# abstract

条件优先网络（CP网络）被广泛用于表达定性偏好。由于用户有时不愿意或不能指定完整的CP网络，因此禁止个性化搜索被有效地进行。在本文中，我们提出一种使用不完整CP网络执行个性化搜索的方法。我们提出了一种用于补充用户CP网络的偏好推荐方案，以提高个性化搜索的准确性。我们进行了广泛的模拟和用户研究，以证明我们的方法的有效性。

# Introduction

随着互联网产品和服务的丰富，服务消费者面临越来越多的选择。面对大量选择的决策总是一个挑战，特别是当决策者有不同的口味或标准时。个性化搜索[1-5]是一种旨在帮助用户根据个人喜好识别所需产品或服务的技术。它引出用户的偏好，以标准模型表示，并让计算机系统自动识别用户感兴趣的产品。

近年来，已经提出了一些具有不同表现力和复杂性的模型来描述用户个性化搜索的偏好。例如，Matthijs等人[6]提出了使用用户完整浏览行为构建用户配置文件的个性化方法。Kim等人[7]提出了一种基于标签的个性化搜索模型，以提高信息检索的准确性和覆盖面。在这些模型中，条件偏好网络（CP-nets）[8-10]是广泛使用的。它不仅可以简洁地表达用户的定性偏好，还能够指定用户偏好的范围。这使得CP网络成为个性化搜索系统的合适模型。

基于CP网络的大多数现有搜索系统[11,12]假设用户应该在搜索之前完全指定他们的偏好设置。这个假设在许多真实的情况下可能不成立。由于偏好引用和描述通常是冗长乏味且容易出错的过程[13]，获取完整的CP网表达用户的偏好并不总是现实的。为了进行个性化的搜索工作，系统必须能够处理不完整的CP网络表达，并充分利用可用信息来检索用户所需的产品或服务。

我们以前的工作中引入了一些相关技术[14,15]，它们尝试利用不完整的CP网络来选择Web服务[14]。此外，我们引入了基于协同过滤[15]的框架来执行偏好互补。在本文中，我们扩展了我们以前的工作，并提出了一个完整的系统，用不完整的CP网进行个性化搜索。首先，我们扩大了从服务选择到对象搜索的方法的应用。其次，我们提出一个新的，完整的框架来管理这种方法的过程。第三，基于协同过滤原理[16]，我们提出了一种偏好推荐方案，用于在个性化搜索过程中逐步补充不完整的CP网络并解决冲突的CP网络。提出了三种新的算法，并将其应用于个性化搜索的不同步骤。最后，进行了新的模拟和用户研究，以比较我们的方法与其他方法。

本文的其余部分组织如下。第2节回顾了CP网络和协同过滤的一些背景。第3节介绍了基于偏好建议的个性化搜索系统。第4节介绍了一套新颖的个性化搜索方法。第5节介绍了实验评估和用户研究的结果。最后，第6节总结了本文，为未来工作提供了指导。

# Background and related work

在本节中，我们简要介绍CP网络和协同过滤。

## 2.1 CP-网

条件优先网络（CP-nets）[8,17]被设计用于表示定性用户偏好。 该模型可以定义如下：

定义1：令V = {X1，...，Xn}是一类对象的属性集合。V上的CP网络是X1，...，Xn（被称为依赖图）的有向图G，其节点用条件偏好表注释，每个Xi∈V由CPT（Xi）表示。每个条件偏好表CPT（Xi）将Xis值的总顺序与Xis父项的每个实例化相关联。

我们使用图1中的示例来说明CP网络的语义。假设爱丽丝计划访问悉尼，需要在互联网上预订酒店。然后，她面对一些选择，例如时间（即周末或工作日），酒店的位置（即城市或郊区）和酒店类型（即经济型酒店或奢侈品 一）。如图1（a）所示，爱丽丝对时间无条件偏好。由相应的CPT指出，她宁愿在悉尼度过周末，而不是平日。然而，爱丽丝对于位置的偏好取决于她什么时候去悉尼。如果是一个周末，她想留在这个城市，因为周末在这个城市举行了许多活动。如果是平日，她更喜欢留在郊区。此外，爱丽丝对酒店类型的偏好取决于她所在的地区。如果她住在城市，她想预订一家经济型酒店，因为她将大部分时间花在商场里而不是在酒店。如果她住在郊区，她更喜欢一家豪华酒店，提供优质的服务。根据爱丽丝参考的CP网络介绍，我们可以引用她的详细偏好图，如图1（b）所示。一个经济型酒店在城市预订了一个周末将是爱丽丝的首选。

在现实世界的设置中，用户可能无法提供他/她的偏好的完整的CP网络介绍，特别是当描述产品或服务的数量很多的属性时。因此，爱丽丝酒店的偏好可能不完整。如图2（a）所示，CPT中的某些字段缺失。在这种情况下，Weekend∧City∧Budget，Weekend∧City∧Luxury和Weekday∧Suburb∧Luxury变得无法比拟（见图2（b））。当CP网络中的优先级规范不完整时，个性化搜索将不太有效，因为可能有太多的候选对用户来说是最佳的。

CP网络模型首先在[8,17]中引入，其中作者提出了该模型的一些核心属性以及用于个性化搜索的算法，例如用于最优结果生成和排名的算法。

近年来，CP网的模式在一些后续工作中进一步扩大[18]。已经提出了各种技术来将CP网络应用于个性化搜索和决策。它们包括执行CP网络启动的程序[19]以及将CP网络与其他决策支持工具集成的方法[20]。最近，提出了以消费者为中心的信息服务组合的基于CP网络的模型[21]。然而，现有的个性化搜索技术很少涉及不完全网络的情况。在本文中，我们提供了具有不完整CP网络的个性化搜索的解决方案。

## 2.2 Recommender system and collaborative filtering

推荐系统是一种旨在寻找特定用户可能感兴趣的产品或服务的技术。它分析用户的个人资料，并通过统计方法预测用户的兴趣。然后，用户的兴趣可以用来识别产品。推荐系统的方法也可以应用于偏好启发。推荐系统中使用的最典型的技术是协同过滤[16,22]。它假设志同道合的用户倾向于具有类似的兴趣，并利用用户偏好中最常见的模式来提出建议。要向用户推荐产品，它首先识别与用户具有相似特性的历史用户配置文件。然后分析志同道合的历史用户的兴趣，并利用他们的共同兴趣选择活跃用户的产品。该方法已成功应用于许多领先的商业网站，如亚马逊和Ebay。为了处理不完整的用户偏好，我们借鉴了协作过滤的想法。我们找到与活跃用户共享类似偏好的历史用户，并使用其偏好来补充活动用户的偏好。

推荐系统[23,24]和协同过滤[16,25-27]已经有很多工作。Zhang等提出了一种用于服务选择的混合推荐系统[28]，其将协同过滤与模糊集技术相结合。然而，**大多数现有方法在项目级别而不是用户偏好级别中工作。换句话说，它通过比较用户对现有产品和服务的评级来识别志同道合的用户。**当产品和服务数量庞大时，这些方法将需要大量的用户评级。在本文中，我们利用直接应用协作过滤对用户偏好的可能性。我们的方法比较用户的偏好（以CP网络表示）来识别志同道合的用户。在偏好水平上工作，这种方法需要更少的信息才能有效。它可以应用于任意大小的产品仓库。

# The framework

根据现有的协同过滤工作[16,26,29]和我们以前的工作[12,14]的观察，相似的用户通常都有类似的偏好和CP网。可以使用他/她志同道合的用户的偏好来预测用户的偏好。因此，我们提出一种基于偏好推荐的完整的个性化搜索系统。该系统的框架如图3所示。

一旦接收到用户的偏好描述，我们的系统首先执行一致性检查。在此步骤中，它检查用户提出的CP网络，以验证其是否包含**冲突**。在一些最近的工作中已经提出了许多用于**冲突检测或CP网络一致性检查的算法[9,10]**，我们略去了这一步的细节。

如果用户的偏好一致，将执行个性化搜索以为用户找到理想的产品或服务。如果结果集的大小可以接受，则个性化搜索就会结束。通常，定义阈值来控制解决方案集中包含的可用服务的数量。阈值通常取决于系统和用户体验的性能。搜索后，用户的偏好将被存储在用户配置文件的存储库中。

如果用户的偏好包含冲突，这些冲突在诱导的偏好图中表示为周期，则将启动冲突删除过程以消除所有冲突（循环）。删除冲突的偏好时，消除参与冲突（循环）的个人偏好，这些偏好的用户最不受支持。修改后的一致CP-nets描述将被传递给个性化选择器以检索用户最喜欢的产品或服务。

为了提高冲突解决的性能，我们将历史用户偏好聚类到几个集群中。每个集群包括具有类似偏好的用户。用于识别类似用户或志同道合的用户的步骤被称为类似用户检测。要测量一组类似用户支持的个人偏好量，执行“偏好投票”。

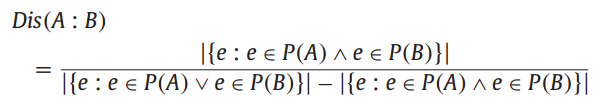
如果用户的偏好设置不正确，则CP网络将不完整。在这种情况下，结果集的大小可能太大，无法让用户使用。类似地，执行偏好投票来提出偏好来补充CP网络。在偏好互补中，候选人偏好将根据投票结果的可能性和敏感度进行评分。将选择具有最高分数的条件偏好并将其添加到当前CP网络中。优先补充是一个增量过程，其中一次将偏好添加到CP网络中。之后，再次执行服务选择以优化结果集。将重复该过程，直到结果集足够小或不能进行更多的互补。

在我们以前的工作中，用于识别类似用户的算法是复杂的，并且在面对大量的历史用户时，偏好投票的成本可能很大。因此，本文改进了偏好互补的方法。首先，我们利用历史偏好预聚类的步骤来降低识别类似用户的计算复杂度。其次，我们允许优先投票返回一个偏好列表的投票。然后，在优先级互补步骤中，顶级K（例如K = 5）偏好将逐个添加到活动用户的偏好。在使用顶级K偏好设置之前，下一轮偏好投票将不被激活。

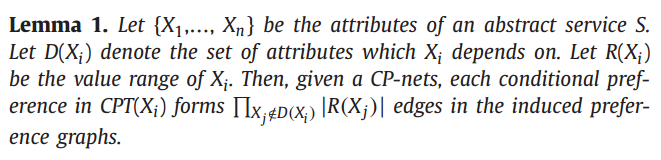
# The methods

在本节中，介绍了基于协同过滤原理的优选推荐方案[16]。这些方法可帮助用户在个性化搜索期间逐步补充不完整的CP网络或解决冲突的CP网络。

如第2节所述，我们使用CP-nets描述用户的偏好。因此，两个用户之间的相似性可以通过它们的CP网络之间的相似度进行测量。这种相似度的直观度量定义如下。定义2（CP网之间的距离）。让A和B是用户请求的对象的两个CPnets。令P（A）和P（B）分别是A和B的诱导偏好图。令e表示偏好图中的边。因此，从B到A的距离计算为：



根据定义，CP网B和CP网A之间的距离是A和B的诱导偏好图之间的重叠大小（如图1（b）所示）除以不重叠的大小部分。为了简化该距离的计算，我们假设不同用户的CP网络共享相同的依赖关系图。这个假设并不一定成立。然而，当用户在CPnets中指定不同的依赖关系图时，我们仍然可以为它们创建一个通用的人工依赖图。这可以通过将它们的依赖关系图连接成一个通用的超级图来轻松实现。然后，可以相应地调整用户的CPT以适应公共依赖图，而不改变其语义。当CP-nets共享一个共同的依赖图时，它们的距离可以从CPT直接计算出来。



属性的值可以是离散的或连续的。如果Xi的属性值是离散的，则R（Xi）是可以分配给Xi的值的集合。否则，连续值将以固定间隔的一系列值替换。| R（XJ）| 是集合R（Xj）的基数。这是R（Xj）的成员数。对于实例，在图1中，D（时间）=∅，Xj∈/ D（时间）= {位置，类型}，| R（位置）| = 2，| R（类型）| = 2。因此，优惠周末？ 平日确定诱导偏好图中的四个边，而偏好周末：城市？ 郊区确定诱导偏好图中的两个边。