基于异构网络分析的商品推荐系统研究

重庆大学硕士学位论文

（学术学位）

学生姓名：董俊平

指导教师：熊庆宇　教　授

专　　业：软件工程

学科门类：工　学

重庆大学软件学院

二〇一五年五月

Study on Items Recommendation System based on Heterogeneous Information Network Analysis

A Thesis Submitted to Chongqing University

in Partial Fulfillment of the Requirement for the

Master Degree of Engineering

By

Dong Junping

Supervised by Prof. Xiong Qingyu

Specialty: Software Engineering

School of Software Engineering of Chongqing University,

Chongqing, China

May 2015

摘 要

随着互联网和移动互联网等技术的不断发展，越来越多的商品交易以电子商务的形式进行，互联网电子商品交易平台的交易记录以前所未有的速度增长并日益累积。面对海量的电子商务交易数据和日益增长的用户购买需求，如何快速、准确地给用户推荐所需要的商品是当前电子商务领域发展所需要解决的关键问题。本文以国家自然科学基金“基于异构信息网络分析的 Web 服务推荐”为支撑，借鉴异构服务网络分析的思想，将其应用于电子商务产品推荐中，提出了基于异构信息网络分析的商品推荐方法。

本文在现有推荐方法、聚类分析和异构信息网络研究的基础上，结合电子商务平台中用户对商品选择的需求，考虑电子商务交易记录中各种参与对象之间的潜在类别关联，构造异构商品网络模型；基于异构商品网络聚类和排序对商品交易数据进行分析，结合一定的推荐策略实现商品推荐。论文从交易数据中买家、卖家、商品、热点词之间的网络关联关系着手进行异构商品网络的提取、构造、分析与处理，充分挖掘商品交易各参与对象之间的潜在类别关联，结合用户需求，提供考虑网络关联和类别区分的商品推荐。

论文主要研究工作包括：

① 结合信息网络异构化发展趋势，基于网络分析、异构网络分析与商品推荐等国内外研究现状及问题分析提出了本文的研究内容与创新点。

② 借助形式化方法研究了异构商品网络描述模型，结合电子商务交易数据特征分析、构造与维护异构商品网络。

③ 基于异构商品网络描述模型，提出了基于异构商品网络分析的商品聚类算法，结合商品网络排序函数与排序模型进行异构商品网络中商品交易各参与对象的聚类分析，实现异构商品网络对象的类别挖掘和类别内的重要性排序。

④ 从商品交易记录中各种对象之间的关系维度出发，针对异构商品网络聚类结果，结合相应的推荐策略提出一种新型的商品推荐模型，分别从推荐思想、推荐流程、算法描述等方面进行详细研究内容的阐述。

⑤ 基于异构商品网络聚类的推荐模型研究，从软件工程的思想出发，从需求分析、系统功能设计、数据库设计和系统实现等方面设计并实现了商品推荐原型系统，从而验证了所提出的推荐模型的可行性。

关键词：异构信息网络，商品推荐，商品聚类，排序函数

ABSTRACT

With the development of the Internet, e-commerce and mobile Internet, a growing number of items trading are in electronic form. The transaction of the Internet electronic items trading platform increase and grow cumulative. Faced with the growing e-commerce transaction data and user requirements, how to quickly and accurately to recommend the goods of meeting the needs of users is a key issue of the current scientific and technological development in the field of e-commerce. In the paper, the National Natural Science Foundation of China, "Web service recommendation based on heterogeneous information network analysis" as the support foundation, and heterogeneous service network analysis as the reference, items recommended based heterogeneous information network analysis is proposed.

Based on the recommendation method, clustering analysis and information network research, it construct heterogeneous items network model which combined with the requirements of the e-commerce platform and considering the potential associations of various categories participating in e-commerce transactions; Next is the network clustering and sorting based on transaction data analysis, combined with a certain recommendation strategies to achieve the items recommendation. From the relationship of buyers, sellers, items, terms, to precede heterogeneous items network, it considere to the relationship of the recommendation between the four items trading process and improve the efficiency of items recommendation.

The details of research works in this paper include:

① Combining the trend of the heterogeneous information network development, research status and analysis the problems based on network analysis, heterogeneous network analysis and recommend merchandise and other domestic and proposed research and innovation point.

② Studies describing heterogeneous information network model with formal methods, combined with characteristics of e-commerce transaction data structure and maintaining of heterogeneous items network.

③ Proposed items clustering based on heterogeneous items network analysis with heterogeneous networks description model. Analysis on heterogeneous items clustering with the Ranking function、Ranking model and NetClus algorithm.

④ From the dimensions of the relationship between items trading record, with the

clustering results, combined with the corresponding recommended strategy recommended electronic goods, items were recommended by a detailed description of the recommended ideas, recommended processes.

⑤ Based on heterogeneous items network clustering result, proceed from the idea of software engineering, design and implement the prototype of items recommendation. Describe the prototype of the design process from requirements analysis, system design and implementation effectiveness analysis.

Keywords: Heterogeneous Information Network, Items Recommendation, Items Clustering，Ranking function.

目 录

1 绪 论

当前人们处于一个信息爆炸的时代，每时每刻都会产生大量不同种类的信息。随着信息的膨胀，信息背后潜在的数据正以极高的速度增长并积累。在大量的数据中，如果不考虑数据之间的关系，整个信息世界就是无数个独立的信息孤岛，在数据与数据之间没有了沟通与交换，信息世界就是一个毫无生机的信息死海。由于信息社会具有了各式各样的复杂多变、种类各异的关系，才使得信息世界拥有了丰富的关联关系，从而形成了如今错综复杂的网络状态，从而支持快速的信息交换和数据传输。随着互联网的迅速普及，电子商务魅力也随之显露，虚拟企业、虚拟银行、网络营销、网上购物、网上支付、网络广告等一大批前所未闻的新词汇正在为人们所熟悉和认可，这些词汇同时也从另一个侧面反映了电子商务正在对社会和经济产生的影响。因此，如何利用电子商务中交易记录的异构性更好的进行商品推荐就成了新的研究课题。论文正是从商品交易中关系的维度出发，基于关系分析和挖掘，提取商品网络数据的网络结构，在异构信息网络分析研究的基础上，对聚类分析、商品推荐进行关系维度的研究。

1.1 研究背景与意义

1.1.1 研究背景

随着互联网的迅速发展，从 20 世纪末开始，一大批互联网企业涌进了人们的日常生活，如国外的 Google、Amazon、Facebook、Twitter，国内的有百度、淘宝、腾讯、京东、网易等。这些互联网企业凭借良好的网络服务和应用改变了人们的生活方式，并极大地提高了人们的生活质量和工作效率。其中，以淘宝和京东等为代表的电子商务平台也以全新的电商模式开启了中国民众网上购物热潮。电子商务在人们的日常经济生活中占的比重随之上升，也得到了越来越广泛的应用。随着电子商务用户的急剧增加，产生的电商数据也是海量之势，如果这些宝贵而又庞大的用户信息和交易数据没有得到合理分析应用，将对公司和企业的领导人决策提供不了持续有利的建议，也会影响电子商务的持续发展。所以近些年来很多电子商务公司都引进了先进的数据挖掘软件挖掘电子商务背后的丰富数据资源，期望通过分析处理这些资源进而挖掘其背后蕴藏着的巨大的有价值信息，提高商品推荐的精确和准确度，使得电子商务更稳定更高效的持续发展。

电子商务网站是企业将产品以特定形式展示在网站上，等待和来访的客户交易。显然这种方法是非常被动的，而且随着互联网的普及和电子商务的发展，电子商务网站的内容随之增加，客户需求也越来越大，网站解决问题的难度也更加

复杂。同时，“one-size-fits-all”的方法不考虑用户的不同需求、偏好、行为特征和人口统计信息等，却在为不同的用户提供相同的方式进行服务，没有考虑用户的个性化需求。如此，导致了客户在浏览电子商务网站时迷失在巨大的商品信息中，找不到符合自己个性化需求的合适商品从而浪费一定的空间与时间，降低了客户购买商品的效率从而影响到电子商务网站的营销份额。基于以上原因，基于个性化服务技术的电子商务网站随之诞生，其目的是为不同的客户提供更具针对性的个性化服务，而电子商务网站中商品推荐系统就是最核心的部分[1]。

随着互联网技术的日渐成熟和电子商务的快速发展，电子商务网站在解决用户日趋复杂的商品选择需求时，其网站布局和推荐方式也变得越来越复杂，使得买家在商品选择过程中迷失在大量的无用商品信息中，而不能选择合适的商品浪费时间。商品推荐系统可以与用户个性化交互，模拟实体商品店销售人员向买家提供商品信息服务推荐，从而帮助用户从大量商品中选择自己最需要的商品并顺利完成整改购买过程。随着电子商务日趋激励的竞争环境发展，商品推荐系统可以有效的保留用户并防止用户流失，从而提高了电商的销售份额。电子商务推荐系统在电子商务领域中具有极好的发展前景和应用范围，已经成为了电子商务中的重要研究内容。但是因为电子商务系统规模的扩大与发展，商品推荐系统也有一系列的挑战及机遇。

同时，信息技术的不断发展和日益增长的数据复杂度，使得由于复杂的信息之间的关联关系而连接在一起形成的各种网络结构。在大量的网络形式中，互联网作为最大的信息网络，对所有的数据对象作为节点，各种关系之间的节点连接形成一个信息网络；互联网作为一个包罗万象的信息网络是由各种信息实体，根据不同的分类标准可以分为不同的子网的信息，例如，通过网页和网页超链接结构，由电子商务参与者组成的电商网络等[2][3]。除了上述几种信息网络形式，关系数据库在连接模式的数据组织可以映射到数据元组作为节点，对元组网络外键关系。其中由电子商务参与者组成的商品网络由于其数据的异构性，可以更好的应用于异构网络分析的方法之中。综合以上背景，如何更好的用异构信息网络进行良好的商品推荐则成为了目前主流推荐方法之外的最新研究。

1.1.2 研究意义

论文结合当前信息网络分析和异构信息网络的研究现状，针对商品推荐中对商品聚类的需求、提高商品推荐准确性的要求提出基于异构信息网络分析的解决方案，从关系的维度出发进行商品排序、聚类和推荐研究，对于提高商品推荐精度，提高异构商品网络资源利用率具有一定的理论研究和实际应用价值。

基于异构网络分析的商品推荐方法可以挖掘商品与各参与方对象之间的潜在类别关联，对于进行商品聚类和商品推荐提供有益的数据基础。考虑商品、买家、

卖家之间的类别相关性进行商品聚类和推荐可以提高为目标用户推荐合适商品的准确度，提高推荐系统的推荐质量，增加电子商务网站的客户黏性，促进网站销售量的增加[4]。异构商品网络将在推荐系统中进行更全面的考虑，特别是在系统参与方三者之间的过程中的商品交换过程，将用户从世界的无限的网络资源和商品环境中独立出来，可大大节省时间和用户的商品采购成本；提高了电子商务网站的用户忠诚度，电子商务网站将更多的浏览者转为商品的买家，可以提高网站的商品销售的能力，为电子商务发展赢得更多的机会。

1.2 国内外研究现状

1.2.1 网络分析

信息网络分析利用信息网络描述数据之间的关系，基于网络节点关系分析的手段进行数据知识挖掘、信息排序与检索。随着信息网络的发展，网络分析的方法解决应用中的相关问题已成为一种非常有效的方法。在之前的研究中 HITS 和 PageRank 是排序算法中较为典型的 Web 搜索排序算法[6]。PageRank 算法在 Google 搜索引擎的成功使用证明了其基于关联分析的排序算法在寻找“突出个体”对象方面具有优势，在该算法中，页面节点和超链接结构包括一个网页的 PageRank 排序算法的网络结构，对网络分析的计算节点的作用下，实现网页在页面搜索空间中的排序[7]。HITS（Hyperlink - Induced Topic Search） 算法则是由康奈尔大学的 Jon Kleinberg 博士于 1997 年提出的，目前，Teoma 一直使用其作为在网页排序过程中搜索引擎的链接分析算法[8]。HITS 算法的思想是通过 Web 页面的网络分析，找到高质量的“Authority”的主题与大量网页的用户查询相关的（在 Authority 页）和“Hub”（页包含很多指向高质量的“Authority”的页面连接到枢纽网页）。上述两种基于信息网络分析的研究方法主要针对 Web 页面的排序和搜索，构造的网络结构是单一的页面组成的网络，是属于同构网络分析研究的范畴。为了对排序从网页对象级别的目标的基础上展开研究 PopRank 和 PageRank 网络调度[9] [10]对象的普及，基于流行度传输因子的算法进行对象的网络排序。面向对象网络结构的排序可以在信息数据对象的层面进行排序重要性计算，提高了排序方法的适用范围。

1.2.2 异构网络分析

由于信息网络结构异构化的发展趋势，异构信息网络的定义在很多研究中被研究者重点指出。研究基于网络结构的分析从单一到异构网络结构的发展，已经成为一个新的研究课题，研究影响较大的包括 Rankclus 和 Netclus [10]。与传统信息网络分析排序算法的应用原理不同的是，在 RankClus 算法中，其重点目标对象的排序方法为异构信息网络，通过提取该目标对象的异构网络特征与潜在的类别关联，分析它们之间的关联关系和类别关联并实现该方法。NetClus 算法则是将排序

的目标对象扩展到 3 个及 3 个以上对象的异构信息网络，考虑不同对象的聚类并在潜在类别分类中实现全部对象的排序操作[14]。

通过以上的分析可以发现，从一个单一的信息网络研究上升到面向对象层次网络对象的分析，排序算法的分析也从具体的 Web 网页转变为面向对象层级的操作；随着信息网络中的对象类别和关系的复杂性的增加，对异构信息网络结构的数据对象的提取，异构信息网络分析方法实现基于数据分类信息是不可避免的；在异构信息网络中，随着数据类别增加，考虑聚类分析的排序是实现清晰类别排序方法研究的有效方式之一。

1.2.3 商品推荐

面对电商平台产品促销的需求，同时为满足潜在用户的购买需要，推荐系统逐渐成为电子商务信息技术的一个热点研究内容，得到了国内外众多研究组织和学者的青睐。自上个世纪末开始，SIGKDD（Special Interest Groupon KDD）、ACM（Association for Computing Machinery）等科研组织或学会组织了一系列电子商务推荐相关的学术会议交流活动。在 1999 年召开的 SIGKDD 会议上，设置了专门的推荐系统研讨组，其研讨主题主要集中在电子商务领域的推荐技术和 Web 挖掘技术等方面；顺应发展形势，ACM 从 1999 年开始，每年定期召开关于电子商务的专题研讨会，在研讨会交流成果中，关于电子商务推荐系统的学术研究占据了相当大的比重;作为人工智能领域最主要的学术会议之一的 IJCAI 会议（International Joint Conference on Artificial Intelligence ，IJCAI），在第 7 届大会上把电子商务和智能推荐作为一个独立的研讨会，凸显电子商务和智能推荐技术的重要程度；在第 24 届 ACM SIGIR 会议上，会议组委会也单独把推荐系统作为一个专门的研讨主题；另外还有一些诸如 PAKM（知识管理应用会议）等国际会议也纷纷加大对商品推荐和推荐系统的关注[15]。大量知名的国际会议的重点关注和专题讨论，将电子商务领域的推荐系统研究、相关技术和产品的研发推到了广大科研工作者的视野，在一定程度上带动了智能化推荐技术的研究。

当前，在国外已有部分大型的电子商务网站采用了 Web 数据挖掘和推荐技术来提高电商平台上商品销售企业的利润。Web 数据挖掘技术在电商平台产品推荐的应用主要体现在客户关系管理（CRM）中，包括用户（消费者）行为分析、站点（网站）自适应、多维营销策略、用户感受改善、客户关系维持等多个方面[16][17]。电子商务推荐系统在国外较为突出的研究包括 IBM 的推荐系统、V5-7820 系统（NEC）、明尼苏达大学的 Schafer 等人的推荐系统、伊利诺伊大学的 Bamshad Mobasher 推荐系统研究、斯坦福大学 Mehmet 等人基于对话的推荐系统等研究。

国内在商品推荐方面的研究起步较晚，但经过几年的发展，已有部分研究取得了一定的成果。如：清华大学的个性化推荐系统 OpenBookmark。

电子商务推荐系统经过前期研究已经取得了大量的研究成果，在科研和产业界都引起了足够重视，但仍然存在如下问题：

① 商品推荐效率与推荐质量的矛盾平衡问题：电子商务推荐系统的推荐结果的准确度和实时效率是一对矛盾。当前大部分推荐技术是以牺牲商品推荐系统的质量为前提保证实时性要求的，同时，也有部分推荐系统推荐质量较高，但效率偏低。如何在推荐系统效率和质量之间寻求平衡点是需要重点研究的内容之一。

② 电子商务推荐系统缺乏具有自适应可扩展能力的体系架构：当前，大部分的电子商务推荐系统都还仅仅是一个单一的商品推送工具，只能基于简单的推荐模型进行初级的商品推荐。但由于电子商务平台网站系统结构的复杂性，网络环境和用户需求的多样性和多变性需要研究新型电子商务推荐系统体系结构，收集多种类型的数据，提供多种推荐模型，满足可自适应、可扩展的推荐需求。

③ 推荐可信度问题：由于整个网络环境的复杂以及信用体系的缺失，面对突如其来的推荐结果，用户往往首先持怀疑态度；电子商务推荐系统为了说服用户选择推荐的产品，需要提高推荐的可信度。目前的电子商务推荐系统只能通过简单的排名、评价评分信息等方式来达到上述目的，如何提高推荐结果的有效性和可信度是需要重点研究的内容之一。

④ 推荐属性单一：当前的商品推荐大多是基于商品的某种属性，如价格、用户评价、销量等信息进行选择和排名，对于在电子商务交易过程中买卖双方和商品属性之间的关联关系和潜在的类别关联考虑甚少。其实在实际交易过程中，某一类别的用户往往会青睐于购买属于他（她）这一类别用户类别人群的商品，而这种类别倾向不仅仅是单纯的商品分类，而是因社会行为、工作性质等各种关系形成的一种“类别”，如有部分人群在喜欢购买苹果手机的同时喜欢逛逛手机贴膜专区，而销售手机的电商同时也喜欢配套销售手机贴膜，这类买家和卖家在潜移默化的买卖关系中就形成了一个类别，这种潜在的类别对于推荐系统来说非常重要。

1.3 研究内容与创新点

1.3.1 主要研究工作

本文就异构商品网络中网络结构特征展开研究，在异构商品网络描述模型、异构商品网络排序、聚类分析、商品推荐等方面展开研究。异构商品网络是一个典型的异构信息网络，本文在异构商品网络分析模型的基础上，在基于异构商品网络的商品聚类、商品推荐等方面展开基于异构信息网络分析模型的应用研究。

根据电子商务推荐系统在全球范围内的快速发展进程，结合数据挖掘技术的相关理论和多种挖掘方法进行了系统的分析比较研究。针对异构商品网络分析的

个性化需求，并在已有的电子商务推荐系统架构的基础上，基于异构商品网络的聚类方法的推荐使用和改进了传统的依照用户评分矩阵的聚类算法，构建并实现了异构信息网络分析的电子商品数据聚类、推荐功能。

① 结合信息网络异构化发展趋势，基于网络分析、异构网络分析与商品推荐等国内外研究现状及问题分析提出了本文的研究内容与创新点。

② 借助形式化方法研究了异构商品网络描述模型，结合电子商务交易数据特征分析、构造与维护异构商品网络。

③ 基于异构商品网络描述模型，提出了基于异构商品网络分析的商品聚类算法，结合商品网络排序函数与排序模型进行异构商品网络中商品交易各参与对象的聚类分析，实现异构商品网络对象的类别挖掘和类别内的重要性排序。

④ 从商品交易记录中各种对象之间的关系维度出发，针对异构商品网络聚类结果，结合相应的推荐策略提出一种新型的商品推荐模型，分别从推荐思想、推荐流程、算法描述等方面进行详细研究内容的阐述。

⑤ 基于异构商品网络聚类的推荐研究，从软件工程的思想出发，从需求分析、系统功能设计、数据库设计和系统实现等方面设计并实现了商品推荐原型系统，从而验证了所提出的推荐模型的可行性。

1.3.2 研究创新点

按照主要研究工作的指导下完成异构商品网络分析模型关键技术研究以及商品推荐应用研究，研究具有以下的创新点：

从电子商品平台参与对象关系的维度出发，提出了基于异构信息网络分析的新型聚类算法。在该算法的指导下，以电子商务平台为基础，提出了基于异构服务网络分析的电子商品推荐聚类算法，基于电子商务平台各参与方对象及关系构建异构电子商务网络描述模型，基于电子商务排序模型构建聚类多维度量，借助网络划分和排序的迭代方法实现电子商品聚类及推荐等相关活动。根据聚类排序结果，结合相应的推荐策略，进行新推荐方法的实施运行工作，应用实验分析新推荐方法的推荐质量，包括推荐性能和推荐准确度等。

1.4 论文组织结构

本文共分六章，第二章是异构商品网络模型的基本概念；第三章对基于异构商品网络分析的商品聚类方法以及对应的聚类函数进行研究；第四章基于异构商品网络分析的手段结合推荐策略对聚类结果进行研究，并结合具体的异构网络形式进行应用研究，通过实验分析对研究方法进行对比分析；第五章是基于异构网络分析的商品推荐原型系统设计与实现，从软件工程的角度实现这个推荐方法；最后则是结论与展望。每章的内容详述如下所示：

第一章 绪论

介绍了异构信息网络分析相关的研究背景、现状以及本文的主要研究内容。

第二章 异构商品网络模型

分析了异构商品网络的结构特征，采用形式化的语言对商品网络和异构信息网络的模型及相关概念进行定义，分析电子商务交易数据特征并建立异构商品网络模型，并实现模型构造与维护并发执行，为论文异构商品网络分析提供了描述模型基础。

第三章 基于异构网络排序的商品聚类算法

基于异构信息网络分析的手段实现新型聚类算法。算法结合异构商品网络特征进行商品聚类算法研究，基于各参与方对象之间的关系及权重实现商品、卖家、买家在不同聚类划分中的排列分布，实现考虑排序的聚类过程。通过实验分析对算法进行评估。

第四章 基于异构网络分析的商品推荐模型

针对异构商品网络模型及商品聚类方法聚类结果排序模型，结合相应推荐策略进行商品的最优推荐，实现更高质量的个性化推荐方法。主要研究内容有分析商品推荐需求、描述推荐方法框架与流程、研究推荐策略并进行相关评估。

第五章 基于异构网络分析的商品推荐原型系统设计与实现

结合软件工程思想，针对基于异构网络分析的商品推荐系统进行系统的需求分析、功能设计、数据库设计及系统实现效果分析等。

第六章 结论与展望

对全文工作的总结和未来研究的展望。

1.5 本章小结

本章对论文研究的背景与意义进行了分析，在网络分析、异构网络分析、商品推荐等方面相关的国内外研究现状进行了阐述，对本文的研究内容和创新点进行了概要介绍，并对论文组织结构以及各部分之间的关系进行全局概述。

2　异构商品网络模型

许多实际应用中蕴藏着大量的与特定关联网络相关的事件或者组成部分，这些网络被称为信息网络，比如，互联网、高速公路网、电网、学术协作网络等等。显而易见的，信息网络是普遍存在的，而且确定了现代信息架构的决定性部分。在信息网络中，异构信息网络是一种包含多种类型节点的特殊信息网络。在分析了异构网络特征与商品推荐需求后，考虑以异构商品网络模型来进行商品的推荐研究。本章重点就异构商品网络模型的定义、电子交易记录提取、异构商品网络的构造与维护等过程为主线用形式化的语言进行详细描述。

2.1　异构信息网络模型

定义 2.1 异构信息网络。给定一个来自T种类型的数据集合$\chi=\{X_t\}_{t=1}^{T}$，其中X_t是属于t_{th}类型的数据集合，如果其中$T\geq 2$，数据集合χ构成的带权重的信息网络$G=\langle V,E,W\rangle$成为异构信息网络，其中W是权重矩阵，以任意边e为输入，边的权重w为输出，可以记为$W:E\to\mathbb{R}^{+}$，通过映射W，$\forall e\in V$存在$w\in\mathbb{R}^{+}$与之对应[18]。如图 2.1 为一个异构信息网络的简单拓扑结构图。

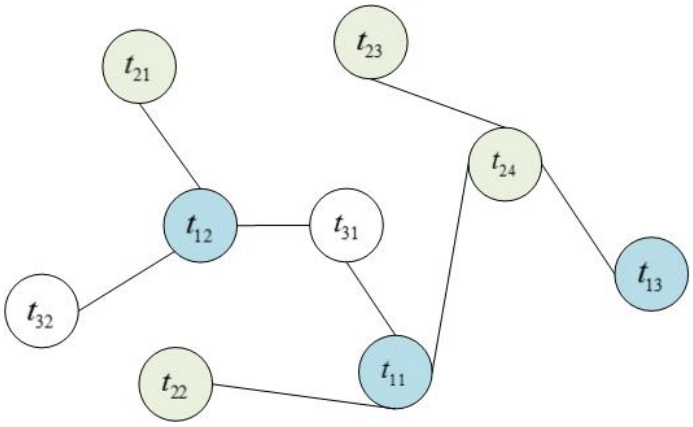

图 2.1　异构信息网络拓扑结构图

Figure 2.1 Topology of the Heterogeneous Information Network

在上图 2.1 所示的异构网络拓扑结构图中存在三种数据类型，数据对象依次为$T_1=\{t_{11},t_{12},t_{13}\}$ $T_2=\{t_{21},t_{22},t_{23},t_{24}\}$，$T_3=\{t_{31},t_{32}\}$。在异构网络中两种对象可能不直接关联，但是通过一种其他对象间接相连。图中可以看到T_1对象之间并不直接相连，但是通过T_2，T_3两种类型可以间接的联系在一起。在很多实际的应用中也可以发现这种模式，例如在一个文献的信息网络汇总，出版的作者与出版会议之间通过出

版文献联系在一起。如上图所示，某些对象类型之间并不存在直接联系。因此，根据研究的内容，描述两种对象之间的关系，成为了异构网络描述的记录方法。这里，以T_1类型为研究对象，可以在以上拓扑结构中提取T_1-T_2，T_1-T_3两种关系类型，可以表示为以下形式：

$$W_{12}=\begin{bmatrix}0&1&0&1\\1&0&0&0\\0&0&0&1\end{bmatrix},\ W_{13}=\begin{bmatrix}1&0\\1&1\\0&0\end{bmatrix}$$

以上矩阵只考虑链接情况，其中 1 表示对象相连，0 表示对象不相连。在具体应用中，不同类型对象的重要性不同，对象之间的关系也不同。例如在异构商品网络中存在，卖家，买家，商品，热点词几种对象的联系。如果研究对象是买家与卖家之间的关系，热点词产生的链接相比之下就不那么重要；如果研究对象是买家品味与热门商品，卖家产生的联系相比就不是很重要。因此，在具体的网络分析中，应该结合研究领域与研究对象，对权重函数W适应性的改变，一些常用的层次分析方法往往用于处理不同对象联系的权重关系。

2.2 电子商务交易数据特征分析

电子商务的覆盖范围较为广泛，通常可以分为三种类型：（1）B2B，即 Business to Business(企业对企业)，国内比较有代表性的有阿里巴巴、慧聪网、铭万等；（2）B2C，即 Business to Customer（企业对个人），跟国内比较有代表性的有天猫、京东、当当、凡客等；（3）C2C，即 Customer to Customer（个人对个人），比较有代表的有易趣网、淘宝网、拍拍网等。本文主要针对 B2C、C2C 混合类型的电子商务网站进行分析，如淘宝网[21][22]。

相对于普通的网络爬虫，针对电子商务网站的爬虫需要对 WEB 页面结果有足够以及 WEB 页面处理知识的了解，针对电子商务网站页面的特点有如下几个：

① 层次结构清晰，页面商品内容按照一定规则进行排布，商品详细信息在页面中位置固定。

② 商品详细页面内容较为固定，变动的多为评价、交易记录、价格等信息。

③ 商品详细页面结构清晰，商品价格、名称、图片等信息总是在固定位置，可通过工具获取得到。

电子商务的主要参与者是卖家，买家。纵然，时下有很多的电子商务平台，卖家仍然是电子商务活动的基本单元。买家则是电子商务活动的动力来源，是电子商务数据中最重要的参与者。在这两者之外，本文把商品中表现出来的热点词汇加以统计，以更好的描述聚类结果。另外可选的异构网络参与者，还有交易平台，商品类别等等。本文试图对电子商务领域卖家，买家，热点词进行聚类分析，

提取以商品为中心的星型网络模式，买家，卖家，热点词与商品相连接，为属性对象。使用这些链接关系，对商品，卖家，买家进行聚类，对热点词进行统计。本文约定大写字母 S 、 B 、 T 、 I 分别代表异构网络中卖家、买家、热点词、商品的对象集合。小写字母 s 、 b 、 t ， i 代表对应对象集合中的对象实例，提取的星型网络模式如下图 2.2 所示：

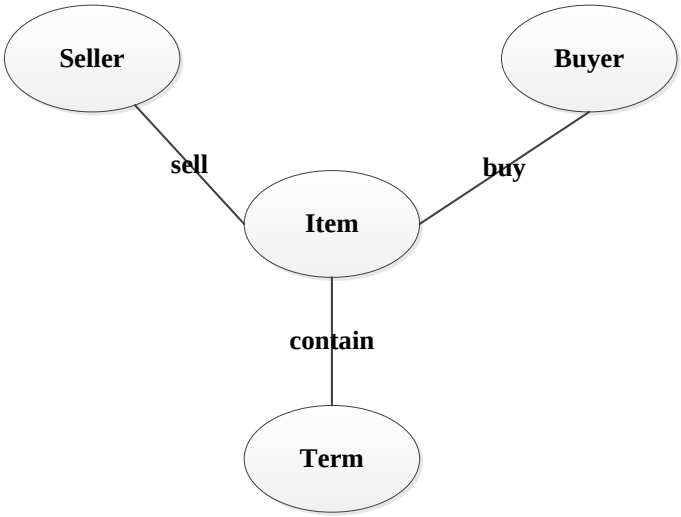

图 2.2 异构商品网络的星型网络模式

Figure 2.2 Star Network Schema of E-Commerce Heterogeneous Network

根据提取的星型结构，对应的电商异构信息网络中有 S 、 B 、 T 、 I 四种对象类型，应该存在 16 种不同类型的联系。但其中某些对象并不存在直接关联，例如买家与热点词不存在联系。在电商异构信息网络中存在的主要联系可分为如下几种:

① 购买与被购买关系（商品-买家关系）

异构商品网络中最重要的关系，买家与商品的关系。商品与买家之间存在购买与被购买的关系。两者的关系可以从电商的交易记录中获得。一般来说，买家与商品之间存在着一对多的关系，买家在信息网络中的重要性能够用购买商品的多少和价值描述。

② 标记与被标记关系（商品-热点词关系）

热点词是从商品描述中提取的。商品与热点词之间存在着标记与被标记的关系，从商品中某个热点词，相当于商品与热点词之间发生关联。一般情况，一个商品中包含多个热点词，一个热点词也可以被多个商品包含。热点词的价值，由关联的商品的价值客观反映。

③ 提供与被提供关系（商品-卖家关系）

卖家与商品之间存在提供与被提供的关系。卖家与商品之间存在多对一的关系，卖家在信息网络中的价值，有它所以同的所有商品的价值的总和表示。属性对象卖家可以有很多的目标对象商品与之相连。

2.3 异构商品网络模型

定义 2.2 异构商品网络。异构商品网络是一个由多种类型节点和节点间关系构成的商品网络。可以表示为 $G=\langle V,E,W\rangle$，其中 V 是网络中多种类型节点的集合，表示为 $V=S\cup B\cup T\cup I$，S 是所有卖家的集合，B 是所有买家的集合，T 是所有热点词的集合，I 是所有商品的集合。E 是网络中所有关系的集合，对于网络 G 中的连个节点 x_i，x_j，如果两者之间存在关系，则必有 $e=\langle x_i,x_j\rangle\in E$；$W$ 是网络中所有对象的关系矩阵，本文中用 w_{ij} 表示对象 x_i，x_j 之间的关系，若 $w_{ij}=0$ 则说明 x_i，x_j 之间不存在联系。

根据定义矩阵 W 可以表示为如下形式：

$$W=\begin{Bmatrix} W_{II} & W_{IS} & W_{IB} & W_{IT} \\ W_{SI} & W_{SS} & W_{SB} & W_{ST} \\ W_{BI} & W_{BS} & W_{BB} & W_{BT} \\ W_{TI} & W_{TS} & W_{TB} & W_{TT} \end{Bmatrix}$$

考虑 W 是一个对称矩阵，并且，属性对象 S、B、T 只与目标对象 I 发生联系，W 可以简记为：

$$W=\begin{Bmatrix} W_{IS} \\ W_{IB} \\ W_{IT} \end{Bmatrix}$$

例如，图 3.4 是包含九个节点的异构商品网络拓扑图，其中 $B=\{b_1,b_2\}$，$S=\{s_1,s_2\}$，$T=\{t_1,t_2,t_3\}$，$I=\{i_1,i_2,i_3\}$，可以提取关系矩阵如下所示：

$$W_{IS}=\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, W_{IB}=\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, W_{IT}=\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

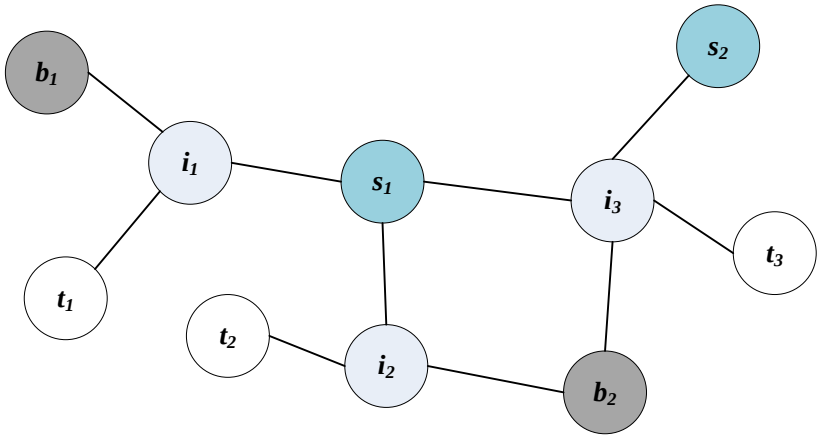

图 3.3 异构商品网络示例

Figure 3.3 Example of E-commerce Heterogeneous Network

2.4 异构商品网络模型构造与维护

异构商品网络的构造提取主要可以分为节点提取和关系提取两大部分。下面分别介绍：

2.4.1 节点提取

本文电子商务异构信息网络来源于某一电商平台的交易记录。在这个数据集中的节点提取的目的是发现异构网络中涉及到的卖家列表（Sellers），买家列表（Buyers），商品列表（Items），关键词列表（Terms）。记录聚类过程中可能用到的属性。提取到如下形式的结果：

$Sellers=\{s_1,s_2,\ldots,s_{|s|}\}$，其中 $s_1,s_2,\ldots,s_{|s|}$ 为卖家实例，$|S|$ 是卖家数量；

$Buyers=\{b_1,b_2,\ldots,b_{|B|}\}$，其中 $b_1,b_2,\ldots,b_{|B|}$ 为买家实例，$|B|$ 是买家数量；

$Items=\{i_1,i_2,\ldots,i_{|I|}\}$，其中 $i_1,i_2,\ldots,i_{|I|}$ 为商品实例，$|I|$ 是商品数量；

$Terms=\{t_1,t_2,\ldots,t_{|T|}\}$，其中 $t_1,t_2,\ldots,t_{|T|}$ 为热点词实例，$|T|$ 是热点词数量

2.4.2 关系提取

根据构造异构商品网络的需要，要在数据集中提取中心对象（Items）与属性对象的多元关系。在本文的情境中主要涉及三种关系：

① 商品与卖家的关系：

$Item-Seller=\{e\,|\,e=\langle i,s\rangle, i\in I, s\in S\}$，若商品实例 i 与卖家实例 s 之间存在联系，则两者的关系可以有图中的一条边 e 描述。

② 商品与买家的关系：

$Item-Buyer=\{e\,|\,e=\langle i,b\rangle, i\in I, b\in B\}$，若商品实例 i 与卖家实例 b 之间存在联系，则两者的关系可以有图中的一条边 e 描述。

③ 商品与热点词的关系：

$Item-Term=\{e\,|\,e=\langle i,t\rangle, i\in I, t\in T\}$，若商品实例 i 与卖家实例 T 之间存在联

系，则两者的关系可以有图中的一条边 e 描述。

2.4.3 关系权重确定

在异构网络中，不同属性对象节点类型与中心对象存在不同性质的关联。与不同对象的不可比性一样，对象之间关联同样是不可比的。在文中同一看待了不同属性对象与目标对象的关系，这种做法没有充分考虑到不同对象关联的性质不同的特点。但对于异构网络，本身存在的不同类型对象已经使用目标对象和属性对象进行划分，实质上是对同构网络分析的很大的突破。

在描述对象间“存在”与“不存在”关联的模式下，0 和 1 就可以很好的描述对象间的连接情况，其中 1 表示对象之间存在联系，0 表示对象间不存在联系。这种模式使用在同构网络中拥有较好的使用效果。但是在异构网络分析中，由于不同属性对象与目标对象的关联程度，对于整个网络的贡献不同，使用简单的 0-1 模式不能区分不同的链接关系。在这个网络中属性对象与目标对象的关系的重要性，可以使用权重描述。例如在异构商品网络中，*Item – Buyer* 关心使用 0.5 权重，*Item-Seller* 使用 0.3 权重，*Item – Term* 使用 0.2 权重，总体的权重之和为 1.这样在网络中，就能更好的使用 *Buyer* 和 *Seller* 节点的重要程度，形成针对这两种类型的异构网络聚类，不至于被 *Terms* 属性产生干扰。

然而，网络中属性对象权重的确定需要采纳相关领域专家的意见。由于本文涉及领域内尚无此类权重关系的权重探究，本文中依旧采用 0-1 表示法来记录网络链接关系。这样网络中的权重关系可以表达如下：

$$w_{x_i,x_j}=\begin{cases}1 & \text{如果} x_i\left(x_j\right)\in S\cup B\cup T\wedge x_j\left(x_i\right)\in I, x_i\left(x_j\right)\text{与}x_j\left(x_i\right)\text{之间存在连接}\\ c & \text{如果} x_i\left(x_j\right)\in T\wedge x_j\left(x_i\right)\in I, x_i\left(x_j\right)\text{在}x_j\left(x_i\right)\text{出现}c\text{次}\\ 0 & \text{其他}\end{cases}$$

很多文章使用的属性对象的相对数量作为网络中权重的判断依据，但 NetClus 算法也是以此为依据的。因此使用，相对数量会在一定程度上影响聚类的效果。

2.5 本章小结

本章主要介绍了异构商品网络模型的构造过程，重点介绍了异构信息网络模型、电子商务交易数据特征分析、异构商品网络模型及模型的构造与维护。从电子商务交易记录的数据爬取，结合异构信息网络模型的定义形成了异构商品网络模型，并且实时维护与构造，为后面的研究工作奠定基础。

3　基于异构网络排序的商品聚类算法

本章主要工作是改进异构网络聚类算法 NetClus，建立适用于商品聚类的异构商品网络排序函数和排序模型，同时分析该异构商品网络中的卖家，买家，商品，热门搜索词之间的关系，提出基于异构网络排序的聚类算法。最后，本文对提出的商品聚类方法进行实验分析，并对一些必要的数据加以说明。这对于传统同构信息网络来讲是一个重大的突破。一方面，兼顾多种对象在网络中的产生的关系，也使得信息的利用效率得到提高，产生更有实际意义的聚类结果。另一方面，在这样的异构信息网络中，聚类结果中表现出，某一具体领域相关的卖家与客户，这为今后的卖家推荐和用户群的划分做了必要的铺垫。聚类结果中商品和热门搜索词的划分，使得对于用户的个性化推荐得以实现。

3.1　异构商品网络排序函数

在信息检索和信息查询过程中，对返回结果进行排序是必要的过程，特别是在大量数据存在的情况下，排序更是必不可少；如何实现大量数据的高效、准确排序是提高信息检索和查询的关键。本节研究基于异构网络分析的商品排序方法，从网络关系的维度出发，重点研究基于异构商品网络分析的排序函数。

3.1.1　排序函数相关定义

排序函数在基于排序的聚类算法中是非常客观的，它不仅为对象提供排序分布去辨识其在聚类中的重要性而且作为特征提取工具去改善聚类质量。在现有基于网络排序思想的指导下，针对信息数据异构形式的结构分析，提出新型的基于异构商品网络分析的排序函数。在第二章中提出异构信息网络的概念，在异构商品网络中，网络对象有多种，传统的基于网络关联分析的排序一般针对同构的信息网络，对于异构商品网络不能直接应用，论文针对异构商品网络的结构特征拟采用特殊的排序函数实现对异构对象的排序，下面对具体的排序函数进行定义和分析[29][30]。

异构商品网络中，对象的全局重要性通过对象权值排序表现。对象加权过程实质是通过分析属性对象链接关系，确定买家，卖家等属性对象的全局重要性。在这一节中，将定义属性对象的加权函数和权值分布概念。其中加权函数将介绍简单加权和权威性加权两种。权值函数，是以链接关系为输入，对象权重为输出的函数，用以计算对象权。权值分布用来描述权重在不同对象间的分布情况。在本节中也将介绍异构商品网络中选取适当加权函数的方法，这一节将涉及一些信息论的基本知识，包括信息熵，相对熵，互信息等概念。下面是在本节中用到的三

个定义：

定义 3.1 排序分布和排序函数。一种类型对象的排序分布 $P(X)$ 是一个离散的概率分布，满足以下约束：$P(X=x)\geq 0(\forall x\in X)$，$\sum_{x\in X}P(X=x)=1$。一个定义在网络 G 上的函数 $f_X:G\to P(X)$ 以网络 G 为输入，权值分布 $P(X)$ 为输出，这样的函数称为排序函数。

定义 3.2 互信息。互信息是一个随机变量包含另一个随机变量信息量的度量。互信息是在给定另一随机变量知识的条件下，原随机变量不确定度的缩减量。考虑两个随机变量 X 和 Y，它们的联合概率密度函数 $p(x,y)$，其边缘密度函数分别是 $p(x)$ 和 $p(y)$。互信息 $I(X;Y)$ 为联合分布 $p(x,y)$ 和乘积分布 $p(x)\ p(y)$ 之间的相对熵。互信息的概念可以用于度量商品网络中属性对象间信息包含的程度。

$$I(X;Y)=\sum_{x\in\chi}\sum_{y\in\gamma}p(x,y)\log\frac{p(x,y)}{p(x)p(y)}=H(Y)-H(X,Y)$$

定义 3.3 正规化互信息（Normalized Mutual Information）。定义如公式 3.1 所示[39]。

$$NMI=\frac{\sum_{x\in\chi}\sum_{y\in\gamma}p(x,y)\log p(x,y)/p(x)p(y)}{\sqrt{\sum_{x\in\chi}p(x)\log p(x)\sum_{y\in\gamma}p(y)\log p(y)}} \tag{3.1}$$

排序常用来显示网络中对象的重要性和关联性，在 PageRank 对象的排序由它出链与入链的数量和链接网页的权重决定，权值越大网页静态的重要性越高。在 HITS 算法中，文档对于查询的排序反应了文档和查询之间的关联程度。在商品网络中，买家，卖家和热点词等对象的排序分布能够决定一个簇的性质，并且反应该属性对象在网络簇中的重要性。

商品网络中不同网络簇中的排序分布是十分不同的。举例来说，服装领域的排序分布实际上比家用电器领域的排序分布更加分散且均匀。在比较好的情况下，排序分布在不同的簇中应形成正交关系。根据一定的独立性假设，可以构建中心对象商品基于排序的生成模型。

现在介绍两种排序函数，并用异构商品网络举例，给出对于 3-类型星型网络中两种排序函数的一些性质。

① 简单排序

简单排序，顾名思义，仅计算对象共同发生的次数，并在类型内正规化。给定一个网络 G，每种属性对象的权值分布定义如公式 3.2 所示。

$$p(x|T_x,G)=\frac{\sum_{y\in N_{G(x)}}W_{xy}}{\sum_{x'\in T_x}\sum_{y\in N_{G(x')}}W_{x'y}} \tag{3.2}$$

其中 x 是 T_x 类型的一个对象。在异构商品网络中，如果使用简单排序计算属性对象权重，买家的权重可以定量的表示为买家购买商品数量权重之和，并在买家对象集合中归一化的结果。卖家和热点词的排序计算同理。

性质 1：给定包含三种类型节点的星型网络 $G=\langle X\cup Y\cup Z,E,W\rangle$ 其中 Z 是中心类型，对于 $\forall z,N_{G(z)}=\{x,y\}(x\in X,y\in Y)$，图 G 使用简单权值，加权 $P(X)$ 和 $P(Y)$ 来估计联合概率 $P(X,Y)$ 的编码误差是 $I(X,Y)$，$I(X,Y)$ 是 X 与 Y 的互信息。

证明：$\varepsilon=\sum_{x\in X}\sum_{y\in Y}p(x,y)\left[\log p(x,y)-\log p(x)p(y)\right]=I(X,Y)$

$=\sum_{x\in X}\sum_{y\in Y}p(x,y)\left[\log p(x,y)-\log p(x)p(y)\right]=I(X,Y)$

从上面简单的网络可以拓展到一般的星型网络，如果一种属性类型对象与其他属性类型对象具有较少的互信息，简单权值适合这种情况。在商品网络中，热点词与买家卖家属性对象类型具有较少的互信息。因此，可以使用简单排序计算热点词的排序分布。

② 权威性排序

对象的权威排序是一个考虑对象间权威性传播的排序函数，对于整个网络的可见性更具代表性。对于一般的星型网络 G，从类型 X 到类型 Y，通过中心对象 X 的权威性得分传播可以用如下公式表示，其中 W_{YZ} 和 W_{ZX} 是两种类型对象节点连接组成的邻接矩阵，必要的时候可以对他们进行正规化（如公式 3.3 所示）。

$$P(Y|T_Y,G)=W_{YZ}W_{ZX}P(X|T_X,G) \tag{3.3}$$

一般情况下，一种类型对象的排序得分可能是多种类型对象排序得分的整合。迭代方法排序分布实际上是利用幂法求矩阵的主特征向量，用以迭代的矩阵反映了两种对象间联系的紧密情况，两种对象可以看作星型网络模式中的一条（或者多条的结合）路径。

性质 2：给定包含三种类型节点的网络 $G=\langle X\cup Y\cup Z,E,W\rangle$ 其中 Z 是中心类型，对于 $\forall z,N_{G(z)}=\{x,y\}(x\in X,y\in Y)$，通过公式 3.3 迭代计算 $P(X)$ 和 $P(Y)$，之后估计联合概率 $\hat{P}(X,Y)=\{\hat{p}(x,y)=P(X=x)P(Y=y),x\in X,y\in Y\}$ 等价于排序矩阵 $M/\|M\|_1$ 代表的联合分布，这个矩阵使得 $\|W_{ZX}W_{ZY}-M\|_F$ 最小。

证明：　$USV^T = W_{XZ}W_{ZY}$ 是 $W_{XZ}W_{ZY}$ 的奇异值分解，U_1 和 V_1 是 U 和 V 的第一列，对应最大奇异值 σ_1。根据埃卡特-杨理论，$M = \sigma_1 U_1 V_1^T = \min_{\tilde{m}} \left\| W_{XZ}W_{ZY} - \tilde{M} \right\|$，其中 $rank(M)=1$。依照权威权值，$P(X) = U_1 / \|U_1\|_1$，$P(Y) = V_1 / \|V_1\|_1$，因此 $M/\|M\|_1 = \frac{\sigma_1 U_1 V_1^T}{\|\sigma_1 U_1 V_1^T\|_1} = P(X)P(Y)^T$，其中 $\|M\|_1$ 是矩阵 M 的 1-范数。

该性质适合简单网络。可以得到一个直观的概念：使用一维排序表示对象关系的条件下，权威排序能够获取网络中最大的构建结构，实际上是星型模型中的一条路径。因此，权威排序在一般情况下比简单排序效果更好。在电子商务数据集中两条经验主义的规则能够作为权威性排序的依据：

① 高排序的卖家能够吸引更多高排序的买家，在一个领域内经营出色的卖家往往能够吸引高消费的用户。

② 高排序的买家更多的光顾高排序的卖家，高消费的用户往往在领域内经营出色的卖家消费。

注意，这些经验规则具有较高的领域依赖性，有赖于相关领域专家的意见和领域内的规则。在这里，可以使用如公式 3.4 所示的迭代公式，使用权威性排序方法对属性对象加权。

$$\begin{aligned} P(S|T_S,G) &= W_{SI}D_{IB}^{-1}W_{IB}P(B|T_B,G) \\ P(B|T_B,G) &= W_{BI}D_{IS}^{-1}W_{IS}P(S|T_S,G) \end{aligned} \tag{3.4}$$

其中 W_{SI} 是卖家与商品的权重邻接矩阵，W_{BI} 是买家与商品的权重邻接矩阵，W_{IS} 和 W_{IB} 分别是以上两者的转置矩阵。D_{IB} 和 D_{IS} 分别是矩阵以 W_{IB} 和 W_{IS} 行和为对角线值的方阵。$W_{SI}D_{IB}^{-1}W_{IB}$ 矩阵能够反映 *Sellers* 对象和 *Buyers* 对象的连接关系。

3.1.2 异构商品网络排序函数选择

根据性质 1 与性质 2，可以为在电子商务数据集属性对象间找到合适的排序方法提供验证。表 3.1 中显示了不同数据规模数据集中，属性对象间的正规化互信息的大小。在互信息的统计过程中，采用了生成可能性最高的 100 个和 1000 个中心对象，并对涉及到的属性对象计算互信息。

表 3.1 不同规模商品网络属性对象间 NMI

Table 3.1NMI of the Atrributes Type in the Item Heterogeneous Network

	S&B	S&T	B&T
Top 100	0.42423	0.30035	0.25282
Top 1000	0.44275	0.25977	0.30693

从表 3.1 中，可以得知在该商品网络中，卖家与买家之间存在较多联系，具有较高的互信息量，分别为 0.42423 和 0.44275，如果使用简单排序的方式将造成比较大的计算误差，因此，计算买家和卖家的排序选择的是权威排序函数。

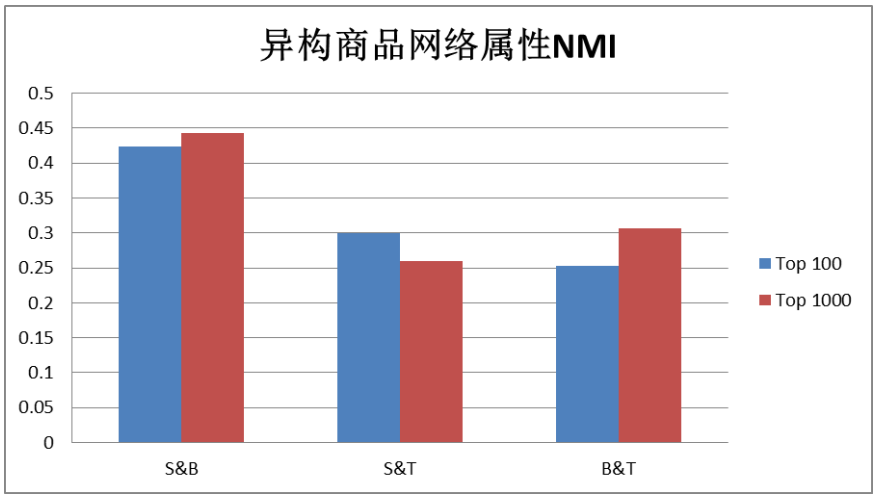

图 3.1 商品网络属性对象间 NMI

Figure 3.1 NMI of the Atrributes Type in the Item Heterogeneous Network

反观，热点词类型，与卖家、买家都拥有较少的互信息，适合选用简单排序的方式。

3.1.3 实例分析

考虑如下两种类型异构商品网络实例，其中包含 8 个商品节点，3 个卖家节点，2 个买家节点。为了探讨在 *items* 类型和 *seller* 类型使用简单排序和权威性排序的差异，可在网络中忽略热点词节点。一个简单两种类型节点的异构商品网络如下图 3.2 所示：

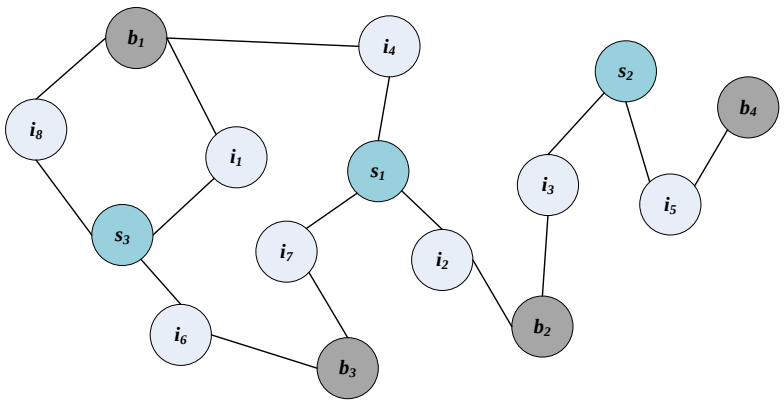

图 3.2 两种类型节点电子商务异构网络

Figure 3.2 Bi-Type of the Heterogeneous e-Commerce Network

提取权重邻接矩阵如下所示：

$$W_{IS}=\begin{bmatrix}0&1&0\\1&0&0\\1&0&0\\1&0&0\\0&1&0\\0&0&1\\0&0&1\\0&0&1\end{bmatrix},W_{IB}=\begin{bmatrix}1&0&0&0\\1&0&0&0\\0&1&0&0\\0&1&0&0\\0&0&1&0\\1&0&0&0\\1&0&0&0\\0&0&0&1\end{bmatrix}$$

为了对比简单排序与权威性排序的效果，对上述网络中 *items*类型对象和 *seller* 类型对象分别使用两种排序方法，排序效果折线图如下 3.3 所示：

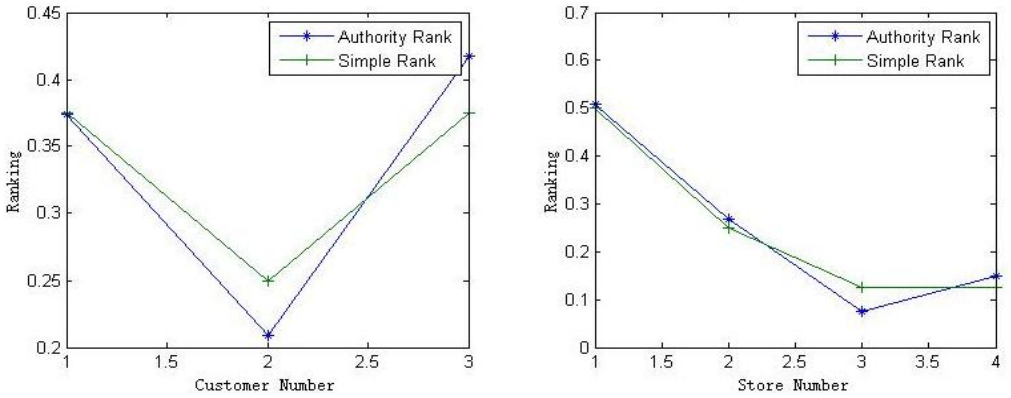

图 3.3 两种排序函数对比

Figure 3.3 Contrasts of Two Ranking Functions

从上图 3.3 可以看出，在具有高互信息的两种类型对象之间。使用权威性排序使得同种对象的中排序的分布更加分散，或者说辨识度更高。在实际的应用中，效果往往更好。

3.2 异构商品网络排序模型

在一些真实的信息网络中优先结合性和相称混合性普遍存在，这意味着在一个网络中如果一个对象拥有更高的度数，也就具有更高产生关系的可能性。在一般的情况下，属性对象在目标对象上的耦合程度越高，该目标对象具有更高与属性对象产生关系的可能性。在本文所用的异构商品网络中，27.43%的买家消费了数据集中 74.84%的商品，数据集中涉及的 60.37%商品出售于 37.29%的卖家。数据集中的统计也成了某些经济学的原理。例如帕累托法则，20%的用户带来 80%的收入；长尾理论中热门商品的交易特征等等。由此，可以做出这样的猜想对象在目

标对象在网络中的重要性，由它连接属性对象的数量决定，在一个网络图中，这一概念对应于节点的度数。

然而，在实际的网络分析中节点的度数并不能很好的反应节点的重要性。PageRank 算法一类的算法并没有把节点的度数（或者链接情况）作为节点全局重要性的直接度量，而是使用了节点的权值。因为在实际的信息网络中，高度数低权值的情况很普遍。在网页作弊算法链接农场中，大量垃圾网页之间存在很高的入链数量，然而，这些网页并不会因此增加与高权重网页的链接数量。在 DBLP 数据集中，在很多低水平会议发表文献，并不能增加在高等级会议上的发表的可能性。在电子商务数据集中，也往往存在这种的虚假买家，为了增加卖家的信誉程度或者是销量，这种的买家往往在比较糟糕的卖家消费，以提高自身在网络中的价值，进而提高卖家的声誉；另外一个例子是，在不同的领域内，销售情况是很不同的，在婚纱礼服的领域内销量或许远远小于服装饰品的销量。因此，权重相对于度数能更好的说明节点的全局重要性。

3.2.1 概率模型

如上所述，高权值买家，卖家，热点词在中心对象 items 上具有更高同时发生的可能性。因而，商品对象的首先被聚类，并用于能够说明当前电商网络划分。为了简化多种属性类型异构网络，试图将属性对象对中心对象生成模型影响因素化[35]。一个可行的办法是使用生成概率描述多种属性对象对中心对象的影响。在这里提出两条假设：

假设 1　在电子商务异构网络G，访问不同买家，卖家和热点词节点等属性对象的概率分布相互独立

假设 2　在电子商务异构网络G，访问同种属性对象类型中的不同属性对象的可能性也是相互独立的。

根据假设 1 在一个异构网络G中随机访问一属性对象的可能性可以表示为公式 3.5 的形式。

$$p(x|G)=p(T_x|G)\times p(x|T_x,G) \tag{3.5}$$

其中$p(T_x|G)$代表在这个网络中访问T_x类型对象的可能性，$p(x|T_x,G)$代表在网络G中T_x类型对象中访问x对象的可能性。例如，在电子商务异构网络中，随机访问一个买家节点的可能性可以表示为$p(b_i|G)=p(B|G)\times p(b_i|B,G)$，$p(B|G)$表示在网络中访问*buyer*节点的概率，$p(b_i|B,G)$表示在所有买家节点中访问$b_i$节点的概率。一般情况下，公式 3.5 中$p(T_x|G)$可以使用$T_x$属性对象和所有属性对象$\cup T_x$数量的比值估计。在同一种属性类型比较对象访问可能性的时候，这个值的作用可以忽略。公式中$p(x|T_x,G)$的值可以使用属性对象权值分布一节中的

方法近似估计。

根据假设 2，在同一种属性对象类型中，访问不同属性对象可能性相互独立，因此，同时访问来自同一属性对象类型的两个属性对象的可能性可以用如下公式计算， 该公式由概率论独立事件的乘法公式可得（公式 3.6）。

$$p\left(x_i, x_j \mid T_x, G\right) = p\left(x_i \mid T_x, G\right) \times p\left(x_j \mid T_x, G\right) \tag{3.6}$$

使用公式 3.5 和公式 3.6，可以计算随机访问属性对象的可能性。在这一前提下，可以计算相关目标对象的生成可能性模型。以电子商务异构网络举例，在网络中，一个中心对象商品 i 由一个买家消费，由一个卖家销售，包括一个或者多个热点词。因此，一个目标对象生成的可能性由多个属性对象决定，例如 $x_{i1}, x_{i2}, \ldots, x_{in_i}$ 其中 n_i 是与这个中心对象相连的属性对象的个数。这样，生成中心对象的可能性可以使用属性对象和与之相关的权重表示。根据之前的独立性假设 1,2，生成中心对象的可能性如公式 3.7 所示。

$$p\left(i \mid G\right) = \prod_{x \in N_{G(i)}} p\left(x \mid G\right)^{W_{i,x}} = \prod_{x \in N_{G(i)}} p\left(x \mid T_x, G\right)^{W_{i,x}} p\left(T_x \mid G\right)^{W_{i,x}} \tag{3.7}$$

其中 $N_{G(i)}$ 是中心对象 i 相连的属性对象的集合，或者说是中心对象 i 的邻接对象，$p\left(x \mid G\right)$ 是任意一个中心对象 i 的关联对象在 G 中的排序，$W_{i,x}$ 是对象 i 和 x 的关系权重，从公式 3.7 可以得知属性对象的生成概率越高，着中心对象生成概率越高。例如，在电子商务异构网络中，如果与中心对象商品关联的属性对象买家和卖家的生成可能性高，中心对象商品就有更高的生成可能性。

在公式 3.7 所示的概率模型中，为了避免可能出现的零可能性的情况，属性对象的相对排序应该使用全局排序进行平滑处理。具体处理使用如公式 3.7 所示。

$$P_S\left(X \mid T_X, G_k\right) = \left(1 - \lambda_S\right) P\left(X \mid T_X, G_k\right) + \lambda_S P\left(X \mid T_X, G\right) \tag{3.8}$$

其中 λ_S 是平滑参数，具体取值视情况而定，在之后的实验中，将对这一参数进行探讨。$P\left(X \mid T_X, G_k\right)$ 是在子网络（网络簇）G_k 中属性对象的相对排序，$P\left(X \mid T_X, G\right)$ 是整个网络中属性对象的排序。

平滑处理是一个信息检索的著名的技术。语言模型中使用平滑技术来处理文档关键字丢失的情况。计算目标对象的生成概率的时候，面临相同的问题。例如，对于一个网络簇中的中心对象商品，可能链接到一个或者多个相对排序为零属性对象，这样这个中心对象的生成概率就为零，它对于每个网络簇的后验概率都将为零，进而不能属于任何的簇。事实上，在初始化簇标记的时候，对象很有可能

被分配到错误的簇中，如果不适用平滑技术，可能将没有机会得到正确的簇。

3.2.2 后验概率模型

① 中心对象后验概率

网络簇中心对象的概率生成模型计算之后，就可以基于概率模型计算每个中心对象的后验概率。这里后验概率的具体含义是，中心对象在一个网络簇中生成的可能，论文采用贝叶斯公式来计算后验概率。

对于任意一个商品节点 i 属于任意子网 G_k 的概率分布可以用公式 3.7 定义的概率模型来计算 $P(i|G_k)$。为了对排序结果进行优化，下面用贝叶斯公式来计算商品对象属于网络 G_k 的后验概率。

$$P(G_k|i) \propto P(i|G_k) \times P(G_k) \tag{3.9}$$

在后验概率模型中，$P(i|G_k)$ 是商品节点 i 在网络 G_k 中的概率分布，$P(G_k)$ 是网络 $P(G_k)$ 的潜在相对大小。在研究中，为了预测网络 $P(G_k)$ 的值，用最大似然估计求解排序分布的最大似然值，其中，似然函数为：

$$\log L = \sum_{i \in I} \log(P(i)) = \sum_{i \in I} \log(\sum_{k=1}^{K} P(i|G_k)P(G_k)) \tag{3.10}$$

然后，基于 EM 算法求 $P(G_k)$ 的局部最优解，EM 求解的过程基于下面两个迭代公式：

$$\begin{aligned} P^{(t)}(G_k|i) &\propto P(i|G_k)P^{(t)}(G_k) \\ P^{(t+1)}(G_k) &= \sum_{i \in I} P^{(t)}(G_k|i)/|I| \end{aligned} \tag{3.11}$$

作为初始条件，设置 $P^{(0)}(G_k) = \frac{1}{K}$。

② 属性对象后验概率

使用中心对象后验概率的计算方法，能够在多维度量模型中衡量目标对象与簇的关系。然而，如何去衡量多维模型中的属性对象呢？根据电子商务异构网络的星型网络模式可知，属性对象与中心对象之间存在联系，而不与属性对象之间发生直接联系。因而，可以试图用与属性对象相连的中心对象的多维向量表示属性对象的多维度量。在商品网络中，试图对卖家这一属性对象聚类的时候，卖家的特征，往往取决于该卖家所经营的商品；同样，买家的特征，往往由买家所消费的商品的特征体现。因而，在电商异构网路中，属性对象 $x \in S \cup C \cup T$ 的后验概率可以使用如公式 3.10 所示。

$$\begin{aligned} P(G_k \mid x) &= \sum_{i \in N_{G(x)}} P(G_k, i \mid G_k) \\ &= \sum_{i \in N_{G(x)}} P(G_k \mid i) P(i \mid x) \\ &= \sum_{i \in N_{G(x)}} p(G_k \mid i) \frac{1}{\left| N_{G(x)} \right|} \end{aligned} \tag{3.10}$$

公式 3.10 的意义说明，属性对象的后验概率等于与之相连的中心对象的后验概率的平均值。例如在电商网络中，卖家的从属于网络簇的后验概率等价于与之相连的商品的后验概率的平均值。

3.3 商品聚类算法

聚类的定义：一个类簇的实体是相似的，不同类簇的实体是不相似的；一个类簇是测试空间中点的会聚，同一类簇的任意两个点间的距离小于不同类簇的任意两个点间的距离；类簇可描述为一个包含密度相对较高的点集的多维空间中的连通区域，他们借助包含密度相对较低的点集的区域与其他区域（类簇）相分离[50]。

当今的商品聚类方法，主要任务是为市场细分和用户细分提供依据，使电子商务平台能够针对性的对买家提供个性化服务，买家能够正在电商平台上找到自己想要的东西[55]。时下电子商务领域内聚类分析的应用可以归纳如下：

① 统计方法观点的聚类，也是传统观念的聚类分析。这种聚类主要依据同种对象间相似性或者距离的衡量，进行全局比较，进而完成聚类。这类方法主要应用在于通过已有数据的分析，如用户消费的方式、内容、水平等，了解数据中存在的规律。

② 机器学习观点的聚类，也称之为概念聚类。在机器学习的聚类中，聚类开始时簇没有标记，之后通过学习方法获得。概念的形成与聚类过程存在一种相辅相成的关系。在电子商务领域内的应用主要是，通过对用户动态的无规律的交易行为的分析，推导出消费客户、消费行为、消费方式的潜在类别关联关系模式，以挖掘倾向购买不同类别商品的特定消费群体。

③ 神经网络方法。神经网络通过学习待分析的数据模式来构造模型，一般可以对隐类型数据构造模型，用于非线性复杂数据。神经网络在电子商务领域内应用并不深入，但具有很好的前景。

异构商品网络聚类的任务是实现网络中买家和卖家的划分，使得同一簇中的买家具有相似的消费行为，同一簇中的卖家将具有相似的经营行为。本文的主要内容也在于理清电商网络中主要对象买家，卖家和商品的关系。

3.3.1 聚类思路

时下流行的聚类算法，K-means 类算法需要比较对象间的相似性；谱聚类算法

是一种图的切割算法，实质上是对网络中节点等同看待的；DBSCAN 类算法，基于空间中对象距离的，本质上也是需要对象直接比较的。这些算法对于同构信息网络是很有效的，但对于异构信息网络往往不能产生理想的将结果，原因在于忽略了对象间本质上的不可比性。

与传统对同一种对象聚类的方法不同，在本文研究中，基于异构商品网络模型，同时对多种不同的商品交易相关对象进行聚类。在聚类过程中，考虑不同对象之间的关系，建立基于网络分析的概率模型，以对象在不同网络划分簇中的概率分布为计算依据，评估对象离各个网络中心的距离，从而实现对不同对象的聚类。聚类算法的思路如图 3.4 所示：

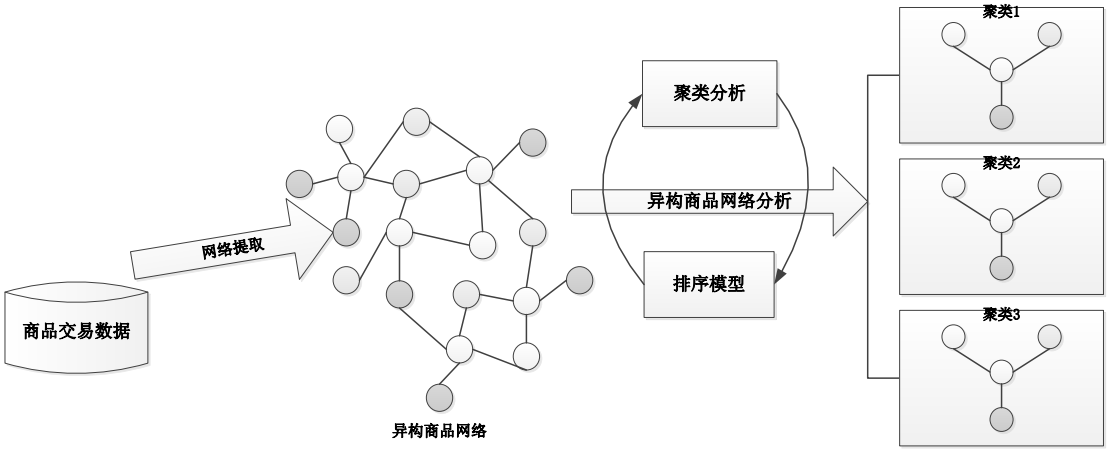

图 3.4 聚类思路

Figure 3.4 Idea of Clustering

异构商品网络聚类的步骤可以简要的表示如下：

① 数据清洗：电商网络中的数据清洗主要涉及一些残缺值的填写和异常值的排出。诸如丢失用户信息的交易记录，丢失卖家信息的商品信息；排除网络中出现次数极低的卖家与买家。

② 异构商品网络提取：根据分析需求在电子商务交易数据集中提取异构商品网络。需要结合电子商务相应的知识和分析需要，提取网络，并定义连接关系的相对重要性。

③ 异构商品网络聚类：是整个流程中最为重要的一步，是整个分析过程的关键。在商品网络聚类实现的过程中，很多具体的应用情况需要考虑，并反复推敲聚类方法的实现方法。同时，也要考虑算法的伸缩性，精准度等问题。

3.3.2 聚类算法流程

根据异构商品网络的定义，结合聚类思路，异构商品网络的聚类问题的主要算

法步骤表达如下图 3.5 所示。

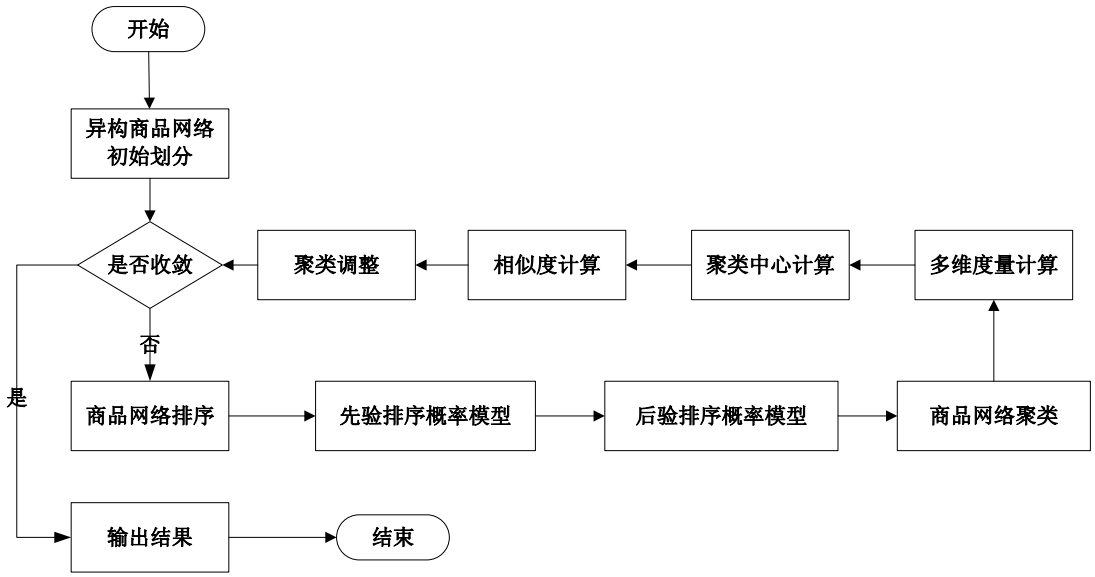

图 3.5 异构商品网络聚类算法框架

Figure 3.5 Framework of the Clustering HCN

如图 3.5 所示，描述了异构商品网络聚类的关键过程，具体描述如下：

① 异构商品网络初始划分：对于商品对象进行初始划分，按照商品对象在新网络中的分布获得包含卖家，热点词等属性对象的初始网络簇，记为 $\left\{C_k^0\right\}_{k=1}^K$；

② 聚类收敛性判断：循环判断聚类是否达到最大迭代次数或满足迭代终止条件；

③ 概率模型构造：构建商品对象在网络簇中概率模型，在该生成模型中包含买家，卖家等属性对象的权值分布和商品对象的生成概率，记为 $\left\{P\left(x\left|C_K^t\right.\right)\right\}_{k=1}^K$；

④ 后验概率模型构造：在概率模型的基础上，计算商品对象的后验概率记为 $\left\{P\left(C_k^t\left|x\right.\right)\right\}_{k=1}^K$；

⑤ 多维模型构建：使用后验概率度量商品对象每个网络划分中的出现的可能性 $(P(1|x),P(2|x),...,P(K|x))$。

⑥ 相似度计算：在每个子网聚类中，基于相似度计算方法计算聚类内部每个对象与每个聚类均值的相似度，根据计算结果决定最接近的聚类作为对象聚类调整的去向，进行聚类调整；

⑦ 重复③和⑥直到簇结构不再发生显著性变化。此时的异构商品网络可记为：$\left\{C_k^*\right\}_{k=1}^K=\left\{C_k^t\right\}_{k=1}^K=\left\{C_k^{t-1}\right\}_{k=1}^K$；

上述几个步骤中涉及了诸多的子过程下面加以解释：

① 簇结构显著性变化判断：对每次迭代过程中的聚类情况计算该次迭代与上次迭代的差距。换句话说，计算簇标记发生变化的对象相对于目标对象总数的百分比。如果百分比小于预先设定的阀值ε则停止迭代，输出聚类结果；另外考虑计算实现过程问题，如果迭代次数大于预先设定的值（Max_ITER）则停止迭代。就本文而言，一个簇内的商品不再发生超过阀值比例的数量变动，即可认为商品聚类完成。

② 加权模型构建：对于能够反映的网络簇性质的属性对象，买家，卖家，热点词，及其之间的联系。分析其中关联信息的多少，使用加权函数，评估对象的重要程度。

③ 多维度量模型构建：多维度量模型是在迭代过程中，针对商品对象而言的。每一个商品可以表示为在K维度量空间(K表示簇数量)中表示为一个K维的向量。每一个维度表示商品在第K个簇所属领域内生成的可能性。

④ 簇中心计算：在K为度量空间下，簇中心也是一个K维向量。电商网络簇中心由它包含的商品对象决定。商品对象与网络簇的距离可以采用余弦距离等方法计算，距离与小，相似度越高，该商品属于这个簇的可能性越大。

⑤ 聚类调整：基于相似度或距离公式，通过对比当前聚类划分中每个商品对象与簇中心的距离，获取与此对象最近的簇，赋值该对象的簇标记为这个簇的序号。

在不同的应用场景中，多维模型和节点加权的模型，应该根据应用场景和聚类需求而定。例如，本文所涉及的异构商品网络中，多类型的卖家表现出不同类型的特征。因而，使用K维度量能够有效的挖掘领域相关的簇。聚类稳定后，每一个维度代表一个领域的特征，进而达到领域划分的目的。

3.3.3 算法伪代码

综合上述过程，异构商品网络聚类算法的伪代码如下：

商品网络聚类算法

过程:异构商品网络聚类

输入:异构商品网络 对象类型$\chi=\left\{X_t\right\}_{t=1}^T$,权重图$G=\langle V,E,W\rangle$

权值函数 SimpleRank()，AuthorityRank()

簇数量 K

输出: K 个包含多种类型的簇，目标类型和属性类型的权值

// Step 0: 初始划分

$\{C_k^0\}_{k=1}^K$ =对商品初始划分

$\{C_b\}$ =背景概率模型

//重复 Step 1-3 直到 epsi＜ε 或者迭代次数超过 Max_Cluster_IterNum

For (iter=0; iter<Max_Cluster_IterNum&&epsi>ε ; iter++)

// Step 1: 为每一个网簇构建基于权值的概率生成模型

If 存在空簇: Goto Step 1

For i = 1 to K

G_i =与 C_i^0 邻接的子图

$$p(i|G_k)=\prod_{x\in N_{G_k(d)}} p(x|T_x,\mathrm{G}_k)^{W_{d,x}}\ p(T_x|G_k)^{W_{d,x}}$$ //商品概率生成模型

End For

//Step 2: 计算每一个商品的后验概率

EM 算法 估计每个簇潜在的簇大小 $P(G_k)$

$P(G_k|i)\propto P(i|G_k)\times P(G_k)$

For i = 1 to K

S_k = 得到 K_{th} 网络划分的质心

End For

//Step 3: 调整目标类型对象的簇标记

For each object d in G_i

For i=1 to k

计算距离 $D(i,S_k)$

End For

赋值 d 的簇标记为 $k_0=\arg\min_k D(i|S_k)$

End For

//Step 4: 在簇内计算属性对象后验概率

$$P(G_k|x)=\sum_{i\in N_{G(x)}} P(G_k|i)\frac{1}{|N_G(x)|}$$

//Step 5: 对属性类型对象聚类

$i=\arg\max_k P(G_k|x)$ 根据后验概率 $P(G_k|x)$ 计算属性对象聚类所属

3.4 实验分析

3.4.1 实验数据及环境

本文实验所用数据集是国内一个外贸电子商务平台（兰亭集势：http://www.lightinthebox.com）的数据集。数据记录时间为 2013 年 4 月 1 日至 2013 年 10 月 1 日。数据集中包括 69292 次交易记录，110899 个交易商品，平台注册用户 547.9 万。本文使用该数据集中提取的两个数据集。

① 数据集 1：对整个数据集中所有领域进行抽样的结果。信息网络中包括 1647

个交易商品节点，200 个买家，74 个卖家节点，163 个热点词汇。

② 数据集 2：在整个数据集上对“Home and Garden”，“Video Game”和“Electronics/Audio/Visual/Photography”三个领域抽样的结果。该数据集中包括 66 个买家节点，363 个交易商品，37 个卖家节点，79 个热点词节点。

数据集 1 将用于之后的参数分析章节；数据集 2 将用于之后的过程描述，精准度分析等章节，实验将从三个方面进行算法性能的测试。

3.4.2 聚类描述

在这一章节中将对实验过程中的一些统计量加以描述，试图用过程统计量描述聚类迭代过程中聚类情况的变化。其中主要涉及以下两个概念：

定义 3.4 相对熵。两个概率密度函数为 $p(x)$和$q(x)$之间的相对熵或 KL 距离定义如公式 3.12 所示。

$$\begin{aligned} D(p\|q) &= \sum_{x\in\chi} p(x)\log\frac{p(x)}{q(x)} \\ &= E_p \log\frac{p(x)}{q(x)} \end{aligned} \tag{3.12}$$

KL 距离能够反应两个概率分布之间的距离。在本文聚类过程中，属性对象每次迭代中具有不同权值分布。背景模型权值分布针对全局构建，对于迭代过程是一个常量。进而，可以使用迭代过程中的权值分布与背景模型的权值分布之间的 KL 距离描述迭代过程中属性对象权值变化。根据本文应用场景，使用平均 KL 距离 $avgD_{KL}$ 概念，定义如 3.13 所示。

$$avgD_{KL}(X) = \frac{1}{K}\sum_{k=1}^{K} D_{KL}\left(P(X\mid T_X, G_k)\|P(X\mid T_X, G)\right) \tag{3.13}$$

下图是数据集 2 中，三种属性对象的 $avgD_{KL}$ 在迭代过程中的变化。如下图所示，三种属性对象的 $avgD_{KL}$ 距离大致呈现随着迭代次数上升的趋势，并且在 10 次迭代左右趋于平稳。

因此，可以推论算法在该数据集聚类问题上具有收敛性，聚类结果能够从一定程度上反映网络中潜在网络模式，如下图 3.6 所示。

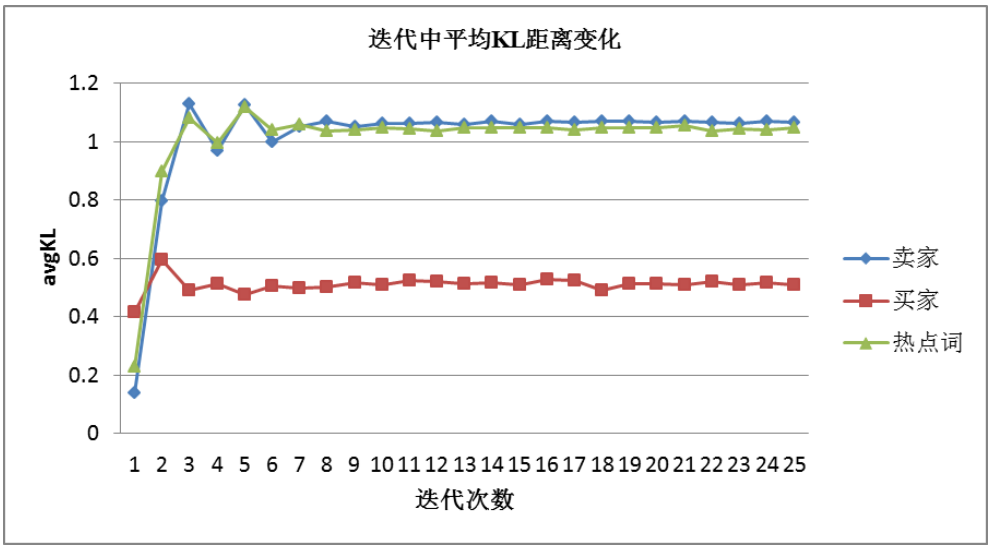

图 3.6 迭代过程中平均 KL 距离的变化

Figure 3.6 Change of Average KL Distance in Clustering

定义 3.5 紧凑度（compactness）。是一个结合了聚类凝聚度和分离度聚类度量概念，结合本文应用场景，异构商品网络聚类的紧凑度表示如公式 3.14 所示。

$$C_f = \frac{1}{|I|}\sum_{k=1}^{K}\sum_{n=1}^{|I_n|}\frac{s(i_{kn},c_k)}{\sum_{k'\neq k}s(i_{kn},c_{k'})/(K-1)} \tag{3.14}$$

其中 I_k 代表第 k 个网络簇中商品的集合，c_k 表示第 k 个网络簇的簇中心。从以上公式可以看出，网络簇中心对象距离从中心越近，同时距离其他簇中心越远，聚类的紧凑度越高。聚类过程中，紧凑度的变化如下图 3.7 所示：

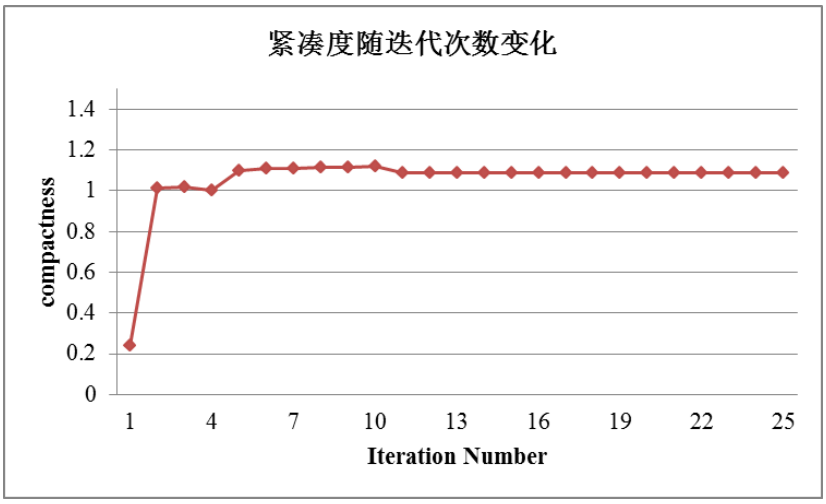

图 3.7 聚类过程中的紧凑度变化

Figure 3.7 Change of Compactness in Clustering

图 3.7 中，在聚类开始的阶段，由于随机的初始划分，聚类紧凑度较低，但随

着潜在网络簇的分离和中心对象的重新划分，每个中心对象更加接近正确的质心。因此，紧凑度呈上升趋势，如下图 3.8 所示。

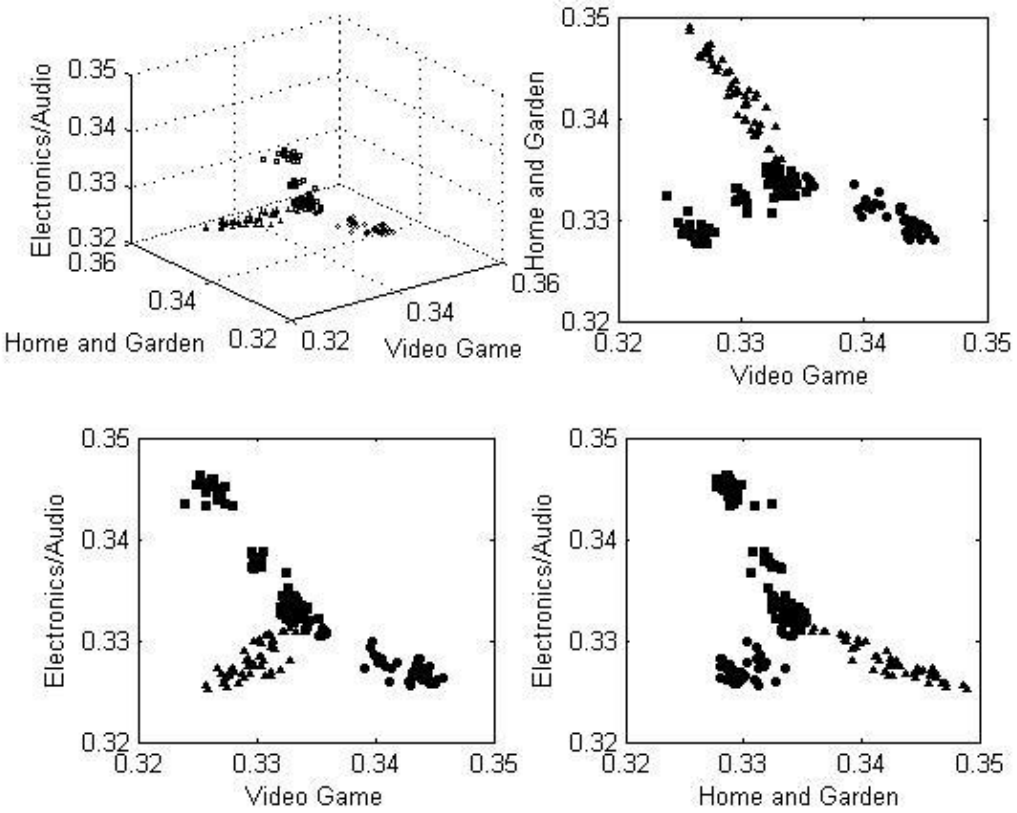

图 3.8 聚类结果空间显示

Figure 3.8 Clustering Result Showed In Multidimensional Space

图 3.8 中，是数据集 2 聚类结果在多维空间中的显示，空间的三个维度分别是聚类结果的三个领域“Home and Garden%”，“Video Game”和“Electronics/Audio/Visual/Photography”。从上图可以看出，在多维度量中，不同簇的商品对象可以明显分开。后面的三幅图是商品在两个维度上的投影，在不同的投影面上，一个类型的商品在一个维度上表现出高相似性，而在其余的维度相似度较低。因此，可知 NetClus 在异构商品网络上有良好的适用性。

3.4.3 精确度分析

① 精度

簇中一个特定类别所占的比例。在异构商品网络中，理想聚类结果中，一个簇只包含一个领域商品。因此，聚类的精度可以看作真实属于该簇所暗示领域数量与簇的大小的比值，如公式 3.15，其中 N_{ij} 代表在簇 j 中真实属于簇 j 所暗示的类别 i 的数量，$\left|I_j\right|$ 是簇 j 的大小.

$$precision\left(C_i, I_j\right)=\frac{N_{ij}}{\left|I_j\right|} \tag{3.15}$$

② 召回率

簇包含一个特定类所有对象的程度。在实际的聚类结果中，一个领域内的商品可能分散到不同簇中。一个簇的召回等价于回收该簇所暗示内商品与商品总数的比值，如公式 3.16 所示。

$$reccall\left(C_i, S_j\right)=\frac{N_{ij}}{\left|S_j\right|} \tag{3.16}$$

其中 N_{ij} 代表簇 j 中真实属于簇 j 所暗示的类别 i 的数量，$\left|S_j\right|$ 代表所属领域所有商品的数量。

③　F 度量

精度与召回率的组合，度量在多大程度上，簇包含一个特定类的对象和包含该类所有对象的程度。定义如公式 3.17 所示。

$$F\left(C_i, S_i\right)=\frac{2\times precision\left(C_i, I_j\right)\times reccall\left(C_i, S_j\right)}{precision\left(C_i, I_j\right)+reccall\left(C_i, S_j\right)} \tag{3.17}$$

对于本文的聚类结果，簇的精度越高，召回率越高，F 度量值越大反映聚类结果越准确。针对数据集 2 使用权威性排序方法得到的聚类的精准度度量如图 3.9。

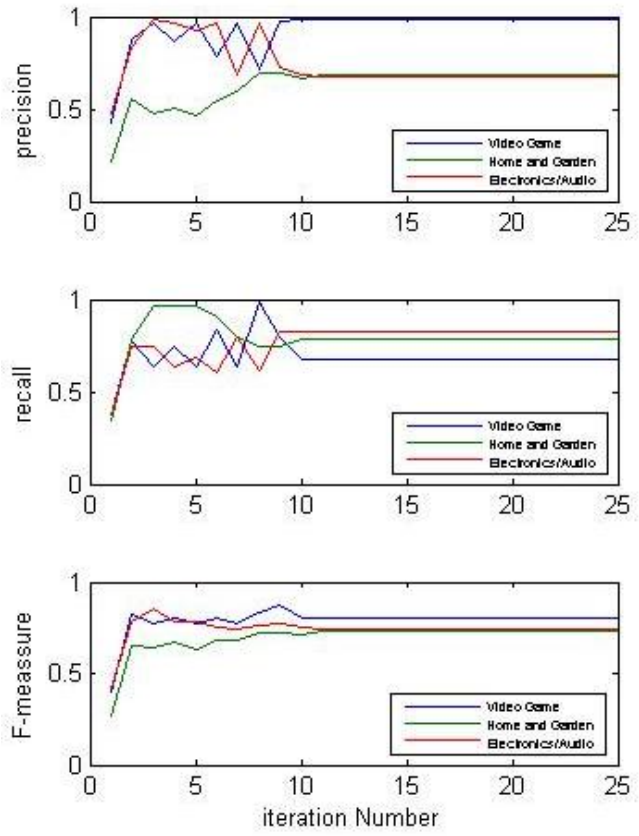

图 3.9 聚类精确度

Figure 3.9 Accuracy of Clustering

从图 3.9 中可以看出，迭代次数在小于 10 次时，伴随着大规模的簇调整，聚类的精准度，召回率，F 度量都发生很大变化。大于 10 次时，聚类趋于稳定。此时的精准度和召回率对于具体类别并不是全局最高的，但整体聚类结果具有很高的准确度。

3.4.4 参数分析

① 最大迭代次数

在商品聚类中，为了得到聚类迭代的最大聚类次数，拟以精度（precision）为聚类度量标准，计算不同迭代次数对应的聚类准确度，具体如图 3.10 所示。

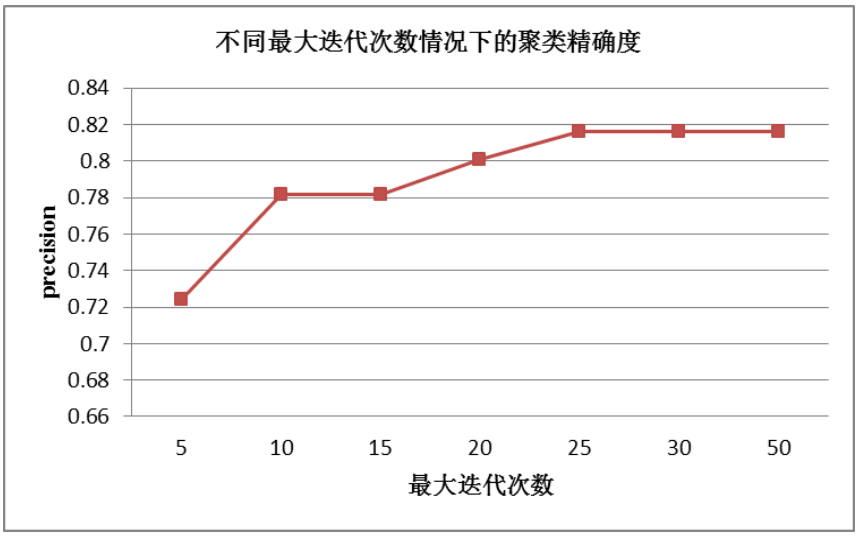

图 3.10 最大迭代次数下的聚类精确度

Figure 3.10 Accuracy of Clustering in the Maximal Iteration

如图 3.10 所示，当最大聚类次数小于 25 时，聚类准确度逐渐增加，说明，在此区间内，随着聚类迭代次数的增加，聚类的效果不断提升；当最大聚类迭代次数为 25 时，聚类准确度达到最大，超过 25 时聚类准确度几乎不变；因此，在论文商品聚类中，选取最大聚类次数 25 为默认的最大聚类次数。

③　平滑参数

在商品聚类的过程中，为了避免零概率事件，使用了加权平滑参数 λ_S，λ_S 反应了属性对象排序过程中对于背景模型的依赖程度。本文对于异构商品网络模型 λ_S 参数选取了 0.0 到 1.0 之间的 11 个值，并使用精准度作为评估标准。实验效果如下图 3.11 所示：

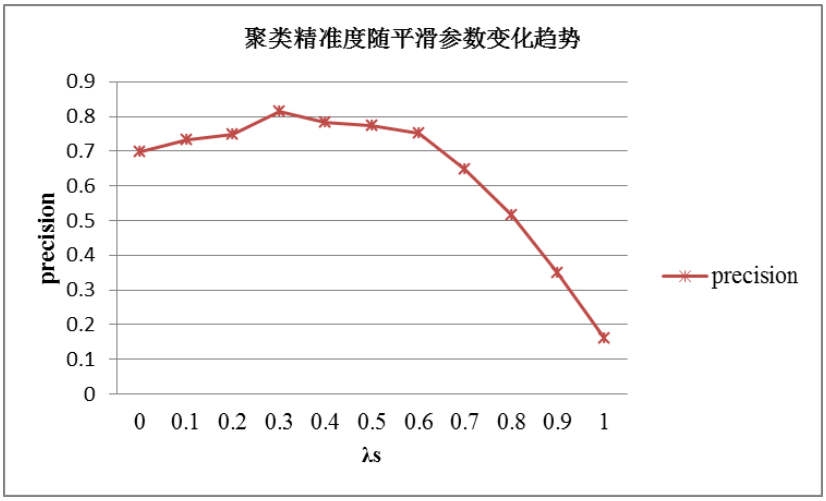

图 3.11 平滑参数对聚类精准度的影响

Figure 3.11 The Affect of the Smoothing Parameter on the Clustering Precision

从图 3.11 中可以看出，在本文的应用场景下，λ_S=0.3时聚类的准确度最高；$\lambda_S \le 0.6$时该参数对精准度的影响较小，并具有较好的聚类效果；$\lambda_S \ge 0.6$聚类精准度急剧下降；λ_S=1.0时，即使用背景模型作为属性对象权值分布时，聚类效果很差，不能形成关于领域的聚类。

3.4.5 时间复杂度分析

异构商品网络聚类算法 NetClus 的时间复杂度有以下几方面组成：

① 计算属性对象，卖家，买家，热点词的全局权重，时间复杂度为$O\left(t_1|E|\right)$；计算中心对象商品的全局权重时间复杂度为$O\left(|E|\right)$。其中$|E|$是商品网络中涉及关系的数量，t_1是使用权威性排序方法中，幂法求主特征向量的迭代次数。

② 计算属性对象的相对权值，时间复杂度是$O\left(t_1|E_k|\right)$，中心对象的时间复杂度$O\left(|E_k|\right)$。其中$|E_k|$是网络簇k中包含的链接关系。对于所有k个网络簇，时间复杂度的和是$O\left(t_1|E|+|E|\right)$

③ 计算商品对象后验概率的时间复杂度是$O\left(t_2(K+1)N\right)$，其中K是聚类数量，N是商品总数量，t_2是EM算法迭代次数。

④ 商品对象簇调整的时间复杂度$O\left(K^2N\right)$。由于每个商品对象使用一个$K-d$向量表示，簇调整需要计算每个商品鱼K个簇的距离，因此时间复杂度为$O\left(K^2N\right)$。

⑤ 计算买家，卖家，热点词的时间复杂度为$O\left(|E|\right)$。这是因为每个属性对象的后验概率由与之相连的中心对象计算，并且属性对象同样使用K维向量表示。因此时间复杂度是$O\left(K|E|\right)$。

综上所述，NetClus 算法在异构商品网络聚类过程中的算法复杂度是$O\left((t_1+1)|E|+t_3((t_1+1)|E|+(t_2+1)N+K^2)N+|E|\right)$,其中$t_3$是聚类调整的最大次数，可以简写如下：$O\left(c_1|E|+c_2N\right)$。

在实验室中，用程序模拟不同数量的网络节点组成的异构商品网络，运行聚类算法，记录运行时间，如图 3.12 是不同规模的网络所耗费的聚类时间：

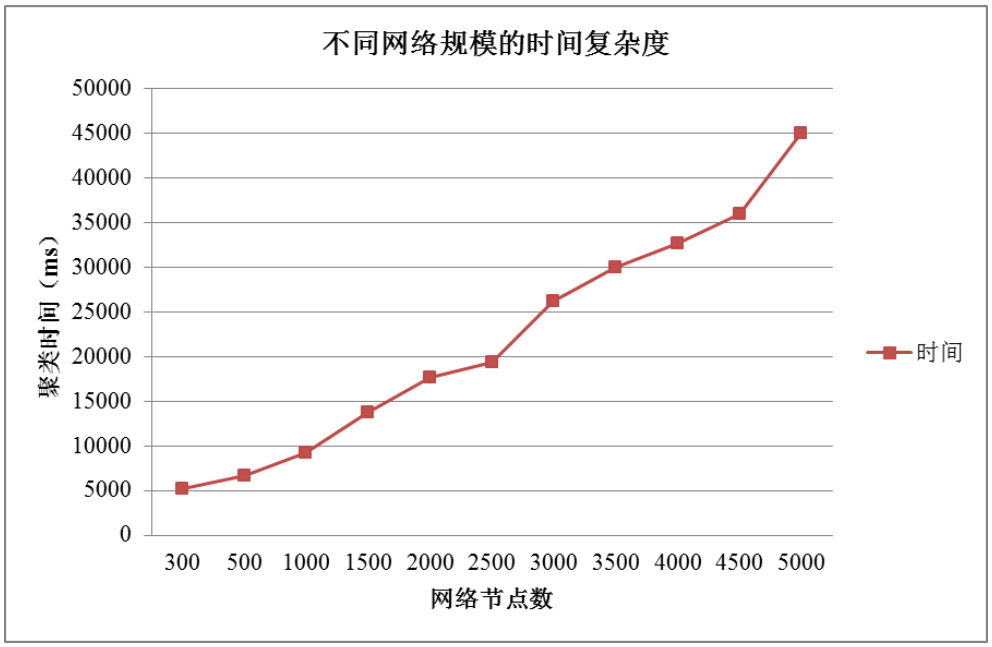

图 3.12 效率分析

Figure 3.12 Efficiency Analysis

如图 3.12 表明，不同规模网络聚类所耗费的时间在同一数量级上，伴随着网络规模的扩大，算法的时间耗费基本呈现线性增长。

3.5 本章小结

本章是基于异构信息网络分析的商品推荐方法研究的关键步骤，将第二章的异构商品网络结合异构商品网络概率模型进行聚类操作，得到收敛的聚类排序结果，为后面进行商品推荐提供了合理有序的数据处理结果。由此，本章主要介绍了异构网络排序模型、商品聚类、聚类评估及相应的聚类算法实验分析。

4　基于异构网络分析的商品推荐模型

商品推荐作为一种缓解电子商务系统数据过载并提高电子商务数据挖掘效率的技术，在过去的几十年里已经形成了较为成熟的理论发展体系。与传统商品推荐方法相比，本文提出的商品推荐方法充分考虑了电子商务中买家、卖家和商品以及商品热点词之间的潜在关联关系，提高了商品推荐中类别属性的推荐效率。下面就以基于异构网络分析的商品推荐方法中的思想、架构、流程、策略等详细阐述该推荐方法的核心研究内容[45]。

4.1　问题提出

商品推荐系统是在电子商务中利用关键技术提高用户推荐需求个性化的有效手段，旨在通过向客户提供其合适的商品的信息与相关建议，或者根据大量相似客户的历史交易记录信息，模拟线下销售人员向客户推荐合适商品并完成交易的过程。而除此之外传统的商品推荐方法，主要是根据用户兴趣模型和用户资源信息模型，这些应用主要来源于用户在交易过程中的历史数据分析而来。推荐系统根据相关的推荐策略结合资源信息的符合度及相似性大小并排序，从而将用户感兴趣的项目分门别类推荐给用户供其选择，作为推荐系统的输出直观呈现给用户作为选择。商品推荐应用的推荐策略主要有基于内容的推荐算法、协同过滤推荐算法、基于关联规则的推荐算法等。然而这些推荐算法会存在推荐质量差、推荐效率低、没有考虑类别属性造成类别混乱等问题。本研究在信息网络分析、数据挖掘等基础上，借鉴了国内外最新推荐方法研究，提出基于异构信息网络分析的商品推荐模型。

4.2　基于异构商品网络的推荐思想

与传统的商品推荐不同，本文提出的推荐模型是基于商品聚类的推荐，推荐过程充分考虑商品以及买卖双方之间的潜在类别关系，突破传统推荐方法大多局限“商品-用户”二维关系计算的模式，从交易记录中买家、卖家、商品以及热点词之间构成的网络关系出发，提取多维关系，建立异构商品网络模型，采用聚类和排序进行商品推荐候选列表的计算，从而为用户推荐具有类别属性的商品。如图4.1 所示，是本文提出的商品推荐方法的基本思想。

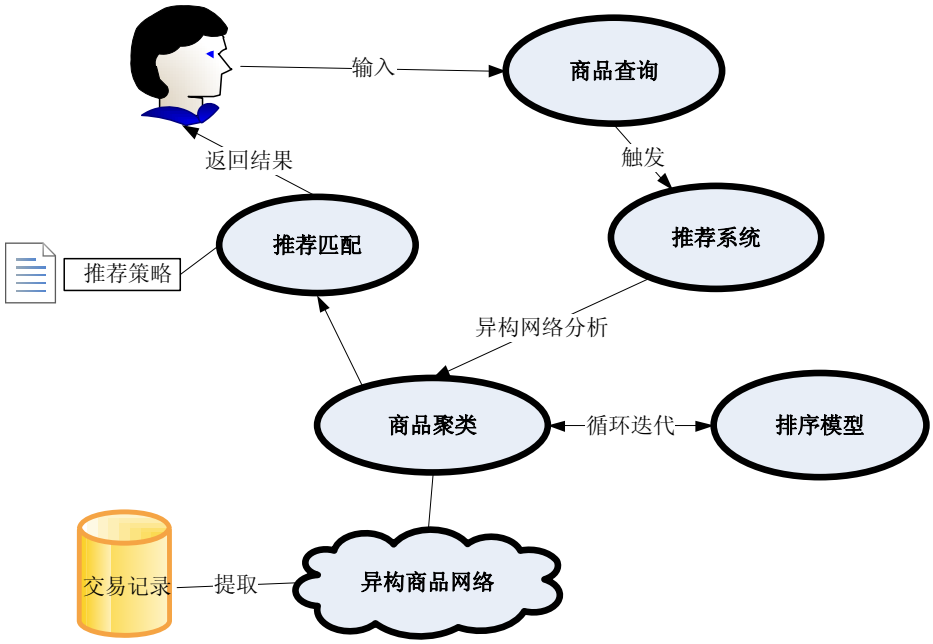

图 4.1　商品推荐思想

Fig.4.1 Item Recommendation Idea

从图 4.1 中可以看出，本研究的灵感思想主要是商务网站中买家在购物行为中的查询行动触发的一系列操作：买家在输入自己查询的关键词后，触发推荐系统的执行；推荐系统在基于异构网络分析的基础上，结合该买家的交易记录进行异构商品网络的提取与维护；结合相应商品聚类并循环迭代排序模型计算出商品聚类结果，获得买家历史交易记录中买家、卖家、商品、热点词的潜在类别关联，分析聚类簇中的关联关系；根据买家查询关键词结合推荐策略进行推荐的匹配动作，最后将推荐结果直观呈现给买家，实现推荐系统的个性化推荐。

4.3　基于异构商品网络的推荐模型描述

4.3.1　商品推荐模型架构

在分析了推荐思想的简单思想后，结合基于异构网络分析的商品推荐研究方法，进行商品推荐方法的框架描述，如下图 4.2 所示。

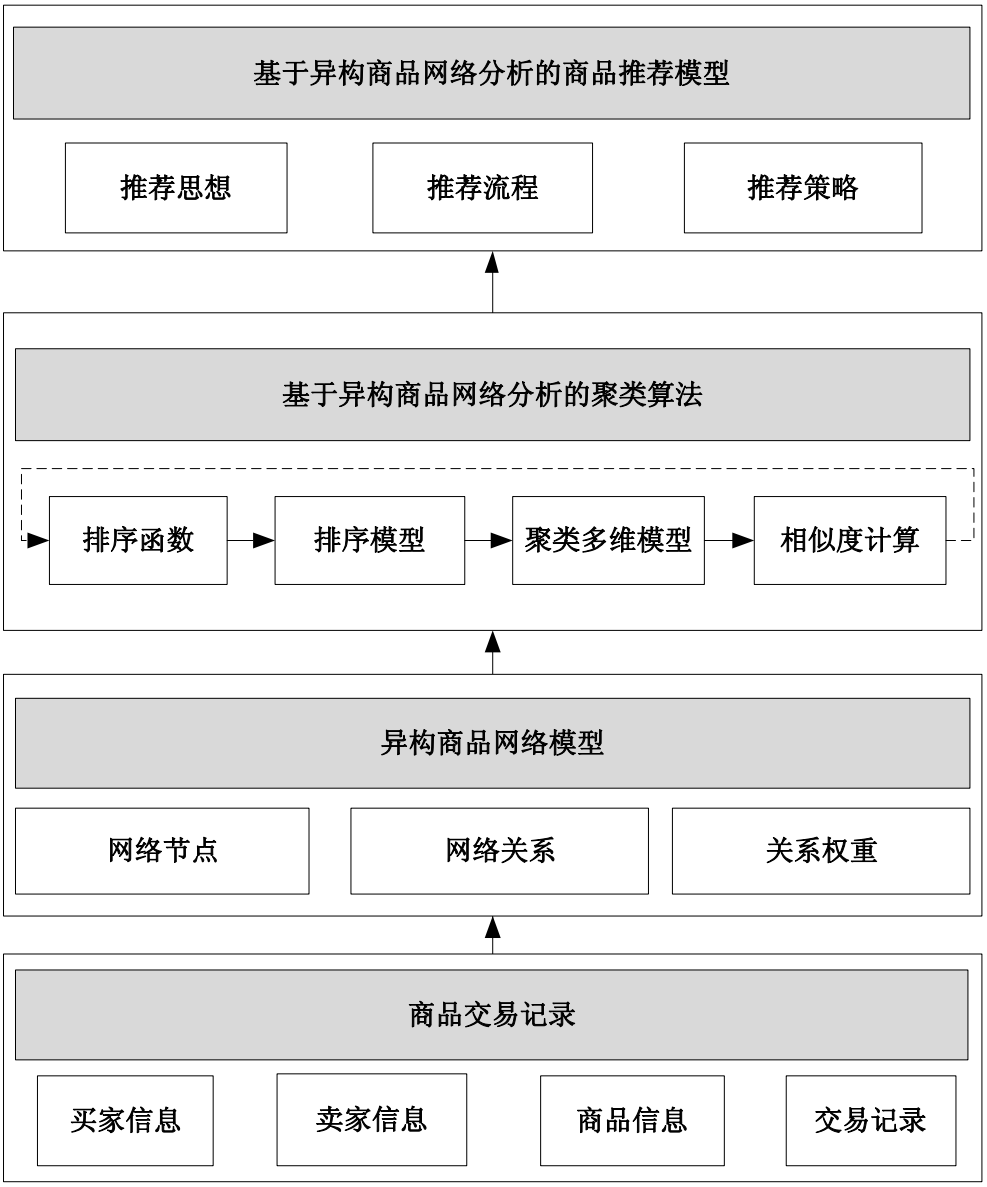

图 4.2 商品推荐架构

Fig.4.2 Item Recommendation Framework

在图 4.2 所示的商品推荐架构中：详细描述了本文提出的推荐模型的架构体系。在该架构中，包含了基于异构信息网络分析的商品推荐模型的全部架构组成部份。首先是商品推荐系统的基础部分商品交易记录，分别是买家信息、卖家信息、商品信息和交易记录的处理过程；然后是异构商品网络模型的提取与构建，结合商品交易记录，经过节点、关系、关系权重等提取，完成异构商品网络的构建；接下来是基于异构商品网络的聚类方法，包含了排序函数（简单排序和权威排序）、排序模型（商品排序模型和后验概率模型）、聚类多维度量、相似度计算等；最后是基于异构商品网络分析的推荐模型研究，分别从推荐思想、推荐流程、推荐策略等方面进行详细的设计与分析。这就是整个基于异构信息网络分析的商

品推荐模型的研究架构，后面的研究都是基于本架构进行的。

4.3.2 商品推荐策略研究

根据 4.3.1 的商品推荐模型架构，推荐策略是推荐过程的重要支撑。推荐策略是商品聚类完成后，根据买家属性进行推荐的根据，可以为买家快速匹配到与查询关键词相关的商品进行推荐活动，使得异构商品网络的推荐得以实现并进行个性化推荐，本文分新用户和老用户来描述用到的推荐策略，详细描述如下所示。

① 新用户

对于系统新用户，本文采用针对新用户的商品的推荐模型，算法描述如下:

针对新用户的商品推荐策略

输入：用户信息，用户查询关键词

输出：商品推荐列表

（1） 针对用户信息结合交易记录，判断该用户为新用户；

（2） 用户查询关键词触发推荐系统；

（3） 结合用户查询关键词，先筛选出与用户查询关键词类似的商品信息；

（4） 针对筛选出的商品信息，进行异构商品网络的提取与构造；

（5） 进行异构商品网络的聚类与排序操作；

（6） 将聚类结果中分类别选取排名较前的商品分门别类，作为商品推荐列表；

（7） 将商品推荐列表按类别推荐给新用户。

② 老用户

对于老用户，本文采用基于异构信息网络分析的商品推荐模型，将整个过程描述如下：

针对老用户的基于异构信息网络分析的商品推荐

输入：用户信息，用户查询关键词

输出：商品推荐列表

（1） 针对用户信息结合交易记录，判断该用户为老用户；

（2） 用户查询关键词触发推荐系统；

（3） 结合用户查询关键词，先筛选出与用户查询关键词类似的商品信息；

（4） 针对筛选出的商品信息，进行异构商品网络的提取与构造；

（5） 进行异构商品网络的聚类与排序操作；

（6） 提取用户历史交易记录中最大兴趣类别；

（7） 按照兴趣排名从聚类结果中选取排名靠前的商品作为商品推荐列表；

（8） 将商品推荐列表展示给老用户。

4.3.3 商品推荐模型流程

使用基于异构商品网络分析的聚类算法将所有注册的用户划分为多个具有相同用户特征的用户聚类簇，同一个项目只出现在一个聚类簇中。按照购买关系的潜在类别关联，提取购买交易记录的关联关系，使得卖家、买家、商品按照自身属性都聚类到相应的簇中。每个项目的聚类簇中的项目都具有极其相似的消费属性，也就是说同一个项目的聚类簇中的项目的行为属性也基本上相同。再根据用户查询关键词，结合聚类簇的结果，按照下面的推荐策略进行商品推荐，可大大提高商品推荐中类别推荐的效率与意义。下面是本文提出的推荐方法的流程描述：

① 用户访问购物系统：用户注册登录系统，提取用户信息，提交购买交易记录，判定是否为新用户；

② 用户查询：用户输入关键词，查询想要的商品 $Keywords$，$Keywords=\{k_1,k_2,...,k_m\}$；

③ 商品初选：推荐系统解析用户输入，从商品库中初选商品备选列表 $Items$，$Items=\{I_1,I_2,...,I_n\}$；

④ 网络提取：以初选出的商品列表中的商品为中心节点，提取历史交易记录中与之关联的买家和卖家、热点词等属性节点，构造异构商品网络模型；

⑤ 商品聚类：结合聚类算法和排序模型进行商品的聚类操作，使得用户交易记录按照异构商品网络模型聚类为相应可供后续操作的簇；

⑥ 推荐策略选择：按照不同的用户类别，选择不同的推荐策略，启动推荐匹配过程；

⑦ 推荐匹配：根据推荐策略，从候选商品中选择最优的实例列表 $\{I_1,I_2,...,I_k\}$；

⑧ 返回结果：将推荐结果返回给用户。

详细的商品推荐流程如下图 4.3 所示。

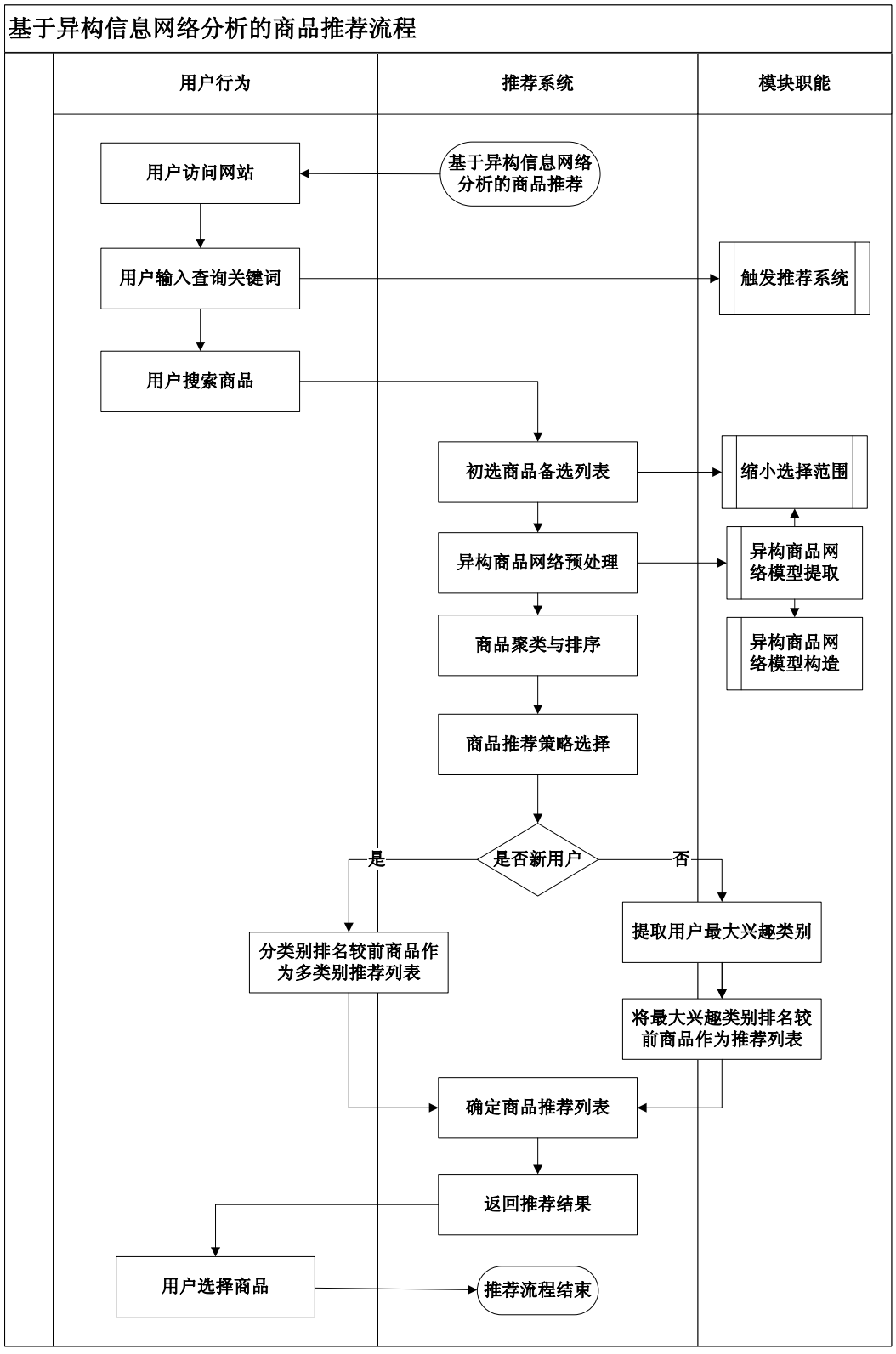

图 4.3 商品推荐流程图

Fig.4.3 Flow Chart of the Item Recommendation

如图 4.3 所示的商品推荐流程中所描述，基于异构信息网络分析的商品推荐流

程主要包括两个模块：用户行为和推荐系统。用户行为模块记录了推荐流程中用户的操作行为，主要包括访问网站，提供用户交易记录，判定用户是否为新用户；按照自己喜好键入查询关键词，从而触发推荐系统的执行；点击“搜索”按钮完成查询过程；推荐系统返回推荐结果后，按照自己喜好选择推荐商品。推荐系统则是推荐的主要部分，负责处理用户的查询请求并分析用户相匹配的商品分门别类进行推荐，主要包括根据用户查询关键词确定初选商品列表，缩小聚类范围，提高系统运行功能；针对初选商品列表结合异构商品网络进行预处理，包括异构商品网络的提取与构造；利用商品聚类和排序模型对异构商品网络进行分析，获得排好序的聚类簇；根据是否为新用户分别结合不同推荐策略进行推荐，如果是新用户，将聚类簇中与用户查询关键词有关联关系的商品簇中排名较前的多类别推荐给用户。如果是老用户，则提取用户的最大兴趣类别，将该最大兴趣类别中排名较前的商品推荐给用户；将商品推荐列表推荐给用户经用户筛选，从而完成了推荐系统的整个推荐过程。

4.4 推荐系统模型案例分析

为了对论文上述提出的基于异构商品网络分析的商品推荐模型进行更加形象化的说明，下面用一个案例阐述整个商品推荐过程和结果。简单起见，假设在某一个电子商务交易平台中有 14 条“苹果”相关的交易记录，分别由 6 个买家用户（$b_1, b_2, b_3, b_4, b_5, b_6$）与 2 个卖家用户（$s_1, s_2$）之间发生买卖关系，具体的交易记录如表 4.1 所示。

表 4.1 交易记录实例

Table 4.1 Transaction Records Example

序号	买家	卖家	商品	热点词
1.	b_1	s_1	i_1（iPhone 5）	iPhone
2.	b_1	s_1	i_3（iPad 3）	iPad
3.	b_2	s_1	i_2（iPhone 4）	iPhone
4.	b_2	s_1	i_3（iPad 3）	iPad
5.	b_2	s_1	i_1（iPhone 5）	iPhone
6.	b_3	s_2	i_4（苹果 MacBook Pro MF839CH/A）	MacBook
7.	b_4	s_1	i_5（苹果 MacBook Air MJVE2CH/A）	MacBook
8.	b_4	s_1	i_3（iPad 3）	iPad
9.	b_4	s_1	i_7 (苹果 MacBook ProMD101CH/A）	MacBook
10.	b_5	s_1	i_5（苹果 MacBook Air MJVE2CH/A）	MacBook
11.	b_5	s_1	i_7（苹果 MacBook Pro MF840CH/A）	MacBook
12.	b_6	s_1	i_5（苹果 MacBook Air MJVE2CH/A）	MacBook
13.	b_6	s_1	i_6（苹果 MacBook Pro MF840CH/A）	MacBook
14.	b_6	s_1	i_7 (苹果 MacBook ProMD101CH/A）	MacBook

按照第二章异构商品网络的定义，实例中的交易各方参与对象（暂时不考虑热点词）可以用如图 4.4 所示的异构网络图表示。

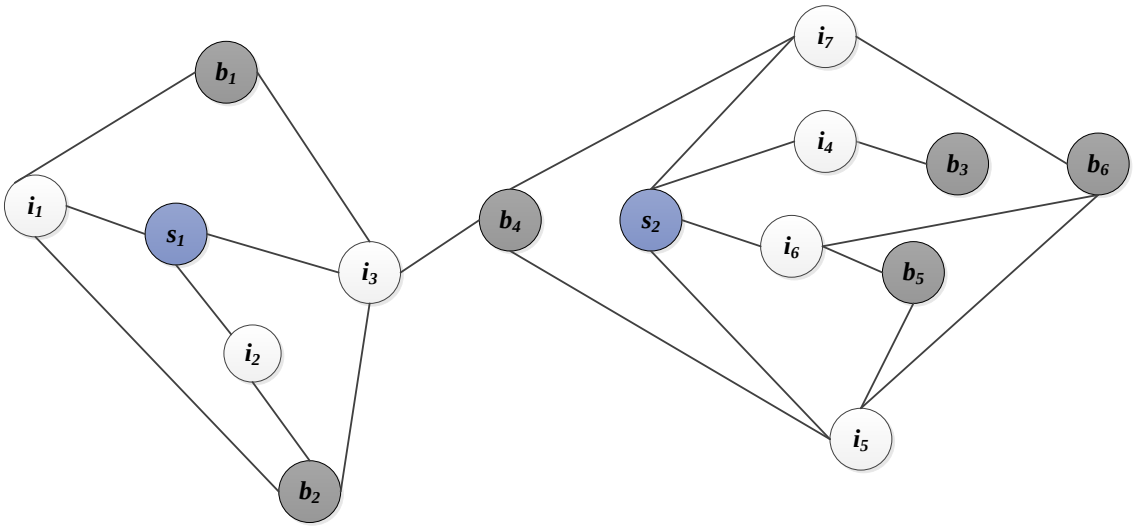

图 4.4 实例异构商品网络

图 4.4 Example of E-commerce Heterogeneous Network

由图 4.4 所示的异构商品网络，按照不同的购买倾向，买家b_1和b_2与卖家s_1所销售的i_1、i_2和i_3分布在一个网络中，属于移动端苹果产品的类别；而同样的，右侧是由卖家s_2以及商品i_4、i_5、i_6、i_7和买家b_3、b_4、b_5、b_6组成的网络，倾向于笔记本系列苹果产品的类别。

下面将实例异构商品网络采用本文提出的推荐模型进行商品推荐，首先按照异构商品网络聚类算法将买卖双方和商品聚为如图 4.5 所示两个类别。

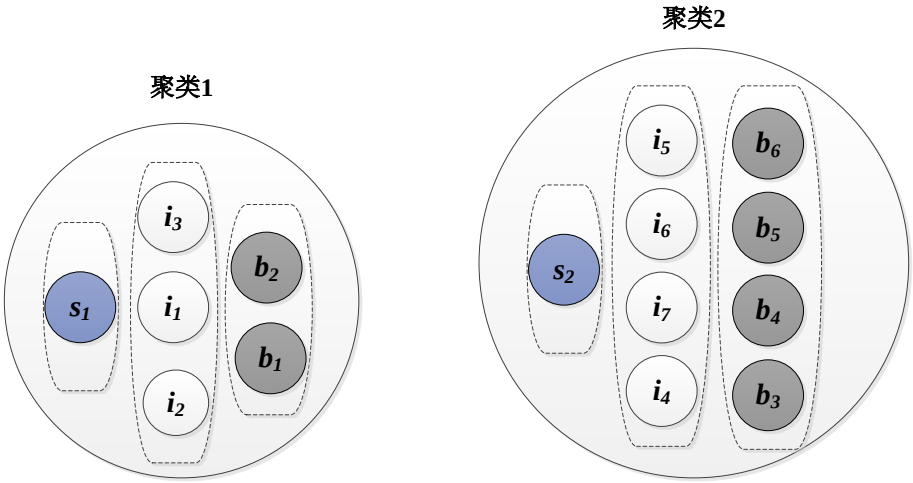

图 4.5 异构商品网络聚类结果

Fig.4.5 Clustering Result of Example Heterogeneous Network

在聚类之后，下面结合不同买家用户，采取不同的推荐策略进行商品推荐，这里模拟两种情况下的商品推荐：新用户商品推荐和有历史记录的用户商品推荐。

① 当一个新用户到达电商平台时，尚未有历史交易记录，假设该用户通过电子商务网站搜索苹果相关的商品；后台通过用户的输入就可以检索到当前平台上

的上述 14 个交易记录；通过构造异构网络、聚类分析，将相关的记录中的商品聚集到如下两个聚类中，并进行了排序：聚类 1={i_3，i_1，i_2}，聚类 2={i_4，i_5，i_6，i_7}，并将两个聚类的结果按照排序概率推荐给该用户，用户可以根据自己的兴趣爱好选择两个类别中不同的商品。

② 当一个有历史记录的用户如b_3访问平台时，后台监测到用户搜索苹果相关的关键词；通过异构网络构造和聚类过程，发现b_3按照网络关系被聚集到和商品i_5、i_6、i_7、i_4同样的类别中，系统按照聚类 2 的商品排序，将商品i_5、i_6、i_7、i_4推荐给买家用户b_3。当然，买家用户b_3上次访问系统后，可能又有新的用户发生交易记录，推荐的结果可能会发生变化，例如可能会推荐一种新型的苹果 MacBook Air 系列产品给b_3。

4.5 推荐模型分析

论文提出的以异构商品网络聚类为核心的商品推荐方法，综合考虑交易记录中各参与方对象之间的网络关联关系进行基于概率分布模型的聚类，以聚类结果为依据进行不同类别区分的商品推荐。该方法在解决商品推荐问题上具有一定的新颖性，但仍然存在一些问题：

① 推荐评估问题：当前的推荐仅在推荐模型思路上从异构网络、聚类分析和排序分布等方面进行新的角度进行创新性研究，但是在推荐结果准确性、高效性等评估还有待进一步深入研究，拟结合买家用户反馈，结合语义识别技术进行后续研究。

② 多方推荐问题：当前的异构网络分析为基础的聚类分析结果的多方应用还不够，当商品、买卖双方和热点词等被聚集到不同的类别并排序之后，如何利用类别关联特征进行商品、买家、卖家之间的多方推荐，如用户推荐、卖家店铺推荐等都是今后需要继续研究的内容。

4.6 本章小结

本章在基于异构商品网络分析的聚类基础上，结合商品交易记录进行商品推荐方法的具体研究。从商品推荐方法思想、推荐方法流程、推荐方法策略等方面，对课题进行深入研究并用商品推荐模型描述了整个推荐过程，清晰客观展现了课题研究思路并为下一章原型系统的实现与设计奠定了理论基础。

5　基于异构网络分析的商品推荐原型系统设计与实现

在分析了基于异构商品网络的推荐方法后，结合软件工程的相关理论，本章着重研究商品推荐原型的设计与实现。从软件工程的角度，考虑并验证基于异构网络分析的商品推荐方法的可行性及创新性。

5.1　异构商品网络推荐系统需求分析

为了实现基于异构信息网络进行商品推荐，本文结合前面研究内容，从软件工程领域进行着手并设计实现该研究内容。需要结合系统需求进行整体的开发与设计，下面详细介绍商品推荐系统的功能需求，如图 5.1 所示。

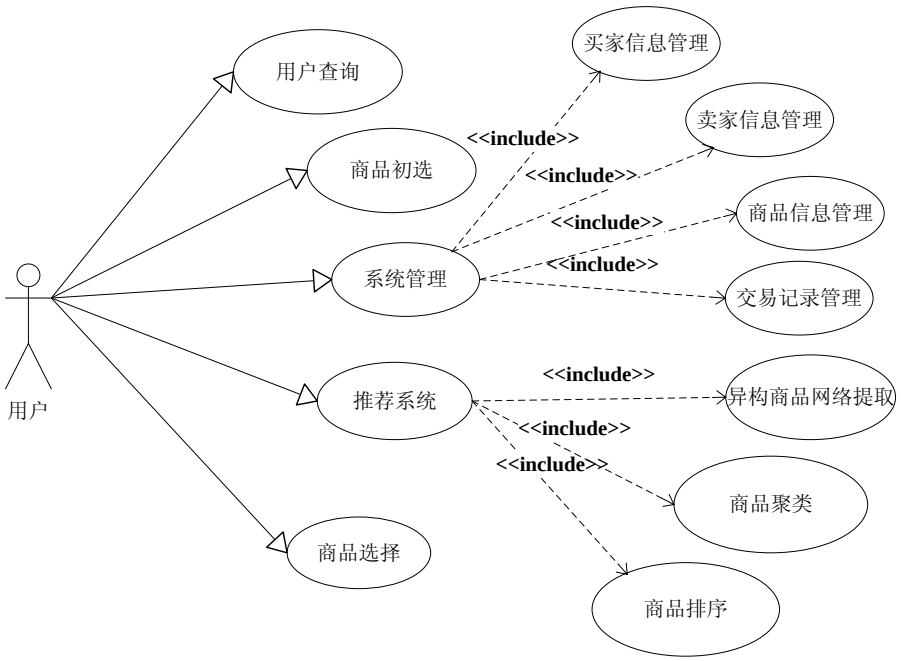

图 5.1 商品推荐用例图

Fig.5.1 Use-case of the Item Recommendation

如图 5.1 所示，推荐系统的主要功能有：用户查询，触发推荐系统执行；系统管理则包括了买家、卖家、商品和交易记录的信息基本管理功能；商品初选是按照查询关键词将系统管理中与之匹配的商品初选出来，缩小选择范围；商品推荐是结合异构商品网络，针对买家商品交易记录进行异构商品网络提取、商品聚类及商品排序操作，完成基于异构商品网络分析的推荐工作；商品选择，买家在推荐系统的商品推荐列表中按照自己喜好进行商品的选择。

具体功能描述如下表 5.1 所示。

表 5.1 系统功能分布

Table 5.1 Function of System

序号	功能项	参与者	描述
1	用户查询	买家	买家按照喜好查询感兴趣商品
2	买家信息管理	系统	管理买家基本信息
3	卖家信息管理	系统	管理卖家基本信息
4	商品信息管理	系统	管理商品基本信息
5	交易记录管理	系统	管理商品交易记录
6	商品初选	系统	按关键词初选商品范围
7	推荐系统	系统	结合查询词与异构商品网络进行推荐活动
8	异构商品网络提取	系统	提取并构造异构商品网络
9	商品聚类	系统	对异构商品网络进行聚类处理
10	商品排序	系统	聚类过程中结合排序函数进行商品排序
11	商品选择	买家	根据推荐系统推荐列表选择商品

5.2 基于异构商品网络模型的系统功能设计

为了更好的将算法应用到实际当中，通过设计原型系统用于模拟真实情况下。原型系统的架构当中有：用户查询商品、系统管理模块、推荐系统模型模块及推荐商品列表展示模块如图 5.2 所示。

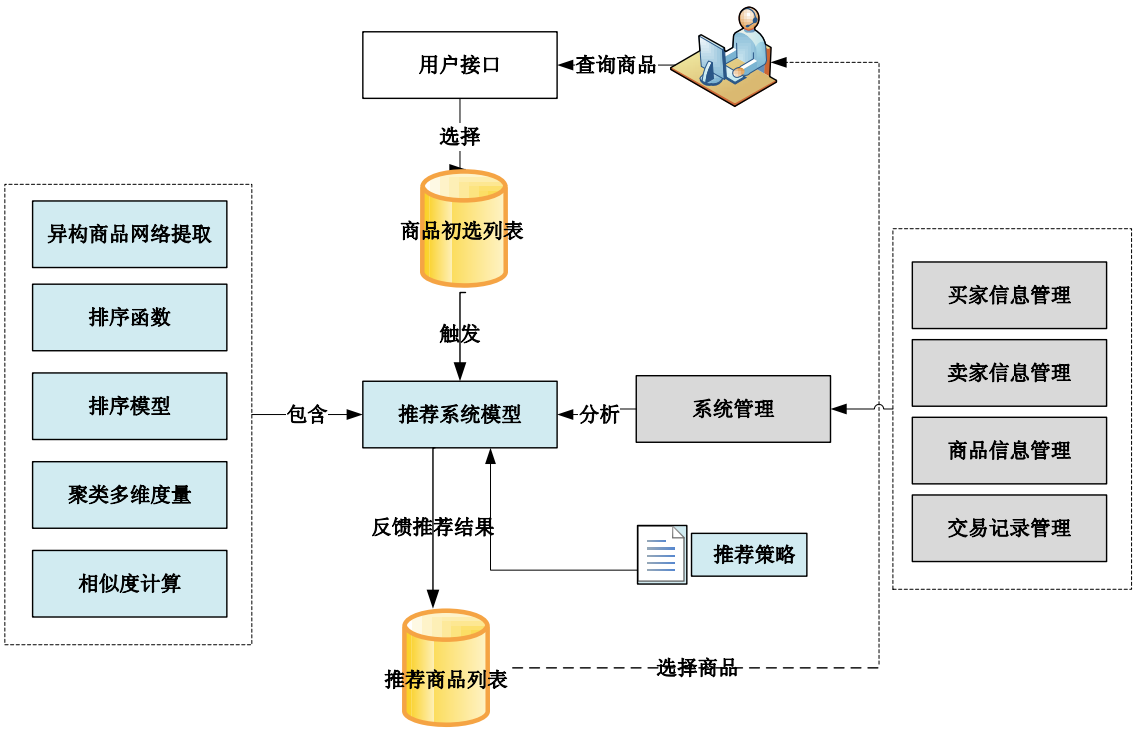

图 5.2 系统原型架构

Fig.5.2 The Architecture of Prototype System

如图 5.2 所示的系统原型架构中所示，系统的设计与实现将选取原型系统作为推荐系统的应用场景，用户查询商品请求进行对商品进行推荐，触发推荐系统进行商品初选列表的执行；结合系统管理模块中买家、卖家、商品、交易记录等进行异构商品网络的具体操作；基于异构商品网络进行商品的排序、聚类多维度量、相似性计算，得出商品聚类簇；根据推荐策略按照新用户和老用户进行分别推荐，形成商品推荐列表；由用户选择商品，则完成了整个商品推荐功能。

下面对系统架构中的各个功能模块做大致的说明：

① 用户接口：用户按照喜好查询商品，提供用户的输入接口，触发推荐系统；

② 系统管理模块：系统中买家、卖家、商品、交易记录等基本信息进行管理，为异构商品网络的提取与构建提供数据基础；

③ 商品初选列表模块：根据用户查询关键词结合系统管理模块中商品信息，进行与查询关键词匹配商品信息的初选，形成商品初选列表，缩小选择范围，提高推荐效率；

④ 推荐系统模型模块：是功能模块中最为重要的部分，包含了将商品初选列表提取为异构商品网络，进行商品排序、商品聚类等操作，形成商品聚类簇；

⑤ 推荐商品列表模块：结合商品推荐策略，将商品聚类簇中的结果按照新用

户和老用户的不同需求进行推荐，供用户选择。

5.3 数据库设计

在商品推荐系统中，商品（Item）、买家（Buyer）、卖家（Seller）和热点词（Term）是主要的参与对象，下面针对原型系统要实现的功能进行数据库设计，分别对如表 5.2 所示的主要关系表进行结构设计。

表 5.2 主要数据库关系表

Table 5.2 Table of the Database Relationship

序号	编码	说明
1.	Item	商品表
2.	Buyer	买家表
3.	Seller	卖家表
4.	Term	热点词表
5.	Item-Term	商品-热点词关联表
6.	Transaction	交易记录表

① 商品表设计

商品表（Item）包括商品的编号、名称、类别、品牌、所属卖家、生产厂商、价格、产地、包装、重量、商品描述等属性，下面具体对商品表的各个字段属性进行详细设计，如表 5.3 所示。

表 5.3 商品表设计

Table 5.3 Table of the Items

字段	说明	类型（长度）	键值	是否为空
ItemID	商品编号	Int(4)	PK	否
ItemName	商品名称	Varchar(20)		否
ItemClass	商品类别	Int(4)		否
ItemBrand	商品品牌	Varchar(20)		
ItemSeller	商品卖家	Int(4)	FK	否
ItemCreator	生产厂商	Varchar(20)		
ItemPrice	商品价格	Decimal		否
ItemAddress	商品产地	Varchar(50)		
ItemPackage	商品包装	Varchar(20)		
ItemWeight	商品重量	Varchar(20)		
ItemDescription	商品描述	Varchar(200)		

②　买家表设计

买家表（Buyer）包括买家用户的编号、用户名、密码、姓名、性别、年龄、身份证号码、电话、邮箱、地址、职业、状态等属性描述，下面具体对买家表的各个字段属性进行详细设计，如表 5.4 所示。

表 5.4 买家表设计

Table 5.4 Table of the Buyers

字段	说明	类型（长度）	键值	是否为空
BuyerID	买家编号	Int(4)	PK	否
BuyerUserName	买家用户名	Varchar(20)		否
BuyerPassword	买家密码	Varchar(20)		否
BuyerRealName	买家真名	Varchar(20)		
BuyerGender	买家性别	Int(4)		
BuyerAge	买家年龄	Int(4)		
BuyerIDCard	买家身份证号码	Varchar(20)		
BuyerPhone	买家电话	Varchar(20)		
BuyerEmail	买家邮箱	Varchar(50)		
BuyerAddress	买家地址	Varchar(100)		
BuyerOccupation	买家职业	Varchar(20)		
BuyerStatus	买家状态	Int(4)		否

③　卖家表设计

卖家表（Seller）包括卖家用户的编号、用户名、密码、姓名、性别、年龄、身份证号码、销售范围、电话、邮箱、地址等属性描述，下面具体对卖家表的各个字段属性进行详细设计，如表 5.5 所示。

表 5.5 卖家表设计

Table 5.5 Table of the Sellers

字段	说明	类型（长度）	键值	是否为空
SellerID	卖家编号	Int(4)	PK	否
SellerUserName	卖家用户名	Varchar(20)		否
SellerPassword	卖家密码	Varchar(20)		否
SellerRealName	卖家真名	Varchar(20)		否
SellerGender	卖家性别	Int(4)		否
SellerAge	卖家年龄	Int(4)		否

字段	说明	类型（长度）	键值	是否为空
SellerIDCard	卖家身份证号码	Varchar(20)		否
SellerAddress	卖家地址	Varchar(100)		否
SellerPhone	卖家电话	Varchar(20)		否
SellerEmail	卖家邮箱	Varchar(50)		
SellerTypes	卖家销售范围	Varchar(50)		否
SellerStatus	卖家状态	Int(4)		否

④ 热点词表设计

热点词表（Item）包括商品相关的热点词，例如“流行”、“青春”、“很好”、“差评”、“还行”等词语，一般反应的是商品在质量、特征、使用效果等的评价方面。为对热点词和商品之间的关系进行描述，下面热点词的特征进行提取，设计其关系属性，如表 5.6 所示。

表 5.6 热点词表设计

Table 5.6 Table of the Terms

字段	说明	类型（长度）	键值	是否为空
TermID	热点词编号	Int(4)	PK	否
TermValue	热点词描述值	Varchar(20)		否

另外，根据用户评价获取商品和热点词的映射，构造了“Item-Term”关联表，如表 5.7 所示。

表 5.7 热点词-商品关联表设计

Table 5.7 Table of the Terms and Items

字段	说明	类型（长度）	键值	是否为空
ID	编号	Int(4)	PK	否
ItemID	热点词 ID	Int(4)	FK	否
TermID	商品 ID	Int(4)	FK	否

⑤ 交易记录表设计

交易记录表（Transaction）对商品在电子交易平台发生交易过程中买卖双方和商品的信息进行记录，已备查询、核对、对账等需要，如表 5.8 所示是历史交易记录表的详细设计，主要包括交易的买卖双方的 ID 号、商品 ID 号、交易时间、交

易金额、交易状态等。

表 5.8 交易记录表设计

Table 5.8 Table of the Items Transaction

字段	说明	类型（长度）	键值	是否为空
TransactionID	交易编号	Int(4)	PK	否
ItemID	交易商品	Int(4)	FK	否
SellerID	交易卖家	Int(4)	FK	否
BuyerID	交易买家	Int(4)	FK	否
TransactionNumber	交易数量	Int(4)		否
TransactionPrice	交易价格	Decimal		否
TransactionTime	交易时间	DateTime		否
TransactionStatus	交易状态	Int(4)		否

⑥　实体关系图

买家、买家、关键词因为商品买卖关系而与商品建立的了关联，产生了关系，下面基于关系型数据库的设计思想，使用实体关系图对主要的表格进行设计，如图 5.3 所示。

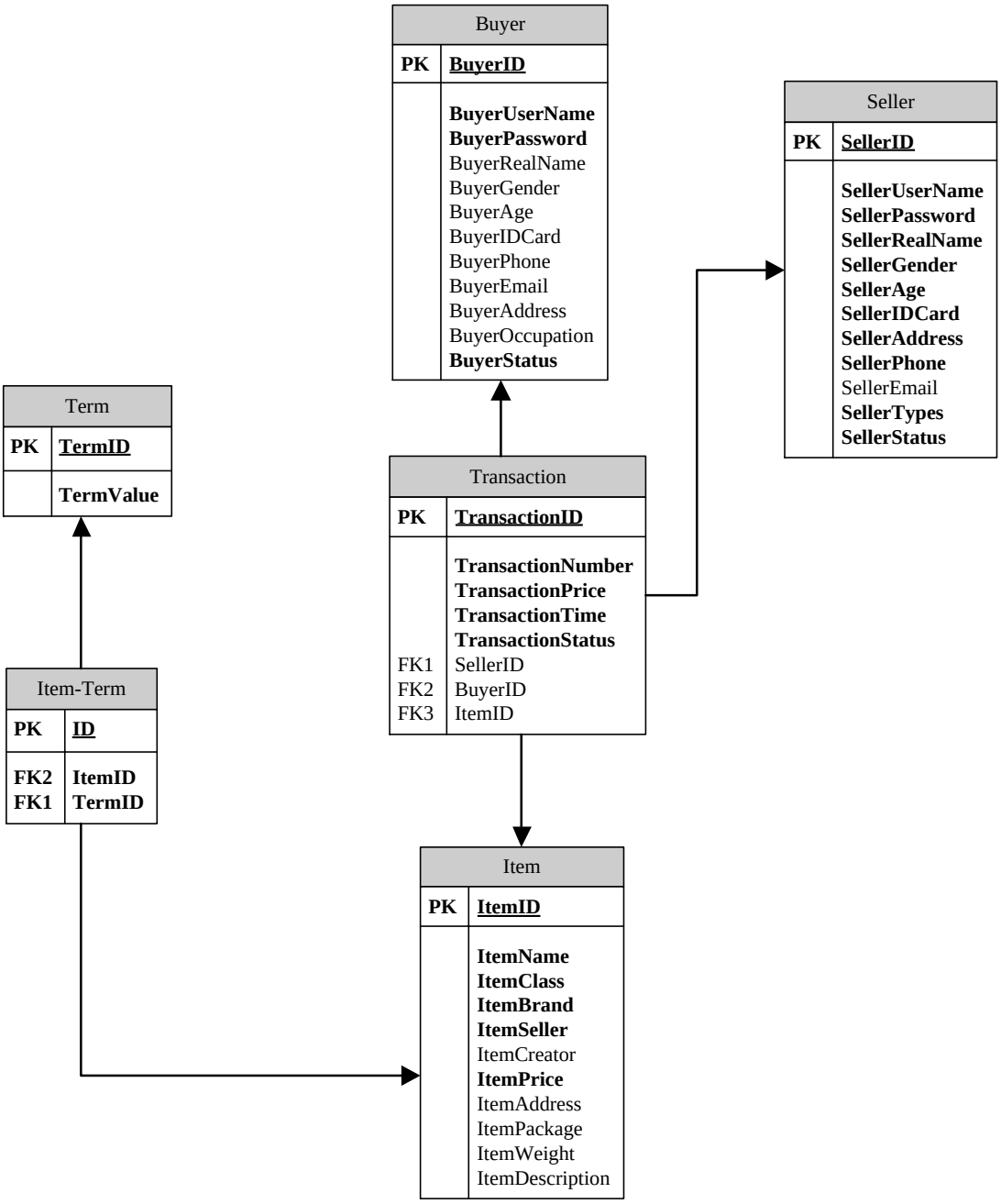

图 5.3 数据库实体关系图

Fig.5.3 Entity Relationship Diagram of Database

5.4 系统实现与效果分析

在分析了系统的需求和功能设计后，结合系统数据库设计进行系统的实现相关工作。基于异构网络分析的商品推荐原型系统是分析买家在购物过程中的交易记录之间的参与方的关联关系，所以需要针对买家的相关信息进行管理，因此需要设计并实现登录注册界面，实现对买家的信息管理功能，下图 5.4 是买家的登录界面。

图 5.4 用户登录

Fig.5.4 Login Page of the User

5.4.1 推荐系统首页

在买家登录后，可以进入推荐商品系统首页。在本文中，基于异构商品网络的商品推荐原型系统用其英文字母首字母组成该系统的图标 HINRS（Heterogeneous Items Network Recommendation System），如图 5.5 所示。

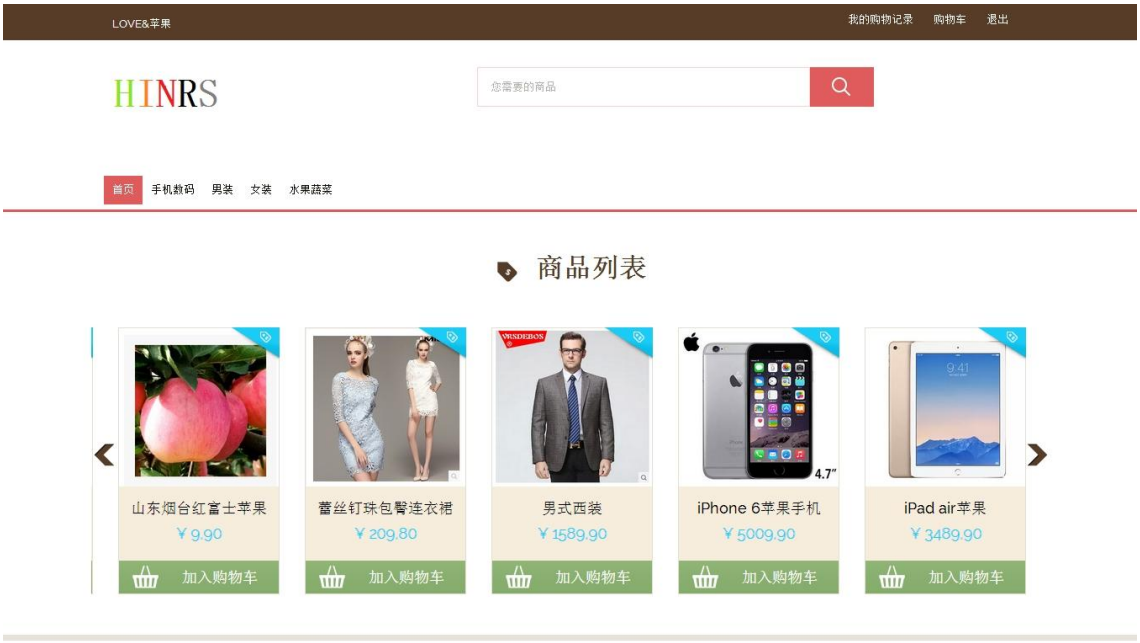

图 5.5 推荐系统首页

Fig.5.5 Front Page of the System

如图 5.5 所示的推荐系统首页，在图左上角有用户基本信息（Love 苹果）；右上角则是该用户的购物记录和购物车，如果不需要登录状态则点击“退出”按钮退出登录；界面正中则是该推荐系统的商品列表，包括“首页”、“手机数码”、“男

装”、“女装”、“水果蔬菜”等常用的品类可供用户浏览及购买。

5.4.2 用户推荐结果页面

当用户在搜索输入框中输入想查询的商品关键词后，HINRS 推荐系统则根据异构商品网络分析来进行商品的相关信息处理，按照该用户是新用户还是老用户按照不同的推荐策略来进行商品推荐。

①　新用户

该用户 ID 为 NewUser，代表其为新用户，则其购物记录为空。这时根据推荐策略将其搜索关键词相关的商品聚类推荐给该新用户 NewUser，如下图 5.6 所示。

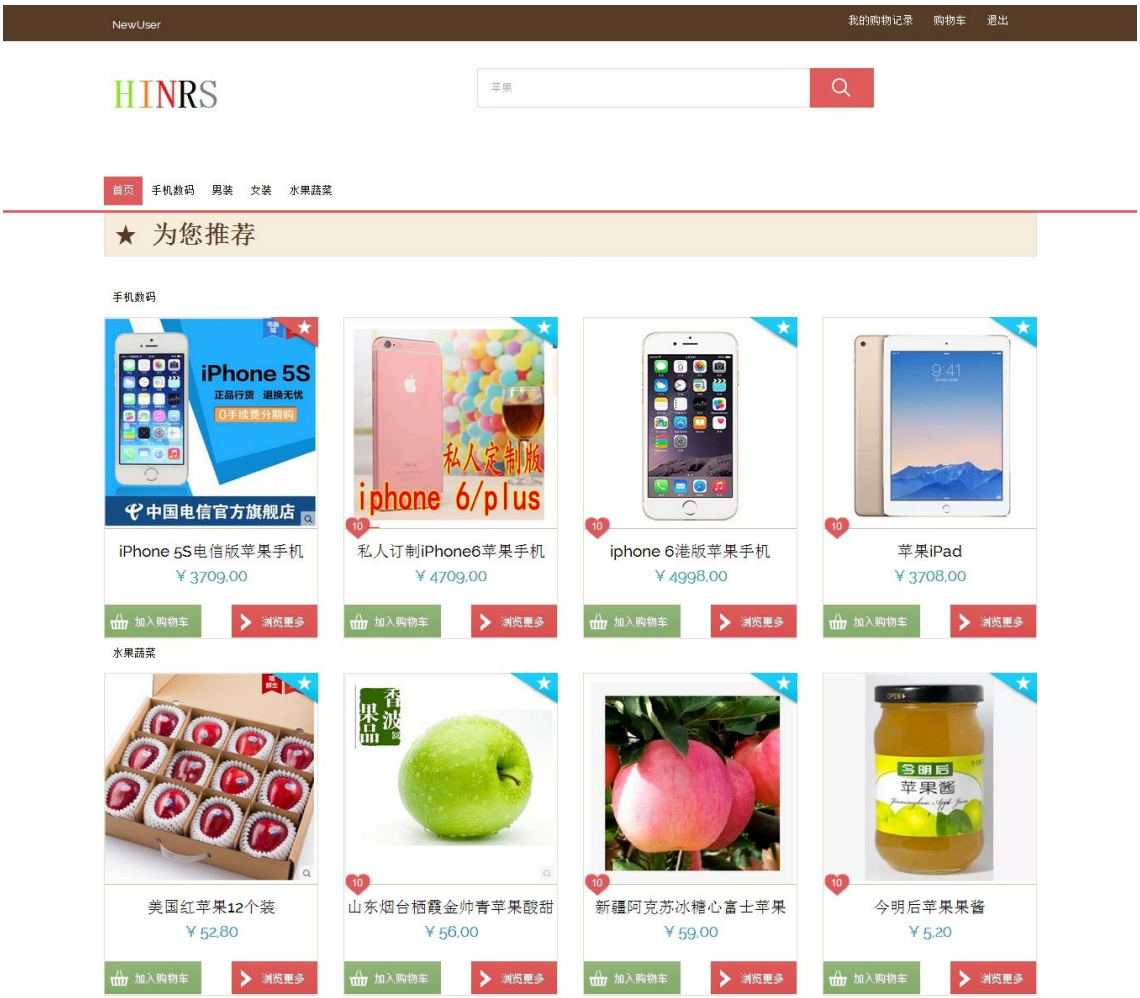

图 5.6 新用户推荐结果

Fig.5.6 Recommedation Results for the New User

在图 5.6 所示的新用户推荐结果页面中，该用户 NewUser 输入查询关键词“苹果”，HINRS 推荐系统根据查询关键词进行聚类分析，将与“苹果”相关的商品取

排名较前并按类别展示给该用户，结果分别是“手机数码”中的 iPhone 5S 和 iPhone 6 及 iPad 等苹果公司相关的数码产品、“水果蔬菜”类别中的“山东烟台苹果”和“新疆红富士苹果”及“苹果果酱”等苹果水果相关的食品。可供用户浏览并选择合适的商品，从而实现了针对新用户的异构商品网络推荐，解决了类别混乱这一问题。

②　老用户

在推荐系统中，如果是老用户，即购物记录不为空时则根据推荐策略进行针对老用户的商品推荐。该用户 ID 为“Love 苹果”，其在搜索框中输入查询关键词“苹果”，HINRS 系统根据推荐策略进行相关推荐界面如 5.7 所示。

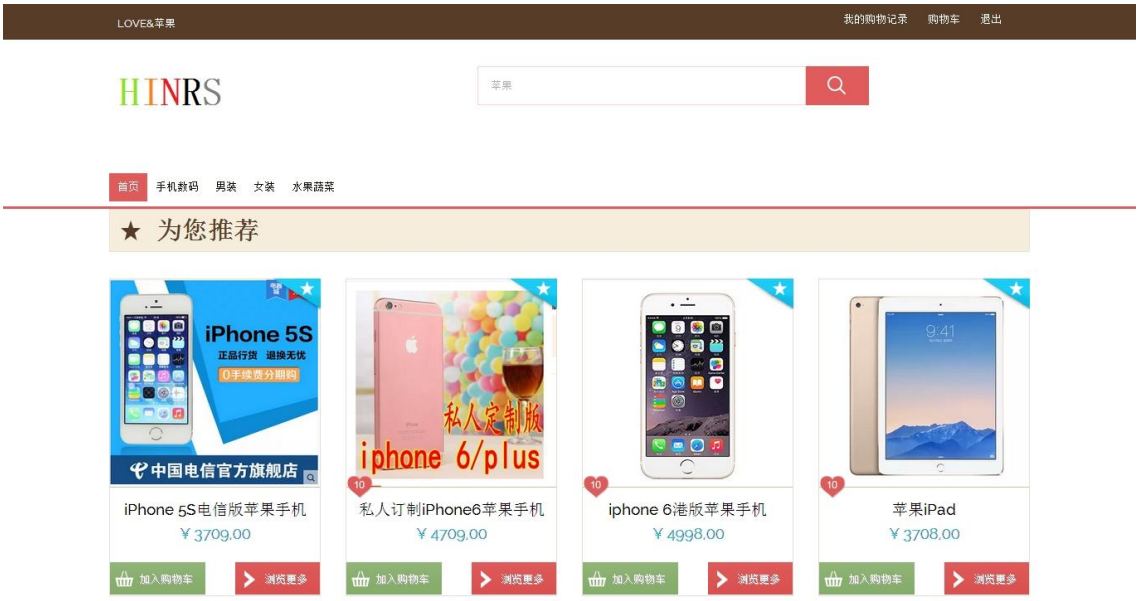

图 5.7 老用户推荐结果

Fig.5.7 Recommedation Results for the Old User

如图 5.7 所示的老用户推荐结果页面中，用户 Love 苹果输入查询关键词“苹果”后触发推荐系统，结合异构商品网络分析其购物记录并结合聚类排序等操作，将推荐结果按照类别排名先后展示给用户 Love 苹果，推荐结果分析其查询关键词“苹果”应该为手机数码类别之中的苹果公司数码产品类别，因此将苹果公司相关数码产品按照排名先后推荐给用户 Love 苹果，供其选择。

5.5　本章小结

本章主要分析了基于异构网络分析的商品推荐原型系统的设计与实现，从软

件工程角度分析并结合具体的操作开发工具，Powerdesigner 进行数据库设计、MyEclipse 和 Dreamviewer 进行系统实现。在实现该系统功能后，进行效果分析，从而验证了本课题的可行性与创新性。

6 结论与展望

6.1 结论

本文针对异构信息网络分析的商品推荐展开研究，在异构商品网络描述模型、异构商品网络排序、聚类分析、商品推荐等方面展开研究。重点针对异构商品网络的构造、聚类与排序、结合推荐策略进行异构商品网络的商品推荐进行研究。主要完成了以下研究工作：

① 结合信息网络异构化发展趋势，基于网络分析、异构网络分析与商品推荐等国内外研究现状及问题分析提出了本文的研究内容与创新点。

② 借助形式化方法研究了异构商品网络描述模型，结合电子商务交易数据特征分析、构造与维护异构商品网络。

③ 基于异构商品网络描述模型，提出了基于异构商品网络分析的商品聚类算法，结合商品网络排序函数与排序模型进行异构商品网络中商品交易各参与对象的聚类分析，实现异构商品网络对象的类别挖掘和类别内的重要性排序。

④ 从商品交易记录中各种对象之间的关系维度出发，针对异构商品网络聚类结果，结合相应的推荐策略提出一种新型的商品推荐模型，分别从推荐思想、推荐流程、算法描述等方面进行详细研究内容的阐述。

⑤ 基于异构商品网络聚类的推荐研究，从软件工程的思想出发，从需求分析、系统功能设计、数据库设计和系统实现等方面设计并实现了商品推荐原型系统，从而验证了所提出的推荐模型的可行性。

6.2 展望

当前的研究主要是针对电子商务交易记录，结合异构商品网络的相关知识进行商品推荐，着重考虑买家与商品之间的关联关系，即从用户角度出发，寻找与用户兴趣度最为匹配的商品。然而基于异构商品网络分析的商品推荐方法，基于异构信息网络模型对商品推荐分析，按照他们之间的类别关联关系可以针对聚类结果中的买家、卖家、商品都可以进行聚类排序，比如买家更倾向于买哪一类别的商品、卖家销售哪一类别的商品。都考虑到了电子商务交易平台上的三个参与方，因此可以具体分析商品推荐过程中买家和卖家各自的需求并进行研究实验来验证。因为篇幅问题，卖家需求问题还没有详细考虑，希望在以后的研究生涯中可以继续针对卖家的倾向性结合异构商品网络的相关研究工作进行延续并有好的成果。

致　　谢

时光飞逝，转眼间三年硕士学业已经进入倒计时，在过去的几年里自己经历了人生的一些挫折与磨砺，特别是研三找工作的时期，让自己更加成熟，为进入社会奠定了基础，相信以后的自己也会更加从容应对各种工作上的困难与机遇。

在这三年的学习中，我最应该感谢的是我的学业及人生导师熊庆宇教授，熊老师以人之父的心态与言行，将自己的经验与知识无私传授给我，让我学习到了课堂上不能接触到的内容；还会支持我的一些抉择，为我的人生画下了添彩的一面；学业上则不忘时刻督促我，并以高标准的要求监督我执行，让我可以意识到自己的不足并能够及时弥补。

此外，还要感谢文俊浩、柳玲、蔡海尼、高旻、曾骏老师的谆谆教诲，在我的学业和工作上都给予了意见与关心，衷心的表达自己对他们的敬意与谢意。还有实验室的崔丽艳、覃梦秋、何波、李飞、陈霞等同学，三年的时间欢笑与痛苦都一起分享过，共同经历了人生这么美好的时期，是我们值得铭记一生的时光。江卓、周魏、王喜宾博士师兄们也对我照顾有加，经常去叨扰还是希望师兄们不要嫌弃。还有一起参与课题研究的唐少雄同学对我整个研究项目的协助。

还有要感谢的就是我的师兄李朋，跟随他学习了五年时间，其间对我的人生指导与呵护，让我的学业生活及未来更加清晰。希望在以后的岁月中，可以执子之手共看人生旅途中的美好景色。

最后，要特别感谢我亲爱的家人，感谢你们对我的养育之恩及呵护之情。严父慈母让我体会到了学业与生活的不同感受，正是你们让我以坚强的心态度过了找工作时那段晦涩艰难的时刻，并给予了我未来生活的信心与力量；还要监督妹妹，好好学习，必上山大，实现姐姐当年的夙愿。

董俊平

二 O 一五年五月 于重庆大学

参考文献

[1] 中国互联网络信息中心,中国互联网络发展状况统计报告[EB/OL]. http://www.cnnic.net.cn/ research/bgxz/tjbg/,2014.1.

[2] Sun, Y., Han, J., Zhao, P., Yin, Z., Cheng, H., and Wu, T. RankClus: integrating clustering with ranking for heterogeneous information network analysis [C]. In Proceedings of EDBT. 2009, 565-576.

[3] Zhang M, Hu H, He Z, et al. Top-k similarity search in heterogeneous information networks with x-star network schema [J]. Expert Systems with Applications, 2015, 42(2): 699-712.

[4] Meng C, Cheng R, Maniu S, et al. Discovering Meta-Paths in Large Heterogeneous Information Networks [J]. 2015.

[5] Beath, C., I. Becerra-Fernandez, et al.. Finding Value in the Information Explosion [J]. Mit Sloan Management Review.2012, 53(4): 18-20.

[6] 刘松钦. 电子商务发展现状及对策研究[J]. 电子商务, 2014 (2): 11-12.

[7] Chung F, Simpson O. Computing Heat Kernel Pagerank and a Local Clustering Algorithm [J]. 2015.

[8] Beebe N L, Liu L. Ranking algorithms for digital forensic string search hits[J]. Digital Investigation, 2014, 11: S124-S132.

[9] Clemons E K, Kauffman R J, Weber T A. Introduction to the Special Section: Economics of Electronic Commerce[J]. International Journal of Electronic Commerce, 2014, 15(1):75-78.

[10] Kumar M, Das D. Scaling Number of Active Links in a Linux Kernel Bond Driver Having Heterogeneous Network Interfaces[J]. Wireless Personal Communications, 2014,76(3): 435-447.

[11] Tzortzis G, Likas A. The minmax k-means clustering algorithm [J]. Pattern Recognition, 2014, 47(7): 2505-2516.

[12] De Prato G, Nepelski D. Global technological collaboration network: network analysis of international co-inventions [J]. The Journal of Technology Transfer, 2014, 39(3): 358-375.

[13] L.A.F. Park and K. Ramamohanarao, Multiresolution Web Link Analysis Using Generalized Link Relations [J]. In Proceedings of IEEE Trans. Knowledge Data Eng. 2011, 1691-1703.

[14] Y. Sun, Y. Yu, and J. Han, Ranking-based clustering of heterogeneous information networks with star network schema [C]. In Proceedings of KDD. 2009, 797-806.

[15] 王小君. 基于内容和网络结构图的个性化推荐算法研究[D]. 华南师范大学, 2010.

[16] X. Li, L. Chen. Recommendations based on network analysis [C]. In Proceedings of 2011

International Conference on Advanced Computer Science and Information Systems, 2011, 9-15.

[17] Z. Zheng, H. Ma, M.R. Lyu, et al. QoS-Aware Web Service Recommendation by Collaborative Filtering [J]. IEEE Transactions on Services Computing, 2011, 4(2):140-152.

[18] Zhou Y, Liu L. Activity-edge centric multi-label classification for mining heterogeneous information networks[C]//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014: 1276-1285.

[19] 国家科学基金会.互联网发展报告 [EB/OL].http://www.nsf.gov/.2013.

[20] 姜峰,范玉顺.基于扩展概念格的 Web 关系挖掘 [J].软件学报,2010,21(10):2432-2444.

[21] 王功聪. 基于内容的网络行为分析[D]. 北方工业大学, 2014.

[22] Ji, H., H. B. Deng, et al. Uncertainty Reduction for Knowledge Discovery and Information Extraction on the World Wide Web [C]. Proceedings of the Ieee.2012, 100(9): 2658-2674.

[23] Usman, M., R. Pears, et al.. A data mining approach to knowledge discovery from multidimensional cube structures [J]. Knowledge-Based Systems .2013, 40: 36-49.

[24] Cavuoti, S., M. Brescia, et al. Data mining and Knowledge Discovery Resources for Astronomy in the Web 2.0 Age [J]. Software and Cyberinfrastructure for Astronomy Ii.2012, 8451.

[25] Boden B, Ester M, Seidl T. Density-Based Subspace Clustering in Heterogeneous Networks [M], Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2014: 149-164.

[26] 池云.异构信息网络的分类研究[J].计算机应用与软件,2014,31(6):330-333.

[27] G. Qi, M. Tsai, S. Tsai, L. Cao, and T.S. Huang, Web-Scale Multimedia Information Networks [J]. In Proceedings of Proceedings of the IEEE. 2012, 2688-2704.

[28] Niu L, Han X, Xu Y. On Charactering of Word-of-Mouth Propagation in Heterogeneous Online Social Networks [J]. Journal of Multimedia, 2014, 9(5): 668-675.

[29] D. Taniar, High Performance Database Processing [J]. In Proceedings of AINA. 2012, 5-6.

[30] 段明秀. 层次聚类算法的研究及应用[D]. 中南大学, 2009.

[31] 陈克寒，韩盼盼，吴健等.基于用户聚类的异构社交网络推荐算法[J].计算机学报,2013,36(2):349-359.

[32] 闵敏. 基于聚类协作过滤的商品个性化推荐系统的实现[J]. 制造业自动化, 2010,1009-0134.

[33] S. Colucci, T.D. Noia, A. Ragone, M. Ruta, U. Straccia, and E. Tinelli, Informative top-k Retrieval for Advanced Skill Management [J]. In Proceedings of Semantic Web Information Management. 2009, 449-476.

[34] Kou Y, Shen D, Xu H, et al. Two-level interactive identification and derivation of topic clusters in the complex networks [J]. World Wide Web, 2014: 1-30.

[35] 崔春生, 吴祈宗, 王莹. 用于推荐系统聚类分析的用户兴趣度研究[J]. 计算机工程与应用, 2011, 47(7):226-228.

[36] 王文东. 模糊文本聚类算法的研究与应用[D]. 西安电子科技大学, 2012.

[37] Y. Ledeneva et al. EM clustering algorithm for automatic text summarization [J].Advances in Artificial Intelligence. Springer Berlin Heidelberg.2011,305-315.

[38] L. Vu, M. Hauswirth, and K. Aberer, QoS-Based Service Selection and Ranking with Trust and Reputation Management [C], In Proc. OTM Conferences (1), 2005, 466-483.

[39] T Y. Accuracy and reproducibility of co-registration techniques based on mutual information and normalized mutual information for MRI and SPECT brain images.[J]. Annals of Nuclear Medicine, 2004, 18(8):659-667.

[40] 王纵虎. 聚类分析优化关键技术研究[D]. 西安电子科技大学, 2012.

[41] 黄鹏. 基于互联网用户特征的商品推荐系统研究[D]. 东华大学, 2014.

[42] F. Farnoud, O. Milenkovic, and B. Touri, A Novel Distance-Based Approach to Constrained Rank Aggregation [C]. In Proceedings of CoRR. 2012.

[43] Y. Wang, Y. Huang, X. Pang, M. Lu, M. Xie, and J. Liu, Supervised rank aggregation based on query similarity for document retrieval [J]. In Proceedings of Soft Comput.. 2013, 421-429.

[44] J. Weston, R. Kuang, C.S. Leslie, and W.S. Noble, Protein Ranking by Semi-Supervised Network Propagation [J]. In Proceedings of BMC Bioinformatics. 2006.

[45] 孙吉贵,刘杰,赵连宇,聚类算法研究 [J].软件学报,2008,19(1):48-61.

[46] P. Li, J.H. Wen, X. Li, SNTClus-A novel service clustering algorithm based on network analysis and service tags [J].Przeglad Elektrotechniczny (Electrical Review), 89(1), 2013.

[47] D. Devlin and B. O'Sullivan, Preferential Attachment in Constraint Networks [C]. In Proceedings of ICTAI. 2009, 708-715.

[48] P. Mahendra, M. Prokopenko, and A.Y. Zomaya, Assortative Mixing in Directed Biological Networks [J]. In Proceedings of IEEE/ACM Trans. Comput. Biology Bioinform.. 2012, 66-78.

[49] 许海玲,吴潇,李晓东,阎保平.互联网推荐系统比较研究 [J],软件学报,2009,20(2):350-362.

[50] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1):48-61.

[51] 池云.异构信息网络的分类研究[J].计算机应用与软件,2014,31(6):330-333.

[52] F. Ricci, L. Rokach, B. Shapira. Introduction to Recommender Systems Handbook [M]. Springer US, 2011, 1-35.

[53] L. Kuang, Y. Xia, Y. Mao. Personalized Services Recommendation Based on Context-Aware QoS Prediction [C]. In Proceedings of ICWS. 2012, 400-406.

[54] 慕春棣, 戴剑彬, 叶俊, 用于数据挖掘的贝叶斯网络[J]. 软件学报 2000,11(5): 660-666.

[55] 时睿. 基于数据挖据的商品推荐系统研究和实现[D]. 上海交通大学, 2013.

[56] Zhang L F, Yang S W, Zhang M W. E-commerce Website Recommender System Based on Dissimilarity and Association Rule[J]. Indonesian Journal of Electrical Engineering, 2014, 12(1): 353-360.

[57] Sun X. Personalized Information Recommendation Based on Web Clustering[C], Proceedings of the 9th International Symposium on Linear Drives for Industry Applications, Volume 2. Springer Berlin Heidelberg, 2014: 511-519.

[58] Li J, Wang X, Sun K, et al. Recommendation Algorithm with Support Vector Regression Based on User Characteristics[C], Proceedings of the 9th International Symposium on Linear Drives for Industry Applications, Volume 3. 2014: 455-462.

附　　录

A. 作者在攻读硕士学位期间发表的学术论文目录

[1] XIONG Q, **DONG J**, WEN J, et al. A Novel Web Service Top-K Query Algorithm Based on Heterogeneous Service Network Analysis[J]. Journal of Computational Information Systems, 2015, 11(1): 387-398. (外文 EI，检索号 20150900580290)

[2] **Junping Dong**, Qingyu Xiong, Junhao Wen, Peng Li. Services Recommendation System based on Heterogeneous Network Analysis in Cloud Computing.Research Journal of Applied Sciences, Engineering and Technology, 7(14): 2858-2862, 2014.（EI 刊源已发表）

B. 作者在攻读硕士学位期间参与的科研项目

[1] 国家自然科学基金面向面上项目：基于异构服务网络分析的 Web 服务推荐研究（项目编号：61379158）；

[2] 教育部博士点基金博导类项目．基于异构网络分析的云服务推荐研究[20120191110028] (2012-2014).

C. 作者在攻读硕士学位期间获奖情况

[1] 研究生国家奖学金　　2014

[2] 重庆大学优秀研究生干部　　2014

[3] 重庆大学研究生 A 等奖学金两次　　2012,2013

[4] 重庆大学优秀研究生　　2013