一、简介

本课题为“文本分类：文献综述”，本文将运用信息检索课程的相关知识及技能，从研究课题的定义及介绍到课题资料收集整理到结论的整个的思维过程进行展示。本课题所使用的工具汇总，见表1-1。

|  |  |  |
| --- | --- | --- |
| 表1-1 研究工具汇总 | | |
| 工具名称 | 主要作用 | 网址 |
| 百度 | 获取领域常识 | <https://www.baidu.com/> |
| 知乎 | 获取小白级研究经验及领域知识 | <https://www.zhihu.com/> |
| mindmaster | 制作思维导图 | <https://www.edrawsoft.cn/mindmaster-free/> |
| CSDN | 获取计算机领域的经验和相关领域知识 | <https://www.csdn.net/> |
| 知网 | 查找学术文献 | <https://kns.cnki.net/kns/brief/default_result.aspx> |
| google scholar | 查找学术文献 | <http://scholar.hedasudi.com/> |
| wos | 学术文献 | <https://apps.webofknowledge.com/> |
| EndNote | 管理文献 | <https://endnote.com/> |
| github | 管理并保存文献 | <https://github.com/> |
| ACM | 计算机领域会议期刊及交流合作 | [https://www.acm.org/search#stq=deep+learning&stp=4](https://www.acm.org/search" \l "stq=deep+learning&stp=4) |
| IEEE | 学会的主要活动是召开会议、出版期刊杂志、制定标准、[继续教育](https://www.baidu.com/s?wd=%E7%BB%A7%E7%BB%AD%E6%95%99%E8%82%B2&tn=SE_PcZhidaonwhc_ngpagmjz&rsv_dl=gh_pc_zhidao" \t "https://zhidao.baidu.com/question/_blank)、颁发奖项、认证（Accreditation） | <https://ieeexplore.ieee.org/Xplore/home.jsp> |

1.1文本分类

文本自动分类简称文本分类（text categorization）,是模式识别与自然语言处理的交叉研究课题。文本分类是在预定义的分类体系下，根据文本的特征（内容或属性），将给定文本与一个或多个类别相关联的过程。文本分类实际上就是将文档通过分类器分别映射到相应的类别列表中。图1-1是文本分类示意图。

|  |
| --- |
|  |
| 图1-1 文本分类系统 |

文本分类的文献综述是介绍文本分类早期研究，重点介绍文本分类各个历史阶段（表1-2）突破性的研究成果及主要研究方法，如机器学习方法、深度学习方法等。最后根据学术前沿展望未来研究方向并得出结论。

|  |  |
| --- | --- |
| 表1-2 文本分类研究阶段 | |
| 研究阶段（时间） | 主要方向 |
| 第一阶段（1958-1964） | 进行自动分类可行性研究 |
| 第二阶段（1965-1974） | 进行自动分类的实验研究 |
| 第三阶段（1975-1989） | 进入实用化阶段 |
| 第四阶段（1990-至今） | 面向互联网的文本自动分类研究 |

1.2文献综述

文献综述实际上就是总结前人工作，了解当前研究水平，分析问题及可能的原因，并进一步探究本领域前沿动态、发展方向，提出自己的观点看法。具体操作路线，见图1-2。

|  |
| --- |
|  |
| 图1-2 文献综述技术路线 |

二、研究框架及信息梳理

2.1思维导图

撰写文本分类的文献综述，首先得明确文本分类整个领域的知识体系及研究脉络，故绘制文本分类的思维导图，见图2-1、图2-2，其中图2-2是在阅读文献后对图2-1的细化与改进。

|  |
| --- |
| 文本分类 |
| 图2-1文本分类思维导图 |
| 深度学习 |
| 图2-2 文本分类思维导图细化及改进 |

2.2文献列表

综述：

Text feature extraction based on deep learning: a review 2017

（引用36次）

基于朴素贝叶斯的文本分类研究综述\_贺鸣 2016

文本分类研究综述 2019 汪岿 刘柏嵩

半监督文本分类综述 2011

中文文本分类方法综述 2019

文本相似度计算方法研究综述 2019

自然语言处理发展及应用综述\_赵京胜 2019

最早自动分类：On relevance, probabilistic indexing and information retrieval(1960)

经典博士论文：Representation and Learning in Information Retrieval（1992）

机器学习系列：

特征选择比较：A comparative study on feature selection in text categorization（1997）

A comparison of event models for naive bayes text classification（1998）

Transductive inference for text classification using support vector machines（1999）

Using maximum entropy for text classification（1999）

Text classification from labeled and unlabeled documents using EM（2000）

贝叶斯反垃圾邮件An evaluation of naive bayesian anti-spam filtering（2000）

Support vector machine active learning with applications to text classification（2001）

Text Classification using String Kernels 2002 （引用次数1627）

Text classification using string kernels（2002）

特征选择：An extensive empirical study of feature selection metrics for text classification（2003）

基于图的：Graph-based text classification: learn from your neighbors（2006）

短文本分类：Short text classification in twitter to improve information filtering（2010）

单词共现：Word co-occurrence features for text classification（2011）

Baseline Needs More Love:On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms

Using TF-IDF to Determine Word Relevance in Document Queries

Chinese Segmentation and New Word Detection using Conditional Random Fields： 2004（引用496次）

A novel probabilistic feature selection method for text classification（2012）

Learning to Classify Text from Labeled and Unlabeled Documents（1998）

A Tutorial on Automated Text Categorisation（1999）

A comparative study of TF\*IDF, LSI and multi-words for text classification 2011（引用399次）

Feature Selection and Feature Extract ion for Text Categorization 1992（引用次数621）

深度学习系列：

Text feature extraction based on deep learning: a review 2017（引用36次）

Adversarial Multi-task Learning for Text Classification（2017）

Deep Learning for Chinese Word Segmentation and POS Tagging

2013（引用226）

Relation classification via convolutional deep neural network（2014）

A method for chinese text classification based on apparent semantics and latent aspects（2015）

基于Word2vec和改进型TF-IDF的卷积神经网络文本分类模型（2019）

文本分类TF-IDF算法的改进研究（2019）

TextCNN：Convolutional Neural Networks for Sentence Classification(EMNLP 2014)

GRU：Document Modeling with Gated Recurrent Neural Network for Sentiment Classification (EMNLP 2015)

RCNN：Recurrent Convolutional Neural Networks for Text Classification(AAAI 2015)

char-CNN，没有使用词向量：Character-level Convolutional Networks for Text Classification（NIPS 2015）

Recurrent Neural Network for Text Classification with Multi-Task Learning (IJCAI 2016)

HAN：Hierarchical Attention Networks for Document Classification (NAACL 2016)

fastText：Bag of Tricks for Efficient Text Classification (EACL 2017)

DPCNN：Deep Pyramid Convolutional Neural Networks for Text Categorization (ACL 2017)

DeepMoji：Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm (EMNLP 2017)

标签量非常多Deep learning for extreme multi-label text classification（2017）

胶囊网络：Investigating Capsule Networks with Dynamic Routing for Text Classification (EMNLP 2018)

RNN-Capsule：Sentiment Analysis by Capsules (WWW 2018)

UMFiT迁移学习：Universal Language Model Fine-tuning for Text Classification（2018）

GCN：Graph Convolutional Networks for Text Classification (AAAI 2019)

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (NAACL 2019)

三、查找并获取信息

3.1知网

关键词检索







3.2 google scholar

直接搜索文章

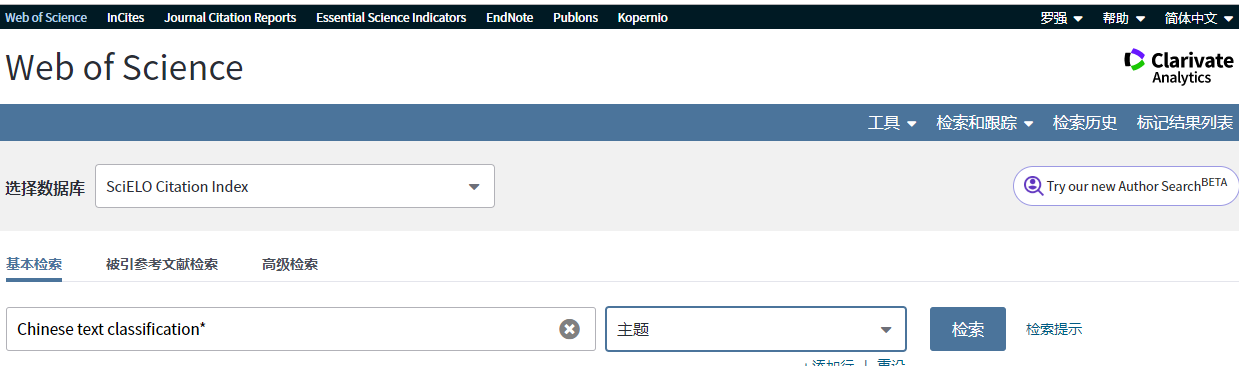


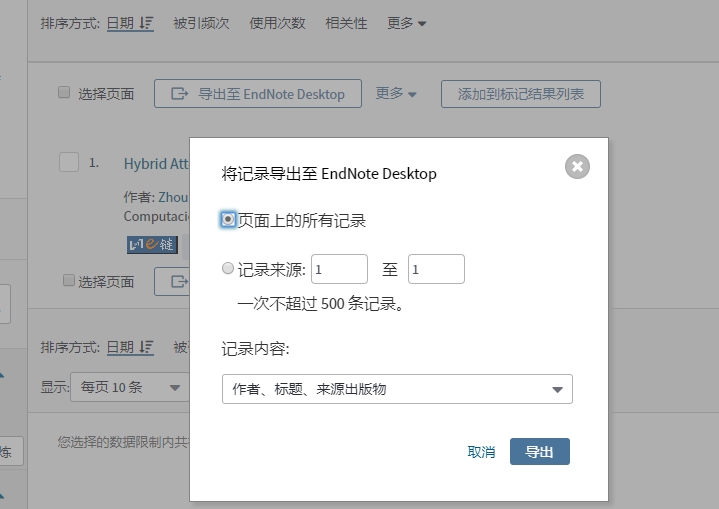




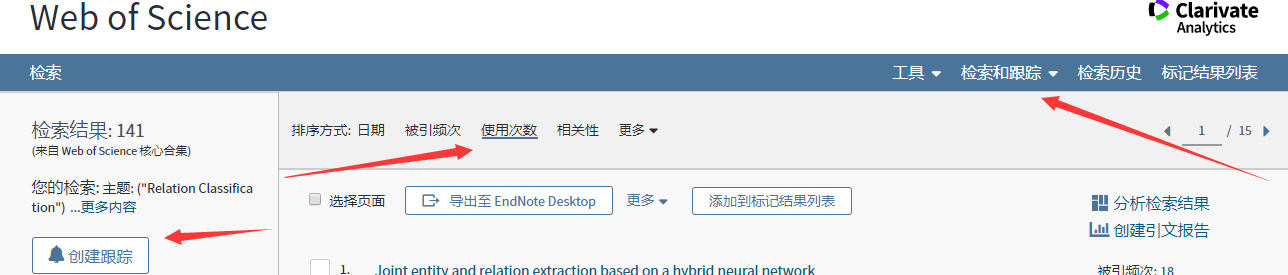
3.3 WOS

①基于更多限定的精确查找





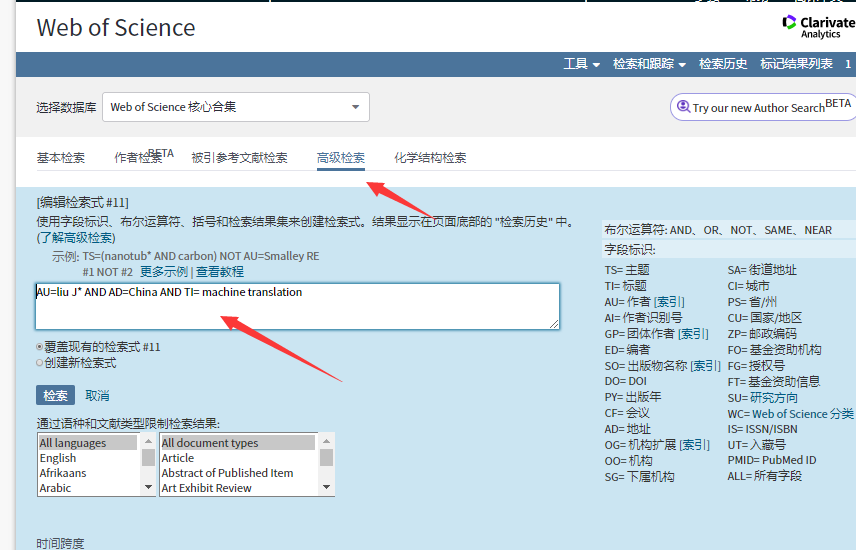
②根据需求设置相关检索限定，找到合适的文献



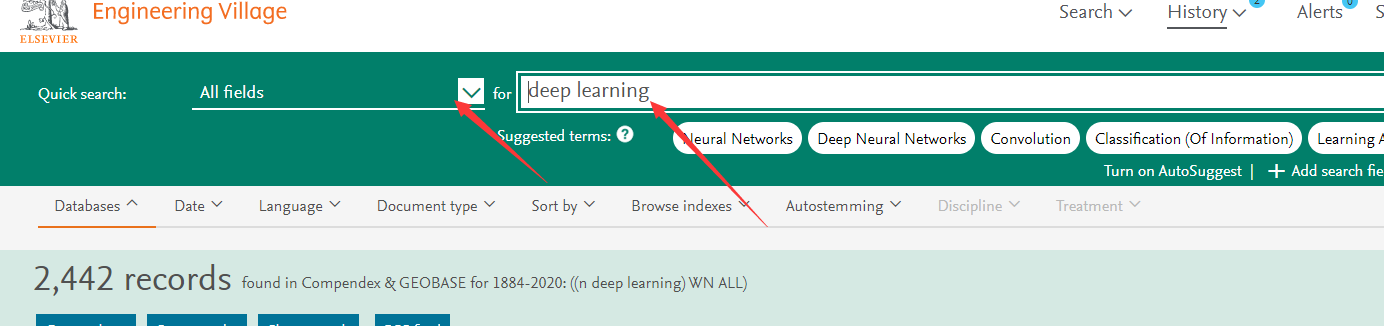
更多检索词的索引方式

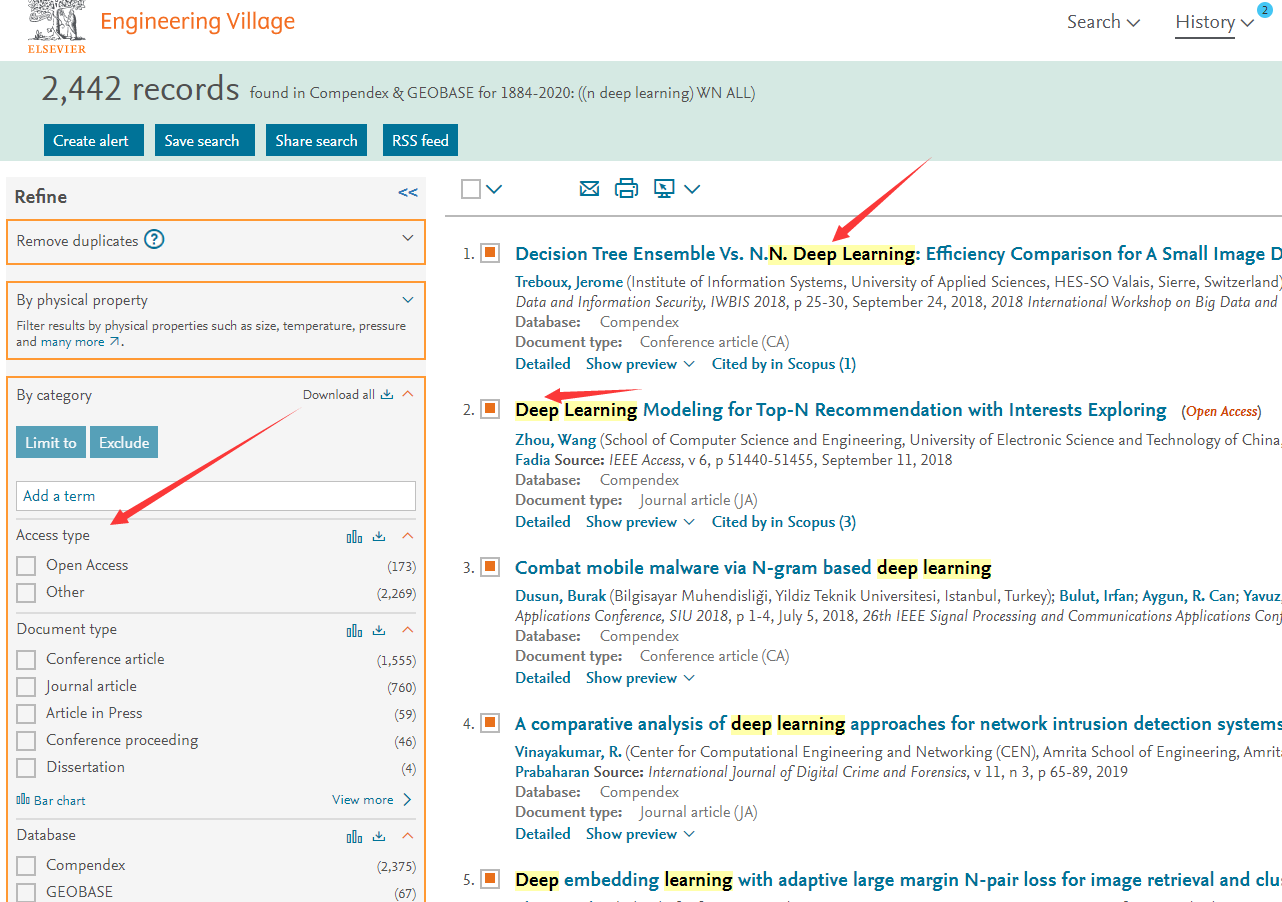


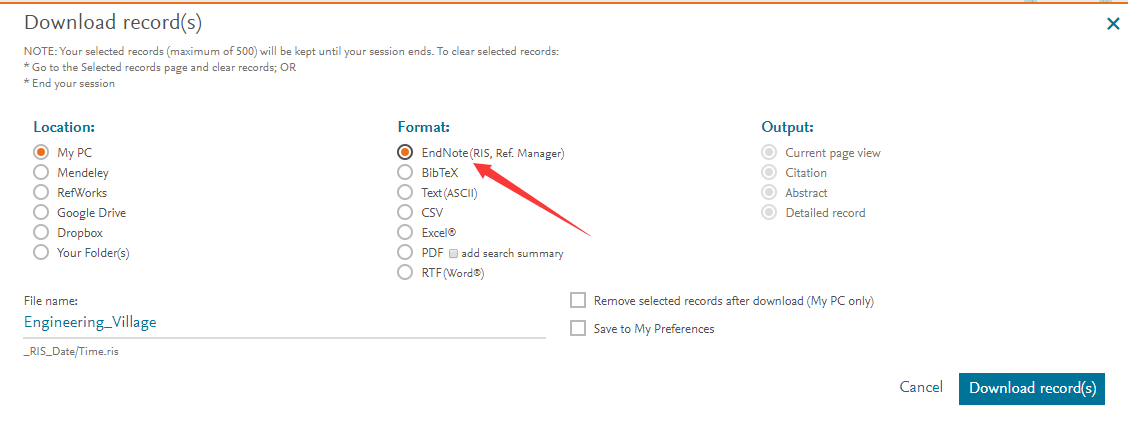
③高级检索



3.4 Engineering Village

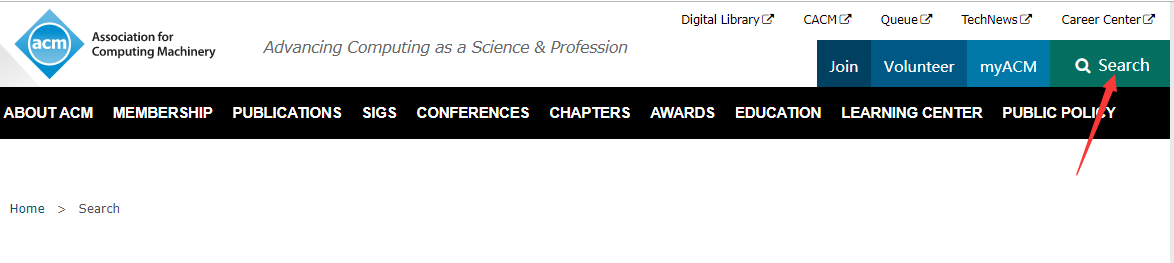


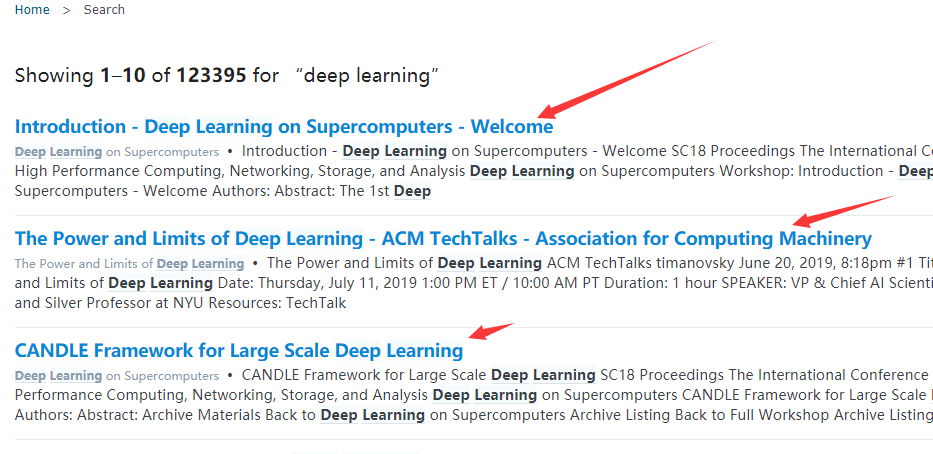




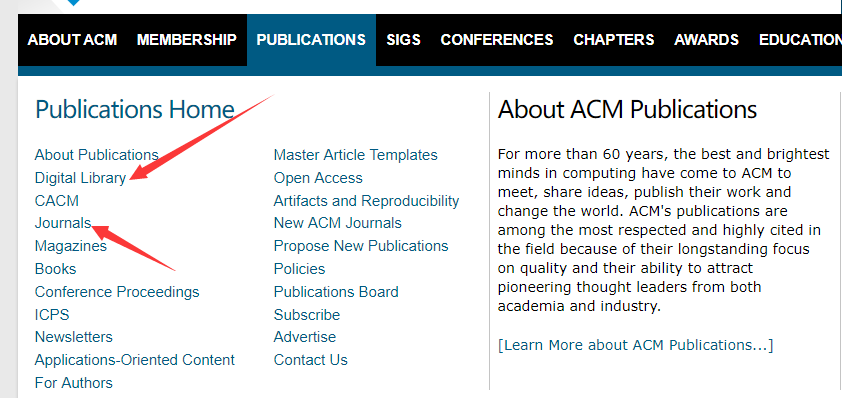
3.5 ACM

①查找会议及领域最新动态

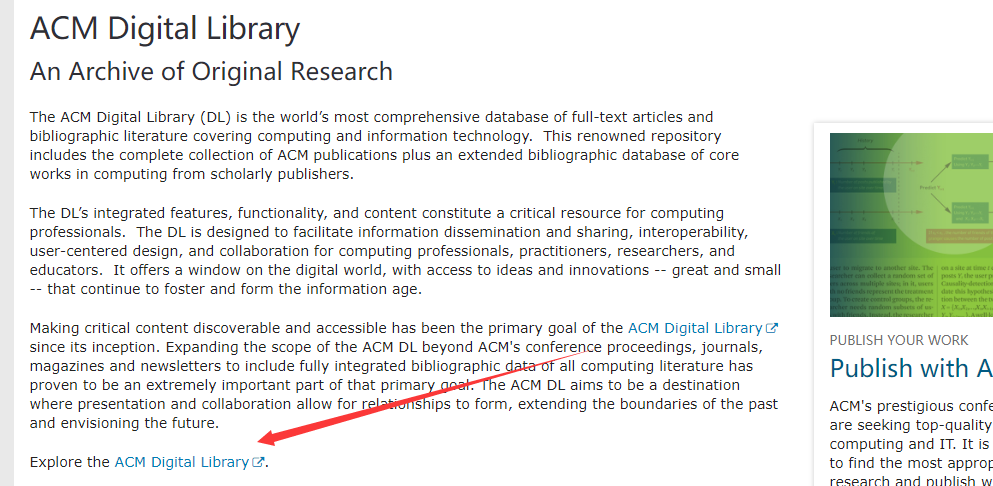




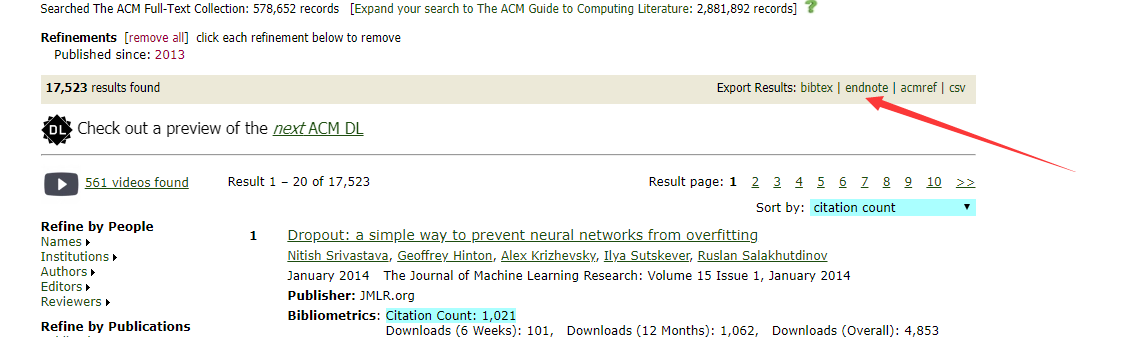
②数据库与期刊



③进入数据库



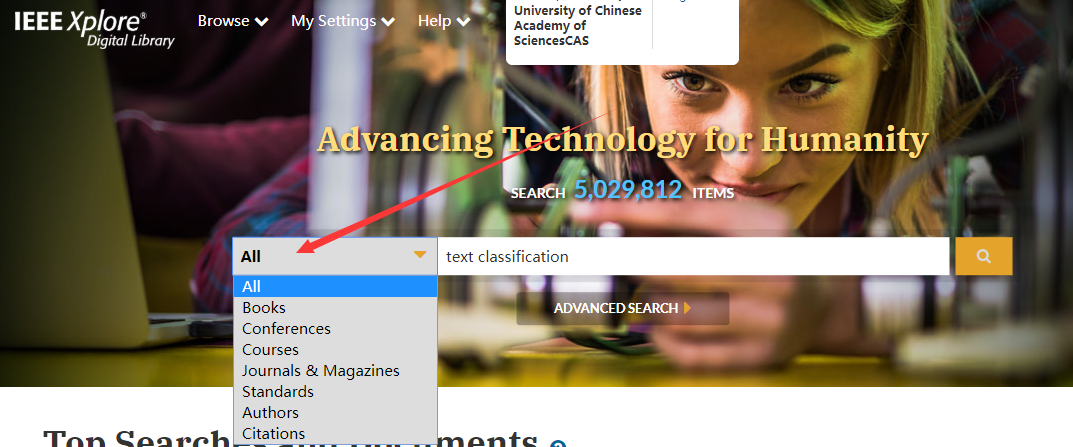
④导入endnote



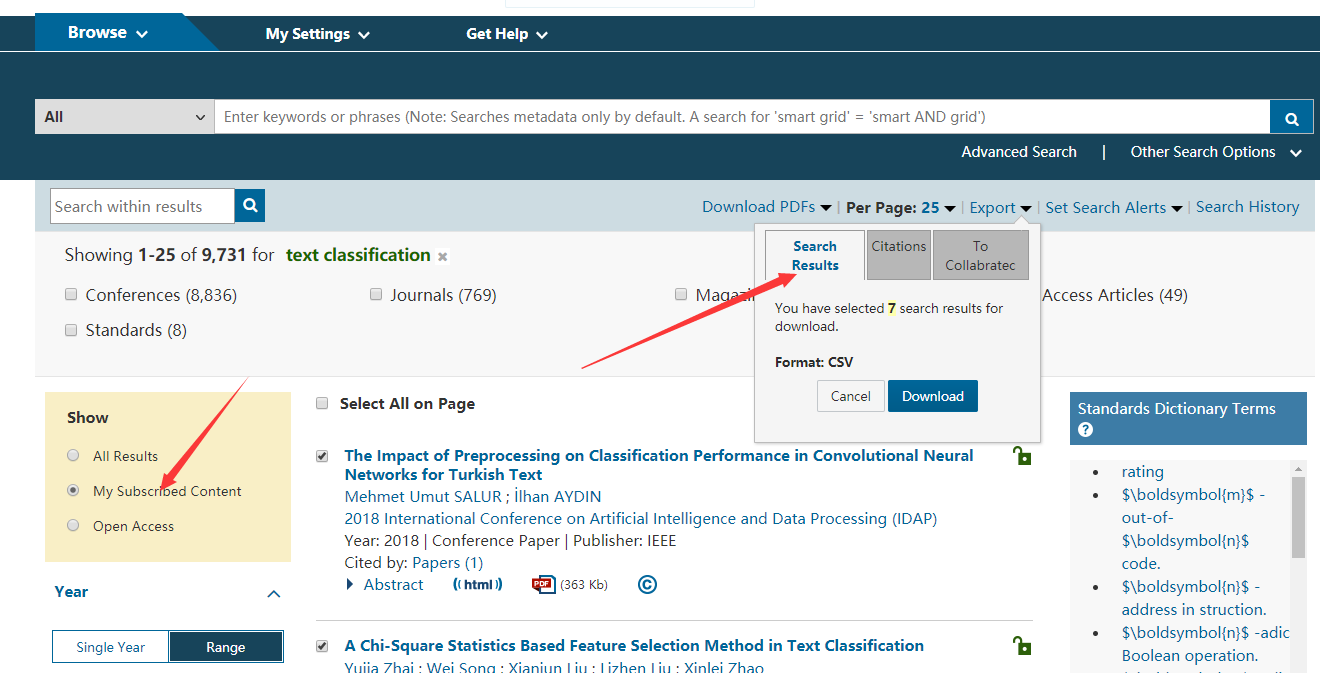
3.6 IEEE

学会的主要活动是召开会议、出版期刊杂志、制定标准、[继续教育](https://www.baidu.com/s?wd=%E7%BB%A7%E7%BB%AD%E6%95%99%E8%82%B2&tn=SE_PcZhidaonwhc_ngpagmjz&rsv_dl=gh_pc_zhidao" \t "https://zhidao.baidu.com/question/_blank)、颁发奖项、认证（Accreditation）等。

①检索



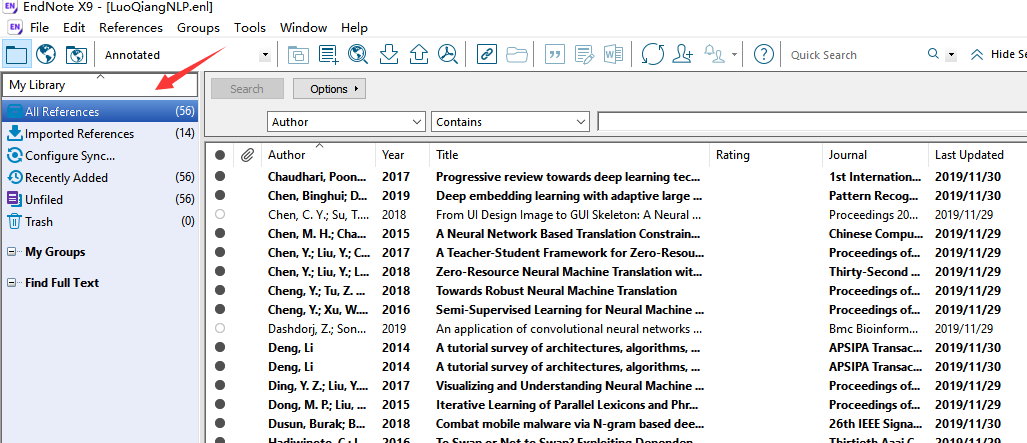
②导出



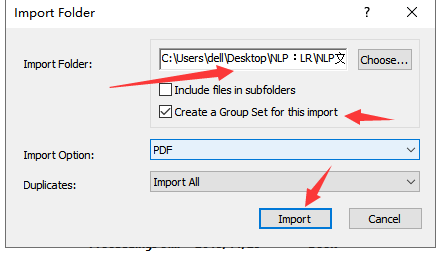
四、管理信息

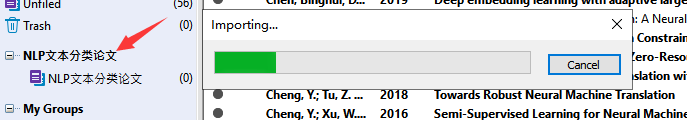
4.1EndNote

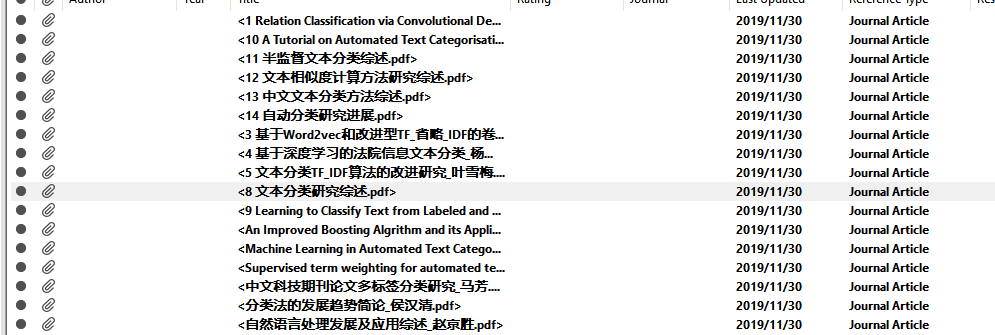
①直接网页或数据库导入



②下载后导入文件夹







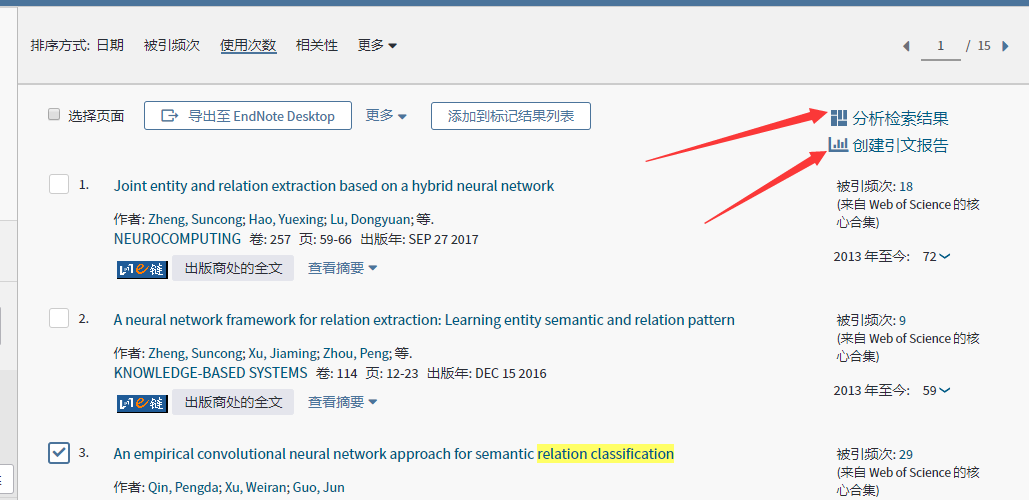
4.2github

动态保存文件，方便共享与各个修改过程的跟踪

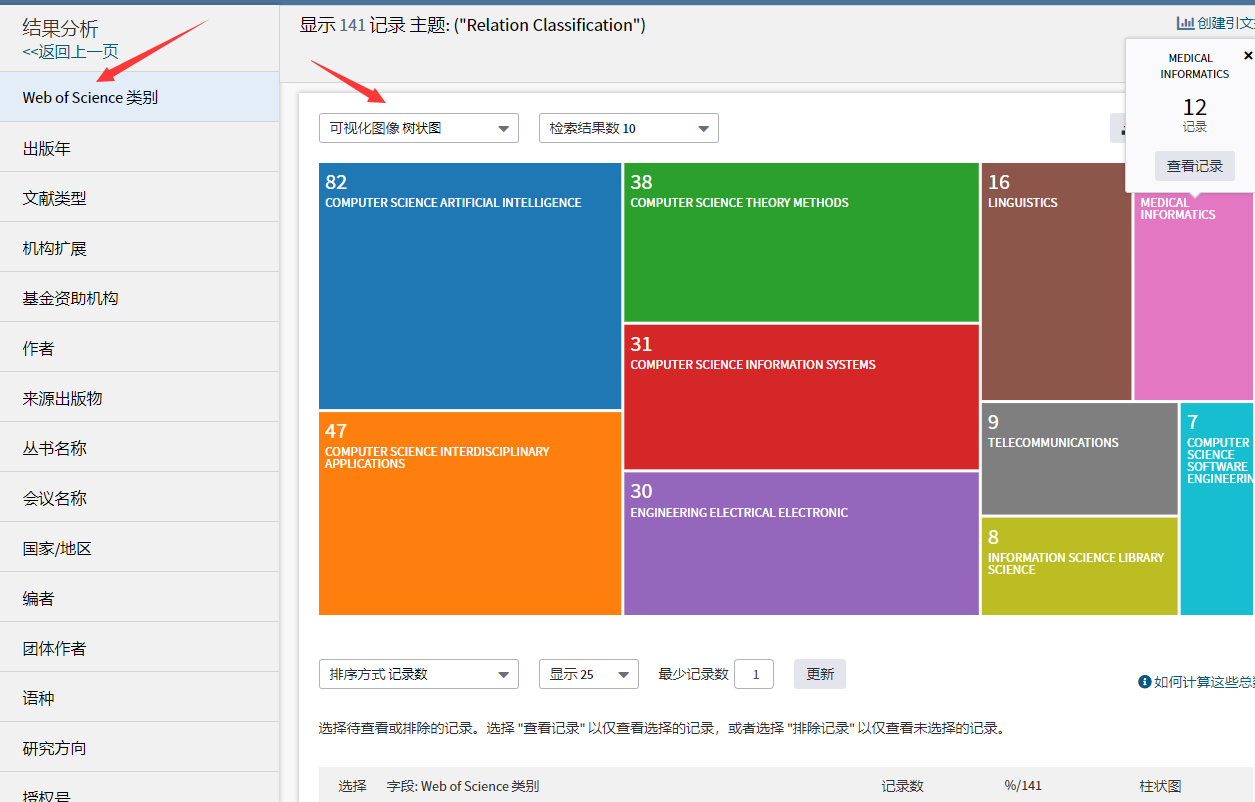
五、分析及利用信息

5.1 利用wos分析引文的相关信息

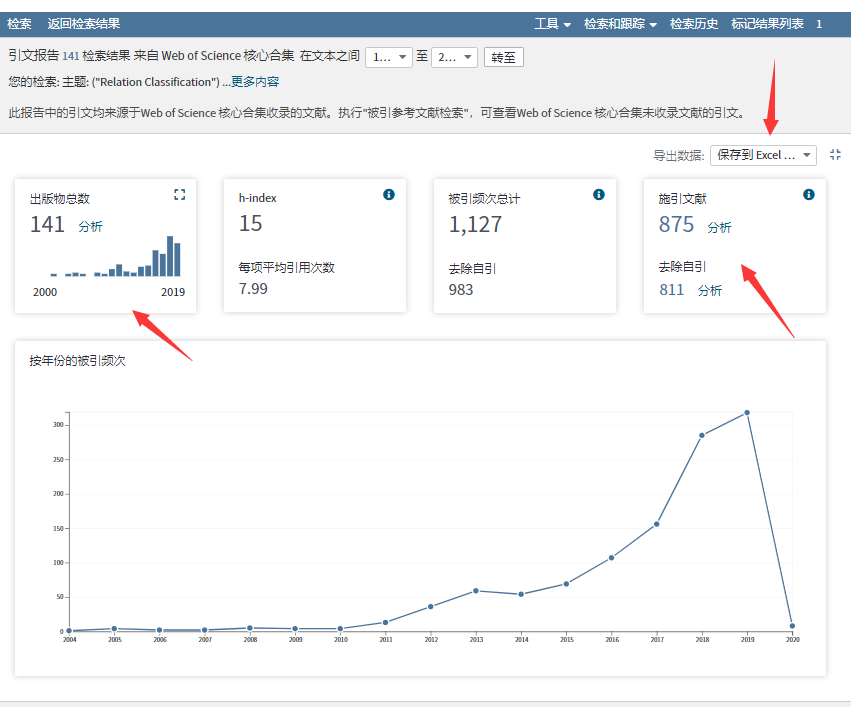
①进入检索文献目录



②分析检索结果

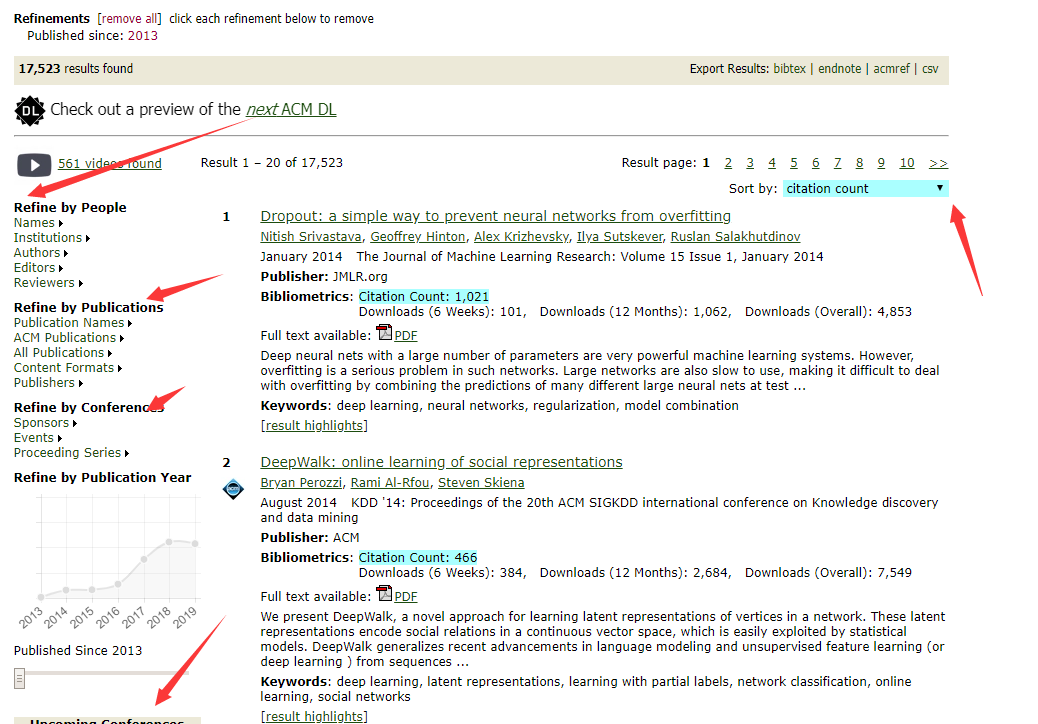


③创建引文报告



5.2ACM

①进行人物、刊物、会议、引用量等的分析



一、研究课题：

文本分类：文献综述

二、课题研究框架及所需信息

形式：用文字、图表、思维导图等

三、查找并获取信息

使用哪些方法、资源及工具

四、管理信息

使用哪些方法、资源及工具

五、分析及利用信息

使用哪些方法、资源及工具