第43卷 第6A期
2016年6月

计 算 机 科 学
Computer Science

Vol. 43 No. 6A
June 2016

基于Copulas框架的混合式查询扩展方法

张书波 张 引 张 斌 孙达明
(东北大学信息科学与工程学院 沈阳 110819)

摘 要 基于语义资料和局部分析的混合式查询扩展可以同时提供具有语义相关性和时效性的扩展结果,但如何有效地混合不同相似度度量指标是尚未解决的问题。提出了一种基于Copulas框架的混合式查询扩展方法,在统一框架内实现了不同类型相似度度量指标的合并。该方法基于语义分析及词语共现分析方法,分别计算扩展词与用户查询词的语义及统计相似概率,进而在Copulas框架下融合扩展词集,选取最高质量的扩展词形成查询扩展。实验结果表明,该方法充分利用了语义及词语共现分析查询扩展方法的优点,有效地弥补了两者的不足,提高了搜索结果的查准率,具有更优的搜索性能。

关键词 信息检索,查询扩展,语义分析,词语共现分析,搜索性能
中图法分类号 TP391 **文献标识码** A

Combined Query Expansion Method Based on Copulas Framework

ZHANG Shu-bo ZHANG Yin ZHANG Bin SUN Da-ming
(School of Information Science & Engineering, Northeastern University, Shenyang 110819, China)

Abstract Hybrid query expansion methods based on semantic and local analysis can provide time-sensitive extension results with semantic correlation. However, how to effectively combine two different kinds of similarity metrics has not been solved. This paper proposd a hybrid query expansion method based on the Copulas framework to implement the combination of different types of similarity metrics. Based on the query expansion methods of semantic and word co-occurrence analysis, the proposed method respectively calculates the semantical and the statistical similar probabilities between expansion-words and the query words submitted by the user. It then selects high quality extension words to obtain the final extension word set. The experimental results show that the method makes full use of the advantages of the two kinds of query expansion methods to improve the precision ratio factor. The method has better search performance.

Keywords Information retrieval, Query expansion, Semantic analysis, Word co-occurrence analysis, Search performance

1 引言

目前,搜索引擎已经成为人们从互联网上获取信息的重要工具。当用户应用搜索引擎检索信息时,如果用户输入的查询词与相关文档中的用词不一致,会出现找不到相关文档的问题,即所谓的词语匹配问题。查询扩展技术[1,2]是解决该问题的有效方法之一,其通过扩展用户查询词提高检索性能。按照扩展词的来源和选取算法不同,查询扩展技术可以分为基于语义分析的方法[3]及基于统计分析的方法[4]两种类型。

基于语义分析的查询扩展方法通过将用户的查询词映射到语义资源(如本体等)中的概念,计算概念之间的相似度,提取相似度大于阈值的概念作为扩展词,使得扩展结果与用户查询词之间具有较强的语义相似性。但由于本体等语义资源的更新并非实时进行,这类方法通常具有时效上的局限性。基于统计分析的方法则通过分析文档[5]或搜索日志[6]中词汇使用的情况来找到与用户查询词在统计意义上最为相关的概念,通常具有较好的时效性,但却忽略了扩展词与查询词之间的语义相关性,检索结果遗漏了部分语义相关的文档。正是由于这两种方法各自存在的优势与不足,目前,许多科技人员开始研究如何混合上述两种方法,以实现同时含有语义相关性与时效性的查询扩展[7-10]。

在以前的混合式查询扩展方法研究中,通常存在一个尚未获得有效解决的问题:基于语义分析和统计方法计算得到的词语语义相似度和词语相关度有着不同的取值范围、分布规律及函数特征,这些方法常用一种简单的权重相加来混合不同的指标。这种简单的混合方法无法真正实现不同类型的度量指标的有效融合,难以取得合理的效果,并且基于经验的权重参数选择方法也极大地限制了这些方法在实际问题中的应用。因此,解决混合式查询扩展方法的相似度指标合并问题是有效应用混合式查询扩展方法必须要解决的关键问题。

针对上述问题,本文提出了一种基于Copulas框架的混合式查询扩展方法。该方法统一以概率方法为基础,研究出扩展词与用户查询词的语义相似概率和词语共现概率计算方

本文受宁夏回族自治区自然科学基金资助项目(NZ13265),中央高校东北大学基本科研专项基金项目(N120804001,N120204003)资助。
张书波(1973—),男,博士生,主要研究方向为信息检索,E-mail:zshubo@163.com;张 引(1985—),男,博士,讲师,主要研究方向为信息检索;张 斌(1964—),男,博士,教授,博士生导师,主要研究方向为信息检索、Web服务;孙达明(1981—),男,博士生,主要研究方向为信息检索。

法，进而在 Copulas 框架下融合两种度量指标，从而选取最高质量的扩展词组成最终扩展词集，实现了在统一框架内的不同类型相似度度量指标的有效合并。实验结果表明，该查询扩展方法可以有效提升搜索引擎的检索性能。

2 相关研究

为充分利用基于语义分析及基于统计分析的查询扩展方法的优点，避免各自的不足，很多研究人员开展了混合式查询扩展方法研究。

文献[7]提出了混合 WordNet 和 KLDLCA 的查询扩展方法。该方法的基本思想是先利用 WordNet 提取扩展词，再用改进的 KLDLCA 公式计算扩展词的权重，选取前 M 个扩展词形成最终扩展词集。

文献[8]利用本体推理与本体中实体相关度得到用户初始查询的候选扩展概念集，同时引入相关参数有效控制候选概念集中扩展词的数量。组合初始查询和候选扩展概念集形成新的用户查询，并利用搜索引擎进行检索，得到初始结果，选择检索结果中前 N 篇文档，利用局部共现分析法计算相关概念与候选扩展词的共现频度。利用筛选函数对候选扩展概念集进行二次筛选，将满足条件的候选扩展词作为最终的扩展词。

文献[9]提出了一种基于语义词典的局部查询扩展方法。该方法利用用户输入的原始查询词集，首次查询获得返回文档结果，利用伪相关反馈方法从中提取与用户查询最相关的扩展词集，再利用《同义词词林扩展版》计算用户原始查询词集和扩展词集的每对词语的语义相似度，删除扩展词集中与用户查询词相似度低的词，获得最终扩展词集。

文献[10]提出了一种基于本体和用户日志的查询扩展方法。该方法先将用户原始查询词匹配到领域本体知识库中的概念，并计算其与领域本体知识库中的同义词、上位词、下位词等的语义相似度，形成初始扩展概念集。为避免加入查询无关词而产生“查询漂移”问题，又结合用户查询日志信息计算出查询词与初始扩展概念集中扩展词间的共现度权值，对初始扩展概念集进行二次筛选，形成最终扩展概念集。

以上基于语义和统计分析的混合式查询扩展方法主要分为两种类型：一类方法仅是在获得初始扩展词集后使用另一种方法进行筛选，该类方法会缺失一部分重要的扩展词，影响搜索性能；另一类方法在合并两组扩展词集时仅将两词集的相似度值进行了简单的累加或按权重进行合计，无法合理地解释合并后的结果，极大限制了这些方法在实际问题中的应用。针对这些不足，本文提出了基于 Copulas 框架的混合式查询扩展方法，应用统一的框架实现了对语义及统计分析相似度度量指标的合并，实现了有效的混合式查询扩展。

3 基于 Copulas 框架的混合式查询扩展方法

为充分利用语义分析和统计分析查询扩展方法的优点，避免两者各自的不足，提高搜索的查询性能，同时避免以往混合式查询扩展方法简单合并不同类型度量指标的低效融合和基于经验选择权重参数的繁杂问题，本文研究出了词语间语义相似的概率计算和词语间统计概率计算方法，并基于 Copulas 框架实现两扩展词集的有效融合，选取了最高质量的扩展词组成最终扩展词集，有效地提升了检索性能。

3.1 基本思想

本文方法的基本思想是先对用户输入的初始查询进行分词、去停用词等预处理。接着，基于语义分析方法，利用本体中概念之间的关系生成语义关系图，以概率方法为基础，计算出用户查询词与概念之间的语义相似概率，选取与用户查询词相似的概念作为扩展词，形成语义扩展词集。基于统计分析方法，利用用户查询词进行搜索，获取与用户查询词相关的最前 N 个文档，计算出词语共现概率，生成统计扩展词集。在上述工作的基础上，利用统一的概率框架 Copulas 融合两种度量指标，选取最高质量的前 K 个扩展词形成最终扩展词集，最后组合用户查询词与最终扩展词集生成新用户查询，进行搜索，实现了在统一框架内的不同类型相似度度量指标的有效合并。该方法的总体流程如图 1 所示。

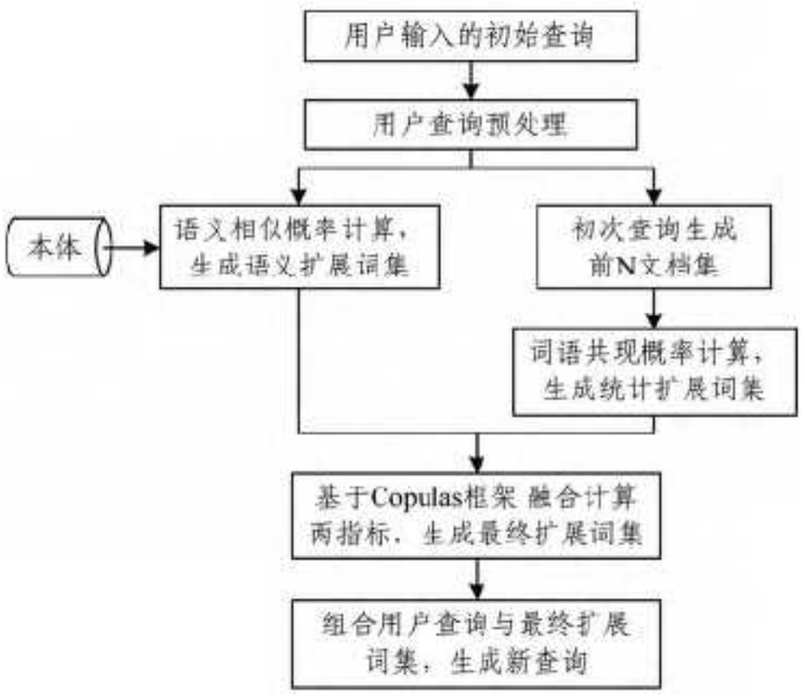

图 1 方法的总体流程图

在上述总体流程中，主要需要解决以下 3 个问题：1）如何基于本体计算词语间语义相似概率，进而生成语义扩展词集；2）如何基于初始检索到的文档集计算词语共现概率，并生成统计扩展词集；3）如何有效融合两扩展词集，并生成用户查询的最终扩展词集。

3.2 词语间语义相似概率计算

在信息检索领域，本体是共享概念模型的明确的形式化规范说明。它明确地、形式化地表达了概念的含义以及概念之间的语义关系，成为提供语义信息的语料库。

经常被用作用户查询语义扩展的本体有知网（HowNet）等。知网作为一个包含词语语义信息及词语间关系的常用词语知识库，对汉语和英语的词使用概念进行语义描述。在知网中一个词可以用几个概念来描述。知网中每一个词语的概念及其描述形成一个记录，每一记录主要包括 5 项内容：NO. 代表词语的顺序编号，W_C 代表词语的中文表达方式，G_C 代表词语词性和拼音标注，E_C 代表词语举例，DEF 代表词语的义原定义。义原是概念的语义最小单元描述，记录了概念之间或概念属性之间的关系。记录中义原加符号代表了词语的语义信息和概念之间的语义关联关系。义原之间主要有上下位关系、属性关系、对义关系和反义关系等，其中最重要的关系是树状层次体系中的上下位关系。

为了充分利用语义资料和局部文档集的优势，弥补语义查询扩展和基于局部分析查询扩展方法的不足，在统一的概

率框架下从语义扩展词集和统计扩展词集中选取最优的扩展词，研究并提出了基于知网的语义相似概率计算方法，用来衡量概念间语义相似度。

利用义原之间的语义关系可以生成语义图，如图 2 所示。

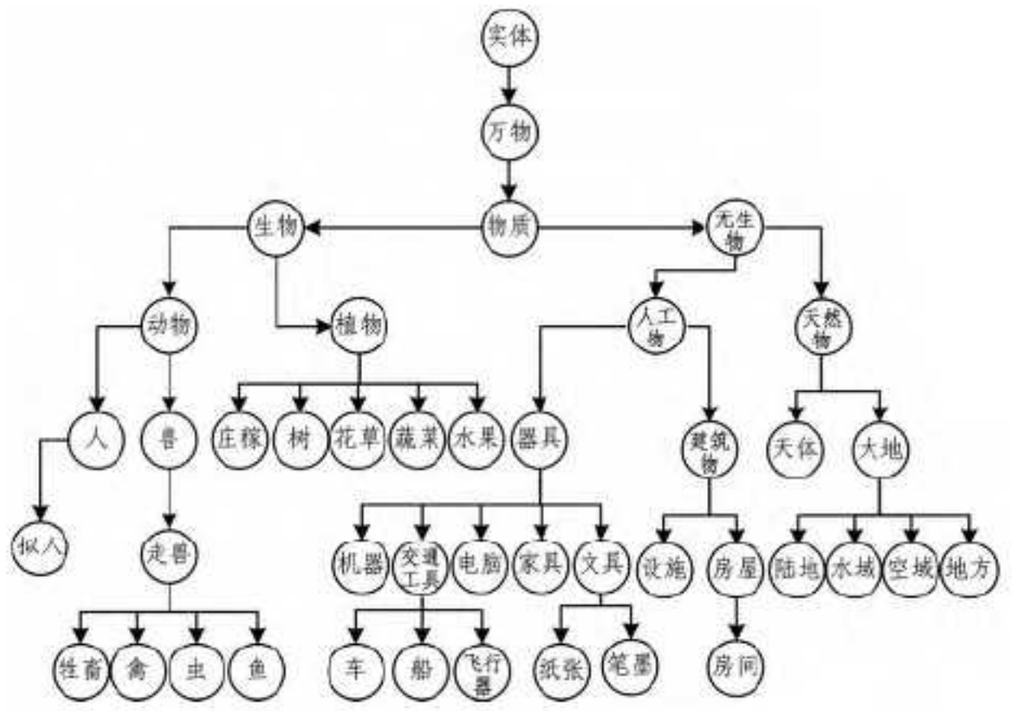

图 2　义原的语义关系图

基于义原语义关系图，可利用个性化 PageRank(Personalized PageRank，PPV) 算法计算概念之间的语义相似概率。在义原图中，用一个偏好向量来表示用户查询词的义原，则对应的 PPV 表示每个概念的义原与用户查询词的义原的相似程度。因此可按照个性化 PageRank 计算概念义原与用户查询义原之间的语义相似概率。

$$r=(1-\alpha)M_r+\alpha V' \tag{1}$$

其中，α 是跳转概率，通常取值为 0.15；V' 为用户的个性化向量，$|V'|=1$，若代表用户偏好的结点有 k 个，则在 V' 中这 k 个结点的值之和为 1，而其他结点的值为 0。式(1)的解就是偏好向量 V' 所对应的个性化 PageRank 向量 $PPV(m)$，它代表了图中每个义原 m_i 与用户查询词义原 q_a 的语义相似概率 $P(m_i|q_a)$，即 $P(m_i|q_a)$ 等价于 $PPV(m)$。

假设定义两个概念 C_1，C_2 的义原集合分别是 $C_1=\{m_{11}, m_{12},\cdots,m_{1n}\}$，$C_2=\{m_{21},m_{22},\cdots,m_{2m}\}$，知网中的概念 c 与用户查询词的概念 q_c 间的语义相似概率计算公式为：

$$P(c|q_c)=P(c|m)\times P(m|q_c) \tag{2}$$

由贝叶斯公式可得：

$$P(c|m)=\frac{P(m|c)\times P(c)}{P(m)} \tag{3}$$

本文认为在知网中 $P(c)$ 和 $P(m)$ 是恒量，则 $P(c|m)$ 的大小与 $P(m|c)$ 大小成正比，因此在式(2)中可以用 $P(m|c)$ 代替 $P(c|m)$。

$$P(c|q_c)=P(m|c)\times P(m|q_c)=\prod_i P(m_i|c)\times P(m_i|q_c) \tag{4}$$

计算可得 q_c 与知网中概念的相似概率集 $S=\{P(c_1|q_c), P(c_2|q_c),P(c_3|q_c),\cdots\}$，选取概率大于零的概念组成语义扩展词集。

3.3　词语间统计相似概率计算

通常，相关的事物常常同时出现，同时出现的事物往往具有相关性。因此，在文档中频繁共现的词对具有统计相关性和相同的主题。词语共现是指不同词语在同篇文档中的一个语境单元范围内共同出现。在大规模语料库的文档中词语共现代表了词语间具有语义关联关系，且词语共现频率越高，说明词语间的语义关联度越强。词语共现分析法是在信息检索领域中运用统计学的方法统计词语同时出现在文档中的频率来挖掘词语之间的相关性的一种方法。该方法被许多学者用来分析大规模文档集中词语之间的关联关系，在搜索引擎中有着广泛的应用，并获得了很高的检索性能。

利用词语在文档集中的共现概率来确定不同词语间的相关性，并选取相关性高的词生成用户原始查询词的扩展词集。

用户查询经分词和去停用词后，形成用户初始查询词集 $Q=\{q_1,q_2,\cdots,q_n\}$。利用用户初始查询词集 Q 进行检索，得到与用户查询相关度最高的前 N 个文档，形成初始文档集 D。文档集 D 中的文档 d 经分词和去停用词后，形成词集 $T=\{t_1,t_2,t_3,\cdots\}$。考虑词汇在文档中的词频的影响，词语 q_i 和 t_j 在文档集 D(d 属于 D 且 d 作为一个语境单元)中的共现概率计算公式为：

$$P(t_j|q_i)=\frac{tf(q_i,t_j)+1}{tf(q_i)+1} \tag{5}$$

其中，$tf(q_i)$ 代表 q_i 在文档集 D 中出现的次数，$tf(q_i,t_j)$ 代表 q_i 和 t_j 在文档集 D 中共同出现的次数。

计算可得 q_i 与文档集 D 中词语的共现概率集 $L=\{P_D(t_1|q_i),P_D(t_2|q_i),P_D(t_3|q_i),\cdots\}$，选取概率大于零的词语组成统计扩展词集。

3.4　基于 Copulas 框架融合两扩展词集

通过计算词语间语义相似概率获得了用户查询词与本体中概念的语义相似程度，通过计算词语间统计概率获得了用户查询词语与局部文档中词语相关的程度。为了从两个扩展词集中按照统一指标选取与用户查询最相关的扩展词，需要统一地描述多维随机向量的累积函数，形成一个唯一的与用户查询相似的衡量指标。而 Copulas 函数正好满足这一要求。Copulas 函数是由 Sklar 首先提出的描述变量间相关性的函数，可用来构造灵活的多元分布函数[11]。Carsten Eickhoff 研究了在许多领域的信息检索中应用 Copulas 解决基于多维质量标准的文档相关性的系统估计必须适应一维的条件的结果排名的问题[12]。

对于一个给定的二维随机向量 $X=(x_1,x_2)$，它具有连续边际：

$$F(x)=P[X_i<x] \tag{6}$$

可映射观察结果到多维数据集：

$$U=(u_1,u_2)=(F(x_1),F(x_2)) \tag{7}$$

则可以用 Copulas 统一描述多维随机向量的累积分布函数：

$$C_{indep}(U)=\exp(-\sum_{i=1}^{2}(-\log(u_i))) \tag{8}$$

利用多维累积分布函数 Copulas 可以把语义扩展词集和统计扩展词集融合成统一的扩展词集，进而选取质量最高的前 K 个扩展词形成最终扩展词集。

对用户查询词 q_i 分别计算出语义相似概率集 S 和词语共现概率集 L。

对知网中与用户查询词 q_i 语义相关的概念 c_j 计算出比 $P(c_j|q_i)$ 值小的概率：

$$F(x_1)=P=\frac{a}{A} \tag{9}$$

其中，a 表示 S 中比 $P(c_j|q_i)$ 值小的概率值的个数，A 表示 S 集中元素的总个数。

同理，对与用户查询词 q_i 共现的词项 t_j 计算出比 $P(t_j|q_i)$ 值小的概率：

$$F(x_2)=P=\frac{b}{B} \tag{10}$$

其中，b 表示 L 中比 $P(t_j|q_i)$ 值小的概率值的个数，B 表示 L 集中元素的总个数。

$$C_{indep}(U)=\exp(\log(\frac{a}{A})+\log(\frac{b}{B})) \tag{11}$$

分别求得同一词项与用户查询词的 $F(x_1)$ 和 $F(x_2)$ 后，按式(11)计算得到 $C_{indep}(U)$ 并将其按降序排列，选取 Copulas 值最高的前 K 个词项作为最终扩展词集。

4 实验与分析

实验的目的是验证采用基于 Copulas 框架的混合式查询扩展方法相比其他同类混合式查询扩展方法的检索性能是否有所提升和改善。

实验数据集采用的本体是知网 HowNet；采用的用户日志是 Sogou 搜索引擎免费提供的搜狗用户日志；采用的语义词典是哈工大信息检索研究室同义词词林扩展版，其包含 77343 条词语；采用的中文分词工具是中科院计算所开发的 ICTCLAS；以百度作为检索平台，以用户初次查询检索获得的前 50 篇文档作为词语共现分析的文档集。

实验采用的主要评测指标是查准率(Precision)和前 N 篇文档的准确率 $Prec@N(Prec@20)$。

查准率是衡量某一搜索系统的信号噪声比的一种指标，即搜出的相关文档数与搜出的全部文档数的百分比。计算公式为：

$$Precision=\frac{|Retrieved\cap Relevant|}{|Retrieved|}\times 100\% \tag{12}$$

其中，$|Retrieved|$ 代表一次检索后所检索出的全部文档的数量，$|Retrieved\cap Relevant|$ 代表一次检索后所检索出的相关文档的数量。

$Prec@N$ 是指对特定的用户查询，利用搜索引擎搜索出 N 篇文档时的准确率。计算公式为：

$$Prec@N=\frac{Relevant}{N}\times 100\% \tag{13}$$

其中，$Relevant$ 为单个用户查询结果中前 N 篇文档中的相关文档数，N 代表单个用户查询检索出的结果中的前 N 个文档。$Prec@20$ 通常反映了搜索结果中前两页的准确率。

按照本文分析的问题特点，从未参与本文研究的学生提出的用户查询集合中选取有意义的用户查询作为实验测试用查询(详见表 1)，这些真实的用户查询充分反映了用户在使用搜索引擎时的真实需求。

表 1 实验测试用户查询

序号	用户查询	同义词举例
1	西红柿	番茄
2	土豆	马铃薯
3	卷心菜	大头菜
4	红楼梦	石头记
5	中秋节	八月十五
6	计算机	电脑
7	手机	无线电话
8	货币	钞票
9	魂不附体	魂不守舍
10	创业板	二板市场

在实验中将本文方法(C-Hybrid)与基于语义词典的局部查询扩展方法(LSR)[9]和基于本体和用户日志的查询扩展方法(OLQEM)[10]进行对比分析，以验证其检索性能。查询性能对比结果如图 3 所示，搜索性能 $Prec@20$ 对比结果如表 2 所列。

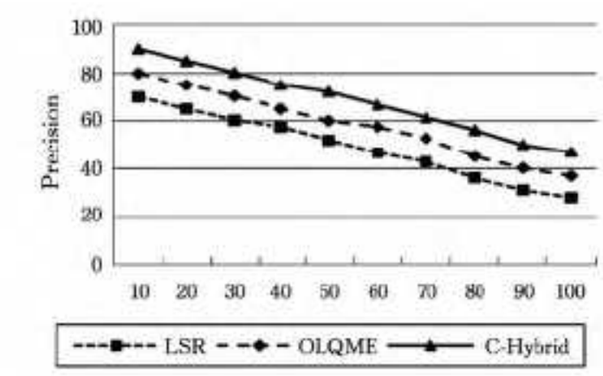

图 3 查询性能对比结果图

表 2 $Prec@20$ 对比结果

查询序号	LSR	OLQEM	C-Hybrid
1	0.65	0.75	0.85
2	0.60	0.70	0.80
3	0.55	0.65	0.75
4	0.55	0.60	0.70
5	0.60	0.65	0.80
6	0.70	0.80	0.85
7	0.65	0.75	0.80
8	0.60	0.70	0.75
9	0.55	0.60	0.70
10	0.60	0.65	0.70

从图 3 和表 2 中可以看出：本文方法比其它两种方法在平均查准率上分别提升了 39.62%，17.27%，用户搜索的平均 $Prec@20$ 分别提升了 27.27 %，12.41%。本文方法比其它两种混合式查询扩展方法的检索性能获得很大提高 这主要是因为基于语义词典的局部查询扩展方法(LSR)先利用用户查询获取 N 个最相关的文档，从其中选取扩展词，接着利用语义词典计算扩展词与用户查询词的语义相似度，去掉语义相似度小于阈值的扩展词，剩下最终扩展词。这种混合式查询扩展方法仅保留了与用户查询共现度和语义相似度高的词，忽略了部分有效的扩展词，具有片面性，如用户输入查询词“土豆”，LSR 生成的最终扩展词集为“马铃薯，土豆，山药蛋，洋芋”，忽略了有效的扩展词“土豆网，杨伟东(土豆总裁)，优酷土豆集团公司”。

基于本体和用户日志的查询扩展方法(OLQEM)先利用用户查询词获取语义相似度大于阈值的扩展词，并从用户日志中选取 N 篇与用户查询最相关的文档，计算扩展词与用户查询词的共现度，并与语义相似度累加，选取大于阈值的词作为最终扩展词。这种混合式查询扩展方法在用户搜索日志中仅包含流行的用户查询扩展词时，则只能选取与用户查询具有语义相似的流行的扩展词，同样忽略了其他部分有效的扩展词，具有局限性。如用户输入查询词“手机”，OLQEM 生成的最终扩展词集为“苹果手机，Iphone4，Iphone5s，Iphone6 Plus 三星手机，小米手机”等，忽略了部分有效的扩展词“小灵通，大哥大，飞利浦手机，老年手机，二手手机”等。

本文方法以概率方法为基础，分别计算扩展词与用户查询词的语义相似概率和词语共现概率，进而在 Copulas 框架下融合两种相似度度量指标，从而选取最高质量的扩展词组成最终扩展词集，实现了在统一框架内的不同类型相似度度

(下转第 496 页)

前端数据库系统中进行 5 次 OLTP 操作，每次的响应时间分别为 2.585s、2.396s、2.624s、2.525s、2.479s，平均时间为 2.522s。

图 5 所示是基于变更数据表建立索引视图的数据加载方案(dataload 方法)运行 5 次的数据加载时间。

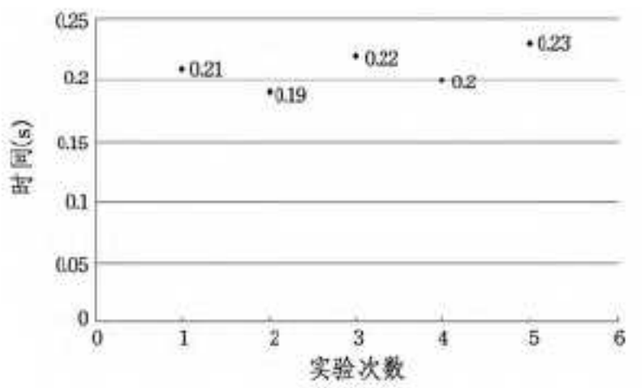

图 5 基于变更数据表建立视图 5 次 dataload 方法的执行时间

由图可见，5 次 dataload 方法的运行时间分别约为 0.21s、0.19s、0.22s、0.20s、0.23s，平均时间为 0.21s。同时在前端数据库系统中进行 5 次 OLTP 操作，每次的响应时间分别为 5.030s、5.203s、4.988s、5.146s、5.075s，平均时间为 5.088s。

通过实验数据可知，两种方法建立索引视图的数据加载方案对数据加载过程本身的效率影响不大。但基于变更数据表建立索引视图的思路，由于基表内容变化频繁，索引视图不断地刷新给前端 OLTP 系统带来了较大的压力，对事务操作的响应时间影响较大[9]。

通过以上分析可知，基于基本信息表建立索引视图的方案由于降低了一些连接和聚集操作的消耗，在数据加载的时间和对源系统的影响上都有所改进。而基于变更数据表建立索引视图的方案中每个周期的数据加载中只需要对一个索引视图进行查询，加载的实时性更好一些，不过由于变更数据表中数据变更频繁，维护索引视图时的数据刷新操作给源系统带来了很大的负担，事务处理系统的运行受到了较大的影响。

结束语 通过实验可以得出如下结论：采用基于基本信息表建立索引视图的方案不仅可以增强数据加载的实时性，还可以减小数据更新过程对源系统的影响。而采用基于变更数据表建立索引视图的方案可以更进一步增强数据加载的实时性，但是会给源系统带来更加沉重的负担。总之，索引视图是实物化的视图，实时的数据加载涉及频繁的连接和聚集操作，采用索引视图可以有效减小这些计算的代价。这一研究结论对动态数据仓库中实时数据加载方法的进一步研究有一定的借鉴作用，同时对特定的场合进行实时数据捕获时有一定的推广应用价值。

参 考 文 献

[1] iHaisten M. Real Time Data Warehouse: The Next Stage in Data Warehouse Evolution[J]. DM Review, 2003

[2] 徐富亮，周祖德. 变化数据捕获技术研究[J]. 武汉理工大学学报(信息与管理工程版)，2009，31(5)：740-743

[3] Ankorion I. Change data capture-efficient ETL for real-time BI [J]. DM Review magazine, 2005(1)

[4] 刘兆强. 基于快照差分的数据源更新检测方法研究及其实现[D]. 广州：暨南大学，2007：13-20

[5] 陆剑锋，张洁. 数据仓库数据更新的研究及基于 Oracle 数据库的开发与应用[J]. 计算机工程与应用，2004，40(26)：384-386

[6] 邹先霞，贾维嘉，潘久辉. 基于数据库日志的变化数据捕获研究[J]. 小型微型计算机系统，2012，3(3)：531-536

[7] 王珊，萨师煊. 数据库系统概论[M]. 北京：高等教育出版社，2009：89-91

[8] SQL Server 视图索引与索引视图指南[OL]. http://database.51cto.com/art/201007/212533.htm

[9] 谭光炜. 动态数据仓库实时数据的捕获及加载技术研究[D]. 贵阳：贵州大学，2015：50-63

(上接第 488 页)

量指标的有效合并。本文方法统筹兼顾了与用户查询词语义相似的词项和共现的词项，全部选取了与用户查询相似概率和共现概率都高的词项和与用户查询相似概率高或共现概率高的词项，有效地合并了语义扩展词集和统计扩展词集，极大地提升了搜索引擎的检索性能。

结束语 基于语义资料和局部分析混合式查询扩展可以同时提供具有语义相关和时效性的扩展结果，但如何有效混合两种相似度度量指标是尚未得到有效解决的问题，本文提出的基于 Copulas 框架的混合式查询扩展方法在统一框架内实现了不同类型相似度度量指标的合并。实验结果表明，该方法充分利用了语义及词语共现分析查询扩展方法的优点，有效地弥补了两者的不足，提高了搜索结果的查准率，比其它混合式查询扩展方法具有更优的搜索性能。

参 考 文 献

[1] Carpineto C, Romano G. A Survey of Automatic Query Expansion in Information Retrieval[J]. ACM Computing Surveys, 2012, 44(1): 1-50

[2] Selvaretnam B, Belkhatir M. Natural Language Technology and Query Expansion: Issues, state-of-the-art and Perspectives[J]. Journal of Intelligent Information Systems, 2012, 38(3): 709-740

[3] 李兴春. 信息检索技术中基于语义的扩展查询研究[J]. 重庆师范大学学报(自然科学版)，2013，30(4)：115-118

[4] Runkler T A, Bezdek J C. Automatic keyword extraction with relational clustering and Levenshtein distances[J]. Institute of Electrical and Electronics Engineers, 2002, 9(2): 636-640

[5] X Jin-xi, Croft B. Improving the effectiveness of information retrieval with local context analysis[J]. ACM Transactions on Information Systems, 2000, 18(1): 79-112

[6] 朱鲲鹏，魏芳. 基于用户日志挖掘的查询扩展方法[J]. 计算机应用与软件，2012，29(6)：113-115

[7] Pal D, Mitra M, Datta K. Improving Query Expansion Using WordNet[C]//CoRR. 2013: 1-18

[8] 王旭阳，萧波. 基于本体和局部上下文分析的查询扩展方法[J]. 计算机工程，2012，38(7)：57-59，69

[9] 吴秦，白玉昭，梁久祯. 一种基于语义词典的局部查询扩展方法[J]. 南京大学学报(自然科学)，2014，50(4)：526-533

[10] 欧阳柳波，谭容哲. 一种基于本体和用户日志的查询扩展方法[J]. 计算机工程与应用，2015，51(1)：151-155

[11] Sklar A. Fonctions de repartition an dimensions et leurs marges [J]. Publ. Inst. Statist. Univ. Paris, 1959, 8(1): 229-231

[12] Eickhoff C, de Vries A P, Collins-Thompson K. Copulas for Information Retrieval[C]//SIGIR'13. Dublin, Ireland, 2013: 663-672