硕 士 学 位 论 文

基于文本挖掘的潜在药物不良反应发现

The Discovery of Potential Adverse Drug Reactions Based on Text Mining

作 者 姓 名：赵明珍

学科、 专业：计算机应用技术

学　　号：21209192

指 导 教 师：林鸿飞　教授

完 成 日 期：2015 年 05 月

大连理工大学

Dalian University of Technology

大连理工大学学位论文独创性声明

作者郑重声明：所呈交的学位论文，是本人在导师的指导下进行研究工作所取得的成果。尽我所知，除文中已经注明引用内容和致谢的地方外，本论文不包含其他个人或集体已经发表的研究成果，也不包含其他已申请学位或其他用途使用过的成果。与我一同工作的同志对本研究所做的贡献均已在论文中做了明确的说明并表示了谢意。

若有不实之处，本人愿意承担相关法律责任。

学位论文题目：_____

作 者 签 名 ：_____ 日期：_____年____月____日

摘　要

生物医学技术的发展为人类提供了大量的药品用以治疗各种疾病。一方面，药物对于治疗人类疾病，改善人类健康水平，延长人类寿命起到重要作用；另一方面，药物不良反应又使得药物在某些情况下对人类身体健康产生严重危害，而这些危害甚至可能是致命的。药物不良反应不仅对个人健康产生危害，而且给整个社会带来巨大的经济损失。如何及时高效而又全面地发现药物所导致的不良反应成为医学界和学术界所关注的热点。

药物流向市场之前必须进行综合的临床试验。但由于其局限性，临床试验不能保证揭露药物所有的不良反应。药物上市后，药物不良事件报告系统成为监控药物安全、发现药物不良反应的主要依托。随着 Web2.0 技术的发展以及互联网广泛的普及，健康社交网站积累了大量的用药者评论，这些用药者评论数据蕴含丰富的药物不良反应信息，为挖掘潜在药物不良反应、监控药物安全提供了新的数据源。

针对药物不良事件报告系统中的数据，本文利用非序列化 Skip-gram 模型，训练生成药物和不良反应的分布式实体向量，利用向量之间的运算来计算药物和不良反应之间的关联性达到挖掘潜在药物不良反应的目的。实验表明，非序列化 Skip-gram 模型生成的分布式实体向量，有效地捕捉了药物和不良反应之间的关联性，可以用于进行潜在药物不良反应的发现。

针对社交网络中的用药者评论数据，本文利用信息熵和字典匹配的方法从用户评论中挖掘药物不良反应。但是，从用户评论中挖掘到的药物不良反应是“潜在”意义的不良反应，尚未得到临床意义上的验证，而验证潜在药物不良反应的真伪性是耗时耗力的过程。因此，本文利用非序列化 Skip-gram 模型，生成生物医学实体的分布式向量，对于药物 d 和不良反应 a，利用所生成的分布式实体向量，尽最大努力发现药物 d 和不良反应 a 之间的关联蛋白质，为生物医学专家最终确定其真实性提供蛋白质级别的参考，从而缩短确定潜在不良反应真实性的时间，实现及时发现药物潜在风险的目标。

关键词：药物不良反应；非序列化 Skip-gram；分布式向量

The Discovery of Potential Adverse Drug Reactions Based on Text Mining

Abstract

With the development of biomedical technologies, a growing number of drugs flow into the market. On the one hand, drugs can treat human diseases, improve human health level and even extend human life. On the other hand, the adverse drug reactions make serious damages to human health in some conditions, and some of them even are fatal. Excepting causing damages to human bodies, adverse drug reactions can cause enormous economic loss. How to find adverse reactions for drugs timely and as many as possible has become the focus of attention of medical professionals and academic professionals.

Before exposing to the public, clinical trials must be designed for drugs to verify their effectiveness and security. However, due to some defects, clinical trials cannot find all adverse reactions of drugs. After flowing into market, adverse drug event reporting systems become the main means of monitoring drug safety and finding potential adverse drug reactions. With the development of technologies related to Web2.0 and the popularization of the Internet, health-related social network websites have collected a huge amount of user comments about drugs from patients. These user comments contain huge amount information about adverse drug reactions. Therefore, health-related social network websites provide another data source for potential adverse drug reaction discovery.

Towards the data from adverse event reporting systems, the paper trains non-sequenced Skip-gram model to generate the distributed vectors for drug entities and adverse reaction entities with these reports. With the entity vectors, the paper computes the relevancy among drugs and adverse reactions and then finds potential adverse drug reactions based on relevancy. The experiment results show that the distributed entity vectors capture the relevancy among drugs and adverse reactions effectively, and can be used to finding the potential drug reactions.

Towards the user comments from social networking websites, the paper finds mentions of adverse reactions from them based on method of information entropy and dictionary matching. However, the adverse reactions found from user comments are potential, not real, adverse drug reactions because they have not gotten clinical verifications. And getting clinical verifications for potential adverse drug reactions is a process requiring much time and effort. Therefore, the paper generates distributed biomedical entity vectors using non-sequenced Skip-gram model. For drug *d* and adverse reaction *a*, the paper tries best to find some proteins, called association-proteins, that can be associated with both drug *d* and adverse reaction *a*

based on the generated distributed biomedical entity vectors. The medical specialists can refer to the association-proteins when verify the reality of potential adverse drug reactions and then reduce the time of the getting clinical verifications, so that they can find the potential drug risks timely.

Key Words： Adverse Drug Reactions; Non-Sequenced Skip-gram; Distributed Vectors

目　　录

1 绪论

1.1 研究背景

药物不良反应（Adverse Drug Reactions, ADRs）是病人为了治疗所患疾病而服用正常剂量的药物时所出现的、与治疗无关且对病人身体健康有害的作用，其中包含生理上的不良反应和心理上的不良反应（如焦虑、狂躁等）。在人类的进化过程中，药物的产生对于治愈各种疾病，改善人类健康，延长人类寿命具有重要意义。然而，事物都具有两面性。药物在治愈人类各种疾病的同时，几乎所有的药物都会产生相应的不良反应，不同之处在于产生不良反应的条件、发生的频率、产生的种类以及相应的危害程度有所差异。

生物医学技术的发展为人类提供了大量的药品用以治疗各种疾病，但另一方面，药物不良反应的出现次数变得越来越多，其危害程度也越来越严重。药物不良反应不仅严重危害民众健康，对整个社会也造成了巨大的经济损失，一直以来是医学界和公众关注的重点之一。Giacomini 等人指出美国每年有超过 200 万的病人由于药物不良反应而住院的，其中大约 10 万人由于药物不良反应而丧生；紧跟癌症和心脏病之后，药物不良反应已成为美国的第四大人口死亡原因[1]。Leaman 等人指出美国每年因为药物不良反应造成的经济损失约为 1360 亿美元[2]。类似的，朱芹英指出，我国每年有 500 万人次的住院是药物不良反应造成的，更严重的是药物不良反应导致大约 19 万人失去生命[3]。其他国家也面临类似的处境，例如：药物不良反应成为瑞典的第七大人口死亡原因[4]。

药物不良反应的发现机制主要包括药物流向市场之前的临床试验和上市后的药物不良事件报告系统（Adverse Event Reporting Systems，AERS）。药物临床试验是药物流入市场之前必不可少的步骤，主要用于检测药物的有效性，发现药物较为严重的不良反应。临床试验中的受试者是经过严格选择和控制的，是在较小范围和特殊群体的人群中进行的药品评价。和整个人类群体相比，临床试验中的受试人员在数量和人群差异程度方面都具有严重不足，加之试验周期较短，导致药物所有的不良反应在临床试验阶段没有被完全揭露，使得存有大量潜在药物不良反应的药物流向市场，对公众健康产生严重威胁。因此我们迫切需要其他发现药物不良反应的手段来成为弥补临床试验的不足。

药物流向市场之后，AERS 系统成为监控药物安全、发现药物安全隐患的重要手段。药物不良反应和药物不良事件是不同的医学概念。相关文献指出[5]：药物不良反应是服用药物后出现的、对病人身体产生意外伤害的反应，其与所服用药物具有因果关系；而药物不良事件与病人所服用药物之间的因果关系尚未得到临床意义上的验证。著名的

AERS 系统主要包括美国食品药品监督管理局（Food and Drug Administration, FDA）下属的安全信息和不良事件报告系统 MedWatch、WHO 下属的不良反应监测系统（The Uppsala Monitoring Centre , UMC）以及欧洲药品管理局（European Medicines Agency，EMA）的不良反应检测系统。AERS 系统收集病人或者医务工作人员在发现药物安全问题时自发提交的药物不良事件报告，并利用特定技术手段对药物不良事件做出评估，从而发现潜在药物不良反应。一定程度上，AERS 系统弥补了临床试验在发现药物不良反应方面的不足。

随着 Web2.0 技术和 PC 的普及，微博、博客、论坛等社交媒体取得蓬勃发展，其中健康社交网络成为病人和医生的交流平台，如 AskAPatient、DailyStrength、MedHelp 等。迅速兴起的健康社交网站成为潜在药物不良反应发现的全新依托平台。在健康社交网络中，用药者可以发表对于某种药物的评论，从而分享自己的用药经历；病人也可以向医生提出自己的疑问，从而得到医生或者具有相似经历的病人的回答。随着时间的推移，健康社交网络中积累了大量的、来自病人的、与药物相关的文本数据，这些文本数据是来自病人的第一手资料，蕴含着丰富的药物不良反应信息。如果能对其进行充分的挖掘，对于及时发现药物不良反应、丰富药品不良反应知识库具有重要的意义。

一方面，健康社交网站积累了大量的药物评论数据，为潜在药物不良反应发现提供了新的平台；另一方面，健康社交网站中的用户评论来自普通互联网用户，是非结构化的文本数据，给不良反应挖掘工作带来巨大挑战。值得庆幸的是，文本挖掘的相关技术为处理上述问题提供了可能性。

随着机器学习和自然语言处理技术的发展，文本挖掘技术在生物医学领域被成功用于解决命名实体识别（Named Entity Recognition，NER）、关系抽取（Relation Extraction）、事件抽取（Event Extraction）等问题，因此生物医学领域出现了很多可用的工具，如 MetaMap 等。另一方面，生物医学领域积累了多种可用的数据本体资源，如 SIDER，为科研人员从事生物医学领域的研究提供了便利。因此，文本挖掘技术和各种生物医学资源使得针对社交网络的潜在药物不良反应发现成为可能。

在本文中，将文本挖掘算法发现的、尚未得到临床医学上验证的药物不良反应称为潜在药物不良反应，其作为最为可能的药物不良反应推荐给医务工作者，为医务工作者及时发现药物的安全隐患并提出预警，改善公众健康做出贡献。

1.2 研究现状

由于其严重危害性，药物不良反应得到各界愈来愈多的关注。除了临床试验等专业手段外，数据挖掘的相关方法也被广泛应用于药物不良反应的发现过程之中，其主要用于药物上市后的阶段。药物上市后，药物不良事件报告系统负责监控药物使用的安全状况，并利用数据挖掘和文本挖掘的相关方法及时的挖掘药物的安全隐患，发现药物的潜在不良反应。近年来，健康社交网站积累了大量来自用药者的药物评论数据，这些评论数据逐渐引起学术界的关注，为潜在不良反应发现提供了新的契机，基于用户评论的潜在药物不良反应挖掘也取得了一定程度的进展。

1.2.1 基于不良事件报告系统的研究现状

AERS 系统收集来自病人或者医生提交的药物不良事件报告。目前，基于药物不良事件报告系统的不良反应发现，大多利用关联规则和统计学的方法，对某种（些）药物和某种（些）不良反应进行安全评估。陈俊玲等人从广东省药品不良反应检测中心获取关于头孢曲松钠 ADR 的 12210 份报告，在经过数据去噪等预处理后，采用关联规则算法，挖掘头孢曲松钠的不良反应与病人的身体特征（如性别、年龄）以及用药途径（如口服、静脉注射）等属性之间的关联性，实验表明，头孢曲松钠的毒副作用与性别、影响系统等因素相关性显著[6]。冯变玲等人采用关联规则的方法从江苏省 2004 年至 2009 年期间的 9640 份心血管疾病患者的药物不良反应报告中挖掘药物、不良反应和用药人群之间的关联关系，共获得 12 个三元组关系，其中 5 个三元组的关联度是一般性的，而另外 7 个三元组的关联性较强[7]。Harpaz 等人利用关联规则算法，以 2008 年 FDA 的自发报告系统在 2008 年所接收到的不良事件报告作为数据集，挖掘由于药物之间相互作用而导致的不良反应[8]。蒋朋利等人利用贝叶斯区间递进神经网络（Bayesian Confidence Neural Network ，BCPNN）、成比例报告比值（Proportional ADR Reporting Ratio，PRR）、伽玛泊松缩减法（Gamma Poisson Shirnker，GPS）等方法从 FDA 不良反应自发报告系统中挖掘他汀类药物与糖尿病、癌症和认知能力下降之间的统计学关系[9]。魏建香等人利用互信息计算药物和不良反应之间联系的紧密度，从 2008 年江苏省药物不良反应数据库中研究 11591 种药物和不良反应组合之间的安全隐患，实验结果表明互信息可以用于药物不良反应检测[10]。Santiago 等人[11]基于药物分子结构相近则不良反应相似的假设，利用药物分子之间的相似性进行药物不良反应推断，并对可能导致横纹肌溶解的药物进行了预测。

但是，以上研究存在几点不足：（1）实验数据相对较少或者数据来自于某一特定区域，病例多样性不足，无法充分揭露潜在药物不良反应；（2）对于潜在药物不良反

应的挖掘主要集中于某种（类）药物与某种（些）不良反应之间的关联挖掘，研究对象相对单一，不能充分利用药物和不良反应之间的共现信息有效的捕捉二者之间的关联度。

1.2.2 基于社交网络的研究现状

随着互联网技术的发展和普及，健康社交网络中积累了大量来自病人的药物评论数据，这些用户评论以文本的形式存在，大多是对公众开放的，并且蕴含丰富的药物不良反应信息，引起医务工作者和科研人员越来越多的关注。

Leman 等人[2]以健康社交网站 DailyStrength 中的药物评论作为实验数据，采用滑动窗口和字典匹配相结合的方法识别评论中出现的“疾病和不良反应”名称，并采用启发式的手段过滤药物适应症，从而识别出用户评论中不良反应名称，实验表明，社交网络中的用户评论数据蕴含丰富的药品安全信息。Leaman 成为最早从社交网络中进行药物不良反应挖掘的学者之一。在 Leaman 的基础上，Nikfarjam 等人[12]从标注数据集中提取用户表达不良反应的语言模式，在这些语言模式的基础上利用关联规则算法提取用户评论中的药物不良反应名称，相较于 Leaman 的结果，其实验结果略有下降，优点是无需构建不良反应词典。Yates 等人[13]从 askapatient.com、drugs.com 和 drugratingz.com 三个社交网站中抓取了 5 种乳腺癌药物的 2500 条用户评论，并对其中的 250 条用户评论进行标注，在此基础上构建 ADRTrace 系统，用于提取不良反应名称，但该系统训练数据集相对较少，并且只考虑了关于乳腺癌的药物，泛化性较差。Bian 等人[14]以 Twitter 作为数据源，首先利用支持向量机（SVM）模型作为分类器，从 Twitter 用户中识别用药者；然后再次利用支持向量机构建分类模型识别不良反应名称，但由于 Twitter 面向通用领域，不是专门的健康社交网站，所以噪音很多，对分类器的影响很大。

虽然上述研究可以从健康社交网络中的用户评论中挖掘药物的不良反应信息，但是所挖掘的药物不良反应尚未得到医学意义上的验证，仍然是潜在意义上的药物不良反应。然而，验证药物不良反应的真实性需要进行大量的临床实验和医学分析，将会耗费大量的人力物力资源，并对药物安全隐患发现的及时性产生严重的影响。如果可以利用文本挖掘的相关技术，发现药物和其潜在不良反应之间的作用机制，从而发现潜在药物不良反应发生的“证据”，将对减少验证潜在不良反应真实性所需的时间、实现及时发现药物的安全隐患产生重大意义。

1.3 本文工作

目前，相关部门和组织主要依托 AERS 系统发现上市药物安全隐患。相关研究的主流手段是利用统计学的相关方法，计算某种（些）药物与某种（些）不良反应之间的统计关联性，从而达到药品安全隐患的检测。然而这些方法的分析对象过于片面，无法从全局出发整体考虑所有药物和所有不良反应的分布情况。受分布式词向量（Distributed Representation）的启发，本文利用语言模型从不良事件报告集中训练生成药物和不良反应的分布式实体向量，由于这些实体向量是从全局中生成的，能够捕捉实体之间的共现信息，可以很好的衡量实体之间的关联度。本文通过向量之间的运算，衡量药物和不良反应之间的关联度，达到挖掘药物的潜在不良反应的目的。实验结果证实基于分布式实体向量的方法可以在 AERS 系统中挖掘药物的潜在不良反应。

在信息技术取得长足发展以及互联网和移动互联网大规模普及的今天，健康社交网站得到越来越多的关注。健康社交网站中积累了大量来自病人对药物的评论数据。这些药物评论是来自病人的第一手资料，蕴含了丰富的药品不良反应信息。本文采用基于信息熵和词典匹配的方法，从用户评论中识别不良反应名称，并依托 SIDER、Semantic MEDLINE 等多种生物医学资源过滤药物适应症和已知的药物不良反应，从而得到潜在不良反应列表。所存在的问题是，得到的潜在药物不良反应尚未得到临床意义上的验证，然而验证其真实性是耗时耗力的过程。本文从 MEDLINE 数据集中，采用语言模型训练生成生物医学实体的分布式向量，对于药物 d 和不良反应 a，尽最大努力挖掘药物 d 和不良反应 a 之间的关联蛋白质，为生物医学专家最终确定其真实性提供蛋白质级别的参考，从而缩短确定潜在不良反应真实性的时间，达到及时发现药物的潜在风险的目的。实验证明，本文所述的方法可以从社交网站中挖掘药物的潜在不良反应，并能够挖掘药物和不良反应之间的关联蛋白质，从而为潜在不良反应找寻“证据”，减少验证其真实性所用的时间，实现及时发现药物安全隐患的目的。

1.4 本文结构

本文使用 4 个章节来阐述基于文本挖掘的潜在药物不良反应发现过程及实验结果。具体安排如下：

第一章，绪论。绪论部分详细阐述了潜在不良反应发现的研究背景，指出不良反应对公众健康和社会经济造成的损害。绪论部分还给出了目前上市药物的安全监控主要依托于药物不良事件报告系统，并指出健康社交论坛成为药品安全监管的新的依托，并对相关研究做了总结；最后，绪论部分给出了本文的工作安排和本文的组织结构。

第二章，相关资源和算法。相关资源和工具主要描述了本文中使用的生物医学数据资源以及所采用的算法。其中，相关资源包括相关的生物医学数据资源和生物医学工具。

第三章，面向不良事件报告的潜在药物不良反应发现，详细讲述了利用词向量的相关技术从不良事件报告中挖掘药物的潜在不良反应的过程，并给出了实验分析结果。

第四章，面向社交网络的潜在药物不良反应发现，详细阐述了从健康社交网站的用户评论中挖掘药物的不良反应信息，并为所挖掘的潜在不良反应信息寻找蛋白质级别的“证据”，最后通过案例验证的方法，给出了关联蛋白质挖掘的结果。

最后给出了全文的总结，总结了本文的研究进展，并对下一步研究工作提出展望和建议。

2 相关资源和算法

2.1 生物医学数据资源

2.1.1 SIDER

SIDER[15]（Side Effect Resource）是常用的药物不良反应数据库，记录了多种药物的不良反应信息，目前的版本为 SIDER2。SIDER2 总共收集了 996 种上市药物和 4192 种不良反应，并从药品说明书以及公开文档中提取了 99423 个（药物，不良反应）对信息。在研究过程中，可以在 SIDER 数据库中查询药物已知的药物不良反应列表，从而实现过滤已知药物不良反应、发现尚未发现的潜在药物不良反应的目的。

在 SIDER 数据库中，不良反应名称已经映射到 UMLS（Unified Medical Language System）的超级叙词表中。其中 UMLS 是美国国立医学图书馆 NLM（The United States National Library of Medicine）开发的一体化医学语言系统。在 UMLS 的超级叙词表中，每个概念都拥有一个唯一的 CUI（Concept Unique Identifier）编号。表 2.1 是 SIDER 数据库中关于药物 trazodone 的部分不良反应记录。

表 2.1 SIDER 中 trazodone 的部分不良反应记录

Tab. 2.1 Some ADRs of trazodone in SIDER

药名	不良反应名称	不良反应 CUI
trazodone	back pain	C0004604
trazodone	diplopia	C0012569
trazodone	psoriasis	C0033860
trazodone	weight gain	C0043094

SIDER 数据库是公开的，科研人员可以从 http://sideeffects.embl.de/download/ 下载其所有的数据文件，在本地进行访问。

2.1.2 Semantic MEDLINE Database

Semantic MEDLINE Database（SemMedDB）[16-18]是生物医学语义关系提取工具 SemRep[19]从 MEDLINE 数据库中抽取的（subject，predicate，object）三元组数据库，其中 subject 和 object 分别表示关系中的主体和客体，predicate 表示主体和客体之间的关系。如三元组（ziprasidone, CAUSES, drowsiness）表示药物 ziprasidone 引起不良反应 drowsiness。

MEDLINE 是美国国立医学图书馆 NLM 创建和维护的全球最大的生物医学文献数据库，收录了生物医学领域 2160 多万篇文献引用，每篇论文引用都包含论文题目、论文摘要、论文的发表日期以及 MeSH 词等信息。MEDLINE 对外是免费可用的，通过在线检索系统 PubMed 可以快速便捷的对 MEDLINE 数据库进行检索和访问。SemRep 是生物医学领域内的关系抽取工具，可以从生物医学文本中识别生物医学实体，并提取这些识别之间的语义关系，以（subject，predicate，object）三元组的形式返回。

SemMedDB 数据库收录了 SemRep 从 MEDLINE 数据库中提取的所有（subject，predicate，object）三元语义关系，截止到 2014 年 4 月，SemMedDB 数据库共包含大约 7 千万个上述语义关系。

2.1.3 MeSH

MeSH[20]（Medical Subject Headings，医学主题词）是美国国立医学图书馆 NLM 创建和持续更新的综合受控词汇表，被广泛应用于生物医学文献的索引、分类和检索。对于 MEDLINE/PubMed 收录的每篇生物医学文献，领域专家都会根据其内容，选择一组可以有效的反应文章的主题的 MeSH 词赋予该文献，在此基础上实现 MEDLINE 数据库的高效索引。MeSH 也被临床试验注册中心（ClinicalTrials.gov）用于疾病分类，在 ClinicalTrials.gov 中，每个临床试验被赋予两组 MeSH 词作为其关键字，一组 MeSH 词表示该临床试验的条件，另一组 MeSH 词表示临床试验中的认为的干预。

表 2.2 MeSH 叙述词 Trazodone 的树编码结构之一

Tab. 2.2 The one of the tree numbers for Trazodone in MeSH

MeSH 叙述词	树编码
Chemical and Drugs	D
Heterocyclic Compounds	D03
Heterocyclic Compounds, 1-Ring	D03.383
Pyridines	D03.383.725
Pyridones	D03.383.725.791
Trazodone	D03.383.725.791.900

在 MeSH 中，主题词（subject headings）又称为叙述词（descriptors）。大多叙述词记录包含简短的定义描述、相关叙述词的链接以及同义词（entry term：款目词）列表。上述额外信息以及分层的结构使 MeSH 词成为一个分类汇编词典，而不是简单的主题词列表。MeSH 中的叙述词是按照分层结构进行组织的，每个叙述词在分层树状结构中可

能出现多次。在分层树状结构中，每个位置都具有一个树编码（tree numbers），代表叙述词的语义标签。由于叙述词可能出现多次，所以叙述词可能拥有多个树编码，例如药物 trazodone 包含两个树编码： D03.383.606.900 和 D03.383.725.791.900。表 2.2 给出了 trazodone 第二个树编码的树形编码结构。

除了叙述词，MeSH 还包括一些标准的限定词（qualifiers），用于对叙述词进行限定。例如，对于主题词 Measles 和限定词 epidemiology，Measles/epidemiology 实现主题词 Measles 范围的限定。此外，MeSH 还包括众多的补充概念记录（Supplementary Concept Records），这些补充概念记录不属于受控词汇表，但是对于每个补充概念，都给出了与其最相近的叙述词，用于 MEDLINE 数据库的检索。

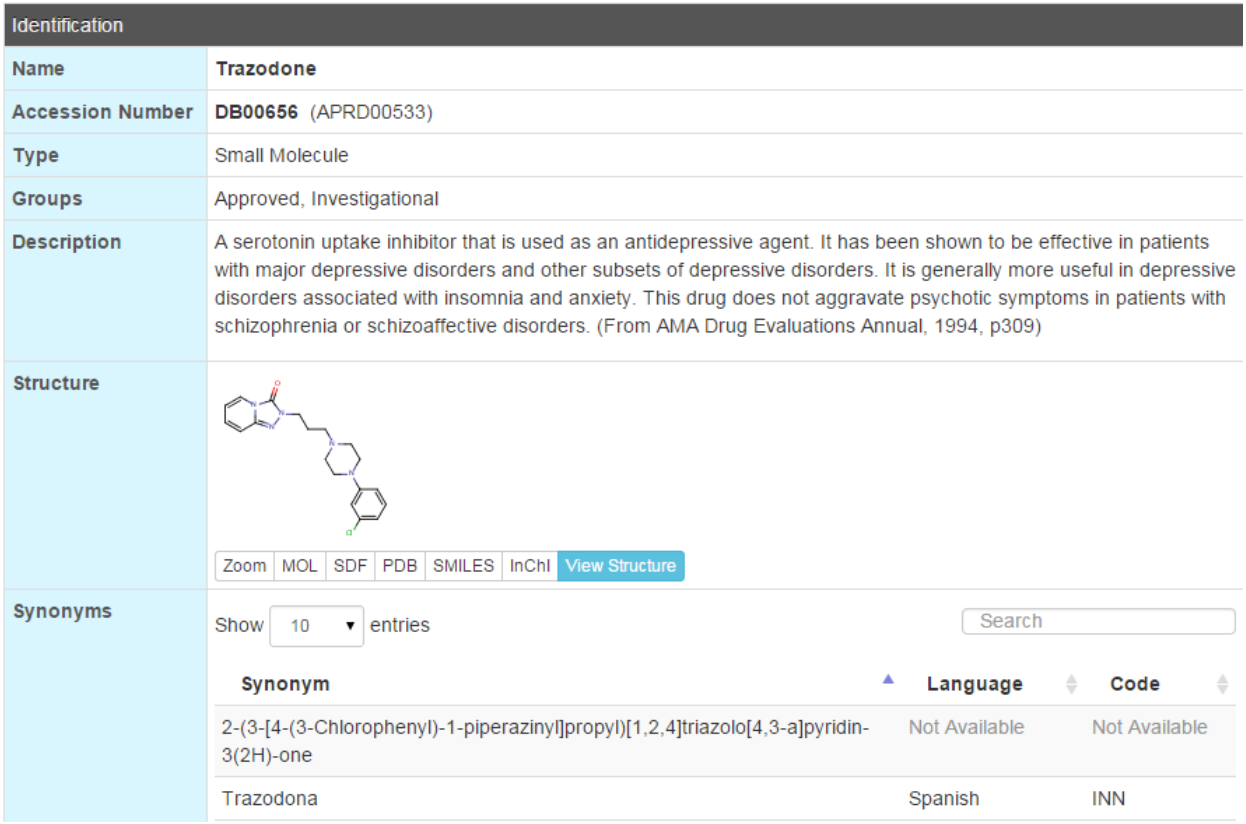
MeSH 提供了在线检索功能，用户可以通过浏览器检索某一 MeSH 词详细信息。此外，MeSH 还提供了 XML 格式的下载文件:，方便用户在本地对 MeSH 词进行检索和访问。

2.1.4 DrugBank

DrugBank[21-24]数据库是由 Alberta 大学创建和维护的、世界上最全的生物信息和化学信息数据库，记录了详细的药物信息和药物靶点信息（序列、结构和通路）。DrugBank 数据库持续更新，目前的版本为 DrugBank Version 4.2，收录了 7759 种药物，其中包括 1600 种 FDA 批准的小分子药物，160 种 FDA 批准的生物技术药物（蛋白质/缩氨酸）、89 种保健用品以及 6000 种实验药物。此外，DrugBank 还将 4282 种非冗余蛋白质序列（药物靶点、酶、载体等）与药物进行关联。在 DrugBank 中，每种药物包含 200 多个维度的信息，如药物名称、药物描述、分子结构、适应症、作用机制、药物之间相互作用等。

由于其范围广泛、引用全面以及描述信息详细等特点，使得 DrugBank 更像一个药物百科全书而不仅仅是药物数据库。因此，DrugBank 被药品制造商、药物化学家、内科医生、药剂师以及公众广泛应用。而且，DrugBank 详实的药物和药物靶点信息使得利用大量现存的药物治疗罕见或者新出现的疾病（药物再利用）成为可能，节省了设计新药物所需的成本，具有重大的现实意义。

DrugBank 数据库对外开放，用户可以通过 DrugBank 提供的在线检索系统访问该数据库，图 2.1 展示了在线检索药物 Trazodone 返回的部分信息，包含药名、药物编号、药物描述以及分子结构等信息。除此之外，DrugBank 还提供了 XML 格式的下载文件（http://www.drugbank.ca/downloads），该 XML 文件包含 DrugBank 数据库的所有数据，用户可以下载该文件到本地，使用 XML 解析工具对其进行检索和访问。



Identification	
Name	Trazodone
Accession Number	DB00656 (APRD00533)
Type	Small Molecule
Groups	Approved, Investigational
Description	A serotonin uptake inhibitor that is used as an antidepressive agent. It has been shown to be effective in patients with major depressive disorders and other subsets of depressive disorders. It is generally more useful in depressive disorders associated with insomnia and anxiety. This drug does not aggravate psychotic symptoms in patients with schizophrenia or schizoaffective disorders. (From AMA Drug Evaluations Annual, 1994, p309)
Structure	Zoom MOL SDF PDB SMILES InChI View Structure
Synonyms	Show 10 entries Search

Synonym	Language	Code
2-(3-[4-(3-Chlorophenyl)-1-piperazinyl]propyl)[1,2,4]triazolo[4,3-a]pyridin-3(2H)-one	Not Available	Not Available
Trazodona	Spanish	INN

图 2.1　Trazodone 在 DrugBank 中的部分信息

Fig. 2.1　Some information about Trazodone in DrugBank

2.2　生物医学工具

MetaMap[25]是由美国国家医学图书馆（National Library of Medicine, NLM）开发的、配置灵活的生物医学实体识别工具，该工具的主要功能是将生物医学文本映射到一体化医学语言系统（Unified Medical Language System, UMLS）中的超级叙词表（Metathesaurus）中。MetaMap 在生物医学文本挖掘中得到广泛应用。MetaMap 在识别文本中的生物医学实体时，会给出映射得分、实体 CUI 以及实体的语义类型等信息。在 MetaMap 的结果基础上，过滤映射得分较低的实体，可以实现高准确率的生物医学实体识别；通过实体语义类型，可以有效的选择感兴趣的生物医学实体（如药物、疾病或症状）；同一生物医学实体可能有多重表达形式，通过实体的 CUI，可以实现生物医学实体的标准化。

图 2.2 展示了 MetaMap 对于短语“head pain”处理后的结果。从中可以看出，MetaMap 从“head pain”中识别出 8 个候选概念（candidates），对于每个候选概念，MetaMap

给出了相似性得分、概念 CUI、概念名称（和通用名称）以及概念的语义类型。例如：对于得分最高的候选概念，其相似性得分为 1000，概念 CUI 为 C0018681，概念名称为“Head Pain”（通用名为 Headache），语义类型为“Sign or Symptom”。最后，MetaMap 将相似性得分最高的概念作为识别结果。

```
|: head pain
|:
Established connection to Tagger Server on localhost.
Processing 00000000.tx.1: head pain

Phrase: "head pain"
Meta Candidates (8):
  1000 C0018681:Head Pain (Headache) [Sign or Symptom]
   861 C0030193:Pain [Sign or Symptom]
   861 C1962977:Pain NOS (Pain NOS Adverse Event) [Finding]
   789 C0234226:Painless [Functional Concept]
   694 C0018670:Head [Body Location or Region]
   694 C1281590:Head (Entire head) [Body Part, Organ, or Organ Component]
   661 C1706305:HEADS (Head - Component of Device) [Medical Device]
   638 C0205096:Cephalic [Spatial Concept]
Meta Mapping (1000):
  1000 C0018681:Head Pain (Headache) [Sign or Symptom]
```

图 2.2　MetaMap 对 head pain 识别的结果

Fig. 2.2　The result of MetaMap output for head pain

2.3　相关算法

2.3.1　非序列化 Skip-gram 模型

Skip-gram[26]模型是 Google 公司在 2013 年发布的开源工具 word2vec[26-28]中的语言模型，用于从大规模非标注的文本数据中学习单词的分布式表述词向量（Distributional Representation，以下简称词向量），从而将单词映射到高维空间中。在单词到高维向量空间映射的基础上，文本数据上的计算可以转化为向量之间的运算，从而降低文本处理的复杂度。由于 Skip-gram 的出色表现，其生成的词向量被广泛用于聚类、同义词发现和词性分析等多种自然语言处理任务。

Skip-gram 模型生成的词向量捕捉了词语之间的语义相似性和句法相似性，而且所生成的词向量还具有语义的线性可加性，例如：向量 $v(\text{"}Russia\text{"})+v(\text{"}river\text{"})$ 与向量 $v(\text{"}Volga\ River\text{"})$ 很接近，向量 $v(\text{"}Germany\text{"})+v(\text{"}capital\text{"})$ 与向量 $v(\text{"}Berlin\text{"})$ 很接近，其中 $v(w)$ 表示 Skip-gram 生成的单词 w 的词向量。

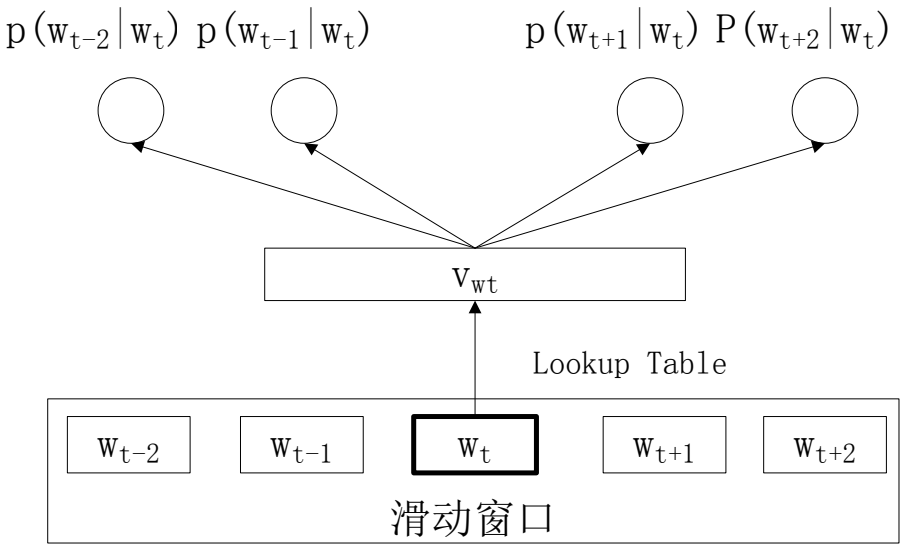


图 2.3　Skip-gram 模型架构图

Fig. 2.3　The Skip-gram model architecture

Skip-gram 模型采用滑动窗口的机制来获取词语之间的共现信息，其基本思路是在当前单词出现的情况下，其相邻单词出现的概率尽可能的大。即假定当前单词为 w_t，滑动窗口的半径为 c，该单词的滑动窗口为 $(w_{t-c},\cdots,w_t,w_t,w_{t+1},\cdots,w_{t+c})$，其目的是使得生成的词向量满足：

$$\max \sum_{-c\le j\le c, j\neq 0} \log p(w_{t+j} \mid w_t) \tag{2.1}$$

其中，条件概率采用 *softmax* 函数来定义，具体的：

$$p(w_b \mid w_a) = \frac{\exp(v_{w_b}^T \cdot v_{w_a})}{\sum_w \exp(v_w^T \cdot v_{w_a})} \tag{2.2}$$

其中，v_w 表示单词 w 的词向量。图 2.3 展示了滑动窗口半径为 2 时的 Skip-gram 模型的架构图，图中为当前单词 w_t 为，“Lookup Table”表示获取单词 w_t 的词向量 v_{w_t}。

假定训练数据为 $w_1, w_2, \cdots, w_T$，Skip-gram 最终的优化目标函数为：

$$\max \frac{1}{T} \sum_{t=1}^{T} \sum_{-c\le j\le c, j\neq 0} \log p(w_{t+j} \mid w_t) \tag{2.3}$$

自然语言是符合语法规则的、有序的单词序列，为了从有序的单词序列中获取词语的语义信息和句法信息，Skip-gram 采用滑动窗口的机制训练分布式词向量，为了便于区分，本文称以滑动窗口为训练机制的 Skip-gram 模型为有序 Skip-gram 模型。

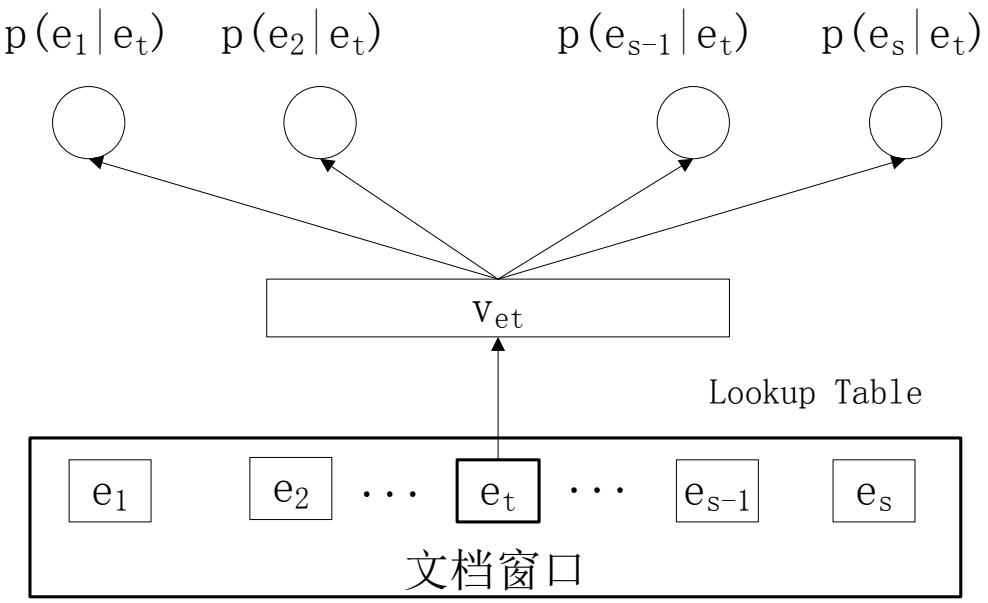


图 2.4　非序列化 Skip-gram 模型架构图

Fig. 2.4　The non-sequenced skip-gram model architecture

在本文任务中，每篇文档通常包含多个生物医学实体，但这些生物医学实体之间是无序的，彼此之间没有逻辑的前后关系。为了从文档集中训练生物医学实体的分布式词向量，本文采用文档窗口代替滑动窗口来训练 Skip-gram 模型。本文称采用文档窗口作为训练机制的 Skip-gram 模型为非序列化 Skip-gram 模型。图 2.4 是非序列化 Skip-gram 模型的架构图。

令文档集合为S，第m篇文档为r_m，文档r_m包含的实体集为E_m，所有实体组成的集合为V。非序列化 Skip-gram 模型的目标函数为：

$$\max \sum_{r_m \in S} \sum_{e_i \in E_m} \sum_{e_j \in E_m, e_i \neq e_j} \log p(e_j \mid e_i) \tag{2.4}$$

其中，条件概率为：

$$p(e_j \mid e_i) = \frac{\exp(v_{e_j}^T \cdot v_{v_{e_i}})}{\sum_{e \in V} \exp(v_e^T \cdot v_{v_{e_i}})} \tag{2.5}$$

2.3.2　信息熵

信息熵（Information Entropy），也被称为熵、平均信息量，是由信息论的创始人香农提出的，用来解决信息的量化问题。“信息量”是个非常抽象的概念。对于一件非常不确定的事情，我们需要大量的信息才能将其弄清楚。因此信息量与事件的不确定性具

有直接关系。由于信息熵可以对信息进行量化，所以在信息论和概率统计中，信息熵通常用来衡量事件的不确定性程度。如果随机变量的信息熵越大，则随机变量的不确定性就越大；反之，随机变量的不确定性越小。

设随机变量 X 是一个取有限值的离散变量，其概率分布为：

$$p(X=x_i)=p_i, \quad i=1,2,\cdots,n \tag{2.6}$$

则，随机变量 X 的信息熵为：

$$H(X)=-\sum_{i=1}^{n} p_i \log_2 p_i \tag{2.7}$$

并且，信息熵满足：$0 \le H(X) \le \log_2 n$。

信息熵在多个领域得到广泛应用，如自然语言处理、风险评估和图像处理。在自然语言处理领域，信息熵常被用于特征选择、新词发现、关键词抽取等任务，取得很好的效果。

2.4 本章小结

本章主要详细描述了本篇论文中使用的相关生物医学数据资源、生物医学工具以及所使用的文本挖掘算法。对于生物医学资源，本章着重介绍了 SIDER、Semantic MEDLINE Database、DrugBank 等知名的数据库，并对其数据格式和使用方法进行了相应的介绍。对于生物医学工具，本章重点介绍了生物医学文本映射工具 MetaMap，并对其输出结果各个字段的含义进行了相应的描述。对于文本挖掘算法，本文着重描述了 Skip-gram 模型、信息熵和点互信息。Skip-gram 模型是语言模型，用于生成单词的分布式词向量，利用滑动窗口的机制从文本数据中训练生成词向量。本文对 Skip-gram 模型的训练机制进行了适当改进，以文档窗口代替滑动窗口来捕捉生物医学实体的相似性，以便满足本文任务的需求。

3 面向不良事件报告的潜在药物不良反应发现

3.1 问题引出

药物临床试验是药物上市前必不可少的步骤，主要用于检测药物对其所治疗的疾病的作用，以及检测药物的安全性。然而由于试验周期短、试验病例少、使用场景单一等缺点，临床试验无法揭露药物所有可能的不良反应，使得具有潜在不良反应的药物流向市场。具有潜在不良反应的药物是医疗卫生的重大安全隐患，不仅对公众健康产生重大威胁，也对社会造成巨大的经济损失。如何监控药物的使用情况并发现药物潜在的安全隐患成为药品制造商、政府部门以及科研机构所关注的焦点之一。

药物流向市场之后，药物不良事件报告系统成为监控药物使用情况，及时发现药物不良反应，保证公众身体健康的重要平台。药物不良事件报告系统收集来自医务工作者以及病人提交的不良事件报告，利用统计学的方法计算药物和不良反应之间的关联性，从而对药物的安全隐患提出预警。在计算药物 d 和不良反应 a 之间的关联性时，大多现有的统计方法只考虑药物 d 、不良反应 a 以及两者之间的分布情况，而没有从全局出发整体考虑所有药物和所有不良反应的分布情况。

本章节在语言模型的基础上，训练生成药物和不良反应的实体向量，据此计算药物和不良反应之间的关联性。语言模型是一种用来计算一个句子合法性和合理性的概率模型。作为一种语言模型的副产品，词向量能够从训练数据集中获取词语的语义和句法信息，常用于计算词语之间的相似性，得到了广泛应用。本质上，词向量的产生是基于词语的共现信息生成的。考虑到在语言模型中的出色表现，本文利用非序列化 Skip-gram 模型，生成药物和不良反应的实体向量，依此计算药物和不良反应之间的关联性，实现潜在药物不良反应发现的目的。

3.2 实验数据

在美国“大数据国家战略”的背景下，美国食品药品监督管理局 FDA 的公共数据开放项目 OpenFDA 逐步向社会开放医疗和健康大数据，其先导项目开放了“381 万份药物不良反应事件报告”。这些报告是 FDA 的自发报告系统在 2004 年至 2013 年期间收集的，主要来自于医护人员或者用药者。

OpenFDA 所提供的药物不良事件报告是一种半结构化的数据，并且向外界提供 API 以便检索和下载这一数据集。截止到 2014 年 11 月，这一公开数据集包含 3,814,223 份报告。每份药物不良事件报告报告都包含病人的基本信息（如性别、年龄、国家）、病人的用药列表以及病人用药后出现的不良反应列表。这些药物和不良反应之

间的因果关系尚未得到临床意义上的验证，是潜在意义上的药物不良反应。 本文利用这一数据集中的用药信息和用药后的不良反应信息，采用非序列化 Skip-gram 模型，生成药物和不良反应实体向量，进行潜在药物不良反应发现。

3.3 研究框架

药物不良事件报告中包含病人的用药列表以及不良反应列表，但是这些药物和不良反应之间的因果关系尚未得到医学上的验证，只是作为可疑药物不良反应而存在。如果对这些可疑药物不良反应进行医学上的一一验证是不切实际的。因此，本文采用文本挖掘的技术从大量的药物不良事件报告中挖掘最有可能的潜在的药物不良反应，从而使医学专家有针对性地验证药物不良反应，缩短确认药物不良反应的时间。图 3.1 是系统流程图，整个系统主要包括不良报告抓取、药名实体识别及标准化、生成分布式实体向量和计算关联度等模块。

药名实体识别及标准化包括药名文本预处理、实体识别和非药名实体过滤三部分。药名文本预处理主要是对非格式化的药名文本做预处理，降低噪音对实验结果的影响；实体识别用于识别药名文本中的生物医学实体，其中大部分是药名实体，并实现实体标准化；非药名实体过滤用于删除非药物实体；计算关联度模块用于计算药物和不良反应之间的关联度，从而对不良反应进行排序，获取最有可能的潜在药物不良反应。

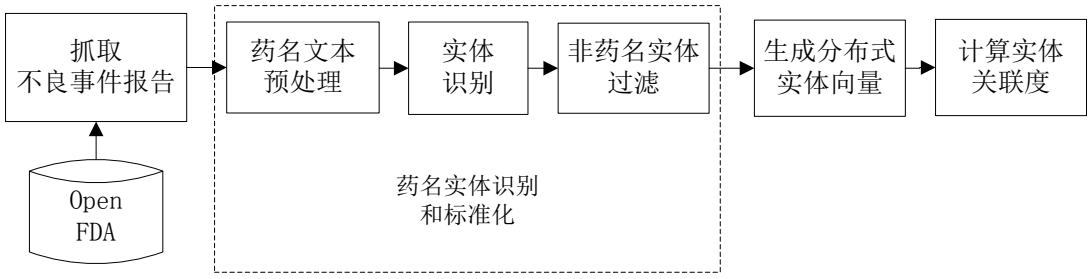


图 3.1 系统流程图

Fig. 3.1 The flow chart of the system

3.4 方法

3.4.1 药名文本去噪

虽然 OpenFDA 以 JSON 格式提供药物不良事件报告，但报告中的“药名域”仍然是非结构化的文本数据，医护人员或病人在提交药物不良事件报告时，对于“药名域”

的表达较为随意（如表 3.1 所示）。本文使用“药名文本”特指出现在不良事件报告中“药名域”中的文本。

药名文本通常包含很多“噪音”，如包含药物的别名、药物的用量以及药物的服用方式等。因此，本文先对药名文本进行预处理，清除药名文本中的噪音数据，以便更好的进行药名实体识别。具体的，预处理包括去除药名文本中[]，（）所包含的内容，以及药物的用法和用量等信息，如表 3.1 所示。

表 3.1　药名文本预处理

Tab. 3.1　Pre-processing of drug text

药名文本	预处理结果
plaquenil /00072602/	plaquenil
nifedipine [nifedipine]	nifedipine
fentanyl patch 100 mcg/hr	fentanyl patch
tylenol (caplet)	tylenol

3.4.2　药名实体识别（MetaMap）

药名文本预处理虽然一定程度上消除了数据噪音，但无法解决药名标准化的问题。如表 3.2 所示，同一种药物仍然存在多种表达方式，药名标准化是指将同一药物的多种表达方式聚为一类。表 3.2 展示了两种药物实体，其在 UMLS 中的 CUI 编码分别为：C0031412 和 C0237417，分别包含 3 种和 2 种表达形式。如果不对药物进行标准化，则会对实验造成严重的影响。

MetaMap 是美国国立医学图书馆（National Library of Medicine， NLM）开发的生物医学实体识别工具，用来识别生物医学文本中的医学概念，并将其映射到 UMLS 的超级叙词表中，广泛地用于生物医学方面的数据挖掘和知识发现的研究中。本文使用 MetaMap 识别不良事件报告中的药名。

使用 MetaMap，可以识别出药名文本中的药名实体，并给出药名实体在 UMLS 中的 CUI（concept unique identifier）。CUI 是 UMLS 超级叙词表中的概念唯一标识符。因此，使用 MetaMap 不仅可以识别药名文本中的药名，还可以对药名进行标准化。

3.4.3　语义类型过滤

虽然 MetaMap 可以识别药名实体并完成药名标准化，但 MetaMap 会从药名文本中识别出多余的实体，这些实体并不是药名。如对于表 3.2 中的“phenytoin sodium cap”，

使用 MetaMap 不仅可以识别出药名“phenytoin sodium”，还识别出实体“Cap”。为了使实验结果更加准确，需要对 MetaMap 的药名识别结果进行过滤。

表 3.2　同一实体多种表达方式

Tab. 3.2　Many names for one entity

药名	CUI
phenobarbital	C0031412
phenobarbital tab	C0031412
phenobarbitone	C0031412
phenytoin sodium	C0237417
phenytoin sodium cap	C0237417

表 3.3　药物语义类型列表

Tab. 3.3　The list of semantic types for drugs

语义类型	语义类型
Amino Acid, Peptide, or Protein	Inorganic Chemical
Antibiotic	Lipid
Biologically Active Substance	Manufactured Object
Carbohydrate	Enzyme
Chemical	Hormone
Chemical Viewed Functionally	Organophosphorus Compound
Chemical Viewed Structurally	Organic Chemical
Clinical Drug	Pharmacologic Substance
Eicosanoid	Substance
Nucleic Acid, Nucleoside, or Nucleotide	Steroid
Hazardous or Poisonous Substance	Vitamin
Neuroreactive Substance or Biogenic Amine	

本章使用基于语义类型的方法对非药名实体进行标记和删除。MetaMap 在完成实体识别的过程中，还给出了每个实体的语义类型。为了获取药物可能的语义类型，本文使用 MetaMap 对 SIDER 数据库中的 996 种药物名称进行映射，共得到 22 种语义类型，但其中 5 种语义类型不能用来标识药物，分别是：“Activity”，“Geographic Area”，“Human”，“Mammal”和“Quantitative Concept”。

另外，本文还添加 6 种语义类型，分别为:“Chemical Viewed Functionally”,“Chemical Viewed Structurally”，“Clinical Drug”，“Substance”，“Steroid”和“Vitamin”。虽然这些语义类型在 SIDER 包含的药物中并没有出现，但仍可以作为药物的语义类型。表 3.3 给出了本文使用的所有的药物语义类型。

使用表 3.3 中的语义类型，对 MetaMap 的药名识别结果进行过滤，完成药名识别和标准化。

3.4.4 生成分布式实体向量

每份药物不良事件报告通常包含多种药物和多种不良反应，图 3.2 展示了一份不良事件报告，该报告完成了药名识别和药名标准化，其中包含三种药物和五种不良反应。但这些药物和不良反应之间的关系是不确定的，所以如何从药物和不良反应的共现信息中挖掘最可能的药物不良反应成为本章研究的重点。本章利用非序列化 Skip-gram 模型生成药物和不良反应的实体向量，通过实体向量之间的相似度，进行潜在药物不良反应发现。

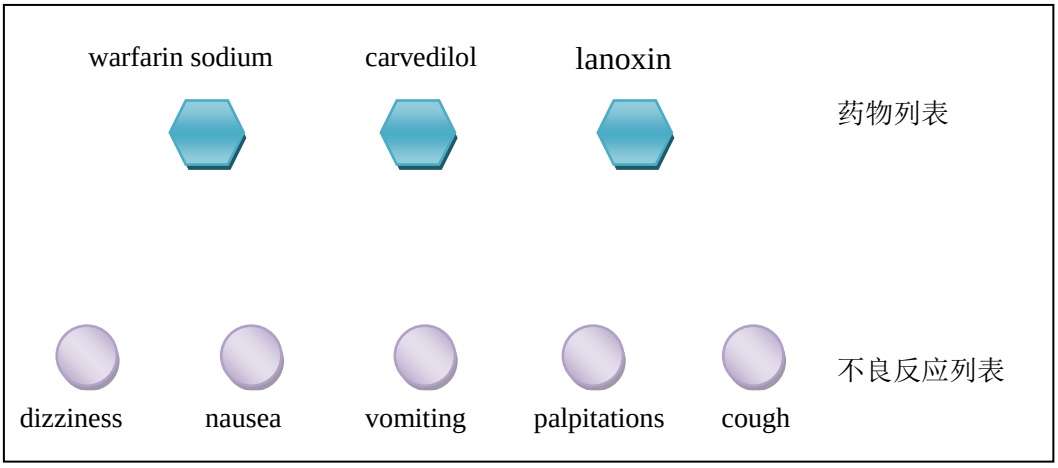


图 3.2 不良事件报告举例

Fig. 3.2 An example of adverse drug event report

非序列化 Skip-gram 模型，利用文档窗口机制捕捉实体的共现信息，从而训练生成药物和不良反应的实体向量。在本章中，每篇药物不良事件报告视为一篇文档，每篇文档包含药物实体列表和不良反应实体列表，在训练非序列化 Skip-gram 模型时，对这两种实体不加以区分。

在本章节中，式（2.4）和式（2.5）中各个符号的含义为：r_m 为第 m 篇不良事件报告，S 表示所有的药物不良事件报告集合；E_m 表示文档 r_m 中的实体列表，包括药物实体列表和不良反应实体列表；V 表示所有的药物实体和不良反应实体集合。

通过在药物不良事件报告集上训练非序列化 Skip-gram 模型，可以得到药物和不良反应的分布式实体向量，以此为基础，就可以计算药物和不良反应之间的关联度，从而实现发现潜在药物不良反应的目的。

3.4.5 计算药物和不良反应之间关联度

在生成药物和不良反应的分布式实体向量的基础上，本章将药物 d_i 和不良反应 a_j 之间的关联度定义为向量之间的余弦相似度，即：

$$sim(d_i, a_j) = \frac{v_{d_i}^T \cdot v_{a_i}}{\left\| v_{d_i} \right\| \cdot \left\| v_{a_i} \right\|} \tag{3.1}$$

其中，v_x 表示实体 x 的分布式向量。

根据药物和不良反应的关联度进行降序排序，就可以得到最有可能的潜在药物不良反应列表。

3.5 实验分析

由于缺乏标准数据集，本文采用案例验证的方法，检验非序列化 Skip-gram 模型在面向不良事件报告的潜在药物不良反应发现领域的效果。下面选取 olanzapine 作为研究对象，分析其潜在药物不良反应发现结果。Olanzapine 是 Leaman 等人[2]研究的对象之一。

Olanzapine（奥氮平）是一种精神类药物，主要用于治疗精神分裂症，及其相关障碍的急性期治疗和巩固治疗；除此之外，对于阳性症状（幻觉、妄想、猜疑等）和阴性症状（言语贫乏、情感冷漠等）均有治疗作用。利用公式（3.1）计算 olanzapine 与不良反应的关联度，并根据关联度对不良反应实体进行排序，从而得到 olanzapine 的潜在不良反应列表。表 3.4 列出了 olanzapine 潜在药物不良反应发现的结果（TOP10）。

在表 3.4 中，olanzapine 的前 3 种潜在不良反应为 schizoaffective disorder（情感分裂性精神障碍）、diabetic coma（糖尿病昏迷症）和 diabetic ketoacidosis（糖尿病酮酸中毒），SIDER 数据库中具有 olanzapine 导致这三种不良反应的记录。

Type 2 diabetes mellitus（2 型糖尿病），又称非胰岛素依赖型糖尿病，主要由胰岛素抵抗以及胰岛素相对缺乏引起，其患者体内产生胰岛素的能力并未完全丧失，临床表

现为多尿、多饮、多食。Girault 等人[29]指出 olanzapine 等非典型抗精神病药会诱发 weight gain 和 type2 diabetes mellitus 等不良反应；文献[30, 31]指出，olanzapine 与 type 2 diabetes mellitus 关系密切。从以上分析可以推断，olanzapine 具有导致 type 2 diabetes mellitus 的风险，而且在 SIDER 数据库中并没有 olanzapine 导致 type 2 diabetes mellitus 的记录（SIDER 中具有 olanzapine 导致 diabetes mellitus 的记录）。

表 3.4 olanzapine 潜在不良反应发现结果（TOP10）

Tab. 3.4 olanzapine's potential adverse reactions(TOP10)

不良反应	sim	SIDER
schizoaffective disorder	0.70	Y
diabetic coma	0.65	Y
diabetic ketoacidosis	0.63	Y
prescribed overdose	0.63	N
type 2 diabetes mellitus	0.62	N
neuroleptic malignant syndrome	0.61	N
hyperglycaemic hyperosmolar nonketotic syndrome	0.61	N
metabolic syndrome	0.60	N
diabetic complication	0.59	N
schizophrenia, paranoid type	0.59	N

注 1）：SIDER 列表示 SIDER 数据库是否存在 olanzapine 导致相应不良反应的记录，N 表示不存在，Y 表示存在。

Neuroleptic malignant syndrome（NMS，神经阻滞剂恶性综合征）是一种服用神经阻滞剂后产生的少见却可能致命的并发症，其临床表现为体温升高、心动过速、精神状态改变和血中肌酸激活酶升高等症状。Sa 等人[32]指出 olanzapine 会诱发 diabetic ketoacidosis 和 NMS。Tripathi 等人[33]详细分析了一位患有双向情感症病人的用药情况，指出病人在服用 olanzapine 后，出现 NMS。Sarıtaş 等人[34]研究了一位患有躁郁症的病人，服用 olanzapine 达 10 年之久，该病人出现 NMS 症状。经过以上分析，我们可以断定 olanzapine 会导致 neuroleptic malignant syndrome 的产生，而且，在 SIDER 数据库中并没有相应的记录。

hyperglycaemic hyperosmolar nonketotic syndrome，又称 Hyperosmolar hyperglycemic state （高血糖高渗状态，HHS），是糖尿病（主要是 2 型糖尿病）的并发症，以严重高血糖、高血浆渗透压、脱水为特点，并可能导致昏迷和死亡。经过上述分析，我们推

断 olanzapine 具有导致 type 2 diabetes mellitus 的风险，而 HHS 是其并发症，所以可以推断 olanzapine 会导致 HHS 的发生。这一推断在生物医学文献中也得到了验证。Endoh 等人[35]通过一个案例，详细分析了 olanzapine 与 HHS 的相关性，指出 HHS 对病人的生命产生严重威胁。在 SIDER 数据库中，并没有相关记录。

Metabolic syndrome(代谢综合征)是指生理代谢层面的心血管危险因子聚集的现象，这些危险因子包括高血压、血脂异常、糖尿病、肥胖以及高尿酸等。Malhotra 等人[36]指出 clozapine 和 olanzapine 等非典型抗精神病药物与 Metabolic syndrome 存在关联；Khalil 等人[37]的研究表明，在多种非典型抗精神病药物中，olanzapine 导致 metabolic syndrome 的几率最大。在 SIDER 数据库中没有 olanzapine 导致 metabolic syndrome 的记录。

综上，在 olanzapine 的前 10 种潜在不良反应中，有 3 种在 SIDER 数据库中存在相应记录，有 4 种在生物医学文献数据库 MEDLINE 中得到验证。对于 diabetic complication（糖尿病并发症），虽然在 SIDER 数据库和 MEDLINE 没有相应记录，但 olanzapine 可以导致 type 2 diabetes mellitus，所以 olanzapine 具有很大的可能性导致 diabetic complication。对于 schizophrenia, paranoid type（周期性精神分裂症），则是 olanzapine 的适应症，但这是由于数据本身的噪音造成的，一定程度上也说明本文的算法可以有效的识别出于 olanzapine 最相近的不良反应实体。

3.6 本章小结

本章利用非序列化 Skip-gram 模型，从大规模药物不良事件报告中利用药物和不良反应的共现信息，训练生成药物和不良反应的实体向量，将药物和不良反应之间的关联度计算转化为实体向量之间的运算。优选的，本文选取 FDA 的公共数据开放项目 OpenFDA 提供的“药物不良事件报告集”作为实验数据。虽然 OpenFDA 以 JSON 格式向公众提供数据，但每份不良事件报告中的药名部分仍然是非结构化的文本数据，本章首先利用 MetaMap 工具识别其中的生物医学实体，并加以药名语义类型和 DrugBank 数据库的辅助，提取每份不良事件报告涉及的药名。然后利用非序列化 Skip-gram 模型，生成药物和不良反应的实体向量，通过实体向量之间的运算，发现药物的潜在不良反应。实验表明，本章所述方法在药物不良事件报告中挖掘药物的潜在不良反应方面具有良好的表现。

4 面向社交网络的潜在药物不良反应发现

4.1 问题引出

随着互联网的迅速普及和各种社交媒体的快速发展，众多健康社交网站逐渐引起网民的关注并取得快速的发展，如 Ask a Patient，DailyStrength，MedHelp 等。在健康社交网站中，网络用户可以根据自己的经历和兴趣建立自己的“病友圈”，分享自己的健康经历，讨论与健康相关的各种话题，例如讨论自己的患病经历和用药体验。经过多年的累积，健康社交网站中积累了大量来自用药者的药物评论数据。这些用户评论是来自用药者的第一手资料，具有充分、及时、传播快等特点。

药物不良反应关乎公众健康，一直以来都是各界的广泛关注焦点，用药安全也越来越受到全社会的重视。健康社交网站中积累的药物用户评论，为发现潜在药物不良反应、改善用药安全提供了一种新的数据源。一方面，社交网络中积累的医疗健康数据蕴含丰富的药物不良反应信息，有待进一步挖掘；另一方面，在社交网络中，用户的用语很随意，经常具有拼写错误和语法错误，这些弊端给社交网络中的文本挖掘带来很大的挑战。到现在为止，从健康社交网站中提取不良反应的研究还相对较少。

基于社交网络可以快速地收集到关于药物的评论数据，可以利用文本挖掘的相关知识挖掘其中的潜在药物不良反应信息，但这些潜在不良反应并没有经过严格医学意义上的检验。因此能否成为药物在医学意义上的药物不良反应尚且需要临床试验的检验和证实。如果可以为所挖掘的潜在药物不良反应找到某些可以解释的原因，例如某种蛋白质，那么就可以大幅度的减少医学领域专家用来确定药物不良反应真实性的时间，这对及时发现药物的安全隐患、改善公众用药安全具有重要意义。

本章节首先利用基于信息熵和字典匹配的方法从健康社交网络的用药者评论数据中识别潜在的药物不良反应。然后在生物医学文献数据库 MEDLINE 的基础上，利用非序列化 Skip-gram 模型，训练生成生物医学实体的分布式向量，据此寻求可以把药物和不良反应联系起来的蛋白质，尽最大努力发现药物引起不良反应的内部机制，减少领域专家最终确定潜在药物不良反应真实性的时间和经济代价，实现及时发现药物安全隐患的目的。

4.2 研究框架

本章节的研究目的是：在文本挖掘技术的基础上，从健康社交网站积累的用药者文本评论数据中，识别药物的潜在不良反应，并尽最大努力为所发现的潜在药物不良反应寻找蛋白质级别的“证据”，从而形成解释潜在药物不良反应的“证据链”。因此，本

章节所述系统在逻辑上包括三部分：数据抓取模块，潜在不良反应识别模块和关联蛋白质寻求模块，如图 4.1 所示。

4.2.1 数据获取模块

本章节所使用的用户评论数据来自互联网，需要使用爬虫相关技术对用户评论进行抓取。本文利用 scrapy（http://scrapy.org/） 程序包搭建网络爬虫，从社交网站中抓取相应的用药者评论。Scrapy 是一个开源的 python 框架，可以根据需求快速实现爬虫的搭建。

4.2.2 潜在不良反应识别模块

首先，该模块利用信息熵和字典匹配的方法，从用户评论数据中识别疾病和不良反应实体；然后使用 Semantic MEDLINE 和 DrugBank 对药物的适应症进行过滤，并利用 SIDER 数据库对已登录的药物不良反应进行标记和过滤，最终得到药物的潜在不良反应列表。

4.2.3 关联蛋白质寻求模块

本章节中，关联蛋白质是指可以把药物与不良反应联系起来的蛋白质，在某种程度上，关联蛋白质可以作为药物引起相应不良反应的原因。如果可以为潜在药物不良反应找到相应的关联蛋白质，对于验证潜在药物不良反应在医学意义上的真实性具有重要的参考意义。本文采用了基于非序列化 Skip-gram 模型的生物医学实体关联度计算方法，并在实体关联度的基础上定义了寻找关联蛋白质的关联紧密度函数，以此为潜在药物不良反应找寻关联蛋白质，为最终确定其真实性提供蛋白质级别的参考。

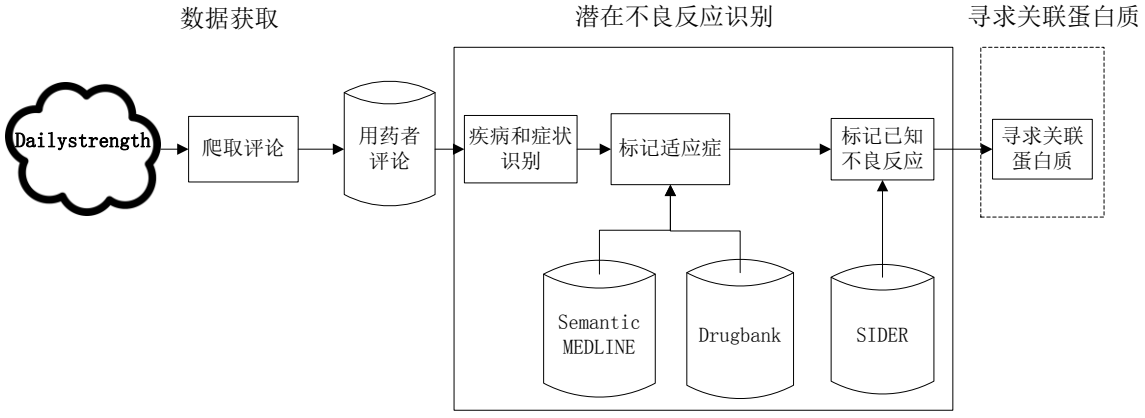


图 4.1 系统流程图

Fig. 4.1 The flow chart of the system

4.3 识别潜在药物不良反应

在临床医学中，疾病名称集合和不良反应名称集合具有很大的交集，如 headache 可以作为疾病的名称，也可以作为药物不良反应的名称。本文的主要研究目的是识别用户评论中的潜在不良反应，所以对于药物的适应症（疾病）和药物的不良反应应加以区分，实现过滤适应症、识别不良反应的目标。在本文中，识别潜在药物不良反应在整体上包括三部分：识别“疾病和不良反应”名称，过滤药物适应症（疾病名称），以及过滤已知的不良反应。

识别“疾病和不良反应”名称，包括构建词典和名称识别两部分；对于药物适应症过滤，本文使用已有的生物医学资源 DrugBank 和 Semantic MEDLINE 进行药物适应症过滤。同时，为了实现识别药物的“潜在”不良反应，需要对已经具有相关记录的不良反应加以标记，并删除。具体的，本文使用药物不良反应数据库 SIDER 标记已知的药物不良反应并加以过滤，从而得到药物的“潜在”不良反应列表。

4.3.1 生成“疾病和不良反应”词典

本文使用的“疾病和药物”词典 *IndSyn* 是基于 SIDER 数据库生成的，该数据库中包含 5719 种不良反应名称和 2669 种适应症的名称。通过合并，得到“疾病和药物”词典 *IndSyn*。由于不良反应和疾病有交集，词典 *IndSyn* 包含 6315 种疾病和不良反应实体名称。

在“疾病和不良反应”实体识别过程中，本文使用词典 IndSyn 完成候选文本到实体名称的映射。

4.3.2 基于信息熵和字典匹配的“疾病和不良反应”实体识别

从用药者评论中识别疾病和不良反应实体，可以理解为：从用药者评论中提取表达疾病和不良反应的文本片段，这些文本片段应该具有高频率和高信息熵的特点。一个文本片段的信息熵越高，说明这个文本片段就越有可能是一个“词”。

信息熵被广泛的用于微博数据中的新词发现和关键词提取，在本文中我们使用信息熵来识别候选疾病和不良反应实体。具体地，假设 s 表示一个文本片段，L 表示 s 在评论数据中的左邻接词集合，R 表示 s 在评论数据中的右邻接词集合。

s 的左信息熵定义为：

$$LE(s) = -\sum_{w \in L} p(w) \log p(w) \tag{4.1}$$

其中 $p(w)$ 表示 w 是 s 的左邻接词的概率。

同理，s 的右信息熵定义为：

$$RE(s) = -\sum_{w \in R} p(w) \log p(w) \quad (4.2)$$

其中 $p(w)$ 表示 w 是 s 的右邻接词的概率。

如果 s 的左信息熵和右信息熵都比较高，那么 s 表示一个词的概率就很大。但是本文的目的是识别用药者评论中的疾病和不良反应实体，而不是对用药者评论进行分词，所以对于信息熵较高的 s 要进行过滤，如果 s 可以映射到“疾病和不良反应”词典 IndSyn 中的某一项，则保留 s，否则去除 s。

本文利用 Jaccard 相似性系数作为文本重叠度函数将文本 s 映射到词典 *IndSyn* 中。具体的，设 $t \in IndSyn$，表示一种疾病或者症状，定义文本片段 s 和词典项 t 的重叠度为：

$$l(s,t) = \frac{|W_s \cap W_t|}{|W_s \cup W_t|} \quad (4.3)$$

其中 W_l 表示对文本 l 进行分词和去停用词后包含的单词集合。$|M|$ 表示集合 M 所包含的元素个数。令 $map(s)$ 表示 s 映射到词典 *IndSyn* 中的项，则 $map(s)$ 定义为：

$$map(s) = \arg\max_{t \in IndSyn} l(s,t) \quad (4.4)$$

如果 $map(s)$!= *NULL*，则表示文本 s 可以映射到词典 *IndSyn* 中。

4.3.3 基于 DrugBank 和 Semantic MEDLINE 的适应症标记

用药者在分享用药经历或者评论某种药物时，不可避免的会提到该药物的适应症或者用药的原因。比如药物 trazodone 的一条评论：“I use this primarily for my sleeplessness”，明确的说明 sleeplessness 是用药的原因，不是 trazodone 的不良反应。所以，应当从识别出的“疾病和不良反应”实体中标记药物的适应症，并将其过滤掉。

药物的适应症可以从药典数据源 DrugBank 中得到。在 DrugBank 中，对于描述药物适应症的数据是叙述性文本，因此是非结构化的数据，例如：trazodone 的适应症为：”For the treatment of depression.”。在本文中，我们使用 MetaMap 从 DrugBank 的适应症描述中识别出相关的生物医学实体，并使用词典 IndSyn 去除非疾病和不良反应实体，从而得到药物的适应症。

除了药物官方文档中标明的适应症，药物经常还有其他的适应症。比如从上述评论中我们还可以看出 trazodone 除了治疗 depression 之外，还可以用于治疗 sleeplessness。所以本文还用 Semantic MEDLINE 对药物适应症做进一步过滤。Semantic MEDLINE 是 SemRep[16] 从 MEDLINE 引用中识别出的三元语义关系（subject-predicate-object）知识库，这些三元组表示 subject 和 object 之间的语义关系为 predicate。例如，如果在 Semantic

MEDLINE 中存在三元语义关系：（trazodone-TREATS-sleeplessness），我们就可以断定 trazodone 可以用来治疗 sleeplessness，从而说明 sleeplessness 也是 trazodone 的适应症。

综上，本文使用 DrugBank 和 Semantic MEDLINE 相结合的方法来标记药物的适应症。

4.3.4 基于 SIDER 的已知药物不良反应标记

经过“适应症”过滤，就可以得到与药物有关的不良反应。然而这些不良反应有很多在药物的官方文档中已有相关记录，本文称其为已知的药物不良反应。SIDER 是从药物官方文档中提取的关于药物不良反应的数据库。本文的目的是识别“潜在”的药物不良反应，所以本文使用 SIDER 数据库对已知的药物不良反应进行标记，并将其过滤掉。

4.3.5 潜在药物不良反应标记

经过适应症和已知不良反应的标记，剩下未标记的“疾病和不良反应”实体就可以作为“潜在”的药物不良反应。

图 4.2 是识别潜在不良反应的详细算法。

输入：药品 d_i 的用药者评论数据集 C_i

输出：“疾病和不良反应”实体名称以及每个实体与药品 d_i 的关系

1：在集合 C_i 中生成文本片段集合 $S_i = \{s \mid LE(s) > 0, RE(s) > 0\}$

2：计算平均频率：$\overline{fre} = \frac{1}{S_i} \sum_{s \in S_i} fre(s)$，其中 $fre(s)$ 表示 s 出现的次数

3：计算平均左右信息熵：$\overline{LE} = \frac{1}{|S_i|} \sum_{s \in S_i} LE(s)$ 和 $\overline{LR} = \frac{1}{|S_i|} \sum_{s \in S_i} LR(s)$

4：生成疾病和不良反应集合：

$D_i = \{l = map(s) \mid fre(s) > \overline{fre},\ LE(s) > \overline{LE},\ \ RE(s) > \overline{RE},\ map(s)!= NULL, s \in S_i\}$

5： for $e_m \in D_i$

6： if e_m 是药品 d_i 的适应症，则将 e_m 标记为”-”

7： else if e_m 是药品 d_i 的已知不良反应，则将 e_m 标记为”+”

8： else 将 e_m 标记为”*”

图 4.2 识别潜在不良反应的算法

Fig. 4.2 The algorithm of recognizing potential adverse drug reactions

4.4 寻求潜在不良反应的证据

不良反应识别完成之后，可以得到药物的潜在不良反应列表。然而所识别出的潜在不良反应尚未经过严格医学意义上的检验，所以能否成为临床意义上的不良反应还需要进一步的检验和证实。检测潜在不良反应的真实性需要进行大量的临床试验和观察，这是一个耗时耗力的过程，不利于及时发现药物的安全隐患。如果可以利用文本挖掘的相关技术，挖掘药物导致潜在不良反应的内部机制，并推荐给医学领域的专家作参考，对于检验潜在不良反应的真实性，改善用药安全，具有重要意义。

蛋白质是生命活动的主要承担者，是生命的物质基础，因此药物的不良反应也多与蛋白质有关。为了便于研究，本文假定药物通过蛋白质引起不良反应。因此，对于上一步发现的“潜在”药物不良反应，本文尽最大努力找到可以把药物和其“潜在”不良反应联系起来的蛋白质，把这些蛋白质作为药物导致不良反应的“证据”，并把这些（药物，蛋白质，不良反应）作为三元组关系，推荐给医学领域的专家作参考，为他们最终确定药物和不良反应的关系减少时间和经济代价。

为了寻求关联蛋白质，本文首先利用 Skip-gram 模型生成生物医学实体的分布式向量，根据实体向量计算实体之间的关联度，在此基础上，根据药物-蛋白质-不良反应三者之间的关联度函数挖掘关联蛋白质。

4.4.1 关联度

对于蛋白质 p，本文利用关联紧密度函数 $f(d,p,a)$ 来衡量其作为药物 d 和“潜在”不良反应 a 的“证据”的可信度。$f(d,p,a)$ 越大，表示蛋白质 p 越能把药物 d 和“潜在”不良反应 a 联系起来，也就表示蛋白质 p 作为“证据”也就越可信，从而蛋白质 p 越能解释药物 d 和“潜在”不良反应 a 的内部原因。本文将其称为关联蛋白质。

具体的，$f(d,p,a)$ 的定义为：

$$f(d,p,a)=\frac{sim(d,p)+sim(p,a)}{1+\left|sim(d,p)-sim(p,a)\right|} \tag{4.5}$$

其中 $sim(x,y)$ 表示实体 x 和 y 的关联度。直观上，如果 $sim(d,p)+sim(p,a)$ 越大，那么关联紧密度 $f(d,p,a)$ 也越大。但是为了防止因 $sim(d,p)$ 或者 $sim(p,a)$ 单方过高而导致的 $f(d,p,a)$ 过高，这里对其使用 $1+\left|sim(d,p)-sim(p,a)\right|$ 进行“平滑”。即：如果 $sim(d,p)+sim(p,a)$ 很高，并且 $sim(d,p)$ 和 $sim(p,a)$ 差异很小，$f(d,p,a)$ 才会高。

4.4.2 基于 Skip-gram 模型的生物实体关联度

传统计算实体 x 和 y 的关联度的方法（如点间互信息 PMI）直接基于 x 和 y 的“共现”情况来计算，如果 x 和 y 经常共现，那么它们之间的关联度也越高。这种方法过于简单，会带来很多噪音。为了更好的衡量实体之间的关联度，本文采用基于 word2vec 的 Skip-gram 模型的关联度计算方法。

为了训练 Skip-gram 模型，本文选择的数据是 MEDLINE 文献引用的 MeSH 词域。生物医学专家为 MEDLINE 中每篇文献都使用某些 MeSH 词进行标注，这些 MeSH 词能很好的描述论文的主题内容，同时，这些 MeSH 词也可以作为共现信息来使用

原始的 Skip-gram 模型是一种语言模型，所需要的训练数据是“序列化”的自然语言文本，然而 MEDLINE 文献引用的 MeSH 词域是 MeSH 词集合，是“非序列化”的数据。所以，在 MEDLINE 文献引用的 MeSH 词域的基础上，采用非序列化的 Skip-gram 模型进行训练，生成 MeSH 词的词向量。

具体的，令 p_i 表示第 i 篇包含 MeSH 词的 MEDLINE 引用，S_i 表示 p_i 的 MeSH 词集合，m_{ij} 表示 p_i 中第 j 个 MeSH 词，$j=1,2,\cdots,|S_i|$，$|S_i|$ 表示集合 S_i 的大小。修改后的 Skip-gram 模型的目标函数为：

$$\max \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{|S_i|}\sum_{\substack{k=1\\k\neq j}}^{|S_i|}\log p(m_{ik}\mid m_{ij}) \tag{4.6}$$

其中 N 表示包含 MeSH 词的所有 MEDLINE 引用总数。

通过上述修改后的 Skip-gram 模型，就可以得到每个 MeSH 词的词向量。MeSH 词的关联度定义为余弦相似度，即：

$$sim(m_1,m_2)=\cos(v_{m1},v_{m2})=\frac{v_{m1}^{T}v_{m2}}{\|v_{m1}\|\cdot\|v_{m2}\|} \tag{4.7}$$

4.5 实验结果分析

针对上述提出的方法，本章共进行了三个实验：实验一主要用于说明基于信息熵的方法可以有效的进行不良反应名称识别，并分析了在社交网络中潜在不良反应发现的结果；实验二用于说明修改后的 Skip-gram 模型在 MEDLINE 数据集上可以捕捉生物医学实体之间的关联度，从而可以用于发现药物和不良反应之间的关联蛋白质；实验三给出了为潜在药物不良反应寻找“证据链”的结果，说明基于 MeSH 词向量的关联度函数可

以有效的发现药物和不良反应的关联蛋白质，为领域专家尽早确定潜在药物不良反应的真实性提供参考依据。

4.5.1 不良反应识别结果

由于在“社交网络中挖掘药物不良反应信息”这一领域缺乏权威的数据集，为了便于比较，本文选择 Leaman[2]所用的健康社交网站 Dailystrength 作为用户评论数据的来源。具体的，在本文中，使用基于 python 的 scrapy 爬虫框架，以 2014 年 6 月 2 日为截止日期，从 Dailystrength 中爬取 600,237 条用药者评论。这些评论中，总共涉及 1075 个健康话题，其中绝大数是关于药物的话题。本文选择评论数最多的 50 种药物作为研究对象。

表 4.1 是与 Leaman 结果的对比情况。从中可以看出，本文的结果与 Leaman 的结果相似性很高，说明本文使用的基于信息熵和词典匹配的不良反应名称识别方法是有效的。其中识别错误主要源于词典 *IndSyn* 中某些名称包含很多停用词。如表 1 中的“not as effective”，去停用词后变为“effective”，从而导致从包含 effective 的用户评论都会识别出该不良反应。另一主要的识别错误是由于词典 IndSyn 中某些名称是由常用词组成的，而且相对较短。如“feeling high”和“effect increase”，在词典映射过程中，feeling 和 feelings 等都会映射到“feeling high”，“increase”和“increased”等都会映射到“effect increase”，从而导致识别错误。

表 4.2 是基于信息熵的方法对于上述 50 种药物的不良反应识别结果。本实验从 50 种药物的评论数据中抽取出 993 个（药物，疾病或症状）关系。总共识别出 265 个适应症关系，其中 DrugBank 标记出 34 个，Semantic MEDLINE 标记出 234 个，所占的百分比分别为 3.4%和 23.3%。对于药物不良反应关系，240 个在 SIDER 种有相应的记录，而 488 个在 SIDER 中并没有记录，所占的百分比分别为 24.2%和 49.1%。这 488 个未记录的药物不良反应就是“潜在”的药物不良反应。

从表 4.2 的结果可以看出，在社交网络中，用药者更倾向于“陈述”药品说明书中未记录的药物不良反应，这也符合实际情况。如果药品说明书中已经说明了某种不良反应，则用户就不会过分“担心”这种不良发应，在心理上甚至认为这种不良反应在某种程度上是“正常”的。相反，如果药品说明书中没有出现某种不良反应，而用药者自身出现了该不良反应，则其更倾向于“寻求”帮助和分享自己的经历。

对于“潜在”不良反应 (d,a)，首先在 Semantic MEDLINE 中检索是否存在三元组 $(d,CAUSES,a)$，如果存在，则说明药物 d 会引起不良反应 a。从 Semantic MEDLINE 中

总共可以为 10 个“潜在”不良反应找到上述三元组。由于篇幅限制，表 4.3 展示了其中 3 个不良反应关系。

表 4.1　与 Leaman 的结果对比

Tab. 4.1　The comparison of the results with Leaman

药物	Leaman 的结果	本文结果
carbamazepine	**somnolence or fatigue(12.3%)**, allergy (5.2%), **weight gain (4.1%)**, **rash (3.5%)**, depression (3.2%), **dizziness (2.4%)**, tremor/spasm (1.7%), **headache (1.7%)**, appetite increased (1.5%), **nausea (1.5%)**	seizures-(29.6%), not as effective*(23.8%), pain-(17.0%), **rash-(10.2%)**, **sleepy*(4.4%)**, effect increased*(3.9%), **weight gain abnormal*(3.2%)**, **dizziness+(2.7%)**, **headache-(2.7%)**,**nausea+(2.7%)**
trazodone	**somnolence or fatigue (48.2%)**, **nightmares (4.6%)**, **insomnia (2.7%)**,addiction (1.7%), **headache (1.6%)**, depression (1.3%), **hangover (1.2%)**, **anxiety attack (1.2%)**, panic reaction (1.1%), dizziness(0.9%)	**insomnia-(18.1%)**, **anxiety*(12.8%)**, not as effective*(10.9%), wakefulness*(10.4%), **sleepy*(7.6%)**, **nightmare-(7.3%)**, **hangover effect*(4.8%)**, feeling high*(4.4%), **drowsiness-(3.9%)**, **headache+(3.4%)**
ziprasidone	**somnolence or fatigue (20.3%)**, dyskinesia (6.0%), **mania (3.7%)**, **anxiety attack (3.5%)**, **weight gain(3.2%)**, depression (2.4%), allergic reaction (1.9%), dizziness (1.2%), panic reaction (1.2%)	not as effective*(33.7%), **sleepy*(15.2%)**, **anxiety+(9.6%)**, **mania-(6.3%)**, **weight gain abnormal*(4.5%)**, hallucination* (4.3%), suicide*(3.9%), feeling high*(2.9%), effect increased* (2.5%), stomach*(2.4%)
olanzapine	**weight gain(30.0%)**, **somnolence or fatigue(15.9%)**, appetite increased(4.9%), depression(3.1%), tremor(2.7%), **diabetes(2.6%)**, **mania(2.3%)**, **anxiety(1.4%)**, **hallucination(0.7%)**, edema(0.6%)	**weight gain abnormal*(54.0%)**, not as effective*(16.9%), **diabetes-(6.4%)**, **anxiety+(5.6%)**, **sleepy*(5.4%)**, **mania-(3.9%)**, **hallucination-(2.7%)**, psychosis-(2.5%), feeling high+(2.5%)

注1）　在本文方法的识别结果中，疾病和不良反应名称使用‘+’，‘-’，‘*’来标记。‘-’表示适应症。‘+’表示已知的药物不良反应；‘*’表示“潜在”的不良反应。

表 4.2 用户评论中“疾病和不良反应”的分布

Tab. 4.2 The distribution of diseases and symptoms in users' comments

标记	计数	百分比
DrugBank 适应症	34	3.4%
Semantic MEDLINE 适应症	231	23.3%
SIDER 不良反应	240	24.2%
“潜在”不良反应	488	49.1%
总计	993	100%

表 4.3 Semantic MEDLINE 中寻求(d,CAUSES,a)的部分结果

Tab. 4.3 The results of finding (d,CAUSES,a) in Semantic MEDLINE

药物	不良反应	PMID	句子
ziprasidone	drowsiness	18211129	In schizophrenia, only olanzapine, **ziprasidone**, and aripiprazole (NNTHs from 5 to 14) **caused** a significantly higher incidence of **somnolence**.
quetiapine	drowsiness	91570956	However, there was a significantly higher incidence of discontinuation due to adverse events and **somnolence** caused by **quetiapine** in bipolar depression than that in schizophrenia or mania.
fluoxetine	nausea	15531007	The incidence of duloxetine-induced **nausea** resembled that produced by paroxetine and **fluoxetine**.

注1） drowsiness 和 somnolence 都是“睡意，困倦”的意思

4.5.2 分布式实体向量

MEDLINE 通过人工标注的方式为每篇文献赋予一些足以描述论文主题的 MeSH 词，每篇引用的 MeSH 词组成一个共现集合。在本实验中，选取 2013 年（含 2013 年）之前的含有 MeSH 域的 2200 万篇 MEDLINE 引用作为基础，并从中抽取出相应的 MeSH 词共现集合组成训练数据集。在训练修改的 Skip-gram 模型时，本文采用 word2vec 的 Hierarchical Softmax 算法，生成的 MeSH 词向量为 100 维。

通过两个案例来说明 Skip-gram 模型生成的 MeSH 词向量可以用来计算 MeSH 词之间的关联度。本文选取点间互信息（Point Mutual Information：PMI）作为对比。

表 4.4 给出了两个模型分别用于求 Headache 和 Snake venoms 最相近的 10 个 MeSH 词的结果。

对于 Headache，Skip-gram 得到词中，6 个是具体的头痛类型，还得到 facial neuralgia 和 trigeminal neuralgia 等与 Headache 语义很相似的词。而 PMI 中只有三个是具体的头痛类型。而且 PMI 模型得到的词中，nausea、spinal puncture（脊髓穿刺）等词跟 headache 是无关的，可见相较于 Skip-gram，PMI 会引入更大的“噪音”。

对于 Snake venoms（蛇毒），两个模型所得到的 MeSH 词都是相关的。Skip-gram 模型得到 4 个是具体的蛇毒类型，1 个毒液的总称，并得到了抗蛇毒素和响尾蛇毒蛋白。只有蝮蛇科、具窍蝮蛇属和竹叶青蛇属与蛇毒的“关联性”差些。而 PMI 模型得到的只有抗蛇毒素、单链蛇毒多肽和爬虫类蛋白质与蛇毒相近，其他的词更大程度上跟“蛇”相近。

通过这两个简单的例子可以看出，经过修改的 Skip-gram 模型可以有效的计算 MeSH 词之间的关联度，并且引入的噪音相对较少。

表 4.4　修改的 Skip-gram 模型与 PMI 模型对比

Tab. 4.4　The comparison between the modified Skip-gram model and PMI model

序号	Headache		Snake venoms	
	Skip-gram	PMI	Skip-gram	PMI
1	migraine disorders	blood patch, epidural	crotalid venoms	snake bites
2	vascular headaches	intracranial hypotension	viper venoms	antivenins
3	cluster headache	vascular headaches	elapid venoms	colubridae
4	headache disorders, primary	spinal puncture	venoms	snakes
5	dizziness	migraine disorders	antivenins	bothrops
6	facial neuralgia	dizziness	hydrophid venoms	disintegrins
7	trigeminal neuralgia	tension-type headache	cobra venoms	viperidae
8	headache disorders	nausea	bothrops	agkistrodon
9	tension-type headache	facial pain	crotoxin	reptilian proteins
10	facial pain	cluster headache	viperidae	bungarus

4.5.3　关联蛋白质

为了使用上述修改的 Skip-gram 模型生成的 MeSH 词向量来计算实体之间的关联

度，对于药物d和不良反应a，需要对其使用 Restrict to MeSH 算法将其映射为 MeSH 词。对于上述 488 个“潜在”不良反应关系，其中 10 个已经在 Semantic MEDLINE 中找到依据，对于剩余的 478 个“潜在”不良反应关系，Restrict to MeSH 算法将其中的 160 个关系成功的使用 MeSH 词来表示。

对于每个“潜在”的药物不良反应关系(d,a)，本文选取$f(d,p,a)$最高的 5 个蛋白质作为把药物d和不良反应a的关联蛋白质。表 4.5 是 trazodone 和 anxiety 关联蛋白质的提取结果。

表 4.5　Trazodone 和 anxiety 的关联蛋白质

Tab. 4.5　The associated proteins of trazodone and anxiety

蛋白质	f(d, p, a)
receptors, serotonin	0.83
5-hydroxytryptophan	0.76
serotonin plasma membrane transport proteins	0.75
receptors, adrenergic	0.75
receptor, serotonin, 5-ht1a	0.74

Trazodone 是一种抗抑郁药，属于 serotonin（5-hydroxytryptophan）受体拮抗剂和再摄取抑制剂。此外，trazodone 也会阻塞 alpha-adrenergic，对 alpha2-adrenergic 有一定的阻塞作用。

下面主要对不良反应 anxiety 做进一步论述。

论文 JA Gingrich[38]指出，“To date several **inactivation mutations of specific serotonin receptors** have been generated producing interesting behavioral phenotypes related to **anxiety**, depression, drug abuse, psychosis, and cognition.”。可以看出 serotonin receptors（5-hydroxytryptophan）与 anxiety 是相关的。

Goldman[39]指出，“HTTLPR (minor allele frequency 0.40) alters **serotonin transporter** function to affect **anxiety**, dysphoria and obsessional behavior, which are assessed in COMBINE and may be related to relapse and addictive behavior.”。可以看出，serotonin transporter 也会影响 anxiety。可以推断，作为特殊的 serotonin transporter，serotonin plasma membrane transport proteins 跟 anxiety 也是相关的。

Shishkina[40]指出，“Brain alpha2-adrenergic receptors (alpha2-ARs) have been implicated in the regulation of anxiety, which is associated with stress.”，说明 adrenergic receptors 跟 anxiety 是有联系的。

通过以上简要分析，本文找到了 trazodone 和 anxiety 之间的四个三元关系组：（trazodone，”serotonin receptors”，anxiety），（trazodone，” 5-hydroxytryptophan”，anxiety），（trazodone，” serotonin plasma membrane transport proteins”，anxiety），（trazodone，” adrenergic receptors”，anxiety），这些三元关系组为 trazodone 和 anxiety 关系的确定提供了参考。

4.6 本章小结

本章旨在从社交网络中提取药物的不良反应，并为“潜在”的不良反应寻求蛋白质级别的“证据”，尽最大努力解释药物和其“潜在”不良反应的关系。

本章首先使用基于信息熵的方法提取用药者评论中的不良反应，并加以词典的辅助，良好的完成了不良反应名称的识别工作。由于本方法是非监督的方法，具有较好的泛化能力。但由于本方法是基于统计的方法，需要的用户评论数应尽可能地多。

然后，本章利用非序列化 Skip-gram 模型生成的 MeSH 词向量，尽最大努力的为“潜在”不良反应寻求蛋白质证据，尝试找到可以把药物和其不良反应关联起来的蛋白质，从而为最终确定药物和不良反应的关系推荐线索。不足之处在于，药名和不良反应名称是 UMLS 超级叙词表中的概念，而修改的 Skip-gram 模型使用的是 MeSH 词，restrict to mesh 算法并不能实现完全映射。在未来的工作中，我们致力于解决这一问题。

综上，由社交网络启动，融合生物信息资源的药物不良反应发现，能够及时发现潜在药物不良反应，并尽最大努力寻求可以把药物和不良反应联系起来的蛋白质，使潜在药物不良反应的检测具有更加实用的价值，对改善人类健康水平、减少经济损失具有重大的意义

结　论

生物医学技术的发展为人类提供了大量的药品用以治疗各种疾病。药物的出现对于改善人类身体健康水平，延长人类寿命具有至关重要的意义。然而，作为药物的一种本质属性，药物不良反应对于人类身体健康和心理健康却产生了严重损害，某些严重的药物不良反应甚至具有致命性。统计数据表明，药物不良反应已经成为多个国家的人口死亡的重要原因。除了对病人的身体和心理产生危害之外，药物不良反应也给病人家庭和社会造成了巨大的经济损失。鉴于不良反应的严重危害性以及药品种类的日益增多，药物不良反应成为医学界和学术界关注的热点之一。

大部分的不良反应在药物的临床试验阶段可以被揭露，并记录在药物的说明文档中。然而由于周期相对较短、受试人群较为单一，临床试验无法揭露药物导致的所有不良反应，导致具有潜在不良反应的药物流向市场。上市后，服用药物的人群变得日益复杂，越来越多的未知药物不良反应逐渐在用药人群中表现出来。相较于药物说明文档中所记录的不良反应，由于其严重的不可控性，这些未知的药物不良反应对于用药者的身体健康产生的危害也更加严重。如果能够及时发现药物的潜在不良反应，从而对药物的安全隐患提出预警，对于改善人类用药状况和提升公众的健康水平具有重要意义。

药物上市后，药物不良事件报告系统成为监控药物使用情况、发现药物安全隐患的重要依托。每份药物不良事件报告通常包含多种药物以及多种不良反应。为了从药物不良事件报告集中充分挖掘药物和不良反应之间的关联信息，本文第三章利用非序列化 Skip-gram 模型，以不良事件报告中药物和不良反应的共现为基础，生成药物和不良反应的分布式实体向量。然后在分布式实体向量的基础上，计算药物和不良反应之间的关联度，并根据关联度进行排序，达到发现潜在药物不良反应的目的。实验表明，基于 Skip-gram 模型的药物不良反应发现方法，可以有效的挖掘药物的不良反应信息，达到发现药物安全隐患的目的。

在信息技术取得长足发展以及互联网和移动互联网大规模普及的今天，健康社交网站受到越来越多的关注。用药者在健康社交网站中分享自己的用药经历和体会，对所服用的药物做出相对真实的评价。随着时间的推移，健康社交网站中积累了大量来自病人的药物评论数据，这些评论数据蕴含丰富的药物不良反应信息。本文第四章利用信息熵和字典匹配的方法从药物评论中挖掘“疾病和不良反应”名称，并利用药物数据库 DrugBank、药物不良反应数据库 SIDER 等生物医学数据资源，过滤药物的适应症以及已有记录的药物不良反应，实现潜在药物不良反应发现的目标。然而，以上步骤所发现

的不良反应仍然是“潜在”的，其临床意义上的真实性尚未得到验证。但是，验证药物不良反应在临床意义上的真实性是耗时耗力的过程，对及时发现药物的安全隐患产生不利的影响。因此，本文第四章以 Skip-gram 模型生成的分布式生物医学实体向量为基础，为潜在药物不良反应挖掘蛋白质级别的“证据”，即挖掘药物和其潜在不良反应之间的关联蛋白质。这些关联蛋白质可以作为生物医学专家最终确定潜在不良反应真伪性的参考，从而尽最大努力缩短潜在不良反应真伪性确定的周期，实现及时发现药物安全隐患的目的，为改善公众的用药状况做出贡献。

在接下来的工作中，我们会从以下方面对现有方法做出改进：1）在 AERS 系统的基础上，利用分布式实体向量，挖掘由于药物相互作用导致的不良反应；2）改善面向社交网络的不良反应识别方法，为所发现的潜在药物不良反应挖掘更多方面的“证据”（如基因、靶点等），从而为验证其真伪性提供更加全面的参考，尽力缩短验证真伪性的时间。

参 考 文 献

[1] Giacomini K M, Krauss R M, Roden D M, et al. When good drugs go bad[J]. Nature, 2007, 446(7139): 975-977.

[2] Leaman R, Wojtulewicz L, Sullivan R, et al. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks[C] Proceedings of the 2010 workshop on biomedical natural language processing. Association for Computational Linguistics: 117-125.

[3] 朱芹英. 我国药品不良反应现状与建议[J]. 青海医药杂志, 2009, (11): 70-71.

[4] Wester K, Jönsson A K, Spigset O, et al. Incidence of fatal adverse drug reactions: a population based study[J]. British journal of clinical pharmacology, 2008, 65(4): 573-579.

[5] 刘新社, 祁秋菊, 董玲莉. 正确认识药品不良事件与药品不良反应的关系 提高监管工作的针对性[J]. 中国药事, 2008, (07): 547-549.

[6] 陈俊玲, 陈冠全, 陈文戈, 等. 关联规则算法在药品不良反应中的数据挖掘研究[J]. 电子世界, 2012, (09): 86-88.

[7] 冯变玲, 杨世民, 贺小红, 等. 药品不良反应多维关联规则挖掘及预警模型构建[J]. 中国药事, 2012, (10): 1076-1082.

[8] Harpaz R, Haerian K, Chase H S, et al. Statistical mining of potential drug interaction adverse effects in FDA's spontaneous reporting system[C] AMIA Annual Symposium Proceedings. American Medical Informatics Association,2010: 281.

[9] 蒋朋利, 吴晶, 孙鹤. 他汀类药物的使用与认知能力下降、糖尿病及癌症之间的关系[J]. 中国临床药理学与治疗学, 2014, (10): 1132-1138.

[10] 魏建香, 孙越泓, 朱云霞, 等. 一种基于互信息的药品不良反应信号检测方法[J]. 南京大学学报(自然科学版), 2010, (06): 705-712.

[11] Vilar S, Harpaz R, Chase H S, et al. Facilitating adverse drug event detection in pharmacovigilance databases using molecular structure similarity: application to rhabdomyolysis[J]. Journal of the American Medical Informatics Association, 2011, 18(Supplement 1): i73-i80.

[12] Nikfarjam A, Gonzalez G H. Pattern mining for extraction of mentions of adverse drug reactions from user comments[C] AMIA Annual Symposium Proceedings. American Medical Informatics Association,2011: 1019.

[13] Yates A, Goharian N. ADRTrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites[M]. // Advances in Information Retrieval. City: Springer, 2013: 816-819.

[14] Bian J, Topaloglu U, Yu F. Towards large-scale twitter mining for drug-related adverse events[C] Proceedings of the 2012 international workshop on Smart health and wellbeing. ACM: 25-32.

[15] Kuhn M, Campillos M, Letunic I, et al. A side effect resource to capture phenotypic effects of drugs[J]. Molecular systems biology, 2010, 6(1): 343.

[16] Kilicoglu H, Fiszman M, Rodriguez A, et al. Semantic MEDLINE: a web application for managing the results of PubMed Searches[C] Proceedings of the third international symposium for semantic mining in biomedicine. Citeseer,2008: 69-76.

[17] Rindflesch T C, Kilicoglu H, Fiszman M, et al. Semantic MEDLINE: An advanced information management application for biomedicine[J]. Information Services and Use, 2011, 31(1): 15-21.

[18] Kilicoglu H, Shin D, Fiszman M, et al. SemMedDB: a PubMed-scale repository of biomedical semantic predications[J]. Bioinformatics, 2012, 28(23): 3158-3160.

[19] Rindflesch T C, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text[J]. Journal of biomedical informatics, 2003, 36(6): 462-477.

[20] Lipscomb C E. Medical subject headings (MeSH)[J]. Bulletin of the Medical Library Association, 2000, 88(3): 265.

[21] Law V, Knox C, Djoumbou Y, et al. DrugBank 4.0: shedding new light on drug metabolism[J]. Nucleic acids research, 2014, 42(D1): D1091-D1097.

[22] Knox C, Law V, Jewison T, et al. DrugBank 3.0: a comprehensive resource for ‘omics’ research on drugs[J]. Nucleic acids research, 2011, 39(suppl 1): D1035-D1041.

[23] Wishart D S, Knox C, Guo A C, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets[J]. Nucleic acids research, 2008, 36(suppl 1): D901-D906.

[24] Wishart D S, Knox C, Guo A C, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration[J]. Nucleic acids research, 2006, 34(suppl 1): D668-D672.

[25] Aronson A R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program[C] Proceedings of the AMIA Symposium. American Medical Informatics Association: 17.

[26] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C] Advances in Neural Information Processing Systems. 3111-3119.

[27] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:13013781, 2013.

[28] Mikolov T, Yih W-T, Zweig G. Linguistic Regularities in Continuous Space Word Representations[C] HLT-NAACL. 746-751.

[29] Girault E M, Toonen P W, Eggels L, et al. Olanzapine-induced changes in glucose metabolism are independent of the melanin-concentrating hormone system[J]. Psychoneuroendocrinology, 2013, 38(11): 2640-2646.

[30] Kusumi I, Honda M, Uemura K, et al. Effect of olanzapine orally disintegrating tablet versus oral standard tablet on body weight in patients with schizophrenia: a randomized open-label trial[J]. Progress in Neuro-Psychopharmacology and Biological Psychiatry, 2012, 36(2): 313-317.

[31] Opp D, Hildebrandt C. Olanzapine-associated type 2 diabetes mellitus[J]. Schizophr Res, 2002, 56(1): 195-196.

[32] Sa Y K, Yang H, Jung H K, et al. Olanzapine-induced diabetic ketoacidosis and neuroleptic malignant syndrome with rhabdomyolysis: a case report[J]. Endocrinology and Metabolism, 2013, 28(1): 70-75.

[33] Tripathi P, Agrawal H, Goyal P, et al. Olanzapine-induced neuroleptic malignant syndrome in a patient with bipolar affective disorder: Does quetiapine holds the solution?[J]. Ind Psychiatry J, 2013, 22(2): 159-160.

[34] Saritas T B, Cankaya B, Yosunkaya A. Olanzapine-Induced Malignant Neuroleptic Syndrome[J]. Turkish Journal of Anesthesia and Reanimation, 2014, 42(5): 288-291.

[35] Endoh I, Kodani E, Yoshikawa M, et al. Olanzapine-Related Life-Threatening Hyperosmolar Hyperglycemic Syndrome: A Case Report[J]. Journal of clinical psychopharmacology, 2012, 32(1): 130-132.

[36] Malhotra N, Grover S, Chakrabarti S, et al. Metabolic syndrome in schizophrenia[J]. Indian journal of psychological medicine, 2013, 35(3): 227.

[37] Khalil R B. Atypical Antipsychotic Drugs, Schizophrenia, and Metabolic Syndrome in Non-Euro-American Societies[J]. Clinical neuropharmacology, 2012, 35(3): 141-147.

[38] Gingrich J A. Mutational analysis of the serotonergic system: recent findings using knockout mice[J]. Current Drug Targets-CNS & Neurological Disorders, 2002, 1(5): 449-465.

[39] Goldman D, Oroszi G, O'malley S, et al. COMBINE genetics study: the pharmacogenetics of alcoholism treatment response: genes and mechanisms[J]. Journal of Studies on Alcohol and Drugs, 2005, (15): 56.

[40] Shishkina G, Kalinina T, Dygalo N. Attenuation of α 2A-adrenergic receptor expression in neonatal rat brain by RNA interference or antisense oligonucleotide reduced anxiety in adulthood[J]. Neuroscience, 2004, 129(3): 521-528.

攻读硕士学位期间发表学术论文情况

1 面向社交网络的潜在药物不良反应发现. 赵明珍，林鸿飞，徐博，郝辉辉. 中文信息学报（录用）. 主办单位：中国中文信息学会、中国科学院软件研究所。中文核心期刊。（本硕士学位论文第四章）

2 一种面向大数据的潜在药物不良反应数据挖掘方法.林鸿飞，赵明珍. 2015 年 3 月 3 日，专利申请号：201510093861.3。（本硕士学位论文第二章）

致　　谢

三年的研究生时光匆匆流走，面试、入学、迎新晚会等场景至今仍历历在目，记忆犹新。三年前，我怀着激动和兴奋的心情来到大连理工大学，走进信息检索研究室，渴望在这里汲取知识的养分，充实自己，提高自己的能力；三年后，我学满毕业，即将离开这座美丽的校园，心中的不舍和留恋唯有即将离开的人才可以体会。在这即将离开的时刻，我怀着难舍的心情，对所有曾经帮助过我、陪伴过我的人表达深深的谢意。

首先我要感谢恩师林鸿飞教授。林老师知识渊博，治学严谨，高瞻远瞩，带领整个实验室走在科研的前沿；在三年的时光里，林老师对我进行了悉心的指导，使我的知识和科研能力得到很大提升。除此之外，林老师积极组织实验室的课外活动，每年的元旦晚会、徒步走活动以及羽毛球比赛成为信息检索实验室的文化标志，在这些活动中，我们不仅增进了师生之间的感情，也收获了快乐。在林老师的熏陶下，自己也喜欢上了羽毛球这项运动。总之，林老师平易近人、豁达乐观、积极进取的风范是我毕生的楷模。

其次，感谢杨志豪老师、王健老师和张益嘉老师，三位老师在生物组组会上给了我们很多指导，对我的成长和学习给了很大帮助；感谢许侃老师一直为维持实验室的日常运转所做的努力，在许老师的带领下，实验室各项活动得以顺利进行。

再次，感谢徐博师姐在科研上对我的帮助；感谢杨亮师兄和刘文飞师兄，两位师兄在实验室工作中给了我很多指导和建议，从两位师兄身上学到很多知识，而且他们踏实认真的工作态度深深地影响了我；感谢马云龙师兄在专利申请方面给予的帮助，马师兄积极负责的态度值得我在今后的工作中学习。感谢赵哲焕师兄在深度学习方面给予我的指导；感谢程亮喜师兄、李浩瑞师兄和李宗耀师兄在科研方面给予的帮助。

然后，感谢杨阳、郝辉辉、杨娅等 IR2012 级全体同学的陪伴和帮助，尤其感谢杨阳同学在研究生生活和学习中对于我的帮助和建议，使我学到很多前沿的知识。

最后，我要感谢我的家人，在漫长的求学过程中，他们给了我巨大的支持和鼓励。

即将毕业之际，感谢所有曾经帮助过我、支持过我、陪伴过我的人，祝你们身体健康，天天快乐。谢谢你们。

大连理工大学学位论文版权使用授权书

本人完全了解学校有关学位论文知识产权的规定，在校攻读学位期间论文工作的知识产权属于大连理工大学，允许论文被查阅和借阅。学校有权保留论文并向国家有关部门或机构送交论文的复印件和电子版，可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印、或扫描等复制手段保存和汇编本学位论文。

学位论文题目：_____

作 者 签 名：_____ 日期：____年____月____日

导 师 签 名：_____ 日期：____年____月____日