知识组织与
知识管理

潜在语义标引(LSI)研究综述*

孙海霞 成 颖

(南京大学信息管理系 南京 210093)

【摘要】 在回顾我国潜在语义标引技术的研究成果基础上,分析、总结我国现有潜在语义标引研究的不足,指出我国潜在语义标引的进一步研究方向。

【关键词】 潜在语义标引 文本处理 信息检索 【分类号】 TP393

Overview of Research on Latent Semantic Indexing

Sun Haixia Cheng Ying

(*Department of Information Management, Nanjing University, Nanjing 210093, China*)

【**Abstract**】 Based on reviewing and analyzing research results on LSI in our country, the authors analyse and summarize existing deficiencies of study on LSI, and outline the research directions of LSI.

【**Keywords**】 Latent semantic indexing Text process Information retrieval

1 引 言

文本中词语之间的关联性以及自然语言中一词多义和异形同义现象的存在,使得传统的基于独立关键词索引匹配的信息检索系统检索出来的结果往往不够准确,会漏掉大量相关信息。为此,学者们积极展开研究,先后提出了词干法(Stemming)、控制词表法(Controlled Vocabularies)等解决方法,但由于这些方法的实质依然是关键词匹配,改进非常有限,从而无法根本解决上述问题[1]。1988 年,Dumais S T. 等人提出了一种新的信息检索代数模型:潜在语义标引(Latent Semantic Indexing, LSI)模型,实现了基于概念的语义检索,较好地解决了上述问题,提高了检索系统的准确率[2]。

2 潜在语义标引 LSI[2]

LSI 假设文本中存在某种潜在的语义结构,这种潜在的语义结构隐含在文本中词语的上下文使用模式中,可通过利用统计方法获得[3]。其核心思想[1]是通过奇异值分解,将文档向量和词(Term)向量投影到一个低维空间,使得相互之间有关联的文献即使没有相同的词时也能获得相同的向量表示。LSI 技术主要包括以下几个步骤:

(1)词 - 文档矩阵的构建与优化

LSI 模型中,文档库是用词 - 文档矩阵(Term - Document Matrix)A_{mn}来表示的。m 为文档库中不同词的个数,一个词对应矩阵 A 中的一行;n 表示文档库中的文档数,每个文档对应矩阵 A 中的一列;a_{ij}表示第 i 个词在第 j 个文档中出现的频率(tf)。Dumais 等人通过实验发现,词加权可以比单纯的 tf 的准确率高出 40% 左右[3],因此,一般使用原始频率值 a_{ij}的优化值。通常,优化值 a_{ij}为局部权值计算函数 L(i, j)和全局权值计算函数 C(i)的乘积。L(i, j)的计算方法通常有词频法、二值法以及对数词频法,C(i)则常用 Idf 以及 Entropy 等方法进行计算,Dumais 的研究表明,L(i, j)使用对数词频法,C(i)使用 Entropy 时效果最好[4-6]。

(2)奇异值分解 SVD - 降维

奇异值分解(Singular Value Decomposition)是 LSI 技术的关键[1]。根据代数知识对 A 进行奇异值分解,使得:

$$A = U\sum V^T$$

其中,U 和 V 分别为与矩阵 A 的奇异值对应的左、右奇

收稿日期:2007 - 07 - 17
收修改稿日期:2007 - 07 - 29
* 本文系南京大学人文社会科学项目"网络环境下异构信息检索标准体系研究"的研究成果之一。

异向量矩阵,且有 $U^TU = V^TV = I$;Σ 是将矩阵 A 的奇异值按递减排列构成的对角矩阵。取 U 和 V 的前 k($k << \min(m, n)$)列,使得:

$$A_k = U_k \Sigma_k V_k^T$$

A_k 是词 - 文档矩阵 A 的近似表示,实际上是一个低维的语义空间,一方面消减了原词 - 文档矩阵中包含的"噪声"因素,凸显了词与文档之间的语义关系,另一方面缩减了词、文档向量维数,使得文档检索效率得以大大提高[4]。U_k 和 V_k 的行向量分别表示词向量和文档向量。k 是语义空间的维数,其大小直接关系到检索质量和检索效率的高低[1],过小会丢失有用信息,达不到区分文献或标引词的目的;过大则捕获不了词与词之间的关联性,达不到语义检索的目的,同时也提高不了检索速度[5]。目前常用的确定 k 值的方法是参考因子分析中 k 值的选择方法的贡献率不等式法[5,6]。

(3)基于潜在语义空间模型的查询

潜在语义检索中,对于用户查询语句(用户可以直接用自然语句进行查询)同对文档的处理一样,根据用户查询语句中各词语的出现频率生成查询矢量 q,其中,第 i 个元素 q_i 的数值表示了第 i 个词语在查询语句中出现的频率。然后将 q 和 V_k 中的每个向量进行比较,计算出它们之间的相似度值。在 LSI 模型中,有 3 种类型的相似度计算[2,7]:词 - 词,词 - 文献,文献 - 文献。不同的相似度计算,所用的向量矩阵是不一样的。最后选择合适的阈值,将符合要求的文献返回给用户。

3 潜在语义标引研究

纵观国内公开发表文献,我国关于潜在语义标引技术的研究较国外要晚上近十年[4,7],公开发表物主要集中在 2000 年后,研究主要集中在计算机领域,图书情报领域的的研究相对较少。

根据 CNKI 以及维普全文数据库的检索,文献[7]首次在国内对 LSI 技术进行了系统的介绍,指出了其在中文信息检索中的广阔前景。目前,LSI 已经被广泛地应用于自然语言处理的各个领域,下面主要依据国内发表的相关文献对 LSI 在信息检索等领域的应用以及研究进展进行述评。

3.1 信息检索

LSI 在信息检索领域的应用,国内除了上文中第二部分所探讨的基本层面上的应用之外,还取得了以下几方面的进展:

(1)语义扩展。顾榕等利用 LSI 技术分析检索结果的聚类中心向量项之间及不同类别向量项与检索词间的语义关系,然后基于 Wordnet 等工具,从近义词、关联词及区分词 3 个角度对初始查询进行扩展,从而有效地提高了信息检索系统的性能[8]。

(2)跨语言检索。林鸿飞等利用 LSI 技术构建潜在语义空间时,无需考虑具体语种的语义和句法,因此可以构建多语种的 LSI 空间,即将不同语言所表达的同一内容映射成 LSI 中的同一个向量,实现跨语言信息检索[9,10]。

(3)个性化服务。信息的剧增使得个性化信息检索越来越重要。在奇异值分解中考虑用户特征关键词的个性权重,利用该含有用户个性化信息的潜在语义关系矩阵处理待检索文献和用户查询语句可以有效实现个性化信息检索服务[11]。

(4)问答系统的改进。通过构建潜在的词 - 句子语义空间,然后在语义空间上实现了问题与答案句子相似度计算,能较好地解决汉语问答系统答案提取中由于同义或多义而导致的"漏提"或"错提"等问题[12,13]。

另外,针对中文信息检索需要进行语词切分和新词识别的特点,将潜语义标引和二元语法短语的信息表示能力结合起来,能够进一步提高中文信息检索的精度[14]。

3.2 分类与聚类

传统文本分类方法大都是通过计算文本向量间的距离而忽略了特征项间的相互影响关系来决定类别的,分类精度不是很理想[15]。通过 SVD 将高维样本文本特征项 - 权重矩阵映射到能够反映特征项向量间语义关系的低维语义空间中,提取的语义特征不仅能够反映文档和词的特征信息,还考虑了文档与类别之间的关联信息,有效地提高了分类的准确率[16,17]。除了将 LSI 与传统的分类方法简单地相结合外,学者们对 LSI 模型自身的改进也展开了积极研究,以更好提高分类聚类的效果[17-19],具体包括:

(1)语义标注。LSI 虽然能够有效捕捉特征项间的语义关联,但它缺乏清晰的表达能力,对特征项进行概念标注能够增强 LSI 的表达能力[6],予以层次分类确定含义。

(2)较小特征值的处理。在 LSI 模型中,一些对分类贡献大的特征项由于其对应的值较小而往往在分解中被过滤掉。对此,曾雪强等提出了一种扩展的 LSI 模型,即在构建文档的语义空间的同时也考虑文档和类别之间的联系。此外,根据特征聚合理论,将对分类贡献相近的特征项归属于同一个新特征项也可以在一定程度上解决此问题[15]。

(3)多类、多层分类模型的研究。一个文本一般具有多个类别属性,在对文档信息和文档类别信息进行语义建模的同时,进一步考虑文档的多类属性信息可实现文本的多类分类[20]。另外,根据潜在语义空间中纬度的统计特性,通过采用不同维度可实现文档在不同概念粒度下的聚类[21]。

3.3 信息过滤

1992 年,Foltz 和 Dumais 首次将潜在语义分析技术 LSA 用于信息过滤中[1]。信息过滤中,用户的兴趣主题

是固定的,而所面对的文献集合则是不断变化的。基于 LSI 的信息过滤,首先要对用户感兴趣的信息实例进行分析,构建出关于用户兴趣主题的潜在语义空间模型,用 LSI 向量表示用户信息需求倾向。新文档到达时,把新文档映射到用户兴趣主题语义空间去,与用户兴趣向量进行匹配来决定取舍[1,22]。山西大学的牛伟霞、张永奎鉴于此思想,成功构建了中文科技文献过滤系统的兴趣主题模型,当 k 取值 100 时,平均准确率比 SVM 方法高出 13.6%[22]。

在邮件过滤系统中,无论是将 LSI 技术与传统的过滤技术相结合[23],还是通过基于 LSI 的文本分类来实现[24],关键是用户主题 LSI 模型或垃圾邮件语义空间的构造[25],然后将新邮件与垃圾邮件向量或用户兴趣主题向量相匹配。

3.4 信息抽取

修辞手法的存在使得简单的词频大小不能表征特征项对文本主题的表现能力,因此主要根据线索词典、词频、词或句子的启发性函数进行模式匹配以摘取文本中重要句子形成摘要的传统方法并不令人满意[27]。潜在语义索引能够利用训练集分析获得的语义知识对自然语言文本进行分析确定文本的主题,从而自动提取文本的概要[1,26]。

在语义空间中计算特征项的权重要考虑特征项对文本主题的表现力以及在整个文本集中使用的模式。不同的特征项、句子和段落对文章主题的贡献是不一样的,因此,基于 LSI 的信息摘要的关键是特征项、句子和段落权重函数的确定,林鸿飞等人分别利用不同的权重函数公式进行了实证研究[26,27]。

将不同语言建立的语义空间中同一事物的概念加以适当的标记索引,通过适当的转换,可实现中西文信息的交叉提取[28]。

3.5 其他领域

(1)图像处理[29-31]。主要集中在图像检索领域。将 LSI 技术引入基于文本检索的系统中,能够有效解决语言多样性带来的问题,即图片标引人员之间、标引人员和用户之间用词不一致问题,提高图片的查全率和查准率。在基于内容的图像检索系统中,实现图像语义处理必须解决两个问题:提供高层语义的描述方式和实现图像视觉特征与高层语义之间的映射。将图像特征看作是文本中的词频,高层语义看成是压缩后的潜在语义空间,LSI 就能够解决上述问题,真正实现图像的语义检索。

(2)语音识别。将潜在语义分析技术能够获取上下文语境信息这一思想应用于语音识别中,可依据所有用来识别历史的词去预测下一个词,从而在识别过程中融入长距离语义信息[32]。任纪生和王作英于 2004 年前后运用新提出的几何加权静态插值方式同三元文法模型相结合,构建了一种新的适于大规模训练的潜在语义分析语言模型[33]。

(3)在认知科学方面,LSI 提供了一种知识进行归纳、表征和应用的理论模型[34]。因语义知识的获取是建立在对大量语料库的统计与分析的基础上的,LSI 模型本身也是一种学习机制[35,36]。

此外,文献[27]、[34]、[36]、[37]对如何实现对文本的理解和知识的获取及在教学中的应用也进行了探讨研究。

4 潜在语义标引的研究方向

经过十多年的努力,学者们在 LSI/ LSA 方面已经取得了巨大的成就,但综合已有的成果来看,为充分实现 LSI 的价值,还有许多问题亟待研究,如:

(1)当前 LSI 主要是通过 SVD 构建语义空间的,但 SVD 算法的时间代价 $O(N^2K^3)$ 和空间代价使得 SVD 算法不太适合于大规模动态变化的数据集[38],因此,一方面需要研究时间和空间代价相对较低的数学模型来支持 LSI 的思想,另一方面,需要研究大规模和超大规模稀疏矩阵奇异值分解的高效算法和面向动态变化的大规模文档库的 SVD 实时更新算法[3,7]。

(2)根据不同处理对象和条件制定最佳 k 值是学者们在实验研究中应予以重视的问题[1],但基于 LSI 的信息检索质量不仅与 k 值的选取有关,也与词 - 文档矩阵 A 本身的建立有关,因此,应加强语料库技术的研究,以使初始词 - 文档矩阵 A 得以优化。

(3)SVD 分解中舍弃了奇异值较小的向量,但恰恰是这些小的奇异值反映了文本的特性,因此,研究如何在不影响压缩语义空间模型下充分发挥反映文本的特性的小的奇异值作用应是今后的一个研究方向[4,39]。

(4)LSI 是通过所有词向量的线性总和来产生文本向量表示文本的含义的,忽略了词语的语法信息和词语在句子中的逻辑顺序,无法考虑句子语法结构中所包含的词语之间更深层次的语义关联信息[26,40]。因而,需要将 LSI 的基本思想和语法信息相结合,提高 LSI 模型对文本内容的把握能力,进而提高文本信息处理系统的性能[41]。

(5)在分类和聚类方面,今后应加强 LSI 与传统分类/聚类技术的优劣势互补研究以提高分类/聚类的准确率和速度[17];同时,也要更多地利用 LSI 技术进行多元和多层分类模型的研究[20,21]。

(6)潜在语义索引缺乏明确的概念解析表达能力,如何建立面向文本处理领域的概念词典,实现 LSI 节点概念标注的支持性研究也应是今后的一个研究方向[20, 39]。

(7)将 LSI 与相关反馈、本体等技术相结合能更好实现个

性化服务,因此,应加强后者的研究以进一步发挥 LSI 的优势,有效实现个性化信息检索服务。

(8)在认知科学方面, LSI 模型缺乏大量重要的经验知识和人类用来构建和应用经验知识的认知能力。因此要加强 LSI 在分析大规模语料库和经验知识表征及利用的能力方面的研究[35,36]。

(8)当前关于 LSI 的应用研究大都集中在文本信息的处理研究上,在图像、音频、视频等多媒体领域中的应用研究相对较少,今后应加强 LSI 在多媒体为领域中的应用研究,如图像、音频、视频等多媒体信息的检索和过滤[25,42]。

另外,在跨语言检索中,应加强建立平行语料库的研究;对于大规模的动态的数据集合,应加强数据库更新技术研究以支持 LSI 的研究;要加强和拓宽 LSI 在其它领域的应用研究[43],如信息判定和预测[44]、信息浏览[40]、P2P 网络[45]等。

参考文献:

[1] Dumais S T. Latent Semantic Analysis[M]. Annual Review of Information Science and Technology, 1989:190 - 230.

[2] Dumais S T, Furnas G W, Landauer T K, et al. Using Latent Semantic Analysis to Improve Rnformation retrieval[C]. Proceedings of CHI'88 Conference on Human Factors in Computing Systems, 1988:281 - 285.

[3] Deerwester S, Dumas S T, Furnas G W, et al. Indexing by Latent Semantic Analysis[J]. Journal of the American Society for Information Science, 1990, 41(6):391 - 407.

[4] 周水庚,关佶红,胡运发. 隐含语义索引及其在中文文本处理中的应用研究[J]. 小型微型计算机系统,2001, 22(2):239 - 243.

[5] 杨梁彬. 文本检索的潜在语义所引法初探[J]. 大学图书馆学报,2003(6):68 - 72.

[6] Dumais S. Improving the Retrieval of Information from External Sources. Behavior Research Methods[J]. Instruments & Computers, 1991, 23(2): 229 - 236.

[7] 冯项云. LSI 潜在语义标引方法在情报检索中的应用[J]. 现代图书情报技术,1998(4):19 - 21.

[8] 顾榕,王小平,曹立明. 一种基于潜在语义分析的查询扩展算法[J]. 计算机工程与应用,2004,40(18):23 - 25,63.

[9] Rehder B, Littman M L, Dumains S T, et al. Automatic 3 - Language Cross - Language Information Retrieval with Latent Semantic Indexing[C]. NIST Special Publication 500 - 240:The Sixth Text Retrieval Conference(Trec - 6),1997:233 - 240.

[10] 林鸿飞,李业丽. 中英文双语交叉过滤的逻辑模型[J]. 计算机工程与应用,2000,36(8):48 - 50.

[11] 杨震,邓贵仕. 基于隐含语义的个性化信息检索[J]. 计算机工程与设计,2003,39(7):90 - 93.

[12] 余正涛,樊孝忠,郭剑毅,等. 基于潜在语义分析的汉语问答系统答案提取[J]. 计算机学报,2006, 29(10):1889 - 1893.

[13] 林鸿飞,丁洪文,杨志豪,等. 基于概念和统计的问答系统实现机制[J]. 大连理工大学学报,2006,46(2):280 - 285.

[14] 刘博勤,丁晓明. 潜语义标引与汉语信息检索研究[J]. 计算机科学,2000,27(3):93 - 95.

[15] 王金凤. 一种基于特征聚合理论和 LSI 的文本分类新方法[J]. 北京理工大学学报(社会科学版), 2004,6(5):16 - 19.

[16] 曾雪强,王明文,陈素芬. 一种基于潜在语义结构的文本分类模型[J]. 华南理工大学学报(自然科学版), 2004(32):99 - 102.

[17] 陈涛,宋研,谢阳群. 基于 IIG 和 LSI 组合特征提取方法的文本聚类研究[J]. 情报学报,2005,24(2):203 - 209.

[18] 李永平,程莉,叶卫国. 基于隐含语义的 kNN 文本分类研究[J]. 计算机工程与应用,2004,40(6):71 - 74.

[19] 何明,冯博琴,傅向华. 基于 Rough 集潜在语义索引的 Web 文档分类[J]. 计算机工程,2004,30(13):3 - 5.

[20] 叶浩,王明文,曾雪强. 基于潜在语义的多类文本分类模型研究[J]. 清华大学学报(自然科学版),2005, 45(S1):1818 - 1822.

[21] 刘云峰,齐欢,代建民,等. 基于潜在语义空间维度特性的多层文档聚类[J]. 清华大学学报(自然科学版),2005,45(S1):1783 - 1786.

[22] 牛伟霞,张永奎. 潜在语义索引方法在信息过滤中的应用[J]. 计算机工程与应用,2001,37(9):57 - 59,62.

[23] 甘勇,陈锁,朱贵良. 基于语义分析的电子邮件过滤系统设计[J]. 微电子学与计算机,2002,19(8):28 - 30.

[24] 杨 清,李方敏. 基于 LSI 和 SVM 分类法的定题邮件过滤研究[J]. 计算机工程与应用,2005,41(32):168 - 171.

[25] 陈华辉. 一种基于潜在语义索引的"垃圾"邮件过滤方法[J]. 计算机应用研究,2000,36(10):17 - 18,35.

[26] 林鸿飞, 高仁璟. 基于潜在语义索引的文本摘要方法[J]. 大连理工大学学报,2001,41(6):744 - 748.

[27] 刘军万,刘飞飞. 基于潜在语义索引的文本结构分析方法的研究[J]. 情报杂志,2004,23(1):56 - 58.

[28] 杨守捷,胡祥恩. 应用潜在语义索引提取信息[J]. 天津大学学报(自然科学版),2002,22(1):48 - 52.

[29] 龚根华,陈恳,万钧. 应用 LSI 实现 Web 图片的索引和查询[J]. 南昌大学学报(理科版),2005,29(4):391 - 395.

[30] 袁磊,曹奎,冯玉才,等. 一种基于 LSI 的图像语义检索技术[J]. 华中科技大学学报(自然科学版),2002,30(2):105 - 107.

[31] 沈玉利,郭雷,任建峰. 基于 LSI 的图像语义检索[J]. 计算机工程与应用,2005,41(22):64 - 65,69.

[32] Bellegarda J R. Exploiting Latent Semantic Information in Statistical Language Modeling[C]. Proceedings of IEEE, 2000(8):1297.

[33] 任纪生,王作英. 一种新的潜在语义分析语言模型[J]. 高技术通讯,2005,15(8):1 - 5.

[34] 郑亚非. 潜在语义分析与篇章理解[J]. 浙江工业大学学报(社会科学版),2006,5(1):70 - 75.

[35] 桂诗春. 潜伏语义分析的理论及其应用[J]. 现代外语,2003,26(1):76 - 84.
[36] 杨守捷,刘曼华. 应用潜在语义分析[J],探析认知科学. 天津大学学报(社会科学版),2001,3(3):238 - 244.
[37] 王慧莉,隋丹妮. 基于潜在语义分析的长时工作记忆在语篇理解中的作用[J]. 北京航空航天大学学报(社会科学版),2005,18(4):94 - 96.
[38] 赵顺,迟呈英. 基于 LSI 和 Rough 集的文本分类研究[J]. 鞍山科技大学学报,2005,28(5):346 - 349.
[39] 王怡,盖杰,武港山,等. 基于潜在语义分析的中文文本层次分类技术[J]. 计算机应用研究,2004,21(8):151 - 154,165.
[40] 林鸿飞,姚天顺. 基于潜在语义索引的文本浏览机制[J]. 中文信息学报,2000,14(5):49 - 56.
[41] 刘云峰,齐欢,代建民. 潜在语义分析在中文信息处理中的应用[J]. 计算机工程与应用,2005,41(3):91 - 93.
[42] 赵明华,游志胜,吕学斌,等. 采用改进得 LSA 模型进行人脸识别[J]. 计算机应用研究,2005,22(10):173 - 174,177.
[43] 盖杰,王怡,武港山. 潜在语义分析理论及其应用[J]. 计算机应用研究,2004,21(3):9 - 12,20.
[44] 李振星,陆大珏,任继成,等. 基于潜在语义索引的 Web 信息预测采集过滤方法[J]. 计算机辅助设计与图形学学报,2006,16(1):142 - 147.
[45] 郭敏,董健全,宋智. 基于 P2P 的隐含予以索引模型的研究[J]. 计算机工程与设计, 2005,26(11):2910 - 2912,2954.

(作者 E - mail:sunyiqin1984@ yahoo. com. cn)

动 态

OCLC 和 Zepheira 联合将重新设计 OCLC 的 PURL 服务

2007 年 7 月 11 日,OCLC 与 Zepheira 公司宣布将展开合作重新构建 OCLC 的 PURL 服务,以更加高效地支持网络数据的管理。

将来开发出来的软件将在开放源代码软件协议下发布,允许 PURL 和 PURL 的基本框架应用于公共或私人的各种应用程序中。PURL 服务在 OCLC 的主持下为世界范围内的图书馆、教育团体、政府、商业、非盈利机构和公民个人提供永久性稳定的 WWW 地址服务已经长达 12 年之久了。PURL 的产生是 OCLC 研究办公室和因特网工程任务组(IETF)统一资源定位符(URL)工作组合作下产生,在动态的变化的网络环境中充当了永久性标识的角色。

PURL 支持为所有的文档和数据命名,用户可以非常简单的参考和共享网络数据中的信息。由于组织的结构、发展方向、优势和机遇的变化都会引起数据产生和所有权归属的变化,因此必须需要一种机制来保证这些信息随之发生变化。PURL 的灵活性和永久性为诸如科研机构、政府机构和图书馆团体等这样重要领域中信息的复用、跨机构和团体的信息交换提供了极大的方便。

Zepheira 是美国语义网技术和企业数据集成方面实力非常强大的公司,它今年将为 OCLC 重新设计和构建新的 PURL 服务,来支持日益增长的 PURL 需求中所面对的灵活性更强以及新特征等方面的需求。新开发出来的服务将会代替 OCLC 主办的位于 purl. org 上的现存服务。

OCLC 和 Zepheira 将来开发出来的新软件将不但为网络化信息资源(如:网络文档)提供永久性标识,也提供对非网络化资源(如:人、组织、概念和科学数据)的永久性标识。另外新软件还会向前迈出重要的一步,那就是提供语义网支持下的网络数据机器处理。更多详情请参考 http://www. purl. org。

(李书宁编译自:http://www. oclc. org/news/releases/200669. htm)

(本刊讯)