分类号 TP311 学号 06060049

UDC 密级 公 开

工学硕士学位论文

关键词抽取的研究与实现

硕士生姓名 罗准辰

学 科 专 业 计算机科学与技术

研 究 方 向 计算机软件与理论

指 导 教 师 王挺 教授

国防科学技术大学研究生院

二〇〇八年十一月

关键词抽取的研究与实现

国防科学技术大学研究生院

The Research and Implementation of Keyword Extraction

Candidate：Luo Zhunchen

Advisor：Prof. Wang Ting

A thesis

Submitted in partial fulfillment of the requirements

for the degree of Master of Engineering

in Computer Science and Technology

Graduate School of National University of Defense Technology

Changsha，Hunan，P.R.China

November，2008

独 创 性 声 明

本人声明所呈交的学位论文是我本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表和撰写过的研究成果，也不包含为获得国防科学技术大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文题目: _____

学位论文作者签名: _____ 日期: 年 月 日

学位论文版权使用授权书

本人完全了解国防科学技术大学有关保留、使用学位论文的规定。本人授权国防科学技术大学可以保留并向国家有关部门或机构送交论文的复印件和电子文档，允许论文被查阅和借阅；可以将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、汇编学位论文。

（保密学位论文在解密后适用本授权书。）

学位论文题目: _____

学位论文作者签名: _____ 日期: 年 月 日

作者指导教师签名: _____ 日期: 年 月 日

目　录

摘　要 i

ABSTRACT ii

第一章 绪论 1

1.1 概述 1

1.2 关键词的应用 2

1.3 关键词抽取面临的主要问题 2

1.4 本文主要工作 4

1.5 本文结构 4

第二章 相关方法介绍 6

2.1 关键词抽取与相关任务比较 6

2.1.1 关键词抽取与自动摘要 6

2.1.2 关键词抽取与信息抽取 6

2.1.3 关键词抽取与自动索引 7

2.2 关键词抽取研究现状 7

2.2.1 关键词抽取中“关键”问题研究现状 7

2.2.2 关键词抽取中“词”问题研究现状 10

2.3 小结 12

第三章 关键词分类问题 13

3.1 关键单词的定义 13

3.2 关键词串的定义 13

3.3 小结 14

第四章 基于分离模型的关键词抽取算法 15

4.1 分离模型的构造 15

4.2 候选关键单词与候选关键词串的生成 17

4.2.1 英文中候选关键单词与候选关键词短语的生成 17

4.2.2 中文中候选关键单词与候选关键词短语的生成 18

4.3 模型的训练与 SVM 学习器 18

4.4 关键词的抽取 20

4.5 小结 22

第五章 分离模型的特征设计 23

5.1 关键单词与关键词串公共特征设计 23
5.1.1 TF×IDF 特征 23
5.1.2 首次出现位置特征 POS 24
5.1.3 TF×IF 特征 25
5.1.4 文档长度特征 NWT 26
5.2 关键单词特征设计 26
5.3 关键词串特征设计 27
5.3.1 互信息特征 27
5.3.2 词串边界参数表特征 28
5.3.3 候选关键词串结尾词词性特征 29
5.3.4 候选关键词串开头词词性特征 29
5.3.5 候选关键词串非结尾词中非形容词非名词的数目 29
5.3.6 候选关键词串所含词数 29
5.4 小结 30
第六章 实验与分析 32
6.1 实验方法 32
6.1.1 分类实验 32
6.1.2 评分实验 33
6.1.3 语料介绍 33
6.2 分离模型与整体模型比较 33
6.2.1 英文中分离模型与整体模型比较 33
6.2.2 中文中分离模型与整体模型比较 36
6.3 关键单词特征与关键词串特征的作用 39
6.3.1 英文中关键单词特征与关键词短语特征实验 39
6.3.2 中文中关键单词特征与关键词串特征实验 41
6.4 与 KEA 的比较实验 43
6.4.1 与 KEA 在英文关键词抽取上的比较 43
6.4.2 与 KEA 在中文关键词抽取中的比较 44
6.5 小结 45
第七章 结束语 46
致 谢 47
参考文献 48
作者在学期间取得的学术成果 52

表 目 录

表 5.1 特征基本信息.....30
表 6.1 语料基本信息.....33
表 6.2 关键单词评分器与整体评分器比较.....35
表 6.3 关键词串评分器与整体评分器比较.....35
表 6.4 综合评分器与整体评分器比较.....36
表 6.5 分类实验训练集中正例与反例的具体数目.....37
表 6.6 分类实验候选关键单词测试结果.....37
表 6.7 分类实验候选关键词串测试结果.....37
表 6.8 Blog 语料整体模型与分类模型比较.....38
表 6.9 英文中相关评分器特征设计.....39
表 6.10 添加关键单词特征实验.....40
表 6.11 添加关键词串特征实验.....41
表 6.12 分类实验添加特征候选关键单词测试结果.....41
表 6.13 分类实验添加特征候选关键词串测试结果.....41
表 6.14 中文中相关评分器特征设计.....42
表 6.15 Blog 添加特征实验.....42
表 6.16 综合评分器特征设计.....43

图 目 录

图 4.1 分离模型的训练与抽取过程.....16
图 4.2 结构风险最小化思想.....19
图 4.3 综合评分器基本原理.....21
图 5.1 词串边界参数表构造过程.....28
图 6.1 关键单词评分器与整体评分器比较.....34
图 6.2 关键词短语评分器与整体评分器比较.....34
图 6.3 综合评分器与整体评分器比较.....36
图 6.4 Blog 语料整体模型与分类模型比较.....38
图 6.5 添加关键单词特征实验.....40
图 6.6 添加关键词短语特征实验.....40
图 6.7 Blog 添加特征实验.....42
图 6.8 输出前 5 个候选关键词比较图.....44
图 6.9 输出前 15 个候选关键词比较图.....44
图 6.10 与 KEA3.0 的比较.....45

摘 要

关键词抽取在自动文摘、信息检索、文本分类、文本聚类等方面具有十分重要的作用。但实际中只要很少一部分文档拥有作者标注的关键词，手工添加关键词是一项繁重的工作。因此非常需要一种方法能够自动抽取关键词。许多学术期刊要求论文作者在论文的第一页列出大约 5 到 10 个关键词，而这些关键词经常是一些包含两个甚至更多词的短语，我们习惯上将其成为关键词串。而且除了论文外，其它大部分文档中的关键词也是一些短语词串，而这部分的关键词抽取是一个很难的过程。

本文提出将关键词抽取分为两个问题进行处理：关键单词抽取和关键词串抽取，并设计了一种基于分离模型的关键词抽取方法。该方法针对关键单词抽取和关键词串抽取这两个问题设计不同的特征以提高抽取的准确性。本文将关键词抽取看成一个有监督学习问题，将每篇文档处理以形成一组词或词串集合，然后通过机器学习的方法对这些词或词串分类，作为关键词的正例或关键词的反例。在特征设计上，我们针对关键单词与关键词串在结构上的不同特点设计了许多特征。比如，通过互信息与词串边界参数表特征提高词串的识别率；根据关键单词与关键词串词性组合的规律，设计了一些语言学特征以提高抽取关键单词与关键词串的效果。

在上述工作的基础上，我们用实验验证了基于分离模型的关键词抽取方法的有效性。实验结果表明，在特征选取相同的情况下，基于分离模型的关键词抽取方法优于基于整体模型的关键词抽取方法。另外我们还验证了针对关键单词与关键词串所设计的特征的有效性。最后我们将针对关键单词与关键词串所设计的不同特征分别添加到分离模型后所形成的关键词抽取器与著名的关键词抽取工具 KEA 进行了比较实验，实验结果显示，我们的关键词抽取器对于关键词的抽取效果好于 KEA。

关键词：关键词抽取，关键词串，分离模型，互信息，词串边界参数表，特征选取，机器学习，语言学特征

ABSTRACT

Keywords are widely used in many applications such as Information Retrieval, Automatic Summarizing, Text Classification, Text Clustering and so on. Only a small minority of documents have author-assigned keywords, and manually assigning keywords to documents is very laborious. Therefore it is highly desirable to automate the keyword extraction process. Many academic journals require their authors to provide a list of about five to fifteen keywords on the first page of each article. Since these keywords are often phrases consisting of two or more words, we prefer to call them key phrases. Most of the keywords from other kinds of documents are also actually phrases, which make the task more difficult.

This paper argues that the keywords extraction can be treated as two problems: extracting key words and extracting key phrases. A keywords extraction method based on separating models was proposed for extracting keywords from the documents. This method develops different features for the two mentioned problems in order to improve the accuracy. This paper also considers the problem of automatically extracting keywords from text as a supervised learning task. We treat a document as a set of words or phrases, which the learning algorithm must learn to classify as positive or negative examples of keywords. Based on the different structure of the key words and key phrases, we develop a set of features. For example using the features of mutual information and parameter table of word-sequence boundary can improve the phrases identification. We also use the part-of-speech rule of key words and key phrases to develop some linguistic features to improve the result of extracting key words and key phrases.

Based on the above work, we run the experiment to evaluate the effect of the keyword extraction method based on separation model. The result shows that, using the same features, the performance of keyword extraction algorithm based on separation model is better than that based on integrated model. In addition we also evaluated the effect of features for key words and key phrases. At last, to compare the work with the famous keywords extractor KEA, we implemented an keywords extractor based on separation model adopting different key words' features and key phrases' features. The result shows that our extractor is better than KEA.

Key Words：keyword extraction, key phrases, separating models, mutual information, parameter table of word-sequence boundary, feature selection, machine learning, linguistic feature

第一章 绪论

1.1 概述

在上个世纪 80 年代的电视剧《霹雳游侠》中，有一部被称为基特(KITT)的汽车，它具有神奇的魔力，具有 20 世纪最受人们认可的一些特征。片中基特(KITT)是一个具有高级语言处理能力并且能够说话和理解语言的智能汽车。

“请别叫我‘汽车’或是‘几只轮子’，我是奈特工业 2000 号”呵呵，KITT 生气了，这样的台词竟然不可思议的清晰的留在我的记忆里。

二十年过去了，现在我们离这样会理解人的语言的智能体还有多远？现在我们到底还应该做些什么呢？我认为像基特(KITT)这样的智能体至少能够通过语言与人类交流。其中包括通过语音识别和自然语言理解来与人类沟通，通过自然语言生成和语音合成来与人类交际。它也能够通过信息检索发现所需要的文本资源在哪里，利用信息抽取技术从文本中抽取需要的信息。最后利用知识进行知识推理[1]。

二十多年来，随着信息时代的发展，信息的表达方式也多种多样。其中以自然语言形式表达信息的文本是一种不可替代的方式。随着网络上文本信息的爆炸式增长，如何提高信息访问的效率成了一个越来越重要的课题。为了对海量信息进行有效地组织、压缩和检索，研究人员在自动文摘、信息检索、文本分类、文本聚类和关键词自动抽取等方面进行了大量研究。但人们想获取自己需要的文本信息不可能从大量的文本中手工得到。因此如何组织、管理、检索文本变得越来越重要。

随着 Google、Baidu 等搜索引擎网站的推广与应用，信息检索技术越来越深入人们的生活。用户通过输入关键词，引擎可以自动返回出现此关键词的网页。看上去搜索引擎似乎无所不能，不但可以从海量信息中获取需要的信息，而且查询时间也很短。事实果真如此吗？它真的无所不能吗？如果用户输入的关键词不够准确或者输入的关键词不在相关网页中出现，那么搜索引擎的检索效果将大打折扣！即便所有的相关网页通过关键词都被检索出来，引擎返回的网页结果排序也是一个亟待解决的问题，因为排在前面的网页是不是与你想要搜索的内容最相关也是一个未知数！

信息抽取不同于信息检索，它首先把信息进行结构化处理，变成表格一样的组织形式，形成一种映射关系。通过输入原始文本信息，得出相关信息点。本质上信息检索的核心技术是采用关键词匹配，而信息抽取则需要更深层次的自然语言理解。虽然它们是两个不同的概念，但两者又是紧密联系的。利用信息抽取技术提高文档检索的粒度与精度，从而整体上提高检索的准确率。当然，信息抽取技术同样可以利用信息检索技术计算一些统计特征，如词频、TFIDF 值等，提高信息抽取的准确率。总之两者相辅相成，缺一不可。

信息检索与信息抽取有一条公共的纽带——关键词。

1.2 关键词的应用

关键词高度概括了文本的主要内容，很容易使不同的读者判断出文本是不是自己需要的。1.1 节提到的在信息检索中利用关键词进行文本或网页的检索，是关键词应用的一个重要领域。当然关键词在文档管理、文本分类和聚类、自动摘要等方面应用也十分广泛[22]。

1. 文档管理

文档管理指文档、电子表格、图形和影像扫描文档的存储、分类和检索。每个文本具有一个类似于索引卡的记录，记录了诸如作者、文档描述、建立日期和使用的应用程序类型之类的元信息。

对于不同领域的文档，诸如电子文档，可以根据领域的不同内容，抽取领域关键词，依次建立关键词索引，提高文档管理的效率。

2. 文本分类和聚类

文本分类和聚类的目标就是将语义相近的文本分成或聚成一类。二者不同之处在于文本分类首先必须定义好类别，然后文档集合按定义好的类别进行分类；而文本聚类不需要预先确定分类的类别，而是根据聚类算法自动聚集成类，聚成多少个类就有多少个类，因此具有一定的灵活性和较高的自动化处理能力。

由于关键词十分精练，因此可以利用关键词以很小的计算代价进行文本相关性度量，从而进行文本分类与聚类。

3. 自动摘要

摘要又称概要、内容提要。摘要是以提供文献内容梗概为目的，不加评论和补充解释，简明、确切地记述文献重要内容的短文。

自动摘要是利用计算机对输入文章自动地进行概要，而关键词抽取往往是许多摘要算法的核心技术。传统的机械摘要方法就是首先抽取关键词，然后以关键词出现的次数计算每个句子的权重，以此判定每个句子是否成为摘要的组成部分。

1.3 关键词抽取面临的主要问题

虽然关键词应用十分广泛，但关键词抽取存在两个基本的问题：

1. 什么是“关键”

在 1998 年版《现代汉语词典》中，所谓“关键”比喻事物最关紧要的部分，或对情况起决定作用的因素。那么对应于文本关键词中的“关键”即为文本中最关紧要、起决定作用的意思。那么如何定量分析文本中词或词序列的“关键”特性是关键词抽取技术中一个重要的问题。

2. 什么是“词”

所谓“词”是语言中最小的、可以自由运用的单位。虽然关键词可以概括文本的主要内容，但大部分关键词都取自文本的词一级，无法发现一些有意义、关键性的短语。短语有许多特点，它比词概括能力更强，信息更加丰富。

另外语言的差异也使得关键词抽取面临不同的问题。汉语不同于英语，汉语文本中的词无天然分隔符，关键词抽取技术大都先依赖词典分词，因此所有词一级的关键词都是词典词，这就造成无法发现一些未收录在词典中的词——未登录词。未登录词基本上可以分为两大部分：

（1） 新词， 即新涌现出来的通用词汇和各行业的专业术语等，如神七、姚黑、超女等。

（2） 专有名词，包括人名、外文译名、地理名称、机构名称等，如刘翔、布什、汶川等。

本文在研究关键词抽取问题中主要对第一种未登录词，即新词的识别进行探讨。第二类词的识别参见命名实体识别的相关工作[2]。

新词的识别是自然语言处理、信息检索和机器翻译等领域的一项基础研究。新词的数量是难以用数字来衡量的，特别是对于某些行业领域，比如对生物科技、信息技术等新兴领域而言，新词将越来越多。如何从庞大且无序的信息中辨别出有意义的新词，也成为当代信息工作的重要研究内容之一。

根据新词的构成方式，新词分为 4 类:

（1） 缩写词：如非典(非典型肺炎)、边警(边防警察)、抗非(抗击非典)。缩写词的构词方式很不规则，使用词中的某个字表示词的含义，但该字的选择有时候只是一种约定或者习惯，所以缩写词的检测是很难的。

（2） 派生词：如垂直化、价值型。这类词有比较明显的词缀语素。

（3） 复合词：如扑杀、现金流。通过复合会产生大量的新词。复合方式多种多样，汉语中很多活跃的字都可能作为复合词的元素，这类词的识别也是非常难的。
单纯词：如肯德基、麦当劳。该类词的意义与单字的意义完全无关，包括音译词等。

就目前而言，新词识别的困难主要在于新词往往被分词系统切分成单字串或者单字与基本词汇的组合；不同于专有名词的识别，新词基本上没有一个比较普遍的规律；对于低频的新词识别尤其困难；对于“旧词新用”，具有词义、用法变化的词语检测也是很难解决的问题。

因此“关键”与“词”两个问题是本文需解决的两个核心问题。

在“关键”方面，如果在进行候选词“关键”性质度量时，把旧词、新词、短语一起

作为候选关键词进行度量，可能因为新词与短语识别的困难，效果不好，也影响“关键”特征的度量。

在“词”方面，通过分词、新词识别、短语识别找到文本中所有可以成为关键词的有意义词序列。新词识别中多数统计的方法能否应用于短语的识别，能否将新词与短语看成一个整体，都是本文需要探讨的问题。

当然，关键词抽取还有许多工作，比如如何度量关键词抽取的好坏；如果把关键词抽取看成分类问题，如何减少训练反例过多的问题；机器学习是否适合关键词抽取；如何利用机器学习进行样本训练等等，这些也是本文在研究中需要考虑的问题。

1.4 本文主要工作

本文提出将关键词抽取分为两个问题进行处理：关键单词抽取和关键词串抽取，设计了一种基于分离模型的关键词抽取方法。该算法针对关键单词抽取和关键词串抽取这两个问题设计了不同的特征以提高抽取的准确性。本文的工作主要分为以下几个方面：

一. 分离模型的设计

根据词与新词、短语的不同特点，将词一级的关键词抽取与短语一级的关键词抽取看成两个不同的问题。利用它们各自的特点，采取基于机器学习的方法将两类关键词分开训练得到训练模型，并以此对相对应的候选项进行是否为关键词的判定。此方法有利于对词一级的关键词与短语一级的关键词开发不同的特征，提高两者抽取的准确率。

二. 新词、短语识别

新词与短语是关键词抽取过程中的难点，本文利用互信息与词边界参数表作为短语一级的关键词抽取的特征，提高此类关键词抽取的效果。

三. 关键词特征的开发

特征的开发是关键词抽取中的重点与难点。我们基于分离模型，针对关键词和关键短语开发相应的特征，提高关键词抽取的效果，是本文的主要工作之一。

四. 实验验证

好的思想需要好的实验方法验证，如何利用实验数据设计一套严谨的实验过程进行实验分析，并从实验结果中发现问题，对于任何科学研究都是不可或缺的。本文将此作为重点，对所做的关键词抽取相关工作，进行了详细的实验验证与分析。

1.5 本文结构

本文的结构如下：

第一章为绪论，概要阐述了关键词的背景、应用及面临的主要问题，同时介绍了本文的主要工作及文章结构。

第二章对相关方法进行研究和分析，主要介绍了关键词抽取与其它类似任务的差异，从“关键”与“词”两个方面叙述了关键词抽取的研究现状，并进行了小结。

第三章主要介绍关键词的分类问题，详细定义了关键单词与关键词串。

第四章主要介绍基于分离模型的关键词抽取算法。分别从分离模型的构造、候选关键单词与候选关键词串的生成、模型的训练与学习器的选择、以及关键词抽取四个方面进行了详细介绍。

第五章主要介绍关键词抽取中特征选择的问题。从关键单词、关键词短语、以及两者的公共特征三个方面论述这些特征对于关键词抽取的意义。

第六章进行了实验分析，验证了分离模型以及特征选取对于关键词抽取的意义，并对实验结果进行了详细分析，最后与代表当前研究水平的关键词抽取工具 KEA 进行了对比实验。

第七章，对全文进行了总结，并对将来进一步的研究工作提出了设想。

第二章 相关方法介绍

本章主要讨论关键词抽取的相关工作。首先对将关键词与相关类似任务进行了比较，然后具体介绍关键词抽取的研究现状。

2.1 关键词抽取与相关任务比较

2.1.1 关键词抽取与自动摘要

许多相关工作发现，从文档中抽取关键句子形成摘要与从文档中抽取关键词的任务类似，如 Luhn[3]、Edmundson[4]、Marsh[5]、Paice[6]、Paice 和 Jones[7]、Johnson[8]、Salton[9]、Kupiec[10]、Brandow[11]、Jang 和 Myaeng[12]等人的工作，但自动摘要比关键词抽取更难。原因在于自动摘不但要抽取或生成摘要句，还对构成摘要的句子的排序有要求，而实际中大多数自动摘要算法都不能很好地解决两个连续抽取的句子之间结合不紧密的问题，难以生成逻辑连贯的摘要。虽然关键词抽取也强调对关键词的准确排序，但是并不需要考虑各关键词之间的联系。另外自动摘还有一个难点就是代词的指代问题，而关键词抽取往往不会抽取文档中的代词。

曾经有一段时间自动摘要主要基于手工制定的启发式规则进行关键句的抽取。这些启发式规则特别适合专业的领域，但如果将这些启发式规则推广到新的领域效果往往不佳。将启发式规则推广到新的领域需要大量的人工参与。因此后来研究者采用了机器学习的方法，通过一批已经人工生成摘要的训练语料进行训练，解决基于启发式自动摘要所遇到的问题[10][12]。基于机器学习的自动摘要方法可以有效解决领域扩展问题，但是机器学习方法仍然需要人工参与，因为训练语料摘要中的每个句子必须完整的出现在文档中，采取的是一种机械文摘的方法。这就意味着现实生活中文档作者撰写的标准摘要可能不适合学习训练。而关键词抽取的一个有利条件就在于作者提供的标准关键词绝大部分出现在相关文档中，可以成为研究和实验的标准。

2.1.2 关键词抽取与信息抽取

另外一个与关键词抽取相关的工作是信息抽取。一个信息抽取系统通常根据预先定义好的指导原则寻找文档中特定的信息。这些指导原则往往针对特定的领域或话题。例如一个关于恐怖袭击领域的话题，信息抽取系统往往详细列出这些指导原则：

（1） 哪个恐怖组织参与了袭击？

（2） 谁是恐怖袭击的受害方？

（3） 恐怖袭击的方式（自杀性爆炸、枪击等等）是什么？

还有许多其它信息可以通过指导原则自动抽取。绝大部分信息抽取系统的建立都需要大量领域专家的参与，耗时耗力。

信息抽取与关键词抽取像一个事物的两端，信息抽取是对专业、特定领域的信息浓缩；而关键词抽取是对普遍的、非特定领域的信息浓缩。两者虽然实现算法上可能有各自的特点，但本质上都是对信息的一种抽象。

2.1.3 关键词抽取与自动索引

与关键词抽取类似的任务还包括索引创建，如 Fagan[13]、Salton[14]、Ginsberg[15]；Nakagawa[16]、Leung and Kan[17]等人的工作。Leung 和 Kan 对相关工作进行了详细的研究并进行了总结。Leung 和 Kan 认为索引主要分为两类：

（1） 为了方便读者浏览的索引，即在一些书籍后面方便查找的索引。

（2） 为了信息检索的索引，即在搜索引擎中建立的索引。

搜索引擎中的索引不适合读者浏览，因为基本上每个文档中的每个词都建立了索引（当然不包括停用词，例如英语中的“the”和“of”等）。方便读者浏览的索引规模较小，因为它仅包含文档中出现过且重要的词与短语。为了信息检索的索引一般只含词，不包括含多个词的短语。因为对于搜索引擎来说，多个词的短语索引的建立对于检索的效果影响甚微[15] [18]，没有必要去发费额外的精力去建立这些索引。

关键词抽取的目的就是方便读者浏览文档，因此方便读者浏览的索引比信息检索的索引更与关键词相关，但是它们又是不同的。一个明显的区别是数量的差别。根据 Nakagawa[16] 的研究发现，通常一篇文档的关键词数量通常为 10^0~10^1 个，而方便读者浏览的索引项数目通常为 10^2~10^3 个。一个关键词抽取算法可以用来抽取方便读者浏览的索引项，反过来利用方便读者浏览的索引项也可以有效的提高关键词的抽取效果。

2.2 关键词抽取研究现状

2.2.1 关键词抽取中“关键”问题研究现状

Krulwich 和 Burkey[19]利用启发式规则抽取文档中重要的词和短语。这些启发式规则主要依据格式和简单结构特点抽取关键词。例如利用候选的关键词是否为斜体字，候选的关键词是否出现在每节的标题中，或者它是否为取首字母的缩写词等等。基于启发式规则的关键词抽取动机是利用这些规则作为特征自动抽取文档关键词。虽然此方法可以自动生成大量的关键词，但它的准确率很低，因此整体效果也不好。

Muñoz[20]利用无监督学习算法判定只含两个词的短语是否为关键词。所谓监督学习就是给定一系列训练样本，其中每个样本都做上了标记，比如说标记出这个样本属于 A 类。

学习的目的是从这些带有标记的样本中学习到一些概念，比如说，什么样的数据对应 A 类而不是 B 类，并且在未来给出新的样本时，能够正确预测新样本的标记。大多数基于统计的学习方法都依赖于大规模训练样本，而自然语言处理中基于此方法需要大规模标注语料。这些语料需要专业标注人员手工标注，通常费时费力价格昂贵，而且目前的语料资源也比较稀缺。那么无监督学习就是给定一系列没有任何标记的训练样本，学习的目的是发现隐藏在这些样本中的某种结构。Muñoz 采取了基于 ART 神经网络的无监督学习算法。但无监督学习算法通常性能不高，虽然 Muñoz 的抽取方法产生了大量的两个词的短语，但准确率同样很低。而且此算法无法对一个词与两个词以上的短语进行判定。

Steier 和 Belew[21]利用互信息发现文档中含两个词的关键词。此方法与 Muñoz 的方法具有同样的局限性，虽然可以自动生成两个词的关键词集合但准确率很低，关键词的构成形式也很单一。不过它们的研究中发现，同样两个词的短语，专业领域计算出的互信息值往往比通用领域高。

Turney[22]与 Witten[23]分别开发了系统 GenEx 与 KEA，这两个系统在关键词抽取的发展史上具有重要的意义。它们首次利用监督学习的方法训练已标注关键词的语料，然后通过训练出的关键词抽取模型对未标注关键词的文档进行关键词抽取，此方法在准确率与召回率上都超越了前人的工作。Turney 利用遗传算法和 C4.5 决策树学习方法设计了系统 GenEx。而 Witten 采用朴素贝叶斯技术对短语离散的特征值进行训练，获取模型的权值，以完成下一步从文档中抽取关键短语的任务。Witten 的工作中发现模型具有领域泛化性，即训练语料与测试语料来自不同领域的关键词抽取效果不会比训练语料与测试语料来自相同领域的关键词抽取效果显著下降。另外 Witten 还开发了关键词特征，即如果某个词或短语经常作为关键词，那么在新的文档中，它也很有可能是关键词。Witten 将此特征进行了实验，结果表明添加此特征比不添加此特征，关键词抽取效果显著提高。但此特征存在明显两个缺点：

（1） 此特征受领域限制较大，无法有效提高训练语料与测试语料来自不同领域的关键词抽取效果。

（2） 添加此特征后，关键词抽取模型需要更多的训练语料才能达到稳定的效果。

Turney 与 Witten 的工作都把文本中连续出现的几个词看成候选关键短语，并未考虑这些词序列是否符合人们习惯性认为的短语形式。

从国内看，由于汉语语言本身的特点，没有显式的词边界，为关键词串自动标引任务又增加了一定的难度，使用最多的一种方法是基于 PAT Tree 结构获取新词，并采用互信息等一些统计方法对文档的关键词进行标引[24]。但获取候选词选用的 PAT Tree，它的建立用计算机实现需要大量的空间消耗，因此还需要进一步深入研究。

李素建等[25]提出的利用最大熵模型进行关键词自动标引的方法，最大熵模型是一个比

较成熟的数学模型，适合于估计事件的概率分布。它是一种概率估计方法，具有较强的知识表达能力，可以综合观察到的各种相关或不相关的概率知识，广泛应用在词性标注、命名实体识别、关键词抽取等自然语言处理领域。

基于最大熵模型的关键词抽取基本原理是：首先对语料进行关键词标引，选取训练数据时，将每一个连续出现的字串作为一个事件。假设有一个样本集合为{(ck_1，Y_1)，(ck_2，Y_2)，…，(ck_N，Y_N)}，每一个 ck_i (1≤i≤N)表示一个进入最大熵模型进行概率估计的候选关键词，Y_i (1≤i≤N)表示该候选项被标引的结果，该结果属于集合{YES，NO}，YES 表示是关键词，NO 表示不是关键词．利用最大熵模型得出在特征限制下最优的概率分布。

最大熵模型中，特征集合的选取与特征参数的估计是一个非常重要的问题，而李素建等的特征选择和特征参数估计时不够准确，造成关键词自动标引应用时不够理想。

王军[26]提出一种用于自动标引文献主题关键词的方法。词表和分类法是传统纸质文献环境下最重要的知识组织工具。它的更新和维护一直依靠手工进行。这限制了它在数字图书馆和网络信息环境下的应用。王军介绍了一种基于统计的、从元数据的标题中抽取关键词并定位在词表中的方法。定位的依据是抽取出的关键词所对应的标引词集的收效性质。但此方法限定从已标引的结构化语料库元数据的标题中抽取关键词。

最近几年有一些研究者在利用词与词之间的关系方面取得了一些进展。研究发现，同许多自然网络、社会网络一样，自然语言也是一个网络。网络是一个由元素及元素间关系组成的系统，系统中的元素为“节点”，各个元素间的联系为“连接”。自然语言在语言、词法、语法、语义、语用的各个层次上都体现出复杂的网络结构。因此如果把文本的词看成是“节点”，词汇间的联系看成是“连接”的话，文本就可以形成一张网。词汇间的互动是有据可循的，它们相互依赖相互联系形成了独特的语法结构，向人们传递了不同的语义信息。按照语言层次的不同，语言网络可分为[27][28]：

（1） 共现网络：文本是由词汇线性组成，很容易联想到的就是共现作为词汇间的联系。限定一个窗口（邻接、句子、段落等），两词汇若出现在同一窗口，则认为它们有关系存在 “连接”，“连接”上的权重表示为共现程度。

（2） 语法网络：文本是由词汇按一定的语法规则组合起来的，依存文法是一种利用词汇信息的语法体系，它认为句子中不同成分（词）之间存在支配与被支配的关系。按照这种关系文本可以构造一张有向图，节点之间的关系就是它们的依存关系。

（3） 语义网络：文本中词汇选择根本的原因在于词汇本身的含义能够表达期望的内容。因此，词汇之间的联系应该存在于语义层。通过词汇间语义联系能够构建文本的语义网。

目前大部分的研究集中在词汇的共现网络上，它能部分反映文本的语法语义关系。但

大部分共现网络的中心结点都是一些语法功能强却没有实际意义的词（如冠词、助词、介词等）。如果在共现网络中去掉这些中心节点，仍然存在两个问题：

（1） 如何选择窗口大小。

（2） 同一窗口下的词汇是不是都有关系。

在依存关系建立起来的语法网络中，由于很多词依赖句子中的主动词，造成语法网络中的中心结点是一些动词，这往往不是关键词。因此许多研究者开始对语义网络进行深入的研究。索红光等[29]提出了利用《知网》知识库构建词汇链的方法。词汇链是由一系列相关词汇组成，最初被用来分析文本的结构。王军首先通过计算词义相似度构建词汇链，然后结合词频与区域特征进行关键词选择。但这种方法只适用于收录在《知网》中的关键词。

石晶[30]基于小世界模型进行文本分割，确定片段主题，进而总结全文的中心主题，使文本的主题脉络呈现出来。为此它首先证明由文本形成的词汇共现图呈现短路径，高聚集度的特性，说明小世界结构存在于文本中；然后依据小世界结构将词汇共现图划分为“簇”，通过计算“簇”在文本中所占的密度比重识别片段边界，使“簇”与片段对应起来；最后利用短路径，高聚集度的特性抽取图“簇”的主题词，采取背景词汇聚类及主题词联想的方式将主题词扩充到待分析文本之外，尝试挖掘隐藏于字词表面之下的文本内涵。

上述研究都是通过加入词与词的关系作为特征，提高了关键词抽取的效果。大多都是普通语言网络的简单应用，未来如何利用语义网络抽取关键词将是本领域研究的重点。

以上工作本质上都是解决关键词抽取中如何度量“关键”特性的问题。至于关键词抽取中“词”问题的解决，国内外的研究也取得了一定的进展，特别是如何从文本中抽取新词、短语等。

2.2.2 关键词抽取中“词”问题研究现状

一种比较常见的研究方法是通过统计 N-gram 词性匹配模式的方法来抽取关键词短语，另外一个相关的研究领域是组块(Chunk)的自动识别，但 Anette Helth[31]指出通过组块自动识别的方法难以获得符合人们习惯的关键词短语，为此她人工总结了 56 个词性匹配模式，用于英文短语一级关键词的自动抽取。

刘远超[32]利用粗集理论在数据泛化和知识约简方面的优势，对人工标注的人民日报关键词短语语料进行了挖掘，从而得到了中文关键词短语的若干构成规则。这些规则可以用于自动关键词抽取，也可以对手工关键词标引进行指导。刘远超与 Anette Helth 工作的相同之处在于挖掘的规则都是基于词性的，区别在于刘远超在挖掘出的规则构成上还包括短语左右相邻词的词性，另外由于规则是从足够规模的真实语料中自动获得的，从而使获取的规则更加客观，完备性也较好。

中文关键词抽取在“词”问题的研究上还存在一个难点，就是新词问题如何解决。目

前，对非命名实体的新词的识别方法，主要分为以规则为主的方法和以统计为主的方法两大类。基于规则的方法是通过标注词典以及组词规则来识别新词，但是该方法的困难是词性的歧义性和语法的灵活性，另外词典中不可能包含所有的中文词，也不能穷尽所有的组词规则。统计方法通过对词共现进行概率统计而实现的。这种方法适用于任何领域，但是它们需要大量的训练语料。文献[33]对新词识别领域的相关工作进行了总结，相关的方法如下：

（1） 词频

新词的一个特点是重复出现。一个新词在文档中通常出现不只一次，尤其在特定领域的新概念，例如：十七大，超级女声，快乐男生等。一个新词在给定文档中重复出现次数被称为词频。

（2） 成词率:IWP(C)

IWP 是一个实数特征，许多汉字可以独立成词或者构成多字词。单个字的 IWP 是指该汉字的成词率，IWP 定义如下:

$$IWP(c) = \frac{C(c,w)}{C(c)} \tag{2.1}$$

其中 C(c，w)是单字 c 作为非独立的词在训练数据中出现的次数，且 C(c)是 c 在训练数据中出现的总数。一个字串的 IWP 值是由组成它的字的 IWP 值的乘积。IWP 值越高，该字串越容易形成新词。

（3） 位置成词率:IWP(c，pos)

一些汉字比其它汉字更倾向于出现在一个词特定位置。例如“性”通常出现在一个词的结尾。而“老”常出现在一个词的开头。因此考虑一个词中的不同位置的 IWP 值，就得到扩展 IWP(c，pos)，pos 是一个汉字在一个词中的位置，有三个可能值:1、2 和 0，分别表示开头、中间和结尾。

（4） 词型模拟:FANA

给定字对(x，y)，如果发生下列情况之一，就称(x，y)具有共同的词干 z:

a) 字串 xz 和 yz 字串是词典词(即，x 和 y 是前缀)。

b) 字串 zx 和 zy 字串是词典词(即，x 和 y 是后缀)。

搜集那些共同词干的数量大于预先设定阈值的词缀对，构成词缀对列表。对于给定的候选 ab， FANA 如下：如果存在一个(a，x)是词缀对，且字串 xb 是词典词，或者(x，b)是词缀对且字串 ax 是词典词，那么 FANA (ab)=1，即为新词，反之不然。例如，给定候选“下岗”，因为(上，下)是一个词缀对（有共同词干，例如:_任，_游，_班等），而且上岗是词典词，所以 FANA (下岗)=1 为新词。

中文不同于英文的一个显著特点就是没有显示的词边界，因此中文分词是所有中文信

息处理工作的基础，也是汉语信息处理的瓶颈。一般来讲，计算机自动分词需要有一部后台词典的支持，但词典中所收录的词是有限的，这时候对未登录词的识别就尤为重要，特别是对新出现词汇的识别更为重要。本文将在第三章详细分析新词与短语的共同点与不同点，并对如何解决问题提出自己的想法。

2.3 小结

虽然国内外有许多关于关键词抽取的相关研究，但以我们掌握的资料，还没有看到有相关工作把词一级的关键词抽取与短语一级的关键词抽取看成两个问题分别解决。

国内外研究关键词抽取的方法很多，但在如何选择特征表示“关键”上本质上不外乎以下三种：

（1） 词频是一种最简单最直观的特征。大部分关键词都是在文本中出现多次的词序列。但是文本中词频最高的序列往往不是关键词，而是一些停用词或由停用词组成。如何利用好词频信息进行关键词抽取是一个重要的问题。另外，通过观察发现本质上词序列的 TFIDF 其实也是一种词频的变相度量。

（2） 词序列出现的位置是关键词抽取可以利用的一个重要特征。大部分关键词出现在标题、摘要、文章第一段、最后一段等的概率大于出现在文本其它地方的概率。

（3） 词性特征，大部分关键词都是名词或名词短语。

虽然研究关键词抽取的学者提出了许多“关键”特征，但本质上都是此三种特征。但以上特征都只考虑词本身的特点。未考虑词与词之间所蕴含的潜在信息对于关键词抽取的重要性。

而在如何确定“词”上同样存在共同的规律：

（1） 相邻词之间的紧密程度。无论是新词还是短语。它们都是一些相邻词的组合，这些相邻词结合的紧密程度大部分高于其它非新词与非短语相邻词之间的紧密程度。

（2） 一些词经常作为新词与短语的首词与尾词。

（3） 短语一级的关键词有其自身的特点与规律，如词性组合特征。如何挖掘与应用这些规律是研究者十分关心的问题。

通过对国内外关键词抽取技术调研发现，关键词抽取还有许多问题没有解决。比如，如何进一步提高文本中短语一级的关键词抽取技术；关键词抽取技术中，抽取新词与短语的共同点与不同点；关键词抽取中特征选择问题；关键词抽取评价标准等问题，这些都是本文需要重点考虑的问题。

第三章 关键词分类问题

关于“关键词”的概念由来已久，命名的方式也多种多样。英文单词“keyword”和“keyphrase”都表示关键词这个概念。这两个单词准确的翻译成中文应该是“关键词”和“关键短语”。实际研究中对于中文“关键词”这个概念也有许多命名法，包括“关键词”、“关键短语”、“关键词短语”、“关键字”等等。同一概念如此多的名称也从一个侧面反映人们对于“关键词”这个概念上应该是一些词还是短语存在许多分歧。鉴于此本文将“关键词”分为两大类，即“关键单词”与“关键词串”。

3.1 关键单词的定义

如果从“关键词”的字面意义上讲，严格意义上的关键词仅含一个词，对应的英文单词为“keyword”。但绝大多数的科学杂志与期刊中的“关键词”大部分是一些短语，这就与“关键词”的字面意思相冲突。因此为了统一与规范，本文将词一级的关键词称为关键单词，也就是说每个关键单词仅包含一个词；而将有多个词构成的“关键词”称为“关键词串”（见 3.2 节）。

如果仅为了从字面意思上区分关键词似乎意义不大，但实际上将关键单词从关键词中剥离出来对于关键词抽取是有意义的。因为词与短语不同，如果把短语比作一条链的话，那么词就是这条链的链结点。从物理结构很容易分析得出，链结点肯定比一条锁链更结实，整体性更好。从自然语言的角度也可以分析出词比短语在结构上更稳定。

英文中的词有天然的边界，很容易判定对象是否是词，而短语没有天然的边界，无法轻易判定一个对象是否是短语。对于关键词抽取这个问题来说，抽取词一级的关键词远比抽取短语一级的关键词来得容易，因为抽取短语一级的关键词首先要判断词串是否是人们通常可以理解的有意义串。当然，中文中的词没有天然的分隔符，大多数关键词抽取方法首先需要分词，但目前中文分词已经到了相当高的水平，远比短语识别准确率高。

因此无论是英文还是中文，单独解决关键单词抽取的问题，复杂性与难度不会比抽取短语一级的关键词高。

3.2 关键词串的定义

关键单词是包含一个词的关键词，那么包含多个词的关键词如何定义？因为在现实生活中，有相当一部分人将关键词称作“关键短语”，为了不与人们的习惯相矛盾，我们将含多个词的关键词命名为关键词短语，即关键词(有时统称为关键短语)包括关键单词与关键词短语。

但汉语文本中词无天然的分割符，而关键词抽取技术大都先依赖词典分词，结果造成一些未登录词被切分成多个词典中的词。本文把这些未登录词以及短语统称为词串。汉语中的关键词则可分为关键单词与关键词串。

汉语中的未登录词与短语有相同的特点，它们在分词时都被切分成由几个词典中的词组成的词序列。与其它词序列相比，词串在相邻词之间结合得更加紧凑。如同 3.1 节讲到的，词串类似于一种链式结构，如果将整条链比作词串的话，那么词就是这条链的链结点。词串的相邻词之间的关系就如同链的相邻结点之间的关系，它们结合得非常紧密。

但未登录词与短语又是不同的，短语有一定的语法结构，而未登录词本质上还是一个词。基于未登录词与短语的相同点与不同点，我们将在第五章特征设计部分详细介绍如何进行词串的识别。

3.3 小结

本节主要介绍了关键词的分类问题。将关键词分为关键单词与关键词串。其中关键单词定义为只包含一个词的关键词，而关键词串分为关键词短语以及一些关键的未登录词。当然关键词串的定义主要针对汉语的关键词抽取问题。因为汉语的关键词抽取首先需要分词，而分词依赖的词典收录的词有限，因此许多未登录词被切分成粒度较小的多个词典词。英文不存在分词问题，它的关键词串仅包含关键词短语，所以本文以后关于英文关键词抽取问题的相关论述中，将关键词仅分为关键单词与关键词短语。

第四章 基于分离模型的关键词抽取算法

我们把关键词抽取看成一个分类问题，即文本中每个候选关键词是属于关键词还是属于非关键词，问题的难点就在于如何正确的区分候选关键词是属于两类中的那一类。机器学习方法为解决这一类问题提供了较好的工具。在机器学习的专业领域中，文档中的候选关键词可以看成样本，学习的问题就在于如何得到一种映射机制，使得样本对应关键词类或非关键词类。机器学习方法可以自动的生成这种映射机制，只要我们提供足够的训练样本，而这些训练样本我们已经人工将其分类，即训练样本中的候选关键词已经标记是属于关键词类还是属于非关键词类。一旦通过学习，这种映射机制被训练得到，那么这种映射机制就能对未标记类别的样本进行类别判定，换句话说，它可以用来对新文档的候选关键词进行是否为关键词的判定。当然通常意义上这种映射机制被人们习惯上称为训练模型。

4.1 分离模型的构造

传统的关键词抽取研究中，关键单词样本与关键词串样本是不加区别的。通过同时对所有标注好的关键单词样本与关键词串样本进行训练形成一个整体模型。然后以此模型来判断其它未标注的候选关键单词与候选关键词串是否为关键词。

然而正如我们在第三章介绍的那样，词串类似一种链式结构，其本身具有一定的结构特点，不应简单地把词与词串等同，而应该把它们分开考虑。正是因为传统的基于机器学习的关键词抽取研究中，将已标注的词样本与词串样本一同训练，使得在训练过程中选取的训练特征必须是两者共有的特征，完全忽略了词与词串的差异性，限制了它们各自合适特征的开发与研究。

基于上述分析，本文将关键词抽取分为两个问题解决，即关键单词抽取与关键词串抽取，针对词和词串的不同特性设计相应的特征，并把关键单词样本集合与关键词串样本集合分别进行学习和训练，以获得不同的关键单词模型与关键词串模型。在应用这两个模型抽取未标记关键词文本中的关键单词和关键词串时，将根据两个不同的模型分别对未标记关键词文本中的候选关键单词与候选关键词串进行判断，如图 4.1 所示：

分离模型具体构造过程如下：

（1） 首先对已标注是否为关键词的候选关键词样本进行分类，根据关键单词与关键词串的定义，将候选关键词样本分为候选关键单词样本与候选关键词串样本。如果候选关键单词被标记为关键词，则此样本为候选关键单词样本正例，否则为候选关键单词样本反例；同样如果候选关键词串被标记为关键词，则此候选关键词串样本为候选关键词串样本正例，反之为候选关键词串样本反例。

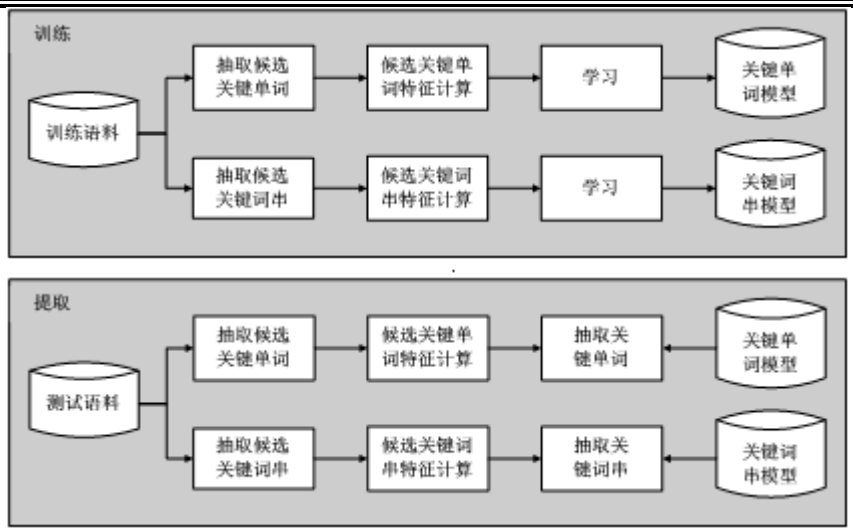


图 4.1　分离模型的训练与抽取过程

（2）　选取训练语料中所有的候选关键单词样本，计算预先定义的适合关键单词抽取的特征。并对计算得到的特征进行离散数值化处理，形成每个训练样本的特征向量。

（3）　选取训练语料中所有的候选关键词串样本，计算预先定义的适合关键词串抽取的特征。并对计算得到的特征进行离散数值化处理，形成每个训练样本的特征向量。

（4）　选取合适的学习器，分别对候选关键单词训练样例集合与候选关键词串训练样例集合进行学习训练，得到关键单词模型与关键词串模型。关键单词模型可以用来对未标记是否为关键词的候选关键单词进行是否为关键单词的判定，关键词串模型可以用来对未标记是否为关键词的候选关键词串进行是否为关键词串的判定。对于未标记关键词的文档，首先按照关键单词与关键词串的定义，将候选关键词分为候选关键单词与候选关键词串两类。

（5）　针对未标记是否为关键词的候选关键单词，根据预先定义的关键单词的特征，计算特征形成特征向量。

（6）　针对未标记是否为关键词的候选关键词串，根据预先定义的关键词串的特征，计算特征形成特征向量。

（7）　利用关键单词模型，对未标记是否为关键词的候选关键单词样本特征向量进行正反例判定。如果特征向量被判定为正例，则此候选关键单词为关键单词，否则为非关键单词。

（8）　利用关键词串模型，对未标记是否为关键词的候选关键词串样本特征向量进行正反例判定。如果特征向量被判定为正例，则此候选关键词串为关键词串，否则为非关键词串。

分离模型不但在词与词串特征选取相同的条件下关键词抽取效果比整体模型好(第六章实验结果证实了这一点)，而且还可以根据词与词串的不同特点设计不同的“关键”特征以提高关键词的整体抽取效果。

4.2 候选关键单词与候选关键词串的生成

候选关键单词与候选关键词串的生成并不是简单的按照关键单词的定义与关键词串的定义。它是一个重要且繁琐的过程，而且中文关键词抽取中的候选关键单词与候选关键词串的生成同英文关键词抽取中的候选关键单词与候选关键词短语的生成细节有明显的差别。

4.2.1 英文中候选关键单词与候选关键词短语的生成

英文关键词抽取中并不是任意的词或者词序列都适合作为候选关键单词或候选关键词短语。因为有相当一部分词或者词序列一般是不会成为关键词的，我们通过调查研究发现下面三条规则可以过滤大部分一般不会成为关键词的词或词序列：

（1） 词序列必须限定在一定的长度，过长的词序列一般不会成为关键词。

（2） 词序列中不容许出现其它非词符号。

（3） 大部分的关键词的首词与尾词一般不会是停用词。

基于以上三条规则，具体生成英文中候选关键单词与候选关键词短语如下：

（1） 对于每个文档首先根据一些固定边界进行划分，这些边界包括标点、括号、数字等。

（2） 选定词长度不超过 4 的所有词序列作为候选关键词。

（3） 过滤其中所有含非词符号的词序列和数字。

（4） 将所有首词或尾词是停用词的候选关键词删除。

（5） 对剩下的每个候选关键词进行 stemming 处理。

（6） 根据关键单词与关键词短语的定义，划分候选关键单词与候选关键词短语。

通过以上候选关键单词与候选关键词短语的生成过程可以过滤大量一般不会成为关键词的词与词序列，减少训练样本数量，提高机器学习模型训练速度。

所谓 stemming 处理，是指将英文单词转换成词干形式。Porter[34]和 Lovins[35]stemming 算法是两个经典算法。Porter 和 Lovins 的 stemming 算法都利用启发式规则删除或转换英文中单词的后缀。另一种 stemming 方法是利用词典处理每一个单词，此词典列出了每个单词出现在文档中的所有可能形式。

Lovins stemmer 与 Porter stemmer 略有不同，它更容易使两个词转换成一个形式，当然 Lovins stemmer 也更容易产生错误[36]。例如 Lovins stemmer 可以将“psychology”与

“psychologist” 转换成同一个“psycholog” 词干形式，而 Porter stemmer 就会将“psychology” 转换成“psychologi” 词干形式，将“psychologist” 转换成“psychologist” 词干形式。但是 Porter stemmer 可以正确的将“police”与“policy” 分别转换成“polic” 与“polici” 词干形式，而 Lovins stemmer 就会将两者转换成一样的词干形式。

我们通过研究发现对于英文的关键词抽取来说，利用 Lovins stemmer 比 Porter stemmer 对于关键词抽取效果更好。因此我们利用 Lovins stemmer 对候选关键词进行 stemming 处理。

4.2.2 中文中候选关键单词与候选关键词短语的生成

因为中文中的词没有天然的分隔符，而大部分中文的关键词都以词为基本元素，因此绝大多数的中文关键词抽取必须首先分词。与英文关键词抽取相同，并不是所有的词或者词序列都适合作为候选关键词，我们通过研究发现英文中适合过滤大量一般不会成为关键词的词或词序列的一般规则同样适合中文。例如英文中的关键词抽取技术在选择候选关键词时，把开头词或结尾词是停用词的候选关键词过滤[23]。我们以同样的方法对中文中候选关键词的选择问题进行了实验，实验结果表明此方法在过滤掉 45%左右的非关键词的情况下，关键词的丢失率不到 1.5%。因此在中文中我们同样采用此方法选择候选关键单词与候选关键词串。具体生成中文中候选关键单词与候选关键词串如下：

（1） 首先对中文文本进行分词处理，使得中文文本分词后形成类似与英文中词与词用空格分开的结构形式。

（2） 对于每个分词后的文档根据一些固定边界进行划分，这些边界包括标点、括号、数字等。

（3） 选定词长度不超过 4 的所有词序列作为候选关键词。

（4） 过滤其中所有含非词符号的词序列和数字。

（5） 将所有首词或尾词是停用词的候选关键词删除。

（6） 根据关键单词与关键词串的定义，划分候选关键单词与候选关键词串。

4.3 模型的训练与 SVM 学习器

利用机器学习的方法解决关键词抽取的问题，需要获得训练模型。此模型可以对未标记是否为关键词的候选关键词进行关键词的判定。为了得到训练模型，我们首先需要选取一批已手工标注关键词的文档作为训练集。按照候选关键单词与候选关键词串的生成过程，对每篇文档生成候选关键单词与候选关键词串，接着将所有文档的候选关键单词作为候选关键单词集合，将所有的候选关键词串作为候选关键词串集合。

对每一个候选关键单词计算设计的特征，并对计算得到的特征进行离散数值化处理形成特征向量。如果候选关键单词在训练集中被标记为关键词，则此候选关键单词将被标记

为候选关键单词集合中的正例，如果在训练集中被标记为非关键词，则此候选关键单词将被标记为候选关键单词集合中的反例。同样对于每一个候选关键词串特征向量的形成也是如此，按照设计的特征计算每一个候选关键词串，离散数值化处理计算得到的特征值形成特征向量。如果候选关键词串在训练集中被标记为关键词，则此候选关键词串将被标记为候选关键词串集合中的正例，如果在训练集中被标记为非关键词，则此候选关键词串将被标记为候选关键词串集合中的反例。利用分类模型的思想，选取所有的候选关键单词样本作为关键单词模型训练样本集合，选取所有的候选关键词串样本作为关键词串模型训练样本集合。

利用机器学习的方法研究关键词抽取问题，必须选择一个合适的学习器。学习器的选择上，Zhang[37]利用 SVM 对关键词抽取取得了很好的效果，我们同样选择 SVM 作为学习器。

支持向量机（SupportVectorMachine）是贝尔实验室研究人员 V.Vapnik[38][39][40]等人在对统计学习理论三十多年的研究基础之上发展起来的一种全新的机器学习算法，也使统计学习理论第一次对实际应用产生重大影响。SVM 是基于统计学习理论的结构风险最小化原则，它将最大分界面分类器思想和基于核的方法结合在一起，表现出了很好的泛化能力。由于 SVM 方法有统计学习理论作为其坚实的数学基础，并且可以很好地克服维数灾难和过拟合等传统算法所不可规避的问题，所以受到了越来越多的研究人员的关注。

SVM 是从线性可分情况下的最优分类面发展而来的，基本思想可用图 4.2 的两维情况说明。图中，方形点和圆形点代表两类样本，H 为分类线，H1，H2 分别为过各类中离分类线最近的样本且平行于分类线的直线，它们之间的距离叫做分类间隔。

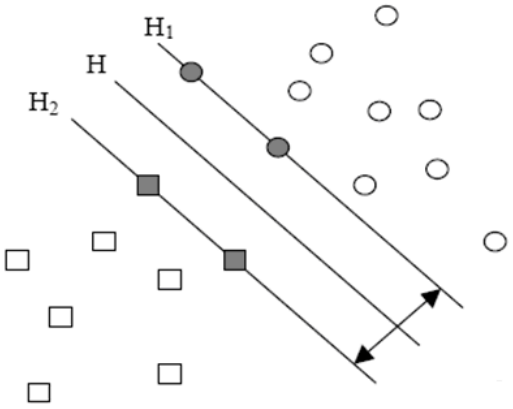


图 4.2　结构风险最小化思想[41]

所谓最优分类线就是要求分类线不但能将两类正确分开（训练错误率为 0），而且使分类间隔最大。推广到高维空间，最优分类线就变为最优分类面。其中最大间隔思想是支持向量机中的一个重要思想，最大分类间隔超平面保证以最大的间隔将两类样例正确分开，这样可以保证学习机在两个方面具有很高的性能，一是最大间隔分类超平面可以容纳

的数据扰动是最大的，二是最大的间隔保证了最优分类超平面最低的 VC 维度从而降低结构风险的界，保证了最大的泛化能力，有效避免“过拟合”问题的产生。

基于 SVM 如此多的优点，我们利用 LIBSVM[42]对关键单词训练样本集合与关键词串集合分别进行训练，以获得关键单词模型和关键词串模型。

LIBSVM 是台湾大学林智仁(Lin Chih-Jen)副教授等开发设计的一个简单、易于使用和快速有效的 SVM 模式识别与回归的软件包，它不但提供了编译好的可在 Windows 系列系统的执行文件，还提供了源代码，方便改进、修改以及在其它操作系统上应用；该软件还有一个特点，就是对 SVM 所涉及的参数调节相对比较少，提供了很多的默认参数，利用这些默认参数就可以解决很多问题。.

通过 LIBSVM 训练得到的模型，无论是关键单词模型还是关键词串模型，都只能分别对新的候选关键单词或候选关键词串进行是否为关键单词或关键词串的二值判断，并不能对候选关键单词或候选关键词串进行“关键”的量化度量。因此，我们修改了 LIBSVM 的部分代码，使得 LIBSVM 训练出的模型可以对新样本成为正例的可能性评分。评分的标准是计算样本到“超平面”的距离。如果在正样本方向，样本离“超平面”的距离越远，则样本的评分越高；如果在负样本方向，样本离“超平面”的距离越远，则样本的评分越低。样本离对于关键单词模型来说，如果新的候选关键单词被关键单词模型评出的分值越高，则候选关键单词越可能是关键单词；分值越低，则越可能是非关键单词。同样对于关键词串模型来说，如果新的候选关键词串被关键词串模型评出的分值越高，则候选关键词串越可能是关键词串；反之，分值越低，则越可能是非关键词串。

通过此方法，我们可以利用改造后的关键单词模型与关键词串模型分别得到关键单词评分器与关键词串评分器，并以此分别对新的候选关键单词与候选关键词串进行评分。当然也可以通过相同的模型改造将整体模型变为整体评分器，此评分器不但可以对新的候选关键单词评分，也可以对新的候选关键词串评分。

4.4 关键词的抽取

基于分离模型的思想，对候选关键单词训练样本集合利用 SVM 进行训练，得到关键单词模型；同样对候选关键词串训练样本集合利用 SVM 进行训练，得到关键词串模型。

对于每篇未标注关键词的新文档，首先自动获得候选关键单词集合与候选关键词串集合。然后计算每一个候选关键单词的特征形成特征向量，利用关键单词评分器对其评分，然后按分值由高到低排序，分值越高，则越可能是关键单词，否则，越可能是非关键单词；计算每一个候选关键词串的特征形成特征向量，利用关键词串评分器对其评分，同样按分值由高到低排序，分值越高，则越可能是关键词串，反之则越可能是非关键词串。

但在关键词抽取的实际应用中，人们往往需要的是能够表述文档内容的关键概念，并

不关心抽取的关键词是关键单词还是关键词串。因此为了找到最合适关键词，需要将候选关键单词和候选关键词串按关键单词模型与关键词串模型对应评出的分数进行统一排序，然后重新按分值的高低选出最可能成为关键词的候选关键单词或候选关键词串。

为此，我们把关键单词评分器与关键词串评分器合成一体，形成了一个综合评分器。综合评分器利用关键单词评分器与关键词串评分器分别对候选关键单词与候选关键词串评分，然后将候选关键单词与候选关键词串按评分的分值由高到低进行整体排序。如果实际应用中需要前 n 个关键词，则利用分离模型关键词抽取算法构造的综合评分器将会把分值最高的前 n 个候选关键单词或候选关键词串输出，以此作为算法输出的关键词结果。综合评分基本原理如图 4.3 所示：

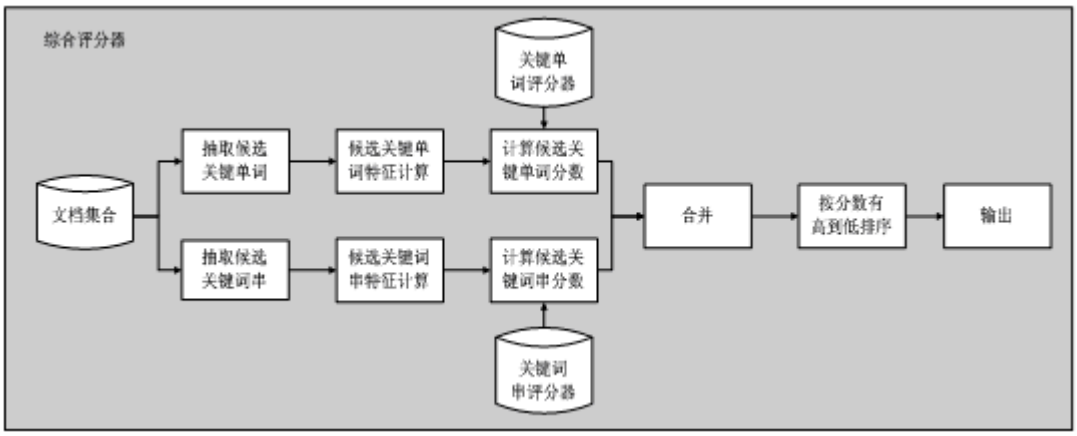


图 4.3　综合评分器基本原理

具体步骤如下：

（1）　首先抽取每篇文档的候选关键单词与候选关键词串。

（2）　按设计的特征，计算每个候选关键单词的特征值形成特征向量；计算每个候选关键词串的特征形成特征向量。

（3）　利用关键单词评分器，输入候选关键单词的特征向量，输出关键单词在文档中的“关键”分数；利用关键词串评分器，输入候选关键词串的特征向量，输出关键词串在文档中的“关键”分数。

（4）　将文档中所有的候选关键单词与候选关键词串合并，并保留“关键”分数。

（5）　将所有的候选关键项按“关键”分数由高到低排序。

（6）　根据用户需要，输出“关键”分数最高前 n 个候选关键项，以此作为文档的关键词。

当然，为了使方法到达最佳的效果，我们在算法中引进了一条规则：如果某个候选关键词是另一个候选关键词的子字符串，且评分器评出的分值低于另一个候选关键词的分值，那么它将不予考虑成为输出的关键词。

4.5 小结

本章主要介绍了基于分离模型的关键词抽取算法。首先从整体上详细介绍了分离模型的构造过程。接着从中英文两个方面，叙述了如何生成候选关键单词与候选关键词串。利用分离模型的基本思想，将候选关键单词训练样本集合与候选关键词串训练样本集合通过机器学习的方式进行模型训练，以得到关键单词模型与关键词串模型。通过对 SVM 的相关介绍与基本原理、特点的分析，阐明了选择 SVM 作为学习器的原因。利用对 LIBSVM 的源代码的修改，将 SVM 训练得到的模型改造成为评分器，能够对候选关键项按分值进行排序，从而选出最佳结果。最后说明了关键词抽取的最终任务是关键词的整体抽取，因此利用关键单词评分器与关键词串评分器构造了综合评分器。

本章主要从理论上阐明了基于分离模型的关键词抽取算法的意义，并介绍了基于分离模型的关键词抽取方法。第五章将给出分离模型的特征设计，第六章对实验验证进行了详细描述。

第五章 分离模型的特征设计

基于机器学习的方法都需要开发一些属性或特征将样本抽象成特征向量，关键词抽取任务同样也需要设计一些特征，以便特征化候选关键词样本，通过训练模型判定是否为关键词。

而传统的关键词抽取方法大多把重点放在解决“关键”的问题上，都是针对性很强的设计一些“关键”特征，以便度量候选关键词的重要性。但是据我们所知，所有的关键词抽取相关方法在设计特征时都有一个特点，它们不将关键词抽取与关键词串抽取看成不同的问题，而是将两者统一，设计一些两者都适合的公共特征。这就造成了基于机器学习方法的关键词抽取设计“关键”特征的局限性，使得许多适合不同问题的特征无法添加到公共特征中。

正如前文的分析，本文认为对于关键词抽取来说，关键单词与关键词串是两个不同的概念。它们在“关键”的度量上其实应该有各自独特的特点，而不是简单的将两者看成一类问题。因此我们利用分离模型抽取算法分别将关键词抽取分为两个问题进行处理：关键单词抽取和关键词串抽取。并且我们以该算法为基础，针对关键单词抽取和关键词串抽取这两个问题分别设计了不同的“关键”特征。

5.1 关键单词与关键词串公共特征设计

虽然我们将关键词抽取看成关键单词抽取与关键词串抽取两个问题，但是传统关键词抽取方法中许多“关键”特征仍然同时适合关键单词的抽取与关键词串的抽取。我们通过对相关工作的研究发现有四个公共特征，既适合两者的关键词抽取，又能在“关键”上很好的度量候选关键词。这些特征包括：TF×IDF、首次出现位置特征 POS、TF×IF、文档长度特征 NWT。其中 TF×IDF、首次出现位置特征 POS 与文档长度特征 NWT 在 Witten[23] 相关工作中取得了好的效果，而 TF×IF 是我们根据前面特征的不足，新引入的特征。

5.1.1 TF×IDF 特征

TF×IDF 特征是对一个词或词序列在一篇文档中的出现次数与这个词或词序列在平时出现次数的一个比较特征。当然这个词或词序列平时出现的次数可以通过度量一定规模的语料出现此词或词序列代替。TF×IDF 的具体计算公式如下：

$$TF*IDF = freq(P,D)\times(-\log_2 \frac{count(P)}{N}) \tag{5.1}$$

freq(P,D)表示词或词序列 P 在文档 D 中的出现的次数，count(P)表示在一定规模的语

料中有多少篇文档出现了词或词序列 P；N 表示一定规模的语料总共有多少篇文档。在第六章的相关实验中我们将用训练集表示一定规模的语料。当然也可以用第三方语料计算每个词或词序列的 TF×IDF 特征值。

公式中的 freq(P,D)表示 TF×IDF 中 TF，这反映出如果一个词或词序列在某一篇文档中出现多次，那么此词或词序列就比其它出现较少的词与词序列更有可能成为关键单词或关键词串。但是在实际的情况中我们发现，一篇文档中经常出现的词或序列往往是一些停用词或一些出现停用词的词序列，而 TF×IDF 中的 IDF 表示词或词序列在一定规模的语料中的文档倒转频率。如果一个词或词序列在平时出现的次数比较频繁，则这个词或词序列成为关键单词或关键词串的可能性就越小；反之，如果一个词或词序列在平时出现的次数比较少，而在某篇文档中出现次数比较多，则此词或词序列成为关键单词或关键词串的可能性就越大。因此我们通过 TF×IDF 的后半部分公式来度量每个词与词序列的文档倒转频率。

通过 TF×IDF 特征的计算公式，我们可以度量每个候选关键单词或候选关键词串的 TF×IDF 特征值，如果候选关键单词或候选关键词串的 TF×IDF 特征值越高，则此候选关键单词或候选关键词串越可能是关键单词或关键词串；反之，如果候选关键单词或候选关键词串的 TF×IDF 特征值越低，则此候选关键单词或候选关键词串越可能是非关键单词或非关键词串。

5.1.2 首次出现位置特征 POS

首次出现的位置特征 POS 是通过比较一个词或词序列在一篇文档中第一次出现时前面有多少个词与其文档所含词数确定的，首次出现位置特征具体计算公式如下：

$$Distance(P;D) = FirstApp(P) / NumWords(D) \tag{5.2}$$

FirstApp(P)表示词或词序列在某篇文档第一次出现时前面已经出现的词数目，而 NumWords(D)表示文档总共包含词的数目，这里主要是为了归一化处理。对于一个词或词序列来说，如果它出现在一篇文档中的标题、摘要、文档的目录、文档的介绍中或者出现在文档的结论部分、最后章节、参考文献中，则此词或词序列成为关键单词或关键词串的可能性比出现在文档其它部分的词或词序列大。因为通常一些专门的关键词标注人员在对一篇未标注关键词的文档进行关键词标注时，往往只注意文档这些部分的词或词序列，它们通常不会完整的阅读整篇文档[43]。而且作者在撰写文档时，其关键的信息基本上也都会出现在这些结构部分。因此，对于关键词抽取来说，更好的方法是对词或词序列出现的具体位置进行详细的分析。比如出现在标题中的词或词序列就比出现在摘要中的词或词序列更有可能成为关键单词或关键词串。但不是每篇文档都有规范的文档结构，而且详细分析每个结构部分的词与词序列会使得关键词抽取算法非常复杂，而将首次出现的位置特征的

作为关键词抽取的特征，不但方法简单，而且实际证明也能取得很好的效果[23]。

通过首次出现的位置特征的计算公式，我们可以度量每个候选关键单词或候选关键词串的首次出现的位置特征值。这个特征值得范围在 0 到 1 之间，如果候选关键单词或候选关键词串的首次出现的位置特征值越接近 0 或 1，则此候选关键单词或候选关键词串越可能是关键单词或关键词串；反之，如果候选关键单词或候选关键词串的首次出现的位置特征值越远离 0 或 1，则此候选关键单词或候选关键词串越可能是非关键单词或非关键词串。

5.1.3 TF×IF 特征

TF×IF 特征是对一个词或词序列在一篇文档中的出现次数与这个词或词序列在平时出现次数的另外一种比较特征。与 TF×IDF 中 IDF 的度量不同，IF 是度量这个词或词序列在一定规模语料中总共出现的次数的倒数，而不仅仅是出现的文档数的倒数。TF×IF 的具体计算公式如下：

$$TF*IF = freq(P,D) \times (-\log_2 \frac{freq(P)}{AllDocsWords}) \tag{5.3}$$

freq(P,D)同样表示词或词序列 P 在文档 D 中的出现的次数，freq(P)表示在一定规模的语料中词或词序列 P 总共出现了多少次；AllDocsWords 表示一定规模的语料总共有多少个词。在第六章的相关实验中我们将与 TF×IDF 的计算方式一样，用训练集表示一定规模的语料。当然也可以用第三方语料计算每个词或词序列的 TF×IF 特征值。

观察 TF*IDF 的计算公式我们可以发现，TF*IDF 中的 IDF 是计算一定规模语料中出现该候选关键单词或候选关键词串的文档数目的倒数，这种计算方式其实对候选关键单词或候选关键词串“关键”的度量是有缺陷的。因为如果某个无意义的候选关键单词或候选关键词串相对集中的出现在一定规模语料的少量文档中，那么此候选关键单词或候选关键词串的 IDF 值会过大，影响关键单词或关键词串的抽取。

因此我们针对 TF*IDF 特征对于候选关键单词或候选关键词串计算“关键”能力的不足，设计了 TF*IF 特征。当然特征 TF*IF 如同特征 TF*IDF 一样也有不足，如果某个有意义的候选关键单词或候选关键词串相对集中的出现在一定规模语料的少量文档中，且出现的次数非常频繁，则此候选关键单词或候选关键词串的 IF 值会过小，同样影响关键单词或关键词串的抽取。因此我们认为特征 TF*IDF 与特征 TF*IF 可以相辅相成，互相弥补对方计算候选关键单词或候选关键词串“关键”性的缺点。

通过 TF×IF 特征的计算公式，我们可以度量每个候选关键单词或候选关键词串的 TF×IF 特征值，如果候选关键单词或候选关键词串的 TF×IF 特征值越高，则此候选关键单词或候选关键词串越可能是关键单词或关键词串；反之，如果候选关键单词或候选关键词串的 TF×IF 特征值越低，则此候选关键单词或候选关键词串越可能是非关键单词或非关

键词串。

5.1.4 文档长度特征 NWT

文档长度特征 NWT 的思想比较简单，它仅仅是对文档所含词数的度量。但它却是对设计关键特征的一个重要补充。我们通过观察 TF*IDF 特征与 TF*IF 特征的计算公式发现，在文档篇幅差距比较大的情况下，长篇幅文档中的候选关键单词或候选关键词串的 TF*IDF 与 TF*IF 特征值普遍比短篇幅文档中的候选关键单词或候选关键词串的 TF*IDF 与 TF*IF 特征值高。这是因为同一个候选关键单词或候选关键词串在短篇幅文档中的词频往往低于长篇幅文档中的词频。因此我们引入文档长度 NWT 特征，克服因篇幅问题造成的计算候选关键单词或候选关键词串“关键”特征的差异。

5.2 关键单词特征设计

关键单词的特征设计只针对候选关键单词，计算候选关键单词的“关键”性。在实际的研究应用中我们发现，通常能针对候选关键单词计算“关键”的特征比较少，主要原因是候选关键单词在结构上比较完整，针对候选关键单词设计的特征同样适合候选关键词串。因此我们将词性特征引入候选关键单词的特征设计。

词性或词类是根据一个词的本意及在短语或句子中所起到的作用划分的。在转换生成语法中，词性称作词汇范畴。词性包括开放性的（经常吸收引进新词）与封闭性的（很少或从不吸收引进新词）。

在处理非母语语言时，词性情况通常很复杂，语言之间差别很大，很容易犯错。比如：西班牙语中，形容词和名词几乎可以互换使用，而英语中如果不曲折变化就几乎绝对不可以。日语有两类形容词（形容词和形容动词），而英语只有一种。汉语有量词，对欧洲人就非常陌生。 而且，很多语言对形容词和副词（或形容词和名词）之间并没有区分。所以要研究词性，必须在某个特定的语言框架下讨论，无法照搬到其它语言上。

我们通过对大量已标注关键词的文档进行词性标注后发现，无论是中文还是英文，绝大多数的关键单词都是名词，基于此我们设计了词性特征 CKWPS，具体计算公式如下：

$$\mathrm{CKWPS}=\begin{cases}1 & \text{如果候选关键单词是名词}\\ 0 & \text{否则}\end{cases} \tag{5.4}$$

如果候选关键单词是名词，就将候选关键单词标记为 1；如果候选关键单词不是名词，就将候选关键单词标记为 0。利用词性标注，将特征 CKWPS 添加到关键单词模型中，提高抽取文档关键单词的效果。

5.3 关键词串特征设计

正如我们在第三章提到的那样，关键词串的抽取在关键词抽取中是一个重点，也是一个难点。不同于关键单词抽取，关键词串抽取不但要解决“关键”的问题，同时还要解决词序列是否有意义的问题。我们在前面把词序列比作一条锁链，锁链的紧密程度不仅仅是依赖链节点的结实程度，更依赖于相连链节点之间结合的紧密性。因此候选关键词串是否是关键词串，不但要考虑词序列整体的重要性，还要考虑词序列内部词与词之间结合的紧密程度。

5.3.1 互信息特征

互信息 MI(mutual information)是统计模型中衡量两个随机变量 X 和 Y 之间关联程度的常用参数，它反映了两变量之间结合的紧密程度，互信息越大说明 X 和 Y 之间的相关性越强，互信息越小说明 X 和 Y 之间的相关性越弱。词序列的互信息定义如下：

$$MI(w_1w_2w_3\cdots w_{n-1}w_n) = Min(MI(w_1w_2), MI(w_2w_3)\cdots MI(w_{n-1}w_n)) \tag{5.5}$$

$$MI(w_{i-1}w_i) = \log\frac{p(w_{i-1}w_i)}{p(w_{i-1})\bullet p(w_i)} \tag{5.6}$$

$$p(w_{i-1}w_i) = \frac{n(w_{i-1}w_i)}{n(w)} \tag{5.7}$$

$$p(w_{i-1}) = \frac{n(w_{i-1})}{n(w)} \tag{5.8}$$

$$p(w_i) = \frac{n(w_i)}{n(w)} \tag{5.9}$$

其中 $MI(w_1w_2w_3\cdots w_{n-1}w_n)$表示词序列 $w_1w_2w_3\cdots w_{n-1}w_n$ 的结合的紧密程度，w_i 表示词，$n(w_i)$表示 w_i 在文本中出现次数，n(w)表示文本中的词数。

词串是一种结合紧密的词序列。如果词序列结合得越紧密，则该词序列越有可能是词串。前面提到词序列类似于一种链式结构，整条链结合的紧密程度并不是由所有相邻链结点子间的平均紧密程度决定，也不是由所有相邻链结点的最强连接强度决定，而是链中最薄弱的环节确定。因此词序列结合的紧密程度理所当然由所有相邻两个词之间互信息的最小值决定。

通过互信息 MI 特征的计算公式，我们可以度量每个候选关键词串的 MI 特征值，如果候选关键词串的 MI 特征值越高，则此候选关键词串越可能是关键词串；反之，如果候选关键词串的 MI 特征值越低，则此候选关键词串越可能是非关键词串。

基于此我们认为通过互信息计算候选关键词串相邻词之间的紧密程度，是判定任意一个词序列是否为词串的有效途径，本文将在第六章予以实验验证。

5.3.2 词串边界参数表特征

词串是由一些连续出现的词典词组成，而词串的串头词与串尾词都有一些共同的特点。比如，经常以副词、助词形式存在的词典中的词很少作为词串的串头词与串尾词，而有部分词典词却经常作为词串的串头词与串尾词。因此，我们据此规律构造了词串边界参数表，近似评估了所有词典词作为词串串头和串尾的可能性。

如果某个词在串头参数表中权值越大，则该词作为词串串头词的可能性越大，权值越小，则该词作为词串串头词的可能性越小。串尾参数表同样如此。图 4 是词串边界参数表的构造过程：

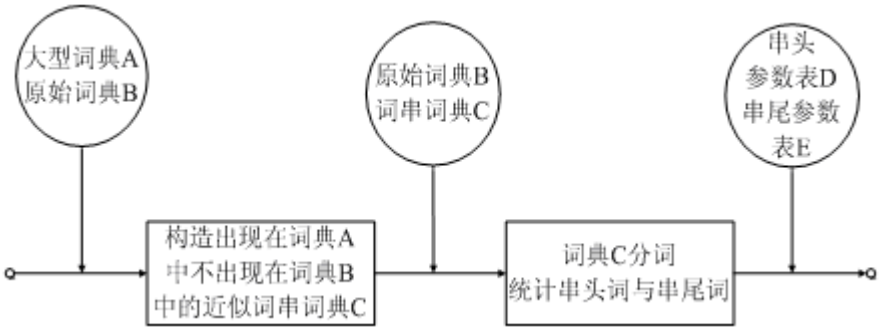


图 5.1　词串边界参数表构造过程

如图 5.1 所示，基本构造流程如下：

（1） 首先把一个拥有 50 多万个词条的词典作为大型词典 A，标准分词器 S 中的词典作为原始词典 B。词典 A 中不仅包含所有词典 B 中的词条，还包含许多人们日常生活经常用到的词串。标准分词器 S 来自我们自己开发的分词器。

（2） 从词典 A 中过滤掉所有出现在词典 B 中的词条，得到近似的词串词典 C。

（3） 利用标准分词器 S 对大型词典 A 进行分词，分词方法采用最长逆向匹配算法。

（4） 统计词典 B 中所有词条作为词典 C 中串头词与串尾词的数目，依次作为词典词的权值，从而生成串头参数表 D 和串尾参数表 E。

利用词串边界参数表识别词串是对单纯依靠互信息识别词串方法的一个重要补充。因为利用互信息识别词串时存在一个明显的缺点，就是抽取文档需要一定的长度，如果文档篇幅较短，往往利用互信息识别词串效果不理想，而利用词串边界参数表识别词串与文档篇幅无关。因此利用词串边界参数表识别词串对关键词串抽取具有重要的意义。

利用词串参数边界表我们设计了两个适合抽取关键词串的特征：串头参数 HB 特征和串尾参数 TB 特征。如果候选关键词串的串头参数 HB 特征值或串尾参数 TB 特征值越高，则此候选关键词串越可能是关键词串；反之，如果候选关键词串的串头参数 HB 特征值或串尾参数 TB 特征值越低，则此候选关键词串越可能是非关键词串。

5.3.3 候选关键词串结尾词词性特征

对于候选关键词串的抽取来说，关键词串的词性组合也是一个重要的信息。我们大量的实验研究发现，大部分关键词串都是名词性短语，为此我们为关键词串模型设计了候选关键词串结尾词词性特征 CKPTWPS，此特征的具体计算公式如下：

$$\mathrm{CKPTWPS}=\begin{cases}1 & \text{候选关键词串结尾词是名词}\\ 0 & \text{否则}\end{cases} \tag{5.10}$$

如果候选关键词串的最后结尾词是名词，CKPTWPS 就等于 1；如果候选关键词串的结尾词不是名词，就将 CKPTWPS 记为 0。利用此特征 CKPTWPS 添加到关键词串模型中来提高文档关键词串的抽取效果。

5.3.4 候选关键词串开头词词性特征

通过大量的实验研究发现，不仅关键词串的结尾词的词性特征是一个判定候选关键词串是否为关键词串的重要特征，而且候选关键词串的开头词同样对于关键词串抽取也有重要的意义。

我们发现大部分关键词串的开头词都是名词或形容词，为此我们为设计了候选关键词串开头词词性特征 CKPHWPS，此特征的具体计算公式如下：

$$\mathrm{CKPHWPS}=\begin{cases}1 & \text{候选关键词串开头词是名词或形容词}\\ 0 & \text{否则}\end{cases} \tag{5.11}$$

如果候选关键词串的开头词是名词或是形容词，则 CKPHWPS 等于 1；如果候选关键词串的开头词既不是名词也不是形容词，就将 CKPHWPS 记为 0。同特征 CKPTWPS 一样，将特征 CKPHWPS 添加到关键词串模型中，借此提高文档关键词串的抽取效果。

5.3.5 候选关键词串非结尾词中非形容词非名词的数目

候选关键词串除了开头词与结尾词对于关键词串抽取是重要的特征外，我们发现关键词串除结尾词外，剩下的词大部分是名词或形容词。因此我们开发了特征 NUM 用以计算候选关键词串中，除了结尾词剩下的词既不是名词也不是形容词的数目。

对于特征 NUM，如果候选关键词串的 NUM 特征值越高，则此候选关键词串越可能是关键词串；反之，如果候选关键词串的 NUM 特征值越低，则此候选关键词串越可能是非关键词串。

5.3.6 候选关键词串所含词数

另外我们将候选关键词串所含词数 LEN 作为抽取关键词串的一个特征。对于候选关键

词串来说，所含词数不同会使它们成为关键词串的可能性不同。

我们通过研究发现，大多数手工标记关键词的工作人员都喜欢使用两个词的关键词串标记关键词[44]，而且许多针对抽取关键词串设计的特征中都没有考虑不同长度的关键词串特征值是否分布在同一数值范围内，因此我们将候选关键词串所含词数 LEN 作为抽取关键词串的特征。

5.4 小结

本节主要介绍了分离模型中特征设计的问题。利用分离模型可以针对关键单词抽取与关键词串抽取分别设计特征的特点，主要从三个方面进行了相关特征设计的论述。

首先从传统的关键词抽取特征设计出发，介绍了既适合关键单词抽取又适合关键词串抽取的公共特征设计，详细说明了特征 TF×IDF、POS、TF×IF 以及 NWT 的基本思想和计算方法，并分析了这些特征对于关键词抽取的优点与不足。

其次针对关键单词的特点设计了特征 CKWPS，用以提高关键单词抽取的效果。

最后重点阐述了关键词串的特征设计。由于词串类似一种链式结构，因此我们设计了判定词串相邻词之间是否结合紧密的互信息特征，以及词串开头词与结尾词是否符合通常人们用词习惯的词串边界参数表特征。从词性角度出发设计了特征CKPTWPS、CKPHWPS、NUM。最后补充了候选关键词串所含词数特征 LEN，以提高文档关键词串抽取的效果。表 5.1 是对相关特征基本信息的总结：

表 5.1　特征基本信息

特征编号	特征名称	特征意义	取值范围	适用模型
(1)	TF*IDF	词频与反转 文档频率的积	(-∞,+∞)	关键单词模型 关键词串模型
(2)	POS	首次出现位置	(0,1)	关键单词模型 关键词串模型
(3)	NWT	文本所含词数	(0,+∞)	关键单词模型 关键词串模型
(4)	TF*IF	词频与反转 频率的积	(-∞,+∞)	关键单词模型 关键词串模型
(5)	LEN	词串所含词数	Int{0~4}	关键词串模型
(6)	MI	互信息	(-∞,+∞)	关键词串模型
(7)	HB	串头参数	Int (0~+∞)	关键词串模型
(8)	TB	串尾参数	Int (0~+∞)	关键词串模型

续表 5.1　特征基本信息

(9)	CKPTWPS	候选关键词 串结尾词词性	0,1	关键词串模型
(10)	CKPHWPS	候选关键词 串开头词词性	0,1	关键词串模型
(11)	NUM	候选关键词串非结尾 词非形容词非名词数目	Int{0~3}	关键词串模型
(12)	CKWPS	候选关键单词词性	0,1	关键单词模型

本节出要从理论与实现方法上介绍分离模型特征设计的问题，特征设计的有效性将在第六章进行实验验证。

第六章 实验与分析

前面介绍了基于分离模型的关键词抽取方法，即把关键词抽取分成关键单词抽取与关键词串抽取两个问题，并从关键单词抽取与关键词串抽取的实际角度出发，介绍了如何设计抽取关键单词与关键词串的相关特征。本章将从实验出发，进行分离模型与整体模型的比较验证，以及特征设计的有效性验证，最后将设计的特征添加到分离模型中与世界上最好的关键词抽取工具 KEA 进行比较分析。在这里我们需要说明的是，本文对中文与英文的关键词抽取都作了相关研究。在实验研究中发现，由于语言具有差异性，并不是所有设计的“关键”特征添加到模型中都能提高关键词的抽取效果。本文在第五章分离模型的特征设计中介绍的特征，有些适合中文的关键词抽取，有些适合英文的关键词抽取，当然有些特征对于中英文都适合，因此实验中，中英文的特征选择也有所不同。

6.1 实验方法

前面介绍了关键单词与关键词串的定义以及分离模型的本质，即把关键词抽取分成关键单词抽取与关键词串抽取两个问题。如何更好地利用分离模型完成关键词抽取任务，我们做了一些探索，提出了两种以分离模型为基础的实验方法：分类实验、评分实验。分类实验的目的主要是针对候选关键项进行是否为关键项的判定，而评分实验主要是对候选关键项按“关键”分数进行排序，计算分数较高的前 n 个候选关键项中有多少个是手工标注的关键词。下面对两种方法的具体实现过程分别进行介绍。

6.1.1 分类实验

选取一批已手工标注关键词的文档作为训练集。同时对每一个文档生成候选关键单词与候选关键词串，并以此作为每一个文档的关键单词候选项集合与关键词串候选项集合。每一个候选项按照表 5.1 计算特征，形成特征向量。如果候选关键单词或候选关键词串属于手工标注的关键词，则为正例，否则为反例。选取所有的候选关键单词样本作为关键单词模型训练样本集合，选取所有的候选关键词串样本作为关键词串模型训练样本集合。选取所有的候选关键单词样本与候选关键词串样本作为整体模型训练样本集合。接着我们利用 LIBSVM 对三个训练样本集合进行训练，获得关键单词模型、关键词串模型、整体模型。

对于新文档，首先自动获得候选关键单词集合与候选关键词串集合。然后对于每一个候选关键单词计算其分别假设其为关键单词，并根据该候选关键单词的特征获得特征向量，最后利用关键单词模型对候选关键单词进行是否为关键单词的判断。候选关键词串同样也如此。而整体模型可以同时判断候选关键单词与候选关键词串。

6.1.2 评分实验

在 LIBSVM 的二分类问题中，新样本的分类是通过模型中的分类器评分判定的。基于 LIBSVM 的实现原理，我们修改了 LIBSVM 的部分代码，使得 LIBSVM 训练出的模型可以对新样本成为正例的可能性评分。

与分类实验中构造训练模型方法一样，我们同样选取一批已手工标注关键词的文档作为训练集构造了关键单词评分器、关键词串评分器、整体评分器、综合评分器。对于新文档中的候选关键单词，计算该候选关键单词的特征并形成特征向量，利用关键单词评分器对其评分，分值越高，该候选关键单词越可能是关键单词；分值越低，则越可能是非关键单词。利用关键词串评分器对候选关键词串评分类似，而整体评分器与综合评分器可以同时对候选关键单词与候选关键词串评分。

6.1.3 语料介绍

我们对 4 个语料进行了关键词抽取实验验证，其中包括中、英文语料，语料基本信息见表 6.1。这些语料都有它们共同的特点，那就是它们都有手工标注的关键词。

表 6.1　语料基本信息

语料名称	语种	描述	平均每篇文档词数目(个)	语料文档数(篇)
Aliweb	英文	网页来自于 Aliweb 搜索引擎	949	90
Journals	英文	论文来自于五个学术期刊	10781	75
CSTR	英文	文档来自新西兰数字图书馆	9007	630
Blog	中文	来自 sohu 博客网站	1129	2096

我们对于每种语料进行关键词抽取实验。因为基于分离模型的关键词抽取方法主要是基于机器学习的方法，因此我们需要对语料进行适当训练集与测试集比例划分，以进行相关的训练与测试。我们将不同语料各自划分成 4 份，其中 3 份作为训练集，1 份作为测试集。

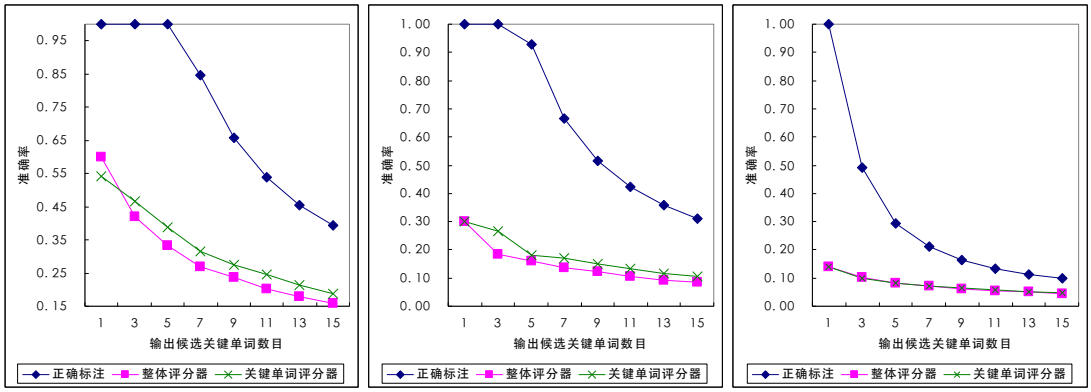
6.2 分离模型与整体模型比较

6.2.1 英文中分离模型与整体模型比较

我们按照 Turney[22]的方法取 Aliweb 语料中 55 篇文档作为 Aliweb 实验的训练集合，剩下 35 篇文档作为 Aliweb 实验的测试集合。Journals 实验训练集合数目为 55 篇，测试集合为 20 篇。由于 CSTR 语料文档数目远远大于 Aliweb 与 Journals 语料，因此按照 Witten[23]的方法，训练集合为 130 篇文档，测试集合为 500 篇文档。我们分别以抽取的候选关键词是否与标注好的关键词的词干形式完全匹配作为抽取正确的判定方式。这种方式与 Turney

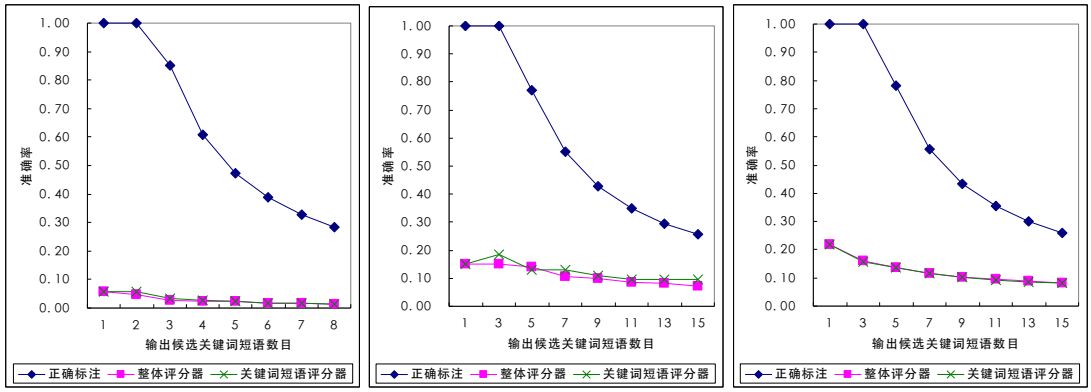
等和 Witten 等采取的评价标准一致。

在英文的关键词抽取中，选取公共特征 TF×IDF 特征与首次出现的位置 POS 特征可以取得关键词抽取较好的效果[23]。因此我们选择这两种特征作为关键词抽取中的“关键”特征，以实现分离模型与整体模型的比较实验。根据分离模型的思想，我们实现了关键单词评分器、关键词短语评分器和综合评分器，根据整体模型的思想我们实现了整体评分器。其中关键单词评分器支持对候选关键单词评分以此抽取关键单词；关键词短语评分器支持对候选关键词短语的评分；而总体评分器与综合评分器既支持对候选关键单词评分也支持对候选关键词短语的评分，并对候选关键单词与候选关键词短语不加区分的按评分高低抽取关键词。在模型的训练过程中由于每篇文档的非关键项数目远远多于关键项数目，使得训练样本的正例与反例极不平衡。为此我们采用了 Chong Huang[45]的方法，随机的在反例样本集合中选取样本，使得训练集中正例与反例的数目为 1:1。为了使分离模型与整体模型关键词抽取方法的实验结果容易比较分析，我们采取了 P@n 截取准确率判断标准。P@n 为抽取前 n 个候选关键词的准确率。n 限制在 15 以下，因为 n 太大没有实际的意义。以此我们得到三种语料的分离模型与整体模型比较实验结果，如图 6.1、图 6.2 与表 6.2、表 6.3:



(a) Aliweb 候选关键单词　(b) Journals 候选关键单词　(c) CSTR 候选关键单词

图 6.1　关键单词评分器与整体评分器比较



(a) Aliweb 候选关键词短语　(b) Journals 候选关键词短语　(c) CSTR 候选关键词短语

图 6.2　关键词短语评分器与整体评分器比较

表 6.2　关键单词评分器与整体评分器比较

	Aliweb (准确率)		Journals (准确率)		CSTR (准确率)	
输出候选关键单词数目	整体评分器	关键单词评分器	整体评分器	关键单词评分器	整体评分器	关键单词评分器
1	0.600	0.543	0.300	0.300	0.140	0.140
3	0.419	0.467	0.183	0.267	0.101	0.099
5	0.331	0.389	0.160	0.180	0.082	0.083
7	0.269	0.314	0.136	0.171	0.071	0.072
9	0.238	0.276	0.122	0.150	0.062	0.063
11	0.203	0.247	0.105	0.132	0.056	0.057
13	0.178	0.213	0.092	0.115	0.051	0.053
15	0.160	0.187	0.087	0.107	0.046	0.049

表 6.3　关键词串评分器与整体评分器比较

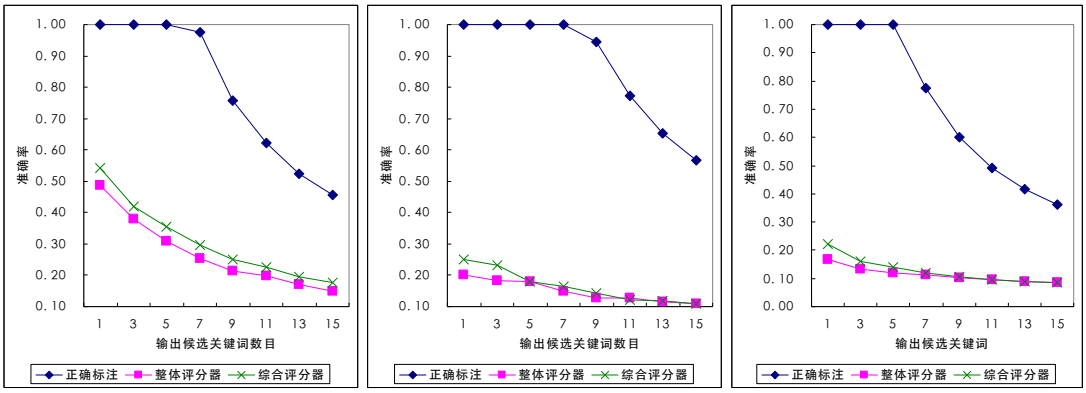
	Aliweb (准确率)		Journals (准确率)		CSTR (准确率)	
输出候选关键词串数目	整体评分器	关键词串评分器	整体评分器	关键串评分器	整体评分器	关键串评分器
1	0.057	0.057	0.150	0.150	0.218	0.218
3	0.048	0.057	0.150	0.183	0.161	0.158
5	0.029	0.034	0.140	0.130	0.137	0.136
7	0.024	0.029	0.107	0.129	0.117	0.115
9	0.022	0.022	0.100	0.111	0.103	0.102
11	0.018	0.018	0.086	0.095	0.095	0.094
13	0.015	0.015	0.081	0.096	0.087	0.085
15	0.013	0.013	0.073	0.097	0.082	0.080

从图 6.1 可以看出在输出候选关键单词数目一定的情况下，利用关键单词评分器提取关键单词的截取准确率普遍高于利用整体评分器提取关键单词的截取准确率。而从图 6.2 可以看出在输出候选关键词短语数目一定的情况下利用关键词短语评分器提取关键词短语的截取准确率略微高于利用整体评分器提取关键词短语的截取准确率。以上说明利用分离模型提取算法分别提取关键单词与关键词短语的效果比整体模型分别对关键单词与关键词短语的提取效果好。

为了进一步分析实际应用情况下的效果，我们不仅对关键单词和关键词短语利用分离模型与整体模型进行了比较，我们还还利用综合评分器与整体评分器对测试集中的语料进行关键词抽取效果（即将关键单词和关键词短语进行同一度量排序）的比较，同样采用截取准确率判断优劣，如图 6.3、表 6.4。

从图 6.3(a)~图 6.3(c)可以看出在输出候选关键词数目小于 7 的情况下利用综合评分器抽取关键词的截取准确率明显高于利用整体评分器抽取关键词的截取准确率。对于 Journals 与 CSTR 语料，在输出的候选关键词数目大于 7 时，综合评分器关键词抽取效果略为优于

整体评分器。



(a) Aliweb 候选关键词　　(b) Journals 候选关键词　　(c) CSTR 候选关键词

图 6.3　综合评分器与整体评分器比较

我们还可以从图 6.3(a)看出 Aliweb 语料无论输出的候选关键词数目为何值（当然数目小于 15），综合评分器抽取关键词的截取准确率都明显高于利用整体评分器抽取关键词的截取准确率，这说明基于分离模型的关键词抽取方法对于篇幅较小的文档有更好的抽取效果。

表 6.4　综合评分器与整体评分器比较

	Aliweb (准确率)		Journals (准确率)		CSTR (准确率)	
输出候选关键词数目	整体评分器	综合评分器	整体评分器	综合评分器	整体评分器	综合评分器
1	0.486	0.543	0.200	0.250	0.168	0.222
3	0.381	0.419	0.183	0.233	0.134	0.160
5	0.309	0.354	0.180	0.180	0.120	0.141
7	0.253	0.298	0.150	0.164	0.113	0.121
9	0.213	0.251	0.128	0.144	0.103	0.107
11	0.197	0.226	0.127	0.123	0.097	0.097
13	0.171	0.196	0.115	0.119	0.090	0.090
15	0.486	0.543	0.200	0.250	0.168	0.222

总的来说，对于英文的关键词抽取问题，无论抽取的候选项是候选关键单词或是候选关键词短语，还是不加区分的候选关键词，基于分离模型的关键词抽取方法抽取关键词的效果都优于基于整体模型的关键词抽取方法。

6.2.2 中文中分离模型与整体模型比较

我们从 Web 网站中抓取了博客网页作为关键词抽取测试的语料。因为每篇博客中都有 tag 标签，可以看成作者手工标注的关键词。我们选取了其中拥有 5 个 tag 标签的中文博客，总共有 2096 篇。每篇博客的平均词数为 1270。由于很多 tag 标签并没有出现在它自己的博客中，因此所有语料总共只拥有 9339 个 tag 标签。我们选取其中 1572 篇博客作为训练集，

剩下的 524 篇博客作为测试集。

由于中英文语言的差异，我们发现在中文的关键词抽取中，我们所设计的所有关键单词与关键词串公共特征对于关键词抽取都可以取得较好的效果。因此为了在进行分离模型与整体模型的比较实验过程中同时验证这些特征对于关键词抽取得作用，我们采取一套不同于英文中分离模型与整体模型比较实验的方法。我们以特征(1)TF*IDF 和特征(2)POS 为基准特征，然后依次添加特征(3)NWT 和特征(4)TF*IF，实验证明特征(3)NWT 和特征(4)TF*IF 对于关键词提取的意义。

我们依然利用 LIBSVM 对训练集中的候选关键单词与候选关键词串按照表 1 选取的公共特征进行训练，同样由于每篇文本的非关键词数目远远多于关键词数目，使得训练样本的正例与反例极不平衡。仍然采用 Chong Huang 的方法，随机地在反例样本集合中选取样本，使得训练集中正例与反例的数目基本为 1:1，具体数目见表 6.5：

表 6.5　分类实验训练集中正例与反例的具体数目

关键单词模型		关键词串模型		整体模型	
正例数目	反例数目	正例数目	反例数目	正例数目	反例数目
5478	5516	1154	1128	6632	6644

按照分类实验的方法对训练样本集合进行训练，我们得到了关键单词模型、关键词串模型、整体模型。然后我们分别利用这些模型对测试集进行分类实验测试，结果如表 6.6、表 6.7：

表 6.6　分类实验候选关键单词测试结果

特征选取(编号表示)	关键单词模型				整体模型			
	正例准确率(%)	反例准确率(%)	整体准确率(%)	整体F1值(%)	正例准确率(%)	反例准确率(%)	整体准确率(%)	整体F1值(%)
(1)(2)	81.513	88.961	88.902	85.047	80.375	89.327	89.255	84.583
(1)(2)(3)	80.091	92.417	92.319	85.771	76.735	93.793	93.657	84.356
(1)(2)(3)(4)	79.750	93.257	93.149	85.930	81.172	92.781	92.688	86.549

表 6.7　分类实验候选关键词串测试结果

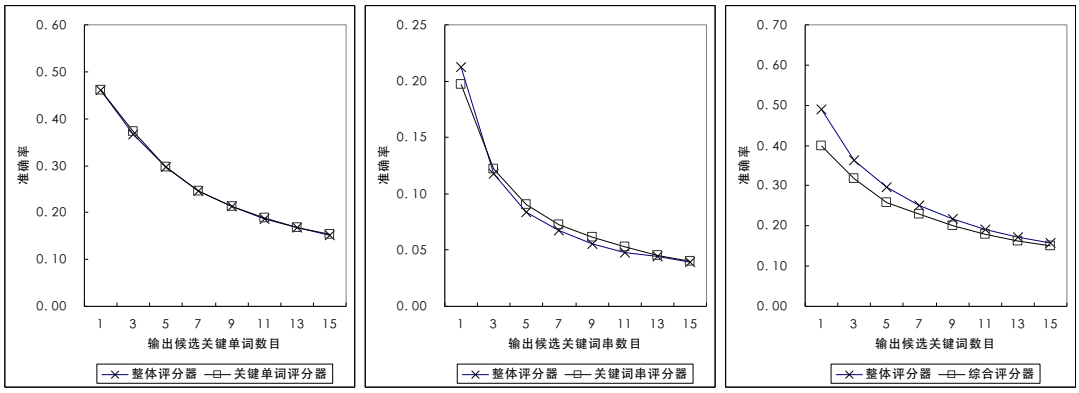
特征选取(编号表示)	关键词串模型				整体模型			
	正例准确率(%)	反例准确率(%)	整体准确率(%)	整体F1值(%)	正例准确率(%)	反例准确率(%)	整体准确率(%)	整体F1值(%)
(1)(2)	72.421	96.668	96.647	82.798	75.368	86.882	86.872	80.712
(1)(2)(3)	71.579	97.285	97.262	82.467	74.947	91.860	91.845	82.540
(1)(2)(3)(4)	77.474	96.669	96.652	86.007	75.368	91.848	91.833	82.790

测试集中包含 1758 个关键单词、218311 个非关键单词、475 个关键词串、540072 个非关键词串。从表 6.6 与表 6.7 中我们可以看出，分离模型比整体模型在候选关键词串测试中，模型在选取相同的公共特征条件下，分离模型比整体模型效果更好，整体 F1 值提

高了 1%~3%。而这种提高在候选关键单词中并不非常显著。实验数据总体上说明，在选取同样特征的情况下，基于分离模型的关键词抽取比整体模型好。

从表 6.6 我们还可以看出，添加特征(3)NWT 时，关键单词模型的整体 F1 值为 85.771%，比不添加特征(3)NWT 时关键单词模型的整体 F1 值 85.047%高。同时添加特征(3)NWT 和特征(4)TF*IF 时，关键单词模型的整体 F1 值为 85.930%，整体模型的整体 F1 值为 86.549%，这比相同条件下不添加特征(3)NWT 和特征(4)TF*IF 时，整体 F1 值高，这说明特征 NWT 与 TF*IF 对于中文关键单词抽取是有帮助的。同样从表 6.7 中我们可以得出这两个特征对于关键串抽取也是有意义的结论。

我们按评分实验的方法以同样的训练集合构造了关键单词评分器、关键词串评分器、整体评分器、综合评分器。同样采取了 P@n 截取准确率判断标准进行分离模型与整体模型在中文关键词抽取的比较实验，需要说明的是，对于中文关键词抽取所有的公共特征都有一定的意义，因此我们选择(1)~(4)特征作为这两模型的特征，并以此构造相关评分器。实验结果如图 6.4、表 6.8：



(a) Blog 候选关键单词　　(b) Blog 候选关键词串　　(c) Blog 候选关键词

图 6.4　Blog 语料整体模型与分类模型比较

表 6.8　Blog 语料整体模型与分类模型比较

	Blog (准确率)					
输出候选关键项数目	整体评分器	关键单词评分器	整体评分器	关键词串评分器	整体评分器	综合评分器
1	0.460	0.460	0.213	0.197	0.490	0.398
3	0.367	0.372	0.118	0.122	0.363	0.319
5	0.297	0.297	0.084	0.091	0.297	0.259
7	0.245	0.247	0.067	0.073	0.250	0.229
9	0.214	0.214	0.056	0.062	0.218	0.201
11	0.187	0.187	0.048	0.053	0.192	0.180
13	0.167	0.169	0.044	0.045	0.171	0.163
15	0.152	0.153	0.039	0.040	0.158	0.151

从图 6.4、表 6.8 我们可以看出在公共特征选取相同的情况下，关键单词评分器抽取关

键单词的效果略好于整体评分器抽取关键单词的效果，关键词串评分器抽取关键词串的效果明显好于整体评分器。虽然对于关键词来说，综合评分器抽取关键词的整体效果不如整体评分器，但基本上处于同一水平。因此我们可以得出结论，在中文关键词抽取中，选取特征相同的条件下，基于分离模型方法的关键词抽取效果与基于整体模型的方法效果相当。

6.3 关键单词特征与关键词串特征的作用

6.3.1 英文中关键单词特征与关键词短语特征实验

我们采取同第 6.2.1 节一样的方法构造了 Aliweb、Journals 和 CSTR 语料的训练集与测试集。并且选取公共特征 TF*IDF 和 POS 作为基础特征，计算方法保持不变。

在英文的关键词抽取中，我们发现统计特征添加到分离模型中对于关键词抽取效果增加不显著，而词性特征对于抽取关键单词与关键词短语的效果有明显改善。因此我们将词性特征添加到分离模型中，以验证这些特征对于关键单词抽取与关键词短语抽取的作用。

对候选关键词的词性标注我们采用 Stanford Tagger [46]。并以此分别计算候选关键单词与候选关键词短语的词性特征。我们以分离模型方法为基础，对训练集进行训练得到训练模型，并将其改造得到了四个评分器分别是：NoPOSWordScorer、POSWordScorer、NoPOSPhraseScorer、POSPhraseScorer。每个评分器能够对对应的候选关键项进行是否为关键项的评分，以此抽取关键项。具体评分特征设计与适用范围见表 6.9：

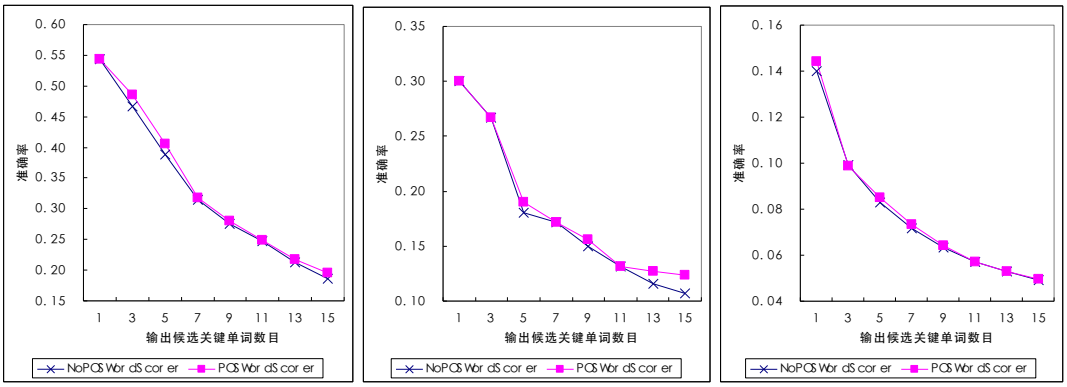
表 6.9　英文中相关评分器特征设计

评分器种类	特征编号	适用范围
NoPOSWordScorer	(1)(2)	候选关键单词
POSWordScorer	(1)(2)(12)	候选关键单词
NoPOSPhraseScorer	(1)(2)	候选关键词短语
POSPhraseScorer	(1)(2)(9)(10)(11)	候选关键词短语

我们同样了采取 P@n 截取准确率判断标准，得到三种语料实验结果如图 6.5、图 6.6，表 6.10、表 6.11。

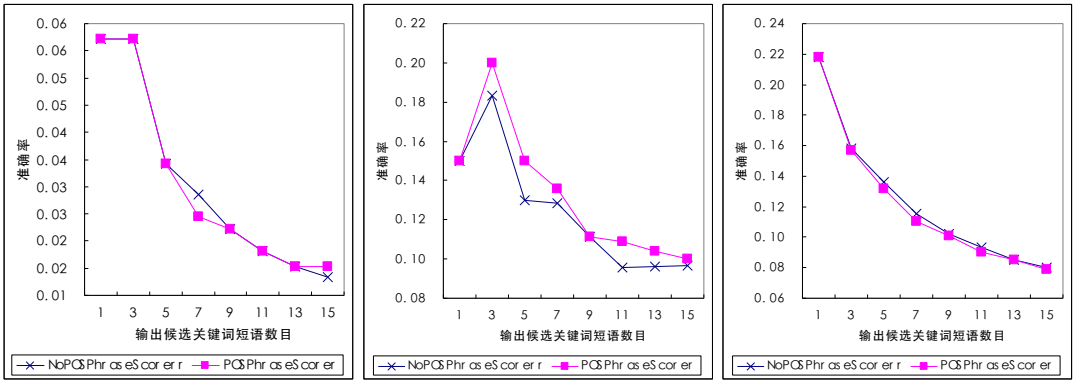
从图 6.5(a)、表 6.10 可以看出，Aliweb 语料输出候选关键单词中，除了输出数目为 1 和 7 时，在输出其它任意数目候选关键单词时，添加词性特征的截取准确率都比不添加词性特征的截取准确率高。虽然图 6.6(a)、表 6.11 显示，添加词性特征效果不明显，但基本上处于同一水平，且在输出数目为 15 时，截取准确率提高 0.02。我们从图 6.5(b)、图 6.6(b)、表 6.10 表 6.12 可以看出，对于 Journals 语料，添加词性特征的评分器比不添加词性特征的评分器，抽取关键项的截取准确率提高明显，其中抽取关键词短语的截取准确率平均提高 0.0086，这说明词性特征对于 Journals 语料提高抽取关键词准确率效果明显。主要原因在

于，Journals 语料都是论文，论文语法规范，关键词词性组合方式较规则。



(a) Aliweb 候选关键单词　　(b) Journals 候选关键单词　　(c) CSTR 候选关键单词

图 6.5　添加关键单词特征实验



(a) Aliweb 候选关键词短语　　(b) Journals 候选关键词短语　　(c) CSTR 候选关键词短语

图 6.6　添加关键词短语特征实验

我们从图 6.5(c)、图 6.6(c)、表 6.10 表 6.11 可以看出，对于 CSTR 语料，添加词性特征的评分器与不添加词性特征的评分器，抽取关键项的截取准确率基本一样，这说明词性特征对于 CSTR 语料提高抽取关键词准确率效果不明显。主要原因在于，CSTR 语料不如 Journals 语料语法规范。

表 6.10　添加关键单词特征实验

	Aliweb (准确率)		Journals (准确率)		CSTR (准确率)	
输出候选关键单词数目	NoPOS WordScorer	POS WordScorer	NoPOS WordScorer	POS WordScorer	NoPOS WordScorer	POS WordScorer
1	0.543	0.543	0.300	0.300	0.140	0.144
3	0.467	0.486	0.267	0.267	0.099	0.099
5	0.389	0.406	0.180	0.190	0.083	0.085
7	0.314	0.318	0.171	0.171	0.072	0.073
9	0.276	0.279	0.150	0.156	0.063	0.064
11	0.247	0.249	0.132	0.132	0.057	0.057
13	0.213	0.218	0.115	0.127	0.053	0.053
15	0.187	0.196	0.107	0.123	0.049	0.049

表 6.11　添加关键词串特征实验

	Aliweb (准确率)		Journals (准确率)		CSTR (准确率)	
输出候选关键词串数目	NoPOS PhraseScorer	POS PhraseScorer	NoPOS PhraseScorer	POS PhraseScorer	NoPOS PhraseScorer	POS PhraseScorer
1	0.057	0.057	0.150	0.150	0.218	0.218
3	0.057	0.057	0.183	0.200	0.158	0.157
5	0.034	0.034	0.130	0.150	0.136	0.132
7	0.029	0.024	0.129	0.136	0.115	0.110
9	0.022	0.022	0.111	0.111	0.102	0.101
11	0.018	0.018	0.095	0.109	0.094	0.090
13	0.015	0.015	0.096	0.104	0.085	0.085
15	0.013	0.015	0.097	0.100	0.080	0.079

总的来说，无论是抽取关键单词还是关键词短语，在添加词性特征后关键项的抽取效果都得到了提高，而且词性特征对于语法较规范的文档意义更大。

6.3.2 中文中关键单词特征与关键词串特征实验

正如我们前面讲到的语言具有差异性。我们发现统计特征和语言学特征对于中文关键词抽取效果都有明显改善，尤其是在关键词串上，添加统计特征和语言学特对于抽取关键词串作用更加显著。

我们将一些新的特征与公共特征添加到关键单词模型和关键词串模型，并与只添加公共特征的关键单词模型和关键词串模型进行了比较实验。实验依然从分类实验与评分实验两个角度进行，与 6.2.2 的实验方法一样，我们选用 Blog 语料得到分类实验结果，见表 6.12、表 6.13：

表 6.12　分类实验添加特征候选关键单词测试结果

特征选取(编号表示)	关键单词模型			
	正例准确(%)	反例准确率(%)	整体准确率(%)	整体 F1 值(%)
(1)(2)(3)(4)	79.750	93.257	93.150	85.930
(1)(2)(3)(4)(12)	81.627	92.090	92.006	86.506

表 6.13　分类实验添加特征候选关键词串测试结果

特征选取(编号表示)	关键词串模型			
	正例准确(%)	反例准确率(%)	整体准确率(%)	整体 F1 值(%)
(1)(2)(3)(4)	77.474	96.669	96.653	86.007
(1)(2)(3)(4)(5)(6)(7)(8)	80.000	95.653	95.639	87.123
(1)(2)(3)(4)(5)(6)(7)(8) (9)(10)(11)	81.053	95.740	95.727	87.781

从表 6.12 可以看出，添加特征(12)CKWPS 后关键单词模型整体 F1 值为 86.506%高于不添加特征(12)CKWPS 关键单词模型整体 F1 值，这说明新特征对于关键单词抽取是有意

义的。

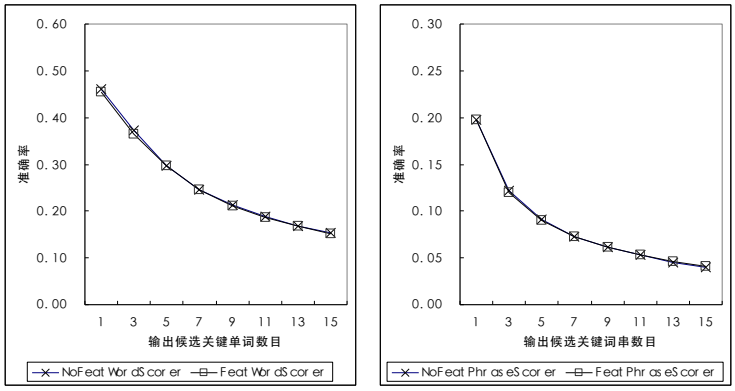
对于关键词串模型，我们发现添加统计特征后关键词串模型抽取关键词串的整体 F1 值为 87.123%，高于不添加统计特征而只添加公共特征的关键词串模型整体 F1 值。添加统计特征与语言学特征的关键词串模型整体 F1 值为 87.781%，高于添加统计特征的关键词串模型。这说明统计特征与语言学特征能够提高关键词串的抽取效果。

对于评分实验我们同样以分离模型算法为基础，对训练集进行训练得到训练模型，并将其改造得到了四个评分器分别是：NoFeatWordScorer、FeatWordScorer、NoFeatPhraseScorer、FeatPhraseScorer。评分器能够对对应的候选关键项进行是否为关键项的评分，以此抽取关键项。具体评分特征设计与适用范围见表 6.14：

表 6.14　中文中相关评分器特征设计

评分器种类	特征编号	适用范围
NoFeatWordScorer	(1)(2)(3)(4)	候选关键单词
FeatWordScorer	(1)(2)(3)(4)(12)	候选关键单词
NoFeatPhraseScorer	(1)(2)(3)(4)	候选关键词串
FeatPhraseScorer	(1)(2)(3)(4)(5)(6)(7)(8)(9)(10)(11)	候选关键词串

我们同样采取与英文中关键单词特征与关键词串特征一样的实验方法，以 P@n 截取准确率判断标准，得到实验结果如图 6.7、表 6.15：



Blog 候选关键单词　　Blog 候选关键词串

图 6.7　Blog 添加特征实验

表 6.15　Blog 添加特征实验

	Blog (准确率)			
输出候选关键项数目	NoPOS WordScorer	POS WordScorer	NoPOS PhraseScorer	POS PhraseScorer
1	0.460	0.454	0.197	0.197
3	0.372	0.365	0.122	0.119
5	0.297	0.297	0.091	0.090
7	0.247	0.246	0.073	0.073
9	0.214	0.211	0.062	0.061

续表 6.15　Blog 添加特征实验

11	0.187	0.186	0.053	0.053
13	0.169	0.168	0.045	0.046
15	0.153	0.152	0.040	0.041

从图 6.7、表 6.15 中我们可以看到无论对于关键单词评分器还是关键词串评分器，添加新的特征对于在输出相同数目的候选关键项时，关键项截取准确率的提高意义不大。但在前面的分类实验中我们发现新的特征对于分类判断是有意义的。这说明新的特征在关键项排序基本不变的情况下，可以大量的过滤非关键项。

6.4 与 KEA 的比较实验

我们基于分离模型的关键词抽取方法支持对关键单词抽取与关键词串抽取添加各自适合的特征的基础上，根据中英文语言上的差别，将表 5.1 的特征添加到分离模型的方法中，并以此与经典的关键词抽取器 KEA 进行了比较。仍然需要强调的是，不是所有的特征适合任意语言的关键词抽取，我们按照 6.1 与 6.2 节提到的特征进行对应的特征添加，具体如表 6.16：

表 6.16　综合评分器特征设计

评分器名称	模型	特征编号	语种
POSScorer	关键单词模型	(1)(2)(12)	英文
	关键词短语模型	(1)(2)(9)(10)(11)	英文
FeatScorer	关键单词模型	(1)(2)(3)(4)(12)	中文
	关键词串模型	(1)(2)(3)(4)(5)(6)(7)(8) (9)(10)(11)	中文

我们以此与 KEA3.0[47]进行抽取效果的总体比较。KEA 是一个基于文本的关键词抽取算法，它是由新西兰怀卡托大学开发。KEA 通过计算每个候选关键词的 TF×IDF 特征与首次出现的位置 POS 特征，利用机器学习算法来判断每个候选关键词成为关键词的可能性。它首选需要通过一批以标注好关键词的文档集合作为训练集进行训练，然后通过 Bayes 算法得到训练模型，以此对新的候选文档的每个候选关键词成为关键词可能性进行评分。通过输入需要的候选关键词数目 n，KEA 自动抽取新文档中评分最高的前 n 个候选关键词。

6.4.1 与 KEA 在英文关键词抽取上的比较

我们设置 KEA3.0 的参数使得抽取候选关键词的长度为 1~4，候选关键词最小词频为 1。输出测试集中每篇文档分数最高的前 5 个与前 15 个候选关键词，并将其与 POSScorer 进行比较，我们仍然利用 Aliweb、Journasl、CSTR 语料进行比较实验。结果如图 6.8、6.9。

我们从图 6.8、6.9 可以看出，在输出候选关键词数目为 5 时，Aliweb 测试集利用 POSScorer 平均输出的关键词数目为 1.83 个，多于 KEA3.0 的 1.68 个；输出候选关键词数

目为 15 时，POSScorer 平均输出的关键词数目为 2.66 个，也多于 KEA3.0 的 2.31 个。而利用 POSScorer 在输出候选关键词数目为 5 时平均输出的 Journasl 测试集的关键词数目为 1.05 个，低于 KEA3.0 的 1.20 个；而在输出候选关键词数目为 15 时，POSScorer 达到 1.90 个，高于 KEA3.0 的 1.85 个。CSTR 测试集中，POSScorer 在输出的候选关键词数目为 5 时，平均输出的关键词数目为 0.71 个，略低于 KEA3.0 的 0.78 个；在输出候选关键词数目为 15 时，POSScorer 平均输出 1.28 个关键词，而 KEA3.0 输出 1.52 个，虽然 CSTR 语料无论在输出候选关键词数目为 5 还是 15 时，都低于 KEA3.0，但对于三个语料来说，POSScorer 与 KEA3.0 抽取关键词的准确率基本上处于同一水平。

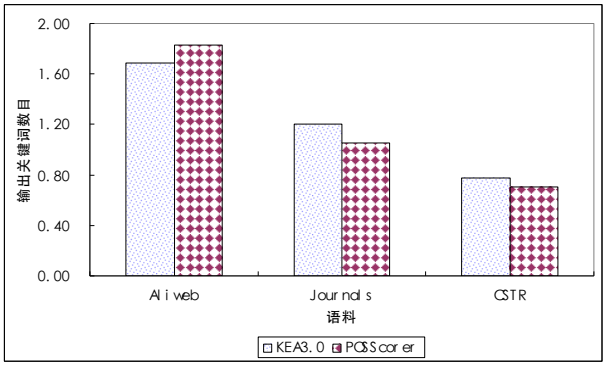


图 6.8　输出前 5 个候选关键词比较图

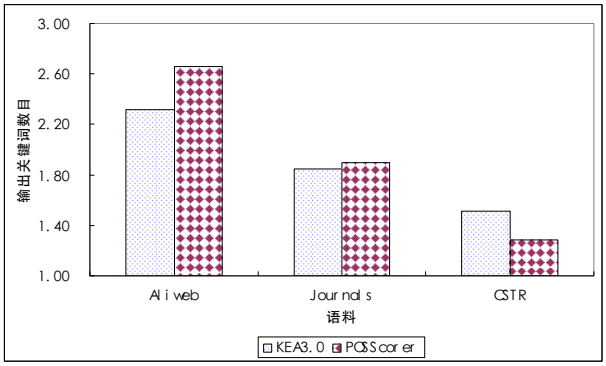


图 6.9　输出前 15 个候选关键词比较图

从上述实验结果我们不仅发现 POSScorer 与 KEA3.0 对于 Journals 语料与 CSTR 语料的关键词抽取效果相当，而且 POSScorer 对短文本的 Aliweb 语料关键词抽取效果更好，未来我们将对短文本的关键词抽取作进一步的深入研究。

另外我们利用三个英文语料抽取 90 篇文档进行显著性实验验证，实验结果表明在置信度为 95%的情况下，抽取 5 个候选关键词时 KEA3.0 好于 POSScorer，抽取 15 个候选关键词时 KEA3.0 与 POSScorer 没有显著差别。

6.4.2 与 KEA 在中文关键词抽取中的比较

我们同样基于分离模型的关键词抽取方法支持对关键单词抽取与关键词串抽取添加

各自适合的特征的基础上，适合中文关键词抽取的所有统计学特征添加到算法中，构造了添加统计学特征与语言学特征的综合评分器 FeatScorer，并以此同样与 KEA3.0 进行在 Blog 语料上关键词抽取效果的总体比较。

我们依然将 KEA3.0 的参数设置为抽取候选关键词的长度为 1~4，候选关键词最小词频为 1。输出测试集中每篇文档分数最高的前 5 个与前 15 个候选关键词，得到结果如图 6.10：

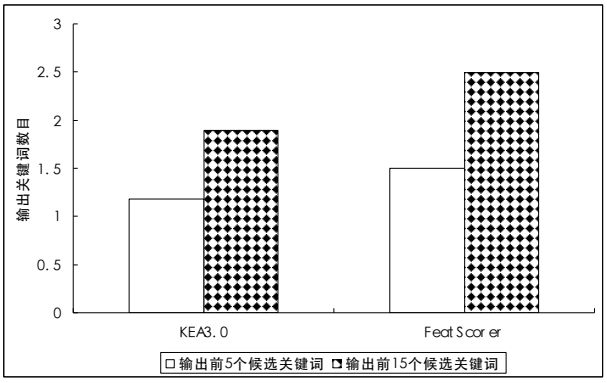


图 6.10　与 KEA3.0 的比较

我们从图 6.10 可以看出，无论在输出候选关键词数目为 5 还是 15 时，综合评分器 FeatScorer 对于关键词的抽取效果都明显好于 KEA3.0。这说明中文中基于分离模型的关键词抽取方法优于 KEA 方法。

另外我们利用三个中文 Blog 语料抽取 90 篇文档进行显著性实验验证，实验结果表明在置信度为 95%的情况下，无论抽取 5 个还是 15 个候选关键词时 FeatScorer 好于 KEA3.0。

6.5 小结

本章主要从实验出发，对关键词抽取的若干问题进行了实验验证并进行了相关分析。主要从分离模型与整体模型的比较；关键单词特征与关键词串特征的作用；与 KEA 的比较实验三个方面进行了验证分析。

本章首先从整体上介绍了实验方法，然后介绍了相关的实验语料。然后遵循实验方法，分别对中英文关键词抽取进行了分离模型与整体模型的实验比较，实验结果显示：在选取公共特征条件一样的情况下，无论是中文还是英文，分离模型对于关键词抽取的效果要优于整体模型。接着本章进行了特征设计的实验，基于分离模型可以分别对候选关键单词与候选关键词串设计不同特征的优点，设计开发了一组适合语言特性与词串组成特性的特征，提高了关键单词与关键词串的抽取效果。最后将基于分离模型的关键词抽取方法与特征设计相结合，构造了综合评分器，并与经典的关键词抽取器 KEA 进行了比较实验，实验结果表明基于分离模型的关键词抽取方法在英文中与 KEA 水平相当，中文中优于 KEA。

第七章 结束语

“到处是水，却没有一滴水可喝”，《娱乐至死》的作者尼尔•波兹曼喜欢用柯勒律治这句话来形容信息过剩时代人们的处境。在信息时代，我们时刻浸泡在信息的海洋里，却又时刻面临信息过剩的苦恼。所谓“乱花渐欲迷人眼”，信息激增的同时也导致了信息过剩和信息泛滥。人们在信息的海洋中要“广、快、精、准”地查找到自己所需要的信息已变得越来越困难。关键词作为信息时代新生产物，能在一定程度上解决信息过剩的问题。

本文从关键词抽取实际角度出发，将关键词分类为关键单词与关键词串。并根据关键单词与关键词串不同特点，提出了基于分离模型的关键词抽取方法。基于分离模型的关键词抽取方法将关键词的模型训练与实际抽取分成两个问题：关键单词抽取与关键词串抽取。基于分离模型的关键词抽取方法能够在关键单词模型与关键词串模型中设计各自不同的特征，以提高各自关键项的抽取效果。

我们认为词串类似一种链式结构，它的完整性不仅依靠其组成部分（词）的完整性，更依赖于词串内部相连词之间结合的紧密程度。因此我们基于词串的特点在关键词串的抽取中设计了如互信息，词串边界参数表等特征，以提高词串的识别率，解决关键词抽取中“词”的问题。

传统的关键词抽取中，有许多“关键”特征早已被研究者发现并应用于实际的抽取过程，如 TF×IDF 特征与首次出现的位置 POS 特征。但我们研究发现这些特征都有一定的局限性，且它们是针对关键词抽取问题整体开发的特征，并没有对关键单词与关键词串抽取加以区别。基于分离模型的关键词抽取方法在特征设计时，除了针对 TF×IDF 特征与首次出现的位置 POS 特征的不足，设计了公共特征文档长度 NWT 特征与 TF×IF 特征，还根据关键单词抽取与关键词串抽取设计了一些适合各自问题的统计与语言学特征。

本文最后通过一系列的实验验证了无论在中文还是英文中，在公共特征选取一样的情况下，基于分离模型的关键词抽取效果好于基于整体模型的关键词抽取。另外我们通过实验验证了一些针对关键单词与关键词串抽取设计的特征，对于关键词抽取有正面的作用。最后我们将开发的基于分离模型的关键词抽取方法与世界上著名的关键词抽取器 KEA 进行了比较，取得了英文中与 KEA 水平相当，中文中好于 KEA 的实验结果。

未来进一步的工作将主要集中在以下方面：

（1） 特征设计上，通过设计更多的特征提高关键词的抽取效果。

（2） 另外在模型训练时，存在正例与反例数量不平衡的问题，这是机器学习正在研究的 UNEVEN 问题，如何解决需要作进一步工作。

（3） 最后未来工作的重点还需要解决分离模型如何合并关键单词与关键词串形成关键词的问题。

致　谢

记得我的导师王挺老师在研一的时候曾对我说过：“年轻就得拼，老了没机会”！两年半的硕士学习中我一直以此为座右铭，不断提醒自己“努力，努力，再努力”！今天我的硕士论文历经几番反复，终于完成了！论文的完成虽然历经波折，但我却料到致谢词会一气呵成。因为这份硕士论文，实在包含了太多人投注的关心和热情。

首先感谢我的恩师王挺教授。记得第一次见到王老师的时候我就被王老师严谨细致、一丝不苟的作风深深吸引。记得当时王老师仔细询问了我本科的学习生活情况，并对我未来硕士生活进行了细致的规划。逝者如斯夫！当初的规划已浓缩于此论文，感谢王老师！

其次我要感谢科大所有老师。没有你们的谆谆教诲也没有我的今天。你们无私的奉献，无微不至的关怀令我感动，谢谢！

另外我要感谢我们自然语言课题组所有的师兄、师姐、师弟、师妹，没有你们的大力支持我无法完成课题的研究。特别感谢刘伍颖师兄，没有您不厌其烦的向我提供帮助，我的工作也不会进行得如此顺利。

我还要感谢六院五队所有的队领导、同学。这是一个优秀的集体！这是一个光荣的集体！有了这个集体我的生活才如此多彩，有了这个集体我的生命才如此完整！

谢谢陆华彪、邹丹、何明，这些我可爱的室友，我最亲的兄弟。是你们陪我度过了无数的夜晚，你们永远是我最好的朋友！

最后，我要感谢我的父母，是你们养育了我，教育了我。没有你们我绝对成不了现在的我。爸爸、妈妈我爱你们！

参考文献

[1] 朱夫斯凯. 自然语言处理综论. 电子工业出版社. 2005

[2] 张晓艳. 命名实体识别研究，国防科学技术大学计算机软件与理论专业硕士学位论文，2005

[3] Luhn HP. The automatic creation of literature abstracts. Journal of Research and Development, 1958.2(2):159~165

[4] Edmundson HP. New methods in automatic extracting. Journal of the Association for Computing Machinery, 1969.16(2):264~285

[5] Marsh E. Hamburger H. and Grishman R. A production rule system for message summarization. In: AAAI-84, Proceedings of the American Association for Artificial Intelligence, AAAI Press/MIT Press, Cambridge, MA, 1984:243~246

[6] Paice CD. Constructing literature abstracts by computer: Techniques and prospects. Information Processing and Management, 1990,26(1):171~186

[7] Paice CD and Jones PA. The identification of important concepts in highly structured technical papers. In: SIGIR-93: Proceedings of the 16th Annual InternationalACMSIGIR Conference on Research and Development in Information Retrieval, ACM, New York, 1993.1:69~78

[8] Johnson FC, Paice CD., Black WJ, and Neal AP. The application of linguistic processing to automatic abstract generation. Journal of Document and Text Management, 1993:215~241

[9] Salton G. Syntactic approaches to automatic book indexing. In: Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics, ACM, New York, 1988:120~138

[10] Krupka G. SRA: Description of the SRA system as used for MUC-6. In: Proceedings of the Sixth Message Understanding Conference, Morgan Kaufmann, and California. 1995

[11] Brandow R, Mitze K, and Rau LR. The automatic condensation of electronic publications by sentence selection. Information Processing and Management, 1995,31(5):675~685

[12] Jang D-H and Myaeng SH. Development of a document summarization system for effective information services. In: RIAO 97 Conference Proceedings: Computer-Assisted Information Searching on Internet, Montreal, Canada, 1997:101~111

[13] Fagan JL. Experiments in automatic phrase indexing for document retrieval: A comparison of syntactic and non-syntactic methods. PhD Dissertation, Cornell University, Department of Computer Science, 1987. Report #87-868, Ithaca, New

York.

[14] Salton G. Syntactic approaches to automatic book indexing. In: Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics, ACM, New York, 1988:120~138

[15] Ginsberg A. A unified approach to automatic indexing and information retrieval. IEEE Expert, 1993. 8:46~56

[16] Nakagawa H. Extraction of index words from manuals. In: RIAO 97 Conference Proceedings: Computer-Assisted Information Searching on Internet, Montreal, Canada, 1997:598~611

[17] Leung C-H, and Kan W-K. A statistical learning approach to automatic indexing of controlled index terms. Journal of the American Society for Information Science, 1997. 48(1):55~66

[18] Croft WB, Turtle H and Lewis D. The use of phrases and structured queries in information retrieval. In: SIGIR-91: Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, ACM, 1991:32~45

[19] Krulwich B and Burkey C. Learning user information interests through the extraction of semantically significant phrases. In: Hearst M and Hirsh H, Eds., AAAI 1996 Spring Symposium on Machine Learning in Information Access. AAAI Press, California. 1996

[20] MuQnoz A. Compound key word generation from document databases using a hierarchical clustering ART model. Intelligent Data Analysis, 1(1): Elsevier, Amsterdam. 1996

[21] Steier AM and Belew RK. Exporting phrases: A statistical analysis of topical language. In: R Casey and B Croft, Eds., Second Symposium on Document Analysis and Information Retrieval, 1993:179~190

[22] Turney P.D．Learning to extract keyphrases from text．National Research Council，Canada，NRC Technical Report ERB-1O57．1999

[23] Witten I.H.，Paynter G.W.，Frank E，Gutwin C. and Nevill C.G． KEA ：Practical automatic keyphrase extraction．In：Proceedings of the 4th ACM conference on Digital libraries，Berkeley，California，US，l999:254~256

[24] Yang Wen—Feng. Chinese keyword extraction based on max duplicated strings of the documents. In：Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval，Tampere，Finland，2002:439~440

[25] 李素建，王厚峰，余士汶，辛乘胜. 关键词自动标引的最大熵模型应用研究. 计算机学报，2004,27(9):1192~1197

[26]王军. 词表的自动丰富—从元数据中抽取关键词及其定位. 中文信息学报，2005,19(6):36~43

[27]R.Ferrer I, Cancho and R.V.Sole. The small world of human language. Proceedings of the Royal Society of London . 2001.2261~2266

[28]刘建毅，王箐华，王枞. 基于语言网络的关键词抽取. 第三届全国信息检索与内容安全学术会议. 苏州. 2007:161~168

[29]索红光，刘玉树，曹淑英. 一种基于词汇链的关键词抽取方法. 中文信息学报，2006,20(6):27~32

[30]石晶，胡明，戴国忠. 基于小世界模型的中文文本主题分析. 中文信息学报，2007,21(3):69~75

[31]Anette Helth. Combining machine learning and natural language processing for automatic keyword extraction. Stockholm: Department of computer and systems sciences, Stockholm University. 2004. 35~38

[32]刘远超，王晓龙，徐志明，刘秉权. 基于粗集理论的中文关键词短语构成规则挖掘. 电子学报，2007,35(2):371~374

[33]Hongqiao Li, Chang-Ning Huang, Jianfeng Gao and Xiaozhong Fan. The Use of SVM for Chinese New Word Identification. In: International Joint Conferenceon Natural Language Processing 2004, IJCNLP2004, Sapporo, Japan. 2004:723~732.

[34]Porter, M.F. An algorithm for suffix stripping. Program; Automated Library and Information Systems. 1980,14 (3) :130~137

[35]Lovins, J.B. Development of a stemming algorithm. Mechanical Translation and Computational Linguistics. 1968.11:22~31

[36]Krovetz, R. Viewing morphology as an inference process. In: Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'93, 191~203

[37]Kuo Zhang, Hui Xu, Jie Tang, Juanzi Li. Keyword Extraction Using Support Vector Machine. In: Advances in Web-Age Information Management, 7th International Conference, WAIM 2006, Hong Kong, China, 2006:85~96

[38]Boser B. E., Guyon I. M. and Vapnik. V. A Training Algorithm for Optimal Margin Classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. Pittsburgh, PA, USA. 1992:144~152

[39]Vapnik V. Statistical Learning Theory. New York:Wiley, 1998

[40]Cortes C., Vapnik. V. Support vector networks. Machine Learning. 1995, 20(3): 273~297

[41]吕玉生. Support Vector Machine 支持向量机 PDF file download at http://icst.nbu.edu.cn/ppt/SVM 课件

[42] Chang C. LIBSVM: a library for support vector machines, 2006. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[43] David, C., L. Giroux, S. Bertrand-Gastaldy, and D. Lanteigne. Indexing as problem solving: A cognitive approach to consistency. In Forging New Partnerships in Information, Medford, NJ, Information Today. 1995: 49~45

[44] Olena Medelyan and Ian H. Witten. Domain Independent Automatic Keyphrase Indexing with Small Training Sets. Department of Computer Science The University of Waikato. 2008

[45] Huang C., Tian Y., Zhou Z., Ling C. and Huang T. Keyphrase extraction using semantic networks structure analysis. In: Sixth IEEE International Conference on Data Mining, Hong Kong, China, 2006:275~284

[46] Kristina Toutanova and Christopher D. Manning. Stanford Tagger: Stanford Log-linear Part-Of-Speech Tagger, 2000. Software available at http://nlp.stanford.edu/software/tagger.shtml

[47] Frank E. KEA: Keyphrase Extraction Algorithm, 1999. Software available at http://www.nzdl.org/Kea/download. html

作者在学期间取得的学术成果

[1] 罗准辰，王挺. 基于分离模型的中文关键词抽取算法研究. 中文信息学报（已录用）

[2] 罗准辰，王挺. 关键词抽取中的分离模型和特征设计. 第四届全国信息检索与内容安全学术会议论文集. 中国，北京. 2008:65~74

[3] 刘伍颖，王挺，罗准辰. 面向多源垃圾信息过滤的直推式迁移学习算法. 2008 中国计算机大会论文集. 中国，西安. 2008（已录用）