学校编码：10384　　分类号____密级_____
学号：24320111152294　　UDC ____

厦门大学

硕 士 学 位 论 文

基于情感和异源异构数据融合的潜在关系发现模型研究

Research of Latent Semantic Discovery Model Based on Sentiment and Heterogeneous Data Integration

张晓霞

指 导 教 师： 吴 清 强 副 教 授
专 业 名 称： 软 件 工 程
论文提交日期： 2014 年 4 月
论文答辩日期： 2014 年 5 月
学位授予日期： 2014 年 月

指 导 教 师：_____
答辩委员会主席：_____

2014 年 4 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于 年 月 日解密，解密后适用上述授权。

（ √ ）2.不保密，适用上述授权。

（请在以上相应括号内打“ √ ”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘要

生物医学数据的数量正在爆炸式地增长，如此海量的数据给医学科学家研究新药带来丰富的理论支持，但研究者们通宵达旦地阅读文献也不及其增长速度，更不用说抽取出隐藏在其中的信息。因此从生物医学数据中自动提取和分析信息的系统变得越来越重要。本论文对科学文献中生物本体间的情感关系表达、潜在关系抽取以及异源异构数据融合三个方面进行研究。

随着信息存储多样化的发展，从单一的数据源中抽取信息有时不能满足科研工作者的知识需求，因此要求异源异构数据能实现集成信息服务，达到跨异构库知识发现的目的。为了解决该问题，本论文研究了基于数据融合和基于结果融合的两种潜在语义分析模型，前者将预处理后的数据源进行集成为一个数据集，然后继续对其进行分析。后者先独立地分析各数据源最后将结果集成。本论文通过实例验证，验证了两种集成方法的可行性和有效性。

本论文利用基于图的半监督学习算法，即标注传递算法，来自动识别出生物实体之间的情感关系表达。目前，大部分研究都采用有监督学习方法，而且通常能取得较好的性能，但是基于有监督学习的关系抽取模型需要大量有标签的训练数据作为样本集，这将需要花费大量的人力和时间，降低效率。而标签传递算法把标签信息从图中的任意一个节点通过加权的各边循环地传递到附近的其他节点，最终达到全局稳定从而推导出未标签节点的标注信息数据，实现当训练数据不足时改善学习性能。

本论文利用基于上下文环境的 ABC 模型去发现潜在关系，该模型能够挖掘多层级实体的潜在关系，从而获得更全面的结果数据。而且本论文跨越传统的数据构建方法，不用疾病-药物之间的关系直接检索，而是采用非相关关系的数据集作为数据源，即疾病-基因、基因-药物之间的关系，从而能够更全面的分析出疾病与药物之间的非相关潜在关系。

关键词：潜在语义分析；情感分析；数据融合

Abstract

The number of biomedical data is growing explosively, such vast amounts of data brings abundant theoretical support for biomedical scientists researching new drugs, but even if they read the literatures day and night, they will not read all, let alone extract hidden information. So, the system of auto-extracting and analyzing information from biomedical data is more and more important. Meanwhile, with the development of biomedical study, the single data source can already not meet the increasing information needs so auto-discovery relationship model from heterogeneous data becomes very important in biomedical domain. The dissertation mainly studies emotional relationships between biological ontologies in biomedical literature, the potential relation extraction, as well as heterogeneous data integration.

With the number of information format stored increasing ,the information drawn from single data source has been already unable to meet the information needs of researchers , thus scientific databases and scientific literature are required to achieve data integration, to discovery knowledge across the heterogeneous database. The dissertation studies two latent semantic analysis models, namely the Latent Semantic Analysis model based on results integration, and Latent Semantic Analysis model based on data integration. The former first analyzes data source , then integrates all results .And the latter integrates intermediate results to a new data set, and then continues analysis . The experiment verifies the feasibility and effectiveness of the two methods.

The dissertation uses graph-based semi-supervised learning algorithm, label propagation method ,to automatically identify the relationship between biological entities. Extracting sentiment relationships between entities from the text automatically is an important direction in the field of text mining. Currently, supervised learning method is used in most of the studies, and usually performance nicely, but a large number of labels are required as sample set of training data, which will cost a lot of manpower and time, so that reducing efficiency. The label propagation method passes tag information from any node in the figure to other

neighboring nodes by weighted edge recurrently, eventually reaching global stability so as to deduce the information data on not label node. And it can improve learning performance when the training data is not enough.

In this dissertation, context -based ABC model is used to discover the multi-level potential relationship entities , and the non-correlation data sets, the relationship of disease-gene and gene-drug, is used as data source instead of traditional construction method, the relationship between disease-drug directly, to analyze more comprehensive potential relationship between disease and drug.

Keywords: Latent Semantic Analysis; Sentiment Analysis; Heterogeneous Data Integration;

目录

Contents

1 绪论

1.1 研究背景

互联网技术在二十一世纪得到迅猛发展，伴随着网络信息的集中式爆炸，人类所接触到的信息量呈指数飞速增长。浩如烟海的文献、资料及数据等信息从原有的书面、有声媒体等形式过渡到网络媒体的形式，人类也从信息量稀缺或者少有的年代过渡到信息量极其庞大的年代。因此面对着杂乱无章且飞速膨胀的海量数据信息，如何最快速并且最有效的筛选出有价值信息并进行自动分析，正是现代各领域实现快速发展的一个瓶颈。

生物医学数据的数量更是正在爆炸式地增长，尤其是科学文献和科学数据库数据。目前，在医学上可用的已有超过 60 万个临床实验报告、超过 150 万个临床案例报告。如此海量的数据给医学科学家研究新药带来丰富的理论支持，但研究者们通宵达旦阅读文献也不及其增长速度，更不用说抽取出隐藏在其中的信息所需要花费的时间和精力，因此从生物医学数据中自动地提取和分析信息的系统变得越来越重要。这些信息能够帮助研究者处理信息、系统地阐述生物模型、提出假设。

挖掘生物实体之间的潜在关系可以给科研工作者提供历史经验支持，更好地辅助其进行科学研究和科研管理。潜在关系主要包括生物实体对的关联强度和“方向性”，关联强度即是指两个生物实体之间的关系大小。挖掘文本中隐藏的潜在语义关系是近年来文本挖掘的一个重要研究方向。由于词的多样性以及语言习惯、人的知识背景的差异，常常同一概念会用不同的词汇来表达。同样，同一个词因语境、使用人的不同，可能表达了不同的含义。潜在语义分析[1]正是在这样的背景下被提出来的，是在现有的信息检索模型的基础上进行改进的。在生物医学文本数据挖掘中，该技术主要是挖掘出科学文献文本中的生物医学实体，然后计算它们之间存在的潜在关系并最终以简单直观的方式展示[2]。但是与许多其他自然科学数据比较，生物医学信息的两大特点—异质性（Heterogeneity）和数据难以量化—更是加大了该领域潜在关系抽取技术的难度。目前，生物医学领域应用较多的是基于共现理论发现隐式关系的挖掘技术。即一般认为，如果两个实

体在同一篇论文中同时出现，往往表明这两个实体之间具有一定的联系[3]。但是，这种方法只能发现有直接关系的生物实体对，而对于多个有间接交互关系的实体对关系发现则是一个挑战。

“方向性”即是指文本的情感关系表达，也就是对某事件文本表现出的是喜欢、中立还是反对。文本情感分析的主要目标是使计算机能够识别本文中人类要表达的情感，即需要建立完善的情感识别模型。国内的自然语言处理领域的研究中，实体的情感关系抽取是文本情感分析的一个重要应用领域。其中大部分研究都采用有监督学习[4] 或基于 Bootstrapping 的半监督学习[5]的抽取模型来抽取实体之间的关系。而基于有监督学习的关系抽取模型需要大量有标签的训练数据作为样本集，这将需要花费大量的人力和时间。基于 Bootstrapping 的半监督学习模型是基于局部一致性假设学习的，即未标签的样本只依据由有标签的样本数据进行训练得到的模型来分类[4]。该方法忽略了未标签样本之间的相似性和关联性，执行分类算法时也没有基于全局一致性，这将造成当训练数据不足及在推导类别边界时不能充分利用无标签数据的信息去挖掘隐藏在数据中的类结构特征。因此，充分运用训练数据获取大量未标签样本数据并达到全局一致性是文本情感分析中亟待解决的问题。

同时，随着科学发展的需要，越来越多的生物医学组织致力于药物发现研究。由于研究人员背景、知识和研究目的的多样性，在研究结果中包含了大量的生物医学实体及其之间的复杂关系数据，且这些研究成果以多种结构方式存储在不同的媒介中。因此，简单的统计分析单一结构的数据源信息已经不能满足生物医学科学家的知识需求。数据融合是对各种异构数据提供统一的表示、存储和管理，这些功能在异构数据融合系统中实现。数据融合屏蔽了各种异构数据间的差异，通过异构数据融合系统进行统一操作，集成后的异构数据对用户来说是统一的和无差异的。因此，异源异构数据融合技术正可以很好的解决这一问题，从多种数据源的异构数据中自动抽取各种关系模型已成为生物医学潜在关系挖掘领域中的重要方向。

生物医学领域涉及到基因序列、蛋白质结构等科学数据，并且仍在飞速增长的各种数据构成了科学数据库和科学文献。前者中的信息数据主要是生物学语言，它需要借助科学文献来解释。后者中的信息数据则比较通俗易懂。相关研究人员

除了关心基因序列的基本信息外，还希望获得能够与其相关的原理解释。这种希望推动了综合研究科学数据与科学文献数据的进程。但是目前的集成方式主要是以数据源融合方式进行数据融合，忽略了多数据源异构数据的集成对系统的多层次影响。

1.2 拟解决问题

生物医学领域知识存在结构复杂和更新速度快的特点。与医学研究相关的文献数据、药品数据、临床实验数据正在海量增长，这些信息和各大数据库数据信息以及互联网丰富的数据资源将会为疾病药物关系的发现和挖掘潜在关系研究提供理论支持，因此自动获取隐藏在其中的生物实体关系成为生物医学领域发展的又一新热点。

但是，目前针对生物医学领域的抽取生物实体中的潜在关系发现研究大多数都仅仅只挖掘单一方面的关系，比如只挖掘疾病-药物生物实体对之间的关系强度大小并没有考虑药物对疾病的情感关系表达，只分析单一结构的数据源数据而忽略了其它不同存储结构的数据信息等。因此，在辅助新药物研究方面，最终得到的信息并不能够全面的衡量生物实体之间的关系，也就不能够给生物医学研究者提供明确的研究方向，达不到减少实验盲目性的目的。所以为了能够给生物医学研究者提供更全面、更精确的生物实体之间的潜在关系做为知识支持，以缩短药物研发的时间，降低研发的成本，本论文将从情感关系表达、相似度两方面研究多数据源异构数据中的生物实体之间的潜在关系。

本论文的研究即是针对潜在语义分析技术在生物医学中的运用，目的是构建出一个基于生物医学潜在语义分析的用于发现药物与疾病潜在关系的模型，该模型可用于构建生物实体链接地图，为生物医学科学家研究疾病与药物之间的关系提供理论基础和实践支持。

综上所述，本论文拟解决的问题主要有：

（1） 如何分析生物医学文献中生物实体之间的潜在语义关系？

本论文拟采用文本挖掘技术挖掘出相关医学文献，再将目标科学文献中的语句进行分词，然后按照生物实体的类别进行概念映射，得到拆分表获取概念所对

应的语义类型，并根据共现理论统计分析出生物实体疾病-基因、基因-药物之间在上下文环境中的关系，进而通过计算相似度得出疾病与药物之间的潜在关系。

（2） 如何对生物医学领域文本进行情感语义分析？

本论文拟将目标科学文献中的语句进行分词，然后按照生物实体的类别概念映射，得到拆分表获取概念所对应的语义类型。用已知生物实体对之间的情感关系表达训练目标数据源中的概念间的正负相关性，从而挖掘出疾病-药物之间的潜在情感语义关系。

（3） 如何对医学科学文献数据与科学数据库数据进行数据融合？

本论文拟采用数据融合和结果融合两种数据融合方式对医学科学文献数据与科学数据库数据进行不同层次的融合，从而挖掘出更精确、更全面的疾病-药物的关系。

1.3 主要研究内容和意义

本论文主要对挖掘多数据源异构数据中的实体间潜在关系进行深入的学习，重点研究生物医学科学文献和科学数据库中生物实体之间的潜在关系强度和情感关系表达，且对其进行数据融合并对结果进行对比分析，构建一个大规模的药物-疾病治疗关系知识库。

本论文将从以下几个方面展开研究工作：

（1） 基于数据融合的潜在语义分析模型：研究基于数据融合的潜在语义分析模型，并按照数据准备、数据预处理、数据分析、数据结果处理流程阐述模型过程和作用。重点阐述基于非相关关系的数据集的构建、基于 LLR 层次归约算法的 C-value 分词技术、基于上下文环境的 ABC 模型的算法思路、数据融合方法模块。并以阿尔茨海默病（Alzheimer disease，AD）为例，采用 PubMed 科学文献集及 DrugBank 数据库数据为数据源，对基于数据融合的潜在语义分析模型进行实验验证，验证本论文方法的可行性和有效性。

（2） 基于结果融合的潜在语义分析模型：研究基于结果融合的潜在语义分析模型，并学习了数据融合、关系传递算法在模型中的应用。以阿尔茨海默病（Alzheimer disease，AD）为例，采用 PubMed 科学文献集及 DrugBank 数据库数据为数据源，对基于结果融合的潜在语义分析模型进行实验验证，验证本论文

方法的可行性和有效性。

综上所述，本论文的研究有以下三方面的意义：

（1）通过挖掘隐藏在生物医学科学文献和数据中的潜在、未知的疾病-药物关系为疾病的新治疗方案或药物新的疗效研究提供历史经验支持，更好地辅助科学家进行科学研究和科研管理；

（2）通过计算机辅助把药物实验的重点引导到最大可能的方向以减少实验的盲目性并减少药物实验的次数，缩短药物研发的时间，降低研发的成本以及提高成功的概率并获得更多的成果。

（3）通过分析归纳总结生物医学科学文献和数据中大量的成功、失败经验，为药物重定向的计算做一个前期筛选，去掉已经验证失败的方向，以减少药物重定向的计算规模，提高计算结果的有效性。

1.4 研究方法

针对论文的主要研究内容和拟解决问题，本论文拟采用文献研究法、专家咨询法、实证分析法等研究方法。

（1）文献研究法。通过网络信息、文献调查等方式收集和调研国内外相关的研究、相关课题的进展，从不同角度进行分析和总结，借鉴其先进经验和思想。梳理潜在语义分析方法、情感语义分析及数据融合的相关理论和方法，分析上述研究在生物医学上应用的可行性。

（2）专家咨询法。本论文的研究需要其他学科的数据作为研究对象，由于为跨学科的研究，所以有关生物医学领域的专业知识及本论文研究产生的生物实体之间的数据信息，拟咨询领域内的专家，判断信息的准确性，并对模型进行修改完善。

（3）实证分析法。选取一种疾病作为案例，使用实证法对模型进行验证及实验结果分析，并对其存在的问题进行改进和完善。

1.5 论文的结构安排

论文共分七章，各章内容如下：

第一章，主要介绍了本文的研究背景及研究意义，分析本论文拟解决的主要

问题，并阐述了本文的主要研究内容和研究方法。

第二章，首先详细介绍了本文研究内容所需要的相关技术及其理论基础。主要有分词技术、潜在关系分析技术、文本情感分析技术及多数据源异构数据融合技术。

第三章，介绍基于数据融合的潜在语义分析模型。按照实验流程分别详细阐述了数据准备、数据预处理、数据分析等过程。且详细介绍了本论文应用到的主要算法，主要包括基于非相关关系的数据集的构建、基于 LLR 层次归约算法的 C-value 分词技术、基于上下文体系的 ABC 模型、基于图的半监督学习算法、数据融合等，并通过实验验证了该模型的可行性。

第四章，进行实例验证基于数据融合的潜在语义分析模型。分析实验结果，验证该模型的可行性和有效性。

第五章，介绍基于结果融合的潜在语义分析模型。分别介绍了三种结果融合方式，并详细阐述了数据融合、关系传递的算法思想，并通过实验验证了该模型的可行性。

第六章，进行实例验证基于结果融合的潜在语义分析模型。分析实验结果，验证该模型的可行性和有效性。

第七章，对论文进行总结与展望。总结了本论文的主要工作及其优缺点，并针对本论文存在的不足对以后的改进工作进行了展望。