EZproxy 日志分析系统开发与应用实践

----北京工业大学图书馆电子资源校外访问日志分析

雷东升　郭振英　周培培　林琳
北京工业大学图书馆
2014年12月20日

主要内容

- 背景
- 可行性分析
- 系统设计与实现
- 问题
- 解决对策
- 小结

背景

- 电子资源成为学校教学科研工作不可获取的资源来源。
- 电子资源校外访问的开通为读者在校外访问图书馆电子资源提供便捷的途径。
- 电子资源校外访问方便读者的同时出现恶意下载资源现象，影响合法读者的访问权利。

背景

- 恶意访问电子资源现象
  - 出于不良目的、动机的恶意下载行为，长期连续的下载大量的数据库中数据。
  - 使用工具等不良下载行文，如在数据库的下载过程中使用专门的工具或者启用多线程方式下载数据。
  - 超过许可范围的不良下载行文，例如有些数据库服务厂商明确规定不允许一次性下载某种期刊同一期半数以上的全文数据。

背景

- 恶意访问电子资源的后果
  - 数据库商：占用了大量的服务器资源，包括通讯带宽资源、计算资源等，影响正常为客户提供服务。
  - 图书馆：恶意下载侵害数据库商的版权，影响了图书馆的声誉。
  - 合法读者：IP段被封掉，导致图书馆的合法用户的访问被中断，合法读者的权利受到侵犯。

可行性分析

- 北京工业大学图书馆从2006年开始启用EZproxy代理服务器，作为校内和校外电子资源的统一入口。EZproxy代理服务器产生的Web日志记录了校内和校外全部读者访问电子资源的详细信息，服务器每天产生一个日志文件，通过对日志的统计分析可以了解各类人员使用电子资源的情况和各种异常使用电子资源的行为。
- 因此开发基于EZproxy日志的分析系统来帮助工作人员统计分析电子资源读者行为的访问是非常必要的。

可行性分析

- 日志特点：
  - 数据量大：100M-200M
  - 包含95%以上的电子资源访问信息
  - 校外访问日志记录清楚，直接记录用户的行为，便于开展分析。

可行性分析

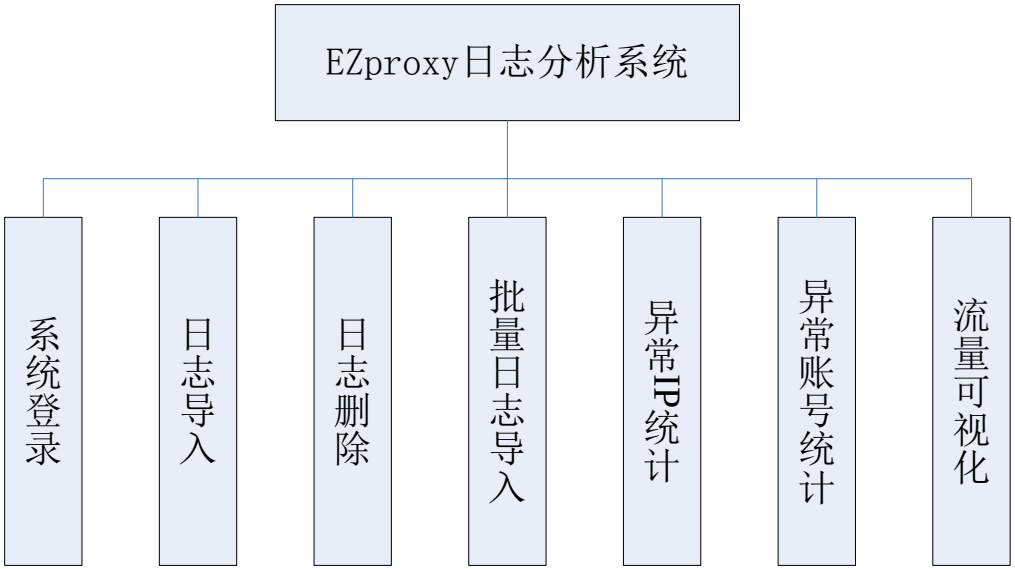
- 日志格式：LogFormat %h %{ ezproxy - session} i %u %t ″ %r″ %s %b

字段	说明
%h	用户访问使用IP地址
%i	代理服务器启用的特殊头标识,为唯一的会话标识,用户登录一次为一个会话
%u	记录用户ID,校内免登录,为Auto, 校外访问为读者证号
%t	访问时间
%r	URL请求
%s	返回状态信息
%b	访问流量

- 日志举例

123.123.250.5 BBisN8vM48mBqCh 05202 [13/Oct/2014:15:03:42 +0800] ″GET http://ac.els-cdn.com:80/S0142961214009120/1-s2.0-S014...669a2 HTTP/1.1″ 206 4827

系统的设计与实现



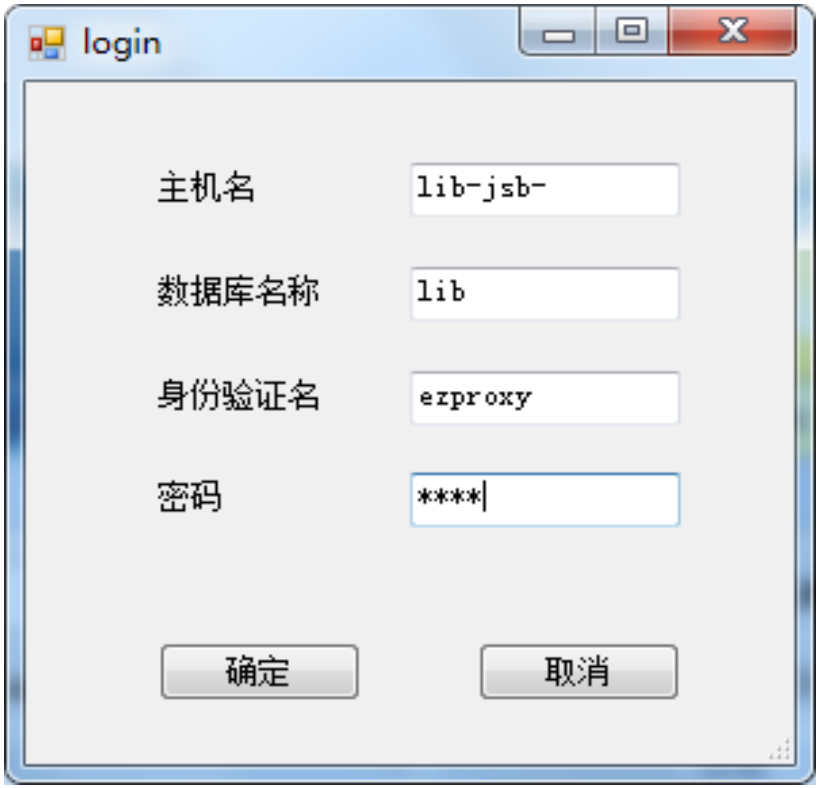
Ezproxy日志分析系统基本功能模块

系统的设计与实现

- 开发平台：
  - 系统基于.NET框架（.NET Framework）的架构设计，
  - 开发平台为Microsoft Visual Studio 2010，
  - 数据库使用SQL Server 2012。

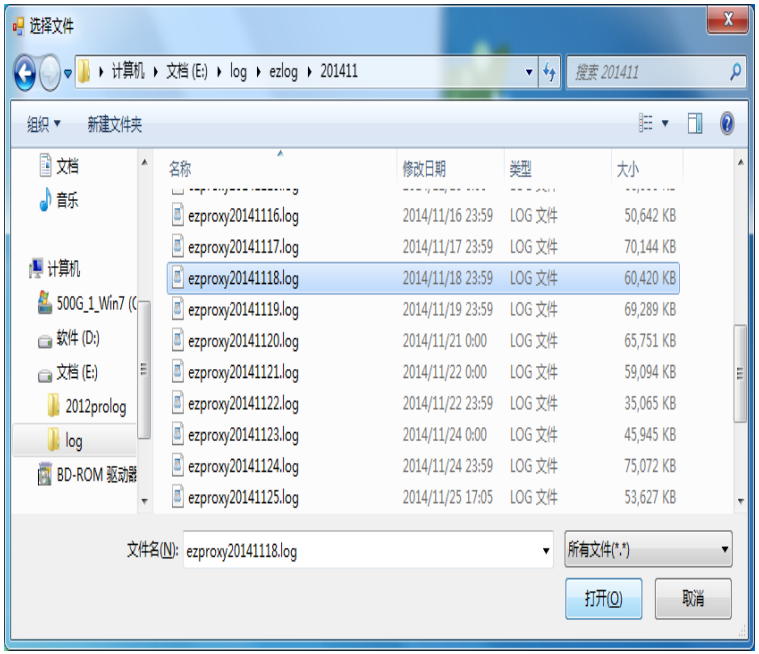
系统的设计与实现

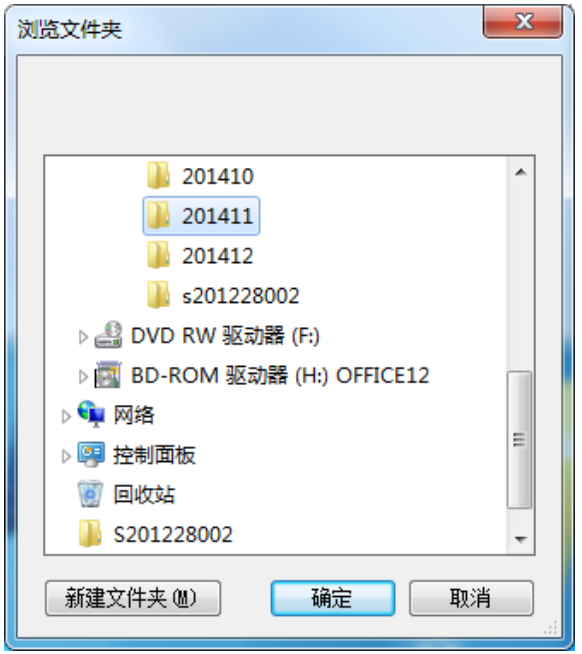
- 系统登陆模块：用户完成用户登录验证功能。



系统的设计与实现

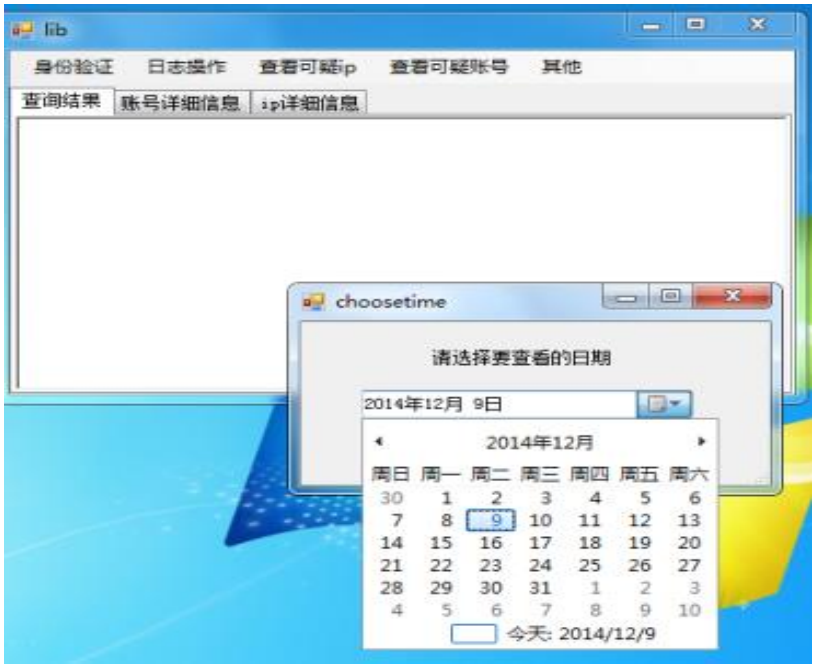
- 日志导入模块：支持单个日志导入和批量日志导入功能。





系统的设计与实现

- 日志删除模块



系统的设计与实现

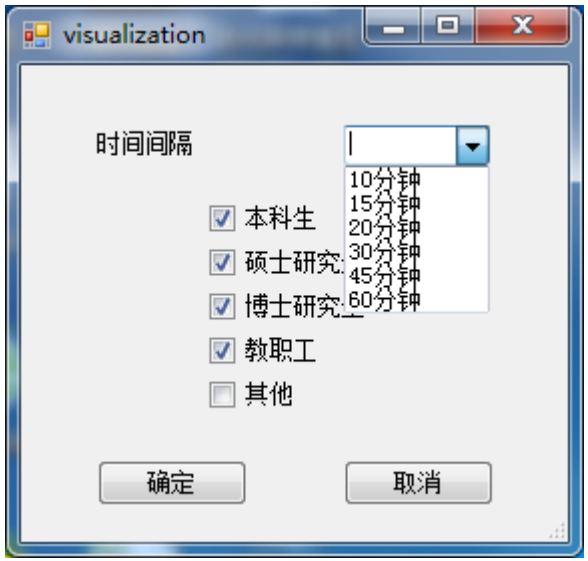
- 异常IP统计功能：按照用户IP地址的下载量进行汇总，统计出下载量过大或者访问次数多的IP地址。

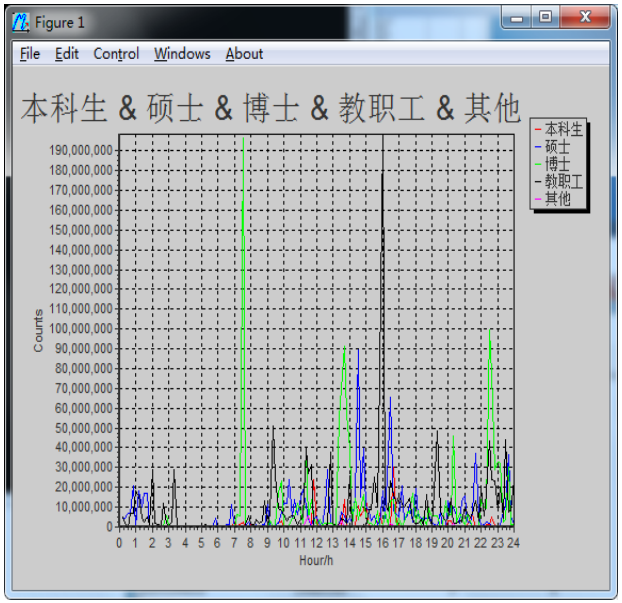
系统的设计与实现

- 异常账户统计功能模块是按照校外访问日志中下载量过大、登陆次数过多、使用IP地址过多进行统计，找出不正常使用的用户。

系统的设计与实现

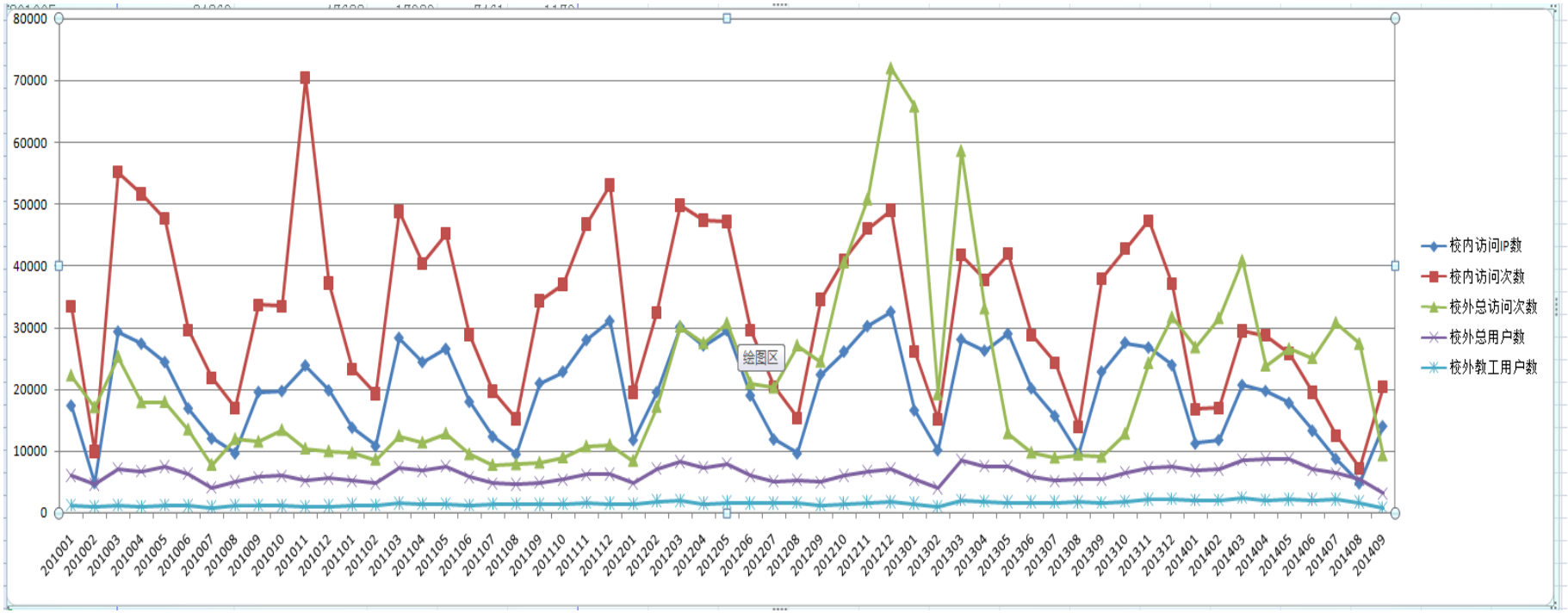
- 流量可视化功能功能模块是对一天的访问数据按照用户类别（教工、博士生、硕士研究生、本科生、其他人员）分时段汇访问量并绘制出来柱状图。





系统的设计与实现

- 电子日志访问统计信息



问题

- 1、校外访问日下载量过大，一般情况下每账户下载量在500M以内，发现少部分账号流量已经达到G，最多日流量达到12G之多。

C4 10938030408

	A	B	C	D	E
1	会话时间	用户账号	总流量	IP数	登陆次数
2	[dbo].[log20130112]	G201007010	12747043404	2	2
3	[dbo].[log20130302]	G201002056	11419271549	498	1003
4	[dbo].[log20130111]	G201007010	10938030408	266	453
5	[dbo].[log20130113]	G201004001	10279019068	496	4697
6	[dbo].[log20130110]	G201007010	10203586743	460	743
7	[dbo].[log20130115]	G201004001	8090001378	834	3182
8	[dbo].[log20130114]	G201004001	7831409135	765	3958
9	[dbo].[log20130314]	G201004006	6932088966	1	1
10	[dbo].[log20130411]	G201001006	6887027255	1423	2473
11	[dbo].[log20130410]	G201001006	6544290546	1259	2287
12	[dbo].[log20140712]	7281	6500027412	546	506
13	[dbo].[log20130116]	G201004001	5965757154	754	1613
14	[dbo].[log20130411]	G201004064	5886959717	771	874
15	[dbo].[log20121211]	G201004001	5876423282	791	1277
16	[dbo].[log20130313]	G201004006	5843700841	1	1
17	[dbo].[log20140902]	7251	5828065917	734	694
18	[dbo].[log20121218]	G201004001	5778978529	659	940
19	[dbo].[log20140825]	5904	5778227901	728	655
20	[dbo].[log20140715]	7281	5595103324	514	406
21	[dbo].[log20130220]	G201002056	5548175536	414	450
22	[dbo].[log20130308]	G201002056	5534027722	542	1323
23	[dbo].[log20121209]	G201004001	5483779952	687	1003
24	[dbo].[log20121230]	G201004001	5300655838	268	483
25	[dbo].[log20130214]	G201002056	5284370984	188	202
26	[dbo].[log20140105]	6624	5255252986	22	28
27	[dbo].[log20140713]	7281	5244567755	537	620
28	[dbo].[log20140714]	7281	5227135046	421	386
29	[dbo].[log20130109]	G201007010	5225809939	423	711
30	[dbo].[log20121213]	G201004001	5133061451	710	921
31	[dbo].[log20130308]	G201004006	5088348748	132	662
32	[dbo].[log20121205]	G201004001	5041697706	604	772
33	[dbo].[log20130102]	G201004001	5017471867	336	457
34	[dbo].[log20140506]	4032	5008260449	700	66
35	[dbo].[log20121121]	G201004001	4919054762	621	970
36	[dbo].[log20130110]	G201004001	4911675305	570	1354
37	[dbo].[log20121207]	G201004001	4858232712	486	800
38	[dbo].[log20121113]	G201004001	4833283250	611	1279
39	[dbo].[log20121114]	G201004001	4822723124	588	1076
40	[dbo].[log20130103]	G201004001	4819546019	332	522
41	[dbo].[log20140323]	7281	4800198074	37	32
42	[dbo].[log20121217]	G201004001	4757069119	702	924
43	[dbo].[log20140903]	6624	4698835300	897	856

Sheet1 Sheet2 Sheet3

问题

- 2、校外访问IP地址多，会话多，单日有多个IP地址，有部分账号的IP地址超过100，最多达到800，会话数也超过100

	A	B	C	D	E	F
1	会话时间	用户账号	总流量	IP数	登陆次数	
2	[dbo].[log20130411]	G201001006	6887027255	1423	2473	
3	[dbo].[log20130410]	G201001006	6544290546	1259	2287	
4	[dbo].[log20130408]	G201001006	4012385737	1234	1813	
5	[dbo].[log20121219]	G201007010	2919394248	1192	1560	
6	[dbo].[log20121220]	G201007010	2880265647	1134	1535	
7	[dbo].[log20130409]	G201001006	4321041772	1043	1559	
8	[dbo].[log20121224]	G201007010	2726060200	1007	1305	
9	[dbo].[log20121221]	G201007010	2461864475	926	1277	
10	[dbo].[log20140903]	6624	4698835300	897	856	
11	[dbo].[log20130115]	G201004001	8090001378	834	3182	
12	[dbo].[log20121223]	G201007010	2648011471	820	953	
13	[dbo].[log20121210]	G201004001	4463638165	807	2205	
14	[dbo].[log20121211]	G201004001	5876423282	791	1277	
15	[dbo].[log20130411]	G201004064	5886959717	771	874	
16	[dbo].[log20130114]	G201004001	7831409135	765	3958	
17	[dbo].[log20130116]	G201004001	5965757154	754	1613	
18	[dbo].[log20140902]	7251	5828065917	734	694	
19	[dbo].[log20140825]	5904	5778227901	728	655	
20	[dbo].[log20121222]	G201007010	2376941871	715	981	
21	[dbo].[log20121213]	G201004001	5133061451	710	921	
22	[dbo].[log20130327]	G201001006	2703422126	706	899	
23	[dbo].[log20130307]	G201002056	1862973377	705	692	
24	[dbo].[log20121217]	G201004001	4757069119	702	924	
25	[dbo].[log20140506]	4032	5008260449	700	66	
26	[dbo].[log20130320]	G201001006	3114767274	700	1749	
27	[dbo].[log20130325]	G201001006	2874926169	699	994	
28	[dbo].[log20121209]	G201004001	5483779952	687	1003	
29	[dbo].[log20130326]	G201001006	2448414500	677	732	
30	[dbo].[log20121016]	G201004001	3428016542	676	752	
31	[dbo].[log20121112]	G201004001	4398559039	668	972	
32	[dbo].[log20121204]	G201004001	4687799546	662	928	
33	[dbo].[log20121218]	G201004001	5778978529	659	940	
34	[dbo].[log20121226]	G201004001	4294032775	656	800	
35	[dbo].[log20121022]	G201004001	3868795649	653	787	
36	[dbo].[log20130407]	G201001006	2779440276	646	1121	
37	[dbo].[log20121212]	G201004001	4266223250	643	869	
38	[dbo].[log20130402]	G201001006	3865664722	639	2064	
39	[dbo].[log20130305]	G201002056	1709231939	637	641	
40	[dbo].[log20130321]	G201001006	3641930591	627	1033	
41	[dbo].[log20121203]	G201004001	3643632693	625	722	
42	[dbo].[log20121121]	G201004001	4919054762	621	970	
43	[dbo].[log20130328]	G201001006	3248511586	620	2038	

Sheet1 Sheet2 Sheet3

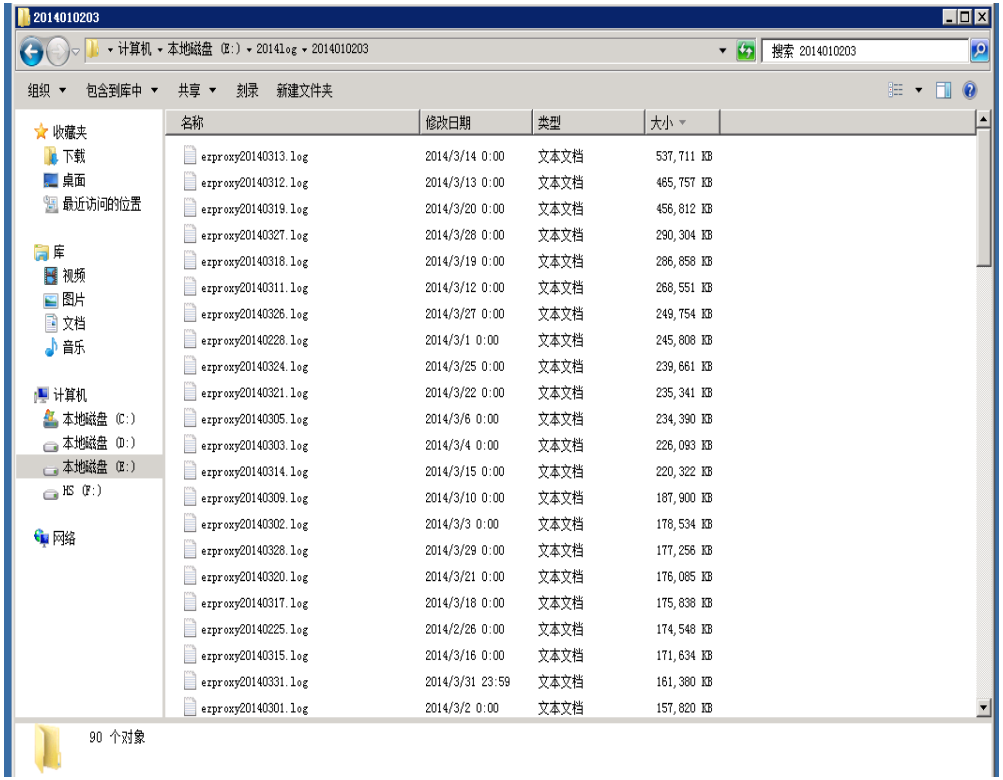
问题

- 3、校外访问单会话多IP，个别账号存在只有一个会话，该会话持续数日没有停止，有多个IP访问。

	A	B	C	D	E	F
1	[dbo].[log20140814]	##	206622493	51	1	
2	[dbo].[log20140815]	##	322181468	53	1	
3	[dbo].[log20140816]	##	48445137	20	1	
4	[dbo].[log20140819]	##	232696467	54	1	
5	[dbo].[log20140820]	##	233132926	48	1	
6	[dbo].[log20140821]	##	287938857	55	1	
7	[dbo].[log20140822]	##	427546698	59	1	
8	[dbo].[log20140823]	##	229281497	23	1	
9	[dbo].[log20140824]	##	201810429	33	1	
10	[dbo].[log20140825]	##	329355959	35	1	
11	[dbo].[log20140826]	##	489525176	165	1	
12	[dbo].[log20140827]	##	132427805	54	1	
13	[dbo].[log20140901]	##	122324898	33	1	
14	[dbo].[log20140902]	##	269912183	41	1	
15	[dbo].[log20140903]	##	122308000	42	1	
16	[dbo].[log20140904]	##	191481724	48	1	
17						

问题

- 4、日志文件大小异常情况，正常情况下每天日志的大小只有200M以内，发现有部分日期日志达到了500M，最高一天2014年3月14日日志达到537M.



解决对策

- 基本解决策略：
  - 每天进行日志分析，发现异常账户，限制该账号的访问权限；
  - 更改管理员口令和服务器口令；
  - 限制一个账号可以登录的IP数量；
  - 限制账号登录时长，限制一个seesion最大活动时长不超过6小时；
  - 流量限制，设置某帐户1小时流量超过500M，挂起24小时。

小结

- 通过使用EZproxy日志分析系统，发现电子资源校外访问异常账户，对异常访问情况分类处理，基本上解决了异常访问导致图书馆电子资源被数据库商临时封掉账号的现象，提高了正常使用电子资源用户的访问效率，保障了图书馆所购买电子资源被安全合法使用。
- 如果EZproxy日志分析系统能够实现实时处理日志，并对异常访问现象报警，有利于及时发现异常访问账户并做出及时的处理，提高系统的可用性。

谢谢！