# Abstract

用户偏好在市场分析中发挥重要作用。在数据库文献中，已经对查询原语进行了广泛的工作，例如可以根据客户所表达的偏好用于产品排名的众所周知的top-k查询。然而，评估产品之间的相似性的基本操作通常是忽略这些偏好而完成的。相反，产品根据其属性在特征空间中描绘，并且通过该空间上的传统距离度量来计算相似度。在这项工作中，我们会根据客户的意见来利用产品的排名，以便将产品映射到以用户为中心的相似度计算空间中。我们识别这种映射的重要属性，导致上下相似性界限，这反过来又允许我们在原始产品空间上使用传统的多维索引，以便执行这些以用户为中心的相似度计算。我们展示了可以有效地执行由常用范围和最近邻查询激励的有趣的相似度计算，同时基于以产品的用户为中心的相似度导出的边界修剪数据集的重要部分。

# 1.INTRODUCTION

对象之间相似度的估计是数据管理中的基本操作。例如，它用于通过Web [1]查找具有相似单词的页面或文档，或者根据他们购买的产品来检测具有异常行为的客户[2]。此外，可以执行相似性计算，用于检测社交网络用户之间的类似对话和评论（即，在Facebook上的评论，Twitter上的推文）[3]。已经提出了许多不同的相似性度量来评估两个数据项之间的相似性，例如欧氏距离和余弦相似度。这样的指标表明，数据项之间的相似性是根据其属性来计算的，而不考虑用户的意见。例如，在业务分析中，产品被表示为由其属性值定义的点。根据所选择的度量，两个产品越接近，它们越相似.

在我们的工作中，我们引入了一个补充以用户为中心的相似度计算方法，其中考虑了用户的偏好。例如，与竞争对手现有产品相比，业务经理想知道其业务产品对客户的影响。她非常重要的是知道哪些产品属于许多不同客户的喜爱的列表。这些知识可以用于专注于具有类似的客户群体的产品，这些客户根据他们的喜好将他们排在高位。然后，可以建立更有效的营销策略，创造出更适合特定客户的产品群。

因此，业务经理应该能够执行一个查询，返回的产品（甚至是从未被评为的产品），这些产品不仅基于他们的特征而且基于用户的偏好。她需要通过用户的眼睛看到数据，同时根据可用的用户偏好进行相似度计算。在我们的框架中，用户的偏好被表示为项目属性的加权因子的向量。为了进行这样的相似度计算，我们利用一种称为反向top-k查询的查询类型[4]。与返回具有特定客户最佳分数的k个产品的top-k查询相反，反向top-k查询的结果是给定产品属于其顶部k集的客户集合。我们的工作也适用于最近在市场上推出的产品或产品处于制造过程的设计阶段，并没有对用户表达意见。

## 实例1

考虑业务分析的一个例子，我们希望根据客户对他们表达的偏好来估计产品之间的相似性。为客户对产品进行排名的常用方法是执行向每个产品分配分数的top-k查询。在典型的设置中，top-k查询使用权重wi 对于每个属性pi 的产品p，并通过评分函数f组合这些权重和属性。每个客户的特定权重集是不同的。在文献[4]中经常使用的分数函数是线性函数fw，对于乘积p，其计算公式为： wi\*pi在不失一般性的情况下，我们假设较低的分数更好。

客户的权重定义与用于表示产品的向量具有相同维度的向量w。因此，它们可以呈现在相同的d维特征空间中，其中d是产品属性的数量。在图1我们可以看到十二个产品p1……p12和w1…w6六个不同客户的加权矢量。该空间的维数为d = 2。如果我们利用欧氏距离来计算产品之间的相似性，那么显然产品p9更类似于产品p7和p12。一个自然的问题是，这种相似性评估对于用户是否是自然的，基于他们表达的偏好，在六个加权矢量中描绘？

没有收到top-k ,reverse topk查询的详细信息（这些将在下一节中介绍），产品p8是最受欢迎的四个客户名单（w1; w2; w5; w6）而这些客户中的三个以及具有加权向量w1，w2，w5的更具体客户也包括产品p9在他们最喜爱的产品列表中。另一方面，p7和p12都不包括在六个客户中的任何一个的最喜爱的列表中。因此，从六个客户的特定集合的角度来看，对于具有加权矢量w1，w2，w5的客户， p8和p9被认为是相似的，而根据其偏好，p9，p7，p12之间没有明显的相似性。

换句话说，产品p9和p8的反向top-k集（对于k = 3）（即，分别位于其前3名喜爱的名单产品p9中的顾客组，以及分别为产品p8的顾客组）包括具有加权向量w1，w2，w5的客户。因此，相同客户在两个或更多产品的反向top-k集合中的共存是其相似性的正面证词。这个简单的例子表明，如果我们将用户偏好考虑在内，产品之间的相似性可能会非常不同，而不是将这些产品视为其特征空间中的各个点，与这些用户偏好的影响隔离开来。

在我们的工作中，我们利用Jaccard系数在反向top-k查询的结果集之间执行相似度计算。还考虑到进一步考虑到产品行列的延伸的相似性概念。我们用我们介绍的两种查询类型来补充这些新的相似性度量，称为类似于众所周知的范围和最近邻查询的u-相似性和m-最近邻查询，但不同之处在于它们通过查看它们来评估产品之间的相似性 反向top-k集。不幸的是，已知反向top-k查询导致高计算成本，这意味着对这些查询的强力评估是不切实际的，即使对于中等数据集也是如此。在这项工作中，我们确定了反向top-k结果的重要属性，使我们能够计算产品之间以用户为中心的相似度的上限和下限。根据这些边界，我们能够使用产品空间上的常规R-tree索引来有效地执行这些查询。我们进一步讨论重新使用先前计算的反向顶部k集的其他优化，以进一步限制被调查产品的相似性。

## 本文贡献

我们介绍一种以用户为中心的相似性搜索的新框架，其基于用户偏好来利用产品的排名来发现类似的产品。

我们定义了以用户为中心的相似搜索的两种新颖的查询类型（u相似度和m-最近邻），识别有效分数界限（第3节）并提出了通过利用派生界限和传统索引结构来修剪搜索空间的有效查询处理算法（第4节）。

我们解释如何使用不同的相似性度量来扩展我们的技术，以更精细的方式捕获以用户为中心的相似度（第5节）。

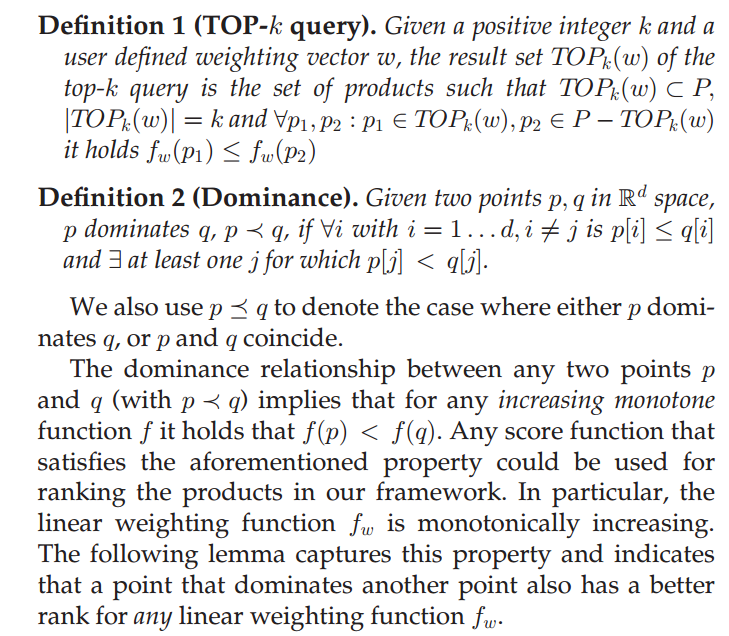
我们显示在处理查询时计算的结果可以被利用来获得更紧密的界限，从而大大提高了查询处理的性能（第6节）。

我们进行详细的实验评估，其演示了以用户为中心的相似性搜索的效率和有效性（第7节）。我们的实验的一个关键发现是，通过利用用户偏好，与传统的相似度计算相比，我们经常获得非常不同的结果。在我们通过真实数据集联系的两个用户研究中，我们发现在大多数查询中，用户倾向于通过我们的技术获得的结果。

此外，我们提出有助于建立相似性框架的定义（第2节），我们讨论相关工作（第8节），并提供结论（第9节）

# 2 PRELIMINARIES

在本节中，我们介绍定义和初步我们利用它来更好地介绍和描述我们的技术。



此外，为了执行我们的例子的相似度计算，我们应该执行反向top-k查询，返回每个客户的权重向量，作为输入的查询产生的产品p属于她的第k个 排名最高的产品。与作为输入特定客户并在产品数据集上执行的top-k查询相反，反向top-k查询作为输入接收产品，并且在所有用户的加权向量集合上执行.

# 3 SIMILARITY & QUERIES

在本文中，我们介绍了以用户为中心的相似度计算方法，其中通过考虑用户偏好来定义两个产品的相似度。如果两个产品满足相同的用户偏好，则这两个产品被认为是相似的，这又意味着它们的反向顶部k集合是相似的。更正式地，假设一组客户C.每个客户ci已经用权重向量wi定义了她的偏好。如前所述，给定一组产品P，如果查询点q是客户ci的top-k查询返回的k个产品之一，加权向量wi将是产品q的反向top-k集合的结果的一部分。

为了对产品的RTOPk集执行相似度计算，我们引入了这些集合的距离（即相似度）的度量。一组众所周知的相似度函数是Jaccard系数。Jaccard系数测量两组之间的相似度，作为其交集大小的一部分除以两组的联合大小。

