刘大力工作进展

1. 前期研发工作

参与CiteXplore系统的研发，主要参与了探索图及探索图元数据存储模块、内容代理与下载代理模块、离线下载模块负载均衡以及客户端的开发工作。

1. 探索图及探索图元数据存储模块

使用HBase + Mongodb实现了探索图及探索图元数据存储系统，保证了探索图数据及元数据在高并发场景下的一致性，实现了探索图数据的版本管理。

1. 内容代理与下载代理模块

使用Kafka实现了模块间的通信，协调了图存储模块，内容提取模块，离线下载模块之间的数据流通，是CiteXplore系统的“交通枢纽”。

1. 离线下载模块负载均衡

使用Haproxy + Lua实现了前端http访问负载以及离线下载模块可控负载策略的负载均衡系统，实现了用户请求下载与查询下载进度功能上的解耦，解决了高并发场景下可能出现的Manager服务器压力过大的问题。

1. 客户端插件开发及前端优化工作

参与开发三大学术搜索引擎客户端插件，使用Firefox的Add-on SDK，分别针对谷歌学术，必应学术，百度学术实现JavaScript注入，并针对三种客户端开发服务器端解析工具，为CiteXplore系统提供了准确的入口数据。并参与前端功能的优化完善工作，实现组用户广播，探索图的打分及评论功能。

1. 研究工作
2. 参与设计实施CiteXplore系统用户行为实验

实验设计为两大类型，一类为学习型，另一类为探索型。

学习型实验中，给定参与者指定的中文任务，要求其在指定时间内在任务范围内进行搜索。我们挑选了有一定复杂度、需要一定搜索量的话题作为任务，使其满足探索式搜索的条件。学习型任务分三次进行，每次时长20～30分钟，次间间隔1～2天。

探索型实验中，我们给定参与者英文摘要，要求其仔细阅读。摘要阅后即焚，并在1～2天后要求其进行摘要内容相关的搜索。这样做主要是为了使实验环境满足用户在进行探索型任务过程中“对整体任务不明确”的特点。探索型任务进行两次搜索，每次时长20分钟，次间间隔1～2天。

在用户的搜索过程结束后的一周时间内，会进行实验报告的整理，报告中会体现出用户得到的搜索结果与结论。对于学习型实验，要求用户根据任务要求，逐条目列出搜索到的信息。对于探索型实验，要求用户列出所有自己认为与英文摘要内容相关的文章标题。

在完成实验报告后，会针对本次实验开展小组讨论，以调查参与者对CiteXplore系统的使用体验。

最后，在经历了足够的时间间隔后，要求用户进行搜索过程的回想，描述自己的搜索过程。两种类型的实验都在CiteXplore系统以及对比系统SearchLog上进行，用户在回想时也分别基于CiteXplore系统以及SearchLog系统。

2016年年底到2017年年初我们做了8组实验，并向SIGIR投稿，得到反馈中提到我们的数据量略少，于是在上学期实验的基础上，本学期预计追加8个人的实验。目前（截止5月14日）第一组4人实验结束了搜索、报告以及小组讨论的部分，5月15日开始将进入回想以及数据分析阶段。5月22日可以开始第二组4人实验。

1. 数据分析工作

分析用户的回想数据，以从中发现CiteXplore系统到底怎样帮助用户更好地进行探索式搜索过程。

分析用户的搜索过程数据，希望从中发现用户搜索过程中的共性，希望能从中找到未来做搜索推荐的思路。

1. 毕业设计进展

毕业设计题目《探索式搜索过程中面向探索方向的多样性查询推荐方法研究》，怎样将时间树这种特定的日志形式更好地结合到可行的推荐方法中，目前还不是特别清楚。但对已有的数据进行分析，得到了一些阶段性成果。

首先对时间树上的查询节点和点击节点分别定义了其关键要素，即查询节点的状态转移性以及点击节点的切题性，并对状态转移进行了大致分类。

其次给出了查询及点击的粒度的大致定义，粒度可能对于未来给出推荐时范围的确定有所帮助。

最后对于时间树上最为特殊的一种结构：query - click - query，做了一些简单的统计分析，但接下来怎样利用这些数据帮助给出推荐还不清楚。

1. 下一阶段的主要工作

对已有数据的分析还在进行中。接下来的主要工作就是在对已有数据进行分析的同时，考虑如何进行推荐，并对CiteXplore系统研发中所参与的各个模块进行技术总结，以巩固技术积累。