

SHANGHAIUNIVERSITY

毕业设计（论文）

**UNDERGRADUATEPROJECT(THESIS)**

**题目:聊天机器人**

|  |  |
| --- | --- |
| **学院** | **计算机工程与科学学院** |
| **专业** | **计算机科学与技术** |
| **学号** | **14122151** |
| **学生姓名** | **王晨悦** |
| **指导教师** | **戴佳筑** |
| **起讫日期** | **2018.01.29 – 2018.06.08** |

目录

[摘要 III](#_Toc514835253)

[ABSTRACT IV](#_Toc514835254)

[第1章 绪论 1](#_Toc514835255)

[§1.1 聊天机器人制作的背景及意义 1](#_Toc514835256)

[§1.2 聊天机器人技术发展现状 1](#_Toc514835257)

[§1.2.1 机器学习算法 1](#_Toc514835258)

[§1.2.2 聊天机器人技术发展现状 2](#_Toc514835259)

[§1.3 聊天机器人的使用 2](#_Toc514835260)

[§1.4 本文组织结构 2](#_Toc514835261)

[第2章 单词的向量表示 4](#_Toc514835262)

[§2.1 SKIP-GRAM概述 4](#_Toc514835263)

[§2.2 代码及训练 6](#_Toc514835264)

[§2.2.1 代码 6](#_Toc514835265)

[§2.2.2 训练过程 6](#_Toc514835266)

[§2.2.3 训练结果 7](#_Toc514835267)

[§2.3 本章小结 9](#_Toc514835268)

[第3章 基于LSTM构建机器人 10](#_Toc514835269)

[§3.1 RNN及LSTM概述 10](#_Toc514835270)

[§3.2 代码及训练 10](#_Toc514835271)

[§3.2.1 代码 11](#_Toc514835272)

[§3.2.2 训练过程 11](#_Toc514835273)

[§3.3 本章小节 12](#_Toc514835274)

[第4章 图形界面搭建 17](#_Toc514835275)

[§4.1 网站整体设计 17](#_Toc514835276)

[§4.2 开发环境简介 18](#_Toc514835277)

[§4.3 代码实现 22](#_Toc514835278)

[§4.4 本章小结 24](#_Toc514835279)

[第5章 总结与展望 26](#_Toc514835280)

[§5.1 本文总结 26](#_Toc514835281)

[§5.1.1 本文的主要工作 26](#_Toc514835282)

[§5.1.2 本文的主要创新点 26](#_Toc514835283)

[§5.2 展望 26](#_Toc514835284)

[致谢 28](#_Toc514835285)

[参考文献 29](#_Toc514835286)

[附录：部分源程序清单 31](#_Toc514835287)

聊天机器人

# 摘要

XXX。

关键词：XXXX，XXX，XXX，XXX

Chat Robot

# ABSTRACT

In recent years, with the rapid development of Micro-blog, the need that users gain the access to information is also a linear growth momentum. The amount of Sina Micro-blog registered users has reached 503 million by 2012. However, the daily flood of Micro-blogs has a serious impact on the quality of information users receive. Thus, how to find content that they are interested in quickly and accurately? Or can we push the information according to the user's interest actively? That is what this paper concerns.

User model is a formal description of users' interests. To establish a precise user model for Micro-blog users, in order to recommend the information they concerned about and contents they are interested in, is the development trend of personalized recommendation. The achievements of this paper can be used for e-commerce, public opinion monitoring, advertising and other areas.

In this paper, Micro-blog user model construction method based on the ontology technology is discussed. First of all, according to each user's micro-blog content analysis, extract the keywords which represent the content information of each Micro-blog; then create user's eigenvectors in order to calculate the user's interest degree; finally match these keywords with the ontology library to create Micro-blog user interest model. The achievements of this paper will lay the foundation of personalized service based on Micro-blogs.

**Keywords:** Ontology, Micro-blog, User model, Interest Degree

# 绪论

本章主要描述了聊天机器人构建的背景、意义等，研究了国内外在机器学习领域科学进展现状，进而调查分析国内外基于机器学习在聊天机器人领域的发展现状，并在此基础上，依托tensorflow框架，运用rnn算法搭建了一个聊天机器人。

## 聊天机器人制作的背景及意义

聊天机器人在本文中是指一个可以根据用户的输入来提供输出的程序，可以根据用户的提问作出一些智能的回答，或者跟用户进行简单的聊天交流，基本概念是模拟人类的对话。聊天机器人的概念诞生于二十世纪四十年代，经过长时间的研究和发展。在几十年的发展过程中，人类对自然语言处理的算法也不断发展，这给聊天机器人的发展提供了必备的基础，也顺理成章地推动了聊天机器人的智能程度变高。但是由于自然语言经过更漫长的时间进化，很难总结出固有的规律，语法也繁杂难以列举，更有很多的习惯用法，使得计算机程序模拟人类对话的难度变得很高。而近年来机器学习算法越来越火热，也给了聊天机器人的发展提供了一条更宽敞的路。机器学习的存在，使得聊天机器人的对话产生并不需要人为总结所有的语法规律，而是可以通过不断地训练，让程序本身就有应有的智能。结合机器学习算法，让计算机程序具备自己的智能，实现模拟人类对话，这就是本文的理论支撑点。

而在应用方面，聊天机器人已经在很多领域创造出了自己的价值。在通讯平台、教育等方面起到了很重要的作用，很多公司也开始使用较成熟的聊天机器人实现简单的客服等工作，而一些大公司，则结合更多的类似语音识别、图像识别技术，创造出能接收更多种形式的输入的机器人，帮助很多需要的人也创造了很大的价值。

## 聊天机器人技术发展现状

从聊天机器人的概念诞生以来，就有大量学者不断地进行研究钻研，近年来，随着机器学习算法变得更成熟和便于应用，聊天机器人的发展也开始走向一个高潮。接下来先具体阐述本文将应用的机器学习算法相关基础，然后阐述聊天机器人的发展历程和现状以及实际各种方面的应用。

### 机器学习算法

机器学习是人工智能的一个分支

### 聊天机器人技术发展现状

聊天机器人的发展我们从1966年的ELIZA说起，由一位MIT的教授joseph所发明。最初创造ELIZA的目的是想让它作为一个陪伴人们聊天的工具，本身ELIZA的功能也只是简单的陪伴。但是事实证明，ELIZA的效果非常好，有聊天记录显示，曾经有来访者跟ELIZA聊天以后发出这样的感慨“你在某些方面挺想我父亲的”，这就是ELIZA的厉害之处。ELIZA的实现并没有涉及很复杂的技术，甚至没有最简单的语义理解，只是简单的将输入和数据库进行匹配，处理的内容也只是文本。ELIZA的工作原理是这样的，它将输入去与关键词匹配，大概提供了两百多个模板，每次匹配到一个或多个关键词就根据相应的模板返回内容。如果没有匹配到关键词，则简单地把输入中的“我”改成“你”返回，这样的规则虽然简单但是非常有效。ELIZA的成功在我看来主要得益于它的规则，一是非常有效的模板和关键词设置，第二是对于没有匹配到关键词的输入的输出规则。而这样的规则，就成了很多现在的产品的基石，比如ALICE、MITSUKU、机器人小冰等，区别就是会增加或者更改更好的模板，以满足更多的功能。但是这样做的缺点很明显，模板数量是有限的，而语言的创造力是无限的，在不理解语义的情况下，仅仅根据模板去回答问题，会非常受限于模板和性能。

之后的机器人就开始对语义进行分析，比如SIRI、MBI的WASTON等等。

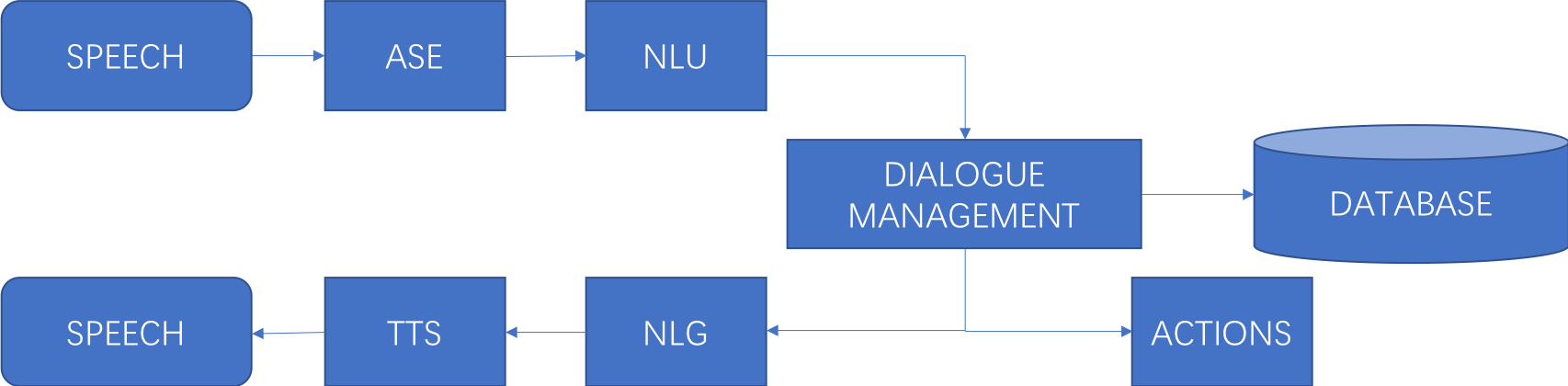
而随着科学发展，我们能收集到的数据量已经大量增加，而处理数据的能力也大大增强，在此基础上，机器人不需要限制在固定的模板上。

总结上面的发展，我们可以看到机器人的架构发生了如下的变化：

最古老的一种，只能处理文本，处理过程是简单的匹配，如图：



下一个阶段出现更多的是这样的架构，它可以在一定程度上理解自然语言，能执行一些比较复杂的动作，之后根据对话管理系统生成一些回答并返回，如图：



而现代的机器人框架则更多结合机器学习算法



旨在构建一个足够强大的神经网络，通过大量的训练来实现智能回答。

## 聊天机器人的使用

我们很容易关注到今年来，很多公司都开始使用聊天机器人，这代表了一种发展趋势，使用更多智能产品来代替人类的工作，解放生产力的同时，提高某些方面的用户体验以及更好地为人类服务。在国内，微信在2016年就开始使用聊天机器人，虽然是处于实验状态，并在中途曾关闭了机器人账号，但随着技术的发展，现状的机器人已经很成熟了。QQ小冰也是另一个例子，不过很显然也是一个实验的状态。百度在人工智能方面投入很大的成本，刚刚上线的APP“简单搜索”也运用了部分聊天机器人的功能，来实现语音的输入到搜索结果的产生的流程优化，提高用户体验。而在国外，聊天机器人的运用则更为领先，谷歌公司一直在人工智能、深度学习等方面有着很好的研究。这几年，无人车技术也开始成为大头公司争夺的高地之一，在无人车上也运用了很多聊天机器人相关技术，用以方便用户的使用。在教育方面，聊天机器人也开始发力，对学习语言的同学来说肯定不陌生翻译器，各种翻译程序在聊天机器人发展迅速的今天，效率和准确度大大提升，而更多的产品走进某些领域的简单应用。因为聊天机器人不需要像人类一样休息，同时在具备大量的训练样本的条件下，可以根据训练样本训练成为一个某领域的“专家”，那么如果在经过样本的提炼和安排之后，可以说各个领域，只要涉及信息的输入和查询的地方，都可以运用聊天机器人来代替人类的繁复工作。

## 本文组织结构

整篇论文分为五章。

第一章介绍了聊天机器人的研究背景、研究意义，引出了聊天机器人实现背后的技术点，并介绍了相关技术的概念和发展。

第二章主要介绍了聊天机器人实现的第一步，对训练样本的选取和处理的过程，包括处理技术的介绍和核心代码部分，总结了需要提高的点。

第三章主要介绍了将第二章提到的训练样本集训练的响亮结果，存储为二进制文件，并传入LSTM算法进行训练的过程，这一步是整个聊天机器人训练和生成的核心步骤，在第三章的训练之后，聊天机器人就具备了自己的智能，可以接受训练集意外的的句子，并通过计算得到合理的输出。

第四章是为聊天机器人搭建了一个图形界面，介绍了主要的框架和服务器环境等等，通过网页的形式展现给用户。便于用户操作和测试。

第五章总结全文，阐述了本文的主要工作难点、重点和创新点，也总结了需要进一步研究和改进的点。

# 单词的向量表示

本章具体描述了基于维基百科的本体构建方法：介绍了本体的基本概念，并引出本文所使用的基于维基百科的本体构建方法。

## SKIP-GRAM概述

本节介绍了本体的基本概念以及目前研究学者常用的四种构建方法。

在计算机领域，1991年开始，研究者们对本体做了多次说明，它表示的含义也更加清晰明确，现在人们一般认为本体论是对概念化对象的明确表示和描述[14]。随着研究者们对本体研究的不断完善，本体的定义有很多种，以下是几种比较有代表性的定义。

1991年，Neches[15]等给出了构成相关领域词汇的基本术语、关系，以及这些词汇外延的规则。1996年，Swartout[16]提出本体是一个知识库结构中术语集合，该结构中的术语是按照继承关系组织起来的，强调了本体中术语(Terms)的重要性。

1993年，Gruber[17]提出本体是概念模型明确的规范说明。1997年，Borst[18]提出本体是共享概念模型的形式化规范说明。1998年，Studer[19]等对上述两个定义进行了深入研究，认为本体是“共享概念模型的明确的形式化规范说明”，它有以下四个方面的含义：

（1）本体是一个概念模型(Conceptualization)，它是指通过抽象客观世界中一些现象的相关概念得到的模型。它表现的含义独立于具体的环境状态；

（2）本体的明确性(Explicit)，本体包含的概念和概念之间的约束都应该有明确的定义；

（3）本体是形式化的(Formal)，意思是本体应该是计算机可处理的；

（4）本体是可共享的(Sharable)，本体中概念、关系、属性的描述是基于标准的、规范的、能被共享的。

本体的定义多种多样，其核心都一样，把本体作为一种描述资源的手段，为不同的主体进行知识交流提供语义基础。在计算机领域中，对本体的研究主要是如何实现这种统一标准，以及本体的构建方法、本体描述语言、本体的管理和本体的应用等内容。

根据本体的语义特性，本体可以应用于不同的领域，致力于提高服务的联想能力和准确性。

总而言之，尽管本体的定义方式多种多样，但本体所包含的基本要素：概念、概念之间的关系等。

目前，本体构建成功的案例很多，根据不同的领域，构建的方法也不一样，现在还没有构建本体的标准。许多研究人员根据经验总结出来了一些方法，1995年，Gruber提出构建本体的五条规则如下：

（1）明确性和客观性：本体应该是背景独立的、客观的，能反映社会真实情况，满足可计算性，具有明确的、客观的形式化语义；

（2）完整性：给出的定义应该是完整的，能表达特定属于的含义；

（3）一致性：只是推理产生的结论与属于本身的含义不产生矛盾；

（4）可扩展性：在扩展本体功能的时候，可以自由添加新的术语而对已有本体的结构和内容不做修改；

（5）最少约束：在满足可能的知识共享需求的基础上本体的约定应该最小。它可以通过只定义通讯所需的词汇或者定义约束最弱的公理来保证。

目前比较普遍的构造特定领域的本体，一般都需要相关领域专家的参与。以下列举了一些在项目实践过程中形成的方法：

IDEF-5方法：IDEF(ICAM Definition Languages)方法是上世纪七十年代由美国空军发明的。在1981年针对集成计算机辅助制造(Integrated Computer Aided Manufacturing，简称ICAM)项目中用于描述企业内部运作的建模方法。最初该方法只是应用于制造业，经过改造后，适用于软件开发。目前已经形成了一系列方法。包括IDEF1X和IDEF0到IDEF14共16套方法，每一套方法都通过建模程序获取某特定类型信息，其中IDEF-5是本体描述获取语言。

骨架法(Skeletal Methodology)：该方法是由爱丁堡大学人工智能应用研究所开发企业建模过程中总结出来的。

评价法：该方法由多伦多大学企业集成实验室，在开发虚拟企业本体工程项目时总结出来的。通过建立制定知识的逻辑模型，用一阶逻辑构造形式化的模型，包括企业设计本体、工程本体、计划本体和服务本体。

七步法：该方法是斯坦福大学医学院提出的基于Protégé本体构建工具的一种领域本体构建方法。一共包括七个步骤：

（1）确定只是本体的专业领域和范畴；

（2）考察复用现有只是本体的可能性；

（3）列出本题中的重要术语；

（4）定义类和类的层次体系；

（5）定义类的属性；

（6）定义类的分面(Facets)；

（7）创建本体实例。

## 代码及训练

以上传统的构建本体的方法仍然存在许多弊端，下面就针对这些不足进行分析阐述，并提出本文所使用的基于维基百科的本体构建方法的优势所在。

### 代码

维基百科(Wikipedia)是一个自由、免费、内容开放的网络百科全书，参与者来自世界各地。这个站点使用Wiki，这意味着任何人都可以编辑维基百科中的任何文章及条目。

维基百科是一个基于Wiki技术的全球性多语言百科全书协作计划，同时也是一部用不同语言写成的网络百科全书，其目标及宗旨是为全人类提供自由的百科全书──用他们所选择的语言来书写而成的，是一个动态的、可自由访问和编辑的全球知识体。

维基百科自2001年1月15日正式成立，由维基媒体基金会负责维持，其大部分页面都可以由任何人使用浏览器进行阅览和修改。因为维基用户的广泛参与共建、共享，维基百科也被称为创新2.0时代的百科全书、人民的百科全书。这本全球各国人民参与编写，自由、开放的在线百科全书也是知识社会条件下用户参与、大众创新、开放创新、协同创新的生动诠释。英语维基百科的普及也促成了其它计划，例如维基新闻、维基教科书等计划的产生，虽然也造成对这些所有人都可以编辑的内容准确性的争议，但如果所列出的来源可以被审察及确认，则其内容也会受到一定的肯定。维基百科中的所有文本以及大多数的图像和其他内容都是在GNU自由文档许可证下发布的，以确保内容的自由度及开放度。所有人在这里所写的文章都将遵循copyleft协议，所有内容都可以自由的分发和复制。

截至2013年1月，维基百科条目数第一的英文维基百科已有415万个条目，而全球所有282种语言的独立运作版本共突破2100万个条目，总登记用户也超越3200万人，而总编辑次数更是超越12亿次。大部分页面都可以由任何人使用浏览器进行阅览和修改，英文维基百科的普及也促成了其它计划。

Wiki一词来源于夏威夷语的“wee kee wee kee”，原本是“快点快点”的意思。在这里“WikiWiki”指一种超文本系统。这种超文本系统支持面向社群的协作式写作，同时也包括一组支持这种写作的辅助工具。

### 训练过程

本体的构建是一个系统性工程，由于没有统一的构建原则、方法，到目前为止，本体工程仍处于相对不成熟的阶段，整个建设过程在以下几个方面还存在很多问题[20]。

（1）构建方法

7种构建方法都是从具体领域本体的开发中总结出来的，应用领域很有限，大多数方法的细节比较粗，相关技术比较少，没有一种方法完全按照生命周期法进行开发。另外，由于没有统一的构建原则作为指导，整个本体的构建过程难以进行规范的管理。

（2）可扩展性

随着领域的不断发展、变化，必然会有更多的领域相关概念和关系引入到核心本体中，需要通过知识的进一步获取、概念的进一步扩充或更改等方式，不断改进和扩展领域本体。但是目前本体的维护和扩展问题还没有得到很好的研究和支持。

（3）共享和重用

领域本体构建的目的是为不同系统提供彼此交流的语义基础。目前，为减少构建本体的工作量，多数的研究均利用现有术语丰富的叙词表作为构建工作的起点。而叙词表到本体的转换还没有一个统一的标准，在转换的过程中各项目使用的描述语言以及描述广度和深度不尽相同，给今后本体之间语义互操作和重用造成了困难。

### 训练结果

维基百科全书于2001年1月投入运行，到2013年1月，英文版的维基百科全书就包含了415万多个条目，全部条目超过了2100万条，具有内容相互独立的282种语言版本。维基百科全书具有与大英百科全书类似的写作风格，相近的准确性，但内容的丰富性已经超过了大英百科全书。其内容不仅包括大量的插图，还包含了大量的多媒体内容，时效性更是印刷版的大英百科全书无法比拟的。与传统百科全书相比，维基百科特点突出：

（1）首先，维基百科始终将自己定位为包含人类所有知识领域的百科全书，而不仅仅是一本词典、在线论坛或其他。

（2）其次，计划也是一个Wiki，允许大众广泛参与。维基百科是第—个使用Wiki系统进行百科全书编撰工作的协作计划。

（3）最后，维基百科是一部内容开放的百科全书，其内容允许任何第三方不受限制地复制、修改及再发布，广泛的参与性，使得维基百科的权威性、中立客观性得到了保证。它方便不同行业的人士寻找知识，而使用者也可以不断增加自己的知识从而充实自己。当然，因为一些特殊的原因，目前中文版和英文版和百科全书也有部分敏感词汇被禁用。

目前，维基百科定期免费提供各个语言版本的所有数据，放到网络上供人下载。其中，数据包中包括有page.sql文件和categorylinks.sql文件，前者记录了所有页面的基本信息，如：页面标题、命名空间、页面长度等，后者记录了各个页面标记的类目信息。[21][22]虽然维基百科网站也提供了树状形式的类目网络供人浏览，但是这个网页的数据存在如下缺陷：

（1）由人工维护，所以数据更新比较缓慢；

（2）迫于部分类目深度过深，所以这个网页把这些类目删除掉，如“生物分类树”；

（3）由于采用简单的深度遍历算法将图输出为树状结构，所以部分类目深度失衡，例如“自然科学”类目就没有子类目。

基于以上原因，我们编程实现了类目网络的自动生成。基于维基百科中文本体的构建主要包括以下几个步骤：

（1）中文繁简转换

目前中文存在两种书写系统——繁体中文与简体中文。一般来自台湾、香港、澳门的使用者使用正体中文（繁体中文），来自中国大陆、新加坡、马来西亚则使用简体中文。[23]作为一个全球华人共同创作的平台，中文维基百科发布的数据中，既有繁体形式的，也有简体形式的，甚至很多在同一篇文章中繁简夹杂。这给我们的抽取工作带来很大的不便：一方面，文本繁简混杂的问题使得我们不能用现有的基于单一文字模式的中文信息处理工具直接分析文本；另一方面，繁简夹杂必然使得我们的研究成果不能很好地得到利用。所以，我们在利用维基百科所提供的简繁对应词表基础上，借鉴MediaWiki 1.4的繁简转换功能的“用字模式”，实现了繁简转换功能。

（2）类目网络清理

在维基页面分类系统中，存在着若干为了方便管理而添加的元类目，例如：“维基百科站务”。[24]因为这些类目所含语义信息较少，所以我们必须清理这些类目。我们剔除所有包含以下关键字的类目：维基，列表，模板，维基人，专题，分类，条目，小作品。另外，从数据库自动生成的类目网络中存在一些孤立点，我们将此类类目也全部清除。在清理之前，中文维基类目之间的直接连接数为21776个，清理后的直接连接关系总数14009个。

（3）识别父子关系

首先界定两个概念：下位词与“Is-a”关系（父子关系）[25][26]。语言学家Fromkin和Rodman[27]认为，下位词是一个一般化词语具体化之后的相关词语集合。例如，深红色，朱红，绯红色都是红色的下位词，而红色就是它们的上位词。同时，红色又是颜色的下位词。因此，下位关系也就是一般化术语（如多边形）和它的具体化实例（如三角形）之间的关系。在计算机科学中，常常将此关系称为Is-a关系[28]。例如，用“红色is a颜色”来描述红色和颜色之间的下位关系。在知识表示和面向对象编程与设计中，在A is a B中，Is-a表示类A是类B的子类，即B是A的父类。换言之，"A is a B"通常意味着概念A是概念B的具体化，概念B是概念A的一般化。举例而言，“水果”是“苹果”、“桔子”、“芒果”等概念的一般化。我们可以说，“苹果is a水果”。下文对这些类目之间的Is-a和Not Is-a关系进行自动识别。

经过以上步骤可以建立一棵基于维基百科的概念知识树，这就是我们维基百科本体知识树的原型。

## 本章小结

本体库的构建是微博用户模型构建的基础，基于所构建完成的类目网络结构，可以生成所有类目的树形结构本体，进而为解决知识的共享和重用问题提供了新思路，也成为当前信息科学领域的研究热点之一。由于传统本体构建方法存在诸多问题，本文中利用维基百科这一新的互动开放的信息交流平台，构建出中文本体库，为后续进行微博用户模型的构建打下了基础。但本体库的完备性将对用户模型构建的准确性影响较大，因此对本体的研究工作仍任重而道远。

# 基于LSTM构建机器人

本章是全文的重点章节，全面阐述了本文的工作内容。首先简单介绍了用户模型的概念及其表示方法，随后将本文的构建过程完整的进行了描述，并将部分系统功能运行的结果作了展示。

## RNN及LSTM概述

在基于内容的个性化推荐中，文本内容是主要的研究对象。自然语言所表示的文本内容需要转化为机器能够识别和可计算的模型才能进行进一步的研究。用户模型并不仅仅是对用户兴趣的准确描述，由于可计算性是它对用户模型的基本要求，也就是说，用户模型不是对用户个体的一般性简单描述，而是一种面向算法的，具有特定数据结构，形式化的用户兴趣描述，它是实现个性化服务的基础和核心。

根据Gerhard Fischer[29]的论述，可以认为广义上的用户模型有如下三种：

（1）用户头脑中的概念模型。这是用户头脑中关于计算机系统及其所应具有的功能的模型，表示了用户对计算机系统的理解和期望，该模型随着用户使用系统经验的增加而不断完善；

（2）设计者的用户模型。设计者头脑中关于用户的模型，是设计者对用户特征的描述，被设计者用来作为系统设计的基础；

（3）计算机系统的用户模型。它是由设计者在设计阶段依据设计者的用户模型用计算机软件构造的，在系统的运行过程中实现的。

由此可以看出，用户建模是这样一个过程：设计者根据用户概念模型调整设计者用户模型，将设计者用户模型用软件的方法转换为计算机系统的用户模型，用户建模的最终目标是计算机系统模型。狭义上的用户模型是指软件系统的用户模型，我们更关心的也是计算机所拥有的关于用户特征的模型。

本文的用户模型是一种对于用户兴趣内容特征的描述和表达，可以收集并提取用户的兴趣偏好，并与本体库进行匹配，进而更好地理解用户的需求和任务，实现个性化的推荐服务。

## 代码及训练

用户模型表示是用户建模的基础，决定了用户模型反映用户信息的能力和可计算能力。本节将讨论用户模型表示的有关方法。

### 代码

用户模型表示方法有很多种，目前使用较为广泛的有关键词表示法、主题表示法和向量空间模型表示法等。

关键词表示法是以用户感兴趣的一组关键词来表示用户模型。例如{篮球，火箭，后卫}。关键词可以由用户自己设定，也可以通过用户的行为、他们在网上留下的信息及其它特征来获取。

主题表示法是关键词表示法的一种改进，这种表示法是以用户感兴趣的信息的主题来表示用户模型的。例如用户对体育、文化、科技感兴趣，则该用户的用户模型表示为{体育，文化，科技}。

向量空间模型表示法则是利用特征词和相应的权值向量来表示用户模型。其基本思想是根据用户感兴趣的文档中各个关键词的出现频率建立特征词及其相应的权值向量来表示用户模型。

虽然存在多种不同的方法来表示用户模型，但常用的用户模型表示方法普遍存在以下两大问题：

（1）缺乏统一标准。仅能被特定的系统所运用，不能在不同系统间实现共享。

（2）缺乏领域知识支持。大部分的表示方式只是对用户兴趣的简单罗列，没有考虑用户兴趣间的关联关系，难以表达用户兴趣的语义内容。

### 训练过程

基于本体的用户模型表示方法使用规范的结构模式描述用户兴趣，把用户兴趣与领域知识的语义概念层次相结合，具有强大的语义表达能力、兼容性和可扩展性，可以很好地解决常用用户模型存在的问题。

本文中使用的本体用户模型是一棵带权的结点树，如**错误!未找到引用源。**所示，图中两结点之间存在父子关系，比如Sports是Football的父节点。节点旁边的数字则是每个兴趣节点的兴趣度大小，也就是用户对这个兴趣的感兴趣程度。

……

……

0.17

0.17

0.58

0.61

0.22

0.81

0.78

0.34

图3‑1本体用户模型兴趣树

基于本体的用户模型表示方法主要有以下几个优势：

（1）充分描述用户兴趣的语义。

用户兴趣中的一个词条往往包含丰富的语义。该方法能描述该词条对应的用户兴趣的层次概念，结合领域本体，使兴趣表示带着丰富的语义。

（2）兼容性和可扩展性。

该方法具有很好的兼容性和可扩展性。根据该方法建立的模型可以方便地移植到其它系统中，也可随着应用的发展，对模型进行扩展。

（3）自适应性。

随着现实世界知识体系的变化以及用户兴趣的变化，该方法能够自动适应这些变化，准确表达用户的当前兴趣。

可以说，基于本体的用户模型表示方法是用户模型表示方法的发展方向，因此本文将实现基于本体的用户模型构建。

## 本章小节

微博用户模型的构建主要包括特征词提取、兴趣度计算以及用户模型的生成。以下将分为三个小节来详细描述构建方法。

首先，我们将获取微博用户信息，理想的用户建模方法应该无需用户主动提供任何信息。自动用户建模就是根据用户在微博平台上注册留下的信息及用户发布在微博平台上的微博信息来构建用户模型。我们可以通过微博平台提供的API，获取到用户的昵称、密码、编号、姓名、关注数、粉丝数、发表微博的数量等个人基本信息以及每个用户在平台上发布过的所有微博信息内容来构建用户模型。

在中文信息处理领域，对中文自动分词的研究已较为成熟，经典的分词方法主要有最大向前匹配法、逐词遍历匹配法、最小向前匹配法等。本文未对分词方法进行深入研究，而是选择直接使用开源的中国科学院计算技术研究所研制的汉语词法分析系统ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System)[31]，该分词系统分词准确率达到97%以上。本文实现的本体构建系统中使用ICTCLAS汉语分词系统官网提供的32位Windows操作系统下的JAVA版本的中文分词和词性标注。

计算机不具有人类的智能，不能像人类一样阅读微博后根据自身的知识和理解能力对微博内容产生理解。因此，在进行微博语义扩展之前首先要将微博转换成易被计算机理解和识别的结构形式。微博的表示要求能够准确有效的表达微博内容，并且还要易于计算机处理。

目前，典型的文本表示方法主要有：布尔模型（Boolean Model）、向量空间模型（VSM）、语言模型（Language Model）、潜在语义索引（LSI, Latent Semantic Indexing）[32]和概率检索模型（Probability Model）[33][34]。这些模型从不同的角度出发，使用不同的方法标注特征词权重和相似度计算等问题。

向量空间模型由Salton等人于20世纪70年代提出，并成功地应用于著名的SMART文本检索系统。VSM被广泛应用于文本分类、文本聚类、信息检索等领域。近年来，在文本挖掘领域向量空间模型已经成为最常用的文本表示方法。VSM是基于这样一个关键假设下提出的，即文档中各词条出现的先后顺序是无关紧要的，每个特征词对应特征空间的一维，他们每一维对于判定文档所属的类别所起的作用是相互独立的。因此，可以把一篇文档看成是一系列无序词条的集合，从文档中选取出n个特征词来表示文本就是形成一个n维向量空间。例如一篇文档中选取三个特征词t1、t2、t3，那么这篇文档就表示为。但是对于整个文档来说，每个特征词对文本的重要程度不同，因此，需要对每个特征词赋予一定的权重。一篇具有n个特征词的文档利用VSM表示方法就可以表示为公式（1）。

 （1）

其中，di是第i篇文档，tij表示第i篇文档的第j个关键词，wij是第i篇文档的第j个关键词权重。关于权重的计算方法有很多种，将在3.3.2兴趣度计算步骤中再作介绍。本文中微博文本表示采用VSM方法。

经过文本分词处理后，需要抽取一定数量的特征词作为向量的各维表示文本。然而，文本中每个特征词对文本主题内容的贡献度不一样，即每个特征词的权重不同，如何准确有效地计算特征词权重成为重要的研究点。在研究最初特征词的权重只有0或者1，如果该特征词在文本中出现过它的权重就设为1，否则设为0。这种方法完全没有体现出在文本中出现的特征词之间对文本主题内容贡献度的差异性，所以这种权重计算方法慢慢被更精确的基于词频统计的方法替代。

常用的权重计算方法有布尔函数、特征词频平方根、WIDF函数及TF-IDF法等。目前使用最为广泛也是本文中所用到的方法就是TF-IDF法，计算方法如公式（2）所示。

**错误!不能通过编辑域代码创建对象。** （2）

其中，**错误!不能通过编辑域代码创建对象。**表示在文本**错误!不能通过编辑域代码创建对象。**中第j个特征词**错误!不能通过编辑域代码创建对象。**出现的次数，N表示文本集中所有文本数，n表示文本集中含有特征词**错误!不能通过编辑域代码创建对象。**的文本数，**错误!不能通过编辑域代码创建对象。**表示文本**错误!不能通过编辑域代码创建对象。**中第j个特征词的权重。根式**错误!不能通过编辑域代码创建对象。**是归一化因子。

将每个特征词在某个用户发表的所有微博中出现的权重相加，就得到了该用户对于该特征词的兴趣度值。

本文中利用基于维基百科的中文本体自动向上扩展建立一个树状层次结构，其中最主要的关系为上下位关系，下一层的同义词集是其父节点的下位关系，反之亦然。上下层的关系也是包含与被包含的关系，下一层的节点包含于其父节点。我们手动建立了一个基于维基百科分类下的分类特征词本体库，共输入有4757条记录，分为财经、IT、健康、体育、旅游、教育、招聘、文化、军事九个大类。

当之前分析出的用户兴趣主题与我们所建立的本体库中的某个词匹配时，说明用户也对该兴趣主题的直接父节点以及祖先节点感兴趣，只不过对它们的感兴趣程度有所差别。基于这个思想，用户模型构建中会自动查找并更新兴趣主题的父亲节点，直至根节点。

通过以上步骤就可以构建出微博用户模型了，包括用户兴趣树和兴趣度两大模块。在用户模型中，用户的兴趣表示为一棵用户兴趣的本体子树，如**错误!未找到引用源。**所示。这棵本体子树包含的信息主要有：

（1）用户的兴趣主题

（2）用户不同的兴趣主题对应的兴趣度

（3）兴趣主题之间的层次结构关系

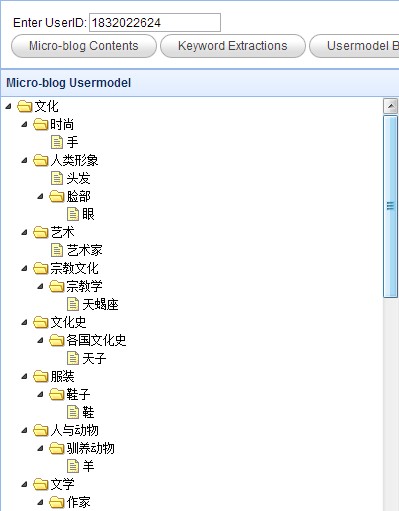
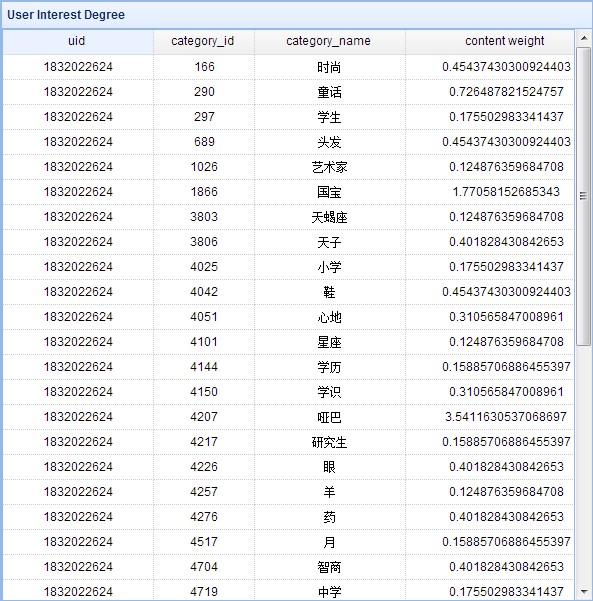


图3‑2用户模型中一个用户的兴趣树

某个用户对于其所有微博中的每个特征词的兴趣度值已经过计算得到，在用户模型构建中，也将这部分内容在界面上显示出来，如图‑3‑3所示。



图‑3‑3用户模型中一个用户对每个微博特征词的兴趣度

本章主要介绍了用户模型的基本概念以及常用用户模型和本体用户模型的表示方法。另外，详细地描述了微博用户模型构建方法的三大步骤，也是本文的最主要内容和贡献。

常用的三种用户模型表示方法缺乏语义和统一的标准。而基于本体的用户模型表示方法不仅能充分描述用户兴趣的语义，还具有兼容性和可扩展性，很好地解决了常用用户模型表示方法存在的问题，是用户模型表示方法的发展方向，因此本文将实现基于本体的用户模型构建。

本文采用基于词频统计的分词方法进行微博文本内容的特征词提取；基于TF-IDF方法计算出微博特征词的权重，并相加得到其兴趣度；最后生成用户模型，显示出用户兴趣树以及用户对每个特征词的兴趣度。

# 图形界面搭建

本章主要介绍微博用户模型构建系统的设计与实现，包括系统整体设计、数据库设计以及系统内各功能模块的设计与界面展示。

## 网站整体设计

本文中设计实现的用户模型评价系统使用的开发工具是Java开发平台Eclipse和关系型数据库MySQL。

Eclipse是一个开放源代码的软件开发项目，专注于为高度集成的工具开发提供一个全功能的、具有商业品质的工业平台。

基于Java的开发平台还有很多，比如JBuilder在以前比较流行，但它是收费的。它的优点在于可以拖拉窗体，比较适合桌面软件的开发。而Eclipse的优点在于它是开源的软件，不收费，并且拥有大量丰富的插件，现在大多数企业都用它来开发，是目前最流行的开发工具之一，所以本文选择Eclipse来进行微博用户模型的构建与展示。

MySQL是一个小型关系型数据库管理系统。与其他的大型数据库例如Oracle、DB2、SQL Server等相比，MySQL自有它的不足之处，如规模小、功能有限等，但是这丝毫也没有减少它受欢迎的程度。MySQL灵活、小巧、便捷，对于一般的个人使用者和中小型企业来说，MySQL提供的功能已经绰绰有余，而且MySQL也是开源软件。所以MySQL非常适合用于轻量级应用的开发，这也是本文选择MySQL作为数据库的原因。

系统整体将分为五个步骤来构建：

（1）分类本体库构建：基于维基百科的分类手动建立一个本体库。

（2）微博特征词提取：收集某一用户所发表的所有微博内容，并对每条微博进行分词操作，提取出特征词，计算得到其权重。

（3）主题兴趣度计算：将每个特征词在所有微博中出现的权重相加，得到这个用户对每个微博特征词的兴趣度。

（4）用户模型生成：将特征词与分类本体库进行匹配，得到这个用户的所有兴趣并形成一棵兴趣树。

（5）系统设计与实现：在系统界面上显示用户的兴趣树以及该用户对每个微博特征词的兴趣度，构建完成一个微博用户的兴趣模型。

系统设计流程图如**错误!未找到引用源。**所示。

用户的微博内容收集

建立本体库

微博内容分词，提取特征词（关键词）

特征词权重的计算

特征词与本体库分类进行匹配

用户所有微博中各特征词权重加和，

得到兴趣度并显示

构建并显示用户模型

显示用户兴趣树

图4‑1系统流程图

## 开发环境简介

数据库表的设计是依据本文第三章所描述的用户模型构建的三大步骤，创建了如下四张表格。

用户发表的所有微博的内容s\_weibo\_content（如**错误!未找到引用源。**所示），包括的字段信息有用户编号uid，该用户发表的微博编号weibo\_id，微博内容content等。

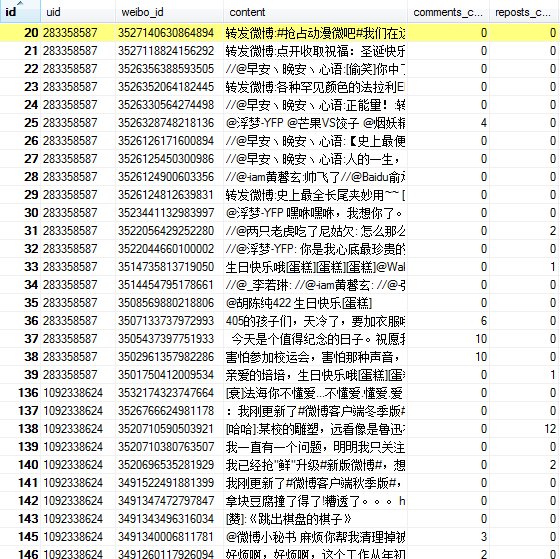


图4‑2微博内容数据表

对每条微博进行分词后所提取得到的特征词s\_feaword（如图4-3所示），包括的字段信息有用户发表的微博编号weibo\_id，从该微博提取到的特征词feaword，以及该词所占的权重weight。



图4‑3特征词数据表

手动创建的基于维基百科的分类本体库s\_category（如图4-4所示），包括的字段信息有分类序号ids，分类名称name，每个分类所对应的父节点序号parent\_id。



图4‑4分类本体库数据表

表4.1分类本体库数据表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **序号** | **姓名** |  | **父节点** | **关系** |
| 1 | 陈寅恪 |  | 0 | 2 |
| 2 | 鲁迅 |  | 1 | 2 |
| 3 | 章太炎 |  | 2 | 3 |

最终构建完成的用户模型s\_user\_profile（如图4-5所示），包括的字段信息有用户编号uid，特征词的分类序号category\_id（即对应表s\_category中的分类序号ids），该特征词的内容权重contentweight等。

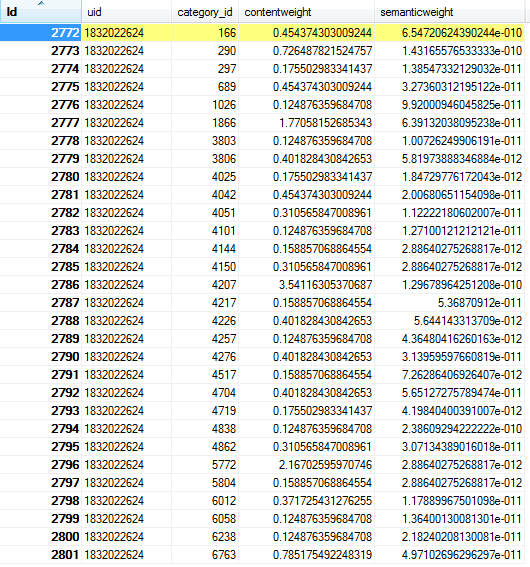


图4‑5用户模型数据表

## 代码实现

系统界面的功能模块主要分为本体的构建与显示、本体的构建与显示和用户模型构建与显示三个部分。

基于维基百科的分类手动建立了4757条分类记录，初步形成了一个可作为研究基础的分类本体库，如**错误!未找到引用源。**所示。



图4‑6基于维基百科的分类本体库

在系统界面上输入一个用户的ID号，就可以展示出该用户所发表的所有微博内容（如**错误!未找到引用源。**所示）以及从这些微博中提取得到的特征词和相应的权重（如**错误!未找到引用源。**所示）。



图4‑7用户发表的所有微博内容



图4‑8用户所有微博中提取的特征词及权重

构建完成的用户模型显示分为两部分，左边为所查询用户的兴趣树，右边为对应计算得到的用户对其所有特征词的感兴趣程度，如**错误!未找到引用源。**所示。



图4‑9用户模型的构建与显示

## 本章小结

本章主要介绍了微博用户模型构建系统的整体设计及各功能模块的展示，简单介绍了开发环境及系统的架构，并展示了部分数据表的内容及演示界面的布局、数据等。

# 总结与展望

本章对全文的主要工作和创新点作了总结，并提出需要进一步研究和改进之处。

## 本文总结

本文针对微博中存在的问题，对用户的微博内容自动进行分析，从而提取用户的兴趣，并建立微博用户模型，为微博信息推荐、舆情监控、微博营销等提供技术支持。

### 本文的主要工作

本文主要研究的是基于本体的微博用户模型构建方法，主要工作内容有以下几个方面：

（1）分类本体库构建：基于维基百科的分类手动建立了4757条分类记录，初步形成了一个可作为研究基础的分类本体库。

（2）微博特征词提取：收集某一用户所发表的所有微博内容，并对每条微博进行分词操作，提取出特征词，计算得到其权重。

（3）主题兴趣度计算：将每个特征词在所有微博中出现的权重相加，即得到了这个用户对每个微博特征词的兴趣度。

（4）用户模型生成：将特征词与之前建立的分类本体库进行匹配，得到这个用户的所有兴趣并形成一棵兴趣树。

（5）系统设计与实现：在系统界面上显示用户的兴趣树以及该用户对每个微博特征词的兴趣度，即构建完成了一个微博用户的兴趣模型。

### 本文的主要创新点

本文基于本体的微博用户模型构建方法创新点主要有以下两项：

（1）实现了利用微博信息分析用户兴趣方法。

（2）实现了利用本体构建用户模型的方法。

## 展望

虽然本文实现了基于本体的微博用户模型构建，但仍存在不少需要改进的地方：

（1）本体的完备性和微博的短文本对用户模型的准确性影响较大。由于手动建立的本体库有限，仅能作为初步研究工作的基础，但为了下一步研究的准确性更上一层楼，本体库的数据需要大量的增加，维护数据库的成本将会变得非常高。而短文本又是微博的一大特点，在140字以内的短文本中提取有效的特征词变得更加不易。

（2）本文对特征词所做的匹配仅限于其内容，但是中文中写法相同却有着截然不同含义的词汇有很多，因而会对用户模型的准确性带来考验。而若是加上语义的因素，难度又将增加不少。

# 致谢

XXX。

# 参考文献

1. 林鸿飞,杨元生.用户兴趣模型的表示和更新机制.计算机研究与发展，200239(7):838-842.
2. 胡学联,潘金贵,李俊,张灵玲. 一个个性化的信息搜集Agent的设计与实现.软件报,2001,12(7):1074-1079.
3. 应晓敏,面向Internet个性化服务的用户建模技术研究. 2003,国防科技大学:长沙
4. 徐振宁,张维明,陈文伟. 基于Ontology的智能信息检索. 计算机科学,20.28(6):21-26.
5. 李勇.智能信息检索中基于本体的个性化用户建模技术及应用. 2002,国防科技学:长沙.
6. FragoudisD.User Modeling in Information Discovery:An overview.InProceedings of Advanced Course on Artificial Intelligence,1999(ACAI99), July,1999, Greece.
7. Balabanovie M, Shoharn Y.Learning Information Retrieval Agents:Experiments with Auto-mated Web BrowsingIn:Proceedings of the AAAISpring Symposium Series on Information Gathering from Heterogeneous, Distributed Environments, March, 1995:13-18.
8. Lieberman H.Letizia. An Agent that Assists Web Browsing.In:Proceedings ofthe International Joint Conference on Artificial Intelligence, Montreal, August, 1995:924-929.
9. Chan P.K.A Non-Invasive Learning Approach to Building Web User Profiles in KDD-99Workshop on Web Usage Analysis and User Profiling, 1999, New York:ACMpress.
10. Schwab L,Kobsa A,and Koychev I. Learning about User from Observation[J].in AAAISpring Symposium on Adaptive User Interface, 2000, Standord,California:AAAI Press.
11. Adomavicius G and Tuzhilin A.Using Data Mining Methods to Build CustomerProfiles.IEEE Computer.Feb 2001:74-82.
12. 余伟. 基于本体的微博客用户行为模型研究[J]. 广东技术师范学院学报, 2010, 31(006): 27-30.
13. 赵岩露, 王晶, and 沈奇威. "基于特征分析的微博用户兴趣发现算法." 电信工程技术与标准化 25.11: 79-83.
14. Uschold M. and Gruninger M., Ontologies: Principle, Methods and Application. Knowledge Engineering Review, 1996, 11(2): p.93-155.
15. Neches R, et al. Enabling technology for knowledge sharing.AI Magazine, 1991, 12(3): p.36-56.
16. Swartout W. Ontologies, 1999. Intelligent Systems and their Applications, IEEE. Issue:1, p.18-19.
17. Gruber T.R.,A Translation Approach to Portable Ontology Specifications .knowledge acquisition, 1993, 5(2): p.199-221.
18. Borst W.N. Construction of Engineering Ontologies for Knowledge Sharing and Reuse. 1997, University of Twente: Enschede.
19. Studer R, Benjamins VR, and Fensel D. Knowledge Engineering: Principles and Methods. Data and Knowledge Engineering, 1998, 25(1-2): p.161-197.
20. 朱晓冰，寇雅楠基于维基技术的本体构建方法探讨图书馆学研究2009.1: p.55-58.
21. 于江生，俞士汝.中文概念词典的结构.中文信息学报.2002年第4期: p.64-68.
22. 由丽凭，杨翠.汉语框架语义知识库概述电脑开发与应用.2007年06期: p.72-74.
23. 张晶，姚建民，赵铁军，李生.基于WordNet和HowNet建设双语语义词典.高技术通讯.2001.12(2-3): p.43-46.
24. 侯汉清，薛春香.用于中文信息自动分类的《中图法》知识库的构建.中国图书馆学报.2005年第5期: p.67-70.
25. 罗志成，马费成，吴晓东，宋倩倩.从维基分类系统构建中文语义词典研究.信息系统学报.2008年02期: p.35-39.
26. 杨玲贤.基于ontology的教学资源知识库构建.计算机与现代化.2009年第11期: p.27-32.
27. Fromkin V, Rodman R. Introduction to Language. London: Holt, Rinehart and Winston, Inc., 1988, 12(1): p.10–16.
28. Brachman R, What IS-A is and isn't: An analysis of taxonomic links in semantic networks [J]. IEEE Computer, 1983, 16(10): p.30–36.
29. Fischer G. User modeling in human–computer interaction[J]. User modeling and user-adapted interaction, 2001, 11(1-2): 65-86.
30. 王金花. 一种利用本体关联度改进的TF-IDF特征词提取方法[D]. 河北大学, 2011.
31. 中国科学院计算技术研究所[EB/0L]. 中文自然语言处理开放平台. http://www.nlp.org.en/project/ project.php?proj\_id=6, 2005-2-2.
32. Scott Deerwester, Susan T Dumais, Furnas W George, et al. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, Vol 41, 1990.
33. 王娟琴. 三种检索模型的比较分析研究—布尔模型、向量模型、概率模型[J]. 情报科学, 1998, 16(3): 225-231.
34. 景玉峰等. 概率检索模型.现代图书情报技术, 1987(1): 29-31.
35. 冀胜利, 李波. 基于SVM的中文文本分类算法[J]. 重庆工学院学报(自然科学), 2008, Vol.22 No.7: 84-87.
36. 崔争艳. 中文短文本分类的相关技术研究[D].
37. 埃克尔著, 陈昊鹏译. Java编程思想[M]. 第4版. 北京: 机械工业出版社, 2007.
38. 李刚. 疯狂Java讲义[M]. 第2版. 北京: 电子工业出版社, 2008.
39. 李刚. 轻量级Java EE企业应用实战:Struts2＋Spring3＋Hibernate整合开发[M]. 第3版. 北京: 电子工业出版社, 2011.
40. Dou Shen, Jian-Tao Sun, Qiang Yang, et al. Latent Friend Mining from Blog Data[C]. Sixth International Conference on Data Mining, 2006: 552-561.
41. Katarzyna Musiał, Przemysław Kazienko, and Tomasz Kajdanowicz. Social Recommendations within the Multimedia Sharing Systems[C]. First World Summit on the Knowledge Society, 2008: 364-372.

# 附录：部分源程序清单

//1. 分词、特征词提取

**publicvoid** Feaword()**throws** Exception

{

List<S\_user> userlist =findusers();

Map<String,Integer> globalmap= **new**HashMap<String,Integer>();//存放全部微博特征词

**int** weibo\_count=0;

**for**(**int** u=0;u<userlist.size();u++)

{

String uid=userlist.get(u).getUid();

List<S\_weibo\_content> weibolist=findByUid(uid);

**if**(weibolist==**null**||weibolist.size()==0)

{

}

**else**

{

weibo\_count=weibo\_count+weibolist.size();

**for**(**int** w=0;w<weibolist.size();w++)

{

S\_weibo\_content sweibo=weibolist.get(w);

//单条微博开始

Map<String,Integer> currentmap = **new** HashMap<String,Integer>();//存放单条微博特征词

SplitWord splitWord = SplitWord.*getInstance*();

splitWord.init();

**int** count=0;

String str=sweibo.getContent();

str=str.replaceAll(" ", "");

String result=**null**;

result=splitWord.split(str);

//System.out.println(result);

String[] rr=result.split(" ");

**for**(**int** i=0;i<rr.length;i++)

{

String[] rr1=rr[i].split("/");

**if**(rr1[1].equals("n"))

{

String noun=rr1[0];

count++;

**if**(currentmap.get(noun)==**null**)

{

currentmap.put(noun, **new** Integer(1));

}

**else**

{

Integer number = currentmap.get(noun);

number++;

currentmap.put(noun, **new**Integer(number));

}

}

}

Iterator iter = currentmap.entrySet().iterator();

**while** (iter.hasNext())

{

Map.Entry<String,Integer> entry = (Map.Entry<String,Integer>) iter.next();

String key=entry.getKey();

**if**(globalmap.get(key)==**null**)

{

globalmap.put(key, **new** Integer(1));

}

**else**

{

Integer number=globalmap.get(key);

number++;

globalmap.put(key, **new** Integer(number));

}

}

}//单条结束

}

System.*out*.println("用户"+uid+"的微博分词结束");

}

System.*out*.println("总微博数weibo\_count="+weibo\_count);

//2. 特征词权重计算

**for**(**int** u=0;u<userlist.size();u++)

{

String uid=userlist.get(u).getUid();

List<S\_weibo\_content> weibolist=findByUid(uid);

**if**(weibolist==**null**||weibolist.size()==0)

{

}

**else**

{

**for**(**int** w=0;w<weibolist.size();w++)

{

S\_weibo\_content sweibo=weibolist.get(w);

Map<String,Integer> currentmap = **new** HashMap<String,Integer>();

SplitWord splitWord = SplitWord.*getInstance*();

splitWord.init();

**int** count=0;

String str=sweibo.getContent();

str=str.replaceAll(" ", "");

String result=**null**;

result=splitWord.split(str);

System.*out*.println(result);

String[] rr=result.split(" ");

**for**(**int** i=0;i<rr.length;i++)

{

String[] rr1=rr[i].split("/");

**if**(rr1[1].equals("n"))

{

String noun=rr1[0];

count++;

**if**(currentmap.get(noun)==**null**)

{

currentmap.put(noun, **new** Integer(1));

}

**else**

{

Integer number = currentmap.get(noun);

number++;

currentmap.put(noun, **new** Integer(number));

}

}

}

Iterator iter2 = currentmap.entrySet().iterator();

**while** (iter2.hasNext())

{

Map.Entry<String,Integer> entry2 = (Map.Entry<String,Integer>) iter2.next();

String key=entry2.getKey();

**double** tf=(**double**)currentmap.get(key)/(**double**)count;

System.*out*.println("count="+count);

System.*out*.println("tf="+tf);

**double** idf=(**double**)weibo\_count/(**double**)globalmap.get(key);

idf=Math.*log10*(idf);

System.*out*.println("globalmap.get(key)="+globalmap.get(key));

System.*out*.println("idf="+idf);

**double** weight=tf\*idf;

S\_feaword ff=**new** S\_feaword();

ff.setSweibo(sweibo);

ff.setFeaword(key);

ff.setWeight(weight);

saveFeaword(ff);

}

}

}

}

//3.用户模型兴趣树的显示

<ul class="easyui-tree">

<s:iterator value="#request['menus']" id="c2">

<s:if test="#c2.parent\_id == 0">

<li data-options="state:'closed'">

<span><s:property value="#c2.name" /></span>

<ul>

<s:iterator value="#request['menus']" id="c3">

<s:if test="#c3.parent\_id == #c2.ids">

<li>

<span><s:property value="#c3.name" /></span>

<ul>

<s:iterator value="#request['menus']" id="c4">

<s:if test="#c4.parent\_id == #c3.ids">

<li>

<span><s:property value="#c4.name" /></span>

<ul>

<s:iterator value="#request['menus']" id="c5">

<s:if test="#c5.parent\_id == #c4.ids">

<li>

<s:property value="#c5.name" />

</li>

</s:if>

</s:iterator>

</ul>

</li>

</s:if>

</s:iterator>

</ul>

</li>

</s:if>

</s:iterator>

</ul>

</li>

</s:if>

</s:iterator>

</ul>

//4. 用户兴趣度计算

public void contentinterest() throws Exception

{

List<S\_user> userlist = findByuid("283358587");

//读取所有用户

if(userlist==null||userlist.size()==0)

{

}

else

{

//读取所有本体类别

List<S\_category> categorylist=ms\_categoryDao.getS\_categoryname();

if(categorylist==null||categorylist.size()==0)

{

}

else

{

for(int u=0;u<userlist.size();u++)

{

//读取某一个用户的所有微博

String uid=userlist.get(u).getUid();

List<S\_weibo\_content> weibolist=findByUid(uid);

if(weibolist==null||weibolist.size()==0)

{

}

else

{

for(int c=0;c<categorylist.size();c++)

{

//判断任意一条微博在每一个分类上的权重，如有则类别兴趣累计

Double wc=(double) 0;

Double ws=(double) 0;

System.out.println(weibolist.get(w).getWeibo\_id()); feaword=findBywid(s\_weibo\_content.getWeibo\_id());//找到微博w的所有特征词记录

List<S\_feaword> feaword=findBywfid(categorylist.get(c).getName(),uid);

if(feaword==null||feaword.size()==0)

{

}

else

{

for(int f=0;f<feaword.size();f++)

{

wc=wc+feaword.get(f).getWeight();

ws=ws+categorylist.get(c).getScd();

}

}

if(wc>0)

{

S\_s\_user\_profile user\_profile= new S\_s\_user\_profile();

user\_profile.setUser(userlist.get(u));

user\_profile.setCategory(categorylist.get(c));

user\_profile.setContentweight(wc);

user\_profile.setSemanticweight(ws);

try

{

s\_UserProfileDao.save(user\_profile);

}catch(Exception e)

{

e.printStackTrace();

}

}

}

}

}

}

}

System.out.println("用户完");

}

//5. 用户兴趣度显示

<table id=*"tt"*class=*"easyui-datagrid"*style="width:*auto*;">

<thead>

<tr>

<thdata-options=*"field:'uid',width:150"*align=*"center"*>uid</th>

<thdata-options=*"field:'s\_category.ids',width:100"*align=*"center"*>category\_id</th>

<thdata-options=*"field:'s\_category.name',width:150"*align=*"center"*>category\_name</th>

<thdata-options=*"field:'contentweight',width:200"*align=*"center"*>content weight</th>

</tr>

</thead>

<s:iteratorvalue=*"#request['plansearch']"*id=*"c2"*>

<tr>

<td><s:propertyvalue=*"#c2.uid"*/></td>

<td><s:propertyvalue=*"#c2.category\_id"*/></td>

<td><s:propertyvalue=*"#c2.category\_name"*/></td>

<td><s:propertyvalue=*"#c2.contentweight"*/></td>

</tr>

</s:iterator>

</table>