

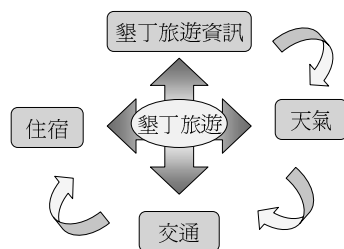
利用全球資訊網之使用者瀏覽 行為探勘網頁流之技術探討

陳武宏 吳宜鴻 陳良弼
國立清華大學資訊工程所

Outline

- Motivation
- Introduction
- SPII Algorithm
- Experiment
- User Interface
- Conclusion & Future Work

Motivation (1)



- 瀏覽目的
 - 計劃週末旅遊，如墾丁旅遊
- 瀏覽目的包含許多的查詢主題
 - 當地旅遊資訊，交通，天氣與旅館
- 查詢主題間具有某種瀏覽次序
 - 當地旅遊資訊→天氣預報→交通→旅館

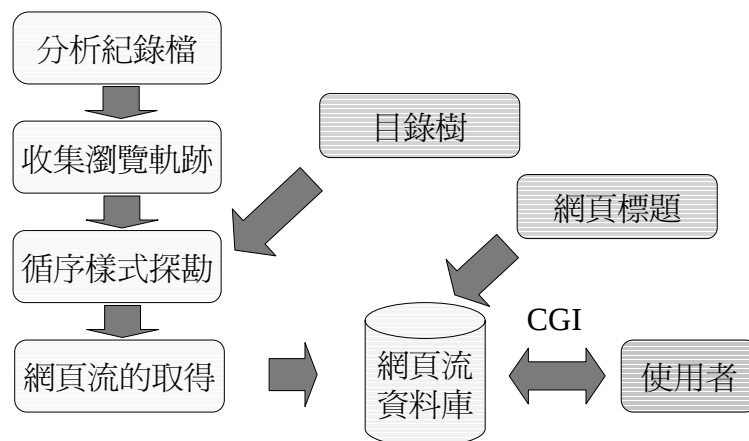
Motivation (2)

- 網頁流 (Webflow)簡單的定義
 - 一群具有瀏覽次序的網頁

- 例子

<http://www.ktnp.gov.tw/> →
<http://www.cwb.gov.tw/> →
<http://www.railway.gov.tw/taiwan/tai17.html> →
<http://www.caesarpark.com.tw/caesar/kenting.htm>

Introduction



Server Log

Access-time Source IP

890984441.324 0 140.117.11.12 UDP_MISS/000 76

ICP_QUERY http://www.yam.org.tw/b5/yam/ - NONE/- -

Destination URL

簡單的描述紀錄檔

(S₁,9,A) (S₂,10,D) (S₁,11,D) (S₂,13,E) (S₃,15,D) (S₃,17,A)
 (S₂,18,E) (S₃,20,E) (S₃,24,E) (S₃,25,S) (S₁,27,E) (S₂,28,A)
 (S₁,30,E) (S₃,32,E) (S₃,34,E) (S₃,36,E) (S₃,38,E) (S₃,40,E) (S₃,42,E) (S₃,44,E) (S₃,46,A) (S₃,49,J) (S₂,51,U)
 (S₃,53,U) (S₃,55,U) (S₃,57,U) (S₃,59,U) (S₃,61,U) (S₃,63,U) (S₃,65,U) (S₃,67,U) (S₃,69,U) (S₂,70,W) (S₂,72,F)...

Navigation Path

- Collect the Browsing Sequences

User	Browsing sequence
S ₁	(A,9) (D,11) (F,24) (E,27) (S,41)
S ₂	(D,10) (E,13) (S,18) (A,28) (Y,45) (U,51) (K,66) (W,70) (F,72)
S ₃	(D,15) (A,17) (E,22) (S,25) (F,42) (A,46) (J,49) (F,56) (S,57)

- Collect the Navigation Paths

[Yan96] user session, timeout

max-gap is 15

Pid	Navigation Path
1	(A,9) (D,11) (F,24) (E,27) (S,41)
2	(D,10) (E,13) (S,18) (A,28)
3	(Y,45) (U,51) (K,66) (W,70) (F,72)
4	(D,15) (A,17) (E,22) (S,25)
5	(F,42) (A,46) (J,49) (F,56) (S,57)

Sequential Pattern Mining

■ SPII Algorithm

- Sequential Patterns discovery using Inverted-list and Intersection operation

■ Inverted List

■ Intersection Operation

■ Processing

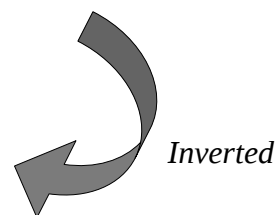
- step I : FS₁
- step II : FS₂
- step III: FS_k, k≥3

SPII Algorithm (1)

■ Step I: FS_1 (1)

Pid	Navigation Path
1	(A,9) (D,11) (F,24) (E,27) (S,41)
2	(D,10) (E,13) (S,18) (A,28)
3	(Y,45) (U,51) (K,66) (W,70) (F,72)
4	(D,15) (A,17) (E,22) (S,25)
5	(F,42) (A,46) (J,49) (F,56) (S,57)

Key	Inverted List (pid, tid)
A	(1,9)(2,28)(4,17)(5,46)
D	(1,11)(2,10)(4,15)
E	(1,27)(2,13)(4,22)
F	(1,24)(2,29)(3,72)(5,42)(5,56)
J	(1,42)(5,49)
K	(3,66)
S	(1,41)(2,18)(4,25)(5,57)
U	(3,51)
W	(3,70)
Y	(3,45)



Inverted
List

SPII Algorithm (2)

■ Step I: FS_1 (2)

Min-sup = 4

Key	Support Count	Inverted List (pid, tid)
A	4	(1,9)(2,28)(4,17)(5,46)
D	3	(1,11)(2,10)(4,15)
E	3	(1,27)(2,13)(4,22)
F	4	(1,24)(2,29)(3,72)(5,42)(5,56)
S	4	(1,41)(2,18)(4,25)(5,57)

SPII Algorithm (3)

■ Step II: FS_2 (1)

Key	Inverted List (<i>pid</i> , <i>tid</i>)
A	(1,9) (2,28) (4,17) (5,46)
D	(1,11) (2,10) (4,15)



Key	Inverted List
A→D	(1,11)
D→A	(2,28) (4,17)

SPII Algorithm (4)

■ Step II: FS_2 (2)

將 FS_1 兩兩做交集

Key	Support Count	Inverted List (<i>pid</i> , <i>tid</i>)
D→E	3	(1,27)(2,13)(4,22)
D→S	3	(1,41)(2,18)(4,25)
E→S	3	(1,41)(2,18)(4,25)
A→F	3	(1,24)(2,29)(5,56)
A→S	3	(1,41)(4,25)(5,57)

SPII Algorithm (5)

■ Step III: FS_k , $k > 2$

針對(k-1)-Prefix相同的sequential pattern做交集

1-prefix: D

D→E	3	(1,27)(2,13)(4,22)
D→S	3	(1,41)(2,18)(4,25)
A→F	3	(1,24)(2,29)(5,56)
A→S	3	(1,41)(4,25)(5,57)
A→F→S	2	(1,41)(5,57)

Intersection Operation

Case I: $FS_k \rightarrow FS_{k+1}$

Freq. 3-Seq.	Inverted List (pid, tid)
D→H→A	(1, 2) (6, 4) (9, 10) (10, 11)
D→H→B	(1, 3) (2, 5) (6, 7) (9, 8) (10, 14)

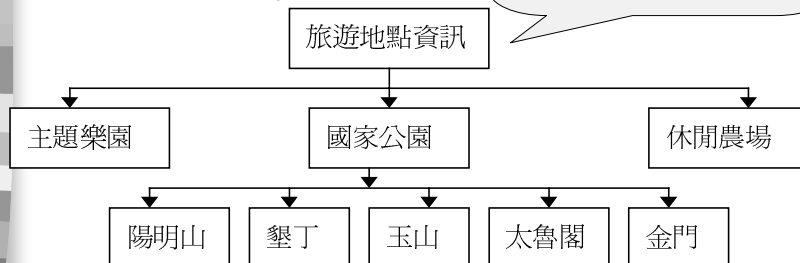
2-Prefix: D→H

Freq. K-Seq.	Inverted List (pid, tid)
D→H→A→B	(1, 3) (6, 7) (10, 14)
D→H→B→A	(9, 10)

Semantic Taxonomy(1)

- Support 太低
- 語意化的網頁流
 - Title
 - Taxonomy

- is-a Concept Hierarchy
- Data exist in any level



Semantic Taxonomy(2)

- 網頁流 (WebFlow)
 - 一群具有瀏覽次序的網頁及其標題或目錄樹上的概念節點
- 關鍵字「墾丁」的語意化網頁流
 - for Title
 - 歡迎蒞臨墾丁國家公園→中央氣象局全球資訊網→墾丁凱薩飯店→臺灣鐵路局--網路訂票
 - for Taxonomy
 - 墾丁國家公園→天氣預測→旅館→鐵路網路訂票



Cross-Level Sequential Pattern Mining

- The same as the SPII Algorithm
 - Max-sup
 - Redundant Sequential Pattern



Experiment

- Pre-processing
 - 去除不必要的檔案，如 gif,jpg,mov,class, mid
 - Get the Navigation Path with one hour
- After Pre-processing
 - 2409 Navigation Paths
 - 614673 Web Pages

站內互連		
81991-81837-81844 25		
81991	Http://www.ftv.com.tw/index.htm	民視新聞資訊網- 線上即時新聞
81837	Http://www.ftv.com.tw/FTVFilter/rindex.htm	民視小品
81844	Http://www.ftv.com.tw/FTVMisc/rindex.htm	民視花架
頁框效應		
183955-183953-183952 30		
183955	Http://140.113.6.80/~u8624535/chat/chatmsg.html	
183953	Http://140.113.6.80/~u8624535/cgi-bin/chataccess.cgi	
183952	http://140.113.6.80/~u8624535/cgi-bin/chat.cgi	
不明查詢主題		
103209-102284-142017 25		
103209	http://home.microsoft.com/intl/tw/	歡迎進入 Internet
102284	http://home.netscape.com/	Netcenter
142017	http://209.150.130.148/hsu/top.html	
一個查詢主題		
124227-124265-124277 25		
124227	http://www.railway.gov.tw/	臺灣鐵路管理局
124265	http://www.railway.gov.tw/taiwan/index.html	臺灣鐵路管理局
124277	http://www.railway.gov.tw/taiwan/tai17.html	臺灣鐵路管理局--網路訂票
多重查詢主題		
207069-90005-90140 24		
207069	http://www.music.com.tw/broadway/top10/top10s.html	飛行音樂網路唱片行 TOP20
90005	http://www.cts.com.tw/channel/news/weather/weather.htm	華視氣象台
90140	http://www.cts.com.tw/educ/lesson/lesson.htm	空中商行專課程查詢

Using the Semantic Taxonomy		
80383-100000203-100000801 124		
100000203	廣播	
100000801	音樂	
80383	奇摩搜尋引擎	
100000708-100000907-100000011-51382 112		
100000708	入口網站	
100000907	新聞	
100000011	金融投資	
51382	HitBOX 入口網站	
100000302-100000702-100000301 126		
100000302	搜尋引擎	
100000702	軟體廠商	
100000301	軟體下載	

探勘時平均所需的時間 (單位: 秒)

Min-sup \ K	k-sequential patterns					Total
	1	2	3	4	5	
1.00% (24)	812.97	13283.76	187.92	4.02	0.16	14288.83
1.25% (30)	805.59	5035.84	20.90	0.25		5862.58

推倒出來的網頁流數量

min-sup \ K	Frequent k-Sequential patterns			
	1	2	3	4
1.00% (24)	810	766	133	12
1.25% (30)	422	191	20	

min-sup count 為 24 與 30 時，Sequential Patterns ($k \geq 2$) 的數量分別為 700 與 173。

專家網頁流 (Expert Webflows)

10000003-10000009-10000005-124277

10000003	Http://www.ktnp.gov.tw/ (歡迎蒞臨墾丁國家公園)
10000009	Http://www.cwb.gov.tw/ (中央氣象局全球資訊網)
10000005	Http://www.caesarpark.com.tw/caesar/kenting.htm (墾丁凱撒飯店)
124277	Http://www.railway.gov.tw/taiwan/tai17.html (臺灣鐵路管理局--網路訂票)

10000010-100000009-10000011

10000010	Http://www.ysnp.gov.tw/ (玉山國家公園管理處)
10000009	新聞
10000011	Http://www.dpp.org (民進黨)

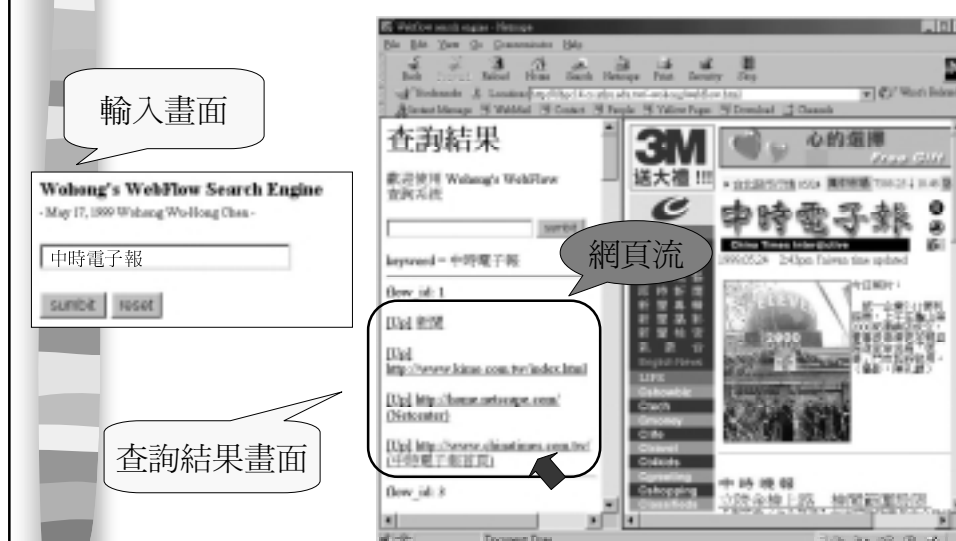
10001106-10000008-100000911

10001106	股市行情
10000008	Http://www.esunbank.com.tw/ (玉山銀行)
100000911	財經新聞

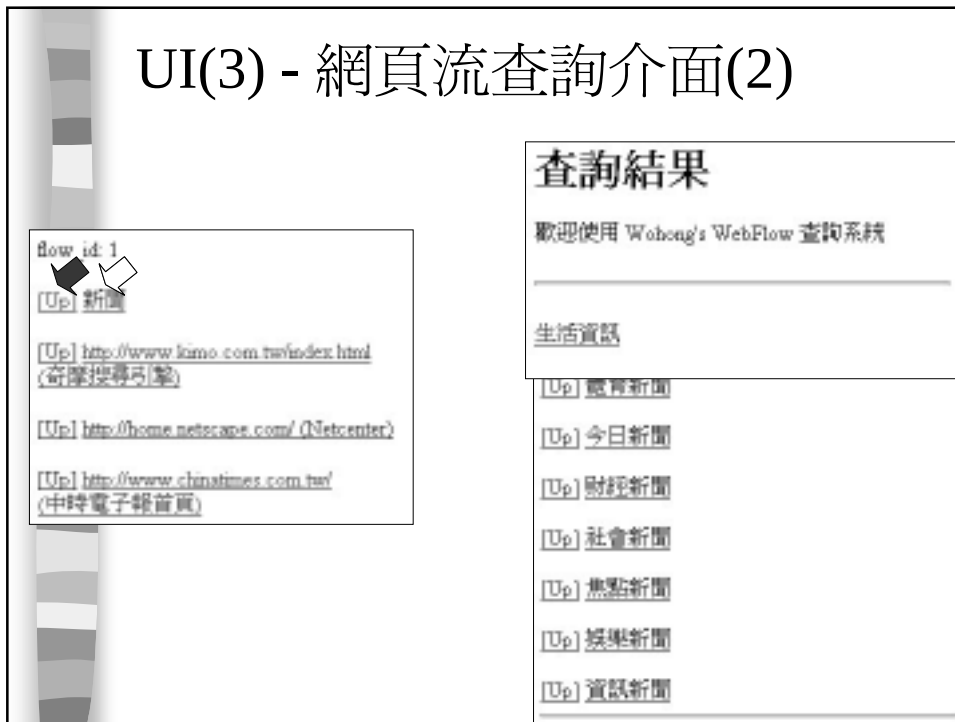
User Interface (1)

- 網頁流查詢介面
- 整合傳統的搜尋引擎

UI(2) - 網頁流查詢介面(1)



UI(3) - 網頁流查詢介面(2)



The screenshot displays a web interface with a sidebar on the left and a main content area on the right. The sidebar contains a box labeled 'flow id: 1' with a cursor icon and a list of links, each preceded by '[Up]'. The main content area has a header '查詢結果' (Search Results) and a sub-header '歡迎使用 Woheng's WebFlow 查詢系統' (Welcome to use Woheng's WebFlow search system). Below this is a section titled '生活資訊' (Life Information) with a list of links, each preceded by '[Up]'. The links in the sidebar are: '新聞' (News), 'http://www.kimo.com.tw/index.html (奇摩搜尋引擎)', 'http://home.netscape.com/ (Netcenter)', and 'http://www.chinatimes.com.tw/ (中時電子報首頁)'. The links in the main content area are: '體育新聞', '今日新聞', '財經新聞', '社會新聞', '焦點新聞', '娛樂新聞', and '資訊新聞'.

flow id: 1

[Up] 新聞

[Up] <http://www.kimo.com.tw/index.html>
(奇摩搜尋引擎)

[Up] <http://home.netscape.com/> (Netcenter)

[Up] <http://www.chinatimes.com.tw/>
(中時電子報首頁)

查詢結果

歡迎使用 Woheng's WebFlow 查詢系統

生活資訊

[Up] 體育新聞

[Up] 今日新聞

[Up] 財經新聞

[Up] 社會新聞

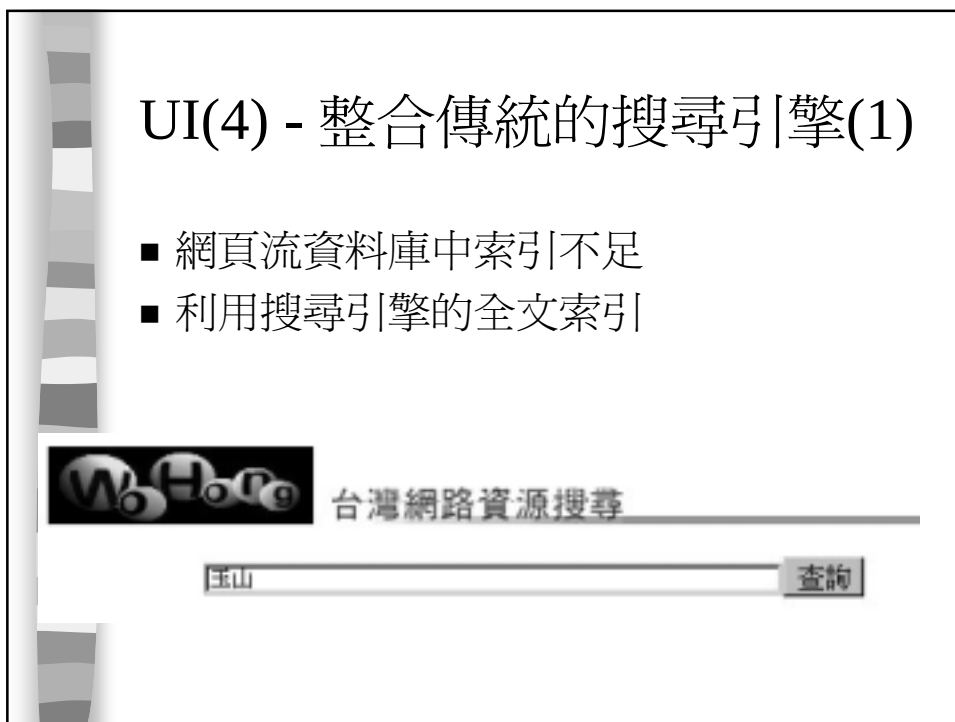
[Up] 焦點新聞

[Up] 娛樂新聞

[Up] 資訊新聞

UI(4) - 整合傳統的搜尋引擎(1)

- 網頁流資料庫中索引不足
- 利用搜尋引擎的全文索引



The screenshot shows a search engine interface. At the top, there is a logo with the letters 'W', 'O', 'H', 'O', 'O' in a stylized font, followed by the text '台灣網路資源搜尋'. Below this is a search bar with the placeholder text '玉山' and a button labeled '查詢'.

W O H O O 台灣網路資源搜尋

玉山 查詢

UI(5) - 整合傳統的搜尋引擎(2)



Conclusion

- WebFlow
- SPII Algorithm
 - Taxonomy
- Expert WebFlow
- Query Interface
- Problem
 - Proxy server log
 - Semantic Taxonomy



Future Work

- Improve SPII algorithm
 - speed, memory
- Incremental
- Taxonomy