Index Structures of User Profiles for Efficient Web Page Filtering Services

Y. H. Wu **A. L. P. Chen**

Database Laboratory
Department of Computer Science
National Tsing Hua University

4/13/2000

Outline

- **Introduction**
- **Related Approaches**
- **Our Approaches**
- **Comparisons**
- **Conclusion**

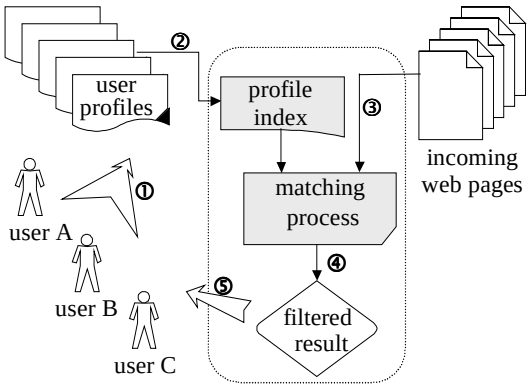
Introduction

- **Motivation**
  - Searching problem on the WWW
    - ☞**search engine**
    - ☞**meta-search engine**
  - The performance may get worse if the number of web pages grows rapidly
- **Goal**
  - Filtering approach
    - ☞**find the matched profiles for each web page**

1

Introduction

- **A Web Page Filtering Service**



2

Introduction

- **An Example**
  - Conjunction of keywords
  - Boolean model
  - Matches
    - ☞ **{P_1, P_4}**

Profile	Keyword
P_1	a b
P_2	a d
P_3	a d e
P_4	b f
P_5	c d e f

Example page
a b c f

3

Related Approaches

- **The Counting Method**
  - Keyword array: inverted lists
  - Profile arrays: TOTAL, COUNT
  - Matching criteria: COUNT=TOTAL

Profile

	TOTAL	COUNT
P_1	2	*2*
P_2	2	*1*
P_3	3	*1*
P_4	2	*2*
P_5	4	*2*

Keyword

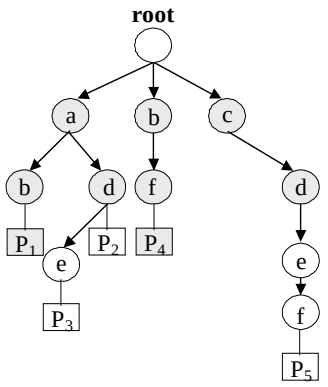
a	P_1 P_2 P_3
b	P_1 P_4
c	P_5
d	P_2 P_3 P_5
e	P_3 P_5
f	P_4 P_5

Example page
a b c f

4

Related Approaches

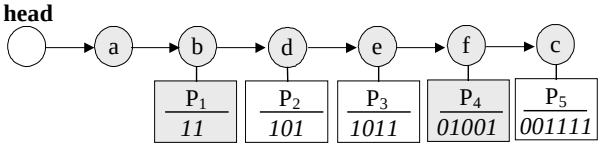
- **The Tree Method**
  - K-nodes: internal nodes
  - P-nodes: leaf nodes
  - External path
    - **root → p-node**
    - **a profile**
  - Matches
    - **root → a → b → P_1**
    - **root → b → f → P_4**



Example page
a b c f

Our Approaches

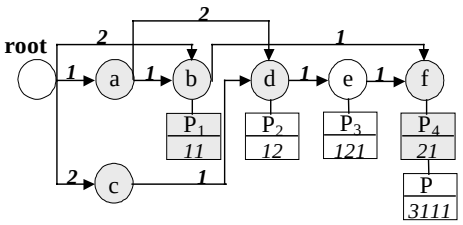
- **Method 1**
  - Index path with path signatures
  - Path signature of the example page: 110011
  - Matches
    - **at node b: P_1 AND 11 = P_1**
    - **at node f: P_4 AND 11001 = P_4**



Example page
a b c f

Our Approaches

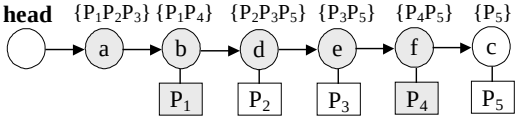
- **Method 2**
  - Index graph with path signatures
  - Matches
    - ☞**root → a → b → P_1: 11**
    - ☞**root → b → f → P_4: 21**



Example page
a b c f

Our Approaches

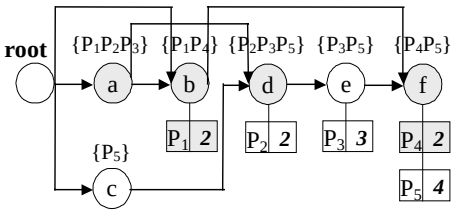
- **Method 3**
  - Index path with profile sets
  - Candidate set
  - Target set ⇐ candidate set ∩ profile set
    - ☞**at node b: T = {$P_1P_2P_3P_4P_5$} ∩ {P_1P_4} = {P_1P_4}**
    - ☞**at node d: T = {$P_2P_3P_4P_5$} ∩ {$P_2P_3P_5$} ={$P_2P_3P_5$}**



Example page
a b c f

Our Approaches

- **Method 4**
  - Index graph with profile sets
  - Path length
  - Matching criteria: matching of keywords $\wedge$ profile id $\subset$ target set $\wedge$ equal path length



Example page
a b c f

Comparisons

- **Notation**

Symbol	Description
P	The set of all profiles
K	The set of all distinct keywords
n	Average number of keywords in a profile
f	Average number of profiles in which a specific keyword is specified
m	Average number of keywords to represent a web page

Comparisons

- **Summary**

Criteria \ Approaches	Counting Method	Tree Method	Method 1	Method 2	Method 3	Method 4
Duplication of information	profile	keyword	no	no	no	no
Sorting of keywords	no	yes	no	yes	no	yes
Storage space	O(|P|+f|K|)	O(n|P|)	O(|P|+|K|)	O(|P|+|K|)	O(|P|+|K|)	O(|P|+|K|)
Insertion/Deletion time	O(nf)	O(nf)	O(n+f)	O(nf)	O(n+f)	O(n+f)
Matching time	O(mf+|P|)	O(mf)	O(mf)	O(mf)	O(mf)	O(mf)
Modification time	O(nf)	O(nf)	O(n+f)	O(nf)	O(n+f)	O(n+f)

11

Conclusion

- **Contribution**
  - Four new methods for profile indexing
  - Comparisons by complexity analyses
  - Efficient web page filtering service
- **Future Work**
  - Prototype system for real data
  - Dissemination and display of the filtered results
  - More predicates for specifying the user profiles

12