【本模版适用于成果形式为实用系统、原型系统、知识库、辅助工具的指南。】

1. **需求概述**

## 需求背景【不超过600字】

【不需要填写】

## 本期目标【不超过200字】

针对网络人物信息数据复杂多样、关系复杂、规模巨大、采集困难的特点，通过研究特点领域专家的有效数据采集方法、特征描述方法、科学评价方法等，综合利用网络爬虫、信息抽取、文本挖掘、机器学习、社交网络分析等技术，根据业务需求，自动从网上采集给定领域的专家信息，从采集的网页数据中自动抽取专家的各项信息，构建一个领域专家库，利用挖掘、关联分析等技术进行专家关系分析、主要观点抽取、科学评价等，为业务中的其他应用提供基础的信息与知识库。

1. **需求说明**

## 功能需求

【详细描述系统总体功能需求，各子系统或模块名称及其功能，逐项定量和定性地叙述功能要求，指明具体指标，每一项需求单独列为一条，不要将多项需求放在一条中描述。其中具体指标允许根据实际情况区分并明确标注“基本要求指标”和“探索要求指标”，“基本要求指标”为要求承担单位必须完成的指标，“探索要求指标”为要求承担单位进行研究探索的指标，但不对指标完成度进行强制要求。】

本课题将根据业务需求，利用网络爬虫自动采集国内主要高校、研究院所的著名研究学者的网络数据。对采集到的人物数据进行信息抽取，并进行结构化存储。利用数据挖掘和关联分析等技术，对专家人物进行特征描述、科学评价和关联聚合分析。利用数据可视化技术，对分析结果进行可视化展示，并提供专家检索和推荐。

## 专家信息自动采集

针对业务需求，开发网络爬虫工具自动采集主要高校和科研院所特定领域的专家信息，数据来源包括大学和研究所的官方主页、专家个人主页、百度百科、学术文献网等。

## 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，构建专家描述指标、评价指标、关联指标，对专家进行综合评价和分析，描述性指标包括：人物的主要属性、论文发表情况、学术关系、专利情况、研究兴趣关键词等；评价指标包括：论文的统计数据、论文的被引用情况、论文的影响因子等；关联指标包括专家之间的主要关系及其关系的强度，关系的类型包括：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、人物与机构的关系等。将处理后的数据在数据库进行结构化存储。

## 专家数据挖掘与分析

分析专家的关系、研究兴趣随时间的变化情况；在各类指标的分析基础上，识别给定领域的主要专家及其研究内容，并利用综合评价模型生成各类专家排行榜；据专家的各种关联关系，包括合作强度、关键词耦合、引文耦合，通过定量分析，提供专家检索、推荐、聚合，对专家检索提供按研究内容检索、按名字检索、按机构检索功能，对文献资料提供按关键字检索、按作者名字检索，专家推荐提供按研究兴趣相似度推荐、按关系强度推荐，专家聚合提供按关系和按研究兴趣的社区发现。

## 专家数据可视化

根据对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。

## 性能需求

【详细说明系统性能需求，包括对该系统的精度、时间特性及灵活性等要求（不限于以上几点，可具体到子系统或模块）。逐项定量和定性地叙述性能要求，指明具体可考核指标，每一项需求单独列为一条，不要将多项需求放在一条中描述。其中具体指标允许根据实际情况区分并明确标注“基本要求指标”和“探索要求指标”，“基本要求指标”为要求承担单位必须完成的指标，“探索要求指标”为要求承担单位进行研究探索的指标，但不对指标完成度进行强制要求。】

## 网络爬虫

网络爬虫能覆盖国内211、985高校给定领域相关专业师资的官方网站、给定领域的中科院相关研究所官方网站，给定领域著名研究学者的个人主页、百度百科网页、以及百度学术和万方等学术网站，并按照业务需求的频率进行网页的更新。

## 数据处理

能对爬取的网页数据进行实时的数据预处理与信息抽取，抽取算法的延迟不高于5秒，抽取的人物属性及各类信息包括：出生日、毕业院校、工作经历、论文列表、主讲课程、专利列表、主持和参与的课题列表、主要研究内容和观点关键词。

## 数据挖掘与分析

能够对10万级以上人物库进行挖掘分析，数据挖掘与分析的更新时间延迟不高于1天。

## 数据可视化

能对10万个以上节点的图数据进行可视化，可视化包括柱状图、饼状图、曲线图、摘要、多维可视化、检索可视化，支持用户的交互。

## 综合管理需求

【说明对该系统的数据管理、故障处理、输入输出数据类型、升级等要求（不限于以上几点，可具体到子系统或模块）。】

**数据库管理**：领域专家系统要求建立一个结构化的数据库管理系统，针对每个专家进行详细信息的记录，方便对专家属性信息，文章发表信息以及言论观点等信息进行特定目标的检索和查询。数据库存储要求跟踪专家科研论文发表动态，言论动态等并进行实时数据更新。

**故障处理**：系统要求全天24小时运行，同时有专业人员迅速反应并快速处理系统故障，以保障用户对系统的使用需求。

**输入输出数据类型**：系统输入为指定的专家人名，输出为对应专家结构化的属性信息，论文发表信息，观点动态信息，专家之间的关联耦合度信息以及专家科研评价信息等。

**升级**：根据用需求，如需升级领域专家系统，需要对专家系统进行严格的版本控制。最新版本的系统需要包含上版系统的所有功能，并按需求更新模块功能以及增加相应的模块功能。

## 接口要求

【说明该系统同其他系统之间的接口、数据通信协议等。】

**系统接口**：领域专家系统和其他模块系统之间存在有数据请求与传输的需求，需要定义接口信息。Web网络信息时代，大多利用HTTP进行数据的请求传输。领域专家系统中，统一采用http的两种请求方式get及post实现数据请求传输。请求接口统一提供请求端IP地址，被请求端服务器地址链接，请求数据类型以及字典形式包装好的请求数据内容。被请求端服务器通过接口统一向请求端返回相应的请求数据。

**数据通信协议**：为提高通信协议的灵活性，支持任意大小规模的网络，能连接所有服务和工作站，领域专家系统与其他系统之间统一采用TCP/IP通信协议进行数据传输。

## 其他要求

【其他需要特殊说明或强调的要求。例如，如果上述需求说明中区分了“基本要求指标”和“探索要求指标”，则在此处添加一段说明文字：“以上需求说明中，标注为“基本要求指标”的为要求课题承担单位必须完成的指标，标注为“探索要求指标”的为要求承担单位进行研究探索并力争达到的指标，但不对指标完成结果进行强制要求。”】

无

1. 验收测试

## 验收标准

【一是说明该系统验收测试环境为实验环境测试还是现网测试。二是说明验收测试的关键功能及性能指标，以及该指标测试时达到什么标准方为可通过验收。】

领域专家系统最终目标是搭建一个能查询获取领域专家属性信息，文章发表信息，言论观点动态信息，专家关联耦合度信息以及专家评价等信息的Web网站系统。该系统验收过程中支持现网环境测试。

课题应达到的指标包括如下信息

* 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，言论观点库；
2. 实现领域专家属性信息构建，言论观点信息抽取，发表文章检索，论文相关性排序，专家推荐等功能
3. 专家关联度评估：
   * 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。
   * 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排序。
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。

1. 按名字和研究内容关键词进行专家检索
2. 根据研究内容和人物关系进行专家推荐
3. 按关系和研究兴趣进行专家社区发现
4. 可视化的网页Web领域专家系统一套：包含2），3），4），5），6）和7）的可视化展示。

* 性能指标

1. 专家知识库中包含专家涵盖给定领域内大部分专家学者；科研论文库包含对应专家近十年发表文章信息；言论观点库包含主要媒体爆料的专家主要言论信息。
2. 专家知识库中构建专家实体数量不低于十万级别，包含专家论文数据量信息不低于十万级别。
3. 专家言论观点信息抽取准确率不低于70%。
4. 论文相关性排序算法MAP不低于80%。
5. 专家属性算法准确率不低于80%。

## 主要测试用例

【对应验收标准，围绕关键功能及性能指标，说明所使用的测试用例，应说明每个用例具体对应的指标及测试方式。】

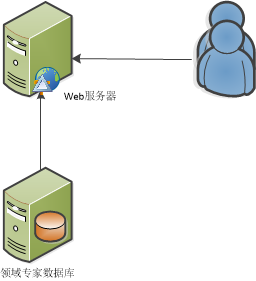
|  |  |
| --- | --- |
| 用例名称 | 指定专家信息展示 |
| 测试内容 | 输入专家姓名，查看专家信息 |
| 测试条件 | 领域专家库系统所有功能开发完毕，正常运行 |
| 测试过程 | 1. 进入领域专家库主页面。 2. 输入框中输入专家姓名。 |
| 预期结果 | 领域专家库系统返回相应专家的属性信息，论文发表信息，论文相关性排序信息，言论观点信息，专家耦合度关联信息，专家科研绩效评估等信息。 |

## 部署运行环境

【说明该系统运行所需的软硬件环境及提供方，并绘制网络拓扑图。】

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |
| 3 | 硬件 | Web服务器 | 1 | 委托方提供 |

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知识库数据存储，对专家发表文章数据信息，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. **课题进度要求**

【说明课题整体进度要求及关键节点要求。包括各阶段工作计划、考核指标、执行进度及提交成果。应至少包含三个阶段：需求调研与分析、开发与上线调试、测试与用户试用。其余关键节点请自行添加。中期检查和终期验收在课题合同中有统一约定，此处如无特殊要求可不体现。】

|  |  |  |  |
| --- | --- | --- | --- |
| **年月** | **工作计划** | **考核指标** | **执行进度及提交成果** |
| 2017年3-2017年4月 | 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。 | 提交需求说明说书 |
| 2017年5月-2017年11个月 | 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。 | 提交课题执行情况，执行纪要 |
| 2017 12月-2018年1月 | 系统优化 | 优化系统各项功能，性能。 | 提交完整领域专家库系统一套 |
| 2018年2-2018年3月 | 测试与用户试用 | 系统单元测试，功能测试，性能测试 | 提交测试报告 |

1. **交付物一览表**

【明确说明课题验收时应提交的所有交付物。填写时请保留模版中包含各类交付物的统一表格，通过在“数量”中填写“0”或其他数字来表达课题所要求的交付物及数量，而不要直接删除。其中 “软件系统（定制）”或“专用设备（研制）”、“验收文档”、“项目成果电子版”为固定要求，其“数量”和“备注”一栏内容不得修改，“软件系统（定制）”应罗列子系统或模块名称及其简要功能。】

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