

Week 3 Sponsor Meeting

Team ADS

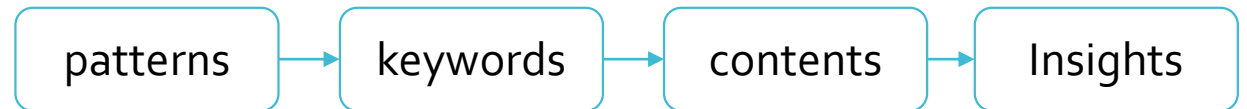
Agenda

- Value extraction from URL
- Work in progress
- Midterm deliverables

Value extraction

❖ Recap

- Systematically find the useful url patterns and manually determine the meaning. Discover rules for user inputs, views and downloads
- Use keywords to identify **search keywords** and **download contents** (all requests to .pdf are considered downloads)



Value extraction

❖ Planned Steps

1. For each domain, summarise all requests to url patterns, starting from the most popular domains
2. For each pattern, find one example. Describe the user action for that example (whether it's a show search result, view item or download item or other).
3. For each pattern, get request counts and user counts (based on sessions).
4. Identify keywords from each pattern and generate regex to look up for content (user inputs)

❖ Execution

1. Database selection and domain mapping:

Identify interested databases based on sponsor preference (and perhaps technical feasibility)

Database	Domains (appearing in sample data)
lawnet	*.lawnet.sg
westlaw	*.westlaw.co.uk *.westlaw.com *.westlaw.co
Ebsco ebooks	*.ebSCOhost.com
MyiLibrary	*.myilibrary.com
ebrary	ebrary.com

Value extraction

❖ Execution

2. Pattern extraction:

- Generalise path and parameter names.
- Path and parameter name set define a pattern

Purposes:

- accurately reflect the URL patterns,
- reduce number of patterns for readability,
- identify book identifiers such as ISBN and doi

Tool:

- Used python **urlparse** library to extract path and parameters

❖ Execution

(continue from last slide...)

However, paths may contain hashed string variables

- replaced them with a placeholder consisting of type

Placeholder	Meaning
\$NUM24	24-digit numerical
\$ALN18	18-char alphanumeric
\$STR26	26-char non-alphanumeric

E.g.

"YnRoX183NzgoMzUxNF9fQU41" → "\$ALN24"
"YnRoX183NzgoMzUxNF9fQU42"

Value extraction

❖ Execution

3. Attach its count and an example to each pattern

pattern	pattern	example			
/\$STR17?	1	https://www.lawnet.sg:443/browserconfig.xml			
/_vti_bin/owssvr.dll?UL&STRMVER&BUILD&CAPREQ&ACT	35	https://www.lawnet.sg:443/_vti_bin/owssvr.dll?			
/lawnet/?	14	https://www.lawnet.sg:443/lawnet/			
/lawnet/c/portal/logout?	1	https://www.lawnet.sg:443/lawnet/c/portal/logoc			
/lawnet/c/portal/logout?referer	50	https://www.lawnet.sg:443/lawnet/c/portal/logoc			
/lawnet/c/portal/render_portlet?currentURL&p_p_lifecycle&_legalresearchresul	8	https://www.lawnet.sg:443/lawnet/c/portal/renc			
/lawnet/c/portal/render_portlet?currentURL&p_p_lifecycle&p_t_lifecycle&p_p_s	22	https://www.lawnet.sg:443/lawnet/c/portal/renc			
/lawnet/c/portal/render_portlet?currentURL&p_p_lifecycle&p_t_lifecycle&p_p_s	155	https://www.lawnet.sg:443/lawnet/c/portal/renc			
/lawnet/c/portal/render_portlet?p_p_static&p_p_lifecycle¤tURL&p_t_life	99	https://www.lawnet.sg:443/lawnet/c/portal/renc			
/lawnet/group/lawnet/\$STR18?p_p_lifecycle&_searchresultpageportlet_WAR_la	12	https://www.lawnet.sg:443/lawnet/group/lawne			
/lawnet/group/lawnet/\$STR18?p_p_lifecycle&_searchresultpageportlet_WAR_la	19	https://www.lawnet.sg:443/lawnet/group/lawne			
/lawnet/group/lawnet/\$STR18?p_p_lifecycle&_searchresultpageportlet_WAR_la	4	https://www.lawnet.sg:443/lawnet/group/lawne			
/lawnet/group/lawnet/\$STR18?p_p_lifecycle&_searchresultpageportlet_WAR_la	2	https://www.lawnet.sg:443/lawnet/group/lawne			
/lawnet/group/lawnet/\$STR18?p_p_lifecycle&_searchresultpageportlet_WAR_la	1	https://www.lawnet.sg:443/lawnet/group/lawne			
/lawnet/group/lawnet/\$STR18?p_p_lifecycle&_searchresultpageportlet_WAR_la	5	https://www.lawnet.sg:443/lawnet/group/lawne			
/lawnet/group/lawnet/\$STR18?p_p_lifecycle&_searchresultpageportlet_WAR_la	5	https://www.lawnet.sg:443/lawnet/group/lawne			
/lawnet/group/lawnet/\$STR18?p_p_lifecycle&p_p_state&_searchresultpageportl	98	https://www.lawnet.sg:443/lawnet/group/lawne			

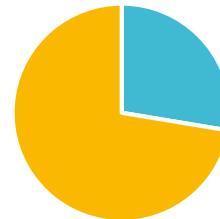
Value extraction

❖ Execution

4. Identify patterns for search, view or download.
Information Gain:

- increased keyword pool for “junk requests”
 - (e.g. 341 of 8098 requests to ebrary.com are font with extensions ttf, woff, eot)

Of sample request



■ Junk ■ Non-junk

❖ Execution

(continue from last slide...)

Information Gain:

- Managed to determine the user action by revisiting requests

<http://www.ebrary.com/lib/smu/docDetail.action?docID=10572555&poo=thailand+%22chinese+diaspora%22&token=967a22b8-85dc-48b5-bfa3-09178cf75492> (from log)



Changed to

<http://site.ebrary.com.libproxy.smu.edu.sg/lib/smu/docDetail.action?docID=10572555&poo=thailand+%22chinese+diaspora%22&token=967a22b8-85dc-48b5-bfa3-09178cf75492>


❖ Execution

(continue from last slide...)

<http://site.ebrary.com.libproxy.smu.edu.sg/lib/smu/docDetail.action?docID=10572555&poo=thailand+%22chinese+diaspora%22&token=967a22b8-85dc-48b5-bfa3-09178cf75492>

Value
extraction

docID=10572555 poo=thailand "chinese diaspora"



MAP OF COCHINCHINA AND CAMBODIA

Library of China Studies : Chinese Diaspora in South-East Asia : The Overseas Chinese in IndoChina (1)

Barrow, Yang C.
Pages: 310
Publisher: LEFAPS
Location: London, UK
Date Published: 2020-2
Language: en

LC Call Number: DS539.C5 -- E37 2812eo
eISBN: 978045721181
pISBN: 9781740761943
Dewey Decimal Number: 305.83
OCLC Number: 706209662

CONTENTS
ACKNOWLEDGMENTS
A NOTE ON TRANSLITERATION
MAPS
INTRODUCTION
1. CHINESE CONGREGATIONS, FRENCH COLONIAL AUTHORITIES, AND THE INDOCHINESE MILIEU
2. CHINESE SUFFRAGE AND CONGREGATIONAL ELECTIONS
3. CONGREGATIONAL LEADERSHIP AND THE POWER OF THE PRESIDENT
4. THE BANKRUPTCY OF CHOLON'S FAMILY, 1871-1913
5. OUR BROTHERS' KEEPERS: MUTUAL AID IN THE THE
6. PRESERVING CHINESE CULTURE: COMMEMORATIVE AND EDUCATIONAL PURSUITS IN THE CONGREGATIONS
7. FINDING THE MIDDLE GROUND: DISPUTE RESOLUTION AND MEDIATION IN THE CONGREGATIONS
8. CONTROL AND CONTESTATION: CRIME, POLICING, AND NARRATION IN THE CONGREGATIONS
9. CHINESE CERIELES
CONCLUSION
NOTES
BIBLIOGRAPHY
INDEX

Library of China Studies : Chinese
Diaspora in South-East Asia : The
Overseas Chinese in IndoChina (1)

❖ Execution

(continue from last slide...)

Database	Domains	Keywords
lawnet	*.lawnet.sg	<ul style="list-style-type: none">•“pdfFileName” for file names (e.g. “[1996] 3 SLR(R) 0371.pdf”)•“contentDocID” for internal content location (e.g. [1985-1986] SLR(R) 0241.xml)•“queryStr=” for query input
westlaw	*.westlaw.co.uk *.westlaw.com *.westlaw.co	*.westlaw.co.uk contains “docguid” but the content is coded with internal ID
Ebsco ebooks	*.a.ebscohost.com *.b.ebscohost.com	“bquery” as parameter name indicates search input
MyiLibrary	*.myilibrary.com	“tid” for title, but it is located with internal ID
ebrary	ebrary.com	“p00” for search queries “docSearch.action?docID=10596700&p00=” is document data for rendering

Value
extraction

Value extraction

❖ Execution

5. Decode query string and content IDs (In progress)

Work in
progress

❖ Execution

1. Tabulate data (i.e. Common Log Format -> CSV)

"uip", "rmtname", "uid", "datetime", "method",
"url", "version", "status", "size", "user-
agent"

+

"domain", "protocol", "port", "type", "keyword"

+

(User demographic data)

Work in
progress

2. Text mining

3. Analysis with MS SQL Data Tool

Deliverables by Midterm

1. Interim report
2. Presentation
3. Tasks:
 1. descriptive analysis report on student data and cleaned log data
 2. Progress on text mining