CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 35 卷第 3 期 2013 年 3 月
Vol. 35, No. 3, Mar. 2013

文章编号:1007-130X(2013)03-0150-05

基于改进互信息的信息检索扩展模型*

涂　伟[1],甘丽新[2],黄乐辉[1],谢志华[2]

(1. 江西科技师范大学文科综合实验中心,江西 南昌 330038;
2. 江西科技师范大学光电子与通信重点试验室,江西 南昌 330038)

摘　要:互信息已广泛应用于信息检索扩展模型中。针对互信息存在倾向于低频词、忽略稀疏数据可能导致负相关的潜在影响的问题,本文将改进的互信息方法应用于信息检索扩展模型中。在五个标准数据集上的实验结果表明,本文提出的基于改进互信息的信息检索扩展模型比基于传统互信息的查询扩展模型具有更优的检索性能。

关键词:查询扩展;互信息;信息检索

中图分类号:TP391　　**文献标志码**:A

doi:10.3969/j.issn.1007-130X.2013.03.025

Expanded information retrieval model based on improved mutual information

TU Wei[1], GAN Li-xin[2], HUANG Le-hui[1], XIE Zhi-hua[2]

(1. Center of Arts Complex Laboratory, Jiangxi Science and Technology Normal University, Nanchang 330038;
2. Key Laboratory of Optic-Electronic and Communication,
Jiangxi Science and Technology Normal University, Nanchang 330038, China)

Abstract: Mutual Information has been widely applied to many expanded information retrieval models. Aiming at problems in mutual information, for instance, being apt to low-frequency words and ignoring negative potential impact led by sparse data, this paper applies improved mutual information in an expanded information retrieval model. Experimental results on the five normal datasets show that the expanded information retrieval model based on improved mutual information outperforms that based on traditional mutual information.

Key words: query expansion; mutual information; information retrieval

1　引言

随着网络信息量的与日俱增,海量信息在丰富了人们的信息来源的同时,也给人们获取信息造成了困扰,用户在获取自己需要的信息资源时需要花费大量的时间和精力,其原因在于用户提供的查询信息过少、查询表达模糊不清等,使得在查询时出现难以克服的问题,即信息迷向、信息过载和词不匹配而造成信息检索的查全率和查准率都较低。查询扩展技术是改善信息检索中查全率和查准率的关键技术之一,倍受学者的重视和关注,并成为近年来研究的热点[1~4]。查询扩展技术即指实现查询扩展的方法和手段,其核心问题是如何设计和利用扩展词的来源。目前扩展词的来源有三种:一是来自初检中认为相关的文档;二是用某种技术如聚类技术、文本挖掘技术等从文献集或查询日志中找出与原查询相关的词作为扩展词;三是来自某种包含词与词间相关信息的资源,这种资源可以是人工生成的,也可以是利用大规模语料通过统计的方

* 收稿日期:2012-04-25;修回日期:2012-07-10
基金项目:江西省教育厅科技资助项目(GJJ11224,GJJ11225);江西省科技支撑计划资助项目(00029511101228076)
通讯地址:330038 江西省南昌市江西科技师范大学光电子与通信重点实验室
Address: Key Laboratory of Optic-Electronic and Communication, Jiangxi Sciences and Technology Normal University, Nanchang 330038, Jiangxi, P. R. China

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

法自动生成,如 WordNet、HowNet 和维基百科 Wikipedia。

词间的相似度或相关度计算早已成为信息领域中的基本问题,是信息检索中的核心问题。查询扩展是提高检索效率的有效方法。无论哪种查询扩展方法,无论哪个检索模型,查询扩展词的选择都非常关键。查询候选词选择法是将相关检索文献中的所有查询词取出来,依据某种重要性标准排序,然后将该排序列表中的前 n 个词加入到查询扩展中。在已有很多研究中[5,6]均采用该策略,即认为:若一个词与查询中某个查询词的相似性越高,则该词与查询主题越相关,该词则越重要,从而被扩展进来。词的重要性标准经常利用词间的相似度或相关度来度量。在现有研究中,词间相关性的度量方法主要有潜在语义索引、协方差、词的共现性和互信息等[7]。

互信息方法是一种常用的词汇间相关度计算方法,它能比较有效地表达词间的依赖程度。然而,互信息在衡量词汇之间的相关程度时也存在一定的缺陷,即:互信息对数据稀疏引起的不准确非常敏感,并且也忽略了负相关的潜在影响,倾向于选择低频词。针对上述问题,钟茂生[8]提出一种改进的互信息方法用于词间的相关关系量化计算,该方法在计算和量化词汇间语义相关关系更为可行。

因此,本文将这种改进的互信息方法应用于 Markov 信息检索扩展模型中。实验表明:基于改进互信息的信息检索扩展模型的检索效果优于基于传统互信息的查询扩展模型的检索效果。

2　基于改进互信息的信息检索扩展模型

Markov 网络是一种有力的不确定性推理的图形工具,它能较好地表示知识关联,在信息检索领域得到了广泛应用[5~7,9]。左家莉[5]首次提出并实现了基于 Markov 网络的信息检索扩展模型,该模型通过对文档的学习,利用传统互信息来度量词汇间的相关性,然后按照词间的相关性排序进行候选词选择,加入到查询扩展进行检索。该模型对于检索性能有一定的提高。本文仍然在 Markov 网络基础上,结合文献[8]中提出的改进互信息的方法来进行词汇相关度量从而用于查询扩展。

2.1　总体结构

基于改进互信息的信息检索扩展模型的总体结构如图 1 所示。该模型的基本思想是:

(1)给定一个标准数据集,利用"词间相关性计算模块"中改进的计算方法,对数据集中的词和词进行互信息相关分析,计算出每对词之间的关联程度,即互信息。得出词与词之间的互信息矩阵。

(2)利用"查询预处理模块"对数据集中的查询集进行预处理,得到查询词集。

(3)运用"候选扩展词选择模块"对"互信息矩阵"与"查询词集"中的查询词进行相关程度匹配和比较,按照某种策略进行候选词选择,从而得到每个查询词的候选扩展查询词表。

(4)利用"查询扩展模块"对提交的每一个查询中的查询词进行扩展,从而检索出相关文档,计算出评价指标的值。

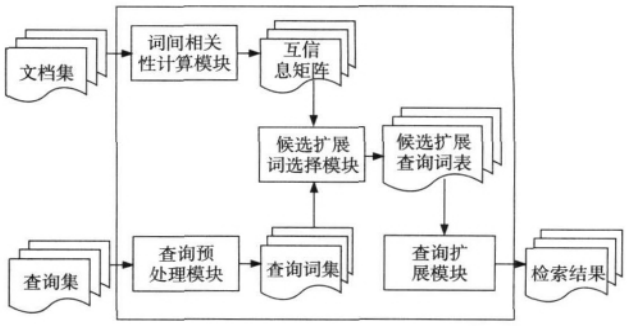

Figure 1　The overall structure of model

图 1 模型总体结构

2.2　模型描述

Markov 网络模型是一种无向图形模型,易于从数据集中获取知识关联。一个 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,$E=\{(x_i,x_j)\mid x_i\neq x_j$ 且 $x_i,x_j\in V\}$,E 中的边表示变量之间的依赖关系。在 Markov 网络中,每个节点 v 条件独立于其邻居节点给定的 v 的非邻居节点的任意节点子集,节点只和其直接相邻节点存在依赖性,即满足 $p(v_i\mid v_j)=p(v_i\mid v_j,(v_i,v_j)\in E)$。为了更好地描述 Markov 网络模型扩展模型,给出它们的形式化描述。

(1)假设 m 为文档集中词的个数,t 表示词。$T=\{t_1,t_2,\cdots,t_m\}$是所有词的集合。

(2)假设 n 为文档集中文档的个数,以 d 表示文档,$D=\{d_1,d_2,\cdots,d_n\}$表示文档集。

(3)文档 d 可以表示为 $d=\{t_1,t_2,\cdots,t_m\}$,其中,$t_1,t_2,\cdots,t_m$ 为 d 索引的词变量。

(4)查询 q 可以表示为 $q=\{t_1,t_2,\cdots,t_m\}$,其中,$t_1,t_2,\cdots,t_m$ 为 q 索引的词变量。

模型分为三层:查询子空间、词项子空间和文

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

档子空间。如图2所示,所有的层构成了一个推理网络。

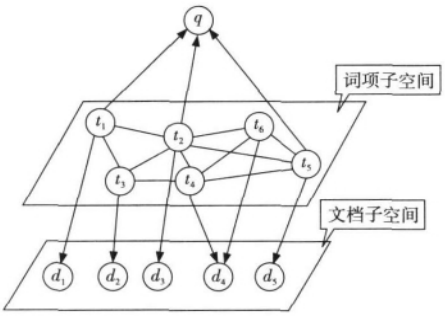

Figure 2 Markov network model

图2 Markov网络模型

2.3 基于改进互信息的词间相关性计算

词间相关性的定量化研究是信息检索中一个重要的基础性工作和核心问题。目前用来度量词与词之间的相关性方法有很多,主要有语义网络、WordNet、Wikipedia、LSI、词的共现性和互信息。互信息已广泛应用于词汇间相关度的计算,它有效地表达了查询词与扩展候选词之间的相关程度[5,10]。词 t_i 与词 t_j 之间的互信息 $MI(t_i,t_j)$ 计算公式如下:

$$MI(t_i,t_j)=\log P(t_i,t_j)/(P(t_i)\times P(t_j)) \quad (1)$$

其中,$P(t_i,t_j)$ 表示词 t_i 与词 t_j 在数据集中同时出现的概率,$P(t_i)$ 表示词 t_i 出现的概率,$P(t_j)$ 表示词 t_j 出现的概率。

然而,互信息在衡量词汇间的相关程度时存在一定的缺陷:互信息对数据稀疏非常敏感,忽略了稀疏数据可能导致互信息为负值的潜在影响,并且倾向于选择低频词。针对上述问题,钟茂生[8]提出了一种改进的互信息方法用于词汇间的相关关系量化计算,其改进的计算公式如下:

$$CMI(t_i,t_j)=\log\left[\frac{P(t_i,t_j)}{P(t_i)\times P(t_j)}\right]/\log\left[\frac{2}{P(t_i)\times P(t_j)}\right] \quad (2)$$

式(2)能避免传统互信息方法中倾向低频词的问题,并且在一定程度上弱化了稀疏数据对相关性的影响。

本文将上述改进的互信息计算公式(2)用于度量词间的相关关系。在互信息计算中,考虑利用词频来计算相应变量的概率(本文以文档为窗口单位),即:

$$P(t_i,t_j)=C(t_i,t_j)/N/,\quad P(t_i)=C(t_i)/N/,\quad P(t_j)=C(t_j)/N \quad (3)$$

其中,$C(t_i,t_j)$ 是指在数据集中,词 t_i 与词 t_j 出现的频率,N 为训练文档集中窗口单元的个数。考虑到词之间的相关强度差异较大,在查询扩展候选词选择中阈值较难设置,同时容易造成候选词列表中词过多或过少的现象,因此本文对互信息值采用如文献[10]方法进行归一化处理,使得 $0\leqslant R(t_i,t_j)\leqslant 1$:

$$R(t_i,t_j)=CMI(t_i,t_j)/\max CMI(t_i,t_j) \quad (4)$$

根据式(3)和式(4),利用"词间相关性计算模块"对数据集中的所有词进行互信息的相关性分析,计算出每对词之间的互信息,从而得到词间的互信息矩阵。

2.4 候选扩展词选择方法

无论哪种查询扩展方法,无论哪个检索模型,查询扩展词的来源是非常关键的。本文通过"词间相关性计算模块"计算出互信息矩阵中词间的互信息之后,根据从查询集提取出的查询词,对每一个查询中的查询词进行候选词扩展。将查询词中的每一个查询词与互信息矩阵中的词进行匹配,然而并不是每一个与原始查询词相匹配的词都有利于检索。因此,设定一个阈值 α,对于每一查询词 q_i,若 $R(q_i,t_j)\geqslant\alpha$,则认为 t_j 与 q_i 相关,则将 t_j 作为候选扩展词加入到 q_i 的候选扩展词列表中 $List(q_i)$。因此,对于每一查询词 q_i,根据词 t_j 与 q_i 的互信息 $R(q_i,t_j)$ 值的大小进行降序排序,得到最终的查询扩展候选词列表 $L(q_i)$。候选扩展词选择算法如下所示:

算法 候选扩展词选择

```
Given a query Q
For each q_i in the Q
  If q_i in the mutual information matrix Then
    if R(q_i,t_j)≥α Then
      sort t_j according to the value R(q_i,t_j) in descend order;
      add t_j into the list of the candidate expand term L(q_i);
    End if
  End if
End for
```

2.5 查询扩展

查询扩展的目的在于将有利于提高检索性能的有用信息加入检索过程中,因此查询扩展词的选取是非常关键的。在查询扩展过程中,本文采用查询词选择法策略,即:给定一个查询,对于一个查询词,若一个查询扩展候选词与该查询词的互信息越大,则认为该词与查询越相关,越有利于检索。由

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

"候选扩展词选择模块"得出每一个查询词都有一个查询扩展候选词列表,该列表按照与其相似度大小来降序排列扩展候选词。因此,本文按照检索的评价指标最优化,对于每一个查询词 q_i,将扩展其候选词列表 $L(q_i)$ 中前 n 个词加入到检索中。查询扩展词与原始查询词重新组成一个新的查询,通过修正原始查询词的权重,重新构造文档和查询之间的相关性。因此给定查询 Q,对任一文档 D_j 和 Q,计算其相关概率。本文采用了 Markov 网络检索扩展模型来计算相关概率:

$$P(D_j \mid Q) \propto \sum_{q_i \in Q}((1-\lambda)P(q_i \mid Q)P(q_i \mid D_j)+ \lambda \sum_{t_k \neq q_i \wedge t_k \in L(q_i)} R(q_i,t_k)P(t_k \mid Q)P(t_k \mid D_j)) \quad (5)$$

其中,$\lambda(0 \leqslant \lambda \leqslant 1)$ 为平滑参数。

在实验中,本文通过同时调整 n 和平滑因子 λ,使模型达到最优检索性能。

3　实验设计与结果

为了评价基于改进互信息的信息检索扩展模型对检索性能的影响,我们采用了未扩展模型(BM)、基于传统互信息的信息检索扩展模型(MIM)和本文提出的基于改进互信息的信息检索扩展模型(IPR_MIM)进行对比。本文选取 adi、med、cran、cisi 和 cacm 这 5 个常用的标准测试文档集进行实验,其中,adi 是信息科学方面的文档集(82 篇文档和 35 个查询);med 是医学方面的文档集(1 033 篇文档和 30 个查询);cran 是航空方面的文档集(1 400 篇文档(含 2 篇空文档)和 225 个查询);cisi 是图书馆科学方面的文档集(1 460 篇文档和 76 个查询);cacm 是计算机科学方面的文档集(3 024 篇文档和 64 个查询)。在相同的数据预处理方式下,比较不同模型的性能差异。在预处理阶段,提取文档中的〈Title〉和〈Body〉部分的内容,去掉了非法字符和数字,大写字母变小写字母,去除停用词,运用 Porter 算法进行词干化处理[5]。

本文采用信息检索系统的一般评价指标 11-avg(在 11 个召回率点(0,0.1,…,1.0)上每一个查询对应精度的平均值)和 3-avg(在 3 个召回率点(0.2,0.5,0.8)上每一个查询对应精度的平均值)。表 1、表 2 分别给出了在 5 个文档集上 3 个检索模型的实验结果。实验以未扩展模型的实验结果为基准,其余两个模型的结果均是在此实验结果上增加的百分比。

Table 1　Experimental results of 11-avg in the standard test set

表 1　在标准测试集上的 11-avg 实验结果

模型	BM(基准)	MIM	IPR_MIM
adi	0.420 2	0.456 1(+8.54%)	0.474 2(+12.85%)
cisi	0.141 8	0.218 9(+54.37%)	0.251 9(+77.64%)
cacm	0.230 0	0.313 6(+36.35%)	0.361 8(+57.30%)
med	0.510 6	0.546 0(+6.93%)	0.548 7(+7.46%)
cran	0.425 5	0.455 7(+7.10%)	0.455 8(+7.12%)

Table 2　Experimental results of 3-avg in the standard test set

表 2　在标准测试集上的 3-avg 实验结果

模型	BM(基准)	MIM	IPR_MIM
adi	0.420 2	0.454 2(+8.09%)	0.499 0(+18.75%)
cisi	0.178 7	0.205 0(+14.72%)	0.233 3(+30.55%)
cacm	0.229 5	0.301 3(+31.26%)	0.340 5(+48.37%)
med	0.510 6	0.553 1(+8.32%)	0.566 8(+11.01%)
cran	0.425 5	0.443 5(+4.23%)	0.456 2(+7.22%)

从表 1 和表 2 可以看出,在所有数据集中,MIM 和 IPR_MIM 整体表现比 BM 的检索效果明显都要好,这说明了扩展后的查询都优于未扩展的原始查询。通过与基于传统互信息的信息检索扩展模型相比较,本文提出的基于改进互信息的信息检索扩展模型(IPR_MIM)总体上表现最好,特别是在 adi、cisi 和 cacm 上检索性能均有较大的提高,说明基于改进互信息的信息检索模型更加有利于提取与原始查询词更相关的查询扩展词,从而在扩展过程中加入更多有用的信息。

在整个实验过程中,有 3 个重要参数需要调整:词间相关性阈值 α、查询扩展词的个数 n 和查询扩展平滑因子 λ。以下分别通过实验说明如何选择上述参数使得最终检索结果达到最优。

为了减少噪音信息和减少查询扩展时的计算量,必须选定合适的阈值。对于 5 个实验数据集,α 取 0.6 比较合适,保证了每个查询词至少有 20 个相关候选词,有利于后面的查询扩展。在查询扩展过程中,对于每一查询词,其查询扩展词的个数 n 应该合理控制,若 n 太小,则在查询扩展中加入的有用扩展信息过少,不利于检索效率的提高;若 n 太大,则容易带来过多的噪音信息,产生主题漂移,而且增加了检索计算量。在实验中,随着 n 的调整,同时调整查询扩展平滑参数 λ。我们采用从 0 到 1,步调为 0.01 来逐步调节 λ。当评价指标 11-avg 和 3-avg 达到最优时,n 和 λ 的取值如表 3 所示。

从表 3 中可知,对于同一个数据集,当 11-avg

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

和 3-avg 达到最优时,n 的各自取值几乎都相同。从 n 的取值可知,真正用于修正的查询扩展词一般在 20 个左右。在数据集 adi 中 n 的值最大,其原因在于 adi 的文档数较少,提取的词信息较少,因此需要加入更多的修正信息。由此可见,n 的取值与文档集的规模密切相关。

Table 3 Value of n and λ based on the optimization of 11-av and 3-avg

表 3 11-avg 和 3-avg 达到最优时 n 和 λ 的值

数据集	评价指标	n 值	λ 值
adi	11-avg	24	0.052
	3-avg	24	0.051
cisi	11-avg	18	0.056
	3-avg	17	0.052
cacm	11-avg	17	0.120
	3-avg	15	0.120
med	11-avg	12	0.056
	3-avg	12	0.063
cran	11-avg	25	0.037
	3-avg	25	0.037

当 n 固定时,通过调整 λ 的取值从而进一步提高检索精度。对于不同的 n,λ 的变化范围不大,而且对于检索效果的影响也比较缓慢。从实验结果可知,最优结果下 λ 取值都比较小,5 个文档集上的 λ 值均不超过 0.12。λ 的取值对于检索结果的影响遵循一定的规律:即与文档集规模有一定的关系。从 3 表中可以看到,在文档集 cacm 上 λ 的值最大,这和 cacm 的文档结构有关,cacm 文档的平均长度比较短,包含的信息量比较少,通过对词权重的修正使得在检索过程中得到很多有益的信息。

4 结束语

本文将改进的互信息方法应用于 Markov 网络信息检索扩展模型中,克服了传统互信息方法中倾向低频词的问题,并且在一定程度上弱化了由于稀疏数据对相关性的影响。通过实验证明了基于改进互信息的信息检索扩展模型在检索性能上优于基于传统互信息的信息检索扩展模型。未来的研究工作主要包括:(1)使用更多的模型在更大的文档集上测试其通用性。(2)为了弥补未登录词由于在单一数据集上无法获取相关语义信息而造成无法查询扩展的缺陷,今后打算考虑利用 Wikipedia 获取更多查询扩展词的信息。

参考文献:

[1] CC. Common criteria for information technology security evaluation,part1:introduction and general model[R]. Version 2.1,August 1999,CCI2 MB2.

[2] Dai Jiahong. Fuzzy cluster-based query expansion[D]. Taiwan:National Sun Yat_sen University,2004.

[3] Fonseca B M,Golgher P B,Moura E S de,et al. Discovering search engine related query using association rules[J]. Journal of Web Engineering,2004,2(4):215-227.

[4] Chen Rui,Zhang Lei, Hu Yan-hua. Model based on semantic information retrieval[J]. Computer Engineering and Applications, 2009,45(26):141-143. (in Chinese)

[5] Zuo Jia-li. Information retrieval model based on markov network [D]. Nan Chang:Jiangxi Normal University,2005. (in Chinese)

[6] Sheng Jun. The research on latent semantic markov network retrieval model [D]. Nan Chang:Jiangxi Normal University, 2006. (in Chinese)

[7] Gan Li-xin. The information retrieval model based on markov concept [D]. Nan Chang: Jiangxi Normal University, 2007. (in Chinese)

[8] Zhong Mao-sheng, Liu Hui,Liu Lei. Method of semantic relevance relation measurement between words[J]. Journal of Chinese Information Processing, 2009, 2(23):115-122. (in Chinese)

[9] Shi Song. Extended information retrieval model based on markov cliques[D]. Nan Chang:Jiangxi Normal University, 2011. (in Chinese)

[10] Cao Ying,Wang Ming-wen Tao Hong-liang. Information retrieval model based on Markov Network [J]. Journal of Shandong University (Natural Science), 2006,3(41),126-130. (in Chinese)

附中文参考文献:

[4] 陈锐,张蕾,胡艳华.基于语义的信息检索模型[J].计算机工程与应用,2009,45(26):141-143.

[5] 左家莉.基于 Markov 网络的信息检索模型[D].南昌:江西师范大学, 2005.

[6] 盛俊.潜在语义的 Markov 网络检索模型的研究[D].南昌:江西师范大学, 2006.

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

[8] 钟茂生,刘慧,刘磊.词汇间语义相关关系量化计算方法[J].中文信息学报,2009,(223):115-122.

[9] 石松.基于 Markov 团的信息检索扩展模型[D].南昌:江西师范大学, 2011.

[10] 曹瑛,王明文,陶红亮.基于 Markov 网络的检索模型[J].山东大学学报(理学版),2006,3(41):126-130.

作者简介:

涂伟(1978-),男,江西安义人,硕士,讲师,CCF 会员(E200027687M),研究方向为信息检索和云计算。**E-mail**:ncsytuwei@gmail.com

TU Wei, born in 1978, MS, lecturer, CCF member(E200027687M), his research interests include information retrieval and cloud computing.

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net