学校代码： 10246
学　　号： 10212010024

硕 士 学 位 论 文

基于相似度计算的药品信息可视化地图的研究与实现

院　　系： 软件学院
专　　业： 计算机软件与理论
姓　　名： 施悦
指 导 教 师： 刘钢
完 成 日 期： 2013 年 月 日

指导小组成员名单

刘钢

徐迎晓

目 录

摘 要

目前医药电子商务行业正迅速发展，但药师等专业人才的供给无法满足该行业的增长速度。对于医药电子商务行业的工作人员，在没有专业医药知识背景的情况下，如何从庞大的药品数据库以及越发快速的药品数据更新中，快速有效的寻找所需要的药品信息，并对顾客实现合理的药品推荐，是我们要解决的问题。

本文针对这个问题，对基于相似度计算的药品信息可视化地图进行了分析研究，通过将庞杂的药品数据库转化为可以直接向客户进行推荐的药品知识库地图，为用户清晰地展示药品之间的关系脉络，使非药师工作人员也能快速有效地根据顾客的需求提供药品的专业化指导和推荐，既提高了经营效率和专业服务水平，又降低了人力资源成本。

本文首先调研了国内外可视化的研究现状，以现实生活中的药品说明书为依据，提出了信息可视化地图的概念，同时，在总结和介绍现有相似度算法的基础上，提出了针对药品的相似度计算模型，并通过实验验证了本文提出的药品相似度算法比传统的相似度算法更加合理，最后以真实医药电子商务网站为背景，设计和实现了基于相似度计算的药品信息可视化地图系统的整体架构和功能模块。

本文所提出的基于相似度计算的药品信息可视化系统能够自动分析药品之间的相似度，将药品之间的关系更加清晰的反馈给用户，从而能使用户更好地了解药品信息，快速有效地找到合适的药品。

关键词

医药电子商务；信息可视化；中文分词；相似度计算；

中图分类号

TP37

ABSTRACT

Nowadays, the medical e-commerce industry is developing so rapidly that the growing pharmacists can't keep up with the industry growth rate. Also, as the huge medicine database is now updating frequently day by day, it seems hard for a medical e-commerce company with lack of professional staffs to provide better medicine guidance and recommendation for customers.

To solve this problem, this paper proposes a medicine information visualization map based on medicine similarity analysis, which can transform the complex medicine database into medicine knowledge base for better understanding the medicine information. This medicine information visualization map can be used to enable the non-professional staff to provide professional medicine guidance and recommendation, thereby not only improve the operation efficiency and professional service level, but also reduce the human resource cost.

This paper first makes a research about the current development of information visualization and some classic tools. Then, the concept of information visualization map is constructed based on the structure of medicine specification. And after summarizing the existing similarity analysis algorithms, a medicine-specific similarity analysis calculation model is proposed, which is proved by experiments to be more reasonable and applicable in medical domain. Finally, a system is designed and implemented based on a real medical e-commerce website to realize the medicine information visualization map, which can analyze the similarity between medicines automatically, then provide more clear relationships between medicines, thereby enable users better understand the medicine knowledge and find the appropriate medicines for customers more efficiently.

Keywords

Medical E-Commerce Information Visualization Map Chinese Segmentation Similarity Analysis

C1assification Code

TP37

第一章 绪论

1.1 论文的研究背景

医药电子商务，是指利用信息技术进行医药相关信息的传播，促进和达成医药类商品的交易[错误!未找到引用源。]。截止至 2012 年底，中国网民总数已超过 5 亿，中国电子商务交易额已达 8.1 万亿元。但是，在医药电子商务领域，截止至 2011 年底，中国已获得《互联网药品交易服务资格证》的企业不超过 100 家。显然，这与电子商务巨大的市场发展前景形成了鲜明对比，因此，医药电子商务的市场拥有巨大的发展前景。

传统渠道的药品流通，通常是从生产企业，到批发企业，到医疗机构(零售药店)，再到患者[错误!未找到引用源。]。在整个过程中，医疗机构和零售药店是最具有发言权的中间环节。医疗机构的医师与零售药店的药师，都是具有相当专业背景的医药类人才，患者在进行购药过程中，有很大一部分的购买决策权都是根据这些医师和药师的推荐进行确定的。

当整个药品的销售环节从实体店转移到了互联网之后，医药电子商务企业应该创新性地建立起符合中国病人求诊问药习惯的经营体系和流程规范，确保患者在享受电子商务优势的同时，也能同样获得原先所拥有的一切临床服务所带来的价值。即使是在网上经营 OTC 和保健类等无需医生处方的医药类产品，消费者也完全有权利在网络消费中获得所有能在实体药店得到的服务，其中最主要的就是向药师的咨询以及对产品的推荐等等。

然而，据有关统计，我国需要 100 万执业药师，但截止到 2012 年 2 月底全国累计只有 20 万人取得执业药师资格，执业药师供需之间仍然存在巨大缺口。事实上，药店配备执业药师，需要的是其丰富的医药专业知识，用以指导病患者购药和用药。医药电子商务在本质上和图书或者家电的电子商务完全不同，医药类产品的销售不仅仅是产品本身的买卖和产品的及时送达，更包括了相关的医药知识和健康理念的传达。经营医药产品的本质在于“为病人带来健康，为病人塑造更高的生活质量”，因此，医药电子商务的经营者必须将这样的价值观和使命感融合到日常经营的每一个环节之中。

本文提出的药品信息可视化地图系统所要解决的问题就是如何将庞杂的药品数据库转化为可以直接向客户进行推荐的药品知识库地图，使非药师工作人员实现药品的专业化指导和推荐。

通过这个系统，对医药电子商务行业的工作人员而言，即使是在没有专业的

医药知识的情况下，面对当前医药电子商务网站后台如此庞大并且日益更新速度相当快的药品数据库，仍然能够在客户提供药品名称或者症状描述之后，快速有效地为其提供合适的可选药品集合推荐，既提高了经营效率和专业服务水平，又降低了人力资源成本。

1.2 论文的研究内容

本文的用户群体主要是医药行业的从业人员。目前，国内医药电子商务平台的医药类产品(OTC 药品、保健药品等，以下简称药品)种类可达上万种，面对药品庞大的数据信息及快速的更新速度，即使是拥有医学背景的人员也无法及时了解药品知识。本文主要通过构建药品信息地图，让用户对药品知识有更好的了解，从而对顾客(往往是患者)进行药品的合理推荐。用户不需要对所有药品都有了解，药品信息地图可以自动分析药品之间的关系，将庞大的搜索结果转化为结构化的药品信息地图，为用户清晰地展示药品之间的关系脉络。即使是在药品更新如此迅速的当今，用户也能高效地从庞大的药品数据库中寻找合适的药品。

本文主要有以下几方面的研究工作：

1、介绍了信息可视化理论的主要内容和典型的可视化工具；研究了目前主要的文本相似度算法。

2、在分析了目前主要的相似度算法的基础上，以现实生活中的药品说明书为依据，建立了药品相似度的计算模型；

3、在前面理论分析的基础上，详细阐述了药品相似度分析的过程步骤，并对其进行实现和效果验证。

4、以真实医药电子商务网站为背景，设计并实现了基于相似度计算的药品信息可视化地图的整体架构和功能模块。

1.3 论文的章节安排

本文就药品信息可视化地图进行探讨，通过对信息可视化和文本相似度进行研究，提出针对药品的相似度计算方法，并进行药品信息地图的可视化实现。本文对系统的关键功能模块进行了设计和实现。

本论文的章节安排如下：

第一章是本文的研究背景，研究内容和意义，主要介绍了医药电子商务行业的发展背景和现状，该行业目前面临的主要问题，以及本系统将要解决的问题。

第二章介绍了相关理论和方法。首先介绍了信息可视化的概念以及具有代表性的信息可视化工具。其次，研究并介绍了目前几类主要的文本相似度算法。

第三章主要设计了药品相似度计算模型。首先，通过分析药品说明书，提出了药品信息地图的概念。其次，根据药品信息地图的定义，提出了药品数据模型，并在数据库中设计相应的药品数据表。最后，在此基础上，建立了药品相似度计算模型，并给出了具体的计算公式。

第四章具体实现并验证了药品相似度算法。首先，分析并介绍了药品相似度分析的整个过程步骤。其次，通过使用中文分词算法对药品说明书进行处理。最后，实现了药品相似度算法，并通过实验进行效果分析和评估。

第五章以真实医药电子商务网站为背景，设计并实现了基于相似度计算的药品信息可视化地图系统。首先，对需求进行了简要分析。其次，从架构设计、功能设计以及交互设计三方面介绍了系统的框架设计。最后，对核心功能模块进行了详细的分析设计和实现。

第六章是对全文的总结与展望。主要分析了本论文提出的药品信息可视化地图的优点以及不足之处，并在此基础上提出了今后在该领域的研究工作的主要方向。

第二章 相关理论和方法介绍

2.1 信息可视化研究

“可视化”(Visualization)其实质是利用计算机的图形图像处理技术，把各种数据信息转换成合适的图形图像在屏幕上展示出来。这一过程涉及到图形学、几何学、辅助设计和人机交互等领域知识[错误!未找到引用源。]。

在 1986 年 10 月，在美国国家科学基金会举办的“图形、图像处理和工作站”讨论会上，第一次正式提出了“科学计算可视化”的概念。1987 年，由 Mccormick 等人[4]所编写的美国国家科学基金会报告《Visualization in Scientific Computing》，对可视化技术领域产生了大幅度的促进和刺激。

随着可视化技术的发展，逐渐形成了一些分类，通常情况下，人们习惯于将可视化分为以下四类：科学计算可视化、数据可视化、信息可视化和知识可视化[3]。本文主要研究的信息可视化理论。

2.1.1 信息可视化

信息可视化(Information Visualization)主要是指利用计算机支撑的、交互的对非空间的、非数值型的和高维信息的可视化表示，以增强使用者对其背后抽象信息的认知[5]。

从定义中不难发现，信息可视化是处于在数据、计算机和用户之间的一类交叉的活动，其核心是数据探索。数据探索主要是为了达到三个目的：1、形象地表达信息，即将抽象数据用可视的方式来表示；2、发现新的知识：3、识别信息在结构、模式、趋势、关系等方面的各种可能的规律。通过探索，人类可以从浩瀚的数据海洋中获取有用的信息，从而在信息中发现知识，从知识中寻求决策。信息可视化帮助人们解决了如何与信息资源之间进行对话的难题。信息可视化可以帮助人们提高和增强认知能力，如以图形的方式表达多维数据，加深用户对数据内在含义的理解。并且还能以直观形象的图像来引导整个检索过程，使检索变得透明，加快了信息检索的速度。

实现可视化信息的检索就是利用可视化技术设法为用户提供一个可视化的环境以支持用户完成信息检索、浏览、挖掘等超出传统的信息系统所能实现的功能[6]。20 世纪 90 年代以来，随着可视化技术的进步，可视化信息检索系统得到了长足的发展。

信息可视化的框架技术还可以分为三种：映射技术、显示技术和交互控制技术。映射技术主要是降维技术，如因素分析、自组织特征图、寻径网(Pathfinder)、潜在语义分析和多维测量等。显示技术把经过映射的数据信息以图形的形式显示出来，主要技术有：Focus+Context、Tree-map、Cone Tree 和 Hyperbolic Tree 等。交互控制技术通过改变视图的各种参数，以适当的空间排列方式和图形界面展示合理的需求数据，从而达到将尽可能多的信息以可理解的方式传递给使用者[7]。

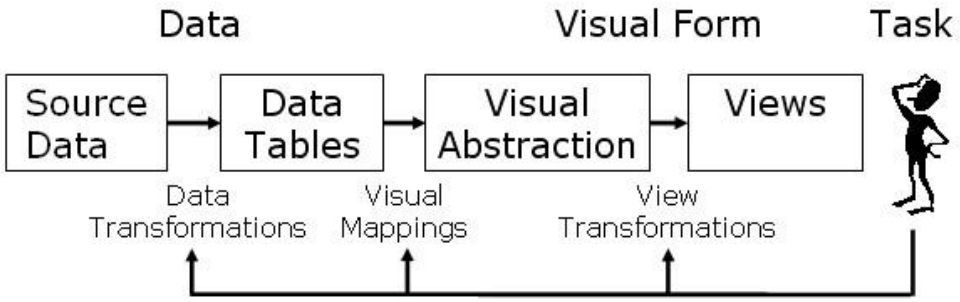

图 1 信息可视化参考模型

信息可视化是从数据到可视化形式再到人的感知系统的可调节的映射过程[8]在信息可视化参考模型[9]中，从原始数据到用户，中间要经历一系列数据变化。图中从左到右的每个箭头表示的都可能是一连串的变换。从用户到每个变换(从右到左)的箭头，表示用户操作的控制对这些变换的调整。数据变换把原始数据映射为数据表(数据的相关性描述)：可视化映射把数据表转换为可视化结构：视图变换通过定义位置、缩放比例等图形参数创建可视化结构的视图；用户的交互动作用来控制这些变换的参数，例如把视图约束到特定的数据范围，或者改变变换的属性等。可视化和它们的控制最终服务于任务。信息可视化要解决的主要问题就是如何实现上述参考模型中的映射、变换和交互控制[10]。

2.1.3 可视化工具

1、信息可视化工具包：Prefuse[11]

Prefuse 是一个可扩展的软件框架,它可以帮助使用 java 语言的开发者开发交互的信息可视化程序。它可以用来建立独立的应用程序，在大型应用中的可视化组件和 web applets。最早的 prefuse 提供 Java 编程语言的可视化框架，prefuse flare 工具包为 ActionScript 和 Adobe Flash Player 提供可视化和动画工具。

支持由表、图、树组成的数据结构，字段的数据索引和选择列的查询，并且高效的利用内存。已存的组件帮助完成布局、颜色、大小和形状设定，变形、动画及更多功能。包括与用户交互和操作的一组库函数。通过一组活动的时序机制

来实现动画。可视化的变形效果，包括物体位置移动和通过空间和语义的放大放小。动态查询过滤数据显示。融合使用了 lucene 文本查询 engines。在位置和动画中利用了物理学中力的摸拟。灵活的多种显示方式，包括概貌+详细的方式和多个显示图。内建类似于 SQL 的语言语句可以针对数据进行行和列的操作。支持数据库的查询结果集合和 prefuse 内部数据的映射。可以利用经过简化的，对开发者友好的 API，建立自定义的过程，用户交互和画图像的组件。

2、超强的数据可视化工具：Processing[12]

Processing 的创始者： Casey Reas 与 Ben Fry 是美国麻省理工学院媒体实验室 (M.I.T. Media Laboratory) 旗下美学与运算小组 (Aesthetics & Computation Group) 的成员。美学与运算小组由著名的计算机艺术家 John Maeda 领导，于一九九六年成立至今，在短时间内声名大噪，以其高度实验性及概念性的作品，既广且深地在艺术及设计的领域里，探索计算机的运算特质及其带来源源不绝的创造性。极少数人能完美结合并平衡艺术家、设计师和计算机工程师的才华于一身，更重要的是 Casey 和 Ben 拥有开放源码的胸襟。

MIT 理工学院媒体实验室成立于 1980 年，现拥有 50 名教授和科学家，下设 33 个研究小组，在读博士和硕士研究生 150 名，每年研究经费为 3000 万美金，其中 75%都来自企业近 150 家公司的赞助。实验室的研究范围为媒体技术、计算机、生物工程、纳米和人文科学。现已成立的研究小组有：分子计算机、量子计算机、纳米传感器、机器人、数字化行为、全息技术、模块化媒体、交互式电影、社会化媒体、数字化艺术、情感计算机、电子出版、认知科学与学习、视觉和模型等。

3、代码可视化工具：Seesoft[13]

贝尔实验室的 Eick 等人在可视化系统 SeeSoft 中实现了一种对上百万行计算机程序进行可视化的方法。SeeSoft 可以用于知识发现、项目管理、代码管理、开发方法分析等领域。SeeSoft 曾经被用于帮助检测大型软件中与“2000 年问题”有关的代码。如今在软件项目管理中 SeeSoft 还将继续发挥重大的作用，比如：版本控制系统，可以跟踪版本时间、程序员代码更新情况、整个项目代码变更情况等等；静态代码分析，比如显示函数调用的位置；动态代码分析，如特征数据。

Seesoft 有四个关键元素：降低显示量，通过颜色显示统计值，直接操作，读取实际代码的能力。降低显示量主要是通过将代码用薄薄的一排排来显示代码文档。每一排的颜色显示主要取决于代码的更新情况。在多数案例中，统计值主要是代码被创建数据。用颜色来描绘代码更新情况的微型图像。运用灵活的交互

技术，用户可以方便地了解代码的更新情况，以及进一步对某段代码进行细致观察，用户通过显示界面进行操作找到感兴趣的部分。为了查看实际的代码文本，用户可打开阅读窗口定位实际代码文本的位置来查看具体的代码情况。每个纵列显示代码文件的大小，文件大小超过一纵列则换一个纵列继续显示。

运用灵活的交互技术，用户可以方便地了解代码的更新情况，以及进一步对某段代码进行细致观察，另外可以通过附加窗口显示具体的代码。Seesoft 一般用于知识发现、项目管理、代码调整以及开发方法分析等领域。

2.2 文本相似度研究

目前，国内外很多学者都在研究文本相似度的计算问题，并且提出了一些解决方案。下面本文将从基于向量、基于语义以及基于属性三个方面介绍国内外的文本相似度研究现状。

2.2.1 基于向量的相似度

G Salton 于 1969 年提出的向量空间模型 VSM(Vector Space Model)[14]，它的基本思想是把文本内容简化为向量空间中以特征项的权重为分量的向量运算，并且它以空间上的相似度表达语义的相似度，通过词频统计和向量降维处理计算相似度。模型中，文本与查询式表示成以词为元素组成的向量，每个词根据词频tf，与逆文本频率idf被赋与一定的权值，然后通过向量元素间余弦角的计算得到文本与查询式间的相似度．其定义如下：

$$q = \left(w_{q_1}, w_{q_2}, \dots, w_{q_m}\right)^T$$

$$d = \left(w_{d_1}, w_{d_2}, \dots, w_{d_m}\right)^T$$

$$sim(q,d) = cos(q,d) = \frac{\sum_{j=1}^{m} w_{q_j} w_{d_j}}{\sqrt{\sum_{j=1}^{m} {w_{q_j}}^2} \sqrt{\sum_{j=1}^{m} {w_{d_j}}^2}}$$

其中，q是查询式向量，d是文本向量，t为文本集中的词语数，w_{q_j}为词在查询式q_j中的权值，w_{d_j}为词在文本d_j中的权值。

基于向量的文本相似度计算方法是最常用的文本相似度计算方法，该方法将要比较相似度的文本根据文本中的词语将文本映射为 n 维空间向量，然后通过比较向量间的关系来确定文本间的相似度，其中最为常用的方法是计算向量间的余弦系数，但传统向量空间模型缺点是模型中各词语间相互独立，无语义上的关系。

为此，广义向量空间模型[15] (Generalized Vector Space Model，GVSM)就利用文本而不是用词来表示词间关系。其相似度计算公式为：

$$\mathrm{sim(q,d)} = \cos(\mathrm{A^T q}, \mathrm{A^T d})$$

其中 A 是$m \times n$的文本矩阵，m是单词数，n是文本集中的文本数。

隐性语义索引模型[16](Latent Semantic Indexing，LSI) 是近年来逐渐兴起的不同于关键词检索的搜索引擎解决方案，其检索结果的实际效果更接近于人的自然语言，在一定程度上提高检索结果的相关性，目前已被逐渐的应用到图书馆、数据库和搜索引擎的算法当中。隐性语义索引模型扩充了广义向量空间模型，描述文本与文本之间的关系，先从全部的文档集中生成一个标引项一文档矩阵，该矩阵的每个分量为整数值，代表某个特定的标引项出现在某个特定文档中次数。然后将该矩阵进行奇异值分解，较小的奇异值被剔除。结果奇异向量以及奇异值矩阵用于将文档向量和带比较文本向量映射到一个子空间中，在该空间中，来自标引项一文档矩阵的语义关系被保留，同时标引项用法的变异被抑制。最后，可以通过标准化的内积计算来计算向量之间的夹角余弦相似度，根据这个值来比较文本间的相似度。计算公式如下：

$$A = U\Sigma V^T$$

$$L = U\Sigma^{-1}$$

$$sim(q,d) = cos(L^T q, L^T d)$$

其中矩阵A可以分解成 3 个矩阵U，Σ，V^T的积，其中矩阵U与矩阵V^T的列向量是正交归一化的，矩阵Σ是对角矩阵，U是$m \times t$矩阵，Σ是$t \times t$矩阵，V是$n \times t$矩阵。

2.2.2 基于属性的相似度

可以看到，相似度计算方法都以余弦角公式为计算基础，在向量模型及其扩展模型中广泛使用。但利用余弦角测试获得向量相似度的方法没有严格的理论根据[17]，潘建红等人[18]试图以属性论为理论依据，以向量为表现形式，建立属性重心剖分模型(Barycenter of the Attribute Coordinate Model，BACM)，通过坐标点与坐标点的距离计算关键词与关键词的相关性，通过坐标点与单纯形的关系计算关键词与文本的相关度，通过单纯形与单纯形的关系计算文本与文本的相似度。在属性坐标系中描述文本向量与查询式向量，确定向量之间的匹配基准，计算匹配距离，从而建立一个新的文本与查询式间相似度评判公式。实验结果表明，此方法具有与空间向量模型相似的效果。因此，利用文本属性重心剖分模型能表达较多的语义信息,它为相似度的处理方法提供了另一种可能。

袁正午等人[19]针对上述基于属性的中心剖分模型在语义信息丢失和效率低

等问题提出了一种改进模型，通过计算查询线与文档单纯形的交点与文档重心点之间的相似度，使得结果保留属性坐标系中文档向量的特征。实验结果表明，该模型的查全率、查准率和 F1 值可以提高 2%~4%左右。

2.2.3 基于语义的相似度

随着因特网技术的发展，网上信息量以指数规律迅速增长.如何准确地获得有价值的网上信息资源；怎样对日益庞大的数据、信息进行处理，快速、准确地找到符合用户需求的信息成为当今信息领域的研究热点。概念是组成信息的基本单位，因此概念之间语义相似度的计算精确度对信息检索的效率起着决定作用。目前，国内外学者在词语的相关性研究方面做了大量工作。

Dekang Lin[20] 于 1998 年提出了一组具有广泛意义的相似度定义，他认为任何两个事物的相似度取决于它们的共性(Commonality)和差异(Differences)，相似度定义如下：直觉告诉我们，对象 A 和 B 之间的相似度与它们之间共性和差异相关，两个对象所拥有的共性越多，则相似度越大，而两个对象之间的差异越多，则相似度越小。当两个对象 A 和 B 是同一个对象时，相似度达到最大。当 A 和 B 无关或独立时，相似度最小。然后他从信息理论的角度给出了任意两个事物相似度的通用公式：

$$Sim(A,B)=\frac{logp(common(A,B))}{logp(description(A,B))}$$

其中分子是描述 A、B 共性所需要的信息量的大小；分母是完整的描述出 A、B 所需要的信息量大小。

刘群等人[21]对 Dekang Lin 的描述做了进一步具体化，认为两个词语的相似度是它们在不同的上下文中可以互相替换且不改变文本的句法语义结构的可能性大小，他们通过分析《知网》的知识描述结构，利用义原的上下位关系计算义原相似度，进而得到词语的相似度。在此基础上，江敏等人[22]进一步考虑了义原的深度信息，并利用《知网》义原间的反义、对义关系和义原的定义信息来计算词语的相似度，在词语极性识别实验中，得到了较好的实验结果。

Ehrig M 等人[23]将相似度计算视为本体映射过程中的一个重要步骤。他们将本体映射定义为：对于给定的两个本体O_1和O_2，映射是指对本体O_1中的每个实体(概念C，关系R或实例I)，我们试图在本体O_2中找到与它有相同或相近语义的对应实体，本体实体的相似度定义如下[23]：

$$Sim:\varepsilon\times\varepsilon\times O\times O\rightarrow[0,1]$$

1）$Sim(e,f)=1$，若两个实体e和f是完全一致的；

2）$Sim(e,f)=0$，若两个实体e和f是不同的而且没有任何共同特征；

3）$Sim(e,f)=Sim(f,e)$，相似度是对称的；

4）$Sim(e,e)=1$，相似度是自反的。

其中，ε是实体集合，O是本体，e，f是实体。

朱礼军等人[24]引入了计算语言学中的语义距离思想来计算领域本体中概念间的相似度。结果表明，该方法可以定量地分析概念、特性之间的相似度，并可以指导基于领域知识本体的语义查询中的概念集扩充和查询结果排序。陈杰等人[25]提出的算法将概念相似度计算分为两层。一层是概念语义初始相似度层．其主要利用概念之间的距离来计算概念的初始相似度。另一层是概念非上下位关系相似度层，其在概念初始相似度的基础上，计算概念通过非上下位关系体现出的相似度。最后通过综合计算，得到领域本体中概念的实际相似度。

李荣等人[26]提出一种综合的概念相似度计算方法，在计算概念相似度时，不仅考虑概念本身的语义，而且考虑概念的属性和上下文结构。对于本体O_1中的一个概念A，计算概念A和本体O_2中所有概念的概念名称相似度，设定阈值，产生概念A的候选概念集。之后只对概念A与候选概念集中的概念计算基于结构、基于属性的概念相似度，再进行相似度的综合，最后通过实例验证该计算方法具有较高的查全率和查准率。

兰美辉等人[27]提出了一种改进的相似度计算模型。该计算模型充分利用了本体层次树的结构特点，不仅考虑了语义距离、层次差、语义重合度，而且考虑了结点密度和边类型，综合这 5 个因素来计算本体概念之间的相似度。通过把结点密度和有向边类型考虑其中，从而更加全面地量化了本体网络中概念结点之间的语义相似度，提高了概念之间语义相似度量化的准确性。

各种文本相似度计算方法均在特定领域取得了良好的效果，但还都存在着缺点与不足，尚需进一步加以改进。

2.3 本章小结

本章主要介绍了相关理论和方法。首先介绍了信息可视化的概念以及具有代表性的信息可视化工具。其次，研究并介绍了目前几类主要的文本相似度算法，为本文后续的系统设计做准备。

第三章 药品相似度计算模型设计

3.1 药品信息地图

目前医药电子商务行业正迅速发展，但药师等专业人才的供给无法满足该行业的增长速度。再加上国内医药电子商务平台的医药类产品种类可达上万种，面对药品庞大的数据信息及快速的更新速度，即使是拥有医学背景的人员也无法及时了解药品知识。对于医药电子商务行业的工作人员，在没有专业医药知识背景的情况下，如何从庞大的药品数据库以及越发快速的药品数据更新中，快速有效的寻找所需要的药品信息，并对顾客实现合理的药品推荐，是我们要解决的问题。

本文针对这个问题，提出了药品信息地图的概念，并对基于相似度计算的药品信息可视化地图进行了分析研究，通过将庞杂的药品数据库转化为可以直接向客户进行推荐的药品知识库地图，为用户清晰地展示药品之间的关系脉络。

所谓的信息地图，就是一种使用地图的形式来描绘信息资源，能够指明信息资源的方位、关联性、具有导航功能的可视化工具[28]。相对于现在常用的网上全文搜索引擎系统，信息地图不仅仅能够指出信息资源的方位，而且能够体现它们之间更具重要性的联系，还能够帮助用户从全局了解信息的分布状况。

Keith V. Nesbitt[29]第一次提出将地图概念应用到不同领域以更有效地显示抽象知识，帮助非专业用户来理解该专业领域的知识。例如构造一个博士论文的信息地图来描绘论文中不同想法之间复杂的互联关系，帮助读者更好地了解论文的整体架构和内容；构造一个商业计划地图来描绘计划中不同事务之间的相互关联；构造一个页面导航地图来帮助学生有效定位到特定的资源；构造一个课程地图帮助学生理解该课程的整体结构。

目前在信息地图领域比较著名的研究团队是卡内基梅隆大学的Dafna Shahaf和微软研究院的 Eric Horvitz[30]，他们起初主要关注新闻领域[31]，生成的信息地图不仅包含了相关的新闻集合，也包含了新闻之间的关联性，让用户更好地理解特定领域发生的新闻故事。同样的，这种信息地图技术也被 Dafna Shahaf 用在了科技文献中[32]，通过生成信息地图，用户不仅能够找到相关的科技文献集合，同时也能了解该领域下科技研究的发展动态。

根据文献[33]提出的药品信息地图定义：

定义 药品信息地图M是(L, V, τ)的集合。其中，τ是相关性阈值，L是相关性大于τ的相关线路集合，$V = (id, name, Fea)$是药品集合。其中id是药品 id，name是药品名称，Fea是药品属性的集合。

对一个信息地图来说，每个节点对应一个药品，包含药品 id，药品名称，药品属性等信息。每条相关线路的相关性都大于τ。药品属性是对药品说明书进行中文分词处理后所得到的一系列属性单词的集合。

该定义中，药品之间的关系由相关性确定，根据文献[33]，相关性的定义只是计算了两两药品之间重合属性的数量之和，对于描述药品与药品之间的关系具有一定的局限性，本文将对药品做进一步的细化分析，即使用相似度计算来描述药品之间的关系。

本文所要实现的药品信息可视化地图将以树状形式显示，即对于每一个药品节点来说，和该药品有关联的其他药品将作为它的子结点呈环绕状显示，而每一个子结点也拥有自己的子结点，以此类推。其中，每个结点包含药品 id，药品名称，药品类型等信息，而每条边则对应药品之间的关系，由药品间的相似度决定。因此，我们对定义做出如下修改：

定义 药品信息地图M是(E, V, τ)的集合。其中，τ是相似度阈值，E是相似度大于τ的药品对，$V = (id, name, type, spec, attr)$是药品集合。其中$id$是药品 id，$name$是药品名称，$type$是药品类型，$spec$是药品说明，$attr$是药品属性集合。

3.2 药品数据模型

根据药品信息地图的定义，我们对数据库做相应的设计，在数据库中设计了两个数据表，ER 图如下：

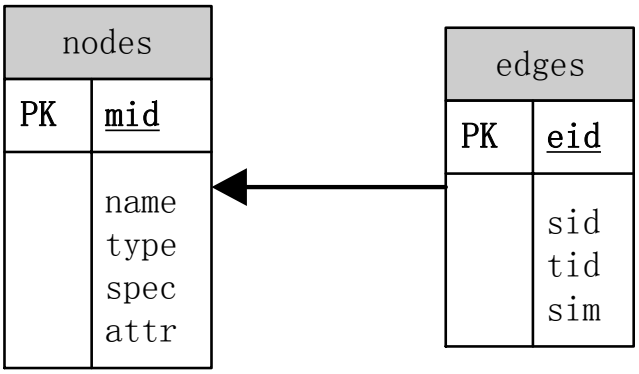

图 2 药品数据库 ER 图

1、nodes 表

nodes 表，即结点表，描述的是药品信息，每条记录包含的信息有：药品 id，药品名称，药品类型，药品说明以及药品属性。其中，药品属性是对药品说明书进行中文分词处理后所得到的一系列属性单词的集合，以逗号分隔，便于后续的相似度计算。

数据序号	数据名称	数据类型	数据说明
1	mid	Primary key，Integer	药品 id
2	name	Char	药品名称
3	type	Char	药品类型
4	spec	Char	药品说明
5	attr	Char	药品属性

表 1 nodes 表

2、edges 表

edges 表，即边表，描述的是药品与药品之间的关系，每条记录包含的信息有：边的 id，该边的两个药品结点各自的药品 id，以及两种药品间的相似度值。其中，edges 表中的 sid 和 tid 分别表示组成一条边的两个药品的 id。sim 值是通过计算 sid 和 tid 两个药品的相似度得到。药品信息可视化地图是通过给出一个相似度阈值，然后从 edges 表中选择 sim 值大于该相似度阈值的所有边绘制而成的。

数据序号	数据名称	数据类型	数据说明
1	eid	Primary key，Integer	边的 id
2	sid	Integer	药品 id
3	tid	Integer	药品 id
4	sim	Double	药品间的相似度

表 2 edges 表

3.3 药品相似度计算

Nenadic 等提出的 Nenadic 算法[34]是一种基于语词特征的术语相似度典型算法。该算法根据语义表达能力的强弱，为术语中心词的匹配分配更多的相似度权重，并基于筛子系数（dice-like coefficient）方法来计算两个术语的语词相似度。如果两个术语相匹配的词语越多，就会得到更多的相似分值；如果它们具有相同的中心词，那么相似度值将会更高。两个术语之间的相似度计算定义如下：

$$LS(t_1,t_2)=\frac{|p(h_1)\cap P(h_2)|}{|p(h_1)|+|P(h_2)|}+\frac{|p(t_1)\cap P(t_2)|}{|p(t_1)|+|P(t_2)|}$$

其中，h_1和h_2是术语t_1和t_2各自的术语中心词，$p(t_1)$和$P(t_2)$分别是术语t_1和t_2的非空子序列。$|p(h_1)\cap P(h_2)|$表示两个术语相匹配的中心词数量，$|p(t_1)\cap P(t_2)|$表示两个术语所具有的相同子序列的数量。

我们认为，药品说明书是载明药品重要信息的法定文件，是选用药品的法定指南。药品说明书的内容包括药品的名称、作用类型、成分、性状、禁忌等重要用药信息，是医务人员、患者了解药品的重要途径。下表是从国内某 B2C 医药电子商务网站下载的一份标准的药品说明书。

【商品名称】	同仁堂感冒清热颗粒
【药品类型】	药品，风寒感冒。
【剂型】	颗粒
【成份】	荆芥穗、薄荷、防风、柴胡、紫苏叶、葛根、桔梗、苦杏仁、白芷、苦地丁、芦根。
【性状】	本品为棕黄色的颗粒，味甜、微苦
【适应症】	疏风散寒，解表清热。用于风寒感冒，头痛发热，恶寒身痛，鼻流清涕，咳嗽咽干。
【用法用量】	开水冲服，一次 1 袋，一日 2 次。
【禁忌】	严重肝肾功能不全者禁用。
【药品说明】	1．忌烟、酒及辛辣、生冷、油腻食物。2．不宜在服药期间同时服用滋补性中成药。3．风热感冒者不适用，其表现为发热重，微恶风，有汗，口渴，鼻流浊涕，咽喉红肿热痛，咳吐黄痰。4．有高血压、心脏病、肝病、糖尿病、肾病等慢性病严重者、孕妇或正在接受其它治疗的患者，均应在医师指导下服用。5．按照用法用量服用，小儿、年老体虚者应在医师指导下服用。6．服药三天后症状无改善，或出现发热咳嗽加重，并有其他严重症状如胸闷、心悸等时应去医院就诊。7．药品性状发生改变时禁止服用。8．儿童必须在成人的监护下使用。
【贮藏】	密封。
【有效期】	3 年

表 3 国内某 B2C 医药电子商务网站下载的药品说明书

根据上表的药品说明书，我们发现，能够反映一个药品重要信息有两部分，一个是药品类型：药品，风寒感冒，还有就是药品说明。其中，药品类型是药品信息的抽象信息，药品说明则是药品信息的具体反映。因此，根据 Nenadic 算法的定义，我们将药品说明视为药品术语，而药品类型视为药品术语中心词。药品之间的相似度取决于药品类型以及药品说明之间的相似度分析。

定义 药品相似度。两个药品之间的相似度是对分别对药品类型以及药品说明进行相似度计算，并对结果进行加权求和所得到的，计算公式如下：

$$sim(m_1, m_2) = \alpha sim_{type}(m_1, m_2) + \beta sim_{spec}(m_1, m_2)$$

其中，α和β分别是药品类型相似度和药品说明相似度的权值，并且满足：

$$\alpha + \beta = 1$$

对两个文本之间的相似度计算我们可以使用向量空间模型（VSM）的相似度计算公式。在 VSM 的文本分类过程中，文本的特征向量与各类代表向量的夹角是决定文档归属的重要依据之一，这些夹角的余弦被称作“相似度”。传统的向量空间模型(Vector Space Model，VSM)中，文本与查询式表示成以词为元素组成的向量，每个词根据词频tf，与逆文本频率idf被赋与一定的权值，然后通过向量元素间余弦角的计算得到文本与查询式间的相似度．其定义如下[35]：

$$q = \left(w_{q_1}, w_{q_2}, \dots, w_{q_m}\right)^T$$

$$d = \left(w_{d_1}, w_{d_2}, \dots, w_{d_m}\right)^T$$

$$sim(q,d) = cos(q,d) = \frac{\sum_{j=1}^{m} w_{q_j} w_{d_j}}{\sqrt{\sum_{j=1}^{m} {w_{q_j}}^2} \sqrt{\sum_{j=1}^{m} {w_{d_j}}^2}}$$

其中，q是查询式向量，d是文本向量，t为文本集中的词语数，w_{q_j}为词在查询式q_j中的权值，w_{d_j}为词在文本d_j中的权值。

定义 药品类型相似度。药品类型向量为$(type_1, type_2, \dots type_m)$，其中$type_j$是对药品类型文本进行中文分词处理以后得到的第j个词语，m是分词处理后的词语总数。我们定义两个药品之间的药品类型相似度为两个类型向量之间的余弦夹角值，计算公式如下：

$$sim_{type}(m_1, m_2) = cos(m_1, m_2) = \frac{\sum_{j=1}^{m} type_{1j} type_{2j}}{\sqrt{\sum_{j=1}^{m} {type_{1j}}^2} \sqrt{\sum_{j=1}^{m} {type_{2j}}^2}}$$

定义 药品说明相似度。药品说明向量为$(spec_1, spec_2, \dots spec_m)$，其中$spec_j$是对药品说明文本进行中文分词处理以后得到的第j个词语，m是分词处理后的词语总数。我们定义两个药品之间的药品说明相似度为两个说明向量之间的余弦夹角值，计算公式如下：

$$sim_{spec}(m_1, m_2) = cos(m_1, m_2) = \frac{\sum_{j=1}^{m} spec_{1j} spec_{2j}}{\sqrt{\sum_{j=1}^{m} {spec_{1j}}^2} \sqrt{\sum_{j=1}^{m} {spec_{2j}}^2}}$$

在分别定义了药品类型和药品说明相似度后，该如何确定α和β是我们要考虑的问题。首先，我们根据药品的源数据可以知道，药品类型是预先定义好的，即由专家实现对其进行类型划分，因此，药品类型的划分本身具有主观性。而药品说明则是从药品说明书上获得的，客观的反映了药品本身所具有的特性。因此，我们可以对药品类型和药品说明进行相似度计算，计算公式如下：

$$sim_{md}(type_{md}, spec_{md}) = cos(type, spec) = \frac{\sum_{j=1}^{m} type_j spec_j}{\sqrt{\sum_{j=1}^{m} type_j{}^2}\sqrt{\sum_{j=1}^{m} spec_j{}^2}}$$

一般来说，如果药品类型与药品说明之间的相似度越大，说明药品类型的划分越合理。也就是说，药品间的相似度可以更倾向于受药品类型相似度的影响；反之，如果药品类型与药品说明之间的相似度越小，说明药品类型的划分越不合理，因此，我们更希望药品之间的相似度更大程度上由客观的药品说明决定。因此，α和β定义如下：

$$\alpha = \frac{sim_{md_1} + sim_{md_2}}{2}$$

$$\beta = 1 - \frac{sim_{md_1} + sim_{md_2}}{2}$$

从公式中可以看出，当分子越大时，α的值越大，β的值越小，即药品类型相似度的贡献越大，而药品说明的贡献越小；反之，当分子越小时，α的值越小，β的值越大，即药品类型相似度的贡献越小，而药品说明的贡献越大。

3.4 本章小结

本章主要建立了药品相似度计算模型。首先，通过分析药品说明书，提出了药品信息地图的概念。其次，根据药品信息地图的定义，提出了药品数据模型，并设计了相应的药品数据表。最后，在此基础上，建立了药品相似度计算模型，并给出了具体的计算公式。

第四章 药品相似度算法实现与验证

4.1 算法流程设计

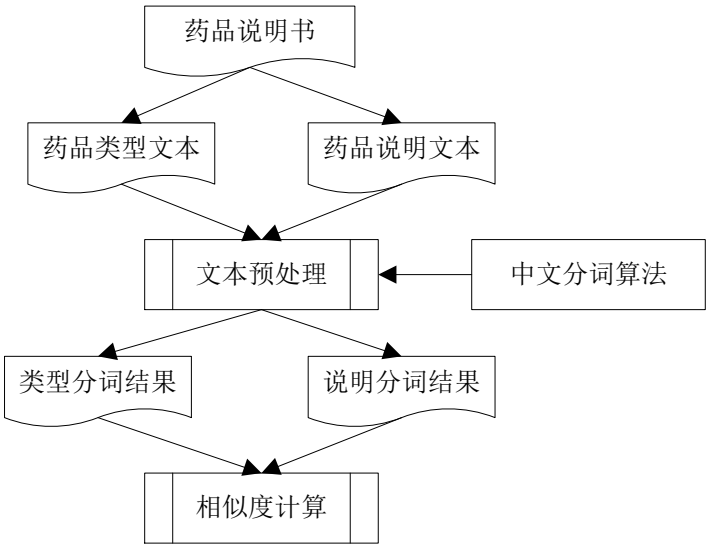

图 3 药品相似度分析流程图

药品相似度分析的整个流程步骤如下：

1、提取药品说明书中的药品类型文本和药品说明文本。

2、对药品类型文本和药品说明文本分别使用中文分词算法进行文本预处理，得到药品类型分词结果和药品说明分词结果。

3、将药品类型和药品说明分词结果代入药品相似度计算模型中进行计算，并得出药品相似度结果。

4.2 文本预处理

在构造药品信息地图之前，对药品数据进行处理是必不可少的一个环节。药品的重要信息存储于药品说明书中，如何从药品说明书中提取药品的属性集合是我们要考虑的问题。

本文主要使用中科院的 ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System)中文分词算法。ICTCLAS 是中国科学院计算技术研究所在多年研究工作积累的基础上，研制出的汉语词法分析系统，主要功能包括：中文分词，词性标注，命名实体识别，新词识别，同时支持用户词典。ICTCLAS 算法采用隐马尔科夫模型，分词速度及分词精度效果均不错，是目前世界上最好

的汉语词法分析器之一[36]。

我们采用 Ansj 中文分词工具对药品说明书进行分词处理。Ansj 基于 ICTCLAS 中文分词算法，比其他常用的开源分词工具的分词准确率更高。它是一款纯 Java 的、主要应用于自然语言处理的、高精度的中文分词工具，目标是“准确、高效、自由地进行中文分词”，支持行业词典、用户自定义词典[37]。

Ansj 算法不仅能够快速高效地获取分词结果，同时，该算法已经实现了用户自定义词典的动态添加删除，也支持从文件加载词典。通过加载药品行业词典，可以得到更为准确的药品属性结果。

我们对药品信息作分词处理，数据库中将药品说明信息存在了 nodes 表中的 spec 字段中，因此，在进行中文分词处理之前，首先需要通过连接数据库将 spec 字段中的说明文本读取出来，然后再作分词处理。最后分词结束后，需要将分词结果以逗号隔开的形式存入 attr 字段中。具体过程如下：

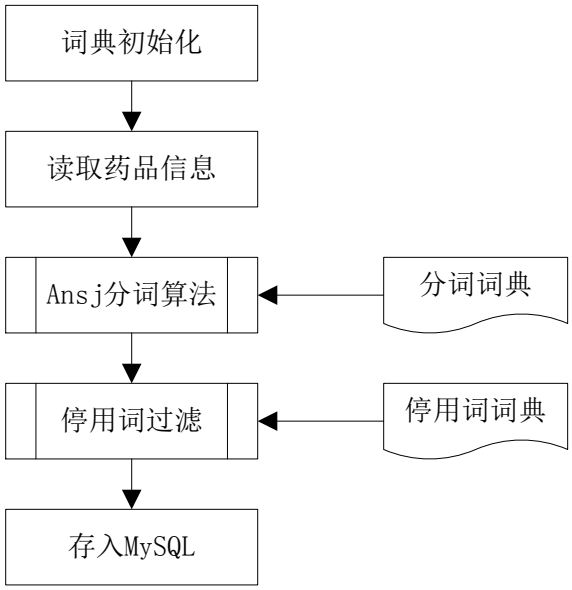

图 4 中文分词处理流程图

首先初始化内部词典，也可以在此过程中用户自定义分词词典。然后连接 MySQL 数据库，从 nodes 表中读取 spec 字段的药品说明文本。系统后台利用 Ansj 分词算法按照分词词典进行分词处理。然后对分词处理后的词语集合根据设置的停用词词典对其进行停用词过滤。最后将分词后的结果以逗号分隔存入 nodes 表中的 attr 字段中作为该药品的属性集合。

虽然我们使用了较成熟的 Ansj 分词算法，但分词结果仍然会遇到一些问题，比如说停用词的处理。停用词是指文本中那些词频过高，没有实质意义的词，或者词频很低，不能代表文本主题的词[38]。

药品说明中虽然不会存在像“你”，“我”等代词，但仍然不免会出现一些常

用词，如连词，冠词等等。词、连词、代词等。像中文中的“和”、“以及”、“中”等都属于药品说明中需要处理的停用词。

停用词处理的常用方法是将文档中词频高于某个数值和词频低于一定数值的词过滤掉，在实际操作中，我们借助了停用词词典来完成停用词的过滤工作。停用词词典是由用户自定义的词典，药品说明经过分词处理后，对照停用词词典，过滤与文本主题无关的停用词。

具体操作流程如下：从分词处理后的词语集合中读入一个词。判断这个词是否存在于停用词词典中。如果该词属于停用词，则删除该词，如果不属于停用词，则读入下一个词。

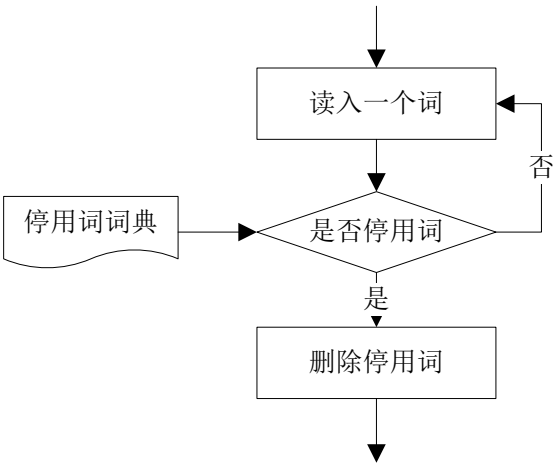

图 5 停用词过滤流程图

4.3 相似度算法实现

在对文本进行分词处理之前，我们事先自定义停用词词典，也可以自定义药品专用词词典，然后使用 Ansj 开源算法进行分词处理，并对分词结果进行停用词过滤，最后将结果转换成为用逗号分隔开的字符串：

```
// 自定义停用词词典
HashMap<String, String> updateDic = new HashMap<String, String>();
HashMap<String,Forest> userForestMap = new HashMap<String,Forest>();
updateDic.put("停用词词典", "Stopword");
updateDic.put("停用词", FilterModifWord._stop);

//自定义药品专用词词典，词典预先存在txt文件中
Forest forest = UserDefineLibrary.makeUserDefineForest(false, "txt");
userForestMap.put("med", forest);
```

```
// 使用Ansj开源分词算法对文本进行分词处理
List<Term> parser =
            ToAnalysis.paser(spec.trim(), userForestMap.get("med"));
parser = FilterModifWord.modifResult(parser);            // 停用词过滤
for(Term term : parser){
    seg += term + ",";
}
```

在对药品文本进行中文分词处理之后，我们需要对分词后的结果进行相似度计算。相似度的计算主要是通过 HashMap 来实现。

HashMap 实际上是一个数组，数组里面的每个元素都是一个链表。每个元素在通过 put 方法放入 HashMap 中的时候，根据该元素自身提供的 hashcode 计算出散列值，该散列值就是数组的下标，同时将新元素放入该数组位置的链表中。

下面将具体介绍通过 HashMap 实现相似度算法的过程，首先我们构造一个 HashMap：

```
Map<String, int[]> SimMap = new HashMap<String, int[]>();
```

这个 SimMap 主要由两部分组成：一个是 String 元素，即相应的药品属性字符串。另一个是长度为 2 的 int 数组元素，该数组分别存放两个药品中该属性出现的频率。

计算两个药品之间的相似度的过程中，读取某一药品的某一个属性，遍历 SimMap 中的 String。如果不存在该属性，put 一个新元素<属性，int[]>，同时该药品下该属性的频率加 1；如果存在该属性，则在该药品下该属性的频率加 1。

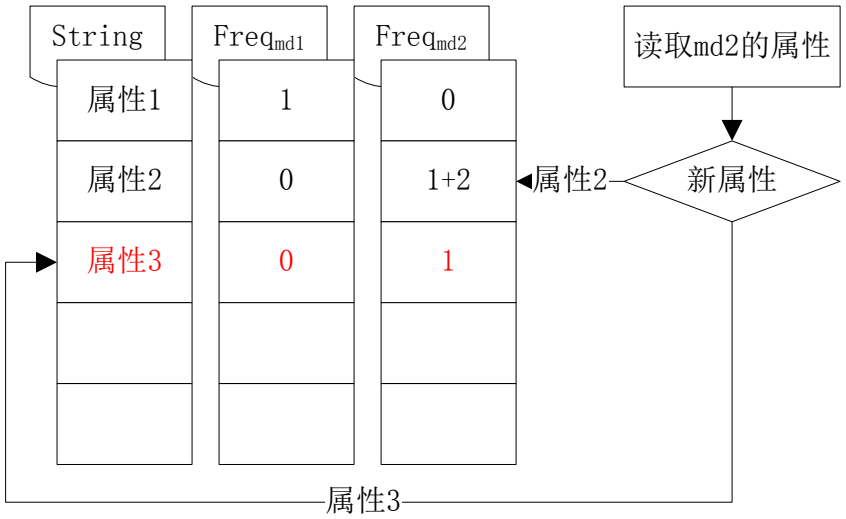

图 6 相似度算法实现流程示例

举例来说，如图所示，左边是 SimMap 的数据结构，String 列储存的是属性例表，$Freq_{md1}$ 和 $Freq_{md2}$ 分别储存的是该属性在 md1 和 md2 中出现的次数。此时，SimMap 中已有数据，分别是属性 1，属性 2。同时，属性 1 在 md1，md2 中出现的次数分别为 1 和 0，而属性 2 在 md1，md2 中出现的次数分别为 0 和 1. 当我们读取 md2 的属性时，判断该属性是否为新属性。假设 md2 的下一个属性为属性 2，即已存在 SimMap 中，那么只要将 md2 下属性 2 的频率加 1，即变为现在 2。假设 md2 的下一个属性为属性 3，即为新属性，那么将<属性 3, (0, 1)>put 到 SimMap 中。

最后，将所有属性的频率信息都存入 SimMap 中以后，我们就可以对两药品之间的余弦相似度进行计算，并根据药品相似度计算模型对其进行加权求和处理：

```
Iterator<String> iterator = AlgorithmMap.keySet().iterator();
double vectorMed1 = 0;
double vectorMed2 = 0;
double denominator = 0;
while(iterator.hasNext()){
    int[] freq = AlgorithmMap.get(iterator.next());
    denominator += freq[0] * freq[1];
    vectorMed1 += freq[0] * freq[0];
    vectorMed2 += freq[1] * freq[1];
}
return denominator / Math.sqrt(vectorMed1 * vectorMed2);
```

4.4 实验分析

我们使用Precision，Recall和FScore来对相似度计算结果进行评价，计算公式如下：

$$Precision = \frac{|Relevant \cap Retrieved|}{Retrieved}$$

$$Recall = \frac{|Relevant \cap Retrieved|}{Relevant}$$

$$FScore = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

其中，Relevant是数据集中人工判定的相似药品对集合；Retrieved是通过相似度算法判定的相似药品对集合。查全率和查准率分别评价算法查找相似药品对的查全性和查准性。FScore则同时考虑了Precision和Recall两个评价指标，能够反映算法的准确性。

为了测试算法的准确性，我们使用药品说明相似度进行对比参照。在实验过程中，我们取 0-0.475 之间以 0.025 为分隔单位的 20 组相似度阈值，分别计算它们的FScore，结果如下图：

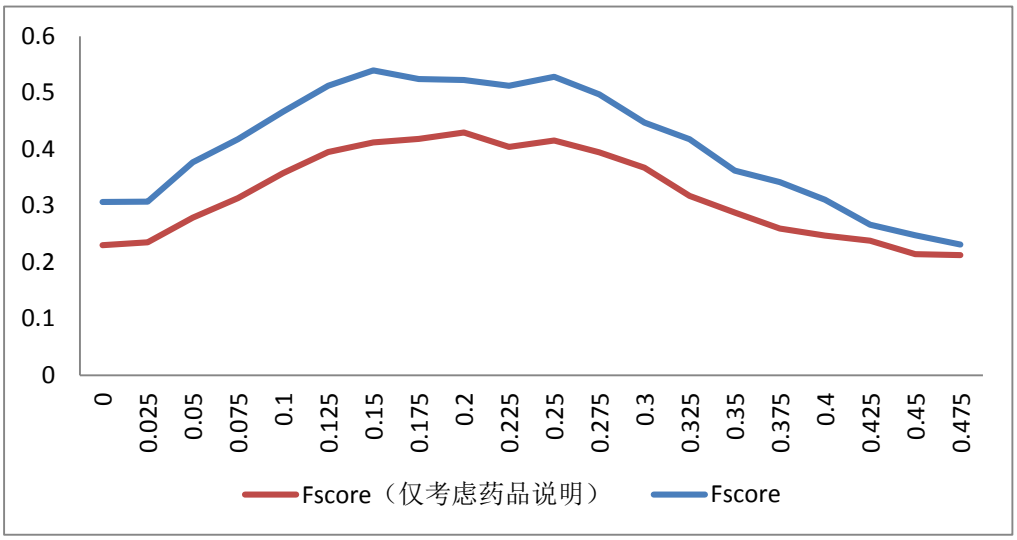

图 7 相似度算法性能比较

我们可以看到，本文提出的相似度算法获得的$FScore$，比仅对药品说明求余弦夹角值的$FScore$要高，也证明了本文提出的相似度算法的准确性更高。

同时可以看到，$FScore$的值随着相似度阈值的变化成曲线分布，最高值在 0.5 左右，而此时，药品间的相似度阈值区间在[0.125,0.25]之间。也就是说，取这个区间的相似度阈值得到的药品信息地图合理性更高。

4.5 本章小结

本章具体实现了药品相似度算法。首先，分析并介绍了药品相似度分析的整个过程步骤。其次，通过使用中文分词算法对药品说明书进行处理。最后，实现了药品相似度算法，并通过实验进行效果分析和评估。

第五章 基于相似度计算的药品信息可视化地图的设计与实现

5.1 系统背景

以国内某 B2C 医药电子商务企业为例，该企业用户的目的是希望在得到顾客的症状描述后，通过关键词搜索能够快速获得该关键词相关的所有药品信息，为顾客对症荐药。

本章将基于药品相似度计算模型实现药品信息可视化地图的原型系统，分别对系统做简要的需求分析，从架构设计和功能设计两方面介绍系统的框架设计，以及对核心功能模块进行详细的分析设计和实现。

通过该系统，用户不需要对所有药品都有了解，药品信息地图可以自动分析药品之间的相似度关系，将庞大的搜索结果转化为结构化的信息地图，为用户清晰地展示药品之间的关系脉络。即使是在药品更新如此迅速的当今，用户也能高效地从庞大的药品数据库中寻找合适的药品。

本章数据来自该企业的后台数据库，共有药品、保健食品、美妆护理、两性、医疗器械等万余种商品数据。单单针对药品信息，就有 2637 种药品信息，分别包括感冒发热、呼吸系统、抗菌消炎、心脑血管等 34 大类药品，其中心脑、消化、筋骨、呼吸、滋补、妇科类药品占总量的 50%左右，如图所示：

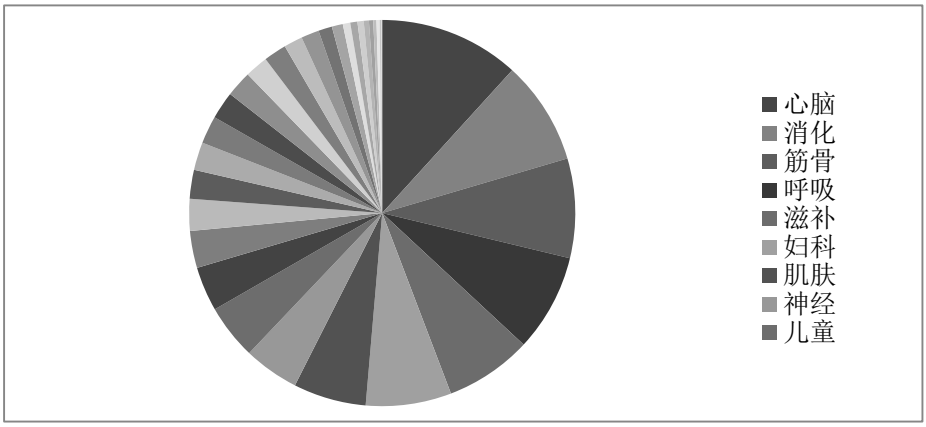

图 8 国内某 B2C 医药电子商务网站的药品分布图

5.2 需求分析

在传统的实体药店里，药品的购药流程比较简单，客户可以直接通过药品名

称向药师获取该药品。而当客户没有办法提供特定药品名称时，可以通过描述症状特征，然后由药师根据其本身所具备的专业医药知识寻找治疗该症状的药品，并为客户提供一系列可选的药品集合供客户进行选择并购买。我们可以看到，在整个流程中，药师的角色至关重要，客户能否购买适合自己的药品完全依赖于药师的专业医药背景和经验。

本文想解决的问题是，对于医药电子商务行业的工作人员，即使是在没有专业的医药知识的情况下，面对当前医药电子商务网站后台如此庞大并且日益更新速度相当快的药品数据库，如何能够在客户提供药品名称或者症状描述之后，快速有效地为其提供合适的可选药品集合推荐。

因此，本文提出的药品信息可视化地图系统主要面向的用户是医药电子商务行业的工作人员。主要解决如何将庞杂的药品数据库转化为可以直接向客户进行推荐的药品知识库地图。整个系统主要实现药品信息可视化，药品检索，药品数据库更新等功能。在实现的过程中，系统主要运用到了中文分词，相似度计算，Prefuse 可视化等技术。系统的主要设计思想是：

1、系统能够自动导入 Excel 药品源数据，该源数据由某医药电子商务网站提供，并能够在后台逐条分析药品记录，对药品信息进行中文分词，相似度计算等处理，最后将处理后的结果存进 MySQL 数据库中。

2、根据 MySQL 数据库中的药品数据，系统能够将药品与药品之间的关系以信息地图的方式在系统主界面进行可视化显示。同时，用户可以通过更改相似度阈值更新药品之间的关系连接，并在系统界面实时更新药品信息地图。

3、系统允许用户与药品信息地图之间的动态交互。用户可以通过移至药品结点查看药品的详细信息，通过点击药品结点，系统会更新该药品成为信息地图的中心结点。

4、系统提供药品检索功能，用户可以通过在搜索框中输入药品关键词，包括药品名称关键词、疾病症状关键词等进行药品检索，系统实时更新检索结果。

5、系统提供药品更新功能。用户可以通过导入新的 Excel 药品源数据表，或者通过选择并删除某一药品，系统会相应更新后台数据库，并在系统界面实时更新药品信息地图。

5.3 系统介绍

5.3.1 系统框架

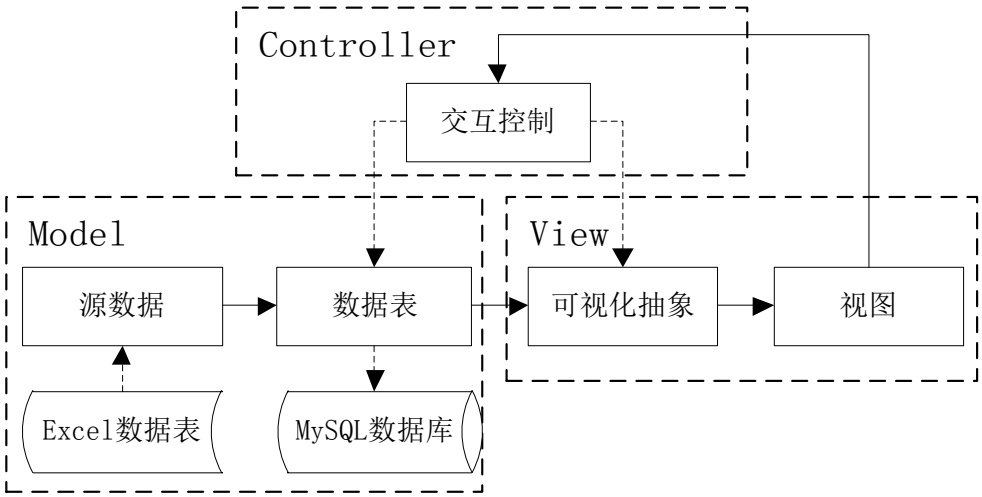

图 9 系统框架图

本文提出的药品信息可视化地图系统采用了 MVC（Model-View-Controller）架构，其中 Model 数据模型由源数据和数据表两个模块组成，View 用户界面由可视化抽象和视图两个模块组成，Controller 控制器由交互控制模块组成。系统的整个框架可以分为以下几个部分：

1、Model 层主要实现对象与数据库中表的连接，完成数据模型的增、删、查、改等功能。主要开发流程包括：系统从 Excel 数据表中获取药品的源数据，然后通过数据处理转换成可以直接用于可视化的数据表。利用 MySQL 数据管理系统来构建数据库，并将数据表存入 MySQL 中。这里所说的数据表是专门为可视化进行服务的，如何构建数据表视具体实现而定。举例来说，可视化可以允许通过不同的图形结构显示，如网状图、树状图等等，为了支持这些结构，在构建数据表时就需要考虑设计相应的数据结构。

2、View 层的主要功能是完成药品信息地图的可视化显示。通过可视化映射，将数据表中的数据转化成可视化抽象。可视化抽象是一个封装的数据模型，包括空间分布、颜色、大小、形状等元素。在可视化抽象中可以定义这些元素的具体数值，通过对可视化抽象里面的数据进行渲染实现最终的交互视图。

3、Controller 层的主要功能是实现用户的动态交互行为，包括拖曳、缩放、移动等。系统捕捉到用户的动作后，将其反馈给 Controller 层，由 Controller 层进行相应的操作，包括对 Model 层中后台数据库的更新操作，或者对可视化抽

象中的元素进行相应更新处理等，最终将处理后的结果返回到 View 层中。

5.3.2 功能模块

本文提出的药品信息可视化地图系统主要实现了后台药品数据库的管理操作，药品信息地图的可视化显示，药品检索，药品数据更新等功能，从而方便用户进行药品信息的检索，实现药品的合理推荐。

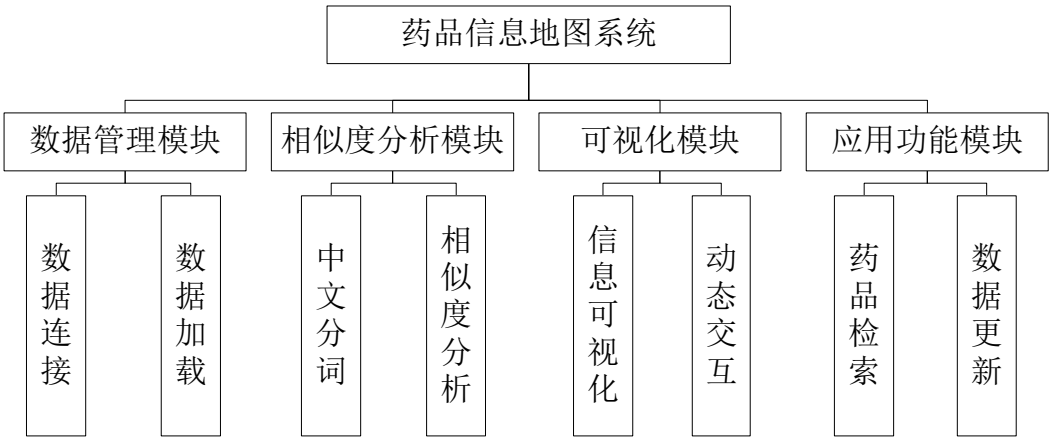

图 10 系统功能模块

1、数据管理模块

该模块主要用于数据库的连接和数据的加载。药品的源数据会预先存放在 Excel 文件中，字段包括药品 id，药品名称，药品类型和药品说明。系统通过 jxl 包连接 Excel 数据库，并逐条导入药品记录，然后通过数据分析模块对药品信息进行处理后将结果通过 JDBC 连接存储在 MySQL 数据库中。

2、相似度分析模块

该模块主要用于数据的分析和处理，首先需要对药品信息文本进行文本预处理工作，也就是本文需要使用到的中文分词处理，然后对分词后的结果进行相似度计算。处理后的结果将存入 MySQL 数据库中，为可视化模块提供基础数据。其中中文分词的主要操作包括去除标点符号、过滤停用词、自定义专业词典等。

3、可视化模块

该模块主要用于信息可视化和动态交互。该模块读取数据库中已处理的基础数据，在系统界面上可视化显示相应的药品信息地图。信息地图以树型结构的形式显示，中心结点被设置在可视化显示的中心位置上，其余药品根据其与中心药品之间的相似度关系呈环状分布。同时，用户可以与系统进行交互行为，包括鼠

标拖曳，地图缩放等等。

4、应用功能模块

该模块主要提供系统的应用功能，支持药品的检索功能：用户通过输入关键词搜索相关药品信息，系统支持用户按药品名称关键词搜索同该药品名称相关的药品集合，以及按症状关键词搜索与该症状关键词相关的药品集合。也支持药品数据的更新功能：用户可以通过打开一个新的 Excel 文件导入新的药品数据，删除药品的中心结点，以及更新相似度阈值来实时更新药品信息地图。

5.3.3 系统交互

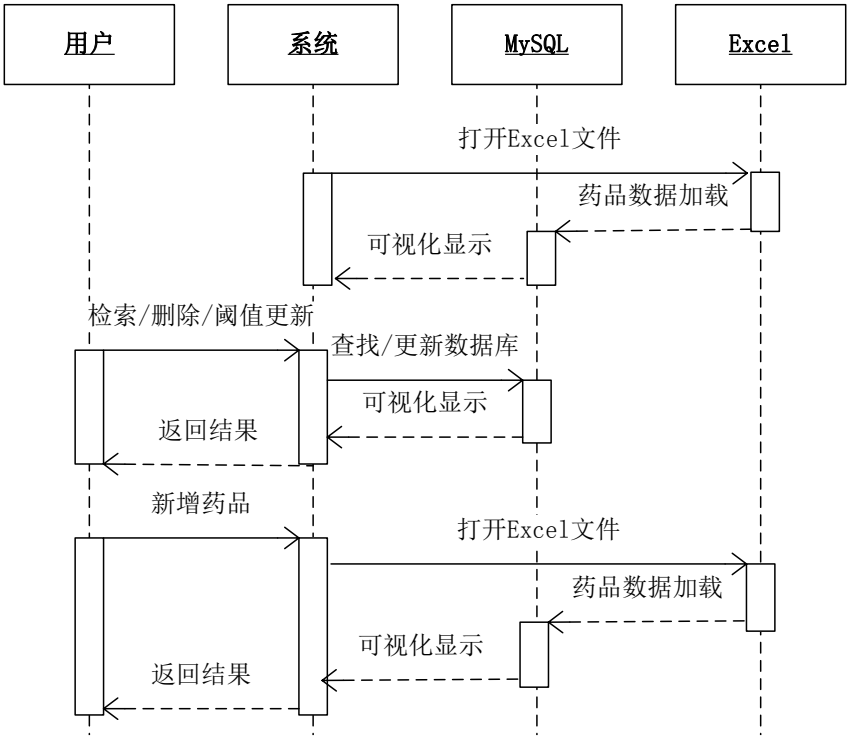

图 11 系统时序图

这里使用时序图来描述用户、系统以及数据之间的交互行为，如图所示：

在数据初始化过程中，系统自动导入 Excel 文件夹中药品源数据，通过对其进行中文分词、相似度计算处理后加载到 MySQL 数据库中，然后在系统界面以药品信息地图的形式可视化显示出来。

在用户与系统的动态交互过程中，对药品的检索、药品的删除以及相似度阈值的更新操作都会涉及到与 MySQL 数据库的连接，其中，药品检索通过关键词检索 MySQL 数据库，获取和关键词相关的药品信息并返回到系统界面。药品删除通过更新 MySQL 数据库，将该药品信息以及同该药品相关的边信息分别从数

据库中删除，然后返回新的药品信息地图。相似度阈值更新则通过实时获取用户修改的相似度阈值，将所有大于该阈值的边信息获取，并相应的更新药品信息地图，返回到系统界面。

在用户进行新增药品的操作过程中，涉及到了系统与 MySQL 数据库以及 Excel 数据库的交互操作。用户通过打开新的 Excel 药品源数据文件，对其中的药品信息进行分析处理后加载到 MySQL 数据库中，其中，不仅仅需要分析该 Excel 文件中药品之间的相似度关系，同时也要更新 MySQL 现有的药品与新药品之间的相似度关系。然后实时的更新药品信息地图，并返回到系统界面。

5.4 数据管理模块

数据管理模块主要负责数据库的连接以及药品数据的加载工作。

根据系统框架中的 Model 层，本文涉及的数据有两种类型，一种是预先存储在 Excel 文件里面的药品源数据，另一种是经过处理后，支持后续可视化操作的存储在 MySQL 数据库中的药品数据表。这两种数据的存储方式不同，因此数据库的连接方式也不同。

药品源数据存储在 Excel 中，因此，本文使用 jxl 包连接 Excel。Jxl 是一个开源的 Java Excel API 项目，通过 Jxl，Java 可以很方便的操作微软的 Excel 文档。对 Excel 源数据进行处理后的药品数据表将存储在 MySQL 中，本文将直接使用 prefuse.data.io.sql 连接数据库。

主要工作流程如下：

1、系统首先连接 Excel 数据表，读入一条药品记录 md，药品信息包括药品 id，药品名称，药品类型和药品说明。

2、系统利用数据分析模块中的中文分词算法对药品说明进行分词处理，得到该药品的药品属性集合。

3、系统读入 MySQL 数据库中 nodes 表里已经存在的每一条药品记录 mdi，取出药品的 attr 字段，即药品属性集合，利用数据分析模块中的相似度算法同 md 的药品属性进行相似度计算。

4、系统判断相似度大小，如果相似度大于 0，则视为两个药品存在关联，为这两个药品之间增添一条边的信息，并存储到 MySQL 数据库中的 edges 表中。信息包括两个药品的 id 号，以及它们的相似度值。

5、系统对 Excel 药品源数据表中的所有药品记录都处理完成后结束。

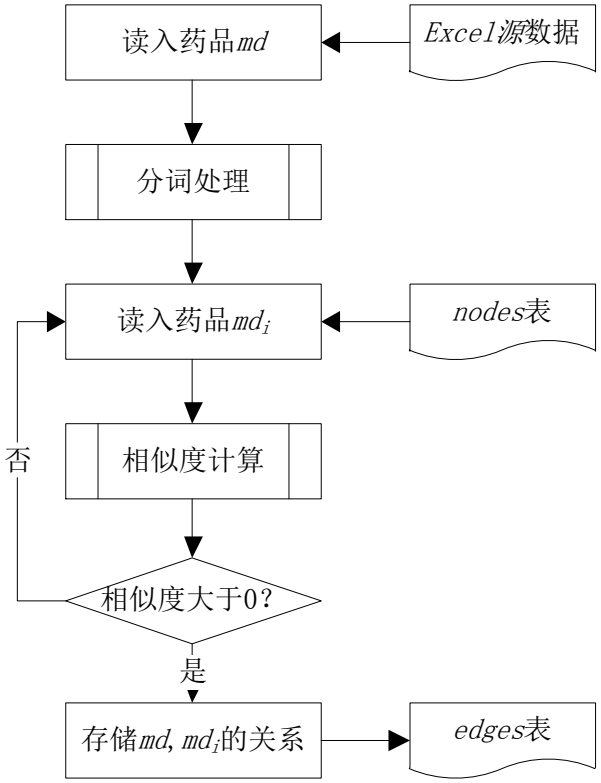

图 12 数据管理模块流程图

5.5 可视化模块

5.5.1 信息可视化

本文使用的 Prefuse 是由美国加利福尼亚州伯克利大学计算机科学分部的 Jeffrey Heer 和 Maneesh Agrawala 开发的交互式数据可视化工具。Prefuse 为数据建模、数据可视化以及用户交互提供了丰富的软件库，可以支持表格、网状和树状图形显示，还具有支持动画制作、动态查询等功能[39]，可用来创建独立的应用程序，或者可视化组件和 Java Applets。Prefuse 可以处理 XML 文件，CSV 文件或者数据库。Prefuse 基于传统的 MVC 架构进行开发，其可视化时需要经过如下处理过程[40]：

1、抽象数据(Abstract Data)。Prefuse 对数据进行可视化的首要步骤是获取数据，并为数据提供了指定的接口和程序，可以显示表、图和多种树形结构。

2、数据过滤。主要是将抽象数据进行提取、转化，使其适用于显示。首先选取要进行可视化的一系列元素。如一个图形或显示在散点图上的重点区域。然后形成一些可视化的属性，如源数据中显示的文字、数字，显示时的坐标点、颜色、大小等等。然后通过 Action 提供为上层组件。

3、数据渲染。即图形绘制的过程，可视化元素通过渲染器绘制到屏幕上，

其中用到上面形成的组件如颜色、位置、大小等等。Prefuse 实现了一些基本渲染器，通过渲染器工厂 RenderFactory 进行管理。

4、交互显示。功能由 Display 组件完成，用于显示 ItemRegistry 中注册的组件。可视化交互功能通过 ControlListener 接口实现．主要是提供对鼠标、键盘的监听功能。

5.5.2 动态交互

在对药品信息地图进行可视化显示以后，我们还为用户提供动态交互操作。包括以下几种操作行为：

1、基本交互行为：指系统仅通过更新可视化抽象中的元素内容进行的基本交互处理，如鼠标拖曳，点击，滚轴放大等，可提供很好的可视化效果和用户体验。同时，为了突出可视化显示效果，用户鼠标移动到所选药品时，该药品会以红色高亮显示；用户输入关键词检索后，检索结果同样会以高亮色显示，以示区别；用户移至某一药品时，该药品的详细信息会自动显示在系统界面上，包括药品名称，药品类型，药品说明，供用户作具体参考。

2、动态更新结点：当用户双击某一药品结点时，该药品会自动移到药品信息地图的中心结点位置，剩余药品根据与其之间的相似度关系呈环绕状分布。通过从中心结点开始的广度优先搜索，由此计算从新的中心结点到每个结点的网络距离，构成新的药品信息地图。

5.5.3 可视化模块实现

<table>
<tr><td colspan="8">MedMapView（药品信息地图可视化）</td></tr>
<tr><td colspan="8">JSpitPane</td></tr>
<tr><td colspan="4">ConstView（创建视图）</td><td colspan="4">FuncCtrl（功能控制）</td></tr>
<tr><td colspan="2">SetGraph（构建图结构）</td><td colspan="2">BasicCtrl（基本控制）</td><td colspan="2">Box</td><td colspan="2">JMenu</td></tr>
<tr><td>NodeTable</td><td colspan="2">InitAction（初始化参数）</td><td>DragControl</td><td>JValueSlider</td><td rowspan="2">JTextArea</td><td colspan="2">menuAction(“新增”)</td></tr>
<tr><td>EdgeTable</td><td>Color</td><td>Font</td><td>WheelZoomControl</td><td>JSearchPanel</td><td colspan="2">menuAction(“删除”)</td></tr>
</table>

图 13 可视化模块

药品信息地图主要由 MedMapView 类来实现，整个视图用 JSpitPane 分成两

个部分，一个部分用来实现药品信息地图的显示和基本操作 ，另一部分用来实现药品信息地图的核心应用功能。

如图所示，ConstView 模块主要用来构建药品信息地图并将其在系统主界面可视化的显示给用户。

1、其中，SetGraph 模块负责同药品数据库连接获取药品详细数据，然后将其转化成图结构，分别导入点和边的信息。

2、InitAction 初始化参数模块来初始化图形的属性参数，包括结点和线条的颜色、字体等。

3、BasciCtrl 基本控制模块控制基本的鼠标动作，包括拖曳、缩放等。

以上是对药品信息地图最基本的构建和控制操作，在此基础上，我们通过 FunCtrl 功能控制模块对该药品信息地图添加相应的功能。

Box 模块将药品检索，相似度阈值更新以及详情显示封装在了一起。

1、通过 JValueSlider 构建一个动态的相似度阈值滑动模块，取值在 0 到 1 之间，用户通过使用鼠标滑动按钮更新阈值，相应的，药品信息地图会随之实时更新，即更新后的药品信息地图反映的药品之间的相似度将大于阈值。

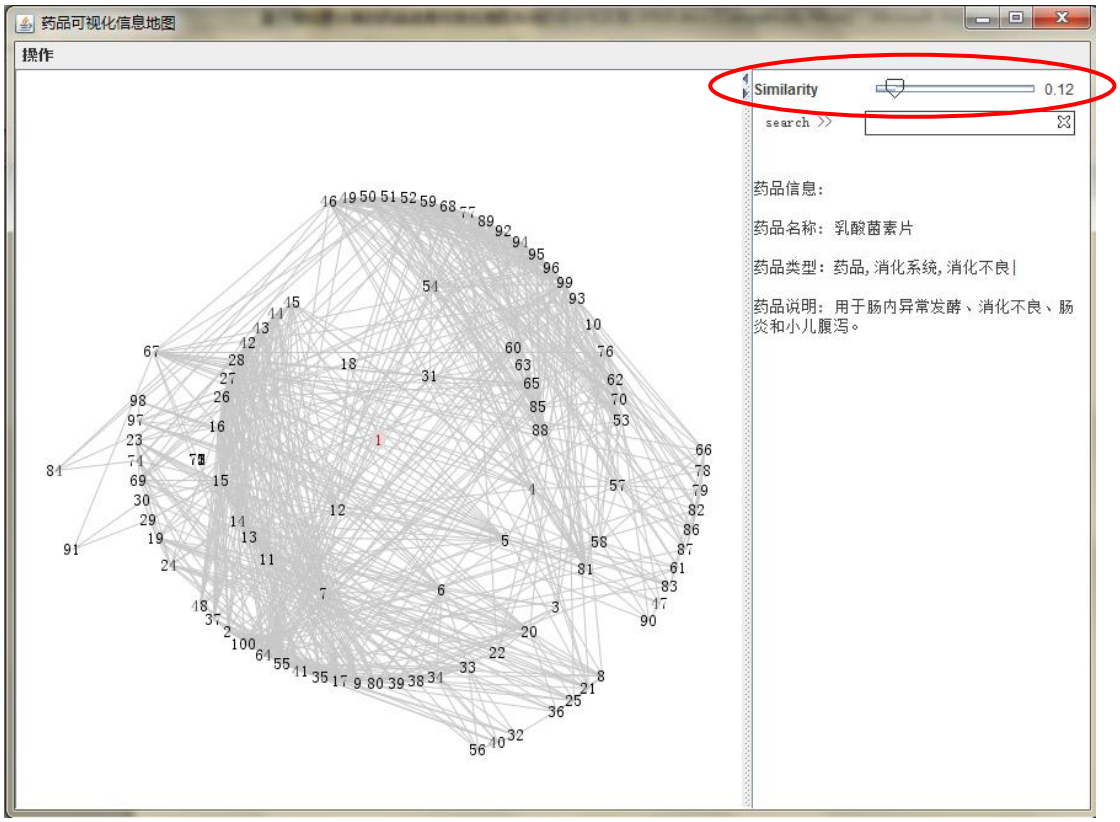

图 14 相似度阈值更新界面

2、JSearchPanel 构建一个搜索框，用户在搜索框中键入药品名字关键词或者症状关键词，通过 Lucene 索引，界面会返回相应的药品结果，并将结果高亮突

出。

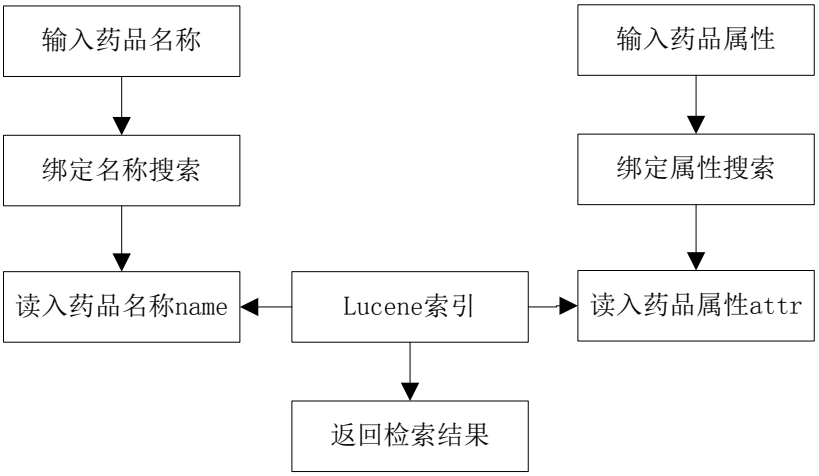

图 15 Lucene 检索流程图

药品检索的过程如下：

1、系统通过绑定 Lucene 的关键词搜索实现两种不同类型的搜索方式。

2、Lucene 分别对药品名称和药品属性建立索引。

3、用户输入药品名称或药品属性关键词(可以是药品功能、症状描述等)。

4、系统对用户请求进行分析，通过 Lucene 索引得到结果集。

5、过滤排序结果，然后将结果作可视化效果处理后反馈给用户。

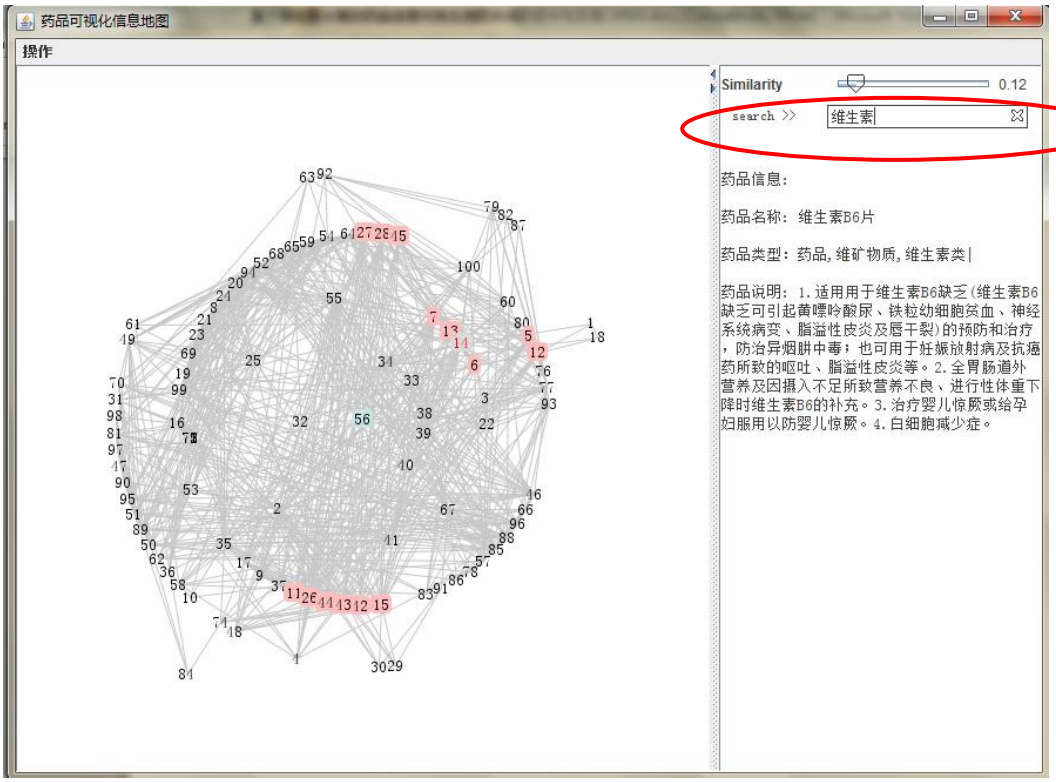

图 16 药品检索界面

3、JTextArea 构建能够显示药品的详细信息的文本框，用户通过将鼠标移至

某一药品结点上方，系统会获取该药品的 id 号，并从后台数据库调用该药品 id 下的药品类型和药品说明，显示在文本框中。

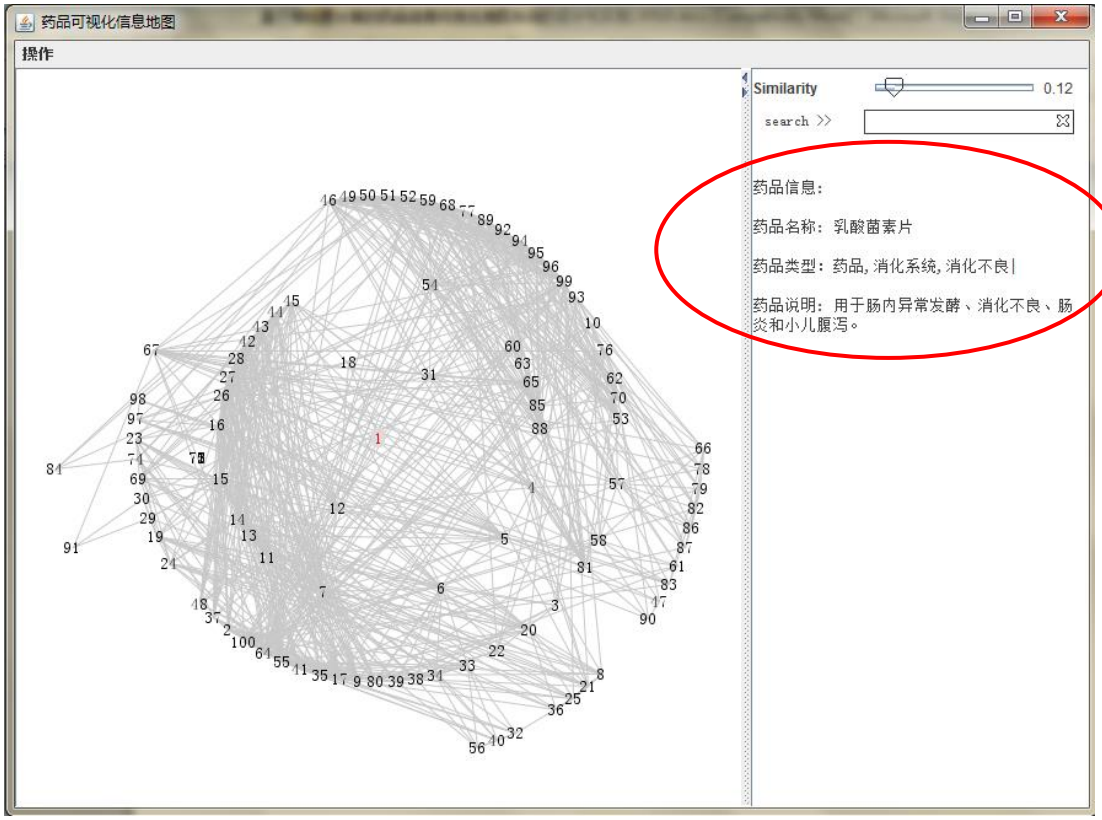

图 17 药品详情显示界面

4、JMenu 模块将对药品数据库的更新操作封装在了菜单栏中。通过“新增”按钮，打开文件管理器，选取 Excel 格式的数据文件，并对其中的药品进行处理更新后更新至 MySQL 数据库中，并实时更新系统界面中的药品信息地图，即更新后的药品信息地图不仅包含了新增的药品结点，同时也更新了新增药品与其他药品之间的关系连线。通过“删除”按钮，系统将处在信息地图中心结点药品从 MySQL 中删除，同样药品信息地图也会实时更新，即更新后的药品信息地图不仅删除了中心药品结点，同时也删除了该药品与其他药品之间的关系连线。

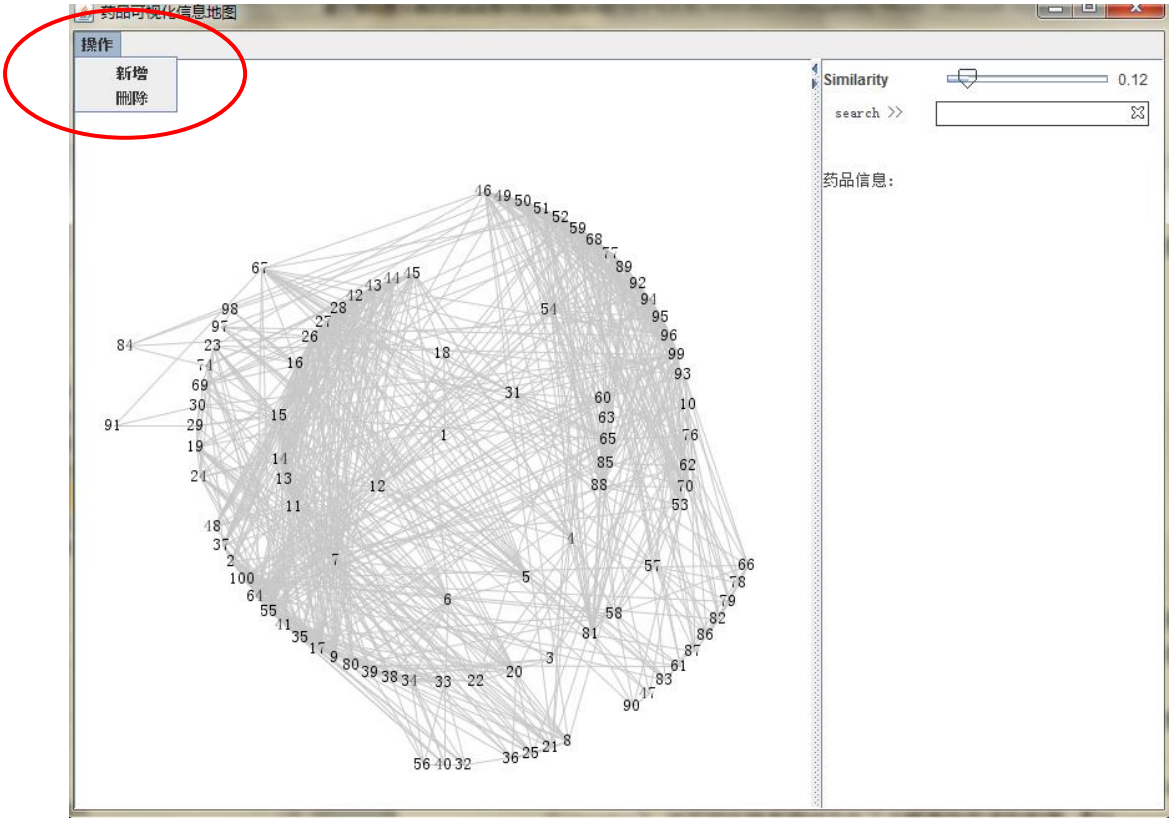

图 18　药品数据更新操作界面

5.6 应用功能模块

5.6.1 药品检索

本文的药品检索功能主要是利用 Lucene 搜索引擎工具包实现的。Lucene 是一个用 Java 写的全文检索引擎工具包，可以方便地嵌入到各种应用中实现针对应用的全文索引/检索功能。Lucene 是 apache 软件基金会的一个子项目，是一个开放源代码全文检索引擎工具包，Lucene 的目的是为软件开发人员提供一个简单易用的工具包，以方便的在目标系统中实现全文检索的功能，或者是以此为基础建立起完整的全文检索引擎。Lucene 源码中共包括 7 个子包，每个包完成特定的功能[41]，如下表所示：

Lucene 包结构功能表	
包名	**功能**
org.apache.lucene.analysis	语言分析器，主要用于切词，支持中文
org.apache.lucene.document	索引存储时的文档结构管理，类似于关系型数据库的表结构
org.apache.lucene.index	索引管理，包括索引建立、删除等
org.apache.lucene.queryParser	查询分析器，实现查询关键词间的运算，如与、或、非等

org.apache.lucene.search	检索管理，根据查询条件，检索得到结果
org.apache.lucene.store	数据存储管理，主要包括一些底层的 I/O 操作
org.apache.lucene.util	一些公用类

表 4 Lucene 包结构功能表

Lucene 功能非常强大，但从根本上说，主要包括两块：一是文本内容经切分词后索引入库；二是根据查询条件返回结果，即索引部分和查询部分。下面结合上述的源码包给出 Lucene 的逻辑结构[42]，从图中我们清楚的看到，Lucene 的系统由基础结构封装、索引核心、对外接口三大部分组成。其中直接操作索引文件的索引核心又是系统的重点。

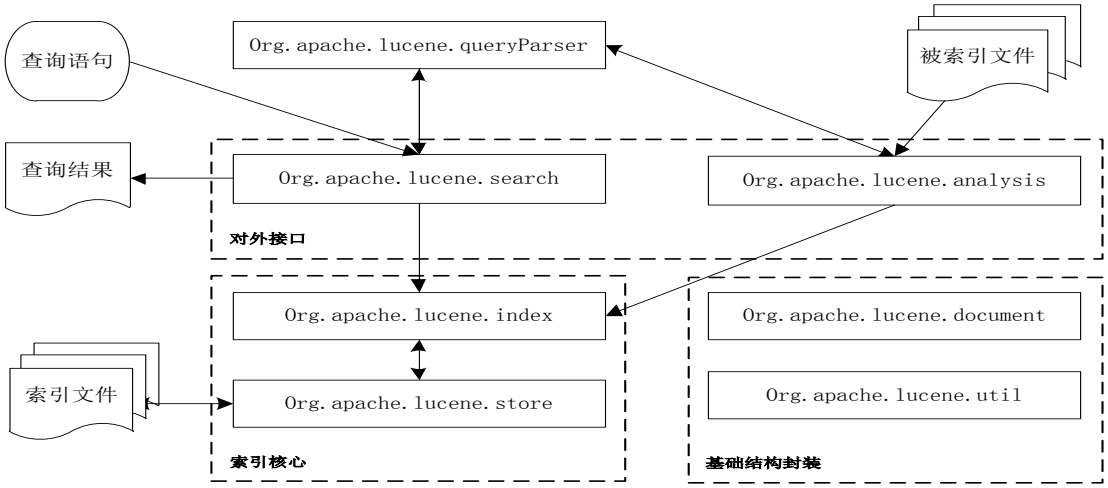

图 19 Lucene 逻辑结构图

5.6.2 数据更新

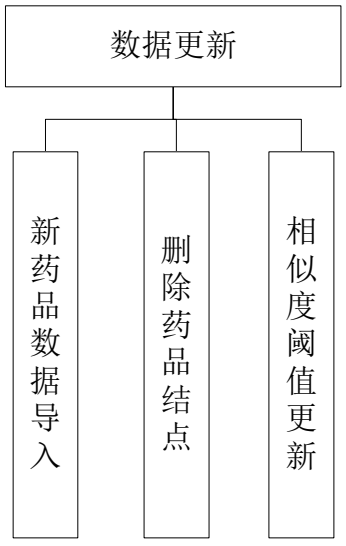

图 20 数据更新模块

数据更新功能主要实现的用户所有可能对后台数据库有更新的操作。即系统会捕捉用户的更新行为，并相应的对后台数据库进行更新，同时反馈给药品信息地图，进行实时更新。

1、新药品数据导入：系统支持用户直接通过界面进行新药品数据的导入工作。用户通过点击菜单栏中的新增按钮打开文件选择窗口，点击相应的 Excel 药品数据文件，系统会在后台自动导入该 Excel 文件中的所有药品记录。并对其进行数据分析处理。最后新的药品记录，以及这些药品记录与原有药品之间的关系会自动更新到药品信息地图中，显示在药品主界面。

2、删除药品结点：系统支持用户对中心药品结点的删除工作。用户通过点击菜单栏中的删除按钮可以删除中心药品节点。系统将获取当前中心药品结点的 id 号，从数据库 nodes 表中删除该药品的记录，并从 edges 表中删除所有和该药品相关的边，同时后台数据库的更新也会实时反应到系统主界面的药品信息地图上。

3、相似度阈值更新：系统支持用户对相似度阈值的更新操作。用户通过滑动屏幕右边的箭头，修改相似度阈值，系统会获取新的相似度阈值，并从后台数据库中的 edges 表中获取相似度大于该阈值的所有边，重新绘制新的药品信息地图。

5.7 本章小结

本章以真实医药电子商务网站为背景，设计并实现了基于相似度计算的药品信息可视化地图系统。首先，对需求进行了简要分析，说明了药品信息地图面向的用户，支持的功能，以及想要解决的问题。其次，从架构设计、功能设计和交互设计三方面介绍了系统的整体框架，本系统划分为数据管理模块，相似度分析模块，可视化模块和应用功能模块四大模块。最后，对核心功能模块进行了详细的分析设计和实现。

第六章 总结与展望

6.1 总结

目前医药电子商务行业正迅速发展，但药师等专业人才的供给无法满足该行业的增长速度。对于医药电子商务行业的工作人员，在没有专业医药知识背景的情况下，如何从庞大的药品数据库以及越发快速的药品数据更新中，快速有效的寻找所需要的药品信息，并对顾客实现合理的药品推荐，是我们要解决的问题。本文针对以上矛盾，对基于相似度计算的药品信息可视化地图进行了分析研究。

本文首先调研了国内外可视化的研究现状，提出了信息可视化地图的概念，同时，在总结和介绍现有相似度算法的基础上，提出了针对药品的相似度计算模型，并设计和实现了基于相似度计算的药品信息可视化地图系统。作者主要工作体现在以下方面：

1、研究并介绍了信息可视化的概念以及具有代表性的信息可视化工具。研究并介绍了目前几类主要的文本相似度算法。

2、通过分析药品说明书，提出了药品信息地图的概念和数据模型。在此基础上，建立了药品相似度计算模型，并给出了具体的计算公式。

3、在前面理论分析的基础上，分析并介绍了药品相似度分析的整个过程步骤。其通过使用中文分词算法对药品说明书进行处理。实现了药品相似度算法，并通过实验进行效果分析和评估。

4、设计并实现了基于相似度计算的药品信息可视化地图的整体架构和功能模块。

6.2 展望

在现有的研究基础上，本文下一步的工作有以下几点：

1、进一步优化药品信息可视化地图。本文提出的药品信息可视化地图主要是通过分析药品本身的说明书进行设计的。在后续的工作中，可以考虑电子商务网站相关的一些数据来进一步优化药品信息可视化地图。

2、同真正的药品电子商务网站集成。从本文的设计和实现可以看出，本文只是一个基于相似度计算的药品信息可视化地图的原型系统，是为了验证模型而独立实现，只是整个医药电子商务网站的一个部分。如何将其整合到现实生活中真正电子商务网站中，实现真正的应用价值是我们今后要考虑和解决的问题。

参考文献

[1] 陈玉文, 沈伟, 闫鸿博. 我国医药电子商务发展的现状及展望. 实用药物与临床. 2006, 9(2): 127-128.

[2] 梁建桥. 我国药品电子商务的现状与思考. 电子商务 E-BUSSINESS JOURNAL. 2009(3): 64-67.

[3] 朱耀华, 郝文宁, 陈刚. 可视化技术简述. 电脑知识与技术. 2012, 08(6): 1402-1407.

[4] Mccormick B H, Defanti T A, Brown M D, et a1. Visualization in Scientific Computing[J]. Computer Graphics, 1987, 12(6): 1103-1109.

[5] Chen Chaomei. Mapping Scientific Frontiers: The Quest Knowledge Visualization[M]. Singapore: Springer-Verlag London Ltd，2003.

[6] 张炯. 可视化信息组织与视频数据库. 情报科学, 2004(2).

[7] 李淑丽. 信息可视化工具的比较研究[D]. 哈尔滨: 黑龙江大学, 2006: 10-13.

[8] Card S Mackinlay J D, Shneiderman B. Readings in information visualization: using vision to think[M]. San Francisco, CA: Morgan Kauffnann Publishers. 1999: 1-34.

[9] Ed H. Chi. A Framework for Information Visualization Spreadsheets. Ph.D. Thesis. University of Minnesota, Computer Science Department. March, 1999.

[10] Card S Mackinlay J D, Shneiderman B. Readings in information visualization: using vision to think[M]. San Francisco, CA: Morgan Kauffnann Publishers. 1999: 85-87.

[11] http://www.civn.cn/p/3547.html. 信息可视化工具包: Prefuse.

[12] http://www.civn.cn/p/5630.html. 超强的数据可视化工具: Processing.

[13] http://www.civn.cn/p/5500.html. 代码可视化工具：Seesoft.罗文华, 阮景, 邱家学. 重视医药电子商务在药品零售连锁价值链中的整合作用. 中国药业. 2005, 14(6): 3-4.

[14] Salton, G. Automatic Information Organization and Retrieval. McGraw-Hill, New York, 1968, Ch.4.

[15] Wong S K M, Ziarko W, Wong P C N. Generalized vector space model in information retrieval. In: Proc the 8th Annual ACM SIGIR International Conference on Research and Development in Information Retrieval. 1985, 18-25.

[16] Deerwester S, Dumais S T, Furnas G W et al. Indexing by Latent semantic analysis. Journal of American Social Inference of Science, 1990, 1(6): 391-407.

[17] Salton Gerard. Developments in automatic text retrieval. Science, 1991, 253: 974-979.

[18] 潘谦红, 王炬, 史忠植. 基于属性论的文本相似度计算. 计算机学报. 1999, 22(6): 651-655.

[19] 袁正午, 李玉森, 张雪英. 基于属性的文本相似度计算算法改进.计算机工程. 2009, 35(17): 4-6.

[20] Lin D. An Information-Theoretic Definition of Similarity[C]. In: Proceedings of the Fifteenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. , 1998. 296-304.

[21] 刘群, 李素建. 基于《知网》的词汇语义相似度计算[J]. 中文计算语言学, 2002, 7(2): 59-76.

[22] 江敏, 肖诗斌, 王弘蔚, 施水才. 一种改进的基于《知网》的词语语义相似度计算. 中文信息学报. 2009, 12(5): 84-89.

[23] Ehrig M, Staab S. QOM--quick ontology mapping //LNCS3298: Proc of the 4th Int Semantic Web Conf. Berlin: Springer, 2004: 683-697.

[24] 朱礼军, 陶兰, 刘慧. 领域本体中的概念相似度研究[J], 华南理工大学学报, 2004, 32(1): 147-150.

[25] 陈杰, 蒋祖华. 领域本体的概念相似度计算. 计算机工程与应用. 2006, 42(33): 163-166.

[26] 李荣, 杨冬, 刘磊. 基于本体的概念相似度计算方法研究. 计算机研究与发展. 2011, 48: 312-317.

[27] 兰美辉, 夏幼明. 基于本体的概念相似度计算模型研究. 曲靖师范学院学报. 2010, 29(3): 67-70.

[28] 周群. 论可视化信息检索系统研究. 情报杂志, 2006, 7:94-96.

[29] Keith V. Nesbitt. Getting to more abstract places using the metro map metaphor. In Information Visualisation ’04.

[30] Shahaf, D., Guestrin, C., and Horvitz, E. Trains of thought: Generating information maps. In WWW ’12, 2012.

[31] Shahaf, D. and Guestrin, C. Connecting the dots between news articles. In KDD ’10, 2010.

[32] Shahaf, D., Guestrin, C., and Horvitz, E. Information maps of science. In KDD ’10, 2010.

[33] 施悦, 申宸, 刘钢, 陈荣华. 医药电子商务的信息地图分析. 计算机应用与软件. 2013.

[34] Nenadic G, Spasic I, Ananiadou S. To insert individual citation into a

bibliography in a wordprocessor, select your preferred citation style below and drag-and-drop it into the document. Automatic discovery of term similarities using pattern mining[c] // Proceedings of International Conference On Computational Linguistics. Taipei, 2002: 1-7.

[35] Salton G. Automatic text processing: the transformation analysis, and retrieval of information by computer. Reading, Pennsylvania: Aoldison-Wesley, 1989.

[36] 刘群, 张华平, 俞鸿魁, 程学旗. 基于层叠隐马模型的汉语词法分析. 计算机研究与发展, 2004, 34(2): 1421-1429.

[37] 孙健. 开源 Java 中文分词器 Ansj. ITeye, 2012[2012-11-05]. http://www.open-open.com/lib/view/open1352527053433.html

[38] 谭琼, 史忠植. 分词中的歧义处理[J]. 计算机工程与应用, 2002(11): 125.

[39] 唐蓓, 夏秋菊. 基于 Prefuse 和社会网络算法的信息检索学科合作网络研究. 图书与情报. 2012, 5: 79-84.

[40] Chu H. Research in Image Indexing and Retrieval as Reflected in the Literature[J]. JASIST, 2001, 52(12): 1011-1018.

[41] Gospodnetic O, Hatcher E. Lucene in action[M].[s.l.]: Manning Publications Co, 2005.

[42] 林碧英, 赵瑞，陈良臣. 基于 Lucene 的全文检索引擎研究与应用[J]. 计算机技术与发展, 2007, 17(5): 186-190.

研究生期间撰写的论文

[1] 施悦, 申宸, 刘钢, 陈荣华. 医药电子商务的信息地图分析. 计算机应用与软件. 2013.

[2] Weidong Zhao, Qinhe Lin, Yue Shi, et al. Mining the Role-oriented Process Models Based on Genetic Algorithm. Proceedings of the Third International Conference on Swarm Intelligence,LNCS7331，Shenzhen, 2012:398-405.

致 谢

在此论文完成之际，谨向给予我指导、关心、支持和帮助的所有老师、同学、亲人和朋友们表示最衷心的感谢！

首先，我要感谢我的导师刘钢老师。刘老师积极工作的态度，和蔼可亲、平易近人的品格都深深地感染我。平时无论是在学业上还是在生活上，刘老师都给予了我很大的帮助，让我能够全身心的在实验室投入工作。在实验室学习的过程中，刘老师不仅为我们提供了宽松的学习环境，还抽出大量时间和我们一起进行了许多有益的讨论和研究，并帮助我确定了论文的研究方向。为此，我感到无比的幸运和感激。

其次，我要感谢徐迎晓老师和赵卫东老师，徐老师在论文上的经验为我提供了不少的指导和帮助，多次阅读我的论文并提出指导性意见，引导我进一步完善论文。赵老师指导我在文献研究上取得了很大的进步，使我的学术研究能力得到了很大的提升。

我还要感谢实验室的林钦和，申宸，王建和张挺同学，在这三年里，他们在生活与学习上给予了我很大的帮助，和你们一起学习工作让我度过了美好的研究生时光。

最后，我要感谢我的父母，他们的支持、理解和鼓励是我永远的精神动力。

论文独创性声明

本论文是我个人在导师指导下进行的研究工作及取得的研究成果。论文中除了特别加以标注和致谢的地方外，不包含其他人或其它机构已经发表或撰写过的研究成果。其他同志对本研究的启发和所做的贡献均已在论文中作了明确的声明并表示了谢意。

作者签名：_____ 日期：_____

论文使用授权声明

本人完全了解复旦大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其它复制手段保存论文。保密的论文在解密后遵守此规定。

作者签名：_____ 导师签名：_____ 日期：_____