基于语义的网络大数据组织与搜索

吴纯青 任沛阁 王小峰
(国防科学技术大学计算机学院 长沙 410073)

摘 要 随着信息技术的飞速发展,网络空间中出现海量异构的数据资源,网络大数据逐渐引起了人们的关注.从网络大数据中发现并获取用户所需的数据资源,需要对网络大数据进行有效地组织管理并进行基于数据语义的相似搜索.为此,需要从网络数据资源中抽取其特征/属性构造高维语义空间,并将数据资源及用户查询信息抽象为语义空间中的特征向量或高维点,进而通过比较特征向量间夹角余弦值或高维点之间的距离来衡量语义相似性.高维索引技术可以对高维语义空间中的数据资源进行有效组织管理,实现基于数据语义的相似性搜索;而降维技术可以消除语义空间维数过高所引发的"维灾"影响.文中对现有的高维数据索引及降维技术进行了系统的综述,然后介绍了现有的基于分布式技术实现高维数据语义相似性搜索的研究工作,最后并展望了未来工作.

关键词 语义空间;高维索引;降维;相似性查询;P2P;大数据
中图法分类号 TP301 **DOI 号** 10.3724/SP.J.1016.2015.00001

Survey on Semantic-Based Organization and Search Technologies for Network Big Data

WU Chun-Qing REN Pei-Ge WANG Xiao-Feng
(*College of Computer, National University of Defense Technology, Changsha* 410073)

Abstract With the development of information technology, massive data resources with heterogeneous structure appear in the cyberspace, which is known as the network big data and has attracted extensive attentions. For mining the useful information from the network big data, it is required to efficiently organize the data resources in the cyberspace and realize the semantic-based similarity search. For an efficient data organization and search, we firstly need to extract the features/attributes of the big data to construct its high-dimensional semantic space, then define the data resources and queries as feature vectors or high-dimensional points in the semantic space, and finally can calculate the semantic similarity by the distance of high-dimensional points or the cosines of the angles between feature vectors. The multidimensional indexes can efficiently organize data resources in the semantic space, realizing the semantic-based similarity search. In addition, the dimensionality reduction technology can avoid the effects of "curse of dimensionality" when the dimensionality of the semantic space is too high. In this paper, the existing multidimensional indexes and dimensionality reduction technologies are reviewed systematically. Moreover, the existing semantic-based similarity search technologies using distributed technology are analyzed, and some suggestions about future research work are discussed.

Keywords semantic space; multidimensional index; dimensionality reduction; similarity search; P2P; big data

收稿日期:2013-05-12;最终修改稿收到日期:2014-04-14.本课题得到国家"八六三"高技术研究发展计划项目基金(2011AA01A103)、国家自然科学基金(61103194)资助.**吴纯青**,女,1964 年生,博士,研究员,博士生导师,主要研究领域为计算机网络与通信.E-mail:wuchunqing@nudt.edu.cn.**任沛阁**(通信作者),男,1985 年生,博士研究生,主要研究方向为分布式网络计算及智能数据处理.E-mail:renpeige@163.com.**王小峰**,男,1982 年生,博士,助理研究员,主要研究方向为可信网络及系统、网络安全及分布智能数据处理.

1 引 言

近年来,随着互联网、物联网、社交网络、云计算等信息技术的发展,网络空间中的数据资源正以前所未有的速度不断地增长和积累,世界已经进入了网络化的大数据(Big Data)时代[1].大数据指在可容忍的时间内用常用软硬件工具无法对其进行有效感知、获取、管理、处理和服务的数据集合[2].IBM、IDC 等权威机构将其特点总结为 4V 特性:规模巨大化(Volume),形式多样化(Variety),生成高速化(Velocity),价值巨大但密度稀疏化(Value).大数据的兴起引起了产业界、学术界及政府机构的高度重视.Microsoft、Google、IBM、Facebook 等国际 IT 巨头广泛应用并推动大数据技术的发展.Jim Gray 提出了数据密集型科学的"第四范式"[3],改变了人们对传统计算科学的看法;而《Nature》与《Science》也相继专刊讨论研究与大数据相关的问题:《Nature》于 2008 年出版专刊"Big Data"[4],介绍了海量数据带来的挑战;《Science》于 2011 年出版"Dealing with Data"专刊[5],讨论了数据洪流所带来的机遇与挑战.此外,美国政府于 2012 年公布了"大数据研发方案"①,该计划被视为美国政府继"信息高速公路"计划后又一重大举措.

网络大数据[6]指"人、机、物"三元世界在网络空间(Cyberspace)中彼此交互与融合所产生并在互联网上可获取的大数据,其特点符合大数据的 4V 特性.首先,网络空间中数据资源飞速增长,数据规模不断扩大,呈现出海量的特性.其次,网络大数据类型丰富多样,呈现出多样化及异构化的特点,网络空间中涌现了大量的结构化数据、半结构化数据及非结构化数据,且非结构化数据的比例不断增长.再次,网络空间中大数据变化更新频繁,常以数据流的形式动态、快速地生成,具有很强的时效性.最后,网络空间中大数据价值巨大但呈现出稀疏性的特点,由于数据量巨大及表现形式多样化,传统的方法技术很难高效发现并获取用户所需的数据资源,实际应用中往往呈现出"数据丰富而知识缺乏"的窘态,网络大数据价值利用密度低.

网络大数据对社会产生了深刻的影响,孕育着巨大的机遇,同时也为有效管理和利用大数据提出了挑战.在当前数据爆炸的大数据时代,针对网络空间中数据规模巨大,形式异构,动态变化,分布广泛等特点,如何高效地组织管理并搜索发现用户所需的数据资源面临着以下 3 个难点问题:

(1)网络空间中数据资源类型繁多,表现形式多样,而形式异构的数据资源可能具有相同或相似的语义信息,传统的基于精确匹配的搜索方法不能有效地获取用户所需的数据资源.为了实现基于语义的智能搜索,需要将网络中海量异构的数据资源统一映射为语义空间中的高维数据,并通过有效手段快速锁定搜索区间,利用相似性搜索方法获取与用户语义相关的数据资源.

(2)随着网络空间中数据资源的日益丰富,语义空间维度急剧增加.当空间维度过高时,在语义空间中实现相似性搜索的性能急剧下降,引发"维度灾难"[7].

(3)由于大数据的 4V 特性,在将网络大数据映射到统一的语义空间并进行基于语义的相似性搜索过程中将占用大量的计算及存储资源.传统的集中式处理方式容易产生性能瓶颈,系统的稳定性及可扩展性较差,不能很好的应对海量的网络数据及复杂的用户需求.

以上 3 个问题可总结为:如何在统一的语义空间中描述网络大数据的语义信息,并对其进行适当的组织划分,实现基于语义的智能搜索发现;如何解决语义空间维度过高所引起的"维灾"问题;如何合理分配数据组织搜索过程中产生的庞大计算及存储开销,提高系统性能.本文对现有工作进行了深入研究,将上述问题的解决方法总结为对应的 3 个方面:高维索引技术、数据降维技术及分布式语义相似性搜索技术,如图 1 所示.高维索引将海量异构的数据资源统一映射到语义空间,并根据给定的用户查询快速确定搜索区间,修剪掉与查询请求语义无关的数据集合,可以实现基于语义的相似性搜索.当语义空间维度过高时,容易引发"维灾"问题,导致高维索

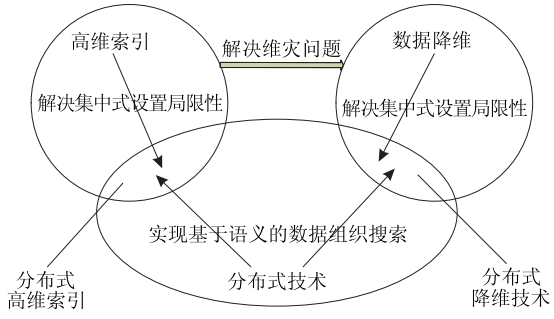

图 1 网络空间中数据组织搜索方法发展趋势

① http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal

引的性能急剧下降，甚至不如顺序扫描. 降维技术通过构造降维映射，获得高维数据集合的低维表示，可以有效消除“维灾”带来的影响. 针对传统集中式设置带来的局限性，分布式高维索引或降维技术可以合理分配计算及存储开销，提高系统的健壮性及可扩展性，有效地应对网络大数据带来的挑战. 本文的贡献主要体现为：

(1) 从向量空间及度量空间角度综述了现有的高维数据索引技术.

(2) 从线性及非线性角度综述了现有高维数据降维方法.

(3) 总结了现有的基于分布式技术的高维数据语义相似性搜索技术，并展望了未来发展方向.

本文第 2 节综述现有的高维数据索引结构；第 3 节系统地介绍现有的高维数据降维方法；第 4 节总结现有的基于分布式技术实现高维数据语义相似性搜索方法；最后一节总结全文并展望未来工作.

2　高维数据索引

高维数据，即高维语义空间中的数据，通常表示为高维点或高维向量. 高维空间内数据通常表现出以下的特点：(1) 分布不均匀. 随着空间维度的增加，数据趋于空间表面分布[8]；(2) 分布稀疏性. 高维数据在空间内分布稀疏，很难对有相似语义的数据信息进行有效的组织管理；(3) 等距性. 由于高维数据趋于空间表面分布，从给定查询点到其近邻点和远邻点的距离随着维度的增加趋于相等[9]；(4) 动态性. 伴随着数据的更新，随时有高维数据进入或离开语义空间；(5) 数据海量性. 语义空间内网络数据量庞大；(6) 时间代价大. 操作高维数据所花费的时间一般远高于传统数据；(7) 不能排序. 无法对高维数据进行有效的线性排序以充分体现其在空间中的相对位置关系.

为了实现高效的基于语义的相似性搜索，学者们提出了多种高维索引结构，用以在相似性搜索过程中修剪掉与给定用户查询语义无关的数据对象，减少搜索空间，缩短查询时间，提高搜索效率. 根据构建高维索引所采用的数据划分标准及相似度量的不同，高维索引可分为向量空间索引结构和度量空间索引结构[10]，如图 2. 两者之间的区别与联系体现为：(1) 向量空间可看作是带有坐标信息的度量空间，在一定条件下可以相互转换. 当在度量空间中只利用一个给定的距离函数获取数据间距离信息时，向量空间则转换成了度量空间；而利用快速映射(FASTMAP)算法可以将度量空间转换为较低维的向量空间；(2) 在相似性查询过程中，度量空间索引仅仅利用基于距离函数的三角不等式性质；而向量空间索引则可以同时利用数据在空间中的位置(坐标)信息. 向量空间索引利用了更多的信息，比度量空间索引具有更好的修剪及搜索效率.

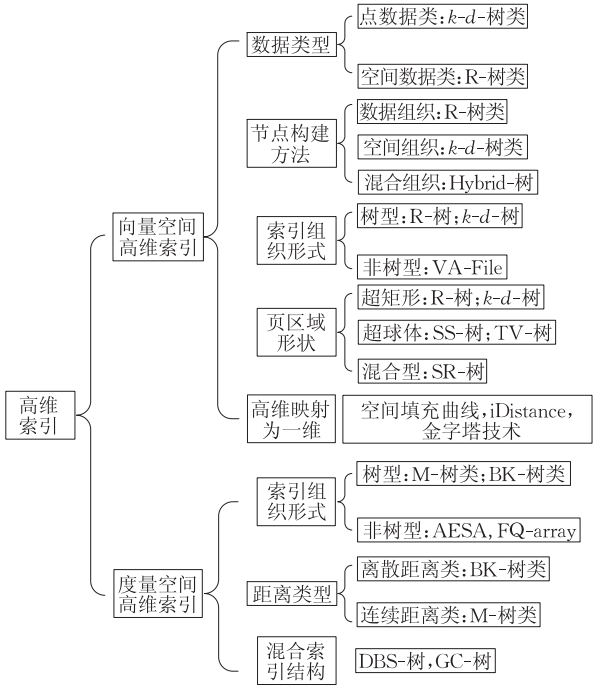

图 2　高维数据索引结构分类

2.1　向量空间高维索引

2.1.1　向量空间高维索引分类

向量空间索引利用数据在空间中的相对位置及距离函数来组织数据，具有不同的分类标准：

(1) 根据所索引数据的类型，可分为点数据类索引和空间数据类索引. 点数据类索引只能处理高维点数据，如 k-d-树[11]；空间数据类索引则可以同时处理点、线及多边形等高维数据，如 R-树[12].

(2) 根据索引节点构建方法的不同，可分为数据组织类、空间组织类以及混合组织类. 数据组织类索引根据数据在空间中的分布组织索引节点，如 R-树；空间组织类将数据空间划分为互相邻接的子空间并将子空间对应为索引节点，如 k-d-树. 混合组织类则同时利用数据组织方法及空间组织方法构建索引结构，如 Hybrid-树[13].

(3) 根据索引组织形式，可分为树型索引和非树型索引. 树型索引中索引节点按照树的形状组织，如 R-树等；非树型索引中索引节点不按照树的形状组织，如 VA-File[14] 等.

(4)向量空间索引将语义上相似(位置接近)的数据聚类为数据页(page),分配到每个数据页的子空间称为页区域(page region).根据页区域的形状,可以分为超矩形、超球体和混合型.页区域为超矩形的索引有 R-树等;超球形的有 SS-树[15]等;混合型的有 SR-树[16]等.

(5)此外,高维映射到一维类索引将高维空间内数据对象按照某种标准映射到一维数据空间,典型的代表有空间填充曲线(Space Filling Curve,SFC)[17]、iDistance[18]方法及金字塔技术[19]等.

2.1.2 典型向量空间高维索引

(1)R-树类.R-树是一种对应空间矩形层次嵌套结构的平衡树,是 B^+-树在空间上的扩展,如图 3 所示.R-树可应用于多维点及空间数据,索引树中的节点对应包含高维数据的最小边界矩形(Minimum Bounding Rectangle,MBR).MBR 是所包含数据的最小近似,且相互之间可以有重叠.R-树具有较好的数据存储效率,但是 MBR 间的重叠导致了数据的重复搜索.

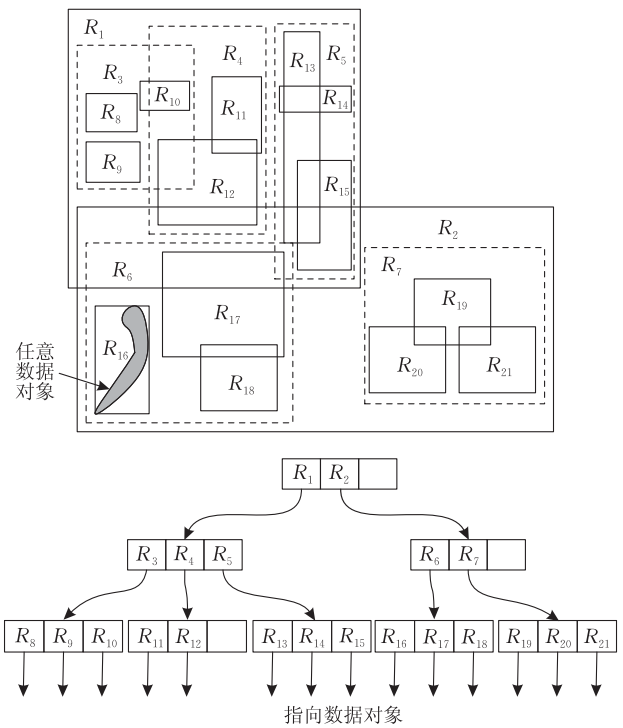

图 3 R-树数据划分及组织示意图

R^*-树[20]是 R-树的扩展,它改进了 R-树的数据插入及节点分裂算法,同时引入了强制重新插入机制,一定程度上提高了搜索性能,但 R^*-树没有有效地降低节点间的重叠程度.R^+-树[21]是针对 R-树中由矩形重叠造成的多重路径搜索问题而提出的,它增加了描述 R-树性能的新特征,并通过分裂的方式使同层节点矩形之间不再存在重叠.X-树[22]基于 R-树引入了超节点的概念,当对溢出的节点执行节点分裂操作而导致较严重的区域重叠时,则扩大溢出的节点使之成为超节点,以便存放更多的数据对象.SS-树利用超球体代替 MBR 作为页面区域,可以有效的减少存储空间,但是在节点分裂过程中不能避免区域重叠现象.SR-树[23]可以看作 R^*-树和 SS-树的结合,同时利用 MBR 及超球体划分高维数据以形成页面区域,从而在更高效的处理相似性查询时避免了较高的区域重叠度.SS^+-树[24]是 SS-树的改进,相较 SS-树,SS^+-树利用更紧密的边界超球体划分高维数据,在执行节点分裂过程中更好的利用了数据的聚类特性,同时利用本地重构规则构建索引树以便减少区域重叠度.CSS-树[25]根据半结构化数据的特点,基于聚类的思想进行索引节点组织及溢出节点的分裂,同时提出了高效的搜索剪枝策略,能有效地提高针对半结构化数据的相似性搜索效率.

(2)*k-d*-树类.*k-d*-树为高维空间中的二叉查找树,可用于点数据相似性搜索.如图 4 所示,*k-d*-树利用超平面分割数据空间,保证了不同节点对应的页面区域间没有重叠,避免了重复搜索.*k-d*-树是基于空间的划分方式,不能充分考虑数据在空间中的实际分布情况,导致很多节点包含很少的数据对象;同时 *k-d*-树高度上不平衡,进一步影响了搜索效率.

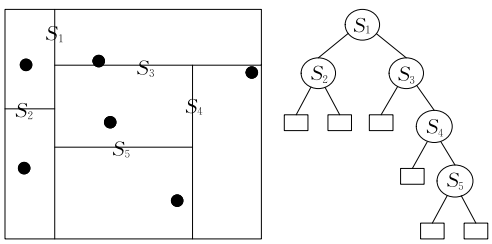

图 4 *k-d*-树空间划分及数据组织示意图

Adaptive *k-d*-树[26]为 *k-d*-树的变种,它在利用超平面分割数据时保证得到的两个子空间内点数据的数目相等,超平面上不一定要包含数据,每个节点内只能包含定量的数据,当溢出时需进行节点分裂.*kd*-B-树[27]采用了 B-树的思想,使其成为平衡的多叉树.当节点溢出时,在节点内部执行强制分割策略,使其变为相邻的子空间.但是这种节点分割方法不能保证最大的空间利用率.

(3)量化近似类.VA-File 首次提出了量化近似的思想,其本质上并不能称为一种索引结构,它提出了利用数据压缩技术以便加速顺序扫描的方法.VA-File 维护两个文件:高维数据的量化近似表示

(以压缩的比特串形式表示)及其精确表示. 空间中的每一维被分配一个比特串近似表示其刻度(比特串的长度决定了近似表示的精度),空间被刻度值划分为大量子区间,空间中的高维点根据其所在子区间的坐标近似表示. 当进行相似性搜索时,首先根据查询点的近似表示进行顺序扫描确定子区间,然后再对子区间内的数据进行精确扫描来确定最终结果.

VA$^+$-File[28]方法是 VA-File 方法的改进,首先使用 Karhunen-Loeve 变换来去除各维之间的相关性,然后根据变换后各维的能量(方差)做不均匀的维数分配. 在保证分配的总维数不变的前提下,能够更精确地估计特征向量;LPC-File[29]同时采用极坐标的方法近似表示高维数据,提高了数据的近似表示精度及相似性搜索精度;VQ-Index[30]用矢量量化替代 VA-File 中的标量量化,有效提高了相似性搜索效率,但矢量量化器需利用历史数据生成. 文献[31]基于高斯混合模型提出了一种矢量量化索引方法,充分利用了各维之间的统计相关性,更加精确的近似表示了高维数据对象. 文献[32]提出了一种并行压缩优先过滤索引 PCPF,通过量化特征向量构建近似向量空间上的高维索引结构,进行空间划分并行构建多个子索引分支,可以显著提升查询匹配效率及精度.

(4)高维映射到一维类. 空间填充利用某种方法对高维空间中的数据近似排序,尽可能确保数据点间的相对位置不变,将高维数据从高维空间映射到一维线性空间,其典型代表有 Z-Order、Hilbert 曲线等,如图 5 所示. 但此类方法灵活性不足,如果对两个不同区域的索引进行组合,至少要对其中一个进行重新编码,且对范围查询的效率较差.

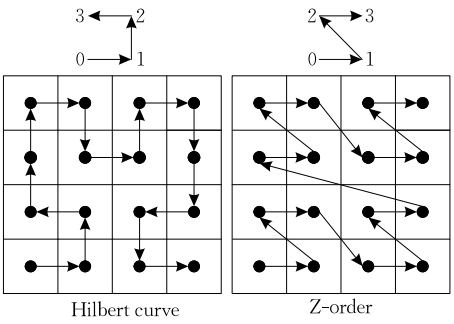

图 5　空间填充曲线示例

iDistance 方法利用数据点相对参考点的距离将空间中高维数据映射到一维空间. 如图 6 所示,该方法首先对空间内高维数据进行聚类,为每一个数据聚集区顺序编号,并为每一个数据聚集区选择一个参考点. 然后根据高维数据所在子空间编号及其到参考点的距离将高维数据映射为一维数值. BC-iDistance[33]通过引入位码来近似表示参考点与数据之间的位置关系,将高维向量压缩为二维向量,并利用特殊的 B$^+$-树组织,查询过程中实现了两层剪枝处理,提高了查询效率. 文献[34]提出了一种基于聚类分解的高维索引,首先基于聚类分解的思想把聚类得到的超球体分解为半径不同的空腔超球体,然后对空腔超球体内数据利用 iDistance 方映射为一维数值,可以有效提高搜索效率.

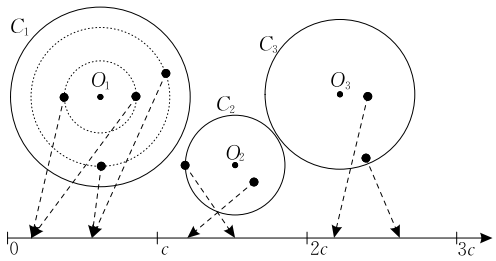

图 6　iDistance 方法

金字塔技术首先将数据空间划分为以中心点为共同顶点的多个高维金字塔,并为每个金字塔确定一个金字塔值,然后根据高维数据所在的金字塔值及其到金字塔底的垂直距离将高维数据映射为一维数值. P$^+$-树[35]首先对高维数据聚类,并将每个数据聚集区规整为超立方体,之后利用金字塔技术的思想索引高维数据,增强了金字塔技术的鲁棒性. iMinMax[36]方法也为一种将高维数据映射为一维数值的索引结构,首先找出某个高维数据所有维度中维度值最大和最小的两个维度,然后根据两个维度值将高维数据映射为一维数据,注意这里的映射也为多对一的映射.

此外,学者们还从不同的角度提出了多种索引结构,如利用超平面划分高维特征空间形成的 Quad-树[37],基于哈希存取方法形成的网格文件类[38]索引结构及 IQ-树[39]等,这里不再详细介绍.

2.2　度量空间高维索引

2.2.1　度量空间高维索引的分类

度量空间高维索引利用满足三角不等式性质的距离函数来实现高维数据的相似性搜索,从不同的角度可以对度量空间高维索引做出不同的分类.

(1)根据索引组织形式,可分为树型索引和非树型索引. 树型索引中索引节点按照树的形式组织,如 BK-树[40]、VP-树[41]等;非树型索引中索引节点不按照树的形状组织,如 AESA[42]等.

(2)根据索引构建过程中所使用的距离函数类型,可分为离散距离类及连续距离类. 离散距离类索

引利用离散距离函数构建索引，距离函数提供有限数量的距离值，如 BK-树；连续距离类利用连续的距离函数构建索引，距离函数提供大量甚至无限的距离值，如 M-树[43]等.

(3)混合索引方法将树状索引与顺序扫描方法相结合，具体实例有 GC-树[44]、DABS-树[45]等.

2.2.2 典型度量空间高维索引

(1) VP-树类. VP-树为一种连续距离类树型索引，采用自上而下的方法来构建，利用数据对象到参考点之间的相对距离修剪查询过程中与给定查询语义无关的数据，不支持数据的更新和删除. VP-树(图 7)递归构造一棵平衡二叉树，首先选择任意数据对象 p 作为根节点(参考点)，同时计算出其他数据到 p 的中值距离 M；然后将与 p 的距离小于 M 的一半数据放入左子树，另一半放入右子树；针对左右子树重复执行上述过程，直到节点中数据数目小于给定的阈值.

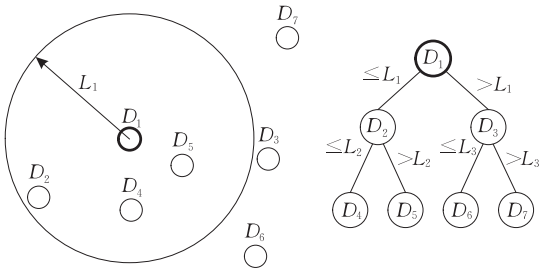

图 7 根节点为 D_1 的 VP-树

针对 VP-树采用二叉树结构所导致的索引高度很高，从而引起大量的距离计算，造成性能下降的问题，MVPT[46]在一个节点中使用多个参考点对数据进行划分，一定程度上提高了搜索效率. VPF[47]将 VP-树的每一层中参考点距离大约为 M 的“中间部分”数据对象挑选出来，并用“中间部分”数据构建第二棵树，以此类推获得具有多棵树的“森林”. 这样当需要搜索“中间部分”数据时，不用同时进入左右子树进行搜索.

(2) BK-树类. BK-树(图 8)属于离散距离类树状索引，索引树的构造过程如下：首先从空间内原始数据集 U 中任意选择一个数据对象 $p\in U$ 作为根节点，然后利用离散距离函数将剩余的数据对象划分为多个子集作为树的分支. 对于每个离散距离 i，定义子集 $U_i=\{u\in U, d(u,p)=i\}$ 为距离根节点为 i 的所有数据元素的集合，并为每一个非空子集 U_i 在 BK-树上建立分支. 然后为所有的非空子集 U_i 循环构造 BK-树直到子集中剩余元素数目为 1 或小于给定的阈值为止，所有将作为子集根节点的数据对象称为支点(pivot).

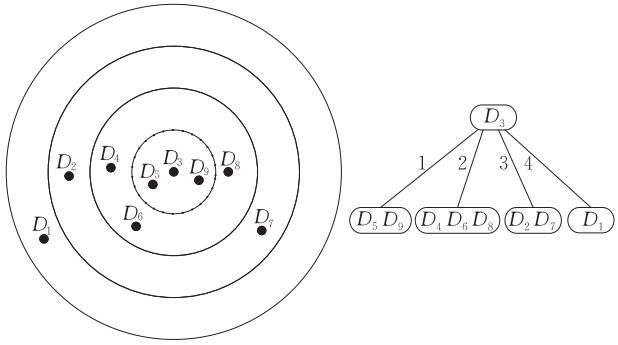

图 8 根节点为 D_3 的 BK-树

FQ-树[48]对 BK-树的结构进行了改进. 与 BK-树不同的是，FQ-树所有的数据元素存储在叶子节点中，且存储在同一层节点中的所有支点是相同的. 该结构可以减少查询时距离比较的次数，代价是树的高度增加. FHQ-树[49]对 FQ-树进一步做了改进——强制所有的叶子节点在同一高度上，这样的设置使我们进行相似性搜索时可以统一考虑同一层的节点，不用额外区分叶子节点和中间节点，提高了搜索效率，但其代价是增加了某些叶子节点的高度. FQA[50]不是树的形式，但具有比 FHQ-树更紧凑的结构. FQA 将 FHQ-树叶子节点中的数据按从左至右的顺序放入一个数组中，并记录下从根节点到数据的高度. 在同样内存的情况下，FQA 可以较 FHQ-树访问更多的支点，从而提高搜索效率.

(3) M-树类. M-树是基于连续距离类的平衡树状索引. 在 M-树中每个节点中首先选取代表元素，然后将靠近代表元素的数据构建成以代表元素为根的子树，代表元素存储其覆盖半径信息. 当进行相似性搜索时，首先比较搜索范围与节点中代表元素的覆盖范围以确定语义相关(范围相交)的子树，然后进入子树中进一步搜索. 当插入数据时，选择距离最近的代表元素对应的子树插入，当引起节点溢出时，需执行分裂操作(图 9). M-树可以支持数据的更新，同时很大程度上减少了相似性搜索过程中的计算量，但是没有处理好节点间的重叠问题.

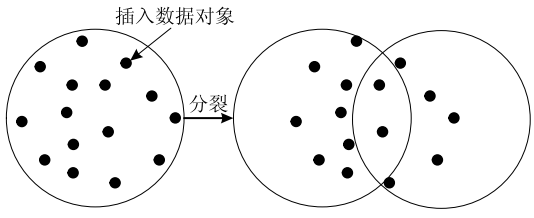

图 9 M-树节点溢出时执行节点分裂

MB^+-树[51]针对节点间的重叠问题，利用 B^+-树和 Block-树索引空间内高维数据，避免了兄弟节

点间的区域重叠. M^2-树[52]使用多种距离函数，可以同时根据数据的多个特征对数据进行检索，实现了高维数据的复杂查询. Slim-树[53]提出了基于最小生成树的节点分裂方法，并通过一个后处理过程最小化索引节点数目及节点间的重叠度. M^+-树[54]提出了一种基于距离及关键维的两步分割策略，根据对相似性影响最大的关键维对数据进行二次过滤，在有效提高数据分割效率的同时降低了索引树的高度. MS-树[55]提出了活动子空间和非活动子空间的概念，并通过一种空间映射方法减少对非活动子空间的访问次数，提高了查询性能.

(4) 混合索引结构. 高维索引能够有效处理聚类或分布密集的数据集，当数据分布特别稀疏时，采用顺序扫描可能更加有效. 混合索引能够区分数据集在空间中的分布状态，针对不同的数据分布针对性的采用顺序扫描或树状索引，能够有效的提高相似性搜索的效率. 如 GC-树首先根据一个密度函数划分数据密集区域及稀疏区域，然后对两者分布采用树状索引及顺序扫描；DABS-树根据数据分布动态调整数据页大小，并通过顺序扫描一个一级目录以解决目录过载的问题；文献[56]对树状索引无法有效处理的数据采用顺序扫描的方法，能够一定程度上解决“维灾”带来的影响. 文献[57]提出一种基于查询采样的高维数据混合索引结构，基于查询采样的方法将分步稀疏的数据从树状索引中分离出来，并对其进行顺序扫描处理，能够有效提高搜索效率.

此外，学者们还提出了多种索引结构，如针对离散距离函数的 BK-树；动态多级树 SA-树[58]；可适用于离散距离函数及连续距离函数的 BS-树[59]等，这里不再详细介绍.

2.3　高维数据索引综合比较

综上，学者们已对高维索引进行了大量研究，提出众多索引结构及其相关算法用以组织管理高维数据，表 1 从数据划分方式、数据节点组织形式、所支持的数据类型及是否支持数据更新等方面对几种典型高维索引结构进行了综合比较.

但是，高维索引并没有在基于语义相似性搜索中得到广泛应用，主要是其性能受到以下因素的制约：(1) 对基于语义的相似性搜索性能不稳定，当搜索范围增大时，获取有用信息的代价迅速上升；(2) 对空间维数较敏感，随着维度的升高高维索引对数据的过滤效果越来越差，当空间维数超过某一“临界值”时，采用高维索引的相似性搜索效率甚至不如顺序扫描；(3) 基于语义相似性搜索容易受到空间“语义无关”维度的影响；(4) 现有的高维索引基本都在某种特定情况下提出(如有的针对聚类数据，有的针对均匀分布数据)，缺乏对数据在语义空间内分布的自适应性，不能很好地满足用户的要求.

表 1　高维索引结构综合比较

高维索引	划分方式	组织形式	数据类型	数据更新
R-树	超矩形	树形	空间数据	支持
SS-树	超球体	树形	空间数据	支持
SR-树	空间矩形＋超球体	树形	空间数据	支持
SS^+-树	超球体＋聚类	树形	空间数据	支持
k-d -树	超平面	树形	点数据	支持
空间填充曲线	保距映射	非树形	点数据	不支持
VA-File	空间划分	非树形	点数据	支持
VP-树	相对距离	树形	点数据	不支持
M-树	相对距离	树形	点数据	支持
iDistance	聚类＋相对距离	树形	点数据	不支持
金字塔技术	空间划分＋相对距离	树形	点数据	不支持
混合索引	树形索引＋顺序扫描	树形	点数据	支持

3　高维语义空间降维

“维度灾难”指在对高维数据组织管理处理过程中遇到的由于数据特征(维度)过多而引起的所有问题，主要表现为以下几个方面：(1) 高维空间数据分布非常稀疏，很难对有相似语义的数据信息进行有效的组织管理，在发现有相似语义的数据信息时，需要访问较大的空间区域；(2) 高维空间中一个给定数据到其最近邻和最远邻的距离在很多情况下几乎是相等的，不能高效地组织和发现与该数据点语义相似的数据信息；(3) 随着维数的升高，高维索引的数据划分效果变差，数据索引节点之间的重叠度随之增大，导致了数据的重复搜索并增加了数据访问路径，从而影响了搜索效率.

此外，高维索引搜索效率受语义空间维度及数据对象内在维度(Intrinsic Dimension)的影响，当在高维空间处理内在维度很低的数据对象时(如在 1000 维空间内查找 2 维平面)，由于受到与内在维度无关维度的影响，导致搜索效率很低. 降维方法是解决“维灾”问题的有效手段，通过有效手段将数据从高维空间映射到低维空间，同时尽可能保持数据集的整体结构和分布不变，从而获得高维数据的一个有意义的低维表示，进而降低基于语义的相似性搜索算法的复杂性，提高搜索效率，如图 10.

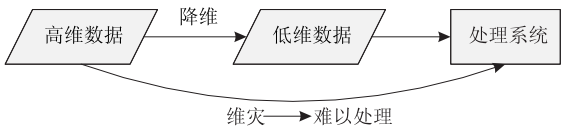

图 10　降维过程

降维方法从数学角度的描述如下：设 $X=\{x_i\}_{i=1}^N \subset R^D$ 为 D 维语义空间 R^D 内数据元素个数为 N 的一个数据集，X 的内在维度为 $d(d \ll D)$，即数据集 X 内的数据元素属于嵌于 D 维语义空间 R^D 内一个维度为 d 的子空间. 降维技术通过找到合适的降维映射 $F: X \rightarrow Y$，将数据集 X 映射为一个新的维度为 d 的数据集 Y，$Y=\{y_i\}_{i=1}^N \subset R^d$，同时尽量保证原数据集的拓扑结构不变，从而可以在较低维空间内对数据对象进行处理，有效地消除了维灾的影响.

目前学者们已经陆续提出了多种降维方法，如潜在语义索引[60]、主成分分析[61]、多维尺度分析[62]以及近年来提出的基于流形学习[63]的算法等. 从不同角度可以对降维方法作如下分类：从降维映射形式角度可分为线性及非线性降维；根据降维过程中是否使用数据中的监督信息可分为无监督降维，有监督降维及半监督降维；从操作数据集范围角度可分为全局方法和局部方法；根据降维过程中特征获取途径可分为特征选取和特征抽取两类. 本文根据降维映射形式将现有降维技术分为线性降维和非线性降维，如图 11.

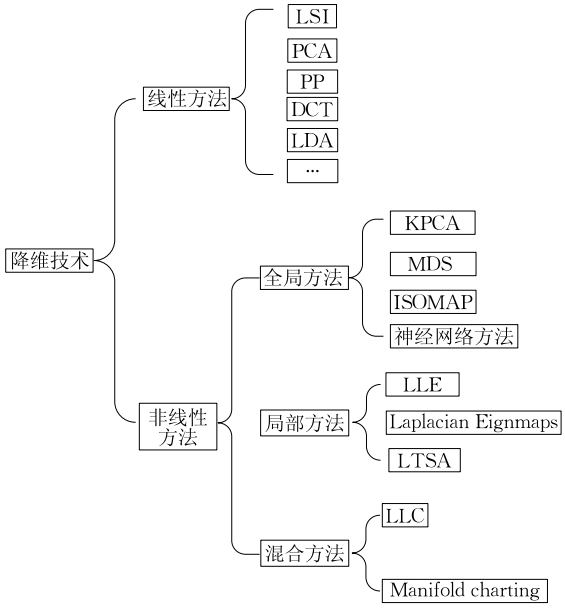

图 11 降维技术分类

形式上，设 $\{x_i\}_{i=1}^N \subset R^D$ 是 D 维空间内数据集合，通过降维映射

$$F(x)=\begin{pmatrix} F_1(x_1) \\ F_2(x_2) \\ \vdots \\ F_d(x_n) \end{pmatrix}=\begin{pmatrix} F_1(x_{11},x_{12},\cdots,x_{1D}) \\ F_2(x_{21},x_{22},\cdots,x_{2D}) \\ \vdots \\ F_d(x_{n1},x_{n2},\cdots,x_{nD}) \end{pmatrix}$$

得到较低维空间 R^d 中的数据集 Y，$Y=\{y_i\}_{i=1}^N \subset R^d$. 若 F 的每个分量 F_i 都只是 X 的线性函数，则称 F 为线性降维；否则，称 F 为非线性降维.

3.1 线性降维方法

线性降维技术以其简单、快速、易于实现、不存在局部极值以及相对有效性的特点得到了十分广泛的应用. 以下介绍几种典型的线性降维方法.

(1) 潜在语义索引(Latent Semantic Index，LSI). LSI 基于向量空间模型(Vector Space Model，VSM)提出. VSM 将高维数据对象描述为特征向量. LSI 对空间内高维数据组成的特征矩阵进行奇异值分解，并舍弃后面绝对值很小的奇异值，进而将高维数据映射到低维空间，同时去除了语义相似性查询过程中的噪声干扰.

LSI 通过矩阵奇异值分解并人为设置阈值舍弃较小的奇异值，阈值的选取将直接影响到搜索的精度；同时由于矩阵分解计算量较大，当数据规模较大或维数过高时，LSI 效率变低. 针对 LSI 不能处理大规模数据的问题，RLSI[64]将根据数据对象的不同“主题”将“特征-对象”矩阵表示为“特征-主题”矩阵及“主题-对象”矩阵的乘积，然后对分解后的矩阵进行并行处理，可以有效提高系统的可扩展性及相似性搜索效率.

(2) 主成分分析(Principal Components Analysis，PCA). PCA 方法将数据映射到一组新变量(主成分)上，并通过设置数据方差的阈值舍弃方差较少的主成分以达到降维的目的. PCA 为一种坐标变换技术，新得到的维度(主成分)为原维度的线性组合，并尽可能反映数据原有的信息. 但在降维过程中，主成分个数需人工选取，选取不当将导致信息丢失.

(3) 投影寻踪(Projection Pursuit，PP). 投影寻踪[65]可以有效地分析和处理服从非正态分布的高维数据. 它基于某种线性组合将高维数据投影到低维子空间中，并寻找能很好地反映原数据特征的最佳投影方向，从而能在低维空间对数据进行分析处理. PP 方法降维过程中由于有大量的点积计算，当空间维度过高时降维效率变差，比较适用于数据量大但维度较低的数据集.

(4) 离散余弦变换(Discrete Cosine Transform，DCT). DCT[66]的基本思想是在一定损失的情况下，采用较少数目的维数来表示高维数据. 其基本原理为对于 D 维数据 $\xi=(\xi_0,\xi_1,\cdots,\xi_{D-1})$，令 $\varepsilon=(\varepsilon_0,\varepsilon_1,\cdots,\varepsilon_{D-1})$ 为 ξ 的 DCT 结果，若 ε 的第 m 项之后均为 0(或绝对值相对首项很小)，则只保留前 m 项，实现

了数据的降维. 高维数据经 DCT 处理后具有保距性，不影响数据间的拓扑结构. 实际操作过程中需要人为控制舍弃部分的阈值，需要根据所能容忍的精度及系统效率综合考虑.

(5) 线性判别分析(Linear Discriminant Analysis，LDA). LDA[67] 也称为 Fisher 判别分析，其主要思想是寻找一个投影矩阵，将高维数据投影到低维空间，同时 Fisher 准则确保了在低维空间中同类数据的区别最小化，异类数据的区别最大化，提高了不同类别之间数据的可分性. LDA 是一种有监督降维方法，可应用于现实中高维数据的分类工作. 但是 LDA 不能灵活调整分解矩阵的大小，当类别中心有重叠时分类效果较差.

3.2　非线性降维方法

现实生活中的真实数据集更多的体现为非线性结构，线性降维方法不能在降维过程中很好的保持数据集的非线性特性. 为弥补线性降维方法的不足，学者们提出了非线性降维方法. 非线性降维方法可以通过对线性降维方法进行非线性扩展(通常采用核技巧或局部方法)来获得，也可以对相应的函数采用神经网络或直接采用最优化方法(如梯度下降法)等来获得. 流行学习是当前流行的一种非线性降维方法，其目标是发现嵌入在高维特征空间中的低维流形结构，并给出一个有效的低维表示. 流行学习的典型代表有等距映射 ISOMAP(Isometric Mapping)[68]、局部线性嵌入 LLE 算法[69]、Laplacian 特征映射[70] 等. 非线性降维方法可分为三种类型(图 3)：(1)全局方法. 降维过程中保留数据集的全局属性；(2)局部方法. 降维过程中保留数据集的局部属性；(3)混合方法. 通过局部线性模型的全局排列实现降维.

3.2.1　全局方法

全局非线性降维技术在降维过程中保留高维数据的全局属性. 典型的全局方法有核主成分分析[71]、等距映射、多维尺度分析及神经网络方法.

(1) 核主成分分析(Kernel PCA，KPCA). KPCA 是 PCA 针对非线性数据集的“核化版本”，其基本思想为将非线性数据集通过核方法映射到一个更高维的特征空间，然后利用传统 PCA 方法实现数据降维. 更高维的特征空间可以通过非线性映射 $\phi:\Phi^N\rightarrow F$ 得到，由于通过 ϕ 映射后的两个点的点积可以用核函数 $k(x,x')=(\phi(x)\cdot\phi(x'))$ 来估计，因此该方法称为核 PCA. KPCA 的性能依赖于核函数的选择，而实际应用中核函数往往无法有效的选择.

(2) 多维尺度分析(Multi-Dimensional Scaling，MDS). MDS 可分为度量 MDS 和非度量 MDS. 度量 MDS 确保降维后低维数据点之间的距离与原数据点之间的距离尽量保持一致，可以尽可能保留数据对象间的相似性. 度量 MDS 方法具有明显的几何意义，降维过程中距离函数的选择具有灵活性. 非度量 MDS[72] 旨在确保降维前后数据对象间顺序关系的一致性，而非数据对象间的相对距离.

(3) 等距映射(Isometric Mapping，ISOMAP). ISOMAP 方法的出发点与度量 MDS 相似，即在保持数据点间相对距离一致的情况下实现数据降维. 但 ISOMAP 采用“测地距离”表示数据对象间的相似性，而度量 MDS 通常选用“欧式距离”. ISOMAP 方法用邻域图(Neighborhood Graph)来表达数据的邻域结构，根据流形的局部欧氏性质，邻域图上的每一条边都可以采用欧氏距离来进行表达；然后用邻域图中的最短路径长度来对未知的全局测地距离进行逼近；最后以逼近得到的这些全局测地距离作为输入运行古典 MDS 算法，将数据重建在其内在的低维参数空间中，可以发现数据内在的全局几何结构及其有意义的内在维. 但是由于估计测地线距离的不精确性，ISOMAP 在未采样的空间可能会得到较差的结果；同时 ISOMAP 具有拓扑不稳定性，可能在邻域图中产生错误的连接；此外，ISOMAP 不产生内在模型，而当其他类似数据需要降维时，必须重新训练.

(4) 神经网络方法. 基于神经网络的降维方法利用神经网络的特性去除高维数据集的冗余特征，将其映射到低维空间，其典型代表有多层自动编码器(Multilayer Auto-encoder)[73]、自组织特征映射(Self-Organizing Map，SOM)[74]、生成建模(Generative Modelling)[75] 等方法.

多层自动编码器为一种具有奇数个隐含层的前向反馈网络，从功能上可分为编码和解码两个网络. 编码网络和解码网络存在一个重合的“瓶颈”层，该层具有最少的神经元，且能够反映高维数据集的本质维数，进而实现数据降维.

SOM 网络分为输入层，竞争层和输出层，基于数据聚类实现数据降维. 竞争层中邻近的各个神经元通过彼此侧向交互作用，相互竞争学习，最终在输出神经元层排列成一张低维的映射图，从而实现数据的降维. 映射图中语义相似的神经元距离较近，语义不相似的神经元则距离较远.

3.2.2 局部方法

局部非线性降维方法基于局部特征保持的思想,即仅仅考虑数据样本附近数据的位置关系,代表方法有局部线性嵌入(Local Linear Embedding,LLE)[76]、Laplacian 特征映射、Hessian LLE[77]等.

局部线性嵌入使人们开始更加关注数据集所蕴含的内蕴特征.其基本思想是在降维过程中尽可能保持邻近点的局部拓扑结构,确保相邻的高维数据在低维空间内同样保持相邻关系.LLE 首先进行近邻搜索,确定高维数据的邻居;然后计算每个高维点的权重,用邻居数据的权值组合表示高维数据;最后,基于本征向量优化技术寻找高维数据的低维嵌入,确保每个数据对象的权重不变.LLE 对流形中短路现象不太敏感,同时克服了局部极小的影响;但对非均匀数据区域表现较差,同样不产生内部模型.

3.2.3 混合方法

混合方法通过全局排列局部线性模型实现降维,具体实现为:首先计算一系列局部线性模型,然后对线性模型进行全局排列.该方法典型代表有局部线性协同(Locally Linear Coordination,LLC)[78]方法及 Manifold Charting[79]方法.

LLC 方法通过计算一些局部线性模型并对这些局部模型执行一个全局排列来实现降维.该方法分为两步:首先通过期望最大化(Expectation Maximization)算法计算一系列关于高维数据的局部线性模型;然后利用扩展的 LLE 方法排列调整这些局部线性模型以得到原数据集的低维表示.LLC 的降维计算代价较 LLE 或 ISOMAP 有了明显的降低;但 LLC 易受到数据集中异常值的影响,同时期望最大化算法易陷入对数似然函数的局部最大值.

此外,学者们还提出了多种降维技术,如线性降维方面的因子分析(Factor Analysis,FA)[80]、独立成分分析(Independent Components Analysis,ICA)[81]等方法;非线性降维方面的主曲线流形[82],最大方差展开(Maximum Variance Unfolding,MVU)[83]等方法,这里不再详细介绍.

3.3 降维技术性能分析

对于某个高维数据集 $\{x_n\}_{n=1}^{N} \subset R^D$ 原始维度为 D,数据集内数据个数为 N,降维后得到低维空间 R^d,d 为目标维数.涉及的参数有邻域图(Neighborhood Graph)中近邻点个数 k,迭代次数 i,稀疏矩阵中非零元素的比率 p,LLC 方法中局部线性模型的数目 m,神经网络中权值的数量 w.表 2 分析了几种典型降维方法的性能.

表 2 降维技术性能分析

降维技术	参数	计算复杂度	内存复杂度
LSI	*none*	$O(d^2D)$	$O(d^2D)$
PCA	*none*	$O(D^3)$	$O(D^2)$
PP	i	$\geq O(ND^2)$	$O(D^2)$
LDA	*none*	$O(D^3)$	$O(D^2)$
KPCA	$k(x,x')$	$O(DN^3)$	$O(N^2)$
MDS	*none*	$O(N^3)$	$O(N^2)$
ISOMAP	k	$O(N^3)$	$O(N^2)$
Autoencoder	*net size*	$O(iNw)$	$O(w)$
LLE	k	$O(DN\log N)+O(pN^2)$	$O(pN^2)$
LLC	k,m	$O(imd^3)$	$O(Nmd)$

现有降维技术为快速有效地处理高维数据提供了一定的便利,但仍存在以下方面的不足:

(1)现有降维方法处理人造数据效果理想,但对现实生活中的实际高维数据集不能实现有效降维.

(2)现有流行学习方法仅实现了位于高维流形上有限数据集的低维表示(点与点的嵌入),但并未建立高维流形空间与对应低维表示空间之间的相互映射关系(集合与集合的映射),这使其无法获得一个新输入高维(或低维)空间数据在对应低维表示空间(或高维流形空间)的映射表示,进而限制了流行学习方法在应用上的扩展.

(3)现有降维方法缺乏自适应能力,对动态增加或变化的数据对象不能实现快速有效的降维,提高降维方法的自适应性成为亟待解决的问题.

4 分布式语义搜索

如上所述,在网络空间实行基于语义的相似性搜索过程中,高维索引技术可以修剪掉大量无用的搜索路径,提高语义空间中的搜索效率;降维技术可以将高维数据映射到低维空间而保留其语义信息近似不变,消除"维灾"的影响并去除语义空间内的噪声干扰.

但是传统的基于集中式设置的网络数据语义搜索机制[84-85]可能引发以下问题:健壮性不高,存在单点失败问题;扩展性较差,用户数量的多少会影响系统性能;信息垄断,服务器对网络中信息数据及搜索过程全权控制,可以随意干预搜索结果;安全隐患,可以对用户行为进行跟踪,容易造成隐私泄露.P2P 作为典型的分布式技术,具有高度的自治性、可扩展性、匿名性、健壮性、负载均衡性及安全性,可以解决集中式设置带来的局限性.

近年来,学者们为了提高基于语义相似性搜索

的性能，分别提出了分布式高维索引技术及分布式降维技术.

4.1　分布式高维索引技术

分布式多维索引综合了多维索引及分布式技术的优点，在语义空间中进行语义相似性搜索过程中通过高维索引技术修剪掉大量无关的搜索路径，并将计算搜索任务交由网络中所有参与者分担，实现了系统的可扩展性并克服了单点失败等问题.

(1) 基于 P2P 实现 iDistance

M-Chord[86]方法将 iDistance 与 P2P 相结合实现了语义空间中的相似性搜索. 该方法首先对高维语义空间内的数据进行聚类，然后利用 iDistance 方法将高维数据对象映射到一个$[0,2m)$的一维关键字区域，保证不同的聚类间的高维数据映射到一维数轴上不同的区间内，确保网络中每个 peer 负责一个区间，并通过维护一个 B^+-tree 来实现数据插入，删除及检索功能. 在相似搜索过程中，首先确定查询的搜索区间，然后在区间内精确搜索得出结果.

R-Chord[87]与 M-Chord 相似，实现了 Chord 与 iDistance 的结合. 区别是 R-Chord 定义了相对位置码(Relative Position Code, RPC)用以实现搜索过程中的进一步数据过滤. 利用 RPC 可以过滤掉下界距离比修剪距离大的数据对象，减少距离计算，提高搜索效果.

MCAN[88]结合 CAN 及 iDistance 技术用来实现语义空间中的相似搜索. 它利用一种基于支点的技术为语义空间中的每个数据对象关联一个坐标，将数据对象映射到一个 N 维向量空间，并将映射后的对象分布于 CAN 网络内，将度量空间中的相似搜索转变为 CAN 空间内的搜索.

SIMPEER[89]. M-Chord 及 MCAN 方法数据预处理阶段(聚类及映射)仍以集中的方式处理，SIMPEER 则完全实现了整个查询过程的分布式处理. SIMPEER 采用一种三层聚类机制(peer, super-peer, routing clusters)：每个 peer 聚类其自身数据，并用 iDistance 索引；super-peer 管理所负责 peer 的聚类信息并利用扩展的 iDistance 方法计算 hyper-clusters，hyper-clusters 用以决定哪个 peer 处理 super-peer 接收到的查询请求；Hyper-clusters 信息在 super-peer 间交互并进一步汇总产生一系列路由簇(Routing Clusters)，路由簇信息保存在 super-peer 层，用来在 super-peer 网络中路由一个查询请求.

SiMPSON[90]也为在高维语义空间中实现相似搜索的一种 P2P 索引结构. 在 SiMPSON 中，任何支持一维范围查询的结构式 P2P 系统可作为其底层拓扑. 首先每个 peer 局部聚类本地数据，然后基于 iDistance 方法将聚类范围映射为 2 个一维索引值(开始索引和结束索引)，并利用一维索引值缩小搜索区间，从而减少搜索代价.

(2) 基于 P2P 实现 R-树

P2PR-树[91]将 P2P 中 peer 组织成分层的树形覆盖网，且将每个 peer 负责的空间数据表示为 peerMBR. 系统中每个 peer 维护一条到根的路径，查询请求必须经过根节点才能达到目的节点. P2PR-树为不平衡树，如果数据是倾斜的，一些 peer 必须维护很长的路径信息，进而会降低搜索性能.

NR-树[92]是 R^*-树的分布式结构. peers 分为 super-peers 及 passive-peers. super-peers 管理一定数量的 passive-peers 形成一个簇. 如果 passive-peers 数量超过一个阈值，簇依据 CAN 的方法执行分裂. super-peers 构成一个 CAN 覆盖网来实现彼此间的通信. 在每个簇中，super-peers 负责构建分布式 R^*-树，当一个 passive-peers 发送一个查询请求，首先将其发送到 super-peers，然后 super-peers 将请求发送给可能包含答案的 passive-peers.

P2PRdNN-树[93]结构与 NR-树类似，唯一区别为 super-peers 间利用主通道以广播的方式传递消息. 但是 super-peers 间以广播形式传递消息将导致系统较高的通信代价.

RT-CAN[94]以云计算为研究背景，将 CAN 及 R-树技术相结合用以研究云系统中索引高维数据的问题. 系统为每个 peer 分配两个角色：存储节点及覆盖节点. 存储节点维护一部分数据并为本地数据构建一个 R-树. 覆盖节点为 CAN 网络中的一个节点，负责 CAN 的一个分区. RT-CAN 的发布策略为：对于一个将要发布的 R-树节点 N，如果其半径小于给定阈值，则将其发送到包含 N 中心的所有簇节点；否则所有与 N 重叠的 CAN 节点在其全局索引中维护 N. 对于一个窗口查询：首先将查询请求路由到包含查询窗口中心的 CAN 节点，然后将该查询递归发送到所有与查询窗口重叠的邻居，存储节点如果收到查询请求则搜索本地索引并返回结果.

(3) 基于 P2P 实现 KD-树

DiST[95]为一种 KD-树与 CAN 技术相结合的分布式多维索引结构. DiST 利用 KD-树划分方法划分语义空间，CAN 网络中每个 peer 负责一个子

空间并创建本地索引.全局索引分布于所有的节点中,每个 peer 仅有局部索引,并根据局部索引路由查询请求.

DKDT[96]将 Chord 与 KD-树相结合用以解决 P2P 系统中的相似搜索问题.该方法首席循环基于 KD-树划分语义空间直到子空间内数据数目不高于给定的阈值,然后利用一个 hash 函数将每个子空间映射到底层覆盖网络.Peers 利用 Chord 协议维护映射到环上的子空间,同时以 DHT 方式及 KD-树方式维护 peers.每个 peer 维护一个称为 Tree Information Base(TIB)的局部数据库.由于 Chord 及 KD-树相互并不匹配,因此在 DKDT 中维护两个结构将导致较大的维护成本.对于一个范围查询,查询发起者利用 TIB 决定树中覆盖查询点的最低节点(peer),然后发送查询到该 peer,接收到查询的 peer 不仅搜索本地数据而且根据其 TIB 将查询转发到其他 peers 进一步搜索.对于 KNN 查询,在范围查询的基础上查询发起者通过维护一个优先队列用以获取最优结果.

m-LIGHT[97]与 DKDT 类似,利用 KD-树划分空间,并为 KD-树中每个叶子节点(子空间)分配一个关键字.通过关键字,每个叶子节点被底层 DHT 中的一个 peer 管理.m-LIGHT 搜索机制与 DKDT 方法类似,其不同之处是 m-LIGHT 在插入,查询及删除数据的过程中利用一个数据感知分裂(data-aware splitting)策略来确保系统负载平衡.

SkipIndex[98]将 SkipGraph 及 KD-树相结合以解决高维索引问题.该方法利用 KD-树划分空间,然后根据分裂记录,KD-树的每个叶子节点(数据子空间)分配一个由 0/1 组成的位串作为该节点的关键字,实现了对子空间的编码;然后根据关键字将叶子节点与 SkipGraph 相关联,则数据的搜索,插入及删除操作可以依据 SkipGraph 协议实施.

此外,学者们还提出了多种不同的基于 P2P 的高维索引技术,如 Squid[99]、Z-NET[100]等基于 P2P 实现的空间填充曲线技术,LINP[101]等基于 P2P 实现的 VA-file 技术等.

4.2 分布式降维技术

分布式降维方法综合了分布式技术及降维技术的优点,可以有效消除"维灾"及集中式架构带来的影响,近年来逐渐引起关注.

(1)基于 P2P 实现 LSI/VSM

pSearch[102]将 VSM 和 LSI 技术应用到 CAN 网络中,其基本思想是为查询请求及所有文档对象建立特征向量,并将其映射到相应的 CAN 地址空间,与查询向量语义相似文档向量在邻近的 CAN 地址空间,查询只需要在限定的邻近节点进行,有效提高了搜索效率.但这种方法存在几个局限性:首先该方法基于 CAN 网络实现,无法适用于其他结构的 P2P 网络;其次由于网络环境中数据维度较高且处于动态变化中,CAN 由于维度的限制不能很好满足有效搜索需求;最后,CAN 对于语义空间的平均划分策略容易导致节点上的负载不均衡.

文献[103]提出一种分级语义覆盖网络,通过创建与网络节点相关联的语义簇来实现语义搜索.分级语义覆盖网络基于 DHT 的思想,利用 LSI 及余弦相似性将语义上相近的节点聚集成簇,能够同时支持基于关键字匹配和基于语义的相似性搜索.相对 pSearch 方法,分级语义覆盖网络可以更好的支持生僻词的查询且查询效率更高.

文献[104]将 LSI 技术与非结构化 P2P 技术结合,实现了语义查询.该方法将网络中的节点分为 super-peer 和 peer,并采用一种三层构造机制:每个 peer 建立自己的词条文档矩阵 $\boldsymbol{A}$;super-peer 将其所负责的 peer 上的资源整合为新的语义词典矩阵 $\boldsymbol{T}$;super-peer 间相互交互信息构造路由矩阵 $\boldsymbol{S}$.当某个 peer 提出查询请求 q,首先将 q 提交给 super-peer,super-peer 通过 $\boldsymbol{S}$ 计算出可能含有结果的若干个 super-peer′,然后 super-peer′根据 $\boldsymbol{T}$ 查询到具体的 peer′,最后确定的 peer′在 $\boldsymbol{A}$ 中查找到与 q 语义匹配的资源并返回给 peer.

FCAN[105]方法通过对关键字向量表示的文本建立索引,并将索引发布到 CAN 网络中,利用索引信息和 CAN 在同一空间的特点,实现结构化 P2P 系统中基于文本内容的查询.FCAN 通过 FastMap 实现了语义空间到 P2P 空间的映射,在保证语义完整性的同时有效地对语义空间进行了降维.

(2)SSW(Semantic Small World)

SSW[106]采用的方法是降低叠加网络的维数,将语义相近的数据对象和 P2P 节点自组织成多个语义簇,然后通过自适应空间线性化方法将高维空间上的语义簇线性化到一维语义小世界上.这种方法克服了 pSearch 中存在的负载过大的弊端,但在降维过程中语义簇丢失了大量相似的语义信息,导致系统准确性降低,同时存在语义空间与 P2P 叠加网络空间维数不匹配的问题.文献[107]提出了一种基于 SSW 在非结构 P2P 上实现语义相似性搜索的,每个节点维护着一系列与其语义相似的节点连

接，可以有效减少搜索过程中的通信代价.

4.3　现有分布式语义搜索技术综合比较

表 3 从空间划分方法、底层拓扑结构、支持的相似性查询方式及是否支持数据更新等方面对几种典型的分布式语义相似性搜索技术进行了综合比较，它们在一定程度上实现了信息语义的相似性搜索，但其实现技术及特点各异.

表 3　分布式语义搜索技术综合比较

方法	空间划分	拓扑结构	查询方式	数据更新
M-Chord	聚类	Chord	范围查询＋KNN 查询	否
R-Chord	聚类＋RPC	Chord	范围查询＋KNN 查询	否
MCAN	CAN 划分	CAN	范围查询	否
SIMPEER	聚类	Super-peer based	范围查询＋KNN 查询	否
P2PR-tree	R-tree	树形	窗口查询	否
NR-tree	*k-d*-tree	CAN＋树形	窗口查询＋KNN 查询	支持
RT-CAN	CAN 划分	CAN	窗口查询＋KNN 查询	支持
DiST	*k-d*-tree	CAN	窗口查询	支持
DKDT	*k-d*-tree	DHT＋树形	范围查询＋KNN 查询	否
pSearch	CAN 划分	CAN	范围查询	支持
SSW	聚类＋分裂	SSW	范围查询	支持

本节分别从基于 P2P 的高维索引及降维技术两方面介绍了实现网络数据语义相似性搜索的方法.这两类方法克服了集中式方法的弊端，提高了系统的健壮性及可扩展性，具有广阔的发展前景.但由于网络空间中数据资源的多样性、异构性、海量性、动态性等特点，导致很难有效地构建网络数据资源的语义空间；另一方面，由于现有高维索引及降维技术本身固有的局限性，导致现有方法不能很好满足人们日益增长的信息需求.

5　结论及未来展望

网络空间中数据资源的爆炸式增长为当今社会带来了宝贵机遇，同时也为如何有效利用网络大数据提出了巨大挑战.在网络大数据 4*V* 特性的背景下，本文研究如何有效组织管理网络空间中的数据资源并实现基于数据语义的相似性搜索，分别从高维数据索引、高维数据降维及分布式语义搜索技术三方面详细回顾总结了现有研究成果，并提出了现有工作在网络大数据背景下的局限性.总体来说，对于网络大数据的研究还处于起步阶段，尚有许多问题亟待解决，将来研究可以重点关注：

(1) 如何准确提取各种网络大数据资源的语义特征信息，确保构造的高维语义空间能够准确反映网络数据资源的语义信息，尽量避免噪声等因素的干扰.

(2) 针对网络大数据资源的语义空间维数高，动态增长，数据分布不规则等特点，提出先进的降维方法以适应网络大数据的特点，在保持数据资源语义不丢失情况下，尽可能降低语义空间维数，消除“维灾”的影响.

(3) 将降维技术与高维索引技术有效结合，在降维的基础上利用高维索引技术，修剪掉大量与给定搜索无关的搜索路径，快速确定搜索范围，进一步提高基于语义相似搜索的速度及精度.

(4) 深化分布式语义搜索系统的研究，针对非结构化 P2P 及结构化 P2P 固有的优缺点，考虑对两者进行综合，取长补短，尽可能减少数据语义搜索过程中所需流量，克服负载不均衡，免费搭乘等不足，进一步提高系统的性能.

(5) 将现有成熟的信息检索技术与基于语义的相似性搜索技术相结合，实现网络空间内智能高效的数据组织及搜索发现.

致　谢　在此，感谢第二届中国互联网学术年会，并向对本文提出宝贵意见的计算机学报李刚老师及各位评审专家表示衷心的感谢！

参　考　文　献

[1] Lohr S. The age of big data. New York Times, 2012, 11

[2] Li Guo-Jie, Cheng Xue-Qi. Research status and scientific thinking of big data. Bulletin of Chinese Academy of Sciences, 2012, 27(6): 647-657(in Chinese)
(李国杰，程学旗. 大数据研究：未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考. 中国科学院院刊，2012，27(6)：647-657)

[3] Anthony J G Hey. The Fourth Paradigm: Data-intensive Scientific Discovery. Microsoft Research, 2009

[4] Big data. Nature, 2008, 455(7209): 1-136

[5] Dealing with data. Science, 2011, 331(6018): 639-806

[6] Wang Yuan-Zhou, Jin Xiao-Long, Cheng Xue-Qi. Network big data: Present and future. Chinese Journal of Computers, 2013, 36(6): 1125-1138(in Chinese)
(王元卓，靳小龙，程学旗. 网络大数据：现状与展望. 计算机学报，2013，36(6)：1125-1138)

[7] Powell W B. Approximate Dynamic Programming: Solving the Curses of Dimensionality. United States: John Wiley & Sons, 2007

[8] Böhm C. Searching in high-dimensional spaces—Index structures for improving the performance of multimedia databases. ACM Computing Surveys, 2001, 33(3): 322-373

[9] Katayama N, Satoh R. The SR-tree: An index structure for high dimensional nearest neighbor queries. ACM SIGMOD Record, 1997, 26(2): 369-380

[10] Liu Fang-Jie, Dong Dao-Guo, Xue Xiang-Yang. Review of high dimensional index structures in metric spaces. Computer Science, 2003, 30(7): 64-68(in Chinese)
(刘芳洁，董道国，薛向阳. 度量空间中高维索引结构回顾. 计算机科学，2003，30(7)：64-68)

[11] Zhou K, Hou Q, Wang R, et al. Real-time KD-tree construction on graphics hardware. ACM Transactions on Graphics, 2008, 27(5): 126

[12] Arge L, Berg M D, Haverkort H, et al. The priority R-tree: A practically efficient and worst-case optimal R-tree. ACM Transactions on Algorithms (TALG), 2008, 4(1): 9

[13] Lee C, Ran B, Yang F, et al. A hybrid tree approach to modeling alternate route choice behavior with online information. Journal of Intelligent Transportation Systems, 2010, 14(4): 209-219

[14] Kriegel HP, Kröger P, et al. Efficient query processing in arbitrary subspaces using vector approximations//Proceedings of the 18th International Conference on Scientific and Statistical Database Management (SSDBM'06). Vienna, Austria, 2006: 184-190.

[15] Yang L, Huang X, Lv R, et al. Performance of SS-tree with slim-down and reinsertion algorithm//Proceedings of the Measuring Technology and Mechatronics Automation (ICMTMA). Changsha, China, 2010: 883-886

[16] Ciferri R R, Salgado A C, et al. A performance comparison among the traditional R-trees, the hilbert R-tree and the SR-tree//Proceedings of the XXIII International Conference of the Chilean Computer Science Society (SCCC'03). Chillan, Chile, 2003: 3-12

[17] Ban X, Goswami M, Zeng W, et al. Topology dependent space filling curves for sensor networks and applications//Proceedings of the INFOCOM. Turin, Italy, 2013: 2166-2174

[18] Qu L, Chen Y, Yang X. iDistance based interactive visual surveillance retrieval algorithm//Proceedings of the Intelligent Computation Technology and Automation (ICICTA). Changsha, China, 2008: 71-75

[19] Berchtold S, Böhm C, et al. The pyramid-technique: Towards breaking the curse of dimensionality. ACM SIGMOD Record, 1998, 27(2): 142-153

[20] Kriegel H P, Schneider R, Seeger B, Beckmann N. The R*-tree: An efficient and robust access method for points and rectangles. ACM SIGMOD Record, 1990, 19(2): 322-331

[21] Sellis T, Roussopoulos N, Faloutsos C. The R+-Tree: A dynamic index for multi-dimensional objects//Proceedings of the 13th VLDB Conference. Brighton, UK, 1987: 507-518

[22] Cai C, Yu T H, Mitra S K, et al. Stack X-tree image coding //Proceedings of the 2000 IEEE Asia-Pacific Conference on Circuits and Systems. Tianjin, China, 2000: 727-730

[23] Katayama N, Satoh S. The SR-tree: An index structure for high-dimensional nearest neighbor queries. ACM SIGMOD Record, 1997, 26(2): 369-380

[24] Kurniawati R, Jin J S. The SS+-tree: An improved index structure for similarity searches in a high-dimensional feature space//Proceedings of the International Society for Optics and Photonics. San Jose, USA, 1997, 3022: 110-120

[25] Yang Jian-Wu, Chen Xiao-Ou. An index structure of semi-structure data set for similarity search. Chinese Journal of Computers, 2002, 25(11): 1219-1226(in Chinese)
(杨建武，陈晓鸥. 半结构化数据相似搜索的索引技术研究. 计算机学报，2002，25(11)：1219-1226)

[26] Greenspan M, Yurick M. Approximate *k-d* tree search for efficient ICP//Proceedings of the 4th International Conference on 3-D Digital Imaging and Modeling. Banff, Canada, 2003: 442-448

[27] Lin H Y, Huang P W. Perfect KDB-tree: A compact KDB-tree structure for indexing multidimensional data//Proceedings of the 3rd International Conference on Information Technology and Applications. Sydney, Australia, 2005, 2: 411-414

[28] Ferhatosmanoglu H, Tuncel E, Agrawal D. Vector approximation based indexing for non-uniform high dimensional data sets//Proceedings of the 9th International Conference on Information and Knowledge Management. McLean, USA, 2000: 202-209

[29] Cha G H, Zhu X M, Petkovic P, Chung C W. An efficient indexing method for nearest neighbor searches in high-dirnensional image databases. IEEE Transactions on Multimedia, 2002, 4(1): 76-87

[30] Tuncle E, Ferhatosmanoglu H, et al. VQ-index: An index structure for similarity searching in multimedia databases// Proceedings of the 10th ACM International Conference on Multimedia. Riviera, French, 2002: 543-552

[31] Ye Hang-Jun, Xu Guang-You. Fast image search using vector quantization. Journal of Software, 2004, 15(5): 712-719(in Chinese)
(叶航军，徐光祐. 基于矢量量化的快速图像检索. 软件学报，2004，15(5)：712-719)

[32] Chen Hui-Zhong, Chen Yong-Guang, Jing Ning, Chen Luo. PCPF: A parallel index for matching the high-dimensional vectors in multimedia databases. Chinese Journal of Computers, 2011, 34(10): 2009-2017(in Chinese)
(陈慧中，陈永光，景宁，陈荦. PCPF：一种面向多媒体数据库中高维向量匹配的并行索引结构. 计算机学报，2011，34(10)：2009-2017)

[33] Liang Jun-Jie, Feng Yu-Cai. BC-iDistance: Bit-code based optimal high-dimensional index. Journal of Chinese Computer Systems, 2007, 28(9): 1647-1651(in Chinese)
(梁俊杰，冯玉才. BC-iDistance：基于位码的优化高维索引. 小型微型计算机系统，2007，28(9)：1647-1651)

[34] Zhang Jun-Qi, Zhou Xiang-Dong, Wang Mei, Shi Bai-Le. Cluster splitting based high dimensional metric space index B+-Tree. Journal of Software, 2008, 19(6): 1401-1412(in Chinese)
(张军旗，周向东，王梅，施伯乐. 基于聚类分解的高维度量空间索引 B+-Tree. 软件学报，2008，19(6)：1401-1412)

[35] Zhang R, Ooi B C, Tan K L. Making the pyramid technique robust to query types and workloads//Proceedings of the 20th International Conference on Data Engineering. Boston, USA, 2004: 313-324

[36] Ooi B C, Tan K L, Yu C, et al. Indexing the edges—A simple and yet efficient approach to high-dimensional indexing//Proceedings of the 9th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Dallas, USA, 2000: 166-174

[37] Mathew R, Taubman D S. Quad-tree motion modeling with leaf merging. IEEE Transactions on Circuits and Systems for Video Technology, 2010, 20(10): 1331-1345

[38] Nievergelt J, Hinterberger H, Sevcik K. The grid file: An adaptable, symmetric multikey file structure. ACM Transactions on Database Systems, 1984, 9(1): 38-71

[39] Berchtold S, Bohm C, et al. Independent quantization: An index compression technique for high-dimensional data spaces//Proceedings of the 16th International Conference on Data Engineering. San Diego, USA, 2000: 577-588

[40] Sebastian T B, Kimia B B. Metric-based shape retrieval in large databases//Proceedings of the 16th International Conference on Pattern Recognition. Quebec City, Canada, 2002, 3: 291-296

[41] Fu A W, Chan P M, Cheung Y L, Moon Y S. Dynamic VP-tree indexing for N-nearest neighbor search given pair-wise distances. The VLDB Journal, 2000, 9(2): 154-173

[42] Vidal E. New formulation and improvements of the nearest-neighbor approximating and eliminating search algorithm (AESA). Pattern Recognition Letters, 1994, 15(1): 1-7

[43] Viet H H, Anh D T. M-tree as an index structure for time series data//Proceedings of the Computing, Management and Telecommunications. Ho Chi Minh City, Vietnam, 2013: 146-151

[44] Cha G H, Chung C W. The GC-tree: A high dimensional index structure for similarity search in image databases. IEEE Transactions on Multimedia, 2002, 4(2): 235-247

[45] Böhm C, Kriegel H P. Dynamically optimizing high-dimensional index structures//Proceedings of the 7th International Conference on Extending Database Technology. Konstanz, Germany, 2000: 36-50

[46] Bozkaya T, Ozsoyoglu M. Distance-based indexing for high-dimensional metric spaces. ACM SIGMOD Record, 1997, 26(2): 357-368

[47] Xu W, Thompson L P, Miranker D P. Empirical evaluation of excluded middle vantage point forest on biological sequences workload//Proceedings of the 1st Workshop on New Trends in Similarity Search. Uppsala, Sweden, 2011: 26-31

[48] Chávez E, Navarro G, Baeza-Yates R, et al. Searching in metric spaces. ACM Computing Surveys, 2001, 33(3): 273-321

[49] Chávez E, Navarro G. A compact space decomposition for effective metric indexing. Pattern Recognition Letters, 2005, 26(9): 1363-1376

[50] Chávez E, Marroquín J L, Navarro G. Fixed queries array: A fast and economical data structure for proximity searching. Multimedia Tools and Applications, 2001, 14(2): 113-135

[51] Yang Q, Vellaikal A. MB^+-tree: A new index structure for multimedia databases//Proceedings of the International Workshop on Multi-Media Database Management Systems. New York, USA, 1995: 151-158

[52] Ciaccia P, Patella M. The M^2-tree: Processing complex multi-feature queries with just one index//Proceedings of the 1st DELOS Network of Excellence Workshop on Information Seeking, Searching and Querying in Digital Libraries. Zurich, Switzerland, 2000

[53] Açar E, Arslan S, Yazici A, et al. Slim-tree and BitMatrix index structures in image retrieval system using MPEG-7 descriptors//Proceedings of the International Workshop on Content-Based Multimedia Indexing. London, UK, 2008: 402-409

[54] Zhou Xiang-Min, Wang Guo-Ren. Key dimension based high-dimensional data partition strategy. Journal of Software, 2004, 15(9): 1361-1374(in Chinese)
(周项敏，王国仁. 基于关键维的高维空间划分策略. 软件学报，2004，15(9)：1361-1374)

[55] Wang Guo-Ren, Huang Jian-Mei, Wang Bin, et al. A high dimensional data indexing technique based on max gap space mapping. Journal of Software, 2007, 18(6): 1419-1428(in Chinese)
(王国仁，黄健美，王斌等. 基于最大间隙空间映射的高维数据索引技术. 软件学报，2007，18(6)：1419-1428)

[56] Chavez E, Herrera N, Reyes N. Spatial approximation + sequential scan = efficient metric indexing//Proceedings of the 24th International Conference of the Chilean Computer Science Society. Arica, Chile, 2004: 121-129

[57] Zhang Jun-Qi, Zhou Xiang-Dong, Shi Bai-Le. High dimensional hybrid index based on query sampling. Chinese Journal of Computers, 2008, 19(8): 2054-2065(in Chinese)
(张军旗，周向东，施伯乐. 基于查询采样的高维数据混合索引结构. 计算机学报，2008，19(8)：2054-2065)

[58] Hjaltason G R, Samet H. Improved search heuristics for the sa-tree. Pattern Recognition Letters, 2003, 24(15): 2785-2795

[59] Kalantari I, McDonald G. A data structure and an algorithm for the nearest point problem. IEEE Transactions on Software Engineering, 1983, SE-9(5): 631-634

[60] Cai D, Guo D, Ji D. Research on optimize technology in latent semantic indexing based on semantic block//Proceedings of the 2009 Chinese Conference on Pattern Recognition. Nanjing, China, 2009: 1-5

[61] Elmansouri R, Elbeqqali O, Ziyati E. Normed principal components analysis: A new approach to data warehouse fragmentation//Proceedings of the 2013 ACS International Conference on Computer Systems and Applications (AICCSA). Ifrane, Morocco, 2013: 1-4

[62] Mackute-Varoneckiene A, Zilinskas A. Multidimensional scaling: Multi-objective optimization approach//Proceedings

of the International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing. Rousse, Bulgaria, 2009: 60

[63] Lin T, Zha H. Riemannian manifold learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(5): 796-809

[64] Wang Q, Xu J, Li H, Craswell N. Regularized latent semantic indexing//Proceedings of the SIGIR'11. Beijing, China, 2011: 685-694

[65] Gao Mao-Ting, Wang Zheng-Ou. A new algorithm for text clustering based on projection pursuit//Proceedings of the 2007 International Conference on Machine Learning and Cybernetics. Hong Kong, China, 2007, 6: 3401-3405

[66] Wu Q, McGinnity T M, et al. Spiking neural network performs discrete cosine transform for visual images// Proceedings of the Intelligent Computing 5th International Conference on Emerging Intelligent Computing Technology and Applications. Ulsan, Korea, 2009: 21-29

[67] Izenman A J. Linear Discriminant Analysis. New York: Springer, 2008

[68] Balasubramanian M, Schwartz E L. The isomap algorithm and topological stability. Science, 2002, 295(5552): 7

[69] Roweis S T, Saull K. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000, 290(5500): 2323-2326

[70] Belkin M, Niyogi P. Convergence of Laplacian eigenmaps// Proceedings of the 20th Annual Conference on the Neural Information Processing Systems. Vancouver, Canada, 2006: 129-136

[71] Hoffmann H. Kernel PCA for novelty detection. Pattern Recognition, 2007, 40(3): 863-874

[72] Cox T F, Cox M A A. Multidimensional Scaling. USA: CRC Press, 2000

[73] Souza M R P, Almeida L R, et al. Combining distances through an auto-encoder network to verify signatures// Proceedings of the 10th Brazilian Symposium on Neural Networks. Salvador, Brazil, 2008: 63-68

[74] Kohonen T. The self-organizing map. Neurocomputing, 1998, 21(1): 1-6

[75] Dai Z, Lucke J. Autonomous cleaning of corrupted scanned documents — A generative modeling approach//Proceedings of the Computer Vision and Pattern Recognition. Providence, USA, 2012: 3338-3345

[76] Roweis S T, Saull K. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000, 290(5500): 2323-2326

[77] Wang J. Hessian locally linear embedding. Geometric Structure of High-Dimensional Data and Dimensionality Reduction. Berlin Heidelberg: Springer, 2011: 249-265

[78] Teh Y W, Roweis S. Automatic alignment of hidden representations. Advances in Neural Information Processing Systems. USA: The MIT Press, 2002, 15: 841-848

[79] Brand M. Charting a manifold. Advances in Neural Information Processing Systems. USA: The MIT Press, 2002, 15: 985-992

[80] Hofmann T, Hartmann D. Collaborative filtering with privacy via factor analysis//Proceedings of the 2005 ACM Symposium on Applied Computing. Santa Fe, USA, 2005: 791-795

[81] Vasilescu M A O, Terzopoulos D. Multilinear independent components analysis//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Los Alamitos, USA, 2005, 1: 547-553

[82] Zhang Z, Zha H. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. Journal of Shanghai University (English Edition), 2004, 8(4): 406-424

[83] Wang J. Maximum variance unfolding. Geometric Structure of High-Dimensional Data and Dimensionality Reduction. Berlin Heidelberg: Springer, 2011: 181-202

[84] Hjaltason G R, Samet H. Index-driven similarity search in metric spaces. ACM Transactions on Database Systems, 2003, 28(4): 517-580

[85] Zhang C, Xiao W, Tang D, et al. P2P-based multidimensional indexing methods: A survey. Journal of Systems and Software, 2011, 84(12): 2348-2362

[86] Novak D, Zezula P. M-Chord: A scalable distributed similarity search structure//Proceedings of the 1st International Conference on Scalable Information Systems. Hong Kong, China, 2006: 19

[87] Yin W, Zhu M, Jiang L. R-chord: A distributed similarity retrieval system with RPCID//Proceedings of the IEEE International Conference on Network Infrastructure and Digital Content. Beijing, China, 2009: 236-241

[88] Falchi F, Gennaro C, Zezula P. A content—Addressable network for similarity search in metric spaces//Proceedings of the Databases, Information Systems, and Peer-to-Peer Computing. Vienna, Austria, 2007: 126-137

[89] Doulkeridis C, Vlachou A, Kotidis Y, et al. Peer-to-peer similarity search in metric spaces//Proceedings of the 33rd International Conference on Very Large Data Bases. Vienna, Austria, 2007: 986-997

[90] Vu Q H, Lupu M, Wu S. SiMPSON: Efficient similarity search in metric spaces over P2P structured overlay networks //Proceedings of the 15th International Euro-Par Conference on Parallel Computing. Delft, Netherland, 2009: 498-510

[91] Mondal A, Lifu Y, et al. P2PR-tree: An R-tree-based spatial index for peer-to-peer environments//Proceedings of the Current Trends in Database Technology Workshops. Heraklion, Greece, 2005: 516-525

[92] Liu B, Lee W C, Lee D L. Supporting complex multi-dimensional queries in P2P systems//Proceedings of the International Conference on Distributed Computing Systems. Columbus, USA, 2005: 155-164

[93] Chen D H, Zhou J J, Le J J. Reverse nearest neighbor search in peer-to-peer systems//Proceedings of the 7th International Conference on Flexible Query Answering Systems. Milan, Italy, 2006: 87-96

[94] Wang J, Wu S, et al. Indexing multi-dimensional data in a cloud system//Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data. Indianapolis,

USA，2010：591-602

[95] Nam B，Sussman A. DiST：Fully decentralized indexing for querying distributed multidimensional datasets//Proceedings of the 20th International Parallel and Distributed Processing Symposium. Rhodes Island，Greece，2006：10

[96] Gao J，Steenkiste P. Efficient support for similarity searches in DHT-based peer-to-peer systems//Proceedings of the IEEE International Conference on Communications. Glasgow，UK，2007：1867-1874

[97] Tang Y，Xu J，et al. m-LIGHT：Indexing multi-dimensional data over DHTs//Proceedings of the 29th IEEE International Conference on Distributed Computing Systems. Montreal，Canada，2009：191-198

[98] Zhang C，Krishnamurthy A，Wang R Y. SkipIndex：Towards a scalable peer-to-peer index service for high dimensional data. Princeton University，New Jersey，USA：Technical Report TR-703-04，2004

[99] Schmidt C，Parashar M. Flexible information discovery in decentralized distributed systems//Proceedings of the 12th IEEE International Symposium on High Performance Distributed Computing. Seattle，USA，2003：226-235

[100] Shu Y，Ooi B C，et al. Supporting multi-dimensional range queries in peer-to-peer systems//Proceedings of the 15th International Conference on Peer-to-Peer Computing. Konstanz，Germany，2005：173-180

[101] Cui B，Qian W N，et al. LINP：Supporting similarity search in unstructured peer-to-peer networks//Proceedings of the 9th Asia-Pacific Web Conference. Huang Shan，China，2007：127-135

[102] Tang C，Xu Z，et al. Peer-to-peer information retrieval using self-organizing semantic overlay networks//Proceedings of the 2003 Conference on Applications，Technologies，Architectures，and Protocols for Computer Communications. Karlsruhe，Germany，2003：175-186

[103] Kim S，Strassner J，et al. Semantic overlay network for Peer-to-Peer hybrid information search and retrieval//Proceedings of the 2011 IFIP/IEEE International Symposium on Integrated Network Management. Dublin，Ireland，2011：430-437

[104] Shen H T，Shu Y F，et al. Efficient semantic-based content search in P2P network. IEEE Transactions on Knowledge and Data Engineering，2004，16(7)：813-826

[105] Wang J，Yang S B，et al. FCAN：A structured P2P system based on content query//Proceedings of the 5th International Conference on Grid and Cooperative Computing. Hunan，China，2006：113-120

[106] Li M，Lee W C，et al. Semantic small world：An overlay network for peer-to-peer search//Proceedings of the 12th IEEE International Conference on Network Protocols. Los Angeles，USA，2004：228-238

[107] Jin H，Ning X，Chen H. Efficient search for peer-to-peer information retrieval using semantic small world//Proceedings of the 15th International Conference on World Wide Web. Edinburgh，UK，2006：1003-1004

WU Chun-Qing，born in 1964，Ph. D.，professor，Ph. D. supervisor. Her main research interests include computer network and communication.

REN Pei-Ge，born in 1985，Ph. D. candidate. His main interests include distributed network computing and intelligent data processing.

WANG Xiao-Feng，born in 1982，Ph. D.，assistant professor. His main interests include trustworthy network systems，network security，distributed and intelligent data processing.

Background

With the development of network big data technology，how to mind the useful data resources from massive data resources have attracted increasing attention. It is a great challenge to effectively organize the massive and heterogeneous data resources and realize the semantic-based similarity search in the cyberspace. Multidimensional indexes can organize the massive and heterogeneous data resources in a high-dimensional semantic space，realizing the semantic-based similarity search；while the dimensionality reduction technology can avoid the effects of "curses of dimensionality". Hence，there is an increasing demand to classify and analyze the existing multidimensional index and dimensionality reduction technologies. The paper reviews the existing multidimensional index and dimensionality reduction technologies systematically. In addition，the existing semantic-based similarity search technologies using P2P methods are analyzed.

This research is supported by the National High Technology Research and Development Program (863 Program) of China under Grant No. 2011AA01A103；the National Natural Science Foundation of China under Grant No. 61103194.