摘要

随着网络中主观性文本的急剧增长，文本情感分析作为一个新兴研究课题，其研究价值和应用价值越来越受到学术界的重视。本文以机器学习领域的理论为依据，设计并实现了一个能够自动化地分析文本情感的系统。

本文首先将文本情感分析分解为两个主要的子问题，即文本的潜在语义分析和文本的潜在语义到文本情感的映射，本文随后使用机器学习领域的理论分别解决这两个子问题，进而完成对文本情感的自动化分析。

# 绪论

## 文本情感分析的研究背景

随着互联网的飞速发展，互联网公司纷纷开始提倡“以用户为中心，用户参与”的开放理念。越来越多的互联网用户不再单纯地从互联网获取信息，转而开始主动创造互联网信息。因此，互联网中出现了大量的博客、论坛和讨论组，并伴随着大量的互联网用户参与，于是产生了大量对于诸如人物、时间和产品等主观性的评论信息，这些评论信息蕴涵着人们各种情感色彩和情感倾向性，如“愤怒”、“兴奋”、“赞扬”、“批评”、“哀伤”、“愉悦”等。基于此，潜在的消费者在购买某个产品或服务时可以参考相关的评论信息从而判断是否购买，政府部门也可以通过浏览公众对新闻事件或国家政策的评论信息来了解舆情。然而，由于越来越多的用户在互联网上发布自己的主观性评论，这类评论信息每天以指数级的速度迅速膨胀，使用人工分析的方法需要消耗大量的人力和时间，成本过高。因此，采用计算机来自动化地分析这些主观性文本表达的情感的技术，称为文本情感分析技术，应运而生，且已经成为目前学术界研究的一个热点。

## 文本情感分析的应用

随着互联网上主观性文本以指数级的速度迅速增长，迫切需要计算机来帮助加工整理这些情感信息，因此，文本情感分析有着重要的应用，下面，本文就文本情感分析的应用现状以及应用前景进行概括介绍了三个主要应用领域，用户评论的分析与决策、舆情监控和信息预测。

### 用户评论的分析与决策

人们在购买某一件产品或服务前，往往都会在网上查询曾购买该产品或服务的人的评论信息，并与其他产品做对比，从而做出最终的决策，是否购买。但是，由于相关的评论信息可能过多，用户没有条件浏览完所有的评论信息，导致最终的决策带有风险。然而，文本情感分析技术可以非常好地解决这个难题。首先可以从互联网自动获取大量的相关评论信息，随后使用文本情感分析技术分析出这些评论信息中的情感，如“衣服很好看”中蕴藏着“赞扬”的情感，“经常掉线”蕴藏着“抱怨”的情感，将这些情感通过统计归纳整理，再展现给用户，方便用户做最终的决策。

### 舆情监控

由于互联网具有开放性、虚拟性、隐蔽性、发散性、渗透性和随意性等特点，越来越多的人们乐意通过互联网来表达其对任何事物的主观看法，网民言论活跃已达到前所未有的程度，不论是国内还是国际重大事件，都能立刻形成互联网舆论。通过互联网来表达观点，传播思想，进而产生巨大的舆论压力，以致于任何部门、机构都无法忽视其影响，可以说，互联网已成为思想文化信息的集散地和社会舆论的放大器。网络信息和社会信息的交融对社会的直接影响将越来越大，甚至关系到国家信息安全和长治久安。因此，社会管理者应及时对这些舆论进行反馈，然而由于互联网上的信息量十分庞大，仅靠人工的方法难以应对网上海量信息的处理。因此，需要依靠情感分析技术自动化地对舆论信息进行监控。

### 信息预测

随着互联网的迅速发展，网络信息对人们生活的影响越来越大。某一个新事件的发生或者网络上对某个事件的热议都在很大程度上左右着人们的思维和行动。比如在金融领域，网络上对某支股票的热议会很大程度地影响金融实践者们的行为，进而进一步影响着股市变化的趋势。再比如，国外总统或议员大选的时候，很多参选者希望通过汇总选民的网络言论来预测自己是否能够获选。因此，信息变得非常重要，同样，由于数据量的原因，无法使用人工分析，需要借助文本情感分析技术自动化地帮助用户通过对互联网上的新闻、帖子等信息源进行分析，预测某一时间的未来状况。

### 其他应用

除了这三个领域外，文本情感分析在其他一些自然语言处理领域也扮演着重要的角色。例如，在信息抽取领域，抽取对象一般是反映客观事实的文本，情感分析技术可以将文本中的主观句和可观句进行分离，从而提高信息抽取的准确率。文本情感分析技术还可饮用于问答系统中，当用户所问问题是情感相关的问题时，该技术可以帮助问答系统提供更真实的答案。此外，文本情感分析技术还可以应用于情感文摘的生成，进而达到汇总归纳的目的。文本情感分析技术的快速发展源于人们改进人机交互的现状的愿望，该技术在以上众多研究领域的应用，使其成为一个非常重要的研究方向。

## 本文工作

本文的工作是灵活使用前文提到的机器学习领域相关的理论，设计并实现一个文本情感分析系统，能够自动化地判断任意一个主观性文本段所表达的主要情感是正面情感还是负面情感，从而达到文本情感分析的目的。本系统的全称为“基于机器学习的文本情感分析系统”（Analysis system of Emotions in Text based on Machine Learning，下文简称“AETML”）。

## 本文结构

本文一共分为五章，详细介绍了AETML所用到的相关理论以及作者所做的设计与实现工作，随后对系统几个关键模块进行了评估，最后对未来工作作出展望。

第一章，绪论。主要介绍文本情感分析的研究背景及意义，并介绍文本情感分析的一些应用，最后阐述本文的工作内容。

第二章，系统需求。本章主要从系统功能设计出发，使用用例图详细描述AETML的系统需求，包括功能性需求和非功能性需求。

第三章，系统设计。本章根据系统需求，对AETML进行概要设计，随后详细介绍对各个模块的详细设计。

第四章，算法设计。并简单介绍AETML所涉及到的机器学习领域相关理论，详细介绍各个模块为完成其功能所运用的关键算法，以及如何运用关键算法达成目的。

第五章，系统实现。介绍系统的开发环境和运行环境，并展示用户界面。

总结和展望部分对整个论文进行总结，指出论文中存在的不足之处，并对下一步作者的工作方向进行展望与设想。

# 系统需求

## 系统用例分析

用户在AETML中可执行文本情感分析功能，即输入一段文本，分析出其中蕴含的情感。AETML所能判定的情感有两种，正面情感和负面情感。用户可通过饼形图的形式得到输入的文本段的正负面，以及AETML所计算的相应的可信度。

文本情感分析用例包含两个主要的子用例，分别是未知文本的潜在语义分析和从潜在语义到情感的数据映射。

先解释潜在语义的概念，定义一个潜在语义向量空间，AETML将所有的文本段所表示的语义视为在这个潜在语义向量空间中的一个向量。潜在语义分析就是根据文本段找到其在潜在语义空间中的向量表示。AETML中定义两种文本，已知文本和未知文本。已知文本指的是在系统投入运行前已有的大量情感正负面已知的文本段，未知文本指的是在系统投入运行后用户输入的情感正负面未知的文本段。

从潜在语义到情感的数据映射，是指根据文本段的潜在语义分布这一特征，判定其表达的主要情感是正面情感还是负面情感。

系统用例图如图2-1所示。



图 2‑1 系统用例图

## 未知文本的潜在语义分析

未知文本的潜在语义分析是指对于未知文本，也就是系统投入运行后用户输入的文本段，在潜在语义空间中找到一个向量用于表示这个文本段。例如，对于文本段“今天天气真好”，假设潜在语义空间为四维向量空间，AETML可能会在潜在语义空间中找到这样一个向量用于表示这个文本段的语义，(0.1, 0.2, 0.3, 0.4)。

未知文本的潜在语义分析用例图如图2-2所示。



图 2‑2 未知文本的潜在语义分析用例图

## 从潜在语义到情感的数据映射

从潜在语义到情感的数据映射，是指根据文本段的潜在语义这一特征，判定其表达的主要情感是正面情感还是负面情感。例如某个文本段的潜在语义为(0.1, 0.2, 0.3, 0.4)，根据这个数据映射，AETML得到其为正面情感。

从潜在语义到情感的数据映射的用例图如图2-3所示。



图 2‑3 从潜在语义到情感的数据映射用例图

## 非功能性需求

1. 可行性分析
   1. 经济可行性：当用于训练的原始语料较多时，需要一定的内存才能完成训练过程，所以计算机的内存需要达到一定的要求才能运行成功。
   2. 技术可行性：AETML所使用的机器学习理论均已在业界得到充分的验证，在理论层面可行。
2. 性能效率分析

语言层面，AETML使用c++，python，php+html+js混合开发。为了提高运行效率，对于时间复杂度要求较高的机器学习过程均使用c++开发；为了提高开发效率，对于时间复杂度要求较低且一次性运行的过程使用python开发；使用php+html+js完成用户交互界面，运行效率极高。

数据结构层面，AETML灵活使用各种数据结构以提高运行效率，比如使用哈希表存储词袋。

1. 稳定性和可读性

AETML中主要过程都是c++编写，且c++代码完全遵守业界编码规范，可读性和稳定性都极高。

## 本章小结

本章首先使用系统用例图将系统功能清晰地表达出来，然后详细阐述了系统的功能性需求，并且从可行性分析、性能效率分析、稳定性和可读性等方面阐述了系统的非功能性需求。

# 系统设计

## 概要设计

为完成文本情感分析的目的，AETML首先对系统投入运行前已知的大量的情感正负面已知的文本段进行潜在语义挖掘，挖掘出每个文本段中的潜在语义的概率分布，并对潜在语义的概率分布与情感正负面间的对应规则进行学习并产生一个模型，该模型能够完成从潜在语义的概率分布到情感正负面的数据映射。模型产出后，对于情感正负面未知的文本段，首先挖掘出其潜在语义的概率分布，然后使用产出的模型根据其潜在语义的概率分布判断其情感正负面，并给出可信度，最终完成文本情感分析的功能。

系统分为四个主要模块，分别为潜在语义挖掘模块、情感规则学习模块、服务器模块和客户端模块。其中，潜在语义挖掘模块分为中文分词模块、已知文本的语义挖掘模块和未知文本的语义挖掘模块；服务器模块分为情感判定模块和数据交互模块；客户端模块分为情感判定结果展现模块和数据交互模块。

系统功能模块图如图3-1所示。



图 3‑1 系统功能模块图

## 潜在语义挖掘模块

### 功能描述

如前文所说，在AETML中，定义一个潜在语义空间，任何一个文本段，一个词所表达的语义都可以使用潜在语义空间的一个向量表示。而对于任何文本段，挖掘出其语义在潜在语义空间中的向量表示是潜在语义挖掘模块的功能。这个过程涉及到中文分词、潜在语义分析技术，将此模块分为三个子模块，分别为中文分词模块、已知文本的语义挖掘模块和未知文本的语义挖掘模块。中文分词模块用于完成分词功能；已知文本的语义挖掘模块负责对情感正负面已知的文本段进行潜在语义挖掘；未知文本的语义挖掘模块负责对情感正负面未知的文本段进行潜在语义挖掘。

潜在语义挖掘模块的数据流图如图3-3所示。



图 3‑2 潜在语义挖掘模块的数据流图

### 中文分词模块

#### 功能描述

在英文的行文中，单词与单词之间是以空格作为自然分界符的，但在中文中，只是汉字、句子和段落能够通过明显的分界符来简单划界，唯独词没有一个形式上明显的分界符。此模块的主要作用是将汉字序列切分成一个个单独的词，也就是将连续的字序列重新组合成词序列。这里需要注意的是，由于AETML使用词袋模型来进行文本情感分析，此模块对于中文文本分词后重新组合成的词序列，使用词袋类型存储。

#### 词袋类型设计

词袋模型在本系统中的体现是指对于文本分词后重新组成的词序列，只保存所有出现的词及其出现次数，并不保存词的先后顺序。

##### 操作需求

在AETML中，对于词袋的主要操作包括：

1. 遍历取词和其出现次数；
2. 存储到文件中；
3. 从文件中恢复；
4. 对文本段进行切词从而构造词袋；
5. 对文件中的文本段进行切词从而构造词袋；
6. 获得词袋中词的总出现次数；
7. 获得词袋中词的种类的数量；
8. 查询某个特定词的出现次数。

##### 存储设计

为了能够高效地实现上节第1、7、8三个操作，AETML使用哈希表来存储词袋。为了提高开发效率，使用C++11提供的unordered\_map来完成哈希表的功能。同时，为了高效地实现第6个操作，词袋类型除了哈希表外还需一个数值类型用于存储总词数。

##### 接口设计

为满足第一个操作需求，即遍历取词和其出现次数，提供接口迭代器；为满足第二个操作需求，即存储到文件中，提供接口dump，此种方式存储的词袋文件格式称为“bag”格式；为满足第三个操作需求，即从文件中恢复，提供接口load，此接口只能接口“bag”格式存储在文件中的词袋；为满足第四个操作需求，即对文本段进行切词从而构造词袋，提供接口construct\_bag\_from\_doc；为满足第五个操作需求，即对文件中的文本段进行切词从而构造词袋，提供接口construct\_bag\_from\_doc\_path；为满足第六个操作需求，即获得词袋中词的总出现次数，提供接口num\_of\_words；为满足第七个操作需求，即获得词袋中词的种类的数量，提供接口size；为满足第八个操作需求，即查询某个特定词的出现次数，提供接口query。

##### 类图

结合存储设计和接口设计，词袋类型的类图如图3-3所示。



图 3‑3 词袋类图

### 已知文本的语义挖掘模块

#### 功能描述

已知文本指的是大量的在本系统投入运行之前已有的情感正负面已知的文本段，经过中文分词模块后此模块得到的是大量情感正负面已知的词袋。已知文本的语义挖掘模块的功能是使用潜在语义分析方法挖掘出情感正负面已知的词袋中的潜在语义的概率分布，此外还要提供每个潜在语义中的词的概率分布供未知文本的语义挖掘模块使用。

#### IPO图

此模块需要输入的是大量的情感正负面已知的词袋、潜在语义空间的维数以及算法最大迭代次数，输出每个词袋中的潜在语义的概率分布，每个潜在语义中的词的概率分布，以及单词表。

IPO图如图3-4所示。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| IPO图 | | | | |
|  | 系统：  基于机器学习的文本情感分析系统  模块：  已知文本的语义挖掘模块 |  | 设计人：张伟  日期：2015.5.29 |  |
|  | 上层调用模块：  无 |  | 可调用下层模块：  中文分词模块 |  |
|  |  |  |  |  |
|  | 输入：  大量的情感正负面已知的词袋；  潜在语义空间的维数；  算法最大迭代次数。 |  | 输出：  每个词袋中的潜在语义分布；  每个潜在语义中的词分布；  单词表。 |  |
|  |  |  |  |  |
|  |  | | |  |
|  |  | | |  |
|  | 局部数据项：  算法最大迭代次数 | | |  |
|  |  | | |  |

图 3‑4 已知文本的语义挖掘模块的IPO图

#### 接口设计

接口设计如下如图3-5所示。

|  |
| --- |
| void plsa**(**  const std**::**vector**<**BagOfWords**>** **&**bows**,** // 情感正负面已知的词袋  int64\_t topic\_num**,** // 潜在语义空间的维数  int64\_t max\_iter**,** // 学习过程的最大迭代次数  std**::**vector**<**std**::**string**>** **\***table\_of\_words**,** // 用于输出，单词表  std**::**vector**<**std**::**vector**<**double**>** **>** **\***prob\_topic\_in\_doc**,** // 用于输出，词袋的潜在语义的概率分布  std**::**vector**<**std**::**vector**<**double**>** **>** **\***prob\_word\_in\_topic**,** // 用于输出，潜在语义的词的概率分布  const std**::**string **&**output\_dir\_path**);** // 日志输出路径 |

图 3‑5 已知文本的语义挖掘模块接口图

### 未知文本的语义挖掘模块

#### 功能描述

未知文本指的是系统投入运行后用户输入的情感正负面未知的文本段，此模块的功能是挖掘出情感正负面未知的文本中的潜在语义的概率分布。此模块首先将文本段传给中文分词模块进行分词，得到正负面未知的词袋形式的文本段。使用其他模块所生成的数据，包括情感正负面已知的词袋及其潜在语义的概率分布和潜在语义中的词的概率分布，运用贝叶斯定理，最终得到情感正负面未知的文本段中的潜在语义的概率分布。

#### IPO图

此模块需要输入的是系统投入运行后用户输入的情感正负面未知的文本段，随后输出其中的潜在语义的概率分布，最终将其提供给服务器模块。

IPO图如图3-6所示。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| IPO图 | | | | |
|  | 系统：  基于机器学习的文本情感分析系统  模块：  未知文本的语义挖掘模块 |  | 设计人：张伟  日期：2015.5.29 |  |
|  | 上层调用模块：  服务器模块 |  | 可调用下层模块：  中文分词模块 |  |
|  |  |  |  |  |
|  | 输入：  系统投入运行后用户输入的情感正负面未知的文本段。 |  | 输出：  该文本段中的潜在语义的概率分布。 |  |
|  |  |  |  |  |
|  |  | | |  |
|  |  | | |  |
|  | 局部数据项：  无 | | |  |
|  |  | | |  |

图 3‑6 未知文本的语义挖掘模块的IPO图

#### 接口设计

未知文本的语义挖掘模块接口设计如图3-7所示。

|  |
| --- |
| void compute\_topic\_distribution**(**  const std**::**string **&**sentence**,** // 情感正负面未知的文本段  std**::**vector**<**double**>** **\***topic\_distribution**);** // 输出： 潜在语义分布 |

图 3‑7 未知文本的语义挖掘模块接口设计图

## 情感规则学习模块

### 功能描述

此模块的功能是学习出从潜在语义的概率分布到情感正负面的映射规则，并产出一个分类模型，而这个分类模型将用于服务器模块的子模块情感判定模块。

该模块的数据流图如图3-8所示。



图 3‑8 情感规则学习模块数据流图

### IPO图

此模块需要输入的是系统投入运行前已有的大量的情感正负面已知的词袋及其潜在语义分布，随后产出一个分类模型，最终将其提供给服务器模块。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| IPO图 | | | | |
|  | 系统：  基于机器学习的文本情感分析系统  模块：  情感规则学习模块 |  | 设计人：张伟  日期：2015.5.29 |  |
|  | 上层调用模块：  无 |  | 可调用下层模块：  无 |  |
|  |  |  |  |  |
|  | 输入：  系统投入运行前已有的大量的情感正负面已知的词袋及其潜在语义的概率分布。 |  | 输出：  分类模型，完成从潜在语义的概率分布到情感正负面的数据映射。 |  |
|  |  |  |  |  |
|  |  | | |  |
|  |  | | |  |
|  | 局部数据项：  无 | | |  |
|  |  | | |  |

图 3‑9 情感规则学习模块的IPO图

### 接口设计

AETML中，本模块应对外提供操作接口应有：

1. 输入大量的情感正负面及其潜在语义分布后产出分类模型的接口；
2. 将分类模型存储到文件中的接口；
3. 从文件中读取分类模型的接口；
4. 销毁分类模型的接口。

接口设计如图3-10所示。

|  |
| --- |
| // 返回值: svm\_model（规则）  svm\_model **\***svm\_train**(**  const svm\_problem **\***prob**,** // 情感正负面已知的词袋及其潜在语义分布  const svm\_parameter **\***param**);** // 算法自身运行参数  // 将svm\_model（规则）存储到文件中  int svm\_save\_model**(**  const char **\***model\_file\_name**,** // 文件路径  const svm\_model **\***model**);** // 要存储的svm\_model（规则）    // 从文件中读取svm\_model（规则）  svm\_model **\***svm\_load\_model**(**const char **\***model\_file\_name**);**  // 由于svm\_model \*所指向的内存不会自动销毁，需要这个额外的接口手动销毁  void svm\_free\_and\_destroy\_model**(**svm\_model **\*\***model\_ptr\_ptr**);** |

图 3‑10 情感规则学习模块接口设计

## 服务器模块

### 功能描述

服务器模块作为服务器的角色，要与客户端模块进行数据交互，从客户端模块获得用户输入的情感正负面未知的文本段，随后对文本段进行情感正负面判定，然后将情感判定结果返回给客户端模块。此模块包含两个子模块，分别是情感判定模块和数据交互模块。

数据流图如图3-11所示。



图 3‑11 服务器模块数据流图

### 情感判定模块

#### 功能描述

此模块的功能是根据情感规则学习模块所得到的分类模型，对潜在语义挖掘模块处理后得到的情感正负面未知的文本段的潜在语义的概率分布进行情感正负面判定。

#### IPO图

此模块需要输入的是系统投入运行后用户输入的情感正负面未知的文本段，通过潜在语义挖掘模块产出的该文本段的潜在语义的概率分布判定其情感正负面，最终将其情感判定结果和判定的可信度返回给服务器模块。

IPO图如图3-12所示。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| IPO图 | | | | |
|  | 系统：  基于机器学习的文本情感分析系统  模块：  情感判定模块 |  | 设计人：张伟  日期：2015.5.29 |  |
|  | 上层调用模块：  数据交互模块 |  | 可调用下层模块：  潜在语义挖掘模块 |  |
|  |  |  |  |  |
|  | 输入：  系统投入运行后用户输入的情感正负面未知的文本段 |  | 输出：  系统投入运行后用户输入的情感正负面未知的文本段经过分类模型处理得到的情感判定结果以及判定的可信度 |  |
|  |  |  |  |  |
|  |  | | |  |
|  |  | | |  |
|  | 局部数据项：  无 | | |  |
|  |  | | |  |

图 3‑12 情感判定模块的IPO图

#### 接口设计

情感判定模块的接口设计如图3-13所示。

|  |
| --- |
| // 返回情感判定结果，是正面还是负面  double svm\_predict\_probability**(**  const struct svm\_model **\***model**,** // 情感规则学习模块所得到的规则  const struct svm\_node **\***x**,** // 情感正负面未知的文本段的语义分布  double**\*** prob\_estimates**);** // 用于输出，一个长度为2的数组，为正面和负面的概率 |

图 3‑13 情感判定模块的接口设计图

### 数据交互模块

#### 功能描述

此模块的功能是监听来自客户端的TCP连接，完成作为c++程序的服务端与基于web的客户端之间的数据交互，包括从客户端获取用户输入的情感正负面未知的文本段，和将文本段的情感判定结果以及判定可信度返回给客户端。

#### IPO图

IPO图如图3-14所示。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| IPO图 | | | | |
|  | 系统：  基于机器学习的文本情感分析系统  模块：  数据交互模块 |  | 设计人：张伟  日期：2015.5.29 |  |
|  | 上层调用模块：  无 |  | 可调用下层模块：  情感判定模块 |  |
|  |  |  |  |  |
|  | 输入：  系统投入运行后用户输入的情感正负面未知的文本段 |  | 输出：  系统投入运行后用户输入的情感正负面未知的文本段经过分类模型处理得到的情感判定结果以及判定的可信度 |  |
|  |  |  |  |  |
|  |  | | |  |
|  |  | | |  |
|  | 局部数据项：  无 | | |  |
|  |  | | |  |

图 3‑14 数据交互模块IPO图

## 客户端模块

### 功能描述

客户端模块作为客户端的角色，要与服务端进行TCP连接，从而进行数据交互，将系统投入运行后用户输入的情感正负面未知的文本段传给服务器端，然后从服务器端获取情感判定结果以及情感判定的可信度，最后将结果以饼形图的方式展示给用户。此模块包含两个子模块，分别是情感判定结果展现模块和数据交互模块。

数据流图如图3-15所示。



图 3‑15 客户端模块数据流图

### 情感判定结果展现模块

此模块的功能是将用户输入的文本段的情感判定结果以及情感判定的可信度，以饼形图的形式展现给用户。假设“今天天气不错”的情感判定结果为正面，且可信度为80%，则应以图3-16的形式展现给用户。

图 3‑16 情感判定结果展现图

### 数据交互模块

#### 功能描述

数据交互模块作为客户端的角色，要与服务端进行TCP连接，从而进行数据交互，将系统投入运行后用户输入的情感正负面未知的文本段传给服务器端，然后从服务器端获取情感判定结果以及情感判定的可信度，然后将结果传给情感判定结果展现模块，从而展现给用户。

#### IPO图

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| IPO图 | | | | |
|  | 系统：  基于机器学习的文本情感分析系统  模块：  数据交互模块 |  | 设计人：张伟  日期：2015.5.29 |  |
|  | 上层调用模块：  无 |  | 可调用下层模块：  无 |  |
|  |  |  |  |  |
|  | 输入：  系统投入运行后用户输入的情感正负面未知的文本段 |  | 输出：  系统投入运行后用户输入的情感正负面未知的文本段经过分类模型处理得到的情感判定结果以及判定的可信度 |  |
|  |  |  |  |  |
|  |  | | |  |
|  |  | | |  |
|  | 局部数据项：  无 | | |  |
|  |  | | |  |

# 算法设计

本章依次罗列并详细介绍AETML的核心理论依据，以及各个模块如何运用这些理论完成各个模块所需的功能。

## 理论依据

AETML能够自动化地分析主观性文本表达的情感，在设计与实现过程中使用了大量机器学习领域相关的理论，其中，本系统的关键理论基础有贝叶斯定理、拉格朗日乘数法、词袋模型、最大期望算法、概率潜在语义分析和支持向量机。本节将简单介绍本系统的关键理论基础。

### 贝叶斯定理

首先定义一些符号，符号A和B代表两个随机事件；P(A)是A的先验概率或边缘概率，之所以称为"先验"是因为它不考虑任何B方面的因素；P(B)是B的先验概率或边缘概率，之所以称为"先验"是因为它不考虑任何A方面的因素；P(A|B)是已知B发生后A的条件概率，也由于得自B的取值而被称作A的后验概率；P(B|A)是已知A发生后B的条件概率，也由于得自A的取值而被称作B的后验概率。

贝叶斯定理是概率论中的一个定理，它进而随机变量的条件概率以及边缘概率分布有关。贝叶斯定理能够告知我们如何利用新证据修改已有的看法。通常，事件A在事件B发生的条件下的概率，与事件B在事件A发生的条件下的概率是不一样的。但是，这两者是有确定的关系的，贝叶斯定理就是这种关系的陈述。

贝叶斯定理的内容是：在B出现的前提下,A出现的概率等于A出现的前提下B出现的概率乘以A出现的概率再除以B出现的概率。

公式形式如下：

通过联系A与B，计算从一个事件产生另一事件的概率，即从结果上溯源。作为一个普遍的原理，贝叶斯定理对于所有概率的解释是有效的；然而，频率主义者和贝叶斯主义者对于在应用中，某个随机事件的概率该如何被赋值，有着不同的看法：频率主义者根据随机事件发生的频率，或者总体样本裡面的发生的个数来赋值概率；贝叶斯主义者则根据未知的命题来赋值概率。这样的理念导致贝叶斯主义者有更多的机会使用贝叶斯定理。

### 拉格朗日乘数法

AETML的理论设计中，多次涉及到条件约束下的最优化问题求解，AETML使用拉格朗日乘数法来解决这个问题。在数学最优化问题中，拉格朗日乘数法是一种寻找变量受一个或多个条件所限制的多元函数的极值的方法。假设存在一个最优化问题，有n个变量和k个约束条件，拉格朗日乘数法引入一种新的标量未知数，称为拉格朗日乘数，将最优化问题转换为有n+k个变量的方程组的极值问题，且这n+k个变量不受任何约束。

任何优化问题都可以写成如下形式：

引入拉格朗日乘数，且大于0，得到拉格朗日方程：

令

分析z(x)的值，若存在大于0，z(x)的值为正无穷，若存在不等于0，z(x)的值也为正无穷，当所有的不大于0且所有的等于0时，z(x)的值等于。于是，可以发现，原优化问题等价于对z(x)的最小化。通过拉格朗日乘数法，将有n个变量和k个约束条件的最优化问题转换为有n+k个无约束变量的最优化问题。

### 词袋模型

词袋模型是自然语言处理和信息检索领域使用的一种文本段的简化表示。因其简单且有效的优点得到了广泛的应用。最近，词袋模型甚至还被应用于计算机视觉。在词袋模型中，文本段中的语法和词的顺序被忽略，仅统计每个词在文本段中出现的频率，并使用这个频率完成对文本段的向量表示。在AETML中，对于用户输入的文本段，将其视为一个词袋进行处理。

### 最大期望算法

在AETML的理论设计中，需要一种方法能够在已知部分相关变量的情况下，估计未知变量的方法。AETML使用最大期望算法解决这个问题。

最大期望算法是用于在概率模型中寻找参数最大似然估计或者最大后验估计的算法，其中概率模型依赖于无法观测的隐藏变量。最大期望经常用在机器学习和计算机视觉的数据聚类领域。最大期望算法经过两个步骤交替进行计算，第一步是计算期望（E），利用对隐藏变量的现有估计值，计算其最大似然估计值；第二步是最大化（M），最大化在 E 步上求得的最大似然值来计算参数的值。M 步上找到的参数估计值被用于下一个 E 步计算中，这个过程不断交替进行，直至收敛。

最大期望算法的完整流程如下：

1. 初始化分布参数
2. 重复直到收敛
   1. E步骤：估计未知参数的期望值，给出当前的参数估计
   2. M步骤：重新估计分布参数，以使得数据的似然性最大，给出未知变量的期望估计

### 概率潜在语义分析

文本和自然语言处理是人工智能和机器学习领域的一个重大挑战。这个领域中的任何进展都会对信息检索、信息过滤、智能接口、语音识别、自然语言处理和机器翻译产生重大影响。这之中一个基本的问题便是使用数据驱动的方式，比如给定文本语料的方式来学习单词的意义和用法。在这样的一个机器学习系统，比如AETML中，最主要的问题是从词汇级别到语义级别的跨越。产生这个问题的原因有两个：（1）一个单词在不同的上下文中有多个语义和用法；（2）同义和近义词，这些词在某个上下文中表示同样或者相近的语义。

概率潜在语义分析（Probabilistic Latent semantic analysis，下文简称PLSA）是处理这些问题的著名方法。PLSA使用词袋模型，在词袋模型中，使用每个词在单个文本段中出现的频率来表示一个文本段，于是每个文本段就有一个高维向量表示，PLSA的主要思想是将这个高维向量映射到一个称为潜在语义空间的向量空间中，从而完成从词汇级别到语义级别的数据映射。

在介绍PLSA之前，需要先了解几个基础的文本段生成模型，包括一元模型，混合一元模型以及PLSA中的文本段生成模型。

#### 一元模型

使用符号w表示一个文本段，且这个文本段一共N个单词。将一个文本段视为一个单词序列，如下

为w中第i个出现的单词。使用表示词的先验概率，那么一元模型中生成文档w的概率为：

一元模型所对应的图模型为

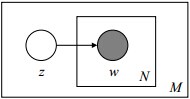


图中被涂色的w表示可观测变量，M表示一共M篇文档。

#### 混合一元模型

混合一元模型中，一个文本段由且仅由一个语义生成。该模型中文本段的生成过程是：给某个文本段先选择一个语义z，再根据该语义生成文档，且该文本段中的所有词都来自该语义。假设有语义，那么生成文档w的概率为

混合一元模型所对应的图模型为



图中被涂色的w表示可观测变量，未被涂色的z表示未知的隐变量，N表示一个文本段中一共N个单词，M表示一共M个文本段。

#### PLSA中的文本段生成模型

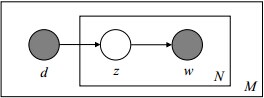
PLSA中的文本段生成模型演变自混合一元模型。在混合一元模型中，一个文本段只由一个语义生成，但在实际情况中，一个文本段往往会有多个语义，只是这多个语义各自在文档中出现的概率大小不一样。

定义一些符号，K表示语义的数量；V表示单词表中词的数量；M表示文本段的数量；表示第i个文本段；表示单词表中第j个单词；表示第k个语义；表示M个文本段中第i个文本段被选中的概率；表示单词表中第j个单词在第i个文本段中出现的概率，这是可以直接计算得到的值，使用在文档中出现的次数除以的总词数便是；表示第k个语义在第i个文本段中出现的概率；表示单词表中第j个单词在第k个语义下出现的概率，与语义关系越密切的词，该值越大。

定义完符号后，PLSA中文本段的生成模型为：

1. 按照概率选择一篇文档；
2. 选定文档后，从语义分布中按照概率选择一个潜在语义；
3. 选定语义后，从词分布中按照概率选择一个词。

所以PLSA中生成文本段的整个过程就是选定文档生成语义，确定语义生成词。PLSA中的文本段生成模型对应的图模型如下



图中被涂色的w和d表示可观测的变量，未被涂色的z表示未知的隐变量。

#### 小结

PLSA所处理的输入是多个文本段，我们使用符号d表示这之中的某个文本段，使用符号w表示单词表中的某个单词，使用z表示某个潜在语义。在PLSA中，有：

P(w, d)，P(d)是可通过直接计算得到的数据。根据上式，可使用最优化方法如最大期望算法得到和，从而得到输入的每个文本段的潜在语义分布。

### 支持向量机

AETML中，需要根据文本段的潜在语义分布推断出其中的情感正负面，这是一个非线性分类问题，AETML使用支持向量机来解决这个问题。本节简要介绍支持向量机的前身逻辑回归，以及支持向量机的主要思想。

#### 逻辑回归

逻辑回归是一种线性分类算法。逻辑回归的输入数据为m个向量，，每个向量是n维的，是n维向量，，以及另外m个数值，。每个有一个对应。逻辑回归的目的的是构建一个映射模型，完成从到的数据映射，从而使用这个模型可以在y未知的情况下，根据x判定y的值。

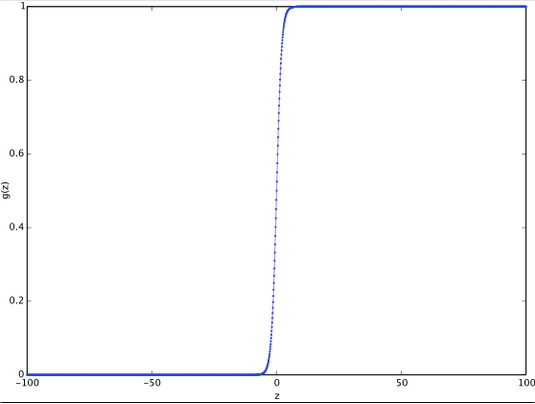
这个过程的图示如下

为了方便，这里讨论二分类的情况，即y的取值只有两种，0或1。

逻辑回归中的映射模型为

式中的w和b是映射模型的参数，图中的训练过程就是确定这两个参数。需要注意的是，w是n维向量，，b是一个数值。式中的g为激活函数，用于将结果映射到0到1之间，激活函数使用sigmoid函数，如下

激活函数的图像如下



由图中可以看出，映射模型产出的结果一定是0到1之间的浮点数，且当大于0时无线趋近于1，当小于0时无限趋近于0。

在介绍对w和b的求解之前，需要对h(x)的效果进行评估，逻辑回归的优化目标是最大化似然函数。

似然函数如下

为了方便，对其取对数，得到新的优化目标

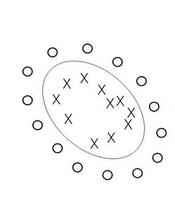
求解使得上式最大的w和b，即完成了训练过程。

对于分布在n维空间的m个数据点，逻辑回归的做法可以认为是在n维空间构建了一个n-1维的超平面来分隔对应y值不同的数据点，对于新的数据点，判断其在超平面的哪一侧，从而判断其对应的y值。

#### 支持向量机简介

上一小节简要介绍了逻辑回归的主要思想。即在n维空间构建n-1维的超平面分隔对应y值不同的数据点。这种做法有一种缺陷，当m个数据点不可能用超平面分隔的时候，即解决非线性分类问题时，逻辑回归作为一个线性分类器，无法发挥作用。

比如，在二维空间中，使用空心圆和叉代表y值不同的m个数据点，假设散点图如下：



此时的超平面是一条直线，显然，对于图中数据点的分布情况，无法使用一条直线分隔。逻辑回归可以将训练数据点映射到高维空间中，从而能够用超平面分隔，如图



但这样直接映射，数据维度会爆炸式增长，计算机无法承受。

AETML使用支持向量机来解决这个问题。支持向量机是90年代中期发展起来的基于统计学习理论的一种机器学习方法，通过寻求结构化风险最小来提高学习机泛化能力，实现经验风险和置信范围的最小化，从而达到在统计样本量较少的情况下，亦能获得良好统计规律的目的。

支持向量机的大致思路与逻辑回归相同，构建一个映射模型，完成从到的数据映射，从而使用这个模型可以在y未知的情况下，根据x判定y的值。如图：

支持向量机与逻辑回归一样，通过构建n-1维的超平面来分隔n维空间对应y值不同的数据点，在超平面一侧是一类数据点，在超平面另一侧是另一类数据点。与逻辑回归不同的是，支持向量机的优化目标不是使得似然函数最大化，而是使得超平面距离最近的数据点的距离最大化。直观来说，好的超平面，也就是分类边界，要距离最近的数据点离远越好，因为这样可以降低分类器的泛化误差。在支持向量机中，分类边界与最近的训练数据点之间的距离称为间隔，支持向量机的优化目标是使得这个间隔最大化。

支持向量机同样面对着逻辑回归同样的问题，超平面如何解决非线性分类问题下的维数爆炸问题。与逻辑回归不同的是，由于优化目标的不同，支持向量机使用核方法可以解决非线性分类问题。具体的支持向量机原理，包括如何使用核方法，在本节的简介中不介绍，后文的算法设计会详细介绍。

## 潜在语义挖掘模块

对于任何一段文本，挖掘出其潜在语义分布是此模块的功能。此模块有两个流程，分别为训练流程和语义分析流程。需要注意的是，语义分析流程依赖训练流程所产出的数据。

### 训练流程

#### 流程简介

此流程的输入是大量的在系统投入运行前已有的情感正负面已知的文本段，此流程的输出是每个文本段中的潜在语义的概率分布和每个语义中的词的概率分布。

训练流程将大量的情感正负面已知的文本段作为语料进行学习，得到每个文本段中的潜在语义的概率分布和每个语义中的词的概率分布。这个过程使用了贝叶斯定理、拉格朗日乘数法、词袋模型、最大期望算法和概率潜在语义分析。

#### 算法详解

首先使用词袋模型，所有文本段都视为一个词袋，忽略其中的词序，仅保留每个词及其在文本段中的出现次数。

然后为了推导方便，定义一些基本符号，N为情感正负面已知的词袋的数量，M为单词的数量，K为低维潜在语义空间的维数，di为第i个情感正负面已知的词袋形式的文本段，wj为第j个单词，zk为潜在语义空间的第k个维度。

我们所要求的是每个词袋中的潜在语义的概率分布P(zk|di)和每个语义中的词的概率分布P(wj|zk)。而我们可以观察得到的数据是每个词袋中的词分布P(wj|di)，在这里使用PLSA模型，有：

定义n(di)表示词袋di中的总单词数量, n(wj, di)表示单词wj在词袋di中出现的总次数，PLSA处理的依据是M\*N个共现数据P(wj, di)，在P(zk|di)和P(wj|zk)已知的情况下，这些数据存在的概率（也称为似然函数）为

为了计算方便，对似然函数取对数，有

继续推演，有

由上式我们可以知道，在似然函数的计算式子中，n(wj, di)和n(di)可观测得到，P(di)在AETML中被设定为1/n。而PLSA的目标则是调整P(zk|di)和P(wj|zk)的值使得似然函数L的值最大化。由于 与所求参数没有关系，且是一个常量，去掉它不影响结果，于是有新的最大化目标：

具体地实现上，AETML使用最大期望算法来完成对上式的最大化。这里将P(zk|di)和P(wj|zk)称为未知的所求参数。而PLSA中存在一个隐藏变量，潜在语义。本文将P(zk|di,wj)，也就是隐藏变量的后验概率称为隐藏的后验参数。最大期望算法的第一步是根据经验初始化未知的所求参数，也就是根据经验初始化P(zk|di)和P(wj|zk)。而具体地实现上，AETML的初始化方法是对于每个未知的所求参数，首先赋予一个大于0小于1的随机值，当所有的未知的所求参数均随机赋值完毕后，进行数据调整使得未知的所求参数满足两个条件：

数据调整的具体做法为

数据调整的具体做法为

完成数据调整后，数据初始化完毕。最大期望算法的下一步称为E步骤，根据当前P(zk|di)和P(wj|zk)的值计算P(zk|di,wj)，再下一步称为M步骤，根据当前的P(zk|di,wj)计算P(zk|di)和P(wj|zk)，然后回到E-Step，循环往复，直到收敛。

E步骤：

计算当前未知的所求参数下的隐藏的后验参数，使用贝叶斯定理，有

M步骤:

在隐藏的后验参数确定的条件下，计算使得似然函数最大化的未知的所求参数P(zk|di)和P(wj|zk)。也就是使得下式最大化：

并且未知的所求参数还有以下约束

使用拉格朗日乘数法处理这种带有约束条件的最优化问题。

引入拉格朗日乘子，得到拉格朗日函数：

求偏导并令其等于0是最优解的必要条件，得到

于是得到在隐藏的后验参数确定的条件下使得似然函数最大化的未知的所求参数P(zk|di)和P(wj|zk)的计算公式

又为了满足约束条件

令

其中

最终得到M步骤的计算公式：

#### 流程图

训练流程的流程图如图4-1所示。



图 4‑1 训练流程流程图

### 语义分析流程

#### 流程简介

语义分析流程的输入是用户使用系统时输入的情感正负面未知的文本段，而输出是该文本段中的潜在语义的概率分布。

语义分析流程以训练流程得到的数据为依据，计算情感正负面未知的文本段中的潜在语义分布，作为AETML服务器模块判定情感正负面的依据。这个过程的主要理论依据是贝叶斯定理。

#### 算法详解

对于情感正负面未知的文本段，仍使用词袋模型，将其视为一个词袋。定义一些符号，M是单词表中的单词数量；K是潜在语义空间的维数；N是情感正负面已知的词袋的数量；zk代表潜在语义空间第k个维度；d代表这个词袋；P(zk|d)代表这个词袋在潜在语义空间第k个维度上的值，也是所求值。

对于一个词袋所代表的潜在语义，AETML将其设定为词袋中所有词的语义叠加，即

上式中，是未知量是，使用贝叶斯定理计算这个未知量，有

是训练流程所计算得到的数据，而使用训练流程中情感正负面已知的词袋中的单词出现频率，的计算式子如下

于是，对一个情感正负面未知的词袋，其潜在语义的计算式子为

#### 流程图

语义分析流程流程图如图4-2所示。



图 4‑2 语义分析流程流程图

## 情感规则学习模块

### 流程简介

流程输入：情感正负面已知的文本段中及其潜在语义分布。

流程输出：潜在语义分布与情感正负面之间的数据映射规则。

此模块依托于潜在语义挖掘模块产生的数据，也就是情感正负面已知的文本段，对潜在语义分布之间与情感正负面之间的规则进行学习，并产出映射规则，传递给服务器模块进行使用。此模块的主要涉及到的理论包括拉格朗日乘数法和支持向量机。

### 算法详解

首先定义一些符号，符号N表示潜在语义空间的维度；符号M表示情感正负面已知的文本段的数量；使用符号表示第i个文本中的潜在语义分布；使用符号表示的第j个维度；使用符号表示第i个文本的情感正负面，如果是正面，则为1，如果是负面，则为-1。

为了方便阐述SVM的思想，假设N=2，那么所有的都是二维空间中的一个数据点，使用空心圆表示代表正面情感的数据点，使用叉代表表示负面情感的数据点，且这些数据点在二维空间中的散点图如图4-3所示。

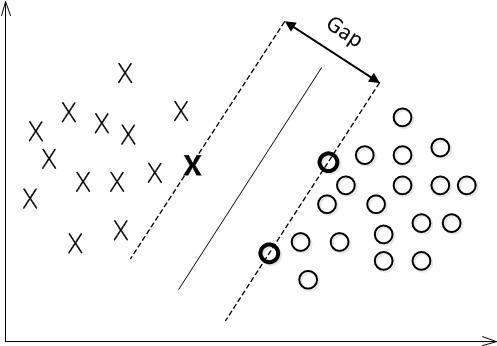


图 4‑3

SVM的目标是在N维空间中构建一个超平面将不同情感的数据点区分，在二维空间中，超平面是一条直线，如图黑色实线，实线一侧是正面情感的数据点，另一侧是负面情感的数据点。SVM的优化目标是使得超平面距离最近数据点的间隔最大化，也就是使得图中的Gap/2最大化。与超平面距离最近的数据点，称为支持向量，图中加粗的两个空心圆和一个叉就是支持向量。

下文中，使用类别1代表正面情感，使用类别-1代表负面情感。回到原问题，对于数据点，需要构造N维空间中的超平面，其方程为

为了方便，设定。则超平面的方程等价于

在超平面上的数据点x满足；在超平面一侧的点满足，这些点被SVM归为一类；在超平面另一侧的点满足，这些点被SVM归为另一类。为了方便，设定满足的数据点的类别为-1；满足的数据点的类别为1。

如上所述，现在SVM的目标是调整w和b的值使得间隔最大化。为了解决这个问题，首先需要计算数据点到超平面的距离，为了方便，令，则就是超平面的方程，且w是这个超平面的一个法向量，设数据点x在超平面的垂直投影点为，有

式中的d就是点x到超平面的距离，由于，于是

于是得到

对于距离来说，我们只关心绝对值，于是点到超平面距离的最终计算式为

对于数据点，由于点到超平面的距离的计算式子中对每个数据点都是同样的值，所以我们先计算所有数据点对应的的最小值，再用其除以就是所有点到超平面的最小距离，根据定义有

此时，我们的目标是最大化分类间隔，也就是最大化。

对于超平面 ，对w和b乘以一个常数c后代表的超平面不变，即和代表同一个超平面，因此，使用这种方法，我们可以使得。

此时，SVM的优化问题变为

由于 与 等价，且，得到新的等价优化目标

这是一个有条件约束下的最优化问题，使用前文所描述的拉格朗日乘数法，引入M个拉格朗日乘子，构造拉格朗日函数

和前面讨论的一样，这里新的优化目标变为

其对偶问题为

由于原优化问题为凸优化，且存在数据点使得，满足slater条件，所以满足强对偶性，即原问题等价为对偶问题，即

所以得到新的优化目标

第一步先求，计算其偏导

令偏导等于0，得到

代入到原式中

优化问题转换为求解

使用SMO算法可以求解出所有的拉格朗日乘子，从而可以得到超平面的方程。

在阐述SMO算法之前，SVM还有一个必须要解决的问题。回到二维空间的例子，情况稍有变化，假设数据点在二维空间的散点图如图4-4所示。

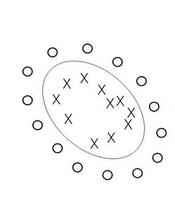


图 4‑4

仅使用一个超平面，也就是一条直线，是无法区分两类数据点的。

SVM将原始数据点映射到更高维空间中，如图4-5所示。



图 4‑5

在高维空间中构造超平面，从而解决这个问题。在实际应用时，若将原始数据直接映射到高维空间，数据维度会呈现爆炸式增长，计算机无法承受。所以，SVM采用核方法在不直接映射的情况下构造超平面。

回想SVM判定数据点类别的方法，对于数据点x，计算，大于0归为类别1，小于0归为类别-1。而根据前文的结论，有

是数据点x与的向量内积。且根据前文结论，如果不是支持向量，那么，所以，有

假设映射到高维空间的映射为，那么判定类别的计算式子为

同理，SVM的优化目标变为

因为在高维空间中，我们只需计算数据点间的内积，如果能够在低维空间完成这样的计算，就能解决维度爆炸问题。

核函数k的定义如下

使用低维空间的数据完成高维空间的向量内积的计算，完成高维映射的功能的同时，避免了维度爆炸问题。不同的映射方式对应着不同的核函数，根据选取的核函数的不同，将原始数据空间映射到了不同的高维空间。

通过核函数映射到高维空间，SVM解决了图中描述的问题，新的优化目标为

在实际应用时，SVM还需考虑离群点的存在，如果不做任何更改使用前文的方法，离群点的存在会使得SVM的结果非常糟糕。

回想原始的优化目标为

为了解决离群点，SVM允许数据点在一定程度上偏离超平面，在原始优化目标上的改动如下

C是人为确定的常数，将原始优化目标修改后，按照前文一样的流程推演，最终得到优化目标

与处理离群点前的优化目标唯一的不同是拉格朗日乘子存在一个上限。

接下来，阐述求解M个格朗日乘子的SMO算法。

首先考虑拉格朗日乘子的三个情况。如果，根据前文结论数据点应为支持向量或者在支持向量后方，即；如果，数据点应在两类别的支持向量之间，即；如果，数据点应为支持向量，即。

SMO首先将所有的拉格朗日乘子初始化为0，从而使得满足约束。接着，SMO选出一个不满足前文三个条件的拉格朗日乘子，然后可选取使得下式最大的

同时更新这两个乘子，使其满足

从而使得更新后仍满足。

分两种情况讨论详细的更新过程。

第一种情况，当时，为了方便，令，k在此步骤中为已知的中间量。在此步骤中，除和之外的拉格朗日乘子由于此步骤不进行操作，视为常数，因此此步骤的优化目标为

由于，得到新的最大化目标

令，在此步骤中为已知的中间量，并去掉优化目标中无关的k项，得到新的优化目标

变换形式

这是一个二次函数，所以可以直接求得使得目标最优化的的值。需要注意的是，这里的还存在一个约束，如下

又有

的下界为,上界为。由于二次函数的性质，可直接计算出最优情况下的。计算完毕后，使用式子计算。

第二种情况，当时，为了方便，令，k在此步骤中为已知的中间量。与前文同理，此步骤的优化目标为

由于，得到新的优化目标

令，在此步骤中为已知的中间量，并去掉优化目标中无关的k项，得到新的优化目标

变换形式

这是一个二次函数，可以直接求得使得目标最优化的的值。需要注意的是，这里的还存在一个约束，如下

又有

的下界为，的上界为。由于二次函数的性质，可以直接求得的值。计算完毕后，使用式子计算。

至此，SMO的一次更新完毕。当所有不满足条件的拉格朗日乘子均更新完毕后，使用下式计算出超平面的w

最后还剩下b值，就可以确定超平面了。

设1类别的数据点中拒与超平面距离最近的点为，则其到超平面的距离为；设-1类别的数据点中拒与超平面距离最近的点为，则其到超平面的距离为。

超平面的选取应使得这两个距离相等，于是有

解得

显然，在w已知b未知的情况下，是使得最小的数据点，是使得最大的数据点。

至此，超平面的参数最终全部确定，将此超平面作为潜在语义分布与情感正负面之间的映射规则传递给服务器模块进行使用。

### 流程图

情感规则学习模块流程图如图4-6所示。



图 4‑6 情感规则学习模块流程图

## 服务器模块

### 情感判定模块

#### 流程简介

流程输入：情感正负面未知的文本段。

流程输出：文本段的情感正负面。

此模块依托于情感规则学习模块产生的数据，也就是潜在语义分布到情感正负面的映射规则，判断文本段的情感正负面。此模块的主要涉及到的理论为支持向量机。

#### 算法详解

首先使用潜在语义挖掘模块，得到情感正负面未知的文本段中的潜在语义分布，从而得到潜在语义空间中的数据点x。

回想情感规则学习模块中所得到的规则的形式为一个超平面

令

对于潜在语义空间中的数据点x，计算f(x)，若大于0，则判定为正面情感，若小于0，则判定为负面情感。

若越大，可认为数据点距离超平面越远，也就是分类可信度越高。所以可以使用作为可信度，如下

通过上式，得到0到1之间的浮点数作为可信度。

#### 流程图

情感判定模块的流程图如图4-7所示。



图 4‑7 情感判定模块流程图

### 数据交互模块

此模块监听作为服务器的计算机的端口，当有客户端进行TCP连接时，从连接中获取情感正负面未知的文本段，并将情感判定结果通过连接传回客户端。

数据交互模块的流程图如图4-8所示。



图 4‑8 数据交互模块的流程图

## 客户端模块

### 数据交互模块

TCP连接服务器，发送情感正负面未知的文本段，并接收情感判定结果，交由情感判定结果展现模块展示给用户。

数据交互模块流程图如图4-9所示。



图 4‑9 数据交互模块的流程图

# 系统实现与评估

## 实现环境

本系统使用的开发平台为Linux机器，操作系统发行版信息为Red Hat Enterprise Linux AS release 4 (Nahant Update 3)， 安装内存为189GB，磁盘总容量为21.6TB。

使用C++、Python、Php、HTML、Javascript作为程序实现语言。使用的c++编译器版本为g++ 4.8.2，使用的Python解释器版本号为2.7.9，后端的开发环境为vim+gdb，前端的开发环境为xampp+sublime。

## 效率评估

在本系统的所有细分模块中，对运行效率要求较高的，有潜在语义挖掘模块的子模块中文分词模块和已知文本的语义挖掘模块，以及情感规则学习模块，其余模块实际应用时对效率要求较低。

### 中文分词模块

使用39MB大小的UTF8编码的语料，共9815个文本段，对其进行分词。

运行时间为7.6s，运行截图如图5-1所示。

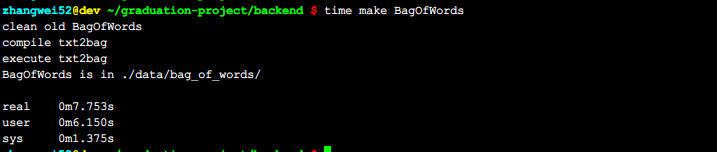


图 5‑1

分词速度为5.03MB/s，系统可接受范围。

### 已知文本的语义挖掘模块

对于39MB大小的UTF8编码的共9815个文本段，对其进行潜在语义分析。

运行时间约为54分钟，运行截图如图5-2和图5-3所示。

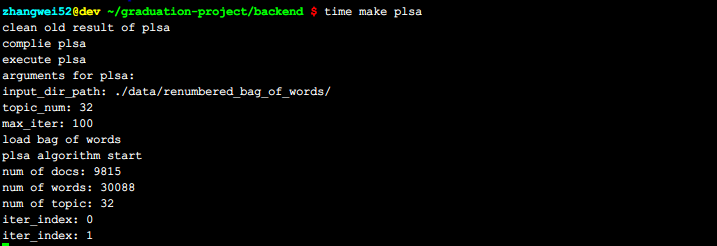


图 5‑2

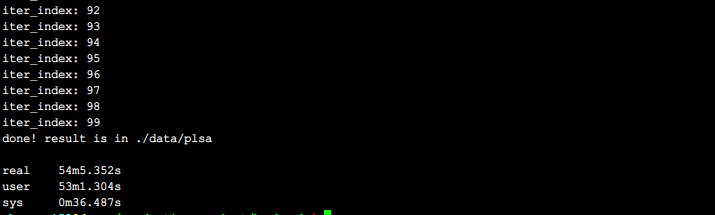


图 5‑3

由于此过程是一次性的，此速度在本系统可接受范围内。

### 情感规则学习模块

本模块处理9815个情感正负面已知的文本段所对应的潜在语义的概率分布，学习从潜在语义分布到情感正负面的数据映射规则，耗时约33秒，运行截图如图5-4和图5-5所示。

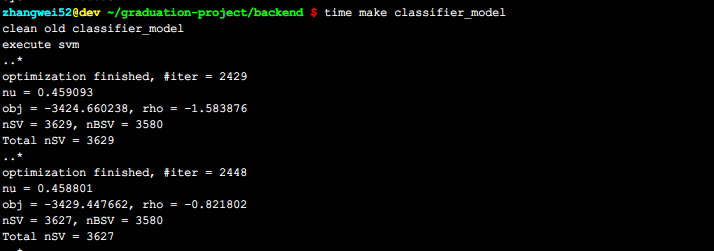


图 5‑4

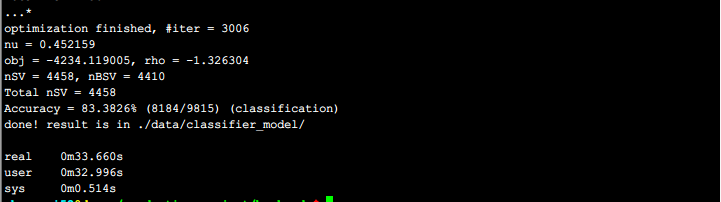


图 5‑5

## 效果评估

### 训练集

使用9815条系统用于情感规则学习模块使用的情感正负面已知文本段作为训练集对本系统进行测试。其中，人工判定为正面且系统判定为正面的文本段有6830条，人工判定为正面且系统判定为负面的文本段有71条，人工判定为负面且系统判定为正面的文本段有2129条，人工判定为负面且系统判定为负面的文本段有785条。

如图5-6所示。

图 5‑6 训练集测试结果

可以得到，准确率为78%。

### 测试集

使用五十条被人工判定为正面情感和五十条被人工判定为负面情感的文本段作为测试集对本系统进行测试。其中，人工判定为正面且系统判定为正面的文本段有48条，人工判定为正面且系统判定为负面的文本段有2条，人工判定为负面且系统判定为正面的文本段有24条，人工判定为负面且系统判定为负面的文本段有26条。

测试集测试结果如图5-7所示。

图 5‑7 测试集测试结果

可以得到，准确率为74%。

## 结果展示

系统提供web操作界面，如图5-8所示。

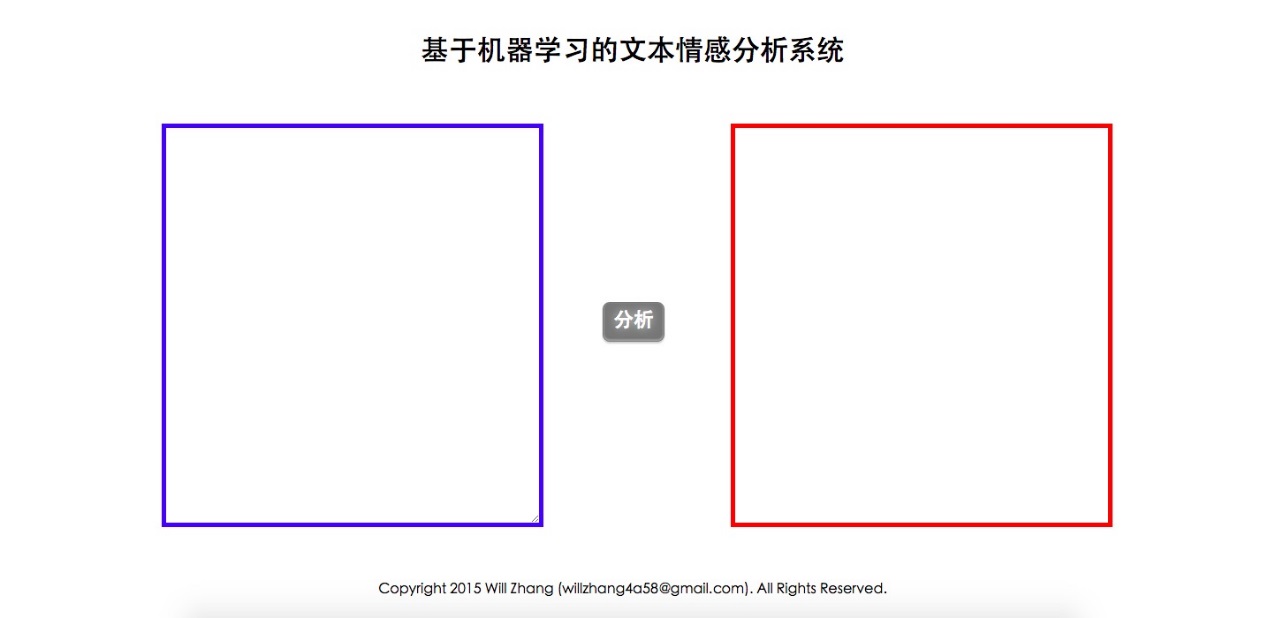


图 5‑8 操作界面

在左侧文本框输入“越来越艰难了”，点击分析，得到文本情感分析结果。如图5-9所示，系统判定为负面情感。且可信度为66%。

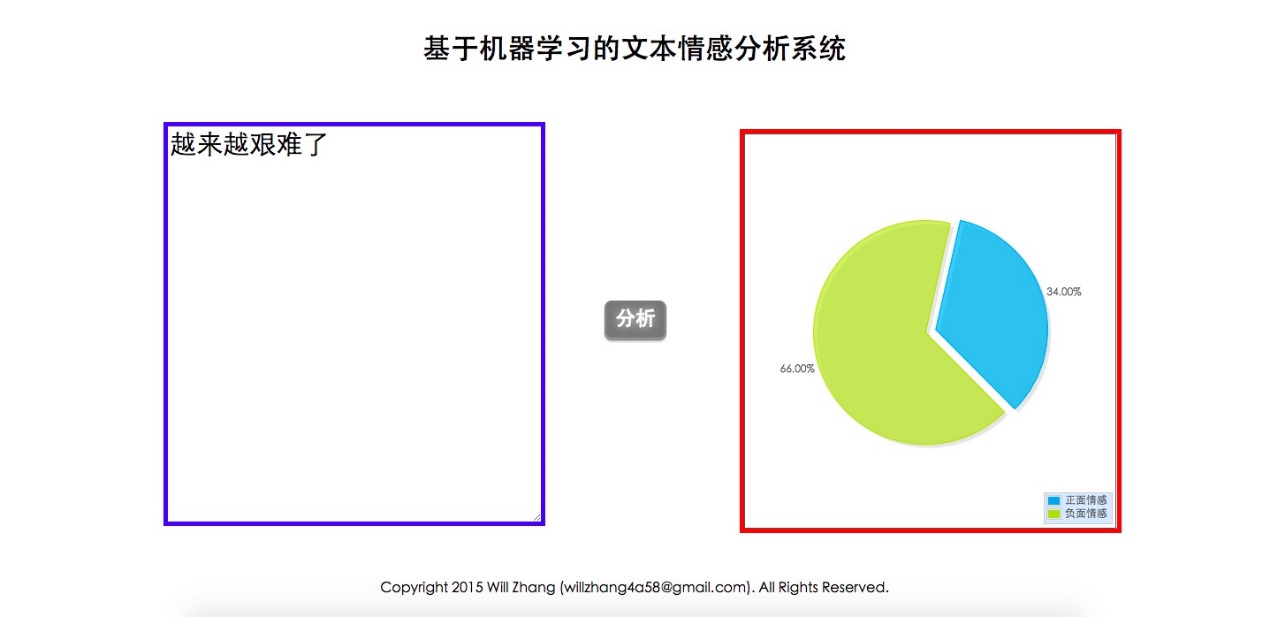


图 5‑9 情感分析结果

# 总结与展望

本文完成了基于机器学习的文本情感分析系统的设计与实现。该系统主要分为潜在语义挖掘模块、情感规则学习模块、服务器模块和客户端模块，系统首先对情感正负面已知的文本段进行潜在语义挖掘，随后学习潜在语义到情感正负面的数据映射规则，将其规则用于判定文本段的情感正负面，使用B/S结构完成一个完整的系统。

在今后的工作中，希望能发现比支持向量机更适合本系统的分类算法，使用比潜在语义的概率分布更有效的数据特征，从而提高情感规则学习模块的准确率，进而提高情感判定的准确率。

总的来说，本系统比较理想的达到了预期的目的，但还有些缺陷，对于情感正负面已知的文本段这一数据的依赖较大，这份数据的质与量对于系统最终的情感分析效果影响极大。为了解决这一依赖，在今后还要继续虚心、努力地学习和钻研。

# 致谢

行文至此，论文已接近尾声。岁月如梭，四年的大学时光也即将结束。离别在即，站在人生的又一个转折点上，心中难免思绪万千，感恩之情油然而生，在此表示我最衷心的感谢。

感谢我的指导老师胡晶晶老师，本文的写作直接得益于她的悉心指点，不厌其烦地帮助进行论文的修改和改进，提出了很多宝贵的意见，为我的工作提供了极大的帮助和支持。

感谢实习期间结识的石立臣博士，在课题立项上给予了极大的帮助；感谢实习期间的导师刘伟硕士，对本文的关键理论提出了丰富的见解和看法；感谢实习时同组的颜冲、杨诚诚、李宗辉硕士，对本系统的性能调优提出了宝贵的意见。

感谢在科技创新基地学习期间给予我极大照顾和帮助的陈杰浩老师、史继筠老师，他们工作勤恳认真，业务水平精湛，让我学习到很多做人做事方面的道理，使我受益良多。

感谢已经毕业的张桉笛师兄、符基高师兄、李雪鹏师兄、刘鹏师兄、白牧师兄、王薄师兄、周扬森师兄和刘乙墨师兄，他们在我大学期间的学习、工作、生活方面给予了我极大的关怀和帮助，让我受益匪浅。

在此感谢北京理工大学软件学院和百度在线网络技术有限公司给予我的悉心栽培，向所有关心、帮助、鼓励过我的老师、同学、同事和亲人表达最诚挚的谢意！

# 参考文献

[1] 赵妍妍,秦兵,刘挺. 文本情感分析[J]. 软件学报,2010,08:1834-1848.

[2] Hofmann T. Probabilistic latent semantic indexing[C]//Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999: 50-57.

[3 ]Platt J. Fast training of support vector machines using sequential minimal optimization[J]. Advances in kernel methods—support vector learning, 1999, 3.

[4] Joachims T. Introduction to support vector machines[J]. 2002.

[5] N. Coccaro and D. Jurafsky. Towards better integration of semantic predictors in statistical language modeling. In Proceedings of ICSLP-98, 1998. to appear.

[6] S. Deerwester, S. T. Dumais, G. W. Furnas, Landauer. T. K., and R. Harshman. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41, 1990.

[7] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. J. Royal Statist. Soc. B, 39:1{38, 1977.

[8] P.W. Foltz and S. T. Dumais. An analysis of information ltering methods. Communications of the ACM, 35(12):51{60, 1992.

[9] T. Hofmann. Probabilistic latent semantic indexing. In Proceedings of SIGIR'99, 1999.

[10] T. Hofmann, J. Puzicha, and M. I. Jordan. Unsupervised learning from dyadic data. In Advances in Neural Information Processing Systems, volume 11. MIT Press, 1999.

[11] T.K. Landauer and S.T. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2):211{240, 1997.

[12] R.M. Neal and G.E. Hinton. A view of the EM algorithm that justies incremental and other variants. In M.I. Jordan, editor, Learning in Graphical Models, pages 355{368. Kluwer Academic Publishers, 1998.

[13] F.C.N. Pereira, N.Z. Tishby, and L. Lee. Distributional clustering of english words. In Proceedings

of the ACL, pages 183{190, 1993.

[14] K. Rose, E. Gurewitz, and G. Fox. A deterministic annealing approach to clustering. Pattern Recognition Letters, 11(11):589{594, 1990.

[15] G. Salton and M. J. McGill. Introduction to Modern Information Retrieval. McGraw{Hill, 1983.

[16] L. Saul and F. Pereira. Aggregate and mixed order Markov models for statistical language processing. In Proceedings of the 2nd International Conference on Empirical Methods in Natural Language Processing, 1997.