中国科学技术信息研究所
论文集（2011）

（上册）

中国科学技术信息研究所　编

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

中国科学技术信息研究所
论 文 集

（2011）

（上 册）

中国科学技术信息研究所　编

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

图书在版编目(CIP)数据

中国科学技术信息研究所论文集.2011/中国科学技术信息研究所编.—北京:科学技术文献出版社,2013.1
ISBN 978-7-5023-7617-8

Ⅰ.①中… Ⅱ.①中… Ⅲ.①科技情报-情报工作-中国-文集 ②信息技术-中国-文集 Ⅳ.①G359.2-53 ②G202-53

中国版本图书馆CIP数据核字(2012)第243064号

中国科学技术信息研究所论文集(2011)

策划编辑:科 文 责任编辑:任昱仰 责任校对:赵文珍 责任出版:张志平

出版者	科学技术文献出版社
地址	北京市复兴路15号 邮编100038
编务部	(010)58882938,58882087(传真)
发行部	(010)58882868,58882866(传真)
邮购部	(010)58882873
官方网址	http://www.stdp.com.cn
淘宝旗舰店	http://stbook.taobao.com
发行者	科学技术文献出版社发行 全国各地新华书店经销
印刷者	北京时尚印佳彩色印刷有限公司
版次	2013年1月第1版 2013年1月第1次印刷
开本	889×1194 1/16开
字数	2656千
印张	89
书号	ISBN 978-7-5023-7617-8
定价	288.00元(上下册)

前　言

中国科学技术信息研究所（简称“中信所”）是科技部直属的国家级公益性科技信息研究机构，定位于为科技部等政府管理部门提供决策支持，为科技创新主体提供全方位的信息服务，努力建设成为全国科技情报领域的共享管理与服务中心、学术中心、人才培养中心和网络技术研究推广中心，在全国科技信息系统中发挥指导和示范作用。

按照国家公益类科研院所的改革要求，中信所于2005年实施了深化改革工作。七年多来，中信所党政领导班子对科技信息工作的发展规律进行了深入思考，结合自身实际情况及相关实践探索，对中信所的改革与发展工作做出了中长期布局与统筹安排。确定了“落实公益定位，坚持科学管理，服务自主创新，建设一流院所”的工作主线。提出了“三步走”的发展思路，即第一步（2006—2008年）是中信所发展的磨合期，其工作重点是新机构、新业务、新机制之间的调整与完善，确立中长期发展规划；第二步（2009—2011年）是中信所发展的整合期，其工作重点是整合与提升全所优质资源和业务工作，进一步夯实发展基础；第三步（2012—以后）是中信所发展的跃升期，在这个阶段，全所的公益研究与服务水平将稳步提升，并形成具有特色的研究与服务产品。

七年多来，中信所启动、开展了一系列基础性、前瞻性工作，为科研人员营造了良好的学术研究环境与学术交流氛围，为科研人员提高学术研究能力、提升学术论文水平奠定了坚实的基础。同时，积极推进资源整合工作，加强了“中国科学技术信息研究所暨国家工程技术数字图书馆网站”、“中国科技情报网”、“中国科技论文统计与分析平台”、“ISTIC专利信息检索与分析平台”和“中国高层次科技人才信息网”等业务支撑平台建设；提出了以事实型数据为基础，综合集成“事实型数据+专用方法工具+专家智慧”的研究方法，明确了中信所必须坚持以事实型数据资源建设为基础，坚持科技情报研究工作的可定量化、科技情报研究成果可重现性的业务发展新思路；全面启动了中国科技论文与引文数据库、ISTIC-中国发明专利数据库和中国高层次科技人才数据库等事实型数据库的建设和质量提升工作，加强了知识挖掘、机器翻译、信息可视化等与科技情报研究工作紧密相关的软件工具和模型的研发工作。七年多来，中信所坚持开放办所，推动开放联合。加强了与国内重点高校的合作，与南京大学、武汉大学、吉林大学等高校签订了合作协议，开展学术研究、人才培养等方面的深度合作；加强了与国际知名同行机构的合作，与汤森路透科技信息集团合作建立

了“ISTIC－THOMSON REUTERS 科学计量学联合实验室”，与美国千年研究所合作成立了“ISTIC－MI 联合研究中心”，与爱思唯尔集团联合创建了“ISTIC－ELSEVIER 期刊评价研究中心”，与德国弗朗霍夫系统与创新研究所（ISI）建立了合作关系，与日本科技振兴机构和韩国科技情报研究所保持着制度化的合作交流。七年多来，中信所持续投入、稳定支持科研人员在图书馆学、未来学、科技政策与管理、信息资源管理、知识工程、自然语言处理、情报学、科学计量学8个重点学科方向上开展学术探索和研究工作。

经过七年多的不懈努力，中信所学术论文产出已经从数量的快速增长期进入到高产状态下的相对稳定期，收录在国际著名检索数据库中的高质量论文呈现大幅增长态势，已经呈现出从数量增长向质量提升的重大转变。这些成绩的取得，归功于科技部的正确领导，归功于全所上下的共同努力和协同配合，归功于我所广大科技信息工作者的拚搏奉献与忘我工作。

现代管理学之父彼得·德鲁克曾经指出，当今世界已由管理的社会变成了创新的社会，当今社会最具有价值的活动无疑是寻找创新的来源。实践证明，创新是社会发展的前提，是社会进步的强大动力。科技创新的灵魂，就在于开放和交流。科技论文是记录、总结科研成果的重要方式，也是科研人员交流学术思想和科研成果，促进科技知识传播，体现科研工作价值和重要贡献的重要载体。科研院所更是只有通过合作和交流才能实现学术研究的互通有无、取长补短，建立良好的科研文化，提高科研效率，融入时代科技发展的潮流。

本论文集编撰收集了中信所科研人员（含硕博研究生、博士后科研人员）2009—2011年以中信所为第一作者单位发表的中文核心期刊论文以及被 SCI、SSCI、EI、ISTP、ISSHP 收录的论文共753篇，各年度论文集均分为上、下两册，共6册，内容主要涉及图书情报研究、战略研究以及其他相关研究等方面，从一个侧面展示了中信所在图书情报基础理论、方法和技术以及科技政策与战略决策、领域分析与研究等学术领域的研究成果。我们衷心希望，这样一部记录我国科技信息事业发展轨迹的重要文集能够在我国科技信息事业发展史上增添浓墨重彩的一笔。

由于编写时间仓促，疏漏和不妥之处在所难免，敬请读者斧正。

中国科学技术信息研究所　所长

贺德方

2013年1月

目 录

上 册

图情研究

A Novel Category Vector - Based Cross Language Text Categorization Method
…… Yingfan Gao Hongjiao Xu Wei Yu Huilin Wang (3)

A Semantic Analyzing Method in the Field of Technological Literature
…… Yao Liu Xuefei Chen Sujian Li Ruijia Wang (13)

Research on Semantic Method of Library Resources' Organizing
…… Yao Liu Xuefei Chen Zhifang Sui Yongwei Hu Qingliang Zhao (19)

Research on the Construction and Assessment of the Technical Control Models in Network Information Communication …… Yao Liu (26)

On Space Analysis and Method Based NSTL for the Original
…… Xiaodong Qiao Changqing Yao Lianjun Zhang Chunyun Hao (32)

PrestigeRank: A New Evaluation Method for Papers and Journals
…… Su Cheng Pan Yuntao Zheng Yanning Ma Zheng Yuan Junpeng Guo Hong Yu Zhenglu Ma Caifeng Wu Yishan (38)

Exploration and Study of Chinese Thesaurus Automation Construction for Digital Libraries …… Wen Zeng (53)

A Gravitation - based Position Model for Visualizing and Clustering Vectored Resources …… Junsheng Zhang Yunchuan Sun Huilin Wang (62)

Calculating Statistical Similarity Between Sentences
…… Junsheng Zhang Yunchuan Sun Huilin Wang Yanqing He (73)

A New Dynamic Method of Machine Learning From Transition Examples …… Xiaodan Zhang (84)

Study on WEB Page Fusion Classification Model …… Xiaodan Zhang Lijun Zhu (89)

A Brief Statistical Analysis of Stomatology Papers Published in International Journals by Researchers From the Mainland of China——Based on Thomson Reuters' Journal Citation Reports for 2009 …… Yuhua Zhang Yuntao Pan Xiao Hong (92)

Chinese Scientific and Technical Information Institutions: Development and Perspective
…… Zheng Yanning (99)

Scientometric Analysis of Physics (1979—2008): A Quantitative Description of Scientific Impact
…… Zheng Yanning Yuan Junpeng Pan Yuntao Zhao Xiaoyuan (114)

Fractional Counting of Citations in Research Evaluation: A Cross - and Interdisciplinary Assessment of the Tsinghua University in Beijing …… Ping Zhou Loet Leydesdorff (122)

Analysis and Evaluation of Chinese Scientific and Technical Vocabulary System –
The New Energy Car Vocabulary System as an Example
…………………… Lijun Zhu Changai Hu Xiaodong Qiao Yunliang Zhang (136)

生物技术相关专利的计量学研究 …………………… 艾志昂 罗 勇 傅俊英 (142)

利用D2R实现关系数据库与关联数据的语义模式映射 …………………… 白海燕 梁 冰 (152)

网络环境叙词表收词新来源刍议 …………………… 鲍秀林 吴雯娜 (161)

数字环境下通用概念获取方法 …………………… 常 春 赖院根 (167)

叙词表无关联词分析与处理 …………………… 常 春 王 星 (173)

专业概念机器辅助分类方法研究 …………………… 常 春 赖院根 (178)

北京市服务外包企业开拓国际市场存在的竞争情报问题及对策建议 …… 陈 峰 赵筱媛 郑彦宁 (185)

竞争情报最佳实践企业的非量化评价方法 …………………… 陈 峰 (190)

科技档案资源对科技创新型企业的竞争情报功能 …………………… 陈 峰 (195)

企业竞争情报工作的评价方法 …………………… 陈 峰 (198)

战略性新兴产业竞争情报需求问题及其求解 …………………… 陈 峰 (203)

国内外属性抽取研究综述 …………………… 丁君军 郑彦宁 化柏林 (209)

基于规则的学术概念属性抽取 …………………… 丁君军 郑彦宁 化柏林 (214)

印度国家软件与服务公司行业协会开展产业竞争情报工作的做法及启示 ………… 丁 培 陈 峰 (221)

我国科技情报科研档案开发利用的思考 …………………… 高凤云 (225)

从哈勃太空望远镜的论文产出看大型科学仪器对科学研究的支撑作用 …………… 高晓培 潘云涛 (230)

查询分类在跨语言检索中的应用研究 …………………… 高影繁 王惠临 徐红姣 (237)

基于跨语言文本分类的多语资源组织方法研究 ………………… 高影繁 徐红姣 于 薇 王惠临 (243)

基于跨语言文本分类的跨语言特征提取方法研究 …………………… 高影繁 王惠临 徐红姣 (248)

基于平行语料的查询翻译词典改进方法研究 …………………… 高影繁 徐红姣 王惠临 (255)

广东省科技论文统计分析 …………………… 郭 红 潘云涛 苏 成 马 峥 (260)

国家自然科学基金资助产出论文计量分析
…………………… 郭 红 潘云涛 马 峥 苏 成 俞征鹿 徐 波 (269)

与国际出版集团合作的中国SCI期刊出版状况分析 …………………… 郭 玉 马 峥 (278)

中国、印度和巴西2006—2010年食品科学领域国际论文的比较研究
…………………… 郭 玉 马 峥 王晓春 (282)

机器翻译系统融合方法及其应用探究 ………………… 何彦青 石崇德 于 薇 张均胜 王惠临 (288)

基于词与短语的多机器翻译系统融合方法研究 …………………… 何彦青 张均胜 王惠临 (295)

回顾与前瞻：写在中国科技信息事业创建55周年之际 …………………… 贺德方 (301)

中国高影响力论文产出状况的国际比较研究 …………………… 贺德方 (305)

图书馆学与情报学研究生职业核心能力构建 …………………… 贺德方 (311)

汉语科技词系统网络特性分析 …………………… 胡长爱 朱礼军 (313)

《信息分析的核心》之信息分析 …………………… 化柏林 (323)

针对学术定义的抽取规则构建方法研究 …………………… 化柏林 刘一宁 郑彦宁 (327)

基于共建共享模式的知识服务系统建设研究——以“中国科技情报网”为例
…………………… 黄东流 张 旭 刘 娅 (334)

中英高校知识产权政策与技术转让比较研究 …………………… 江 山 张杰军 赵 捷 (341)

基于客户价值的信息用户流失预测研究 …………………… 赖院根 刘砺利 (347)

基于生存分析的信息用户流失研究与实证 …………………… 赖院根 刘砺利 (352)

期刊论文与专利文献的链接研究 …………………… 赖院根 (358)

网络环境下的信息用户分析体系研究 …… 赖院根　周杰（365）

网络环境下的信息用户流失分析 …… 赖院根　刘敏健　王星（371）

信息用户生存时间分布研究：以 NSTL 为例 …… 赖院根　曾建勋　王星（378）

面向对象数据库中的频繁模式发现 …… 李琳娜　张志平（384）

基于复杂结构数据聚类的推荐系统 …… 李琳娜　张志平（392）

多国 OA 期刊资源整合平台的构建及其权益管理功能的引入
——基于下一代 NSTL 体系的国际合作 …… 李颖　梁冰　乔晓东（398）

日本科学技术振兴机构纵览及其科技信息流通促进体系 …… 李颖　乔晓东　深泽信之（404）

NSTL 网络服务系统新进展及下一步建设计划 …… 梁冰　乔晓东　李颖　郝春云（411）

电子出版物呈缴系统信息资源长期保存标准化框架研究 …… 刘春燕　张爱霞　汪东波　沈玉兰（415）

数字环境下国际信息资源标识的新进展 …… 刘春燕　沈玉兰　张薇　李再扬（421）

国内外知识组织体系标准的现状及发展趋势研究 …… 刘华（428）

基于文献计量的国内外信息与文献国家标准对比研究 …… 刘华（435）

院士信息资源共享服务系统门户网站建设的思考与实践 …… 刘鹏飞　尹盛鑫（441）

网络环境下叙词表编制系统中的并发机制探讨 …… 刘伟　周杰（446）

2005—2008 年我国在社会发展主要研究领域的 SCI 论文情况分析 …… 刘娅　付晟鑫　许端阳（451）

从国际科技合著论文状况看中国环境领域国际科技合作态势 …… 刘娅（461）

我国科技信息机构知识产权信息服务工作现状调查与分析 …… 刘娅　雷孝平　黄东流（474）

学术定义抽取系统实现及实验分析 …… 刘一宁　郑彦宁　化柏林（479）

基于网络关系整合的竞争态势分析方法 …… 刘志辉　赵筱媛　杨阳（486）

研究专业演化图谱及其应用研究 …… 刘志辉　郑彦宁（492）

后金融危机时代中国科技期刊出版新特点及竞争策略分析 …… 马兰（501）

换个角度看中国科技期刊出版体制改革 …… 潘云涛　武夷山（504）

转型期我国科技期刊的发展路径 …… 潘云涛（507）

基于互联网的科技项目基本信息整合建库及应用初探 …… 彭洁　赵辉　刘润达　齐娜（511）

NSTL 的关联数据构建与应用场景设想 …… 乔晓东　白海燕　梁冰（517）

地方科技文献共享平台数据库建设情况分析 …… 屈宝强　吴家喜　赵伟　刘圣君（525）

复杂网络中重叠社团发现的算法探讨 …… 屈丽娟　屈宝强　王玙（533）

医学学术文献资源服务网站评价指标体系的用户调查研究 …… 阮以旻（541）

基于网络信息共享的信息质量管理的认识 …… 宋立荣（547）

我国科技信息资源共享中信息质量管理机制探讨 …… 宋立荣　李思经　赵伟（553）

基于语义网络的中文词汇链构造方法 …… 宋培彦　杨代庆（561）

用户交互式叙词表更新机制研究 …… 宋培彦　张峻铭（567）

全媒体时代广播资源整合与共享刍议 …… 谈黎红　邹晶（572）

广东省科技期刊评价指标分析 …… 王菁婷　马峥　刘清海　潘云涛（575）

首都医科大学主办医学期刊文献计量学分析 …… 王丽　马峥　张俊敏　潘云涛（586）

学术期刊的引用认同分析——以 Scientometrics 为例 …… 王丽　袁军鹏　王娜　潘云涛（594）

基于 Wiki 技术的标准术语库的设计与实现 …… 王莉　梁冰　郝春云　朱晓华　白海燕（602）

我国高层次科技人才信息资源建设现状调查与分析
…… 王运红　吴晓莉　彭洁　赵伟　吴家喜（610）

自然科技资源信息共享与实物服务的互动过程模型研究 …… 王运红　彭洁　吴霞　赵伟（616）

基于 HTTP 协议的 OA 期刊元数据动态收割研究 …… 吴广印　苏学　甘大广（622）

期刊论文元素标引规范研究 …… 吴琳　程煜华（627）

基于 DDC 的《英文超级科技词表》范畴体系构建研究
——以工程技术为例 …………………………………………………………………… 吴雯娜　王　星(632)
仅根据 Proximity 数据构建向量空间模型的方法 ………………… 徐　硕　乔晓东　朱礼军　郭怀恩(643)
科技信息资源内容监测与分析服务平台概况 …………………… 徐　硕　乔晓东　朱礼军　张运良(651)
多任务最小二乘支持向量回归机及其在近红外光谱分析技术中的
应用研究 ……………………………………………… 徐　硕　乔晓东　朱礼军　安　欣　张录达(659)
试论微博客在图书情报机构应用 ……………………………………………………… 许素全　熊三炉(664)
一种断点续传的多线程新闻组抓取方法及存储结构
…………………………………………………………… 杨代庆　王志苹　王　星　刘敏健　常迎春(669)
我国博硕士学位论文的服务模式及版权问题探索 …………………………………… 杨少萱　褚汉启(675)
2009 中国自然科学论文作者的简要统计分析
……………………………………………………… 杨志清　潘云涛　张玉华　马　峥　苏　成　郭　红
徐　波　俞征鹿　袁军鹏　俞良行　张　梅　王小琴(682)
基于 XML 元数据交换任务的国际科技系统间互操作——以 JST 亚洲科技门户电子期刊和
ISTIC 精品期刊为例 …… 姚长青　李　颖　严莹莹　焦淑娟　乔晓东　加藤齐史　加藤多惠子(693)
基于 Android 平台的移动电子政务系统 …………………………………………… 叶小榕　邵　晴(700)

图情研究

A Novel Category Vector – Based Cross Language Text Categorization Method

Yingfan Gao[1], Hongjiao Xu[2], Wei Yu[3], Huilin Wang[4]
(*Institute of Scientific and Technical Information of China*
{[1]*gaoyingf*, [2]*xuhj*, [3]*yuwei*, [4]*wanghl*} *@ istic. ac. cn*)

Abstract: Due to the globalization on the Web, monolingual Text Categorization can be reformulated as a cross language TC task To establish a practical English – Chinese CLTC system, a feature translation method and a fast text categorization algorithm based on a novel Category Vector Space Model are proposed in this paper Provided a Chinese – English bilingual dictionary in scientific and technological fields, parallel corpora was employed to append translation probability value to bilingual dictionary so as to disambiguate translation results The experiment results show that the CLTC system which was established by method in the paper is practical and valuable The performance of Cross – Language text categorization system exceeds that of Mono – Lingual text categorization system and the result is exciting.

Keywords: Translation Probability; FTCC Algorithm; Cross – language Categorization Effectiveness

1 Introduction

Text categorization (TC) is an important method for information organization. The purpose of TC is to classify documents into some predefined categories automatically. TC algorithms, such as KNN, SVM, Naive Bayese and so on, are very mature by now and have entered the period of practicality, But in recent years because of the growth in the popularity of the Web, many companies and organizations were required to manage documents in different languages. So it is essential to solve the problem of classifying multi – language documents. Moreover, the mapping among different classification systems of multi – language is another conundrum. The automatic mapping of multi – language thesaurus is very representative. the Cross – Language Text Categorization (CLTC) can provide techniques to extend existing automatic classification systems in one language to new languages without requiring additional intervention of human experts. So CLTC has significance both theoretically and practically.

In CLTC, the system is trained using labeled examples in a source language, and it classifies documents in a different target language [1]. The difference between CLTC and Mono – Lingual text categorization (MLTC) lies in different language space for trainmg documents and test documents. The three most common translation methods are parallel corpus – based translation method, machine learning software – based translation method, and dictionary – based translation method.

Text Feature Translation is one of the major issues to CLTC. "Feature" means words or phrases in documents [2], After feature translation, the documents from source language would be transformed into target language and the words or phrased in translated documents could satisfy the conditional independence assumption among different features. So text categorization algorithms based on conditional independence assumption could do well. A fast text categorization algorithm is proposed in the paper. The algorithm is on the basis of a novel Category Vector Space Model (CVSM) and is called FTCC (Fast Text Categorization based on

CVSM) algorithm.

On the basis of a Chinese – English bilingual dictionary and large – scale parallel corpora in scientific and technological (S & T) fields, establishing a practical cross – language text categorization system is our goal with high speed and acceptable accuracy. This paper is organized as follows. In section 2, we will introduce Cross – language Feature Translation Method. In Section 3, we will discuss FTCC algorithm. In section 4, Experimental procedure and the experimental evaluations will be harrated and listed. The conclustions are given in section 5.

2 Cross – Language Feature Sellection Method

2.1 Related Work

To MLTC, "Feature selection" means using ceratin evaluation function to assign each original feature item marks, sort all the features according to its evaluation result and select the optimum feature as feature subset. To CLTC, there are two possible ways for feature selection one is done before translation, the other is done after translation. Feature selection after translation has been employed by many researchers [4 – 8] Using different translation resource, researchers translate source language documents into target language After that, feature selection in target language would be done. Few researchers [3] [9] use the method of feature selection before translation. Bel N. et al. [3] translated the words which occur only in a category and the method was very practical, low cost, but poor accuracy. Bel ' s method was representative, but their translation dictionary was very limited in scale and translation method was too simple to disambiguate the resutl of translation. The efficient feature translation methods proposed in the paper focus on translation disambiguation on the conditions of a Chinese – English biligual dictionary and large – scale parallel corpora in S&T fields.

2.2 Maximum Forward Matching for Multi – Word Phrases Translation

Using a bilingual dictionary is the better approach when no commercial MT (Machine Translatin) system with an established reputation is available, especially in S&T fields. We can translate document easily by replacing each term of document with its translation equivalents appearing in a bilingual dictionary or a bilingual term list. The bilingual dictionary available is in S&T fields and there are many multi – word phrases in it. So, identifying multi – word concepts of documents and matching them with dictionary are the main tasks of translation. The method we used is *Forward Maximum Matching* (*FMM*) . FMM [10] methods appear in Chinese word segmentation programs usually. In that cases, Chinese input would be segmented into single word first, the according to Chinese dictionary, the single words would be assembled to phrases. To English, there are spaces among the words, so segmentation is not necessary.

In this paper, the bilingual dictionary in S&T fields includes many multi – word phrases. So FMM method we used is forward matching multi – word phrases in documents according to the bilingual dictionary. The main idea is like the FMM in Chinese word segmentation. But before FMM process, the Chinese documents here will need word segmentation first by some kind of open source software (we use Stanford segmenter [11]developed by the Stanford Natural Language Processing Group) and maintain the smallest size of segmentation. The English documents will not need word segmentation. Moreover, before matching, some preprocessing steps should be done, including lowercase transforming, stopwords filtering, stemming, and lemmatizing, et al. After preprocess above, FMM process the treats the Chinese documents and the English documents alike and that ' s the main difference between FMM in this paper and FMM in reference. Owing to the limitation of space, the more details will be abbreviated here

Ballesteros [12] pointed out problems with this method as follows:

- Specialized vocabulary not contained in the dictionary will not be translated.
- Dictionary translations are inherently ambiguous and add extraneous information.

For applying dictionary – based translation effectively, it is important to resolve ambiguity of word sense since several translations with a different meaning are usually listed in each entry of bilingual dictionaries. We proposed a novel and efficient translation disambiguation method based on bilingual dictionary and parallel corpus.

2. 3 Identifying OOC Words

Given a finite vocabulary, the presence of out – of – vocabulary (OOV) words is inevitable. OOV words constitute a major source of error in document translation. The vocabularies in bilingual dictionary are limited with respect to magnanimity literature. To solve the question, we use large – scale parallel corpora to complement information.

"Translation" here means mapping term statistics from one language to another, not simply replacing the terms themselves Translation probabilities can be estimated from parallel corpora. With sentence – aligned parallel corpora, the freely available GIZA + + toolkit [13] can be used to train translation models. GIZA + + produces a representation of a sparse translation matrix and many words, forms. or expressions are incorrect or unacceptable. So we remove these words first according to some prearranged rules. For example, if a digit and a word are linked together without space or words with different language are linked together without space, there would be wrong. Then, after data cleaning, each word' s translation probabilities would be recomputed according to the principle of normalization. Moreover, to further reduce the translation noise, we set threshold of translation probabilities. Considering larger scale of the sparse translation matrix produced by GIZA + +, we have employed the random sampling method to get the threshold of translation probabilities. Experimental results show that threshold value 0. 3 is appropriate for filtering most of translation noise. Owing to the limitation of space, we would not talk about treatment of details here.

Coming from parallel corpora, translation probability dictionary could solve the OOV problem to a certain extent and would be beneficial supplement to the bilingual dictionary in S&T fields.

2. 4 Bilingual Dictionary with Probability Value

There are multi – translation entries for a word (or a phrase) in the bilingual dictionary usually and it' s difficult to distinguish the different translation. The method in section 2. 3 couldn' t be used to improve the bilingual dictionary available completely. Because the bilingual dictionary that we use is in S & T fields, there are many multi – word phrasese in it and each multi – word phrase might be of multi – translation entry. The method in section 2. 3 use GIZA + + toolkit and could produce translation probability only at word level, not at phrase level. We have to ruse the parallel corpora to make the phrases' probability value. The new method could be described as follow:

Input: Sentence – aligned parallel corpora; bilingual dictionary

Output: Bilingual Dictionary with probability values

Steps:

- Take an English term ET from the English – Chinese bilingual dictionary, and simultaneously take one or multi – Chinese translation entries CT_l ($l = 1, 2, \ldots, k$) corresponding to ET where k is the number of CT to ET
- According to sentence – aligned parallel corpora, we get the number $n_{CT_l \mid ET}$ which is the occurrence times between CT_l and ET in English – Chinese parallel sentence pairs.
- When ET is translated to CT_l, we could compute the translation probability value according to formula as follows:

$$P\ (CT_l \mid ET) = \frac{n_{CT_l ET}}{\sum_{l=1}^{k} n_{CT_l ET}} \qquad (1)$$

- In the Chinese – English direction, the method is like steps above. But the Chinese word segmentation by open source software should be done before the steps above.

We consider, to a word or a phrase in dictionary, the different probability value of translation entry reflects its translation habit in reality, The bigger the translation probability value, the higher the possibility that the word or the phrase is translated to the translation entry. And these appended information all comes from the parallel corpora and could help us reduce the ambiguity of feature translation.

3 Fast Text categorization Algorithm based on CVSM

3.1 Related Work

TFIDF classifier is based on the Rocchio relevance feedback algorithm and uses TFIDF word weights. There are a number of algorithms in this family which differ in their selection of word weighting method and similarity measure [14] . The basic idea of the algorithm is to represent each documented as a vector in a vector space. Each dimension of the vector space represents a word selected by the feature selection process described above.

The values of the vector elements for a document are calculated as a combination of the statistics TF (w, d) and DF (w) . The *term frequency* TF (w, d) is the number of times word w occurs in documented. The *document frequency* DF (w) is the number of documents in which the word w occurs at least once, The *inverse document frequency* IDF (w) can be calculated from the document frequency. The wight of word w_1 in documented could be calculated as TF (w_1, d) . IDF (w_1) . The TFIDF algorithm learns a class model by combining document vectors into a prototype vector for every class. Prototype vectors are generated by adding the document vectors of all documents in the class. To classify a new document d', the cosine of the prototype vector of each class with the vector of d' is calculated. The new document is assigned to the class with which its document vector has the highest cosine.

3.2 Category Vector Space Model

TFIDF classifier combines document vectors into a prototype for every class [15] . We consider, the constructing method of category vector of TFIDF classifier is defective. The distribution information of the features within a category and among categories could not be reflected completely. We propose a Category Vector Space Model (CVSM) to resolve the problem. CVSM considers that.

- Within a category, a feature's importance is up to two factors. One is feature's occurrence frequency in the category The other is the feature's distribution among documents of the category. The higher the frequency, the more important the feature; the more uniform the distribution, the more important the feature.
- Among categories, the more uneven the feature's distribution, the stronger the feature' s resolving ability to a certain category.

  According to CVSM, we can draw the conclusion as follows' .
- Within a category, the higher the frequency *f* and document frequency *df*, the more important the feature.
- Among categories, when a feature occurs only the category, it will have the strongest resolving abiliey to the contegory on the busie of the theory of unforation eneropy.

Note' In the following paragraphs, "feature" will be replaced by "term" We think "term" would be more comprehensible.

3.3 Weights of CVSM

Provided n categories such as C_1, C_2,..., C_n. k_1 represents the number of documents in category C_i (i=1, 2,... n). t_1, t_2,..., t_m are *m* terms in training documents set. f_y indicates the number of timese the term $t_1$occurs in category C_J, df_{ij} indicates document frequency the term t_1 occurs in category C_j.

- Within – category weight

The within – category weight I_{ij} that the term t_{ij} occurs in category C_1 could be measured by f_{ij}, df_{ij} or combination of the both. In the paper, we choose df_{ij} as the metric for measuring the importance of term t_{ij} in category C_1 and it would be recomputed according to the principle of cosine normalization as follows:

$$I_{ij} = df_{ij}/\sqrt{\sum_{1=1}^{m} df_{ij}^2} \quad (2)$$

- Among – categories weight

The entropy of term t_i in all categories could be computed as follows:

$$H_i = -\sum_{J=1}^{n}\frac{f_y}{F_1}log\frac{f_y}{F_1} \quad (3)$$

According to the properties of entropy, H_1 is between 0 and *logn* $(0 \leqslant H_1 \leqslant logn)$. So H_1/logn would be between 0 and 1 $(0 \leqslant Ht/logn \leqslant 1)$. When term t_1 occurs in all categories with the same probability, Hol/lorgn would dose tol. when term occurs only in a certain antegory, H. /logn would close to O. Set E to 1 – Hi/logn. So when term ti olcurs in all cateyories with the same probability, E would be equal to 0. When term t_1 occurs only in a certain category, E would be equal to 1 E is called *weight factor among categories* in the paper and could be formulated by:

$$E_1 = 1 - \frac{H_1}{logn} = 1 + \frac{1}{logn}\sum_{J=1}^{n}\frac{f_y}{F_1}log\frac{f_y}{F_1} \quad (4)$$

- Total weight of term t_1 in category vector

Total weight of term t_1 in vector of category C_J could be formulated as follows’

$$W_{ij} = Z_{ij} \times E_i (i = 1,2,...,m; j = 1,2,...,n) \quad (5)$$

3.4 FTCC Algorithm

The Fast Text Categorization based on CVSM (FTCC) algorithm is proposed in the paper. In FTCC, the vector representation method of new document which is to be classified would be the same as that of the category vector. Like TFIDF classifier, cosine similarity could be still used to measure the distance between the vector of new document and category vector. After sorting according to the distance above, the category that is the shortest distance to the new document would be the requested result The form of the distance is as follows:

$$\mathrm{sim}C_{JX} = \cos(C_J, d_x) = \frac{C_j \cdot d_x}{\| C_J \| \, \| d_x \|} = \frac{\sum_{i=1}^{m} W_{ij} W_{1\lambda}}{\sqrt{\sum_{i=1}^{m} W_{IJ}^2}\sqrt{\sum_{i=1}^{m} W_{1\lambda}^2}} \quad (6)$$

Where d_x represents the vector of document to be classified and C_J represents category vector. $W_{i\lambda}$ and W_{iJ} are the vector weight of the word t_i in document vector d_x and category vector C_j.

In order to increase the computing speed of sim$C_{i\lambda}$ and reduce the processing overhead of data sparsity, a condensed vector representation method such as $\{t_1, W_{i\lambda}\}$ is proposed here, where t_1 represents term which occurs in the document to be classified and $W_{i\lambda}$ is the weight for the component of d_x corresponding to t_1. So, most terms which appear in category but don’ t appear in the document to be classified will be ignored. The process of computing sim$C_{i\lambda}$ is to search W_{ij} in C_i according to t_1. That is to say, the key stage of FTCC is to search a certain term in a m – dimensional space where m is the number of terms in training documents set. In the worst – case scenario, searching would be done *logm* times. If vectors could be saved using hash table, searching speed would be quicker.

4 Experiments

4.1 Evaluation Method

The most popular index for evaluating the CLTC system is the cross – language categorization effectiveness η, which is the ratio of average precision of CLTC system to MLTC system.

$$\eta = \frac{Q_{CLTC}}{Q_{MLTC}} \times 100\% \quad (7)$$

Where Q_{CLTC} is the average precision of CLTC system and Q_{MLTC} is the average precision of MLTC system. MLTC system here is considered as the baseline system for measuring CLTC system.

The reference availble of CLTC didn’ t mention whether there should be uniform test collection for CLTC and MLTC We consider that comparing CLTC and MLTC in the same test collection would be more convictive. To MLTC, the recognized test method is 10 – fold cross validation. The method divided the data set into 10 parts, 9 of which would be training set in turn and the remaining 1 of which would be used for testing. The average of 10 experimental results could be used for estimating the accuracy of categorization algorithm. The 10 – fold cross validation could assure there' s no overlap between training document set and testing document set for each experiment and the average of 10 experimental results is very fair.

To CLTC, the source language and target language are different. So to ensure that the experiments are fair. we would first divide documents which are translated from Chinese to English (C – E) into 10 parts randomly as training sets. Then the 10 testing sets of MLTC experiments would be used to test these training sets respectively Each training set would be tested on 10 testing sets individually. So there are 100 times TC experiments in all. The average of 100 experimental results would be the final results. We name this evaluation method 100 – *fold cross validation.*

In order to keep the consistence of testing method between MLTC and CLTC, there would be 10 training set and 10 testing set for MLTC too. Training classifier on each training set and testing the documents on 10 testing set respectively. 100 times experiments would be donefor MLTC too and this would be the *baseline* experiment used for comparing with CLTC.

4. 2 Experimental Corpora

To do Cross – Language Feature Translation, a Chinese – Enghsh bilingual dictionary with approximate 450,000 Chinese terms and 460,000 English terms[1] are available m the S&T fields. Moreover, about 1,100,000 sentence – aligned parallel corpora[2] are used to produce bilingual dictionary with probability values.

To CLTC experiment, experimental documents have been got from WanFang DATA company (www. wanfangdata com cn) and all are dissertation abstracts in the S&T fields. According to the Chinese Library Classification method, we selected 9 categories including Aero, agriculture, architecture, chemistry, energy, environment, machine, miliary and mine. There are 546 Chinese documents and 517 English documents in experimental document set[3] and could he shown in Table 1:

Table 1 Experimental Corpora

Category Language	Aer – o	Agriculture	Architecture	Chemistry	Energy	Environment	Machine	Military	Mine
Chinese	42	17	46	27	138	28	107	16	96
English	53	17	71	23	136	28	104	21	93

To MLTC experiment, the ratio of training set to testing set is 75% to 25% in the paper. To CLTC, after cross – language feature translation, documents translated from C – E would be extracted randomly, the number of extracted documented would be equal to that of training set in MLTC experiments, and these extracted documents could be the training set of CLTC experiments. The random extraction experiment here would be done for 10 times independently. The Macro – F1 and Micro – F1 index would become the global measurement index in the paper.

4. 3 Baseline Experiments

According to the algorithm in section 3, we ran the experiment in test corpora in section 4. 2. The experiment results of 100 – fold cross validation could be shown in Table 2. And comparison between FTCC algorithm in section 3. 4 and TFIDF algorithm in section 3. 1 would be listed in Table 3. With the aim of space saving, only the average result of 100 – fold cross validation experiments would be shown.

4. 4 CLTC Experiment 1

In this experiment, Chinese is the source language

[1] The copyrights of the English – Chinese bilingual dictionary are owned by ISTIC

[2] The copyrights of the Sentence – aligned parallel corpora are owned by ISTIC

[3] The interested reader could get my experimental corpora by sending an email to me and it' s free