

内部

国家 242 信息安全计划 响应文件

指南编号：17 指 153

指南名称：国内行业专家画像库
系统

北京航空航天大学
2018-04-08



目 录

第一部分 基本文件.....	3
第一章：磋商响应函.....	4
第二章：磋商报价一览表.....	7
第三章：法定代表人授权书.....	9
第四章：联合体授权委托书和合作协议(不涉及).....	13
第二部分 商务及资格证明文件.....	14
第五章：供应商概况.....	15
第六章：供应商资格声明文件.....	16
第七章：★信用记录证明.....	18
第八章：★供应商营业执照（副本）或事业单位法人证书复印件.....	20
第九章：★供应商组织机构代码证及税务登记证复印件.....	22
第十章：★备案成功通知书.....	23
第十一章：供应商质量管理体系认证证书.....	26
第十二章：供应商近三年相关项目业绩证明文件.....	27
第十三章：★依法缴纳税收和社会保障资金的证明.....	29
第十四章：★谈判供应商参加政府采购活动前三年内在经营活动中没有重大违法记录的书面证明.....	31
第十五章：★近一年银行提供的资信证明或近一年经审计的财务报表.....	33
第十六章：★保密资质证明.....	35
第十七章：★具有履行合同所需的设备和专业技术能力的声明.....	37
第十八章：★供应商控股股东名称、控股公司的名称和存在管理、被管理关系的单位名称说明.....	39
第十九章：★供应商没有为本项目提供过整体设计、规范编制或者项目管理、监理、检测等服务的投标人的书面声明.....	41
第三部分 课题申请书.....	43
第二十章：课题申请书.....	44
一. 基本信息.....	47
二. 课题组主要研究人员.....	48
三. 课题情况.....	54
四. 声明.....	78
第二十一章：申报课题信息简表.....	79



第一部分 基本文件



第一章：磋商响应函



磋商响应函

致：国家计算机网络与信息安全管理中心

根据贵方为 242 课题委托开发(17 指 153,国内行业专家画像库系统)项目采购服务的磋商邀请, 签字代表()经正式授权并代表供应商(北京航空航天大学, 北京市海淀区学院路 37 号北京航空航天大学)参加贵方组织的本次采购活动, 并对此项目进行响应。

在此, 签字代表宣布同意如下:

1、我方同意在本项目磋商文件中规定的响应文件递交截止时间起 90 天内遵守本响应文件中的承诺且在此期限期满之前均具有约束力。

2、我方承诺已经具备《中华人民共和国政府采购法》中规定的参加政府采购活动的供应商应当具备的条件:

- (1) 具有独立承担民事责任的能力;
- (2) 具有良好的商业信誉和健全的财务会计制度;
- (3) 具有履行合同所必需的设备和专业技术能力;
- (4) 有依法缴纳税收和社会保障资金的良好记录;
- (5) 参加此项采购活动前三年内, 在经营活动中没有重大违法记录。

3、我方承诺我单位及申请团队已注册为国家 242 信息安全专项合作单位及团队。

4、针对涉及国家秘密课题的申请, 我方承诺已经具备国家保密局认可的涉密资质或通过采购人专项保密审查。

5、我方承诺我单位法人及主要业务、技术人员均无犯罪记录。

6、提供供应商须知规定的全部响应文件, 包括响应文件正本 1 份, 副本 0 份, 磋商报价一览表正本 1 份, 以及提供包括响应文件全部内容的电子版 1 份。

7、按磋商文件要求提供和交付的服务的响应报价为捌拾伍万元整。

8、保证忠实地执行双方所签订的合同, 并承担合同规定的责任和义务。



9、保证遵守磋商文件的各项规定。

10、我方完全理解贵方不一定接受最低价的响应或收到的任何响应。

11、我方愿意向贵方提供任何与本项目有关的数据、情况和技术资料。若贵方需要，我方愿意提供我方做出的一切承诺的证明材料。

12、我方已详细审核全部响应文件，包括响应文件修改书（如有的话）、参考资料及有关附件，确认无误。

13、我方将严格遵守《中华人民共和国政府采购法》的有关规定，若有下列情形之一的，将被处以采购金额 5%以上 10%以下的罚款，列入不良行为记录名单，在一至三年内禁止参加政府采购活动；有违法所得的，并处没收违法所得；情节严重的，由工商行政管理机关吊销组织机构代码证。构成犯罪的，依法追究刑事责任：

- (1) 提供虚假材料谋取成交的；
- (2) 采取不正当手段诋毁、排挤其他供应商的；
- (3) 与采购人、其它供应商工作人员恶意串通的；
- (4) 向采购人工作人员行贿或者提供其他不正当利益的；
- (5) 未经采购人同意，在采购过程中与采购人进行协商磋商的；
- (6) 拒绝有关部门监督检查或提供虚假情况的。

与本项目有关的一切往来通讯请联系

地址：北京市海淀区学院路 37 号北京航空航天大学

传真：010-82317850

电话：82316682 电子邮件：yolixs@buaa.edu.cn

供应商代表签字：_____

供应商名称：北京航空航天大学（盖章）

日期：2018-04-08



第二章： 磋商报价一览表



磋商报价一览表

指南名称：国内行业专家画像库系统

指南编号：17 指 153

磋商报价（元）	其他 声明
大写：捌拾伍万元整 小写：850000.00	

注： 1、以上报价是供应商为完成本项目、达到磋商文件第三章项目需求的所有费用，已包括所有相关税费。

2、此表应按“供应商须知”的规定密封标记并单独提交，同时在装订成册的响应文件中仍应提供本表。

供应商代表签字：_____

供应商名称：北京航空航天大学（盖章）

日 期：2018-04-08



第三章： 法定代表人授权书



法定代表人授权书

本授权书声明：注册于中华人民共和国的(北京航空航天大学)的在下面签字的(徐慧彬、校长)代表本单位授权的在下面签字的(张小明、助理教授)为本单位的合法代理人，就指南编号为 17 指 153 的(国内行业专家画像库系统)的响应（含转为其他方式），以本单位名义处理一切与之有关的事务。

本授权书于 2018 年 04 月 10 日 签字生效，特此声明。

法定代表人签字：_____

被授权人签字：_____

供应商盖章 北京航空航天大学(盖章)

（附：被授权人身份证复印件）





五世版木



第四章： 联合体授权委托书和合作协议 (不涉及)



第二部分 商务及资格证明文件



第五章： 供应商概况



第六章： 供应商资格声明文件



资格声明

国家计算机网络与信息安全管理中心：

为响应你方 2018-04-10 的 242 课题 17 指 153 磋商邀请，下述签字人自愿参与本项目，提供磋商文件中规定的国内行业专家画像库系统，提交下述文件并声明全部说明是真实的和正确的。下述签字人将就下述文件中存在的虚假或不真实内容对采购人承担法律责任。

(1) 我方的资格声明，各有 1 份正本，0 份副本。

(2) 下述签字人证明本资格文件中的内容是真实的和正确的。

下述签字人知道，采购人可能要求其提供进一步的资格材料并同意按采购人的要求提交。

供应商

名称（盖章）：北京航空航天大学

地址：北京市海淀区学院路 37 号北京航空航天大学

传真：010-82317850

邮编：100191

授权签署本资格文件

签字人姓名、职务（印刷字体）张小明、助理教授

签字：_____

电话：82316682



第七章：★信用记录证明



当前位置：首页 > 政府采购严重违法失信行为记录名单 >



政府采购严重违法失信行为信息记录

zhf-wentian.com

企业名称：北京航空航天大学

执法单位：

处罚日期：

至

查询

重置

序号	企业名称	统一社会信用代码 (或组织机构代码)	企业地址	严重违法失信行为 具体情形	处罚结果	处罚期限	处罚日期	公布日期	执法单位
没有该企业的记录									
本次查询的企业：北京航空航天大学									
本次查询的时间：2018年04月02日 14时12分									

提示：本平台信息依据《关于报送政府采购严重违法失信行为信息记录的通知》（财办库[2014]526号）发布，如有疑问请联系具体执法单位。

版权所有 © 2018 中华人民共和国财政部



第八章： ★供应商营业执照（副本）或 事业单位法人证书复印件



第九章：★供应商组织机构代码证及税务登记证复印件



第十章：★备案成功通知书



国家 242 信息安全专项合作单位及团队的书面证明

致：国家计算机网络与信息安全管理中心

我单位已成功注册为国家 242 信息安全专项合作单位，李舟军老师负责的信息安全与智能信息处理实验室团队已成功注册为合作团队，具体信息如下：

单位名称：北京航空航天大学

单位备案号：242HZDW01700

团队名称：信息安全与智能信息处理实验室

团队备案号：242HZDW01701

团队负责人：李舟军

特此声明。

团队负责人：

合作单位(盖章)：北京航空航天大学



知情同意书

同意张小明作为信息安全与智能信息处理实验室（242HZDW01701）的成员申请国内行业专家画像库系统（17 指 153）课题。

团队负责人签字：_____

申请人签字：_____

日期： 2018-04-08



第十一章： 供应商质量管理体系认证证书



第十二章： 供应商近三年相关项目业绩 证明文件



供应商近三年相关项目业绩

项目名称	项目单位	业主联系人/电话	项目总价
面向特定领域的知识抽取与政局风险预测关键技术研究	国家计算机网络与信息安全管理中心	赵忠华/15210631358	130
面向局域网的网络地图信息库系统构建	国家计算机网络与信息安全管理中心	赵忠华/15210631358	50

- 注：1.供应商应将其近 3 年内承担的相关项目业绩填入本表中。
 2.供应商上述项目业绩须提供包括但不限于合同或委托函作为证明文件
 3.供应商业绩及证明文件应如实提供，如有虚假将导致其响应被拒绝。

供应商名称（盖章）：_____

供应商代表签字：_____

日 期：_____



第十三章：★依法缴纳税收和社会保障 资金的证明



事业单位不需提供

事业单位不需提供



第十四章：★谈判供应商参加政府采购 活动前三年内在经营活动中没有重大违 法记录的书面证明



无重大违法记录书面声明函

致：国家计算机网络与信息安全管理中心

我单位参与编号为 17 指 153 号项目投标，我单位郑重声明：我方参加本次采购活动前三年内，在经营活动中无重大违法活动记录，在北京市行政区域内未受到处理并且不在投标限制期限内，符合《政府采购法》规定的供应商资格条件。我方对此声明负全部法律责任。

特此声明

供应商：（盖章）



法人代表或授权代表：（签字或盖章）

（签字或盖章）



第十五章：★近一年银行提供的资信证明或近一年经审计的财务报表



资 信 证 明 书 (正 本)

Certificate of Creditworthiness (original)

编号: 京B00207361

日期: Date: 2018 04 03

致: 国家计算机网络与信息安全管理中心

北京航空航天大学委托我行对其在我行的结算记录开立证明书。经确认, 其具体情况如下:

自2017年04月02日至2018年04月02日(即该日我行营业终了结账时)止, 该单位无违反我行结算制度规定的行为。

仅此证明, 下无正文。

银行签章
Bank's Seal

有权签字人
Authorized Signature

验证码: [0354B0
F48020]

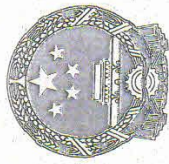
提示: 阅读本证明书时请同时阅知证明书背面“声明”。

Note: Please read this certificate in conjunction with "Statement" on the back of the certificate.



第十六章： ★保密资质证明



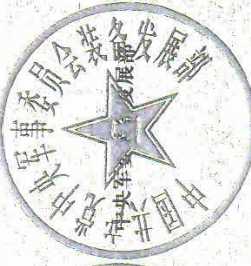
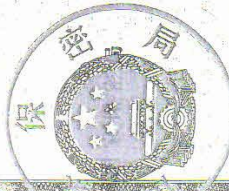


武器装备科研生产单位
一级保密资格证书

(副本)

编号: BJA17005

有效期至: 2022年 6月 15日



单位名称: 北京航空航天大学
注册地址: 北京市海淀区学院路37号
法定代表人: 徐惠彬

编号 18-01-001

此件为内部文件, 仅用于

242101010101

不可复印, 传真和上网!

说明: 1. 保密资格证书自颁发之日起生效;

正本和副本具有同等法律效力。

2. 保密资格证书不得伪造、涂改、出借、转让, 如发生丢失应及时向发证机关报告。

3. 保密资格证书有效期满后, 即行作废。



第十七章：★具有履行合同所需的设备 和专业技术能力的声明



有履行合同所需的设备和专业技术能力声明函

致：国家计算机网络与信息安全管理中心

我单位参与编号为 17 指 153 号项目投标，我单位郑重声明：我单位具有履行合同所需的设备和专业技术能力，符合《政府采购法》第二十二条规定的供应商资格要求。我方对此声明负全部法律责任。

特此声明

供应商：（盖章）

法人代表或授权代表：（签字或盖章）张小明



第十八章： ★供应商控股股东名称、控股公司的名称和存在管理、被管理关系的单位名称说明



供应商隶属关系声明

致：国家计算机网络与信息安全管理中心

我单位参与编号为 17 指 153 号项目投标，我单位郑重声明：我方属于事业单位，隶属于中华人民共和国工业和信息化部，我方对此声明负全部法律责任。

特此声明

供应商：（盖章）

法人代表或授权代表：张小明（签字或盖章）



**第十九章：★供应商没有为本项目提供
过整体设计、规范编制或者项目管理、
监理、检测等服务的投标人的书面声明**



书面声明

致：国家计算机网络与信息安全管理中心

我单位参与编号为 17 指 153 号项目投标，我单位郑重声明：我方非为本项目提供整体设计、规范编制或者项目管理、监理、检测等服务的投标人。我方对此声明负全部法律责任。

特此声明

供应商：（盖章）

法人代表或授权代表：（签字或盖章）张小明



第三部分 课题申请书



第二十章： 课题申请书



指南编号：17 指 153

申请书编号：

密 级： 内部

国 家 242 信 息 安 全 计 划

课题申请书

课题名称： 国内行业专家画像库系统

申 请 人： 张小明

依托单位： 北京航空航天大学

通讯地址： 北京市海淀区学院路 37 号北京航空航
天大学

邮 编： 100191

联 系 人： 李舟军 (82338247)

传 真： 010-82317850

电子邮箱： yolixs@buaa.edu.cn

申请日期： 2018-04-08

国家计算机网络与信息安全管理中心



填 写 说 明

- 一、 请严格按表中要求填写各项。
- 二、 凡是机密、秘密或内部课题,在申请书封面标上“机密”、“秘密”或“内部”,无密课题不标注。
- 三、 申请书编号由国家计算机网络与信息安全管理中心填写。
- 四、 申请书文本中外文名词第一次出现时,要写清全称和缩写,再出现同一词时可以使用缩写。



一. 基本信息

课题 申请 人情况	姓名	张小明				性别	男			
	文化程度	博士				职称	中级			
	最后学位	博士研究生学位				所学专业	计算机科学与技术			
	出生年月	1980-12-07				手机	13810643192			
	联系电话	82338247				电子邮件	yolixs@buaa.edu.cn			
课题 组 情况	(人) 高级 职称	(人) 中级 职称	(人) 初级 职称	无职称 (人)	(人) 研究生 学历	(人) 本科 学历	(人) 大专 学历	博士 生(人)	硕士 生(人)	本科 生(人)
	0	2	0	13	15	0	0	2	8	5
依托 单位 情况	名称	北京航空航天大学								
	通讯地址	北京市海淀区学院路 37 号北京航空航天大学					邮编	100191		
	单位性质	事业单位					参加单位总数	1		
	所在地区 (省、自治区、直辖市、计划单列市)	北京市海淀区								
	所属部门 (国务院各部、委、局及其机构)	计算机学院								
其它 主要 联合 申请 单位	单位名称				参加形式			单位性质		
申报日期		2018-04-08								



二. 课题组主要研究人员

序号	姓 名	性 别	出生年月	身份证号	学位	职称	学历	专 业	研究方向	为本课题工 作时间（人 年）	在课题中职务（组 长、副组长或成员） 及分担的任务	所在单位
1	张小明	男	1980-12-07	43041919801207 3417	博士 学位	中级	研 究 生	计 算 机 科 学 与 技 术	数 据 挖 掘	1	组 长、理 论指 导	北 京 航 空 航 天 大 学
2	姚燕青	女	1983-02-15	13013319830215 3028	博士 学位	中级	研 究 生	计 算 机 科 学 与 技 术	信 息 安 全	1	副 组 长、理 论指 导	北 京 航 空 航 天 大 学
3	黄斐然	男	1987-12-10	43052219871210 0010	硕士 学位	无	研 究 生	计 算 机 科 学 与 技 术	数 据 挖 掘	1	成 员、专 家检 索	北 京 航 空 航 天 大 学
4	闫昭	男	1988-09-02	14010519880902 0554	硕士 学位	无	研 究 生	计 算 机 科 学 与 技 术	自 然 语 言 处 理	1	成 员、专 家观 点分 析	北 京 航 空 航 天 大 学
5	周志博	男	1991-01-01	43052119910101 2618	硕士 学位	无	研 究 生	计 算 机 科 学 与 技 术	数 据 挖 掘	1	成 员、专 家信 息采 集	北 京 航 空 航 天 大 学
6	李朝卓	男	1991-03-03	37152219910303 001X	硕士 学位	无	研 究 生	计 算 机 科 学 与 技 术	数 据 挖 掘	1	成 员、专 家影 响力 分 析	北 京 航 空 航 天 大 学



7	吴侯	男	1992-04-27	110107199204270034	硕士学位	无	研究生	计算机科学与技术	自然语言处理	1	成员、专家观点分析	北京航空航天大学
8	刘昊岩	男	1994-07-20	150421199407200037	硕士学位	无	研究生	计算机科学与技术	自然语言处理	1	成员、专家社会关系分析	北京航空航天大学
9	刘赞	男	1994-08-03	522130199408032419	硕士学位	无	研究生	计算机科学与技术	数据挖掘	1	成员、专家信息采集	北京航空航天大学
10	陈琚	男	1994-08-22	430424199408224838	学士学位	无	研究生	计算机科学与技术	数据挖掘	1	成员、专家影响力分析	北京航空航天大学
11	杨德剑	男	1994-08-24	341125199408248716	学士学位	无	研究生	计算机科学与技术	数据挖掘	1	成员、专家检索	北京航空航天大学
12	周逸坤	男	1994-09-30	440508199409302612	学士学位	无	研究生	计算机科学与技术	数据挖掘	1	成员、专家影响力分析	北京航空航天大学
13	陈昊	男	1994-10-14	120103199410142611	硕士学位	无	研究生	计算机科学与技术	数据挖掘	1	成员、专家影响力分析	北京航空航天大学
14	张博	男	1995-03-16	320721199503162059	学士学位	无	研究生	计算机科学与技术	数据挖掘	1	成员、专家关系分析	北京航空航天大学
15	罗雅文	女	1995-06-17	43100219950617	学士	无	研究	计算机	数据挖掘	1	成员、专家信	北京航空



				0029	学位		生	科学与 技术			息采集	航天大学
--	--	--	--	------	----	--	---	-----------	--	--	-----	------

计：15 人年



2.1 课题组长资历情况(所学专业、学历, 从事过的主要研究工作及所负责任, 与所申请课题有关的研究成果和获发明专利情况, 在国内外主要刊物上发表论文情况, 开发的主要产品及其市场情况等)

张小明, 男, 1980 年生, 博士, 北京航空航天大学计算机学院讲师。

2003 年获国防科技大学计算机科学与技术专业学士学位, 2006 年获国防科技大学计算机软件与理论硕士学位, 2011 年获北京航空航天大学计算机应用技术博士学位。研究工作主要集中在机器学习、文本挖掘、社交媒体挖掘等方向。2009 年 9 月至 2010 年 9 月, 受国家留学基金委公派至澳大利亚 Queensland 大学, 在 HengTao SHEN 教授指导下进行联合培养。在次期间, 研究工作主要集中在图像标签信息量与关联度的计算、跨媒体搜索的图像标注、特征融合的图像标注与事件检测。2012 年留任于北京航空航天大学, 主要从事大数据挖掘、知识抽取等领域的研究, 近年来以第一作者身份共发表学术论文 20 余篇, 部分研究工作已经发表在《IEEE Transaction on Multimedia》、《IEEE Transaction on Cybernetics》、《WWW Journal》、《Signal Processing》、《Journal of Multimedia Tools and Applications》、《Journal of Intelligent Information System》、《软件学报》等重要国际和国内期刊以及 IJCAI2015、ICMR2015、SDM2014、AAAI2013、WAIM2013、DASFAA2011、ICADL2011、APWeb2011、WAIM2011、WISM2010、ADMA2009 等国际会议上。

2011 年以来, 参与和主持了多项自然科学基金项目、军口 863 项目、242 信息安全项目等。在基于文本挖掘与大数据分析方面积累了丰富的经验。领导的团队设计并研发出了文本精准推荐系统、基于局域网的地图库系统、面向社交网络的舆情监控与分析系统。其中文本精准推荐系统为支持网络信息与安全管理中心相关处室具体业务而开发, 该系统根据用户提交的事件文档或关键词组推荐事件相关的其他文档, 并且支持用户对推荐结果进行反馈, 并根据反馈信息进一步提高推荐文档的准确度。基于局域网的地图库系统为解决客户应用暴露于互联网的问题而开发, 该系统包含了世界主要城市以及中国乡镇级的地图数据, 实现了多种样式的地图展示, 提供 API 以支持用户的应用。面向社交网络的舆情监控与分析系统主要功能包括: 面向 web2.0 的数据深度采集、社交网络热点话题检测与监控、话题中关键人物分析、话题传播趋势分析与预测、微博转发评论关系的动态分析与展示、话题关注点变化分析、敏感信息自动检测、舆情预警、面向专题的自动回复等。该系统在存储方面使用了 Hadoop 的 HBase 分布式列存储, 分布式计算方面使用了 MapReduce 计算模型, 在页面前端方面应用了最新的 HTML5, ajax, highchart 等热门 Web 技术。课题组在地图呈现、服务器端推送、大数据可视化、个性化页面布局等方面积累了大量经验。

近期发表的主要论文目录:

1. Zhang, Xiaoming, Z. Li, and X. Hu. Learning Geographical Hierarchy Features via a Compositional Model. IEEE Transactions on Multimedia 2016 , 18 (9) :1855-1868.
2. Xiaoming Zhang, Xia Hu, Zhoujun Li: Learning Geographical Hierarchy Features for Social Image Location Prediction. IJCAI 2015: 2401-2407.
3. Xiaoming Zhang, Shufan Ji, Senzhang Wang, Zhoujun Li, Xueqiang Lv. Geographical Topics Learning of Geo-tagged Social Images. IEEE Trans. Cybernetics 46(3): 744-755 (2016).
4. Xiaoming Zhang, Senzhang Wang, Zhoujun Li, Shuai Ma. Landmark Image Retrieval by Jointing Feature Refinement and Multi-modal Classifier Learning," IEEE Trans.



Cybernetics, DOI: 10.1109/TCYB.2017.2712798.

5. Xiaoming Zhang, Zhoujun Li, Xueqiang Lv, Xiaoming Chen. Integrating Multiple Types of Features for Event Identification in Social Images. *Journal of Multimedia Tools and Applications*, first online: 06 January 2015.

6. Senzhang Wang, Honghui Zhang, Jiawei Zhang, Xiaoming Zhang, Philip S. Yu, Zhoujun Li: Inferring Diffusion Networks with Sparse Cascades by Structure Transfer. *DASFAA* (1) 2015: 405-421.

7. Xiaoming Zhang, Xiaoming Chen, Yan Chen, Senzhang Wang, Zhoujun Li, Jiali Xia: Event detection and popularity prediction in microblogging. *Neurocomputing* 149: 1469-1480 (2015)

8. Yan Chen, Xiaoming Zhang, Zhoujun Li, Jun-Ping Ng: Search engine reinforced semi-supervised classification and graph-based summarization of microblogs. *Neurocomputing*. 152: 274-286(2015)

9. Xiaoming Zhang, Qiao Deng, Zhoujun Li: Friendship Link Recommendation Based on Content Structure Information. *WAIM* 2015: 486-489

10. Xiaoming Zhang, Zhoujun Li, Senzhang Wang, Yang Yang, Xueqiang Lv: Location Prediction of Social Images via Generative Model. *ICMR* 2015: 275-282

11. Chaozhuo Li, Fang Wang, Yang Yang, Zhoujun Li, Xiaoming Zhang: Exploring Social Network Information for Solving Cold Start in Product Recommendation. *WISE* (2) 2015: 276-283

12. Haijun Zhang, Zhoujun Li, Yan Chen, Xiaoming Zhang, Senzhang Wang: Exploit Latent Dirichlet Allocation for One-Class Collaborative Filtering. *CIKM* 2014: 1991-1994

13. Xiaoming Zhang, Xiaoming Chen, Zhoujun Li: Fusing Multi-type Features with Distance Metric Learning for Social Image Application. *ICIMCS* 2014: 403.

14. Senzhang Wang, Sihong Xie, Xiaoming Zhang, Zhoujun Li, Philip S. Yu, Xinyu Shu: Future Influence Ranking of Scientific Literature. *SDM* 2014: 749-757 (3)

15. Xiaoming Zhang, Zhoujun Li, Wen-Han Chao, Jiali Xia: Popularity Prediction of Burst Event in Microblogging. *WAIM* 2014: 484-487

16. Yan Chen, Jichang Zhao, Xia Hu, Xiaoming Zhang, Zhoujun Li, Tat-Seng Chua: From Interest to Function: Location Estimation in Social Media. *AAAI* 2013.

17. Haijun Zhang, Chunyang Liu, Zhoujun Li, Xiaoming Zhang: Collaborative Filtering Based on Rating Psychology. *WAIM* 2013: 655-665

18. Haijun Zhang, Xiaoming Zhang, Zhoujun Li, Chunyang Liu: Collaborative Filtering Using Multidimensional Psychometrics Model. *WAIM* 2013: 691-697

19. Xiaoming Zhang, Heng Tao Shen, Zi Huang, Yang Yang, Zhoujun Li: Automatic Tagging by Exploring Tag Information Capability and Correlation. *WWWJ*. 15(3), pp 233-256, 2012.

20. Senzhang Wang, Zhoujun Li, Chunyang Liu, Xiaoming Zhang, Haijun Zhang: Training Data Reduction to Speed up SVM Training. *Applied Intelligence* .

21. Senzhang Wang, Fang Wang, Zhoujun Li, Xiaoming Zhang: Who are More Influential Spreaders? Users with Broader Social Circle or with More Friends? *CGC*



2012: 586-593

22. Senzhang Wang, Zhoujun Li, Xiaoming Zhang: Bootstrap Sampling Based Data Cleaning and Maximum Entropy SVMs for Large Datasets. ICTAI 2012: 1151-1156.

23. Yan Chen, Zhoujun Li, Liqiang Nie, Xia Hu, Xiangyu Wang, Tat-Seng Chua, Xiaoming Zhang. A Semi-Supervised Bayesian Network Model for Microblog Topic Classification. In Proceedings of COLING 2012.

24. Xiaoming Zhang, Zi Huang, Heng Tao Shen, Zhoujun Li. Probabilistic Image Tagging with Tags Expanded Based on Text-based Search. In proceeding of DASFAA 2011.

25. Xiaoming Zhang, Heng Tao Shen, Zi Huang, Zhoujun Li. Tagging Image with Informative and Correlative Tags. In Proceeding APWeb2011.

26. Senzhang Wang, Fang Wang, Yan Chen, Chunyang Liu, Zhoujun Li, Xiaoming Zhang: Exploiting social circle broadness for influential spreaders identification in social networks. World Wide Web 18(3): 681-705 (2015)

27. Senzhang Wang, Zhoujun Li, Chunyang Liu, Xiaoming Zhang, Haijun Zhang: Training data reduction to speed up SVM training. Appl. Intell. 41(2): 405-420 (2014).

28. Xiaoming Zhang, Zhoujun Li, Wen-Han Chao: Improving image tags by exploiting web search results. Multimedia Tools Appl. 62(3): 601-631 (2013)

29. Xiaoming Zhang, Xiaojian Zhao, Zhoujun Li, Jiali Xia, Ramesh Jain, Wen-Han Chao: Social image tagging using graph-based reinforcement on multi-type interrelated objects. Signal Processing 93(8): 2178-2189 (2013).

30. Xiaoming Zhang, Zhoujun Li, Wenhan Chao. Tagging Image by Merging Multiple Features in an Integrated Manner. Journal of Intelligent Information Systems. 39(1), 87-107, 2012.

31. 张小明, 李舟军, 巢文涵: 基于增量型聚类的自动话题检测研究. 软件学报, 23(6), 1578-1587, 2012.

2.2. 课题组长及课题组主要成员目前承担其他科技计划课题情况

1. 北京市自然科学基金面上项目, 4182037, 基于多模态学习的社会图像情感分析技术研究, 2018.01-2020.12, 25万, 主持。

2. 软件开发环境国家重点实验室开放课题, SKLSDE-2017ZX-19, 社交媒体数据特征向量学习及应用关键技术研究, 2017.03-2018.12, 主持。

3. 国家自然科学基金青年基金项目, 61202239, 基于跨媒体数据挖掘的社会图像事件分析与标注, 2013/01-2015.12, 25万元, 已结题, 主持。

4. 中央高校基本科研课题, YWF-13-T-RSC-072: 社交网络中热点话题预测与人物关系分析, 2013/01-2013/12, 20万元, 已结题, 主持。

2.3. 副组长资历情况



姚燕青，女，博士，讲师，硕士生导师。2015年7月毕业于北京航空航天大学，获计算机软件与理论专业博士学位。2012年至2014年国家公派美国纽约大学柯朗数学科学研究所计算机系进行联合培养。现任职于北航计算机学院信息安全与智能信息处理研究室。研究领域为密码学（如基于非完美随机源的密码学、基于格的密码学、弹性泄漏非延展密码编码学）和人工智能的数学基础。近年来发表相关学术论文10余篇，在CRYPTO 2015（美密会，国际密码学领域排名第1，北航第1篇）、Indocrypt 2015、ICITS 2013、Fuzzy Sets and Systems(Q1区)、IET Information Security、Computers & Electrical Engineering等重要国际会议和期刊上发表高质量论文10余篇。主持国家自然科学基金青年基金项目、软件开发环境国家重点实验室自主研究课题、中央高校基本科研项目等。此外，担任多个重要的国际会议和期刊的评审人。讲授计算机学院本科生课程《密码学基础》，研究生课程《信息安全的数学基础》和《密码学》。

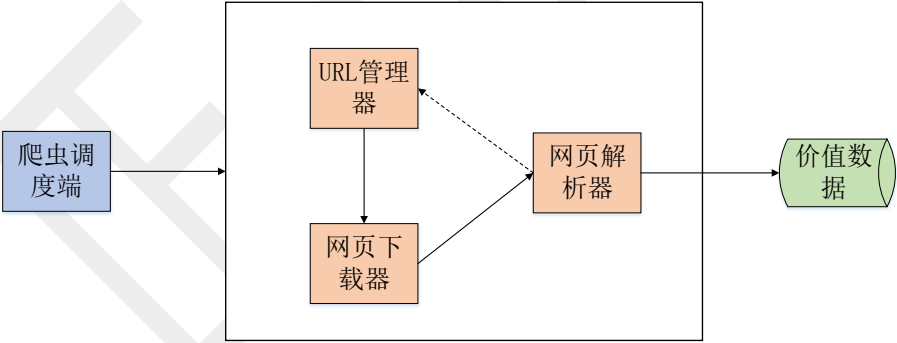
三. 课题情况

3.1. 与课题有关的国内外现状、发展趋势以及专利申请和授权情况

与本课题相关的研究内容包括专家信息自动采集与更新、专家信息抽取、专家领域影响力分析、专家社会关系分析、专家观点分析、专家信息快速检索与可视化。下面分别从这几个角度，介绍国内外研究现状及发展动态。

1、专家信息自动采集与更新

网络爬虫是某种依据事先制定的规则方法，让计算机自动分析并抓取互联网网页信息的程序或脚本。网络爬虫模块如下：



爬虫模块图

目前主流网络爬虫主要依靠 Python 脚本语言编写，采用 urllib2^[1]爬虫技术。urllib2 技术是一个凭借 URL 来抓取对应网页数据的 python 模块，以 urlopen 函数的形式提供便于使用的接口，能够使用不同协议来获取网页信息。

给定一条初始 URL 于调度端，网络爬虫借此种子 URL，通过网页下载器来爬取对应页面的数据，同时解析器^[2]承担从初始网页数据中提取需要的文本信息的任务，然后搜索页面中包含的其他 URL 作为下步将要爬取的页面，存储于 URL 管理器中，最后将解析得到文本信息以结构化的方式存储下来。

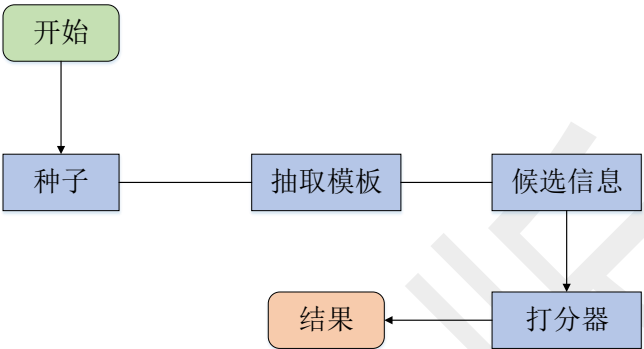
2、专家信息抽取



信息抽取^[3]的定义为从自然语言文本中抽取指定类型的指标、关系、事件等事实信息，并形成结构化数据输出的文本处理技术。

抽取专家信息包括专家的评价指标（人物的主要属性、论文发表情况、学术关系、专利情况、研究兴趣关键词）、描述指标（论文的统计数据、论文的被引用情况、论文的影响因子）和关联指标（论文合作关系、研究兴趣相似关系、共同参与课题的关系、同事关系、同学关系、人物与机构的关系）。

目前主流的抽取方法是给定某一类别的指标实例，从网页中抽取同一类别其他指标实例。基本思路在于种子词与目标词在网页中具有相同或者类似的上下文（包括网页结构和上下文）。因此需要首先利用种子词提取模板，随后利用模板提取更多同类指标。处理该问题的主流框架为：



信息抽取流程图

指标抽取主要包括两种方法：一种是基于规则的抽取，一种是统计的抽取。

基于规则的指标抽取方案^[4]依赖于规则模板，需要在人工总结现有的信息规则的基础上制定模板，优点是简单有效，缺点是非常依赖经验，可扩展性差。

基于统计的指标抽取方案都是以某些机器学习技术为核心，比如分类器、条件随机场、频繁模式挖掘和关联分析等。张巧^[5]等提出一种基于弱监督学习的属性抽取技术，在个人主页中，部分人物指标的相邻位置存在一些触发词，用以描述该指标，在 bootstrapping 方法中，这样的词被称为前导词。可以通过少量的前导词，即种子模式，根据已有信息，持续迭代更新相关指标特征以及模式，以此来发掘更多的模式特征。贾真^[6]等采用了一种基于弱监督学习的属性抽取技术，分类器采用的是最大熵模型。他在传统的 n-gram 模型上，改进成为 n-pattern 模型，两者区别在于，n-pattern 中的项不一定连续，缓解了部分数据稀疏的问题。

3、 专家领域影响力分析

著名的 Google PageRank^[7]算法是一种根据网页之间相互的超链接数量来进行网页排名技术。PageRank 算法最初仅是应用在搜索引擎中，用来计算网页排名，但随后，部分学者将 PageRank 算法引用到社交网络中，认为 PageRank 算法是社交网络用户个体影响力分析的基础算法^[8,9]。

在学术领域，经典的科研专家的学术影响力评价指标包括文章引用次数、H 指数、I10 指数、G 指数等。随着 Web2.0 技术及社交网络的发展，这些评价体系的权威性受到了冲击，Priem 等提出一种多元化科研人员评价体系 Altmetrics^[10]，使用更多的社交网络参数来进行学术声望评判。Altmetric 评分根据不同社会化媒体引用文献的次数，一般包括新闻报纸、博客、微博，论坛等。

基于社交网络的用户影响力分析模型，国内外各大学术社交网络中针对其科研用户也推



出各类影响力排名算法,由清华大学唐杰等人开发的 Aminer^[11]研究人员社会网络,通过分别统计科研人员的文章数量、引用数量、H 指数、A 指数、G 指数等信息,生成专家统计信息雷达图,并可分别通过上述指数进行专家排名。

4、专家社会关系分析

关系抽取是信息抽取研究中的一个重要环节,它的作用是抽取两个实体之间的语义关系。实体关系抽取是篇章理解的关键技术,同时对于信息抽取、问答系统、机器翻译等都有着非常积极的意义。

现有主流的实体关系抽取技术分为有监督的学习方法、半监督的学习方法和无监督的学习方法三种:

1)、无监督的关系抽取方法最早是由 Hasegawa^[12]在 2004 年的 ACL 会议上提出的。无监督的学习方法假设拥有相同语义关系的实体对拥有相似的上下文信息。因此可以利用每个实体对对应上下文信息来代表该实体对的语义关系,并对所有实体对的语义关系进行聚类。

2)、有监督的学习方法将关系抽取任务当做分类问题,根据训练数据设计有效的特征,从而学习各种分类模型,然后使用训练好的分类器预测关系。该方法的问题在于需要大量的人工标注训练语料,而语料标注工作通常非常耗时耗力。

3)、半监督的学习方法主要采用 Bootstrapping 进行关系抽取。对于要抽取的关系,该方法首先手工设定若干种子实例,然后迭代地从数据中抽取关系对应的关系模板和更多的实例。

总体而言,无监督方法不需要人工的干预指导,具有良好的领域移植性,在处理大规模开放领域数据时具有其他方法无法比拟的优势。但当前无监督方法判别实例的相似性仅用到实体对的上下文特征,对其他包含语义信息的特征未加利用,同时聚类方法的选择具有盲目性,致使无监督方法的准确率和召回率都比较低。

半监督方法结合了有监督和无监督两种方法的优势,不需要过多的训练语料,并且可以发现种子集中不包含的关系实例,因此成为关系抽取研究的热点。很多学者将在其他领域表现不俗的算法或思想引入到弱监督方法中,使其更加完善,如 Enrique^[13]引入主题模型 (Topic Model) 来计算模式隶属于某类关系的概率, Mike Mintz^[14]尝试使用知识库 FreeBase 来代替手工选取种子集等。但弱监督方法的两个主要问题:如何减少选取种子集的人工干预以及如何控制扩展过程中的噪声并没有完全解决,这也限制了该方法的性能。

与其他两种方法相比,有监督的学习方法能够抽取更有效的特征,其准确率和召回率都更高。因此有监督的学习方法受到了越来越多学者的关注,本课题也将采用有监督的学习方法进行研究。监督学习的方法解决实体关系抽取问题主要采用指导性机器学习方法,主流的指导性学习方法有基于特征向量的方法^[15,16]和基于核函数的方法。其中,基于核函数的方法,直接以结构树为处理对象来计算它们之间的相似度,再使用支持核函数的分类器进行关系抽取。然而,基于核函数方法的一个致命的缺点是训练和预测的速度太慢,不适于处理大量的数据。基于特征向量的方法^[17,18]是将关系样例进行特征抽取并将其表示为特征向量,然后通过机器学习的方法来训练关系实例,该方法能够取得比较好的效果。本课题也是拟采用基于特征向量的方法来实现文本中的关系抽取。

5、专家观点分析

观点信息抽取,又称为情感信息抽取,旨在抽取情感文本中有价值的情感信息,一般包括评价词语、评价对象、观点持有者、评价短语等信息,它是情感倾向分析和观点检索的基础。现有的观点抽取主要包括基于统计的方法、基于机器学习和基于图模型的方法。

基于统计的方法统计学的方法往往依赖于文章的表层特征。比如依据句子在段落中的位



置,段落在文章中的位置,词频与逆文本词频的大小,句子与标题的相似度等特征评估句子的重要性。虽然原理看似简单,但实现结果却有很高的准确率,甚至超过后来很多更复杂的算法。近几年学术界进一步提出了基于整数线性规划的方法以及次模函数最大化的方法,可以在句子选择的过程中同时考虑句子冗余性。

在统计方法基础上发展的机器学习方法比如决策树模型、隐马尔科夫模型,条件随机场模型,神经网络等等。例如 Conroy 基于隐马尔科夫模型对词汇之间的关联性,基于相互的依赖关系进行计算。机器学习方法主要关注如何将文本摘要问题转化为机器学习问题,其主要策略是选取文本特征、确定使用的机器学习方法以及在语料库上进行学习训练,最终得到模型。

基于图模型的方法基于图模型的方法的基本思路是把文章的句子或者段落作为一个分析对象,每一个分析对象作为图中的一个点,点与点之间的关系通过寻找两个分析对象是否在某个特征上相似或者重合来确定是否连接,如果判断有关系,则在两点之间添加一条边。建立完基础的图之后,通过图模型上的迭代算法(Page Rank, Text Rank, HITS)来迭代计算图中各个节点的权值,按照权值的大小排序之后选择权重大的分析对象(句子或者段落)作为结果。

相对于国外文本自动分类的研究而言,国内对于文本摘要的研究起步较晚,上世纪 80 年代,上海交通大学王永成教授最早开始研究自动文本摘要技术,1997 年开发出了中文文献自动摘要系统。其关键技术有三:构造关键词词典;从文献的有关部位中选择包含关键词句子作为候选文摘句,考虑候选文摘句中关键词数、距离大小、句子的位置特征等特征,计算的到句子的权重,从大到小排序后,按照文摘篇幅的大小来选择权值相对较大的句子作为摘要句;而后对摘要句句进行通顺化处理得到最终的文摘。目前,性能较好的包括多种基于图模型的自动文本摘要方法,另外还有压缩式摘要方法,在比较式摘要、跨语言摘要、演化式摘要等多类型摘要任务上有长足进展。在学术文献摘要方面,则分别提出基于有监督学习和整数线性规划模型的演示幻灯片的自动生成方法。

观点倾向分析的主要目标是识别目标对象所表达的观点总体倾向,并预测观点的类型。主要包括基于词典或规则的方法和基于监督学习的方法。

基于词典或规则的文本情绪分析方法是一种人工参与的方法,其基本思想是按照人工构建的情绪词典或者情绪转移或组合规则对文本的情绪类别进行标注。该类方法基本上是在情绪词典的基础上,构建合理的约束(如情感绑定、主题相关性等),以提高对具体领域情感分析的能力。如文献^[19]提出基于组合范畴语法的方法来分析文本情绪。通过总结情绪类别与词汇间对应关系,依据上下文转移规则,建立情感类别转移规则和对应的词汇库。

采用基于监督学习的方法,可以让机器自己学习模型参数,判断观点的基本情绪,如“喜悦”、“悲伤”、“惊奇”等,不需要人工手动构建和更新规则库,降低了维护成本。其基本思路是对特定领域的基本情绪类型进行规定,并在已有的情绪语料库的基础之上,依据情感词汇的数量、极性等信息提取语料特征,并选择合适的分类算法如 SVM、朴素贝叶斯等,训练分类模型。

特征提取是观点倾向分析中最重要的一环,好的特征可以充分反映出文本的内在结构,是提高识别准确率的重要基石。除了情感词汇等特征外,一些其他特征如颜色、表情符号(Emoticons)、话题标签等也可能会反映并影响人的情绪。如文献^[20]构建“颜色-情绪-主题”词典,并证明在该词典的基础上,考虑情绪和颜色间的相关性对新闻评论、微博和儿童故事的情绪分析有一定的作用。文献[21]和文献[22]分别将表情符号、话题标签特征考虑到网络舆情平台的情绪分析中,取得了较好地预测效果。



6、 专家信息快速检索与可视化

目前中文知识检索策略的研究主要集中在：知识库的站点定位，动态约束性概念网络，知识检索相关技术（如本体论、多 Agent 技术、数据库自由标引、网络数据挖掘、动态约束性概念网络技术），从知识检索的主体角度改善知识检索的不足。如金燕等分析目前 Web 检索中存在的问题，提出本体论驱动的知识检索方法。王娜从知识检索过程中涉及的用户主体的角度提出了基于多智能主体协作的知识检索，从用户知识、专家知识、领域知识角度提出了基于智能主体的多维学习方法^[23]。肖红琳指出实现全文数据库自由标引是实现高效知识检索的重要方法，并从构建知识概念库、引入智能搜索引擎技术机制、动态约束概念网络、知识组织方法以及本体对全文数据库自由标引方法进行了探讨^[24]。金燕等分析了网络数据挖掘的技术在 Web 知识检索中的应用：挖掘网络信息源，形成知识库；挖掘用户（普通用户和专家）在访问 Web 时留在服务器上的记录，发现用户知识检索的规律^[25]。陈芳将网络信息资源的知识检索分为找知识站点库和提取相关知识条两步，并提出了网络信息资源知识检索的实现方法：加权关键词匹配定位知识站点；从定位的站点中利用动态约束性概念网络找出具体知识。

可视化展示的研究在工程技术上已经相当成熟，很多现有的插件都提供了非常漂亮的可视化实现方法，如 Highlight 以及百度公司的 echarts 等。本课题中设计的可视化展示都可直接使用这些插件完成漂亮的可视化展示。

- [1] Tinh T T M, Tin C N T, Mạng Q T, et al. Dive Into Python-Chapter 11. HTTP Web Services[J]. Tailieu Vn.
- [2] 陈晓慧, 陈荣国, 卫文学. 基于网络爬虫的 Web 服务抓取解析器的设计与实现[J]. 地理信息世界, 2010, 08(3):64-68.
- [3] 张丙奇, 姜吉发. 企业相关信息抽取技术与系统实现[J]. 微电子学与计算机, 2004, 21(1):1-6.
- [4] 丁君军, 郑彦宁, 化柏林. 国内外属性抽取研究综述[J]. 情报科学, 2011(5):793-796.
- [5] 张巧, 熊锦华, 程学旗. 基于弱监督学习的主页人物属性抽取方法[J]. 山西大学学报(自然科学版), 2015, 38(1).
- [6] 贾真, 杨燕, 何大可. 基于弱监督学习的中文百科数据属性抽取[J]. 电子科技大学学报, 2014, 43(5).
- [7] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: Bringing order to the web[R]. Stanford InfoLab, 1999.
- [8] Tunkelang D. A twitter analog to pagerank[J]. The Noisy Channel, 2009.
- [9] Haveliwala T, Kamvar S, Jeh G. An analytical comparison of approaches to personalizing pagerank[R]. Stanford, 2003.
- [10] Bhue S, Singh K, Biswal S K. Altmetrics: Article Level Metrics Makes Easy for User Community[J]. 2016.
- [11] 唐杰. AMiner[EB/OL]. (2006-09-06)<http://aminer.org>.
- [12] Hasegawa. T, Sekine. S, Grishman. R. Discovering Relations among Named Entities from Large Corpora[C]. In Proc. of ACL-2004, 2004: 415-422.
- [13] Enrique Alfonseca, Katja Filippova, Jean-Yves, et al. Pattern Learning for Relation Extraction with a Hierarchical Topic Model[C]. Proceedings



of the 50th Annual Meeting of the Association for Computational Linguistics, 54-59.

[14] Mike Mintz, Steven Bills, Rion Snow, et al. Distant supervision for relation extraction without labeled data[C]. Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, 1003-1011.

[15] 毛小丽. 基于特征向量的实体间语义关系抽取研究[D]. 重庆: 重庆大学, 2011.

[16] 黄鑫. 基于特征向量的中文实体间语义关系抽取研究[D]. 苏州: 苏州大学, 2009.

[17] TYMOSHENKO K, GIULIANO C. Semantic relation extraction using Cyc [C] //Proc of the 5th International Workshop on Semantic Evaluation. Stroudsburg, PA: Association for Computational Linguistics, 2010:214-217.

[18] GIULIANO C, LAVELLI A, PIGHIN D, et al. Kernel methods for semantic relation extraction [C] //Proc of the 4th International Workshop on Semantic Evaluations. Stroudsburg, PA: Association for Computational Linguistics, 07:141-14

[19] Smith P, Lee M. A CCG-Based Approach to Fine-Grained Sentiment Analysis in Microtext[J]. Aaai Spring Symposium, 2013.

[20] Volkova S, Dolan W B, Wilson T. CLex: a lexicon for exploring color, concept and emotion associations in language[C]// Conference of the European Chapter of the Association for Computational Linguistics. 2012:306-314.

[21] Purver M, Battersby S. Experimenting with Distant Supervision for Emotion Classification[C]// Conference of the European Chapter of the Association for Computational Linguistics. 2012:482-491.

[22] Mohammad S M, Kiritchenko S. Using Hashtags to Capture Fine Emotion Categories from Tweets[J]. Computational Intelligence, 2015, 31(2):301-326.

[23] 王娜. 基于多智能主体协作的知识检索. 现代情报, 2007(12) : 152-155.

[24] 肖红琳. 论全文数据库自由标引和知识检索. 内蒙古科技与经济 2006(13) : 77 - 79.

[25] 金燕, 张玉峰. 网络数据挖掘及其在面向 Web 的知识检索中的应. 现代图书情报技术, 2003(6) : 55 - 57.

3. 2. 课题的研究意义（包括必要性和可行性）

随着以计算机网络、人工智能为代表的信息技术的迅猛发展, 知识这种无形的资源在各个领域显现出巨大的活力, 成为推动经济增长的重要生产要素。社会已经从工业经济时代发展到了知识经济时代。个人和组织在长时间的工作实践中, 积累了海量的知识资源, 如何合理地开发与利用这些知识, 为全社会创造出尽可能多的价值成为一个值得研究的课题。

国务院颁布的《国家中长期人才发展规划纲要》中明确指出, 为了加强人才工作基础性建设, 推荐人才工作信息化建设, 必须建立人才信息网络和数据库, 为此政府需要积极支持社会各方力量建设完善面向市场的各类专业人才数据库和人才管理系统。同时, 纲要还提出,



在体制机制创新中必须创新人才评价发现机制,要完善人才评价标准,注重靠实践和贡献评价人才,将评价人才和发现人才相结合。领域专家库系统的建设是符合我国人才发展战略规划种“建设人才信息网络,评价人才与发现人才相结合”指导思想的重要应用研究。专家间的各种关联关系能够对相关领域专家资源的深度聚合与可视化揭示,能够为科研工作和科研工作和科研管理部门提供相应的信息服务和决策支持。领域专家库系统的建成有助于各领域研究者加深对同行的了解,也有助于科研管理部门和社会大众把握各领域研究人员分布情况和遴选相关领域的专家,能为各科研机构 and 高校的人才选拔,招聘以及项目团队组建提供定量依据。

必要性: 在网络和数据库技术迅速发展的今天,建设一个专家库管理系统是非常有必要的。(1) 易用性: 以档案、资料等形式手工管理人才库,不但工作量大,而且极易出现差错。系统的建立,提高了专家人才管理的效率和精度,有利于人才储备。(2) 时效性: 当发生涉及自然灾害、事故灾难、公共卫生、社会安全等方面的重大突发公共事件时,政府决策部门需要在第一时间履行应急预案,并迅速找到相关领域专家,商讨对策。通过信息系统的使用,职能部门可以按图索骥,找到专家,避免贻误处置相关事件的时机。(3) 开放性: 不论是在平时工作、生活中,还是在遇到突发事件时,社会公众都需要一个和专家交流的信息平台,以解决各种切身实际问题。专家应发挥其专业咨询作用,面向全社会宣传科学知识,提出专业意见,保证信息的通畅。

可行性: 专家是在某一领域积累了大量理论知识或实践技能的人。俗话说:“隔行如隔山”,由于行业的细化,人们不可能深入了解其他领域的各种专业技能,但很多现实问题的复杂程度超乎想象,并非集中于某一特定领域,而是需要综合运用多学科、多行业的知识加以解决。互联网中存在包括专家在内的海量知识信息,这些信息大多以非结构化,零散的形式存在,没有一个统一的,可供快速查阅,了解专家科研动态的系统。当前,伴随着以计算机网络、人工智能为代表的信息技术的迅猛发前,大数据,分布式计算,数据挖掘技术的技术解决了许多以前遗留下来的问题,这些都为建立一个统一的专家库系统提供了可能。专家库系统的建立将为业内外人士,政府提供更丰富,详实的科研人才信息,为科研管理部门提供更多的信息服务和决策支持。

3.3. 课题的研究内容、拟采取的研究方法和技术路线

本课题拟收集整理各行业专家信息,建设国内行业专家系统,除了需要涵盖专家的专业背景、学术和行业影响力等数据外,更侧重于分析专家的社会关系、参与的活动、参加的组织、观点、倾向性等,为调研专家背景提供数据支撑和决策辅助。具体研究目标与内容以及采取的研究方法和技术路线如下:

1、 研究目标与内容

(1) 专家信息自动采集与更新

针对业务需求,在网上寻找包含专家信息的数据来源,并开发相应的网络爬虫工具,能够自动采集专家信息,数据来源包括但不限于大学和研究所得官方主页、专家个人主页、百度百科、学术文献网站以及其它发布专家信息的系统等。专家信息需要周期性的更新于扩充,保证专家库的动态性与时效性。

(2) 专家信息抽取

针对爬取到的人物网页信息,利用信息抽取技术自动抽取专家的各项信息,包括:人物的基本属性(包括但不限于姓名、性别、年龄、籍贯、单位、职称、职务、从事行业等)、



论文发表情况、学术关系、专利情况、研究兴趣关键词等。

(3) 专家领域影响力分析

识别给定领域的主要专家及其研究内容,并构建专家影响力评价指标,对专家在领域内影响力进行评价,并对影响力评价方法的科学性进行解释说明。

(4) 专家社会关系分析

针对专家的过往经历、参与的活动、参加的组织等信息对专家社会关系进行分析。如论文合作关系、相似研究兴趣关系、共同参与课题的关系、同事关系、同学关系、人物与机构关系等。通过挖掘这些信息,构建专家的社会关系网络,对专家的领域和群体关系进行社区分析,对专家之间形成的小群体关系进行发现识别,并挖掘网络中的核心人物。同时,能够对专家之间的关系进行分析,发现不同专家之间潜在的联系关系。

(5) 专家观点分析

挖掘设计专家的新闻报道、言论著作等信息,从中抽取专家的观点,专家观点主要包括专家的学术观点、主观意见、社会事件参与评价等等,据此构建专家的观点库,同事,针对专家的社会关系,对专家联系较多的其它人员进行挖掘,分析其关系圈,以及关系圈的观点,对专家观点库进行补充。在专家观点库基础上,分析专家对特定领域的观点倾向性。

(6) 专家信息快速检索与可视化

提供专家库检索功能,通过关键词对专家信息进行检索查询,并返回查询结果,支持按行业、研究内容、人名、机构名等字段进行检索的功能,支持多个字段组合检索名,支持根据专家关系进行检索。支持选定部分专家作为重点关注对象,并可进行分组管理,以跟踪订阅形式对专家的信息、观点等持续追踪。根据专家的各项属性信息、统计信息、关系网络、检索与挖掘分析结果进行可视化展示。

2、研究方法与技术路线

(1) 专家信息自动采集与更新

①收集专家个人信息主页:考虑到不同领域的专家个人信息页面的格式大多有所不同,需要预先收集统一展示该领域全部或部分专家个人信息的网页地址,使得网络爬虫能从一个统一的入口进入,依次爬取每个专家的个人信息。

②编写网络爬虫规则:爬虫定位至具体某个专家的个人信息页面时,预先分析该页面的网页 Html 源代码,在爬虫解析器部分编写相应的解析函数,使得爬虫能够自动从夹杂 html 标签和无关噪音数据的初始网页数据中,提取出包含专家个人信息的文本数据。

③结构化存储专家信息:爬虫解析器还承担着将整个包含专家个人信息的文本数据,粗略切割成包含不同指标的分段。利用一定的匹配规则,预先按不同指标分割成不同的语句段,并保存在 MySQL 数据库中。最后将爬取该专家个人信息的原网页地址 URL 一并保存下来。

④更新专家信息:由于预先存储了之前爬取专家个人信息的网页 URL 地址,设定一个固定间隔时间段,启动相应的爬虫,重新爬取一遍该网页,重复前三个步骤,以此更新专家个人信息。

(2) 专家信息抽取

①获取源语料并清洗:利用网络爬虫自动获取相关人物的个人简历,相关信息等。对粗糙的网页数据,预处理建立匹配指标模板,粗略筛选出包含主要信息的文本,主要目的是为了去除网页内部的隐含的噪音,例如广告等。然后将初始文本数据存储下来。

②标注训练语料:人工标注训练语料工作量太大,研究自动化方案。从分利用网页中结构化的信息框中的属指标信息,借此对文本进行标注。步骤为:搜寻信息框中是否包含某个指标,如果这个指标出现在了文本中的某个句子里,并且这个句子也包含了人名,那么这条

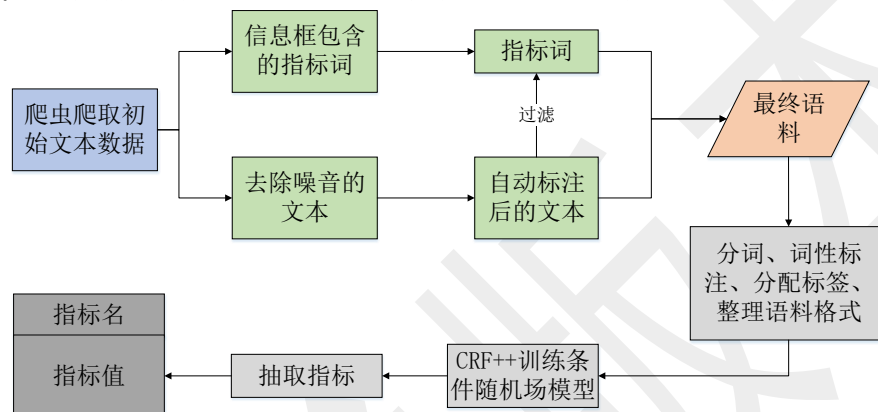


句子就是一条符合条件的训练语料。

③标注语料质量提升：为了进一步提升自动标注的语料的质量，研究基于熵的特征排序技术，设计过滤函数提取目的词，最后基于目的词对自动标注的语料进行过滤，以此筛去那些没有描述指标信息却被标注出来的语句。

⑤利用条件随机场训练抽取模型：把从非结构化文本中抽取专家人物指标这一问题，当作序列标注任务来求解，即实际上是使模型学习对特定人物指标的触发词和指标属性边界的相关特征。本文将使用条件随机场（CRF，Conditional Random Field Algorithm）技术来训练模型，并自动抽取专家人物指标。

⑥存储抽取的专家人物指标：将抽取好的专家人物指标以结构化的方式存储。本文将利用 MySQL 数据库来存储结构化的指标数据，利于后续研究时查阅和使用。



信息抽取技术路线图

(3) 专家领域影响力分析

①将专家影响力的影响因素分为学术影响参数 A 与社交网络影响参数 S 两部分。

②将专家学术声誉计算参数定义为 A。科研领域学术影响因子确定为基本影响参数与合作者影响参数两部分。基本影响参数引入专家的 H 指数、总影响因子和所发文章总数三个参数。基本影响参数的具体公式定义为：

$$A_b(i) = \frac{H(i) * I(i)}{P(i)}$$

其中 H 为 H 指数，I 表示影响因子总和，P 表示作者所发文章数量。该公式表示，如果在发的论文总数相同的情况下，作者的 H 指数和总影响因子越高，说明该专家的科研声望值越高。

同时将合作者影响力参数引入到专家影响力指标中，认为文章合作者的影响力会对专家的影响力有较大的影响。最终确定公式为：

$$A(i) = A_b(i) + \sum_{k=1}^H \left(\sum_{j \in U(i)} \frac{U(j)}{L(j)} \right)$$

式中 U 表示合作专家的影响力值，L 表示该专家在文章中的署名位置，第一作者为 1，第二作者为 2，以此类推。

③引入 PageRank 算法进行社交网络专家影响力分析。

根据 PageRank 公式，模型中的 S 参数定义如下：

对于不同社交平台（如微博、公众号、领英等），首先计算专家的粉丝关注情况值



$$R(i) = c \sum_{j \in B(i)} \frac{R(j)}{F(j) + \lambda}$$

R 表示要计算的专家粉丝关注情况值, B 为该专家的粉丝数(被关注数), F 表示关注数, c 为常数。该模型认为拥有越多高影响力粉丝的专家, 那么该专家的学术影响力值也就越高。考虑到专家在社交平台发布消息的互动情况, 计算

$$T(i) = \sum_{j \in C(i)} \frac{G(j)}{N(j) + \lambda}$$

T 表示某专家的用户评论情况值, G 表示好评数, N 表示差评数, C 表示所有评论数。该公式指出, 用户对某专家的好评越多, 该专家的影响力越高; 差评越多, 该专家的影响力越小。

确定社交网络影响参数为专家关注情况与评论情况之和:

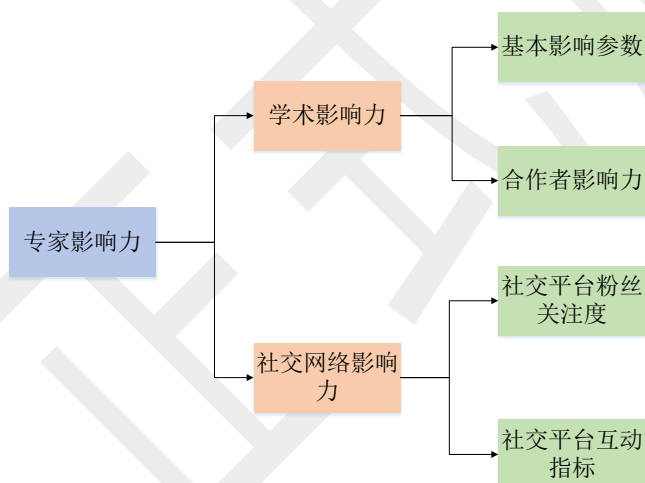
$$S(i) = R(i) + T(i)$$

④定义专家影响力模型 AS 公式为:

$$AS(i) = aA(i) + bS(i)$$

AS 表示专家影响力评价价值, S 表示专家社交网络影响参数值, A 表示学术影响参数值。a 与 b 表示两类影响因子权值, 且 $a+b=1$ 。

基于社交网络的专家影响力的具体计算模型如下:



专家影响力评价模型

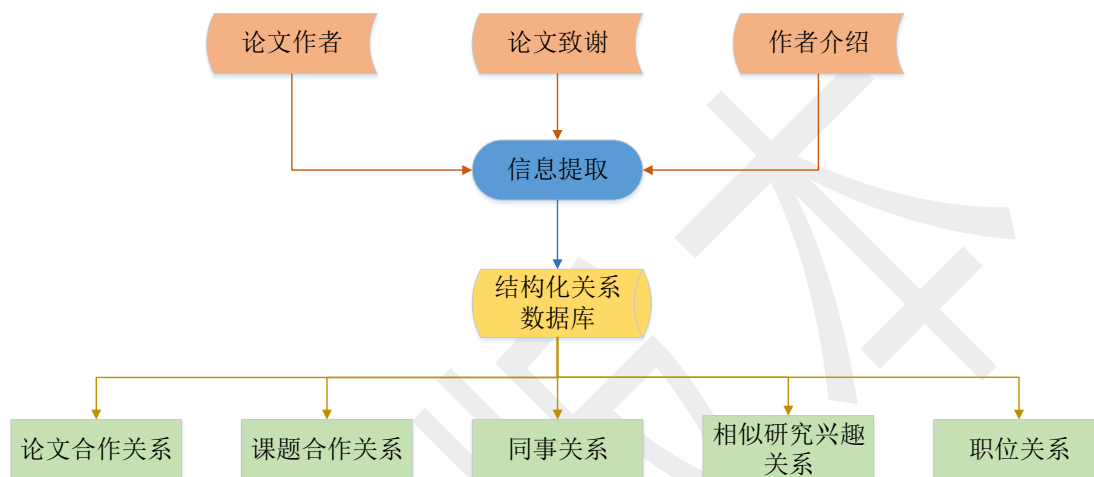
(4) 专家社会关系分析

专家社会关系主要指专家之间的论文合作关系、课题合作关系、同事关系、相似研究兴趣关系以及职位关系等。这些关系信息直接或间接地存在于论文作者、论文致谢、作者介绍、研究方向等数据信息中。通过挖掘并提取相应信息用于构建结构化的知识库, 利用结构化的知识库对专家间的社会关系进行分析。

论文作者一般包含多个人名, 这些人名直接反映出了专家之间的合作关系。另外, 有些专家之间可能存在通过第三方作者相互合作的现象, 这种间接合作的合作关系可通过构建专



家关系网络,利用多层网络结构侧面反映专家间的间接合作关系。论文致谢中包含了论文课题的研究来源,其中的课题编号尤为重要,可以利用课题编号编制网络,出现在同一个课题编号网络中的专家视为课题合作关系。期刊论文中一般包含着作者简介,其中包含与专家相关的,丰富的信息,包括专家所在单位、研究兴趣等。利用关键词匹配,触发词或者句法分析等知识抽取技术方法对这些信息进行抽取,构造结构化的数据信息存于数据库。分别利用研究方向,单位等信息建立专家网络。在同一个研究方向网络中的专家视为具有相似的研究兴趣关系。在同一个单位网络中的专家视为具有同事关系信息。



专家社会关系分析路线图

对作者简介部分的处理,主要采用信息抽取技术识别其中的研究方向,工作单位等信息。由于系统关注的信息较少,拟直接采用传统基于触发词,基于规则或者依存句法分析的方法。这些方法在面对分类较少的情况下往往表现出相对不错的性能,完全可适用于系统应用中。

(5) 专家观点分析

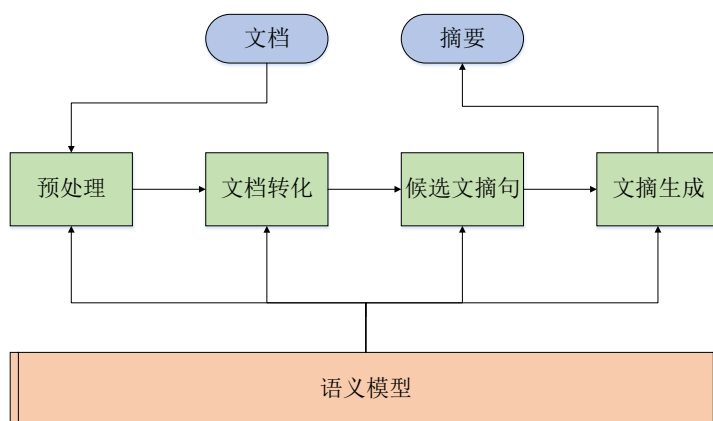
观点抽取主要包括自然语言语料中的人物识别和文本摘要两部分。

对于人物识别,命名实体识别(Named Entity Recognition, NER)的主要任务是识别出文本中的人名、地名等专有名称和有意义的时间、日期等数量短语并加以归类。将识别出的人物与专家数据库中的物作比对,确定所识别人物的身份信息。若数据库中存在重名,则可以根据地名等信息唯一确定身份。否则从语言语料中去除。

对于文本摘要,基本思路为:在语义模型的基础之上进行文本摘要的分析,将预处理后的文档结合语义空间的特点,转换成一个数学描述的模型;同样依据每个基本分析单位在语义空间上的关联建立联系并计算其重要程度,从而选择出能够代表文档的基本分析单位作为候选,最后结合文本摘要任务的其他要求比如篇幅、冗余信息、连贯性等特点生成最终的文本摘要作为输出。

具体来说,假设文档D中句子集合为S,需要根据算法确定每个句子的权值大小,根据选择的特征(基于选择句子所含词语构建统计学特征,基于主题模型方法选择句子所含主题特征)来计算句子的权值大小。然后确定文章主要表达几个思想,其次确定每个句子包含多少某个思想的词语或者主题,计算出文章表达几个思想分别占有多大的比重,提取到这些特征之后,再来确定每个句子的权重,句子权重由两点决定:1)这个句子表达了几个思想,2)表达的思想在文章中所占的比重是多大。

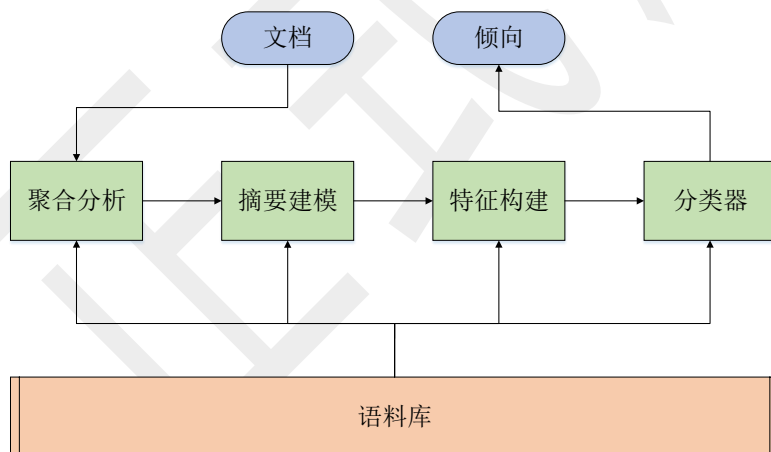




观点抽取模型图

文摘算法中的另一个问题是对于冗余信息的排除。对于文档 D 的两个句子 S_i 、 S_j 他们表达的意思基本相同，包含的思想个数相同，但是长度不同。文摘在保证质量（包含尽可能多的文章思想）的同时，还是考虑到对于冗余信息的去除，以及文摘易读性。所以需要过长或者过短的句子都进行惩罚，一个可行的方法是首先求出文章中所有句子长度的中值（不选择均值的原因是为了避免过长句和过短句的影响），以此确定文摘句长度基准，然后对过长句和过短句进行惩罚，影响到句子的权值大小。

对于观点倾向分析，其基本思路为：在语料库的基础上，对专家观点形成的摘要在领域、话题等维度进行聚合分析，过滤重复观点，对某一特定主题表述出清晰明确的情感倾向。结合背景知识，对摘要信息建立数学模型，描述词汇句子在向量空间的表示。然后，抽取合适的特征集合如关系对、信息增益、点互信息等来刻画文本，使得不同类别的文本特征集具有类内同质，类间异质的特点。选择已知成熟的分类算法对观点类别构建分类模型，输出情感类别。



观点倾向分析模型图

特征构建过程中只考虑二元关系<评价词汇，评价对象>是不够的。意见和评论往往具有一定的场景限制，这样的条件限制使评论变得准确和丰富。因此，抽取类似的条件信息，能够从整体上看待观点倾向性，更有利于进一步挖掘观点的倾向性信息。本课题用图来表示倾向性信息抽取的结果，如用结点表示评价词汇和评价对象，有向边刻画倾向条件及程度等。利用图的结构化等特点设计相应的编码器，将图信息转化成相应的特征向量。

(6) 专家信息快速检索与可视化

本课题中的知识检索主要从知识库中检索出关系抽取构建的三元组。知识检索策略的研



究主要集中在：知识库的站点定位，动态约束性概念网络，知识检索相关技术（如本体论、多 Agent 技术、数据库自由标引、网络数据挖掘、动态约束对外性概念网络技术），从知识检索的主体角度改善知识检索的不足。明确关系的取值和属性，整合人物对和关系的信息行为三元组表达式。当文本与查询的三元组表达式中已知部分匹配一致时，则得到查询未知部分的一个取值。由此可以实现类似 QA 功能又可以完成精确的信息检索。可视化展示主要以图表的形式直观地展示关系三元组。现有的可视化技术已经相当成熟，展示的方式也是多种多样。本课题中关系三元组可视化展示主要以百度 echats 中的力导向图为基础，根据实际需求进行更改。

3.4. 课题应达到的技术水平、技术指标和提交的成果形式

■ 功能指标

- 1) 构建领域专家知识库，包括专家基本属性库，科研论文库。
- 2) 实现领域专家属性信息抽取，专家观点分析，专家信息检索，论文相关性推荐等功能
- 3) 专家社会关系分析：
 - 收集领域专家之间的潜在关联信息，包括研究领域、科研机构、论文合作发表信息、专利合作、项目合作信息等，构建专家社会关系网络。
 - 利用潜在关联信息进行专家关联度计算，利用关联度排序算法对专家关联度进行分析，发现不同专家之前潜在的联系信息。
- 4) 专家科研绩效评价：
 - 构建专家科研发表论文数量，被引用程度，被关注程度，高被引论文数，影响因子数，成果利用率等信息。
 - 利用相应的评价指标对专家科研及影响力进行评估和排名。
- 5) 可视化的网页 Web 领域专家系统一套：包含 2)、3) 和 4) 的可视化展示。

■ 性能指标

- 1) 专家数据手机规模
基本要求指标：专家知识库中专家数量不低于一万级别，包含专家论文数据量信息不低于十万级别。
- 2) 专家信息抽取准确率
基本要求指标：专家基本属性信息抽取准确率不低于 60%。
- 3) 专家信息检索速度
基本要求指标：百万级记录检索响应速度不高于 3 秒。
- 4) 专家信息更新速度
基本要求指标：专家信息至少每个月自动检测更新一次，并支持通过人工指定部分专家进行直接更新。

■ 提交成果形式

序号	交付物类型	交付物名称及说明	数量	备注
1	软件系统	领域专家库系统一套，主要包含以下模块：	1套	需提交独立运行系统或程序模块



	(定制)	<ul style="list-style-type: none"> ● 数据爬虫模块 ● 数据处理模块 ● 数据挖掘与分析模块 ● 数据可视化模块 		和相应的源代码,如果是程序模块,还要提交与宿主系统的兼容测试报告
2	专用设备 (研制)			
3	专利	领域专家库构建相关专利	1	需提交专利受理通知书
4	软著			需提交软著登记证书
5	研究报告	领域专家库构建技术报告测试报告		可根据实际情况提出研究报告成果要求
6	验收文档	按照中心统一要求提交全套验收文档,例如:验收申请书、研制总结报告、需求分析说明书、概要设计说明书、详细设计说明书、测试方案、自测报告、验收测试报告或成果试用效果证明、用户手册、涉密科研课题保密工作自查表(保密课题提供)等。	1套	按照中心统一要求进行打印、胶装、签字、盖章。
7	课题成果电子版	包含所有课题成果的电子版,例如软件系统的源代码、全套验收文档、验收答辩 PPT 等。	1套	以光盘形式提交
8	其它			

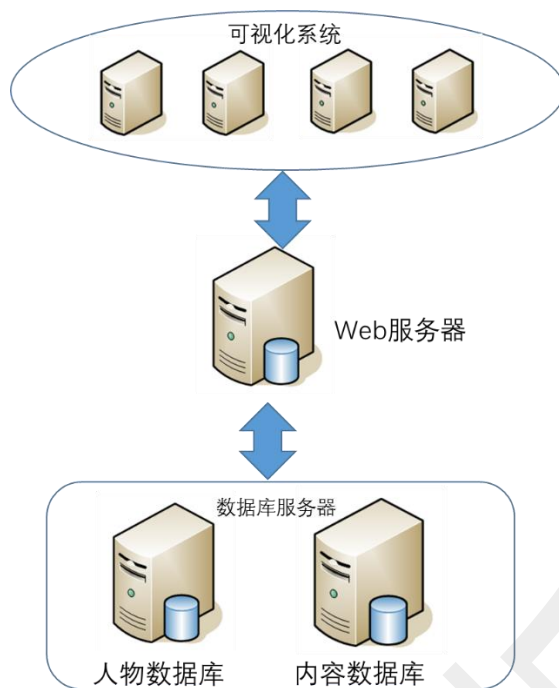
3.5. 初步的验收测试环境（给出网络拓扑图、各节点机的功能）

1). 验收测试环境的硬/软件配置清单

序号	类别 (软件/硬件)	功能	数量	是否移交给合同签订委托方 (是/否/委托方提供)
1	硬件	数据库服务器	1	委托方提供
2	硬件	专家分析服务器	1	委托方提供
3	硬件	可视化系统	1	委托方提供

2). 验收测试环境的网络拓扑图、各节点机的功能描述





根据课题的研究目标及内容，验收测试环境网络拓扑应主要分为三类节点机：

- 1) 数据库服务器节点：主要负责领域专家基本属性信息、及专家相关的数据的存储，对专家的个人信息、关系信息、记忆发表文章数据信息和专家言论观点信息等数据进行存储。
- 2) Web 服务器节点：领域专家系统前端界面可视化展示；通过获取数据库服务器的数据，对数据进行属性抽取，观点抽取，检索排序，关联聚合分析，科研绩效评价处理等得到的数据进行可视化展示。
- 3) 可视化系统：对分析和处理结果进行可视化展示，包括检索结果、观点分析结果、关系分析结果等的可视化。

3.6. 进度安排与考核指标

时间点	内容	考核指标
2018.4 — 2018.5	需求调研与分析	调研课题研究内容，国内外研究现状，研究可行性，明确需求
2018.6 — 2018.12	开发与上线调试	完成数据库构建，信息抽取，挖掘算法实现，系统功能开发，可视化展示
2019.1 — 2019.2	系统优化	优化系统各项功能，性能
2019.3 — 2019.4	测试与用户试用	系统单元测试，功能测试，性能测试



3.7. 单位情况

3.7.1. 现有工作基础（包括所在单位已具备的人力、物力基础）

（1）承担单位情况及研究基地

本课题的承担单位为北京航空航天大学计算机学院，计算机学院拥有中国科学院院士 2 名，教授 30 多名，拥有计算机科学与技术一级学科博士点授予权，拥有计算机科学与技术国家一级重点学科。建有软件开发环境国家重点实验室、北京市计算机新技术重点实验室、北京市网络技术重点实验室，拥有先进的网络、计算、存储和图书情报等方面的基础设施，具有良好的实验条件和环境保障。

课题组所在的信息安全与智能信息处理研究室拥有多名青年教师和博士后，15 名博士生，20 余名硕士生，拥有共 150 平米的 5 个实验室，共有 40 多台微机，Dell 高档服务器和 Dell 的 PowerVault MD3000i 磁盘阵列各一台。课题组所在的软件开发环境国家重点实验室也具有有良好的实验条件和环境。

（2）课题组技术基础

本项目的学术骨干承担了多项自然科学基金和北京科技支撑计划项目，在数据挖掘、自然语言处理、社交网络分析等领域开展了前沿性的研究工作，取得了一系列具有国际水平的研究成果，培养了一批优秀人才，形成了一支富有创新精神的研究队伍，构建了高水平实验研究平台，建立了良好的国际合作关系，为本项目的研究打下了坚实的基础。

在数据分析、数据可视化与信息检索方面，申请人所在科研团队自 2005 年起，就已开展社会网络数据挖掘、文本检索的相关研究，承担了国家自然科学基金、国家 863 计划重点课题、国防预研和国际合作等国家级研究项目多项，发表学术论文 200 余篇，包括著名国际期刊 TKDE、information Science、WWWJ、IEEE Transaction on cybernetics，和著名国际学术会议 SIGIR、SIGKDD、AAAI、IJCAI、AAAI、CIKM、SDM、DASFAA 等。本项目的研发团队由青年教师、博士生、硕士生组成了搭配合理的研究梯队，参与人员在相关方向已有 1-3 年以上的研究基础。

（3）课题相关技术积累

近几年来，课题组开发了一系列与本课题相关的软件系统，主要包括：

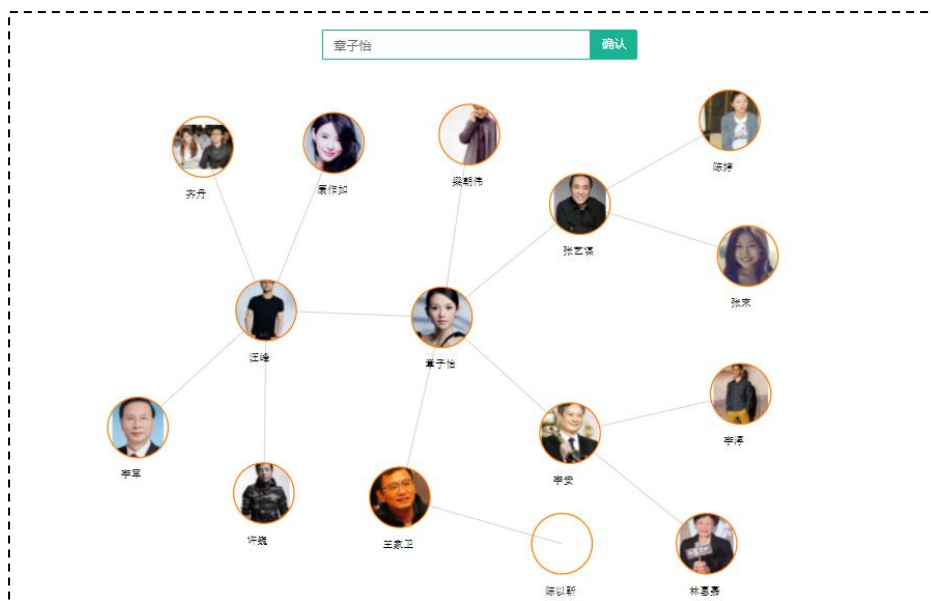
1) 面向特定领域的知识抽取与政局风险预测系统（2017 年 6 月-2018 年 6 月）

系统主要实现对人物相关网页语料进行包括属性、关系、观点在内的知识信息抽取，另外利用特定国家或地区的暴乱或抗议数据进行政局风险预测。在特定领域，给定一段非结构化文本，能够自动抽取出文本中的所有实体、实体属性、以及实体间的关系。其中实体包括人名、地名、机构名等。针对特定的人物，抽取出该人物对某个事件的观点或立场。

实体属性抽取包括性别、国籍、党派、职务、职业、出生日期、出生地、父母、子女、夫妻、岳父母、学历、专业、毕业学校、种族、工作地、工作单位、宗教等；实体关系抽取：人-机构工作关系、人-机构毕业关系、人-地名出现关系、人-地名出生地关系、人-人亲属关系、人-人社会关系。观点抽取指针对指定人物，抽取人物针对某个事件的观点或立场。如抽取美国总统、美国国会议员对 TPP、南海、台湾等问题发表过的观点和立场。

政局风险预测通过获取国家或地区的长时间段内的抗议、袭击等事件信息，主要包括数量，地点，群众舆论等建立模型，利用时间序列的方法进行未来一段时间内的风险预测，为国家或地区的安全提供信息支撑。





关系网络展示

中国、人民、一带一路、国家、经济、世界、中方、双方、中华民族、讲话、精神、成果、目标、利益、贡献、社会、制度、能力、理念、中共、事业、大陆、全球、体系、会议、方向、香港、两岸关系、社会主义、中美关系、原则、世界经济、中美、国际、国际关系、民族、政治、我国、方面、理论、思想、科技、工作、体制、国防、国际、机遇、倡议、进程、基础、政府、形势、国家主席、地区、中共中央总书记、小康社会、责任、侧结构性改革、水平、方式、中央、报告、要求、正义、领土、疆域、孙中山先生、美国

找到相关结果约3116个:

G20领导人警惕民粹主义兴起
 "发展为了人民、发展依靠人民、发展成果由人民共享。"习近平表示。
 FT中文网 <http://www.ftchinese.com/story/001069247> 2016-09-06 05:05:04

习近平向世界发出一份绿色发展的邀请函
 习近平强调发展清洁能源再一次彰显中国走绿色发展道路的决心。
 四月网 <http://news.m4.cn/2016-06/1310654.shtml> 2016-06-21 11:11:43

习近平致信祝贺第39届国际标准化组织大会召开
 习近平强调，中国将积极实施标准化战略，以标准助力创新发展、协调发展、绿色发展、开放发展、共享发展。
 宣讲家网 <http://www.71.cn/2016/0912/909754.shtml> 2016-09-12 20:08:00

观点抽取展示

2) 社交网络舆情监控与预警系统（2012年5月-2014年12月）

本系统是一套针对境内外微博、论坛、博客、新闻等网站的实时信息，集舆情信息的收集、分析、监控、预警与引导为一体的网络舆情分析与预警系统。主要功能包括：

面向 web2.0 的信息自动爬取：不仅可以爬取各大新闻网站等静态网页，还可以爬取微博网站等动态交互式数据，可以根据用户定制的主题实现深度爬取等。

舆情热点话题发现与分析：系统实时地发现社交网络上发生的重大事件引起的或用户发起的具有重要影响的话题，并对检测出的话题进行多维分析，地理分布及变化、热度变化、关注点变化、情感变化等。

舆情信息传播路径分析：自动分析舆情信息从源发点到当前节点所经过的传播路径，包括信息传播和人物关系双层路径图。

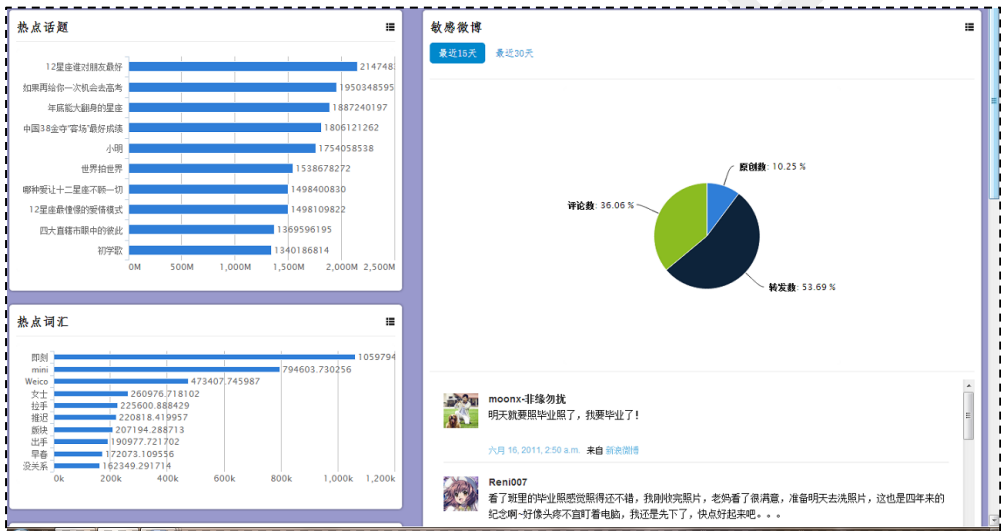
热点与敏感词汇发现：自动发现舆情话题中突发的热点词语，并且利用模式匹配即自然



语言处理技术自动识别敏感词汇及其各种变形、伪装等。

舆情实时监控：对用户自定义的事件或重大事件进行跟踪，对事件背景和发展状态进行实时分析，对触发用户定制的预警条件的敏感信息及时进行预警，通过邮件、飞信等手段通知用户及时做出反应。

系统在 hadoop 分布式平台下，综合利用 hive 和 hbase 等组件，实现海量社交网络数据处理，利用 baidu 地图实现话题地理分布的动态展示，利用 Highchart 展示热点话题的热度变化趋势、情感变化趋势，利用 D3.js 展示博主的社区关系以及话题的网络拓扑结构，这些工具对 HTML5 都有很好的支持。界面展示考虑了与用户的互动，例如根据用户的选择、登录信息展示相应的信息，并且根据最新爬取的网络数据对展示内容实时进行更新，给予文本聚类、文本摘要等技术自动生成舆情摘要的报告。系统主要的界面如下几幅图所示。

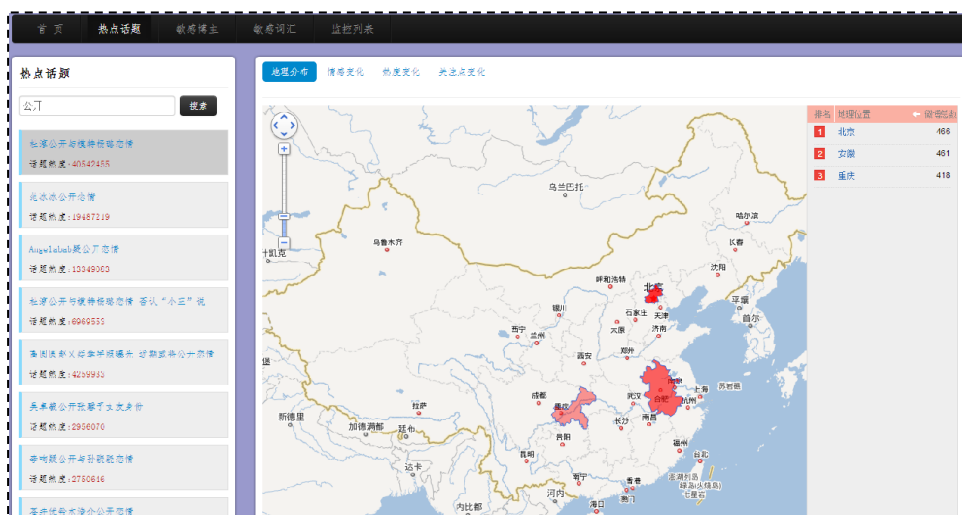


热点话题列表界面



话题热度趋势的分析





话题热度在全国的分布图



微博的传播路径图

3) 基于用户反馈的文本检索系统 (2015 年 4 月-2016 年 3 月)

给定包含大量新闻信息的文本语料以及查询关键词,根据关键词对新闻语料计算相关度并进行排序。用户显示反馈相关文档信息和不相关文档信息,根据反馈信息再次进行相关度排序,循环这一过程直到用户满意。我们应用组合式反馈技术来增强反馈效果,以提高检索结果的排序质量、从而更准确地获取用户的查询意图。



基于相关反馈的文本排序系统				
选择文件 未选择任何文件		+ - 中国 与 爆炸		
上传		确定		
序号	标题	内容	相关度	相关与否
1	[近年来中国频发粉尘爆炸事故]	[摘要]本溪煤矿发生煤尘爆炸，死亡1549人。8月2日拍摄的昆山爆炸事发现场。此次江苏昆山市开发区中荣金属制品有限公司汽车轮毂抛光车间...	0.21183	是/否
2	[瓦斯爆炸三人遇难]	[原标题：瓦斯爆炸三人遇难成都在建隧道发生爆炸追踪截至25日晚，成都“2·24”瓦斯爆炸事故已致3人遇难。目前，隧道内勘查、搜救工作仍在紧...	0.17463	是/否
3	[中国兵器集团公司发生爆炸已造成一死...]	[摘要]中国兵器集团公司发生爆炸今日上午，中国兵器工业集团辽沈工业集团发生爆炸。工作人员称系安装炮弹引信时爆炸，已造成一死一伤。返...	0.17208	是/否
4	[波士顿爆炸已致179人伤亡中国留学生1...]	[中新网4月17日电据外媒报道，当地时间16日，美国波士顿马拉松比赛接连发生两起致命爆炸袭击事件，警方表示，目前爆炸事件伤亡人数已升至...	0.16919	是/否
5	[昆山工厂爆炸前曾被通知整改员工多不...]	[原标题：昆山工厂爆炸前曾被通知整改员工多不知粉尘会爆炸37岁的宋成强很幸运，尽管他没有意识到。8月2日，一场突如其来的爆炸降临昆山...	0.16558	是/否
6	[中国女留学生在波士顿爆炸中一昏迷...]	[爆炸中昏迷的中国女留学生周月舒。@zzxb27/图失踪的中国女留学生吕令子，据新华网消息，中国驻纽约总领事馆16日向记者证实，一名中国女...	0.15943	是/否
7	[中国矿业大学一实验室爆炸]	[原标题：中国矿业大学一实验室爆炸信息时报网《扬子晚报》报道，昨日12时40分许，位于江苏徐州的中国矿业大学南湖校区化工学院一实验...	0.15866	是/否
8	[中国兵器工业集团一工厂爆炸致1死1伤]	[原标题：中国兵器工业集团一工厂爆炸致1死1伤据天都市报报道昨日上午10时05分许，中国兵器工业集团所属辽沈集团发生爆炸，造成1...	0.15472	是/否
9	[中国兵器工业集团一工厂爆炸致1死1伤]	[据天都市报报道昨日上午10时05分许，中国兵器工业集团所属辽沈集团发生爆炸，造成1死1伤。经初步了解，事故系辽沈集团兼压药工...	0.15472	是/否
10	[波士顿爆炸嫌犯被锁定习近平悼念...]	[4月17日晚，沈阳市民在街头用蜡烛悼念16日在波士顿马拉松爆炸案中不幸遇难的中国沈阳籍女留学生。红圈中即为疑似沈阳市民街头悼念美国...	0.15117	是/否
关键词：中国、爆炸、发生、事故、人员、美国、受伤、记者、飞机、报道、工作、粉尘、波士顿、现场、时间、调查、马来西亚、医院、公司、国家				

基于用户反馈的信息检索系统

4）湖南移动网络和互联网服务质量监控系统（2011 年 11 月——2012 年 10 月）

该系统主要用于移动、联通、电信等互联网运营商实时监测网络和互联网应用层的服务质量。系统可以测量网络层和应层的大多数常见协议。系统使用 ExtJs、JFreeCharts、百度地图、Flex、HTML5 等作为前端展示技术，后台使用分布式 Mysql 集群和自行设计的并行算法处理了每天采集的海量数据(每天原始数据总量达到了 1 千万，单表超过了 180 万)。

另外系统还在 HTML5 SVG 基础上实现了网络拓扑图绘制工具，借助 SVG 的帮助，拓扑图能够在 IE、Firefox、Chrome 等主流浏览器上使用。系统可以在 web 上展现网络探针、交换机、路由器、服务器等上百个组成的复杂网络拓扑图，同时实现了节点状态、链路状态的实时监测状态（打点、连线、最短路径等功能）。

5）新农村合作医疗信息管理系统（2008 年 10 月——2009 年 11 月）

这是为湖南长城医疗科技有限公司开发的新农合医疗信息管理系统，该系统采用 J2EE 的 Struts + Hibernate + Spring 框架，表现层采用基于 Ajax 技术的 ExtJs 技术，界面美观。该技术平台集成了权限控制、数据库连接管理、事务管理、通用通信框架、及通用数学计算框架。

团队依据卫生部《新型农村合作医疗信息系统基本规范 2008》实现了包括参合管理、补偿管理、基金管理、会计核算、查询统计、监测分析、业务公示、配置维护、基线调查、方案设计与测算、体检管理、参合群体分析、疾病信息分析和数据整合 14 大功能部分，还开发了门诊统筹子系统的 web 版本和移动版本。

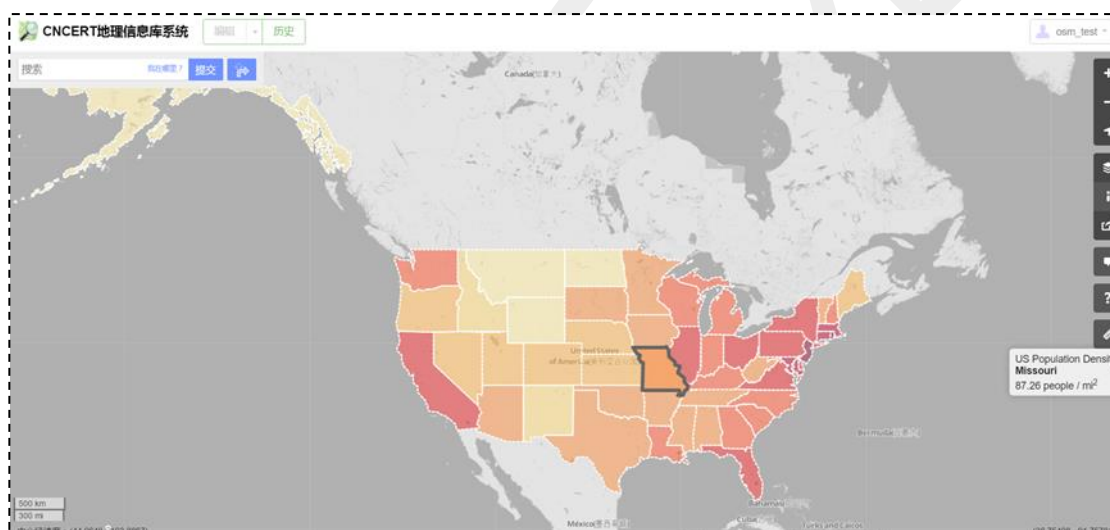
6）局域网地图库系统（2016 年 3 月——2017 年 4 月）

面向局域网的网络地图信息库系统实现了一个面向局域网的网络地图信息库系统，系统包含全世界各级行政区的多有地名及经纬度信息，POI 点数量超过 40 亿条，实现全世界所有大城市的中文名称显示。系统提供方便快捷的地图数据修改 GUI，可以方便地理信息数据的实时修改。另外还提供功能丰富的地图 API，使得开发人员能够使用统一的数据格式方便地实现热点图、动画路径、图层遮罩等功能。





地图展示界面



基于地图 API 的可视化开发

3.7.2. 合作单位任务分工（若无合作单位可不填写）

3.8. 课题经费预算

3.8.1. 经费预算表

单位：万元		
科目	预算金额	备注
（一）直接费用	85.00	



1、人员费	29.40	
(1) 高级研究人员	6.00	高级研究人员 5 名,按 0.25 万/月计算,计 $2 \times 12 \times 0.25 = 6$ 万元
(2) 其他研究人员	23.40	其他研究人员 13 名,按 12 个月, 0.15 万/月计算,计 $13 \times 12 \times 0.15 = 23.4$ 万元
2、设备费	10.80	3 台服务器,每台 2 万,需要设备购置费 6.0 万
(1) 购置	10.80	
(2) 试制	0.00	
3、国际合作与交流费	6.00	
(1) 课题组成员出国考察费用	6.00	出国参加会议或学术交流, 2 人次, 每人 3 万元
(2) 外国专家费用	0.00	
4、业务费	33.40	
(1) 材料费	8.00	8 用于购置课题研究过程中需要的相关材料;
(2) 燃料及动力费	0.00	
(3) 外协测试化验与加工费	0.00	
(4) 出版物/文献/信息传播/知识产权事务费	8.00	用于支付论文版面费、资料复印费、专利费等;



(5) 会议费及差旅费	17.40	学术会议 0.60 万/人次×20 人次= 12 万元；交通费每人月 300 计算，计 0.03 万/人月×15 人×12 月=5.4 万元。
5、修缮费	0.00	
6、其他直接费用	5.40	用于项目数据标注、数据采集费、以及专家咨询费等
(二) 间接费用	0.00	
1、现有仪器设备使用费	0.00	
2、房屋占用费	0.00	
3、直接管理人员费用	0.00	
4、其他间接费用	0.00	0.00
(三) 协作研究支出	0.00	
协作支出 1	0.00	
协作支出 2	0.00	
协作支出 3	0.00	
合 计	85.00	

3.8.2. 经费预算说明

(对各科目支出的主要用途、与课题研究的相关性及测算方法、测算依据进行详细分析说明。未对支出进行分析说明的，一般不予核定预算)

1、人员费

本课题共有中高级研究人员5名，根据国家关于人员费方面的财政标准，人员费按照0.25万/月计算，高级研究人员的人员费共计为 $2 \times 12 \times 0.25 = 6$ 万元。

本课题共有其他研究人员13名，根据国家关于人员费方面的财政标准，整个研究周期按照12月来算，每名其他研究人员的人员费按照0.15万/月计算，其他研究人员的人员费共计为 $13 \times 12 \times 0.15 = 23.4$ 万元。

因此，人员费的预算共计为 $6 + 23.4 = 29.4$ 万元。

2、设备费

需要购置6台PC机，按照每台PC机0.8万元计算，需要设备购置费4.8万元。还需要购置3台服务器（1台web服务器，2台数据库服务器、）作为工作平台，以每台2万计算，需要设备购置费6.0万。

两项设备费的预算共计为 $4.8 + 6.0 = 10.8$ 万。



3、国际合作与交流费

课题组成员出国参加国际会议或学术交流, 2人次, 每人3万元。

因此, 国际合作与交流费的预算为6万元。

4、业务费

本项目需要8.0万元的材料费, 用于购置课题研究过程中需要的相关材料;

需要出版物/文献/信息传播/知识产权事务费等方面的费用8.0万, 主要用于支付论文版面费、资料复印费、专利费等;

需要17.4万的会议费及差旅费, 其中包括: 课题组拟参加的相关学术会议和调研, 预算为0.60万/人次 \times 20人次= 12万元; 在市内交通费, 按每人月300计算, 预算为0.03万/人月 \times 15人 \times 12月=5.4万元。

业务费的预算共计为8.0+8.0 +17.4= 33.4万元。

5. 其他直接费用

项目数据标注、数据采集费、以及专家咨询费等其他未列费用共计为5.4万元。

整个课题的预算为85万元。



四. 声明

申请人承诺：

我保证申请书中所有信息真实可靠。如有失实，本人将承担相关责任。

课题申请人（签字）：_____

年 月 日

依托单位承诺：

同意申报，并已对申请人的资格和申请书内容进行了审核。如有失实，本单位将承担相关责任。

申请单位签章：_____

依托单位法定代表人（签字）：_____

年 月 日



第二章：申报课题信息简表



申报课题信息简表

(2018-04-08)

指南编号	课题名称	申请单位	申请人	申请经费 (万元)	密级	申请人联系 方式	申请人 Email	联系人	联系人 手机	联系人 Email	团队名称	团队备案号
17 指 153	国内行业 专家画像 库系统	北京航空 航天大学	张小明	85.00	内部	138106431 92	yolixs@b uaa.ed u.cn	李舟军	823382 47	lizj@buaa .edu.cn	信息安全与 智能信息处理 实验室	242HZDW01701



五世版木

