Recommending trusted online auction sellers using social network analysis

使用社交网络分析推荐值得信赖的在线拍卖的卖家

摘要

目前，大多数在线拍卖网站用于推荐卖家的信誉系统是极度简单的，并没有考虑到一些卖家勾结从而虚假地增加他们的评级。本文提出一个推荐系统，采用交易关系去计算卖家的推荐等级。我们证明了交易历史形成的网络结构可以用于去揭露这种潜在的卖家的勾结行为。

从结构角度出发，我们关注交易者之间的关系而不是他们的属性值，采用两个社交网络指标k-core和中心加权算法创建一个collaborative-based 推荐系统。这个系统能够显示一个账户的勾结风险。我们通过将本系统与真实世界的黑名单数据对比，发现本系统能够筛选出76%黑名单账户，这些黑名单数据由一个主流在线拍卖网站定期公布。这个系统能够提供警告早于官方公布的黑名单数据几个月，来防备可能的卖家勾结行为，因此本系统能够合并在用于推荐卖家的信誉系统之中。

1 Introduction

信誉系统建立在一个简单的机制上，就是报告买家对卖家的评价的累积结果，评价可能是积极的、中等的或消极的，并附带评论。它利用一个私人订购系统去限制只有在交易完成之后，买家才能对卖家评论。过去成功的私人订购系统是基于对地理位置接近和频繁重复交易的验证，或者担保人存在，或者资源在附近的交易。然而，在线拍卖市场的特性允许注册匿名的大量账户来作弊，避免消极评价。一些研究结果也表明信誉机制对公众是没有益处的，因为在线拍卖交易都是先付款再发货，所以买家易受骗。

当前在线拍卖的信誉系统有3点缺陷：

1. 卖家与买家之间信息不对称
2. 网上假名滥用信用声誉的道德危机
3. 交易完成之后不给消极反馈的趋势

这些缺陷为恶意的卖家提供了一个低风险的环境，同时也给买家追回损失带来了困难。而传统的推荐系统，无论是基于内容的推荐系统，或者协同推荐系统，都是无法解决这些缺陷带来的问题。本文的推荐系统通过引入社交网络分析去分析累积评分的底层结构和它相对应的交易网络，通过k-core和中心权重过滤出恶意的卖家。

2. Review of the literature

2.1 Online auction

不像传统的拍卖方那样负责配送和拍卖物品的鉴定，网上拍卖网站只充当平台提供者，避免在任何交易中充当中间人。所以网上拍卖网站的收费也远远低于其他拍卖机构，因为进入或者离开拍卖网站的成本较低，也促进了欺骗和勾结。

2.2 Recommendation systems

推荐系统从推荐数据的来源分类，可以分为3类

1. 基于内容的推荐系统，这类系统基于物品的相似性，推荐与用户之前的喜好相类似的物品。
2. 协同推荐系统，这类系统基于相似用户的共同品味。
3. 混合推荐系统，前两者结合

以上两种系统采用的技术可以是基于启发式的，也可以基于模型的。基于启发式的技术通常研究交易行为的上下文信息(contextual information)去评定等级。基于模型的技术是基于概率统计模型或机器学习模型，包括贝叶斯分类、线性回归、聚类、决策树和人工神经网络，去创造各个层次的推荐级别。

两种推荐系统都缺乏对虚假账户的考虑，给蓄意欺诈留下了很大空间。

2.3 Social network analysis (SNA)

SNA试图量化社区中成员之间的交互作用，从而描述社区的结构和成员。这种方法被开发和使用在社会学领域已经数十年，社会科学家采用这种方法去调查社会的集体动态。SNA的基础成员是行动者或结点，作为行动的源，和链接作为结点之间的关系。这些结点可能是个人、组织或团体，这些链接可能是一个或多个类型的关系或者行动之间的共同特征。本文使用SNA作为理论框架，考虑到两个主要SNA度量，centers 和 k-core.

SNA关注联系，group和network是两个基本的关系类型。Group是关系紧密、范围小和低流动性的。Network是关系稀松、大范围、联系类型多样和高流动性的。关系是资源交换的渠道，在网上拍卖中就是交易(transaction)。联系能通过距离、密度、频率和中心性来度量，其中中心性是network和group的结构指标，体现了个体或结点相对于其他结点的地位。Closeness是全局中心性的度量，Betweenness指明了结点相对于其他结点的地位。本位使用一种中心性度量Centers找到子网络的中心，这些子网络是从原始交易网络中提取出来的。Centers概念遵循robbery算法，比邻居结点度大的节点从相连的结点偷权重。在这个过程的开始阶段，每个结点的初始权重的计算是基于它的度数，然后每个结点和邻居结点比较，stronger结点从weaker结点偷权重，偷的权重越多，就越stronger。本文称这种度量为centered-weights或CW value。

子群(subgroup observation)观察是本文的重心，子群能够提供有价值的信息去理解子群对整个网络的影响。子群能够通过component、clique、k-plex和k-core得到。K-core计算见令一个文档。不同k-core等级之间的结点减少的比率显示了group的结构细节。高k-core结点的数量越多，这些结点凝聚力越强。Core/periphery(C/P)率是k-core成员数量除以与k-core相邻边缘结点的数量。本文首先尝试在推荐系统中使用社交网络，这个方法使克服在线拍卖欺骗成为可能。

3. Building collaborative recommendation system with relationship-based heuristic technique

从交易历史记录导出的k-core指标可以显示出账户交易关系的凝聚性和信誉结构。换句话说，经常活跃的账户将会形成子群，这些子群能够被k-core指示器捕获。这么做的目的是将恶意商家和正规账户分离。本文的研究可以从以下四方面进行阐述，数据采集、关系构建、SNA度量构建和基于对虚假账户识别的召回率和准确率进行测试。

3.1 Data source and data collection process

我们收集用户时变的图形数据(pattern data)作为协同过滤分析的来源。我们收集的数据来自台湾主流的在线拍卖网站主机，雅虎台湾，它占了75%台湾在线拍卖市场份额。使用雅虎台湾的另一个原因是它定期公布确认的欺诈账户，而我们测试一个账户是否是欺诈的就要用到这个公布名单。本文收集了第一次黑名单公布以来到2005年5月26日的账户的信誉历史。2042

在收集完黑名单账户之后，我们继续研究欺诈账户的交易行为模式，通过以月为基础重建账户交易的交互模式。重建黑名单账户的交易网络是基于账户的评价日志的。采集来的数据最终形成3个数据库，黑名单账户的资料信息，过去评价的交易日志和附带评论的额外的拍卖信息。除了黑名单账户之外，从交易日志中获得其他112491个账户。

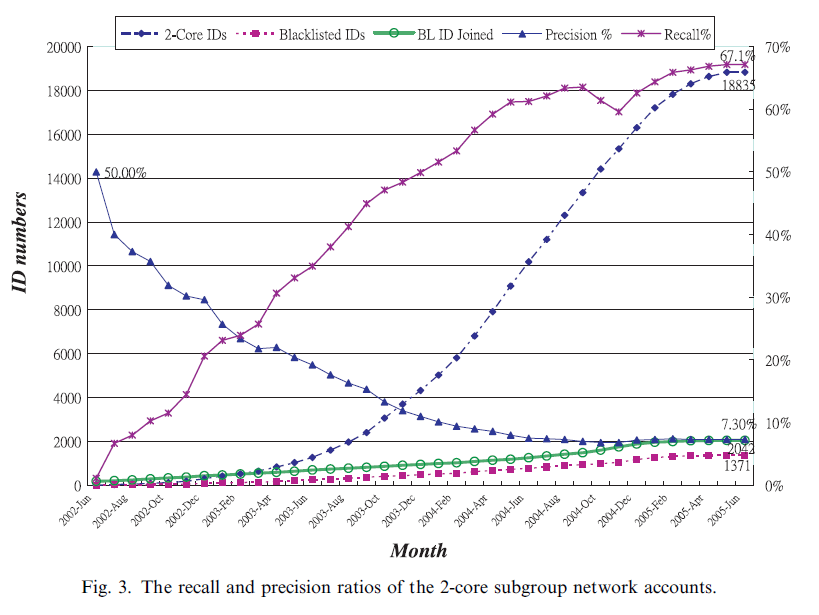
3.2 Constructing the interaction structure from transactions

为了简化计算过程，我们以至少与两个黑名单账户交易的账户为目标，构成2-core子群。删除92233个账户，这些账户只交易了一个。我们分析剩下20258个账户的社交网络指标，从而跟踪他们之间的网络联系，从2001年10月只有11个账户和6笔交易，到2005年7月20258个账户和53788笔交易。

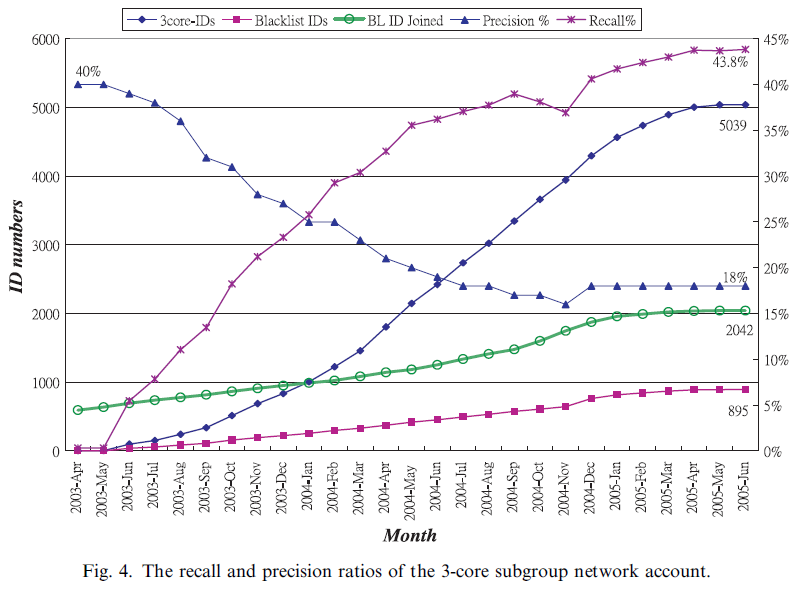
接着，对于每个月的交易网络，生成网络的2-core子群来确定网络的聚集子群。



随时间和账户数量的发展，准确率不断降低，召回率不断提升。



鉴于此，为了得到好的推荐效果，进一步处理是有必要的。《Detecting online auction inflatedreputation behaviors using social network analysis》说有两种方法提升准确率，一是高level的core网络，例如3-core或更高。另一种方法是计算子群网络的C/P率。这里首先尝试第一种方法，采用10-core网络，得到结果如下：



准确率虽有提升，但是召回率下降，type II 错误率增加。接下来尝试第二种方法同样无效。因此，本文采用另一种方案，中心性度量(centrality measurement)，得到结果如下：

