潜语义标引 文语信息检索 关键词 (25)

计算机科学2000Vol. 27№. 3

查全率

93-95

潜语义标引与汉语信息检索研究

The Application of PSI in Chinese Information Retrieval

刘博勤 丁晓明
(西南师范大学 重庆400715)

G354.4

Abstrat On the space-based information retrieval, Latent Semantic Indexing (LSI) extracts the patent semantic structure from the document database according to statistical information in the term-document matrix, and forms the semantic sub-space, so as to improve the query precision. This paper investigates on the application of the 2-gram noun phases into LSI for Chinese information retrieval. Tested with the TREC Topics, the R-precision improved by 10.9%.

Keywords Information retrieval, Indexing, Semantic indexing, Chinese

1 引言

典型的传统信息检索系统，如布尔逻辑模型、向量空间模型，根据用户提供的查询条件，依据关键词的匹配或向量空间的相似系数，返回相关查询结果。对于相同的概念，使用不同的词汇表示，如同义词或近义词，或同一词汇在不同的语言环境中拥有不同的语义，即一词多义，因此基于语词匹配的查询方法，其准确性和完整性都不够理想。尽管同义词词典的使用，在一定程度上，提高了信息检索的查全率(recall)，但却降低了查询的精度，且在实际应用中，需要不断更新同义词库，才能满足系统不断变化的要求。

在自动信息标引过程中，利用语法、语义分析方法，抽取短语结构进行文档标引，但取得的实验结果并不理想[1]，因为短语的特征过于单一，检索的查全率较低。况且，汉语语法、语义分析更加复杂，利用深层处理进行自动语义标引，其效率和精度都有待提高。然而基于统计的浅层自然语言处理方法，如词性标注、同义词聚类等，在信息检索中的应用，取得了较好的效果。

潜语义标引(LSI)根据统计方法，获得数据库文档潜在的语义概念空间结构，利用概念标引取代关键词标引[2~4]。潜语义标引，假设语词的用法具有潜在的关联关系，而且这种关系被语词的用法变化所模糊。将关键词-文档关联矩阵进行奇异值分解，从中抽取出语词的用法特征，以消除语词用法变化所造成的影响。实验证明，潜语义标引比原有的语词标引具有更准确的文档内容表达能力。

潜语义标引，与传统的信息检索技术相比，其信息检索的精度有了一定的提高。但在现有的潜语义标引研究中，没有考虑关键词自身的特征对潜语义标引的影响。本文根据汉语的特点，分析了词性和词组对潜语义标引的影响和作用，并介绍了基于词性信息的2元语法短语的潜语义标引在信息检索中的应用。

2 潜语义标引

Furnas 等[2~4]将奇异值分解应用到文档的语义特征抽取中，将文档的关键词向量空间转换为语义概念空间，并在降维的语义概念空间中，进行相关信息查询，称之为潜语义标引。潜语义标引方法，增加了下列语义信息：根据文档数据库的全局信息，将文档数据库向量空间影射到维数较低的正交语义概念子空间。在传统关键词的权重计算中，每个关键词是相互独立的，而 LSI 假设相关关键词是性线依赖的。LSI 降低向量空间的维数，克服关键词用法变化所造成的影响，保留文档的语义信息。

LSI 假设文档的语义相似性与语言用法模式相一致，从这些模式中能够抽取出相关语义信息。LSI 的数学基础是关键词-文档矩阵 X 进行奇异值分解 SVD (Singular Value Decomposition)，矩阵 X 可唯一分解为 $T_0S_0D_0^T$，如图1所示。其中 T_0为"左奇异值向量"，D_0为"右奇异值向量"，其列向量都两两正交；S_0为奇异值对角矩阵，其值按照降序排列。

与 X 相对应，矩阵 T_0中，具有类似用法的关键词，对应的向量也相似，用法不同的关键词对应的向量也不同。

$$\underset{t\times d}{X} = \underset{t\times r}{T_0} \times \underset{r\times r}{S_0} \times \underset{r\times d}{D_0^T}$$

图1 关键词-文档矩阵的奇异值分解

假设矩阵 X 的秩为 r，则 S_0恰好 r 个对角元素，都为正值，矩阵 T_0具有下列特性：1)T_0矩阵保持了原有文档的相似性；2)删除 T_0矩阵中每个向量最后几个成

员，将极大地提高文档相似性排列次序。

从 T_0、S_0、D_0向量中，取前 k 个向量，删除其它 r－k 个向量，X 中文档向量的内积变化将最小[9]。假设删除后 r－k 个分量得到的向量矩阵分别为 T、S、D，由 $T\times S\times D^T$ 恢复的矩阵 X′ 保留了原文档数据库的重要频率特征，同时降低了语词用法变化对信息检索精度的影响。将文档空间转化为概念空间，可提高信息检索的精度。如图2所示

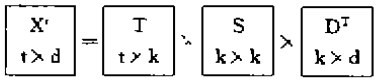

图2　关键词-文档矩阵的恢复

k 的取值与文档数据库的大小相关，一般在100～300之间。根据实际文档数据库的大小，选择合适的 k 值。为了进行信息检索，用户的查询条件作为伪文档向量，需要转化为 k 维概念空间向量，在语义子空间中，进行文档相似比较。用户的查询条件可按照下式转换：

$$\bar{q}=q^TTS^{-1} \qquad (1)$$

其中，q 为用户查询向量，$\bar{q}$为 k 维空间向量，在降维语义概念空间中，计算查询条件与文档向量的相似性，根据相似系数的大小，排列相关文档，作为检索结果。

3　基于词性信息的 n 元语法短语潜语义标引

与西文信息检索相比，中文信息检索有其独特的一面。汉语语词间没有分隔符号，语词的切分歧义和新词的识别需要借助于上下文语法、语义信息。信息检索的精度与语词切分的准确性密切相关。在许多信息检索中，基于单字、n 元语法单字串、切分语词等单元进行标引[5～7]。基于单字和 n 元语法单字的标引，在 TREC-5和 TREC-6的测试结果中，虽然基于 n 元语法单字的查询精度与基于语词切分方式的查询精度相近，但 n 元语法单字标引所需要的空间为基于语词方法的 n 倍，且查询的精度要低。为了减少系统的存储空间，提高信息检索的精度，在语词切分的基础上，进行语词的词性标注，采用语词标注和 n 元语法的语词标注相结合的方法，可克服基于单字的 n 元语法标引的不足，利用语词的词性信息和搭配关系，可提高信息表示的准确性。

3.1　语调切分及词性标注

汉语语义信息的最小单位是语词，所以语词标引是智能汉语信息检索的基础。语词切分存在切分歧义，切分的歧义有两种：交集型和组合型，其中交集型切分歧义约占90%[11]。在本系统中主要利用语词频率和词性搭配信息，消除切分歧义。与其它自然语言处理相比，汉语信息检索对切分精度的要求并不十分严格，可根据语词的位置关系检索出相关信息。

在信息检索中，不同词性的语词所表示的语义信息并不相同，名词、动词、形容词、副词等所表示的语义信息，比其它词性语词要多。基于词性信息的语词标引，更能够准确地表示文档的信息内容，提高自然语言查询、相关信息反馈查询、信息过滤和分发的准确性。

语词词性标注采用马尔科夫模型。以《现代汉语语法信息词典详解》对语词属性的描述和清华大学语料库词性标记集为基础[12,13]，将词性划分为26个大类，82个子类，25个标点符号，共计107个词性细类。使用标注语料，开放语料词性标注的准确率为96%。其中词性的搭配信息反馈用于语词切分的歧义消除。

3.2　基于词性信息的2元语法标引

利用简单的语词表达文档的内容，通常是不合适的，因为单个语词语义含义较广，语词的组成通常是偶然的，不能够准确地区分文档内容。更好的方式是利用有意义短语词组标引文档的内容，特别是，如果这些短语在该领域中表示重要的概念。例如“希望工程”是 TREC 中文信息检索测试库中的第十九个主题词[14]，其中既不是“希望”，也不是“工程”关键词本身很重要。“希望”、“工程”在文档数据库中出现的频率较高，它们的文档频率倒数(idf)权重太低。在包含成千上万文档的数据库中，短语关键词是十分必要的。

目前利用完全的语法、语义分析获取短语进行文档标引，其效率无法适应大规模动态信息检索的要求。纯粹的2元语法标引，关键词数量为 N^2，N 为语词数。且并非所有2元语法标引都表示正确的语义内容。因此，在语词切分、词性标注的基础上，进行2元语法的标引，成为信息检索标引的一条切实可行的途径。

在 TREC 测试中，基于潜语义标引的信息检索、过滤系统，没有利用关键词的语义表示能力。如词性以及短语标引，提高关键词的内容表示能力。将潜语义标引中的关键词特征统计信息同关键词的语言属性结合起来，克服潜语义标引中关键词语言信息的不足。

基于语词的2元语法标引，增加相邻语词词组，作为关键词。如语句“北京市国有企业主要经济指标大幅增长”的二元语法关键词为：“北京市”、“市国有”、“国有企业”、“企业主要”、“主要经济”、“经济指标”、“指标大幅”、“大幅增长”。然而并非所有二元语法关键词都反映文档的语义信息。如上述二元关键词中“市国有”、“企业主要”、“指标大幅”等属于不同的语法结构单元，组成错误语词搭配关系，可利用词性信息，减少非法短语标引数。

基于词性信息的2元语法标引，主要标引由名词＋名词、动词＋名词、形容词＋名词以及由助词连接的上述结构短语。2元语法标引，并非基于语法结构分析基础之上的短语抽取，会存在不合法的短语标引，但对信息的检索精度影响较小。

3.3　短语权重计算

在进行信息检索前，利用潜语义标引方法，对文档数据库进行标引。关键词权重采用 tf * idf 计算[10]。短

语所表示的语义信息，比语词所表示的信息要更具针对性，在语词标引的基础上，增加短语标引，2元语法短语关键词的本地权重 plw_k 计算如下：

plw_k = ln(短语关键词 i 在文档 k 中出现的次数)+1/2(ln(关键词 w1 在文档 k 中出现的次数)+ln(关键词 w2 在文档 k 中出现的次数))

其中 w1、w2分别表示2元语法短语的第一个词和第二个词。权重的规一化，相对于所有语词和短语关键词来计算。

4 基于词性信息的 n 元语法潜语义标引实验

测试的主题信息从 TREC 所提供的54个主题中[14]选择10个文档数据库所包含的主题内容，分别是 CH2、CH3、CH4、CH6、CH16、CH18、CH19、CH38、CH41、CH53。测试的数据资料为95年1、2、3月份新华社新闻，共计13398条。根据主题内容描述、事先判定，哪些文档与给定的主题相关，哪些不相关，从中筛选出相关的 N 个文档。测试要求是根据给定主题的主题名称、相关的关键词、主题内容描述等因素，查询并返回 N 个相关文档列表及相关评分。根据相关文档数，计算查询的精度。

利用平凡词表，包含120个平凡语词，删除常用的平凡语词，实际标引的语词数为29995。采用稀疏矩阵 Lanczos 迭代算法进行关键词-文档关联矩阵奇异值分解[9]。在 P166、128M 内存微机上，k 取值为309时，上述文档数据库奇异值分解迭代次数为839，迭代时间为2小时10分。

查询条件的形成由主题名称、相关的关键词以及主题内容描述(手工选择)三部分内容组成，三部分的权重比例分别为2:2:1。对主题的名称、关键词、描述内容三部分，进行切分和2元语法短语抽取，再进行规一化处理，产生的伪文档向量作为查询条件。将查询条件，按照式(1)，转换成 k 维语义子空间向量，在语义子空间中，进行查询-文档的相似性计算。按照相似系数的大小，排列输出结果。

查询包括基于语词级的潜语义标引和基于词性信息的2元语法的潜语义标引，及基于传统向量模型的相关结果。采取等量查询方式，即返回的结果为数据库中的实际相关文档数，R 精度值可同时反映出系统的查询精度和查全率，比较不同标引方式的查询精度。

测试结果表明，基于词性信息的语词2元语法的潜语义标引，与基于语词的潜语义标引相比，系统的相关检索精度提高了10.9%。如主题 CH19，"希望工程"，在基于语词的潜语义标引查询中，许多与"希望工程"无关的"工程"建设项目内容，被作为结果返回。这是由于"工程"在某些文章中出现的频率较高，使其文章获得较高的相似值。

另外一些主题的查询精度变化不大，如主题 CH38，"中国野生动物保护形势"，由于特定的语词和词组仅出现在相关的文档中，使用短语标引，实际上没有提高相应检索的精度。

此外，结果分析发现，词典中没有的新词，如"京九铁路"中的"京九"，容易影响查询的精度，主要因为向量空间模型中，没有利用语词的上下文位置关系，查询所需要的信息。2元语法标引能够部分解决新词识别问题。

结论 在向量空间模型中，潜语义标引能够将文档信息和用户的查询要求，从语词空间转换为语义子空间，从而按照语义信息进行信息的检索，与传统的向量空间相比，克服了一词多义和一义多词对信息检索精度的影响。基于词性的2元语法潜语义标引，针对中文信息检索需要进行语词切分和新词识别的特点，将潜语义标引和2元语法短语的信息表示能力结合起来，进一步提高了中文信息检索的精度。基于语短的潜语义标引，需要进一步提高汉语短语，特别是名词短语的识别能力，以便更准确反映文本的语义信息。

参考文献

1 Strzalkowski T. Natural Language Information Retrieval. Information Processing & Management, 1995, 31(3): 397～417

2 Furnas G W, et al. Information retrieval using a singular value decomposition model of latent semantic structure. In: Proc. of SIGIR. 1988. 465～480

3 Deerwester S, Dumais S T. Indexing by latent semantic analysis. Journal of the Society for Information Science, 1990, 41(6): 391～407

4 Berry M W, Dumais S. Using Linear Algbra for Intelligent Information Retrieval. CS-94-270, Dec., 1994

5 Leong M K, Zhou H. Preliminary Qualitative Analysis of Segmented vs Bi-gram Indexing in Chinese. In: Text Retrieval Conference(TREC-5). NIST, Gaithersburg, Maryland, 1996

6 Kwok K L, Grunfeld L. TREC-5 English and Chinese retrieval experiments using PIRCS. In: Text Retrieval Conference(TREC-5). NIST, Gaithersburg, Maryland, 1996

7 Kwok K L, Grunfeld L. TREC-6 English and Chinese retrieval experiments using PIRCS. In: Text Retrieval Conferenc(TREC-6). NIST, Gaithersburg, Maryland, 1997

8 Merialdo B. Tagging English Text With a Probabilistic Model. *Computational Linguistics*, 20(2): 155～168

9 Berry, M. SVDPACKC User's Guide. CS-93-194, 1996

10 Salton G. Introduction to Modern Information Retrieval, McGraw-Hill, 1983

11 刘源. 信息处理用现代汉语分词规范及自动分词方法. 清华大学出版社, 1994

12 余士汶. 现代汉语语法信息词典详解. 清华大学出版社, 1998

13 黄昌宁. 语言信息处理专论. 清华大学出版社, 1996

14 NIST. 中文信息检索主题1, 2. TREC@NIST. GOV