第35卷 第10期 计 算 机 工 程 2009年5月
Vol.35 No.10 Computer Engineering May 2009
·软件技术与数据库· 文章编号：1000—3428(2009)09—0049—02 文献标识码：A 中图分类号：TP311.52

基于潜在语义分析的跨语言查询扩展方法

闭剑婷，苏一丹
(广西大学计算机与电子信息学院，南宁 530004)

摘 要：针对传统查询扩展方法存在的问题，提出一种基于潜在语义分析的跨语言扩展方法，利用聚类提高扩展文本集合的精度，并用潜在语义分析实现无需翻译的查询扩展，减轻翻译歧义带来的影响。实验结果表明，该方法能够获得较好的性能。
关键词：潜在语义分析；查询扩展；跨语言

Expansion Method for Language-crossed Query Based on Latent Semantic Analysis

BI Jian-ting, SU Yi-dan
(College of Computer & Electronic Information, Guangxi University, Nanning 530004)

【Abstract】 Aiming at the problems existed in traditional method of query expansion, an expansion method for language-crossed query based on Latent Semantic Analysis(LSA) is proposed, which uses clustering to improve accuracy of expansion word sets, and uses LSA to implement query expansion without translation. The impact of error in translation is eliminated. Experimental results show this method achieves better performance.
【Key words】 Latent Semantic Analysis(LSA); query expansion; language-crossed

1 概述

在跨语言信息检索中，很多研究者都借鉴单语言的扩展方法对查询式进行扩展，其中以两步伪相关反馈法(two-stage pseudo relevant feedback)的使用最普遍[1]。该方法用源语言查询式检索出排列好的源语种文档，在前面 n 篇文档中抽出 m 个出现频率最高的词作为查询扩展，并将新的查询式用双语词典进行翻译构成目标语言查询式，但该方法存在以下2个缺陷：(1)当前 n 篇文档与查询式无关时，会连带造成扩展偏差；(2)由于目前机器翻译的不成熟而产生翻译歧义问题。

针对上述问题，本文利用 k-means 算法进行聚类，提高扩展文档集合的相关度，利用潜在语义分析在处理文本时的语言无关性，直接从扩展双语文本集合中找出与源查询语言相似度最高的目标语种词汇作为扩展词，从而减轻翻译歧义对目标语言查询式的影响，提高查询准确率。

2 基本思路

2.1 双语潜在语义空间的建立

潜在语义分析(Latent Semantic Analysis, LSA)是种通过对大量文本进行分析，自动获取和表示词的语义信息，这种潜在语义信息是词语在上下文语境信息的总和，不依赖于特定的语种和单词的具体形式，而是由文本集合中单词的整体使用模式所确定。因此，能够实现完全自动的多语言处理[2]。

奇异值分解(Singular Value Decomposition, SVD)是最早提出也是目前普遍使用的典型 LSA 空间的构造方法。通过对文本集的词-文本矩阵的奇异值分解计算，并提取 k 个最大的奇异值及其对应的奇异矢量构成新矩阵来近似表示原文本集的词条-文本矩阵。

先根据具体的语料形成原始的 $m\times n$ 词条-文本矩阵 $\boldsymbol{A}=(\alpha_{ij})_{m\times n}$，其中，$\alpha_{ij}$ 要考虑局部权值 $L(i,j)$ 和全局权值 $G(i)$ 2个方面因素：

$$\alpha_{ij}=L(i,j)\times G(i) \tag{1}$$

实验证明以对数词频法取局部权值和以 Entropy 方法取全局权值得到的检索效果最好[3]。

$$L(i,j)=\mathrm{lb}(tf_{ij}+1) \tag{2}$$

$$G(i)=1-\sum_j \frac{p_{ij}\,\mathrm{lb}(p_{ij})}{\mathrm{lb}ndocs} \tag{3}$$

其中，$p_{ij}=tf_{ij}/gf_{ij}$，tf_{ij} 和 gf_i 分别表示词 i 在文档 j 和整个文档库中的出现频度；$ndocs$ 为文档库中的文档总数。由于要处理的语料是双语语料，因此分别形成中、英文的词条-文本矩阵：

$$\boldsymbol{C}=(c_{ij})_{mc\times n} \tag{4}$$

$$\boldsymbol{E}=(e_{ij})_{me\times n} \tag{5}$$

其中，mc 是中文词条数；me 是英文词条数；n 是文本数。为形成双语潜在语义空间，将2个矩阵拼接：

$$\boldsymbol{M}=\begin{bmatrix}\boldsymbol{C}\\ \boldsymbol{E}\end{bmatrix} \tag{6}$$

其中，$\boldsymbol{M}$ 为 $(mc+me)\times n$ 矩阵，再对 $\boldsymbol{M}$ 进行 SVD 分解，这里假设 $m=mc+me$：

$$\boldsymbol{M}_{m\times n}=\boldsymbol{U}_{m\times m}\boldsymbol{\Sigma}_{m\times n}(\boldsymbol{V}_{n\times n})^{\mathrm{T}} \tag{7}$$

其中，$\boldsymbol{U}$ 是 $m\times m$ 的正交矩阵，称为 $\boldsymbol{M}$ 的左奇异向量；$\boldsymbol{V}$ 是 $n\times n$ 的正交矩阵，称为 $\boldsymbol{M}$ 的右奇异向量；$\boldsymbol{\Sigma}$ 是 $m\times n$ 的对角矩阵，

基金项目：国家自然科学基金资助项目(60564001)
作者简介：闭剑婷(1981－)，女，硕士研究生，主研方向：人工智能，智能信息检索；苏一丹，教授
收稿日期：2008-11-10 **E-mail：**bijianting415@126.com

对角元素为 $\lambda_1,\lambda_2,\cdots,\lambda_r$，且 $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_r > 0$。根据因子分析理论和具体实验来选取 k 值，依据给定的阈值，选取前 k 个最大的因子，使 k 满足以下的贡献率不等式：

$$\sum_{i=1}^{k} \lambda_i / \sum_{j=1}^{r} \lambda_j \ge \theta \tag{8}$$

其中，θ 为包含原始信息的阈值，可取为 40%, 50%,…。贡献率不等式用以衡量 k 维子空间对原始空间的表示程度，但是这个数值可能很大，不便于规模的控制，考虑向量到向量运算的速度和存储空间的限制，一般 k 值取 100~300 之间。

选取适合的 k 值后，分别对 $\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V}$ 矩阵进行 k 截取，即截取 $\boldsymbol{U}$ 的前 k 列；$\boldsymbol{\Sigma}$ 的前 k 个最大的奇异值；$\boldsymbol{V}$ 的前 k 列。最后形成 A 的相似矩阵 $\boldsymbol{M}'$：

$$\boldsymbol{M}' = \begin{bmatrix} \boldsymbol{U}_k^c \\ \boldsymbol{U}_k^e \end{bmatrix} \boldsymbol{\Sigma}_k \boldsymbol{V}_k \tag{9}$$

其中，$\boldsymbol{U}_k^c$，$\boldsymbol{U}_k^e$ 分别为 $mc \times k$，$me \times k$ 矩阵，代表 k 维中文、英文词典向量矩阵；$\boldsymbol{V}_k$ 为 $n \times k$ 文本矩阵。

2.2 原始查询的同义词扩展

对包含 m 个查询词的原始查询 $Q_0 = \{t_1 t_2 \cdots t_m\}$，每个词需要找到它们在 k 维词典中的词向量。若该词在词典中，则存在对应词向量，否则用 fold-in 方法构造新词的 k 维词向量 $\boldsymbol{t}_k$：

$$\boldsymbol{t}_k = \boldsymbol{d}\boldsymbol{V}_k \boldsymbol{\Sigma}_k^{-1} \tag{10}$$

其中，$\boldsymbol{d}$ 为该词在各个文本中的出现频度向量。

由此得到 m 个词向量，通过计算向量间的相似度，进行同义词扩展，从词典中找到与原始查询各词具有最大潜在语义相似性的词。实验证明，采用 *pearson* 相关系数作为向量相似度计算公式较余弦相似度更适合本文所用语料，并且经过抽样统计，*pearson* 值大于 0.7 的两词向量，其词义相似的概率更大，所以，选取 0.7 作为同义词扩展的 *pearson* 阈值。经过此扩展后，源语言的查询式扩展为 $Q_1 = \{t_1 t_2 \cdots t_m t_{m+1} \cdots t_{m1}, m \le m1 \le 2m\}$。

2.3 LSA 跨语言扩展

将上述查询式 Q_1 在双语空间中按下述步骤进行跨语言扩展：

(1)将查询 Q_1 同样用 fold-in 方法构建成 k 维文本向量 $\boldsymbol{q}_k$；

(2)在已经建立好的双语空间文本矩阵 $\boldsymbol{V}_k$ 中计算各文本与 $\boldsymbol{q}_k$ 的相似度，找出相似度最大的前 T(T=500)个文本；

(3)对这 T 个文本进行 k-means 聚类[4]，找出和 $\boldsymbol{q}_k$ 最近的簇，该簇中的所有文本形成查询扩展文本集合 $ED = \{d_1 d_2 \cdots d_t, t < T\}$，且 ED 中的所有文本均为双语文本；

(4)对文本集合 ED 中的所有目标语言词汇用 Entropy 式(3)进行全局信息计算，找出最大的 *me* 个词，记为 $Q' = \{q_1 q_2 \cdots q_{me}\}$。

这 *me* 个词是目标语言词汇没有经过翻译，且由源语言的查询式产生，不存在翻译歧义的影响，将这 *me* 个词作为目标语言查询的一部分。

2.4 翻译

尽管源语言查询式 Q_1 在翻译过程中可能存在歧义，但由于 Q_1 是和用户查询意图最接近的词语组合，应该将其应用到目标语言查询式中。本文借用已经形成的双语词典，用双向翻译模型对 Q_1 进行翻译[5]，得到由翻译形成的目标语种查询式 $Q_2 = \{t_1' t_2' \cdots t_m' t_{m+1}' \cdots t_{m1}', m \le m1 \le 2m\}$。将 Q' 和 Q_2 结合，得到最终的目标语言查询式 $Q = \{t_1' t_2' \cdots t_m' t_{m+1}' \cdots t_{m1}' q_1 q_2 \cdots q_{me}\}$。

2.5 查询词的权重

查询扩展后，不同的词在查询中的重要性不同。因此，查询扩展要考虑的另外一个问题是扩展后各个词在新查询中的权重分配。

大多数的查询扩展方法都直接采用 Rocchio 式(6)来计算扩展后的新查询 Q_{new} 中每个词项 q 的权重：

$$Weight(q \mid Q_{new}) = \alpha \cdot Weight(q \mid Q) + \beta \cdot \frac{\sum_{D \in S} Weight(q \mid D)}{n} \tag{11}$$

其中，$Weight(q \mid Q)$ 为查询词 q 在初始查询 Q 中的权重，通常直接使用 q 在 Q 中的频度来表示；$Weight(q \mid D)$ 为查询词 q 在文档 D 中的权重，其值与所采用的检索模型具有一定的关系；n 为局部文档集中的文档数；α 和 β 为 2 个大于 0 的可调参数。

然而，上述的 Rocchio 公式没有把扩展时对扩展词的评分考虑进去，在进行查询扩展时，对扩展词的评分值从另外一个角度也反映了扩展词的重要程度。基于这种考虑，采取如下简单的权重分配方法：

$$Weight(q \mid Q_{new}) = \alpha \cdot Weight(q \mid Q) + \beta \cdot \frac{Score(q)}{MaxScore} \tag{12}$$

其中，$Score(q)$，$MaxScore$ 由最后的目标语言查询式 Q 中各个词项的来源决定。$Q = \{t_1' t_2' \cdots t_m' t_{m+1}' \cdots t_{m1}' q_1 q_2 \cdots q_{me}\}$，$\{t_1' t_2' \cdots t_m'\}$ 是由源语言原始查询 $Q_0 = \{t_1 t_2 \cdots t_m\}$ 中各个词项翻译而来(记为第Ⅰ类)；$\{t_{m+1}' t_{m+2}' \cdots t_{m1}'\}$ 是同义词扩展后翻译而来(记为第Ⅱ类)；$\{q_1 q_2 \cdots q_{me}\}$ 是经 LSA 跨语言扩展而来(记为第Ⅲ类)。

若 q 属于第Ⅰ类，取 $Score(q) = MaxScore$；若 q 属于第Ⅱ类，$Score(q)$ 取其在同义词计算中得到的 *pearson* 值，$MaxScore$ 值取 1；若 q 属于第Ⅲ类，$Score(q)$ 取其在扩展文本集合 ED 中的全局信息值(*Entropy* 值)，$MaxScore$ 为 $Score(q)$ 的最大值。

3 实验设计

3.1 数据集和测评方法

实验所选用的数据集是从中国知网 CNKI 系列数据库中收集的计算机(372 篇)、哲学(350 篇)、文艺(390 篇)、农学(360 篇)4 类文章的中英文摘要。

每类取 70%左右(计算机 260 篇、哲学 245 篇、文艺 273 篇、农学 252 篇)建立 LSA 双语空间，用于之后的查询扩展和翻译模型。

主要测评指标是 MAP(Mean of Average Precision)，并以 Prec@10 和 Prec@20 作为辅助测评指标。*MAP* 表示查询集中每个查询准确率的算术平均值。

Prec@X 表示某个查询 Q 在检索出 X 篇文档时的准确率，Prec@10 和 Prec@20 在搜索引擎中通常反映在第 1 页和第 2 页搜索结果的准确率。

3.2 查询扩展规模

扩展规模包括 2 个方面：(1)扩展文本集合的规模；(2)扩展词的个数。本文中的扩展文本集是由第 1 次检索的结果聚类而得，其规模大小和聚类算法有关，这里不对该问题进行讨论。由于本文中近义词扩展的词条数相对于 LSA 跨语言扩展的词条数较少，因此只讨论通过 LSA 跨语言扩展的规模，即 *me* 的大小。若 *me* 过小，则没有达到扩展的目的，*me* 过大，则有可能引入噪声，检索出无关文档。

图 1 为实验后得到的 *MAP* 和扩展词数的关系图，从图中

(下转第 53 页)