北卡大学教堂山分校为期 6 个月学习报告

武汉分馆 仇华炳

1、基本情况

学习时间：2011 年 8 月---2012 年 1 月
学习单位：北卡大学教堂山分校信息与图书馆学院
指导老师：Brad Hemminger 副教授, Jane Greenberg 教授
主要参加项目：ISB 项目/DARi 项目

2、主要工作介绍

在如期到达北卡大学教堂山分校之后，与指导老师 Brad 和 Jane 分别进行了详细的沟通和交流， 根据我的学习目的和计划，对我在北卡大学教堂山大学近 6 个月的学习和工作进行了认真安排。期间，我主要参加的工作和活动如下：

2.1.学习参观工作

为了尽快熟悉信息与图书馆学院的基本情况，在指导老师的带领下，参观了信息与图书馆学院的图书馆，ODUM 实验室（Odum Institute, the "nation's oldest multidisciplinary social science university institute." 提供统计知识操作培训）。在信息与图书馆学院的图书馆中，我参观了他们的群组研讨学习室，学习室配备了多种现代化的会议设备；参观了该院期刊图书馆藏，包括图书馆硕博士论文馆藏。图书馆内收藏了自 1963 年以来的硕博士论文印本，并且建立了索引目录，其 1999 年以后的硕士论文采取电子收缴，并且可以查看部分内容。

2.2.参加的主要项目

2.2.1 ISB 项目

项目目的：通过对 Apache Server Logs 的分析，研究用户如何使用 facet 在带有 facet 功能的 catalog 中来查找资料的（行为模式）的特点；研究分面查找功能是否能够有效（或者能再多大效率上）帮助用户找到他们想要的资料；研究其它不为人注意的现象。

分面 facet 是指事物的多维度属性。例如一本书包含主题、作者、年代等分面。而分面搜索是指通过事物的这些属性不断筛选、过滤搜索结果的方法。可以将分面搜索看成搜索和浏览的结合。分面搜索作为一种有效的搜索方式，已经被用在电子商务、音乐、旅游等多个方面。

用户对 catalog 的每一次操作行为都会留下一条日志信息，日志量是十分大的，一条日志记录是形如如下的形式，

```
71.70.185.34 - - [09/Mar/2011:00:13:53 -0400] "GET
/search?Ntk=Keyword&Ne=2+200043+206472+206590+11&N=206432&Ntt =boston+globe HTTP/1.1" 200
40035 "http://search.lib.unc.edu/search?Nty=1&Ntk=Keyword&Ntt=boston+globe" "Mozilla/5.0
(Macintosh; U; PPC Mac OS X 10.4; en-US; rv:1.9.0.7) Gecko/2011021906 Firefox/3.0.7"
```

对文本日志本身，通过人工解读，会有下面一些问题：

1). 哪些日志对应的是一个用户；
2). 哪些是一个用户发出的一次连续的会话 session；
3). 无法判断用户是否找到了他想要的；
4). 日志条数是海量的，仅仅通过人工读阅，工作量太大；

因此我们采取了可视化化的方法进行解读。

2.2.1.1 可视化

日志中的各个部分所代表的意思如下：

Definition	Log Segment
Remote Host IP	71.70.185.34
Date and Time	[09/Mar/2011:00:13:53 -0400]
Request	“GET /search?Ntk=Keyword&Ne=2+200043+206472+206590+11&N=206432&Ntt=boston+globe HTTP/1.1″
Status Code	200
Bytes Sent	40035
Referring Page	“http://search.lib.unc.edu/search?Nty=1&Ntk=Keyword&Ntt=boston+globe″

Request 中可能出现的各个参数的解释如下：

参数	Meaning
N	unique ID of facet applied to query
No	the offset, used for pagination
Ntk	the index for text searching (Author, Keyword, etc.)
Ntt	text search terms applied to query
Sort	overrides the default sort order for displaying results
Ne	request to expand a facet group
view	request to switch result view from brief to full, or vice versa
Nty=1	(this is a constant that you should always include)

用户对 catalog 的操作模式是有限的，因此可以用有限种类的 action 来描述这些模式。一种 action 就是一种用户与 catalog 的交互。可以通过两条相邻的日志中 URL 的变化来得知用户在这之间进行过什么操作并把它划入某类 action 之中。

比如上表中的 Request 中的参数是 Ne=2+200043+206472+206590+11&N=206432，表示用户展开了几个 facets（2+200043+206472+206590+11），并且点击了 206432 这一个 facets。通过统计和人工识别分析，一共可以统计出 23 种 action，为了便于对 action 进行分析，我们进一步对相似的 action 进行归并，最终把所有 action 分为十类，之后通过程序操作自动识别出 action 所属于的分类。action 分类如下表所示：

细分 action 编码	粗分 action 编码	细分 action 编码	粗分 action 编码
BeginSimpleText	TextSearch	AddFacet	ModifyFacets
SingleTermText	TextSearch	RemoveFacet	ModifyFacets

细分 action 编码	粗分 action 编码	细分 action 编码	粗分 action 编码
MultipleTermText	TextSearch	RefineYears	ModifyFacets
MultipleFieldsText	TextSearch	NextPage	NextPage
BooleanTextSearch	TextSearch	SortResults	SortResults
EmptyTextSearch	TextSearch	Refresh	Refresh
SwitchTextField	TextSearch	ViewRecord	ViewRecord
BeginAdvancedSearch	AdvSearch	FollowupAction	FollowupAction
OpenFacet	ShowHideFacets	BeginNewTitlesSearch	OtherSearch
CloseFacet	ShowHideFacets	SelectNewTitles	OtherSearch
ShowMoreFacet	ShowHideFacets	BeginCallNumberSearch	OtherSearch

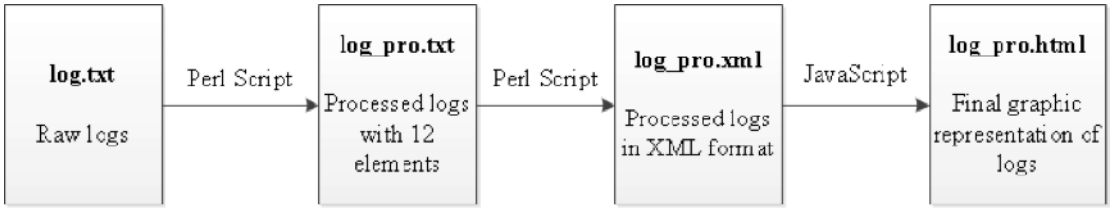
在用程序对日志进行处理的同时，还要取出或者通过程序计算日志中和该 action 对应的其余 10 个参数元素，这些元素是我们进行可视化呈现所必须的有用信息，也是后续进行用户行为统计分析所必须的统计样本。

还以之前的日志举例，这 12 个元素内容如下表展示：

元素	参数	参数值	描述
1	**细分 action 编码**	AddFacet	细分 action 编码
2	**粗分 action 编码**	ModifyFacets	粗分 action 编码
3	**Searchterm**	boston+globe	文本查询检索.从“Ntt”获取
4	**Field**	Keyword	查询所用字段种类. 从“Ntk”获取
5	**Facet**	206432 (Format:Online)	选择打开的 Facet 值. 从 “N”获取
6	**Refine**	-	用起始年做精炼. 从 “Nf” 获取
7	**Expand**	2+200043+206472+206590+11 (Availability+Location+Format+Subject+Publication Year)	展开来的 facets.从 “Ne”
8	**More**	-	用户点击“show more” Facet group. 从 “more”获取
9	**Record**	-	已查看过的记录.从 “R”
10	**Time**	65	当前 action 与下一个 action 直接的时间间隔 (秒). 从 date& time stamp
11	**Visit_num**	822	Session number 计算

12	**Visit_step**	3	Step number within a session 计算

总体来说，我们对日志数据的可视化处理可以用如下简化流程图表示：



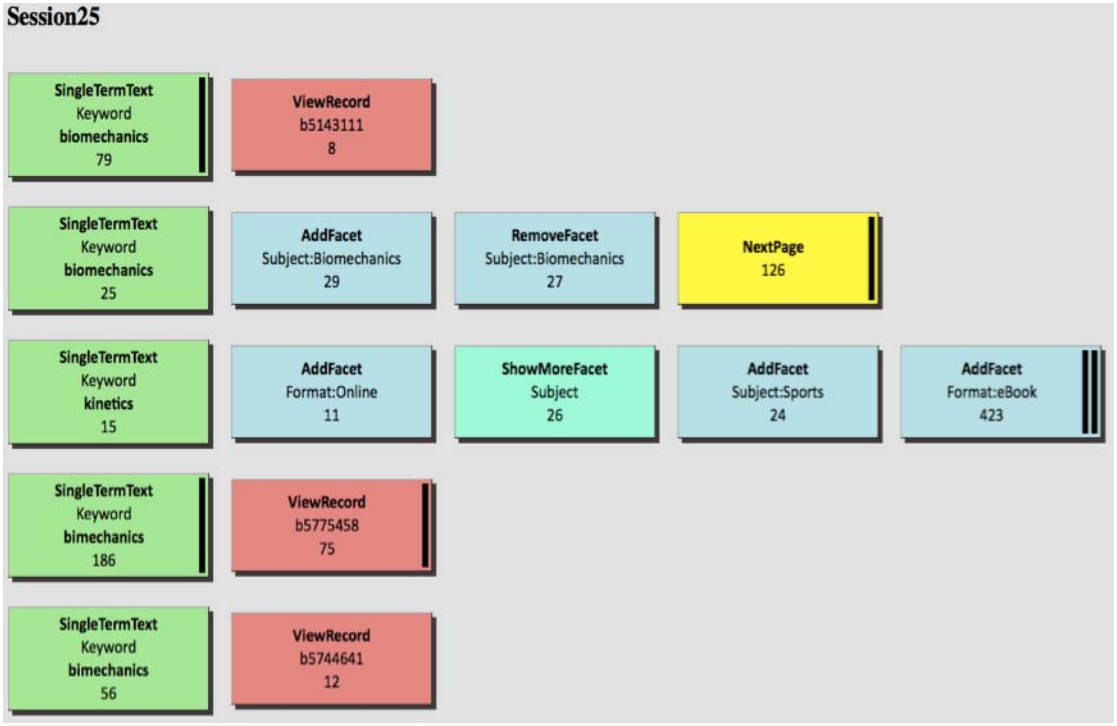
首先，对日志进行若干的程序预处理之后，提取所需要的 12 个元素，之后把这些元素按照一定的格式封装到 XML 中，然后使用脚本语言对 XML 进行处理，并且呈现到 html 页面上。

在可视展现的细节上，采用了使用不同颜色文字区分不同 action，如用黑竖线的数量来表示用户停留的时间长短等等。

粗分的 10 种 action 用 10 种不同颜色的矩形方框表示，再在同样颜色的矩形方框标上不同的文字内容以进行细分 action 显示，如下图

Action code	**Shape**	**Variables displayed in the rectangle**
TextSearch	Green rectangle	action, field, searchterm, time
AdvSearch	Light green rectangle	action, time
ShowHideFacets	Light blue rectangle	action, expand, time
ModifyFacets	Blue rectangle	action, facet, time
NextPage	Yellow rectangle	action, time
SortResults	White rectangle	action, time
Refresh	Gray rectangle	action, time
ViewRecord	Red rectangle	action, record, time
FollowupAction	Light pink rectangle	action, time
OtherSearch	Dark green rectangle	action, time

对一个会话过程的可视化展现，如下图所示。大量的会话便可以同时展示出来，供查看。



通过视觉上的扫视大量日志的色彩分布规律，如矩形的布局(颜色、文字内容)，每一个会话的时间长度等信息，图书馆员和研究人员可以对用户的行为有一个感官上直觉上的感受。比如说：

1)大量的绿色文本框相对于少量的红色文本框，可以从直觉上判断文本检索比 facet 检索的频率要高，用户使用得更频繁。

2)用户在每一排最后一处 action 上要花费更多的时间停留，说明用户在进入其他路径(如重新检索、点击进入某条记录)之前会花费相对较长时间去阅读、挑选或者思考。

2.2.1.2 数据分析：

我们在可视化展示后，我们希望了解用户是如何使用 facet 来帮助自己进行查找的。我们首先提出了几个问题，然后通过 SAS 软件结合数据库和程序进行了分析。

问题如下：

1）用到 facets 的频率会是多少？

2）通过对日志的分析，能否判定用户检索 catalog 的模式能够自然的被划分为某些查找模式？

3）facets 功能会对用户的查询行为模式产生显著影响吗？大多数 facets Search 是从 text Search 开始吗？

4）当 facets 在会话中使用了，那么这些会话的后续会有一定的模式（如相似的步骤等）吗？

通过对 13 万多个有效 session 进行分析，得出一个 session 中平均的 action 数是 9；出现最多的 action 是 MultipleTermText 和 ViewRecord，这些依然是传统 catalogue 中使用最多的 action；所有的 facets 操作 action 只占总 action 数的 6%（回答了第一个问题），比预期要低得多。也发现了一些请求不是来自于人，是来源与机器，网络爬虫。不过我们在后来做了一些处理，来清除这些数据。

2.2.1.2.1 研究方法：

1).聚类分析法

为了解答第二个问题，使用了聚类分析。按照会话之间的相似性进行聚类，换句话说：被聚类到一个组里面的会话具有相似性，反之则无。为了节省计算时间，使用的是 10 种粗分类的 action，同时为了避免人为的主观分类因素的影响。混合使用了层次分析法和非层次分析法。

结果: 通过聚类分析，发现了八个聚类簇，归纳出 8 类出现的比较多的操作模式：

1)SimpleTextSearch: 总的操作很少，只是输入文本，然后查找；

2)DetailedTextSearch:多数 action 平均分布于输入文本和 view record；

3)InDepthTextSearch: 点 next page 很多；

4) AdvancedSearch: 使用高级检索比较多；

5) FacetTextSearch: 多数 action 平均分布于文本检索和 facet；

6) FollowupSearch: 深入点击每条记录中的其他链接；

7) RSS group: 使用 RSS；

8) Roberts Outliers: 机器或者爬虫。

2).马尔科夫分析

为了回答第三个问题，对所有含有 facets 的 action 的连续性进行多阶马尔科夫聚类，通过卡方分布检验得到最相关的阶层为 2。就是说当前的操作和前面 2 步的操作相关性高。找出连续 3 个 action 操作的序列中重复比例最高的前 10 类。在这 10 类之中，TextSearch--TextSearch—TextSearch 和 ModifyFacets--ModifyFacets—ModifyFacets 是重现率最高的 2 类，但是加起来也不到所有模式的 12%。

因此认为，加入了 facets 的 catalog 并没有按照之前预料的 facets 会对用户的查询行为模式产生显著影响而产生特定的模式。

通过转移矩阵分析（一阶马尔科夫分析）：在产生 ModifyFacets action 前的一步中 textsearch 占 38.61%，产生 ShowHideFacets action 前的一步中 textsearch 占 36.59%

因此，得出 textsearch 是使用 Facets 的最可能起始点。

3).MRP 分析

为回答第四个问题，对所有含有 facet 操作 action 的所有会话使用最大重现分析(maximal repeating patterns)，发现了 54 种包含 facet 的重复率最高的 action 序列。这些序列通过人工分类分为三大类：

1) facet search-> viewing record;

2) text search-> facet search -> viewing record;

3) repeating (2) twice.

2.2.2 DARi 项目

DARi (The Data-at-Risk Initiative)项目是 MRC(UNC 大学 SILS 学院元数据研究中心)的一个孵化中的小项目。

DARI 项目的目的是要创建一个 inventory 清单列表，用来收集各种濒临灭亡或者失效的科学数据资料的描述信息以及没有充分描述的科研数据（比如说松散开来的老的文稿，实验

室笔记本，没有数字化的老照片，过去保存在磁带或者磁碟中的科研数据文档，等等任何应该做长效保存却没有做的科学数据或资料），最终提交交给相关部门进行保护。

DARi 本身并不存储这些数据，而是要收集并定位这些数据，对这类数据集中进行描述汇总，期待引起有关方面的重视，未来开展保护工作。目前项目组使用开源软件 Omeka 构建了一个在线提交濒危数据的平台，并设计好了数据条目的描述模版，供科研人员编辑提交濒危数据的元数据信息。

我在项目组中的主要工作：

1).参加 DARi 项目中分学科领域的用户调研（比如 DARI Focus Groups Planning 文件）调研对象主要是 UNC 校内各个学院的研究人员，调研方式是网络调研。

2).参与调查问卷问题与选项的讨论，设计工作。

3).参与平台界面的设计讨论工作。

2.2.3 其他工作活动介绍：

1）参加了 MRC 5 annual 的学术会议，还参加了分会场讨论；

2）参加 SILS 建院 80 周年的系列学术活动；

3）在 SILS 做报告介绍国科图的基本情况；

4）旁听 Human Information Interactions ，Organization of Information 等课程；

5）参观访问（县,市,高校等）多所图书馆，了解美国的图书馆文化和 IC 布局。

3.总结与主要的收获

在 UNC 大学为期 6 个月的学习期间，通过参加一些项目和课程学习，以及与指导老师之间的交流和学习，开阔了眼界，对今后的工作有很大帮助。

1).务实的工作态度。

我所参与的 2 个项目组都是非常务实的团队，每次会议，每次交流，都直入主题，开门建山，效率很高，相关人员实事求是的介绍本项目的进展、成绩和存在的问题。就算是 MRC 五周年的庆祝活动，也没有任何花哨的准备，邀请了国内一些同行，并且完全以一次学术会议的形式举办，纯学术型的。

2).保持开放的思想。

学院的老师同学很注中参加学术沙龙，头脑风暴等学术讨论活动，深入探讨，交换意见；他们还很积极地参加各类学术会议，制作海边，宣传自己的研究工作，不断寻找合作伙伴。他们更多的时候是通过兴趣做一件事，先从小做起，不断努力完善，不断发展做大，获得资助。

3). 有很强的时间意识。

工作计划和日程安排完善而有条理，会议时间的讨论确定等充分利用在线日历日程工具，十分科学民主。充分高效利用时间，在中餐的时候就能够聚在一起开展一次学术沙龙活动。

4). 科研的严谨态度

比如，对于用户行为模式，我们可能会想当然的通过讨论或者直觉来给出结论，而 Brad 他们在研究前只是客观的进行少量的假设，然后用大量数据分析的方法并结合用户访谈观测来证实或者推翻假设，再得出可能的结论。

5).个人掌握了一些工具和语言和方法。

为了顺利参与 ISB 项目，在美期间我自学了 Perl 语言，并且为项目写了部分处理程序。通过和项目组成员交流和自学，掌握了统计软件 SAS 基本的操作。

了解到许多信息查找行为的研究都遵循一定的研究方法：首先建立理论或者假设模型，通常会从心里学的角度出发，然后进行行销或者组织用户行为(如有偿的用户调研，录音录像)，最后收集数据做统计分析，通常是通过回归分析和要素分析来测试之前的模型。某些时候假设会被证实，某些时候假设被证伪。

这次美国学习的机会难能可贵，在美国的学习开拓了我的视野，学习了新的知识，认识到了自己的不足，激励我在今后的工作中，努力做好本职工作，长期坚持学习研究的态度。

最后感谢中国科学院国家科学图书馆群星计划的支持！