang Mingwen, Wang Xi. Extended information retrieval model based on Markov network [J]. Journal of Tsinghua University: Natural Science Edition, 2005, 45 (Suppl): 1847-1852 (in Chinese)
(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(增刊): 1847-1852)

[11] Gan Lixin, Wang Mingwen, Zhang Huawei. Markov network information retrieval modle based on cliques [D]. Nanchang: Jiangxi Normal University, 2007 (in Chinese)
(甘丽新, 王明文, 张华伟. 基于团的 Markov 网络信息检索模型[D]. 南昌: 江西师范大学, 2007)

**Tang Wanning**, born in 1987. Master. His main research interests include information retrieval & data mining.

**Wang Mingwen**, born in 1964. PhD, professor and PhD supervisor. Member of China Computer Federation. His main research interests include information retrieval & Text classification & machine learning.

**Wan Jianyi**, born in 1974. PhD, professor. Her main research interests include big data pricessing (wanjianyi@yahoo.com.cn).