基于隐马尔可夫模型的查询扩展方法

矫 健 张仰森
(北京信息科技大学智能信息处理研究所 北京 100192)

摘 要 对查询进行扩展的目的是找出查询中的潜在语义,确定用户意图,进而构造更适合于搜索引擎检索的查询语句,以提高检索的准确率。提出利用隐马尔可夫模型预测查询中的潜在语义的方法,该模型在大规模用户查询日志上进行训练。由该模型预测出的扩展语句查询的准确率较词共现扩展、同义词扩展等方案均有明显提升。

关键词 隐马尔可夫模型,扩展查询,查询日志
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.12.036

Query Expansion Method Based on Hidden Markov Model

JIAO Jian ZHANG Yang-sen
(Institute of Intelligent Information Processing, Beijing Information Science & Technology University, Beijing 100192, China)

Abstract Automatic query expansion has been a main technique to improve retrieval performance by identifying the potential intentions of the users. In this paper, a method to identify the potential intentions of the users based on hidden Markov models was proposed. The model is trained with large amount of query logs provided by Sogou laboratory. Experiments show that the proposed method has significant improvements in retrieval accuracy for query expansion than other methods.

Keywords HMM, Query expansion, Query logs

1 引言

准确率是搜索引擎赖以发展的主要问题,导致搜索引擎准确率低的因素有很多,可以概括为两个方面:一个是用户自身不能对所需要的信息进行有效描述并利用恰当的关键词进行检索;另一个是搜索引擎难以通过检索词准确地获取用户意图并返回与之最相关的信息。针对这些问题,研究人员试图通过用户反馈的方式分析用户的查询特征,为用户和搜索引擎之间建立一座桥梁。由于用户查询日志记录了用户与搜索引擎之间的交互信息,因此针对查询日志的分析和挖掘成为当前研究的一个热点。这些研究分析了查询词的长度、频次,查询的时间分布,查询修改等用户行为特征,但是将用户查询与用户点击的页面做关联分析的研究却很少。本文将对两者进行深入的分析,找出用户查询词与用户点击的页面间的潜在联系,并通过这种联系进行扩展查询。

扩展查询作为搜索引擎提升性能一种方式,其主要思路是给定用户查询 $Q=\{t_1,t_2,\cdots,t_k\}$,将查询中的一个或多个关键词 t_i 进行替换或者扩充,以此来增大检索范围,提高检索的召回率和准确率。例如,当用户输入"西红柿炒鸡蛋"进行查询时,可以将该查询扩展为"西红柿炒鸡蛋 or 番茄炒鸡蛋 or 番茄炒鸡蛋做法"。从上例可以看出,扩展查询需要解决的问题包括:概念的表述问题("西红柿"和"番茄"指向同一事物,但表述方式不一样,也有可能是同一词汇的多义问题,例如"苹果"可以是一种手机、一种水果、一部电影等)以及用户潜在意图的挖掘问题(用户输入"西红柿炒鸡蛋"是希望知道如何烹制这道菜)。本文将两个问题统一成预测用户潜在意图的问题,并提出了利用隐马尔可夫模型预测用户意图并扩展查询的解决方案。

2 相关工作

自从20世纪60年代第一个检索系统问世以来,研究者就试图利用扩展词表扩展查询来改进查询的准确率和召回率。目前来看,扩展查询的方法主要有两类,即基于规则和基于统计。基于规则的方法指对于特定的查询模式制定相应的查询规则执行查询。例如,当用户输入的是一个地名,则优先返回该地点的地理位置及交通路线等信息,又如,当用户输入"天气"时,首先利用IP地址等信息获取用户位置,然后查询该位置的天气信息返回,而当用户输入的是"地点+天气"时,则查询用户输入地点的天气。利用规则进行扩展查询的优点是直观、准确而且高效。缺点也很明显,即查询规则需要人工制定,成本过高且适用范围较窄。

基于统计的扩展查询试图利用自动构造的扩展词表来对查询中的词进行扩展,根据对扩展词表的不同理解,又可以将扩展词分成同义词、搭配词、关联词等。这些研究的重点多集

到稿日期:2014-01-26 返修日期:2014-04-20 本文受国家自然科学基金(61070119,61370139),北京市属高等学校创新团队建设与教师职业发展计划项目(IDHT20130519),北京市教委专项(PXM2013_014224_000042,PXM2014_014224_000067)资助。
矫 健(1987-),男,硕士生,主要研究方向为中文信息处理,E-mail:949967253@163.com;张仰森(1962-),男,教授,主要研究方向为中文信息处理、人工智能,E-mail:zys@bistu.edu.cn。

中于扩展词表的自动构造以及扩展后的概念相关度计算。根据构造扩展词表所用的资源分类，这些研究主要分为两个方向，一个是从现有的词典、百科、文档等知识库中挖掘扩展词：文献[1-5]根据词典中词与词之间相互解释等特点来构造词向量，并利用向量空间模型计算相似度，从词典中获取同义词表的方法。文献[6,7]介绍了从文本或者百科词典中获取同义词表的方法，这些方法利用模式匹配的方式来收集可能的同义词的集合，然后根据词与词之间共现的关联关系构造图，并利用图的分割理论将图划分为若干区域，每个区域代表一个同义词集。文献[8]介绍了一种利用搜索引擎获取同义词表的方法。该方法首先构造一个候选同义词集，然后利用搜索引擎搜索该同义词集中的词，并用逐点互信息方式计算词与词间的相似度，相似度高的被最终确定为同义词。文献[9,10]给出了同义词的一种特殊形式-缩略语的自动提取方法。文献[11,12]给出了结合《知网》进行扩展查询的方法，这种方法首先从《同义词林》等词典中查找出用户查询中的同义词，然后利用《知网》将得到的同义词与用户查询进行语义相似度计算，选取语义相似度高的扩展词进行扩展查询。文献[13,14]介绍了利用本体进行扩展查询的方法，其核心思想是利用本体知识库的层次关系将用户查询中的概念进行上位或下位扩展，例如将“汽车”下位扩展成“四轮汽车”和“三轮汽车”，然后利用语义词典或者本体库自身计算每一个扩展词与原查询的语义相似度，选取相似度高的进行扩展查询。上述研究的进行均需要一个相对完善的语义知识库的支持，而包括《同义词林》、《汉语语义词典》、《知网》在内的众多语义资源都是人工构造的，成本高且更新慢，据此自动构造的同义词表也面临着利用率低、准确率不高等问题。另一个研究方向是利用用户反馈信息进行扩展词表挖掘，其核心思想是从用户选定的相关文档中挖掘出与用户查询关联性强的词作为扩展词。扩展词表的挖掘方式多与前述文献相似，区别在于从用户反馈中获取扩展词更具针对性，从 TREC 会议的研究结果上看，这类研究可以显著提高检索的效果。同时也有研究指出，利用用户反馈得出的扩展词表稳定性较差，严重依赖于用户反馈的效果。通常其需要与利用语义知识库获取的扩展词表结合使用(即先利用语义知识库获取的扩展词表扩展查询，然后从用户反馈的相关文档中进一步挖掘扩展词表的方式)才能取得更好的效果。

前人的研究多将注意力集中于用户查询中的词，即对用户查询中的单个词进行替换、扩展，而较少考虑用户通过查询语句反映出的潜在语义，即利用整个查询语句而不是查询中的某个词实现对用户查询意图的预测。基于这种想法，本文提出了一种基于隐马尔可夫模型的扩展查询方法。该方法首先利用隐马尔可夫模型进行用户意图预测，然后利用预测结果进行扩展查询，以提高扩展查询的准确率。

3 基于隐马尔可夫模型的查询扩展方法

3.1 用户查询分析

用户利用搜索引擎进行查询的过程如图 1 所示，当用户获取到所需要的页面或者认为无法获取到所需页面时会终止查询，否则，用户会继续点击其它页面或者修改查询词重新搜索。搜索引擎返回的相关页面列表的形式为：页面 Title+/n+查询词所在的部分页面内容+/n+页面的 url+其他信息。用户判断某一个页面与其输入的查询是否相关主要依赖于该页面的 Title 和部分页面信息，其中 Title 占主要部分，即用户自行判断其查询语句与搜索引擎返回的 Title 之间的相关度，只有认为某一 Title 与其查询相关时，才会点击相应的链接。为了对用户查询进行深入分析，我们对搜狗实验室提供的 3 个月的用户查询日志进行了统计，该日志包含了用户查询时间、查询词、用户点击的连接等信息，其具体格式为：

[访问时间+\t+用户 ID+\t+[查询词]+\t+用户点击的 url 在返回结果中的排名+空格+用户点击的顺序号+\t+用户点击的 url]

一个具体的例子为：

[08:08:10 \t 3901556559932498 \t [bbc 中文网主页] \t 5 7 \t www. hackarea. com/Article/free/freematerial/20 0607/2364. html]

它表示 ID 号为 3901556559932498 的用户在 08:08:10 输入了查询语句“bbs 中文网主页”进行搜索，该用户点击的 url 在搜索引擎的返回结果中排名第五，且这是该用户针对这一查询点击的第七个页面。

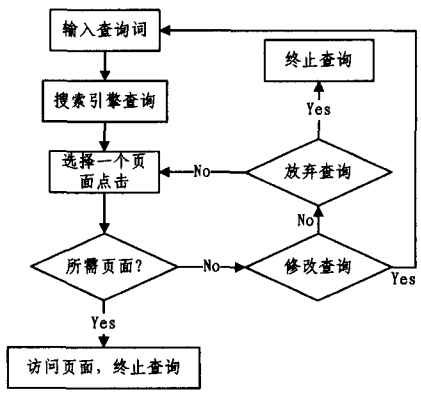

图 1 用户查询流程图

通过对上述查询日志的统计，我们发现：用户查询的平均长度约为 3.167 字，其中查询长度为 3 和 4 的比例占到了总体的 48.1%，这意味着大多数用户仅仅通过一个到两个词进行了查询，而用户首次查询即成功的概率不足 40%，文献[15-17]中的相关统计结论与我们的结果基本一致，这就说明，仅仅利用用户输入的 2 到 3 个查询词无法准确地定位到用户需要的页面，对用户查询进行扩展十分必要。如果将任意一个不同的用户查询以及用户点击的页面的 Title 看成是图中的结点，用户的点击看作是关联两者的边，则可以将上述用户查询日志转化成一个二部图，其形式如图 2 所示。

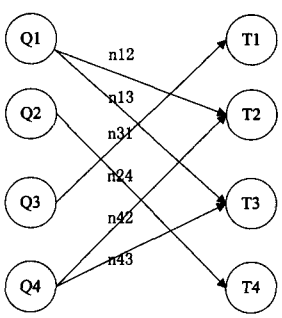

图 2 查询-Title 二部图

图中两点之间有一条边表示用户查询了 Q_i 并且点击了 T_j，边的权值代表点击次数。例如有 100 个用户输入 Q_1 进

行搜索，并且点击了标题为 T_2 的页面，则 n_{12} 的值为 100。由此我们可以得到如下结论：

1. 边的权值越高意味着其所在的查询与 Title 间的相关程度越高；

2. 对于某一个 Title 结点，指向它的查询结点越多说明它表达的含义越模糊；

3. 对于某一个查询结点，它指向的 Tilte 结点越多说明它的指向性越不明确。

我们希望利用图 2 中查询与 Title 间的映射关系将用户查询意图的预测转换成解码问题，即认为用户输入的查询是一个有序的词的序列，而用户点击的页面的 Title 也是一个有序的词的序列。当用户输入某一查询时，利用隐马尔可夫模型将该查询解码成最可能的 Title 形式，由搜索引擎去搜索该 Title，例如，用户输入“人民日报语料”时，将该查询扩展成“语料下载”。然后由搜索引擎去搜索“人民日报语料 or 语料下载”，从而提高查询的准确率和召回率。

3.2 数据准备

为了获取模型的训练数据，按照如下步骤进行数据准备：

1. 数据爬取：利用爬虫程序爬取用户查询日志中 url 对应页面的 Title 和 meta 信息。meta 提供了页面的元信息，例如针对搜索引擎和更新频度的描述和关键词等。由于查询日志的日期(2008 年 6 月、2007 年 3 月、2006 年 8 月 3 个月共 9.5G)距现在较远，一部分链接已经失效或者被重定向，因此我们只获取了保持原 url 的页面信息。

2. 去除噪声：实验初期，我们尝试仅利用查询量大的查询词构造训练语料，令人失望的是实验结果比预期的要差，分析所获得的扩展查询语句，我们发现问题的原因在于搜狗查询日志中包含大量的推荐查询，例如，搜狗推荐的查询关键字“北约军演”，用户只是感兴趣才点击，并没有明确的查询意图，所有与之相关的内容都可能被用户点击，造成严重的查询意图漂移。鉴于这一部分数据所占的比重较大，查询意图预测结果不如人意，我们在后续的实验中做了去噪声处理。通过分析，我们发现这些查询具有：导向大型新闻网站(搜狐、凤凰网等)、查询数量相对集中(1500-3000 次)等特征，根据这些特征，我们剔除了这部分数据。

3. 训练语料抽取：选取图 2 中边的权值大于 500 的查询-Title 对作为训练数据集。

4. 对查询和 Title 进行分词、去除停用词等操作。分词采用 NLPIR 系统实现，同时加入我们自行收集以及搜狗实验室提供的网络词汇表作为用户词库以加大分词的粒度。

3.3 模型训练

隐马尔可夫模型是一个五元组，可以记为：$\mu=(S,K,A,B,\pi)$，其中 S 是状态的集合，K 是输出符号的集合，π 是初始状态分布概率，A 是状态转移概率，B 是符号发射概率。在本文的系统中，S 是所有可能被预测的词的集合，K 是用户输入的查询中的词的集合。

3.3.1 数据对齐

我们将用户输入的查询 $q(t_1,t_2,\cdots,t_n)$ 看成是观察序列，每一个词 t_i 都是一个观察值，将该查询对应的 $Title(d_1,d_2,\cdots,d_m)$ 中的每一个词 d_j 看成是一个状态。模型需要获取任意一个状态 d_j 被预测成某一个观察值 t_i 的概率。这就要求训练语料中每一个状态与其对应的观察值是一一对应的，而语料中的数据并不是对齐的，例如用户查询为“北大”，其对应的 Title 为“北京大学 Peking University”。为了实现观察值和状态的一一对应，我们采用扩展的基于最长公共子串的序列比较算法—Needleman/Wunsch 算法来实现两个序列的匹配，其具体方式如下：

定义 $LCS(Q,T)$ 表示字符串 Q 和字符串 T 的最长公共子串的长度，并且有如下公式：

$$LCS(i,j)=LCS(t_1,t_2,\cdots,t_i,d_1,d_2,\cdots,d_j)\ 0\leqslant i\leqslant n,0\leqslant j\leqslant m \tag{1}$$

$$f(t_i=d_j)\ \text{then}\ LCS(i,j)=LCS(i-1,j-1)+1 \tag{2}$$

$$\text{if}\ (t_i\neq d_j)\ \text{then}\ LCS(i,j)=\text{Max}(LCS(i-1,j-1),LCS(i-1,j),LCS(i,j-1)) \tag{3}$$

利用上述公式，将所有的 $LCS(i,j)$ 计算完毕后，可以得到一个最长公共子串长度的矩阵，利用如下步骤对矩阵进行回溯，得到两个序列之间的一个匹配。

1. 将回溯起点定位在矩阵的右下角。

2. 回溯单元格，至矩阵的左上角：若 $t_i=d_j$，则回溯到左上角单元格；若 $t_i\neq d_j$，则回溯到左上角、上边、左边中值最大的单元格，若有相同最大值的单元格，则按照左上角、上边、左边的顺序回溯。若当前单元格是在矩阵的第一行，则回溯至左边的单元格；若当前单元格是在矩阵的第一列，则回溯至上边的单元格。

3. 根据回溯路径，写出匹配字串：若回溯到左上角单元格，则将 t_i 添加到匹配字串 Q，将 d_j 添加到匹配字串 T；若回溯到上边单元格，则将 t_i 添加到匹配字串 Q，将 NV 添加到匹配字串 T；若回溯到左边单元格，则将 NV 添加到匹配字串 Q，将 d_j 添加到匹配字串 T。

这里，我们引入了空值的概念，记为 NV。它代表观察序列或者状态序列中的隐含语义。上例中的查询和 Title 将最终被对应成：“北大”-“北京大学”，“NV”- “Peking University”。

3.3.2 参数估计

将用户查询和其对应的 Title 对齐后，就可以利用有指导的学习方法对模型的参数进行估计。这里，我们利用极大似然估计的方式。

$$\overline{\pi_i}=\frac{s_i\ \text{作为第一个预测词出现的次数}}{Title\ \text{的总数}}\times 100\% \tag{4}$$

该式表示的是预测词 s_i 在预测序列中第一个出现的概率。

$$\overline{a_{ij}}=\frac{Title\ \text{中从预测词}\ s_i\ \text{到预测词}\ s_j\ \text{的次数}}{Title\ \text{中从预测词}\ s_i\ \text{到其他预测词(包括}\ s_i\ \text{和}\ NV\text{)的次数}}\times 100\% \tag{5}$$

该式表示从预测词 s_i 到预测词 s_j 之间的转移概率。

$$\overline{b_j}(i)=\frac{Title\ \text{中预测词}\ s_j\ \text{对应的查询词是}\ K_i\ \text{的次数}}{Title\ \text{中预测词}\ s_j\ \text{出现的次数}}\times 100\% \tag{6}$$

该式表示预测词 s_j 预测的查询词是 K_i 的概率。

3.4 查询扩展

对于用户输入的关键字，首先进行分词，去停用词等预处理，得到查询序列 $q(t_1,t_2,\cdots,t_n)$，然后利用 Viterbi 算法进行

扩展，具体过程如下：

1. 设置计数量 $\delta_t(i)$，它表示在确定了前 $t-1$ 个查询词的预测词后，将查询词 K_t 预测成 s_i 能取到的最大概率。

2. 初始化 $\delta_1(i)=\max[\pi_i \cdot b_i(K_1)]$ (7)

3. 利用如下递推关系进行动态规划搜索，确定最优的预测词序列：

$$\delta_{t+1}(i)=\max[\delta_t(j) \cdot a_{ji}] \cdot b_i(K_{t+1}) \quad (8)$$

4. 记录在搜索过程中确定的预测词序列，该序列即是扩展词序列。

5. 如果在搜索过程中出现多个预测词对应的 δ_t 具有相同值的情况，则分别沿不同路径进行搜索，得到多个预测词序列。

由于在对齐过程中引入了空值(NV)的概念，在进行预测前同样需要对查询词序列进行空值的添加，否则，模型将退化为仅利用同义词表进行同义替换的扩展查询方式，达不到潜在语义预测的目的。为此，我们需要确定一种合理的空值添加方案来将一个或多个空值添加到查询词序列的适当位置。这里我们根据词在查询中的位置进行分类：查询中的第一个词称为查询的首词，最后一个词称为查询的尾词，其他的词称为中位词。如果查询中只有一个词，那么它既是首词又是尾词。例如查询“北京大学 研究生部 官方网站”，“北京大学”称为查询的首词，“官方网站”称为查询的尾词，而“研究生部”称为查询的中位词。确定了词在查询中的位信息后就可以进行空值的添加，具体方式为：首先，统计每一个词作为首词、尾词和中位词的频率，然后根据统计结果设定阈值，当某一个词主要做首词时，则只在该词后添加空值；如果主要做尾词，则只在其前面添加空值；如果主要做中位词，则查看其与前后词的接续强度，接续强度低于某一阈值时添加空值。

4 实验及结果分析

4.1 查询预测实验

根据 3.1 节中对用户查询日志的统计，用户输入查询的平均词数是 2.27，而用户点击的 Title 的平均词数是 5.41，这就需要对用户查询添加 2 到 3 个空值。由于用户输入的查询词数量多为 1 到 3 个，而当查询词数量超过 3 个时，可以认为用户输入的查询意图已经比较明确，则不需要添加空值，直接利用隐马尔可夫模型进行预测即可。我们分别设计了查询词数量为 1 到 3 个的空值添加方案，利用人工统计的方式随机检测了 100 个查询的预测的准确率。

对于单词查询，添加位置有两个，利用前面介绍的方法，阈值设为 0.5～0.9，实验结果如图 3 所示。

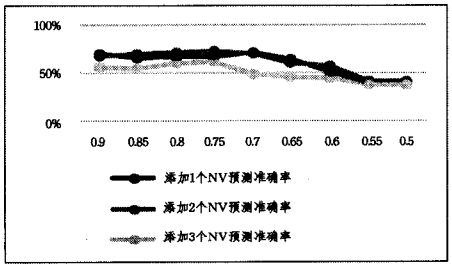

图 3 单词查询预测准确率

从图 3 可以看出，对于单词查询，添加 1 个和 2 个空值的预测准确率要明显高于添加 3 个空值，这主要是由于模型中空值的不确定性过大，因此当多个空值连续出现时，预测过程中误差较大；而如果只添加 1 个空值，更多的时候由于 2 个词的表意依然不够明确导致效果不如 2 个空值。通过这个实验，我们得出的结论是：添加空值的时候，尽量避免连续的空值出现，而且数量以 2 个为宜。

接下来，我们实验了查询词数量为 2 和 3 的情况，对于二词查询，我们拟添加 2 到 3 个空值。其具体规则为：如果查询的首词阈值超过 0.75，则在查询的开头添加 1 个空值；如果尾词阈值超过 0.75，则在查询结尾添加 1 个空值。计算查询中 2 个词之间的接续强度，即当 2 个词的互信息低于 10 时，则在查询 2 个词中间添加 1 个空值，查询的前后各添加 1 个空值。添加完成后保证总词数不超过 5，如果多于 5，则按照连续空值不超过 2 个的规则删除多余的空值。对于三词查询，我们拟添加 1 到 2 个空值，具体方式与二词查询类似，这里不再详述。部分二词查询预测实验结果如表 1 所列。

表 1 部分二词查询预测结果

查询词	添加空值结果	预测结果
梦幻西游 私服	梦幻西游 NV 私服 NV	梦幻西游 在线 私服 百度知道
欧洲杯 直播	NV 欧洲杯 NV 直播	2008 欧洲杯 视频 直播
条件概率 公式	条件概率 NV 公式 NV	条件概率 NV 百度 百科
玉石 鉴定	NV 玉石 鉴定 NV	NV 玉器 鉴别 方法
高考 答案	NV 高考 NV 答案	2008 高考 真题 答案

表 1 选择了比较有代表性的几组数据进行展示，从该结果可以得出如下结论：利用隐马尔可夫模型进行用户意图预测的方法具有很强的时效性，用于预测的语料必须及时更新。由于我们利用的是 2008 年的查询日志进行模型的训练，因此预测结果均为 2008 年的相关信息。另一方面，在实际操作过程中空值添加的位置和数量直接影响了预测效果，而纯粹利用规则的方式添加空值并不能很好地反映数据的特征，因此如何对用户查询日志进行深层次的挖掘以获取查询词之间的关联关系的问题还有待于进一步的研究。

4.2 扩展查询实验

利用隐马尔可夫模型进行查询意图预测的目的是进行扩展查询，为了验证扩展查询的效果，我们设计了 4 组对比实验：直接查询、词共现扩展查询、同义词表扩展查询和利用隐马尔可夫模型扩展查询。实验数据为从搜狗查询日志中随机抽取的 10 条查询量超过 2000 的查询和 10 条查询量低于 20 的查询。对于每一条查询，分别利用不同的扩展方式在搜狗搜索引擎上进行检索，从返回结果中选择前 50 个非重复文档，并人工标注出其中的相关文档。然后统计不同扩展方式的准确率。其中，共现词表利用 PMI-IR 算法从搜狗查询日志中获取，该算法在我们的另一项研究中有详细介绍[18]，此处不再详述。同义词表从《知网》同义词集和哈工大《同义词林》提取。

实验结果如图 4 所示：该结果记录的是 10 条查询准确率的平均值，折线表示的是查询量低于 20 次的查询，柱状图表示的是查询量高于 2000 次的查询。从实验结果可以看出，对于高频查询，进行同义扩展的效果并不理想，这是因为像“百度”、“搜狐”这样的高频查询词具有明确的指向性，没有同义词进行扩展。而对于低频查询(如“北大”)，属于部分用户查

(下转第 188 页)

tio in fuzzy rough set theory with application to tumor classification[J]. Applied Soft Computing,2013,13(1):211-221
[25] 徐菲菲,魏莱,杜海洲,等. 一种基于互信息的模糊粗糙集分类特征基因快速选取方法[J]. 计算机科学,2013,40(7):216-235
[26] 曾安平,李天瑞,罗川. 高斯核模糊粗糙集中对象集变化时近似集增量更新方法研究[J]. 计算机科学,2013,40(7):173-177
[27] Klement E P, Mesiar R, Pap E. Triangular norms[M]. Kluwer Academic Publishers,2001
[28] Yeung D S,Chen D G,Tsang E C C,et al. On the generalization of fuzzy rough sets[J]. IEEE Transactions on Fuzzy Systems, 2005,13(3):343-361
[29] Fung L W, Fu K S. An axiomatic approach to relational decision-making in a fuzzy environment[C]// Zadeh L A,Fu K S,Tanaka K,et al. , eds. Fuzzy Sets and Decision Processes. New York: Academic Press,1975:227-256
[30] Qian Y H,Wu W Z,Dang C Y. Information granularity in fuzzy binary GrC model[J]. IEEE Transactions on Fuzzy Systems, 2011,19(2):253-264

(上接第 171 页)

询时的缩略用法,这样的词进行同义扩展的空间较大,因此性能提升较大。共现词扩展则正好相反,与高频查询词共现的词与查询词具有很强的关联性,扩展查询效果较好;而与低频查询词共现的词与查询词不具备强关联性,利用其进行扩展反而引起语义的迁移。由于隐马尔可夫模型在进行用户意图预测的过程中,结合了同义词扩展和共现词扩展的优点,因此在两种环境中查询的性能均有一定提升。同时,利用隐马尔可夫模型扩展也具有共现词扩展的不足,即在训练语料量较少的情况下,性能提升有限,部分查询扩展查询后准确率甚至不如未扩展的准确率。

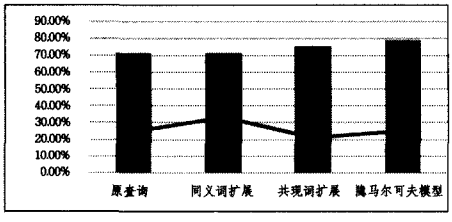

图 4 扩展查询准确率

需要特别指出的是,利用隐马尔可夫模型进行扩展查询的优势在于其意图预测。例如在我们的实验中,用户输入的"清华"被扩展成了"清华大学 官网"。这一扩展查询代表了大多数用户的查询意图,其返回的清华大学官方网站链接排在搜索引擎返回页面的首位,而其他的返回页面由于受到"官网"这个扩展词的限制,相关度急剧下降,这使得我们的实验结果在一定程度上有所损失,但并不影响将最相关的页面返回给用户的初衷。

结束语 扩展查询是提高搜索引擎性能的重要手段之一。本文给出了一种利用隐马尔可夫模型进行扩展查询的方法,该方法综合了基于同义词表扩展查询和基于共现词表扩展查询的优点,利用大规模用户查询日志进行模型训练,实现了对用户查询中潜在意图的预测。由该方法得到的扩展查询模型在中高频查询词中取得了较好的效果,而对于低频查询词则仍需进一步改进。

由于用户输入的查询词数较少,因此本文提出的基于最长公共子串的对齐方式系统开销较少,这让模型在大规模查询日志上进行训练成为可能。

参 考 文 献

[1] Crouch C J. A cluster-based approach to thesaurus construction [C]// Eleventh International ACMSIGIR Conference on Research and Development in Information Retrieval. 1988:309-320
[2] Blondel V D,Senellart P P. Automatic extraction of synonyms in a dictionary [R]. Presented at the Text Mining Workshop,2002
[3] Salton G. The Smart Retrieval System-Experiments in Automatic Document Processing[M]. New Jersey,USA:Prenvtice Hall. Inl,1971
[4] 陈建超,郑启伦,李庆阳,等. 基于特征词关联性的同义词集挖掘算法[J]. 计算机应用研究,2009,26(7):2517-2519
[5] Schutze H,Pedersen J. A co-occurrence-based thesaurus and two applications to information retrieval[J]. Information Processing and Management,1997,33(3):307-318
[6] 吴云芳,石静,金彭. 基于图的同义词集自动获取方法[J]. 计算机研究与发展,2011,48(4):610-616
[7] Matsuo Y,Sakaki T,Uchiyama K,et al. Graph-based clustering using a Web search engine [C]// Proc of EMNLP. 2006:542-550
[8] Turney P D. Mining the Web for synonyms:PMI-IR versus LSA on TOEFL[C]// Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001). Freiburg,Germany, 2001:491-52
[9] 崔世起,刘群,林守勋,等. 中文缩略语自动抽取初探[C]// 孙茂松,陈群秀. 自然语言处理与大规模内容计算. 北京:清华大学出版社,2005:53-58
[10] 谢丽星,孙茂松,佟子健,等. 基于用户查询日志和锚文字的汉语缩略语识别[C]// 孙茂松,陈群秀. 中国计算语言学研究前沿进展. 北京:清华大学出版社,2009:551-556
[11] 田萱,杜小勇,李海华. 语义查询扩展中词语-概念相关度的计算[J]. 软件学报,2008,19(8):2043-2053
[12] 熊桂喜,王开锋. 基于语义的查询扩展研究[J]. 微计算机信息, 2008,24(30):177-178,187
[13] 杨清琳,李陶深,农健. 基于领域本体知识库的语义查询扩展[J]. 计算机工程与设计,2011,32(11):3853-3856
[14] 李海芳,史俊冰,段利国,等. 一种基于含糊同义词的查询扩展方法[J]. 计算机应用与软件,2011,28(12):41-43
[15] 余慧佳,刘奕群,张敏,等. 基于大规模日志分析的网络搜索引擎用户行为研究[J]. 中文信息学报,2007,21(1):109-114
[16] 岑荣伟,刘奕群,张敏,等. 基于日志挖掘的搜索引擎用户行为分析[J]. 中文信息学报,2010,24(3):49-54
[17] 窦志成,袁晓洁,何松柏. 大规模中文搜索日志中查询重复性分析[J]. 计算机工程,2008,34(21):40-41,44
[18] 张泽伟,矫健,张仰森. 基于 PMI-IR 的联想词表构造方法研究[J]. 计算机技术与发展,2014,24(6):140-144