

教师接收学生实习计划表

实习题目	搜索引擎中关键词精确检索与近似检索技术		
指导教师	陆嘉恒	职 称	副教授
指导学生人数	5	所需机器台数	5
指导计划（实习内容、实习目标、实习计划与实习要求）			
<p>1. 实习内容:</p> <p>从上个世纪中叶开始，人类逐渐开始使用计算机管理数据，数据管理也走过了人工管理，文件系统和数据库系统三个阶段后日趋成熟，但是随着互联网的飞速发展和信息量的爆炸增长，海量异构数据源使得传统的数据管理方式开始有些力不从心了，新的问题不断涌现，具有代表性的当属数据源的多样性带来的数据有效检索的困难。例如今年“512 汶川大地震”之后，谷歌，腾讯和搜狗等许多互联网公司纷纷行动，开设网上寻找亲人的论坛。但是这些论坛的数据格式各不相同，如何建立一个统一的检索平台，让丢失亲人的用户一下子能够检索各个网站的异构数据源显然是一个重要的研究课题。传统的数据集成技术周期长，效率低，不能满足用户的需求。所以，面对现实应用中的多个异构数据源，促使我们探索一种有效便捷的面向互联网搜索引擎的数据检索技术：这就是——基于关键词的精确检索和近似检索技术。</p> <p>关键词检索是信息发现最有效的查询方式，其最大优点是简单性，用户无需掌握复杂的查询语言，也无需了解任何数据的具体组织方式，即可轻松提交查询。关键词近似检索非常适用于异构数据源。因为面对没有一个统一模式定义的数据集合，检索系统无法支持类似 SQL 的结构化的查询语句。而关键词精确检索和近似检索就成为一个有效的查询手段。</p> <p style="text-align: center;">本次毕设的具体内容包括：</p>			

- **建立能够具有扩展性的描述多种异构数据源的关键词检索语义** 传统的关键词检索语义仅仅能针对一种数据类型，但是面对海量异构数据源，过去的单一语义不再适用，如何建立能够有效描述异构数据源中各种数据类型的数据模型与检索语义，将是一个非常具有挑战性的问题。
- **定义统一的关键词检索的排序策略** 传统信息检索技术中排序主要是根据 TF/IDF 和 PageRank 权重。但这里的最大的难点就是 TF/IDF 值不再是最高标准，而文档的结构特征和文档之间的关联关系等一些在多个异构数据源形势下产生的新标准可能会扮演非常重要的角色。
- **设计合理的基于中英文字符的近似匹配函数** 该函数必须解决两个问题。首先不同实体在不同数据源中有不同的称谓，需要考虑同义替换，其次，传统的编辑距离适用于英文输入方式，对于中文的拼音输入法并不合适，需要考虑同音替换。
- **建立支持字符近似检索的数据索引机制** 该索引机制需要解决基于数字签名的哈希技术中的误报值比较频繁的缺点，能够支持高效的字符近似检索。
- **搭建可用的关键词检索的原型系统** 将各个模块的成果综合起来，搭建一套可用的、能够支持对各个数据源的搜索引擎的原型系统。

2、实习目的：

通过本次毕设，同学们将有以下的收获：

- (1) 参与团队研究，提高自己的研究水平。
- (2) 实际动手开发搜索引擎，为将来工作打下良好基础
- (3) 完成毕业设计和论文，取得学士学位。
- (4) 发表一篇中文或英文论文，为将来进一步深造做准备。

3、实习计划：

2008年10月-2008年11月，进入研究研究小组，并开始进行相关研究的资料查询；

2008年12月-2009年1月，计划研究内容，开始创新思维和探索。

2009年2月-2009年4月，主要是撰写论文和编程实现新的算法。

2009年5月，完成毕业论文和毕业设计答辩。

4、实习要求：

学生要求具有独立思考和分析问题能力，关键在于愿意参与到我们的研究中来，并贡献自己的聪明才智。

理想的目标是：整个团队开发一个新的搜索引擎系统。学生每个人完成自己的毕业论文，并发表一篇中文或英文的期刊论文。

请感兴趣的同學访问陆老师网站：www.jiahenglu.net 或发Email jiahenglu@gmail.com 了解更多信息。