维科号 (2026-01-07)

Zilliz 出海业务负责人乔丹：向量数据库破研发瓶颈，AI 赋能范本转移｜2025 极新 AIGC 峰会演讲实录

2025 年 12 月 26 日，【想象·2025 极新 AIGC 峰会】在上海浦东浦软大厦成功召开。Zilliz 出海业务负责人乔丹先生在会上做了题为《向量数据库对研发范本转移的影响》的演讲，从非结构化数据特点、大模型幻觉解决到向量技术应用场景，深入解析了向量数据库如何重构 AI 研发的底层逻辑。

Zilliz 出海业务负责人 乔丹

乔丹重点提到以下几点：

“AI 业务中，非结构化数据向量化，是目前最为常见且成熟的数据处理手段之一。”

“幻觉有多种表现形式，如在日常生活中我们能直观感知到的，就是 AI 产出了错误的答案。”

“万物皆可向量化，”

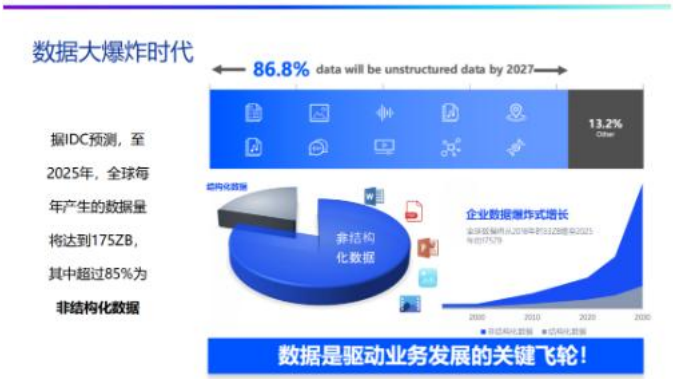
以下为乔丹演讲原文，经极新整理，希望能给大家带来收获。

01 数据治理挑战

“非结构化数据其实都是可以通过向量来进行表征的”

首先我们如果要给它一个简单的定义，除了传统标量形式（比如一个字段一串字符）之外，视频、音频、图片这类数据，我们定义为非结构化数据，而这些非结构化数据其实都可以通过向量来进行表征。

我们试想，每天接收的各种信息中，除了文字数据，很多都是通过视频、音频等形式获取的。其实非结构化数据在我们生活中的占比远比各位想象的要高，这张饼状图可以很好地揭示了我们日常信息收集中的信息占比，非结构化数据显然处在相对主导的位置。当然在计算机领域，或者在数据治理领域，非结构化数据的应用其实还处于方兴未艾的早期状态。



我们的使命就是专注于解决非结构化数据相关的问题。这里我们做一个简单的数学理解，结合最早的解析几何知识，我们可以把生活中很多事物标定为二维、三维乃至无数维坐标系中的一个点。

现在以三维为例，比如有两个单词，“面包” 和“bread”。“面包”可在向量空间中用一组特征向量（如 xyz123）表征，而在传统关键词搜索中，很难直接将“面包”与“bread”匹配——传统搜索仅能匹配“面”“包”这类字面重合的关键词，无法感知二者的语义关联。但如果把它们映射到几何框架中，“面包”是 123，“bread”是 124，在向量空间中，我们可以计算它们之间的相对几何关系和距离，进而得到二者的相关性。

这就是为什么我们可以用一种简单的几何方法，将以前无法匹配和关联的非结构化数据关联起来。当然这只是一个简单例证，如果我们能把这些维度进行百倍、千倍甚至万倍的拓展，一串几何字符所能囊括的信息会远超我们的想象。

02 模型可靠性危机

“幻觉有多种表现形式，在日常生活中我们能直观感知到的，就是它产出了错误的答案”

某知名厂商的大模型，之前的能力可以通过一个问题来验证：单词 school books 有几个 o？这是个很简单的问题，但之前一些版本的大模型给出的回答是有两个，这显然和人眼观察的实际情况不符，正确答案应该是有 4 个 o。

这种情况不只是国内存在，海外也一样。还会自作聪明地补充了这些字母分别出现在哪些位置，但它给出的位置也是错误的佐证。

不过如果追加提问进行纠正，模型有时候是能够反省的。在纠正之下，模型会再进行一次计算，最终得出正确的答案。

这种现象叫什么？有个很专业的名词，叫 Hallucination，也就是幻觉。这其实是个非常哲学化的概念，当我们把大模型当作一个交流对象时，它给出的那些并非是基于事实的回答，而是幻觉。

幻觉可以有很多种表现形式，但在日常生活中我们能直观感知到的，就是它产出了错误的答案。这些其实都是很小的问题，但试想如果使用者是一名学者，正在进行严谨的学术研究，2023 年我们用旧版本模型做了一次简单测试，没有任何上下文，直接提问：上海市 GDP 排名前三的是哪个区？ 模型给出的答案是浦东新区、武汉新区、杨浦区。先不管浦东新区和杨浦区是不是前三，我们能确定的是，武汉新区根本不属于上海，这显然也是出现了幻觉。

但此时我们该如何克服这种现象？其实这就涉及到技术领域老生常谈的方法“检索增强生成”，也就是我们俗称的 RAG。简而言之，我们会在操作中

针对性弥补这一弊端，方法很简单：在提出问题的同时，人为插入一个知识库，为大模型提供对应数据（比如上海下属各区的实际 GDP 数据），随之而来大模型给出的回答就是正确的。这就是一个非常简单的 RAG 雏形，能帮助大家在使用大模型处理文档或生活中的问题时，既利用它的优势，又避免它对真实信息的干扰。

但同时，有些场景下并不会这么顺利，因为我们可能没有现成的知识库，这时候该怎么做？答案也很简单：需要在给大模型的提示词（prompt）中加上“如果没有答案就不要瞎编”的要求。当大模型接收到这个信息后，比如面对“上海市 GDP 排名第三的区是哪个”这类问题，若现有知识库信息无法判断，它就会如实回应，还会给出一些相关性解释，总而言之，它最终不会给出误导性的结论式表达，避免对实际生活中的操作产生重大偏差影响。

如果不想纠结复杂的 IT 概念，可以简单理解：当我们把这类优化措施封装在后台，以及封装在用户端或业务端的各个交互环节时，就产生了各种各样的 RAG 演化和变种，这也是我们现在强调的 AI 在终端或业务端创新的重要方面。

03 技术应用瓶颈

“万物皆可向量化”

2022 年的时候，还有很多人把大量的经济成本以及团队精力投入到发掘创意上面，但显然模型的增长能力对我们而言是比较有挑战性的。而 RAG 能让我们以一种相对轻度、便捷的方式，解决很多切实的业务问题。
那么在这个环境中，向量数据库起到什么作用？可以理解为，在与大模型的沟通当中，所有语言内容的底层其实都不是一串规则化的标量，而是语义化的向量，语义即向量。所以当你要大规模地为大模型插入知识库时，其底层依托的其实就是向量数据库。

由此我们可以产生一个应用场景迁移的思考：向量数据库会在哪些方面起到作用？横向上，在搜索、推荐系统、大模型、风控等场景都能发挥作用，横轴可以无限延伸；纵向上则对应非结构化数据类型，这类数据其实都可以被向量化。两者交叉会产生无数的应用场景赋能。

这些能力其实都是日常可以用到的。比如大家在 A 电商平台进行购物，你觉得某样东西特别贵，去 B 电商平台拍张照搜索，会发现同款商品价格比其他地方便宜 90%。这是怎么实现的？其实就是把两张图片的向量特征提取出来，再进行比对，计算它们在坐标系里的某种算法下的最合适的近邻关系，我们就找到了最具性价比的商品。

这是商业场景的应用，刚才也提到了分子药研发，我们可以把分子结构进行向量化。我们服务的客户里也有材料类型的企业，甚至在自动驾驶领域，随着越来越多的多模态方案出现，相关技术如何辨别不同数据之间的差异，都可以借助向量数据库来实现。