Personalized Information Delivery on the Static and Mobile Web

Dik Lun Lee
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
Nov 2, 2009

Objectives of this Talk

- Traditional IR vs. mobile IR
- Information Push as the default information access model
- Estimating user interests via search engine clickthroughs

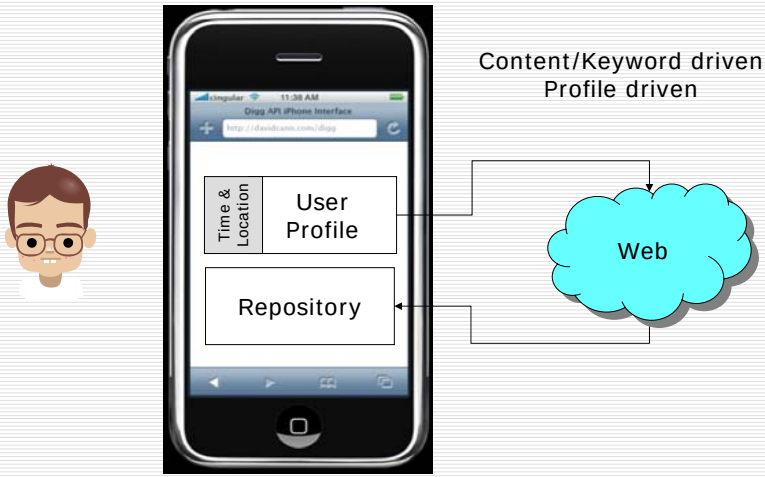
Web Search vs. Mobile Search

- Simple mobile search model
  - Shrink the desktop/web search onto a mobile device
  - Voice I/O, auto-completion (Google Suggest), query suggestion, aiming at reducing the user I/O effort
  - Vertical search services to cater for common mobile search
    - Route, restaurant, directory search
  - Yahoo Go!, Google Mobile
- Proactive model
  - Up-to-date and relevant information are pushed to mobile device, replacing explicit requests by local browsing
  - Make possible by large local storage and high bandwidth
  - Require profiling user interests and context awareness
  - Best-effort suggestions

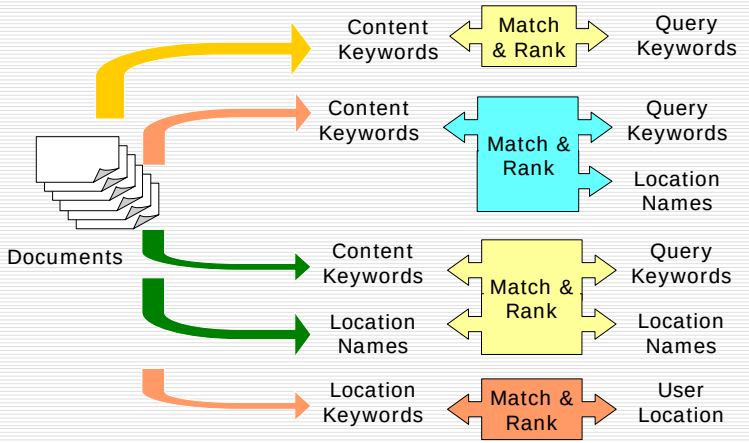
Proactiveness: While you are shopping…

- Do you want your mobile devices to be loaded with useful coupons, store information and sales items?
- What about a bookstore offering a discount on a book that you browsed on Amazon yesterday?
- What about the time for the next bus that you take every day?
- … …

User Profiling: Online vs Mobile



Location-Based Search



What does the user really want?

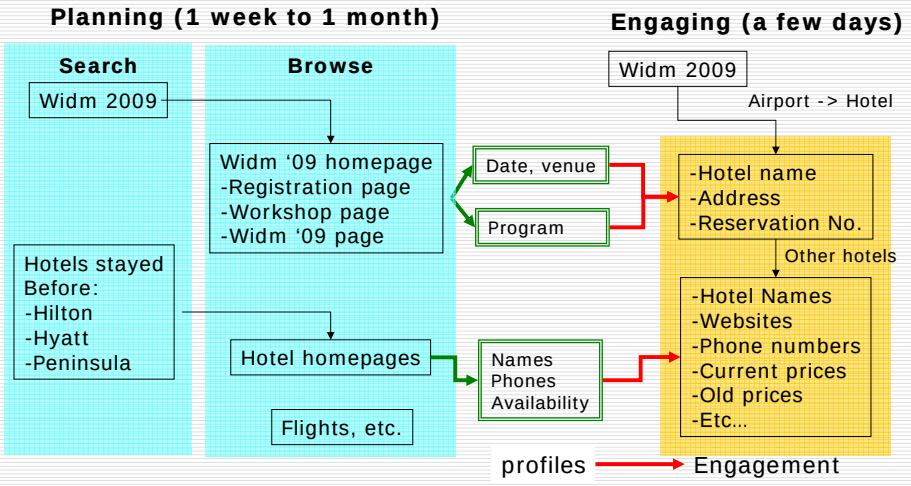
User Profiling as a Universal Requirement

- Web/desktop search, mobile search, pro-active or passive, knowing the user interest is very important
  - More relevant search results
  - Suggest relevant queries
  - Display related information
- Question: how to collect, derive, represent, utilize and refine

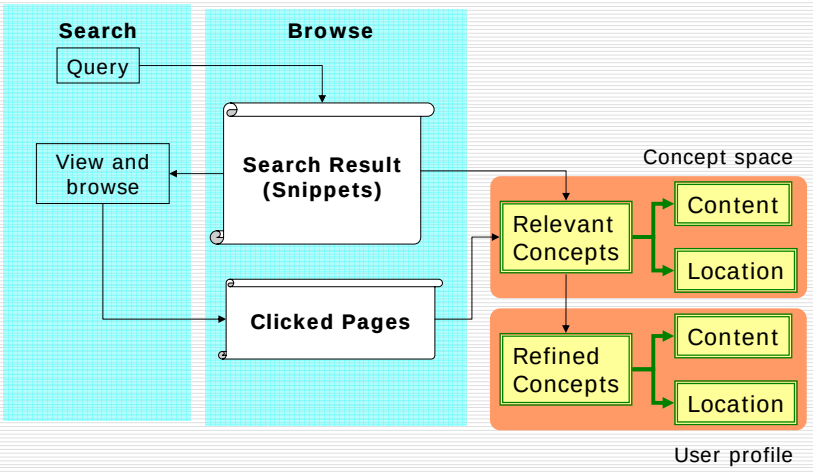
User Profiling: Online vs Mobile

- Comprehensive profiling
  - Online tracking: search and web browsing
    - Predictive of future events and needs
  - Mobile tracking
    - Predictive of local interests (both temporal and spatial) and action items
    - Location semantics: semantic location modeling

User Profiling – An Example



User Profiling – Concept Extraction



Clickthrough Data

Doc	Clicked	Search results
d_1	√	Apple Computer
d_2		Apple – Quicktime
d_3		Apple – Fruit
d_4	√	Apple - Mac
d_5		History of Apple Computer
d_6		Apple Mac News
d_7		Apple tree
d_8	√	Apple – Support
d_9		AppleInsider

- Preference mining: Given the clickthrough data, what is the user interested in?

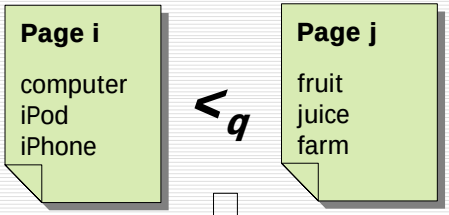
Inferring User Preferences (Joachims)

- Assumption: Users read the results from top to bottom, click on relevant results and skip non-relevant results
- E.g., the user clicked #1, #4 and #8, we can infer that #1, #4 and #8 are relevant while #2, #3, #5, #6 and #7 are non-relevant
- It cannot infer if #9 and #10 are relevant or not since it is not sure if the user has examined the items below the last click
- Instead of a relevant vs non-relevant decision, the following user preferences can be inferred:
  - #1 over #2, #3, #5, #6 and #7
  - #4 over #2, #3, #5, #6 and #7
  - #8 over #2, #3, #5, #6 and #7
  - no further preference can be concluded

Result list:

1. **Apple Store √**
2. Apple - QuickTime
3. Apple - Fruit
4. **Apple .Mac √**
5. www.applehistory.com
6. Adam Country Nursery
7. Apple cookbook
8. **Apple Support √**
9. ... ...
10. ... ...

From Page Preference to Concept Preference



[computer, iPod, iPhone] $<_q$ [fruit, juice, farm]

Feature vector / User profile

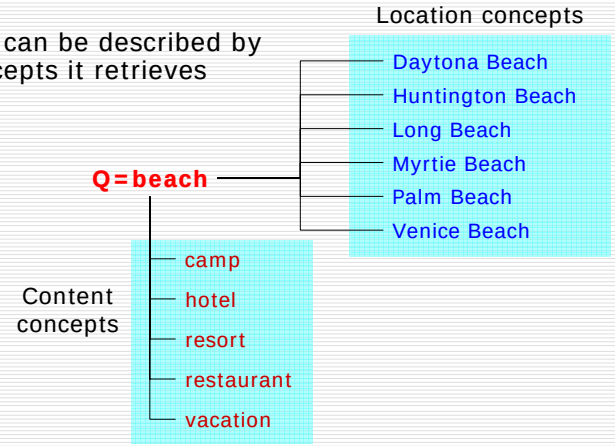
a_j	computer	iPod	iPhone	fruit	juice	farm	...
weight	1	1	1	-1	-1	-1	0

Now we know concepts are used to profile a user's interests

How to know if a concept is content or location related?

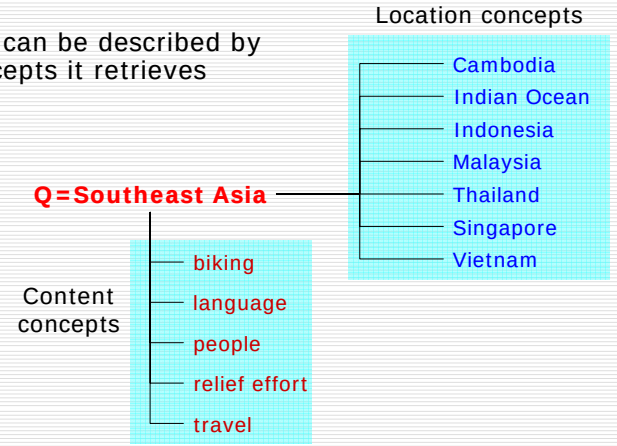
Example: Location Query

- A query can be described by the concepts it retrieves



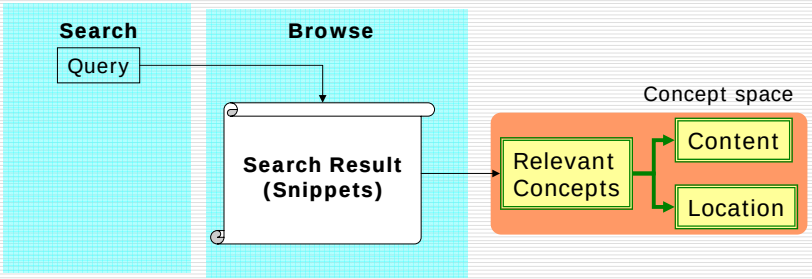
Example: Location Query

- A query can be described by the concepts it retrieves



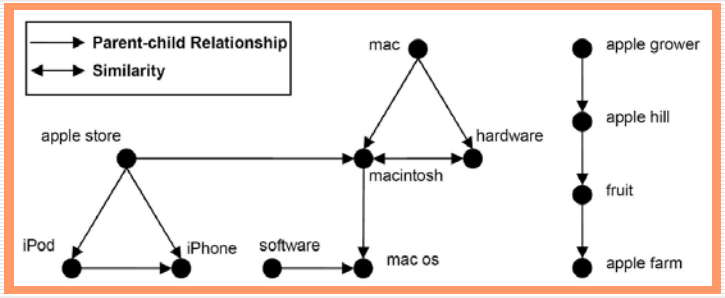
Concept Extraction

- The longest sequence of words appear in > *n* snippets.
  - Snippets are considered by the search engine as the most important document segment relevant to a query
  - Identify longest meaningful phrases in the snippets

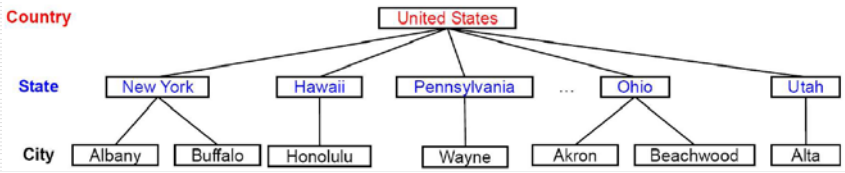


Concept Ontology

- Content concepts are organized into hierarchy
  - Similarity(x,y) => x and y coexist in the same snippets *m* times
  - Parent-Child(x,y) => x coexists with many concepts, including y but not vice versa



Location Ontology



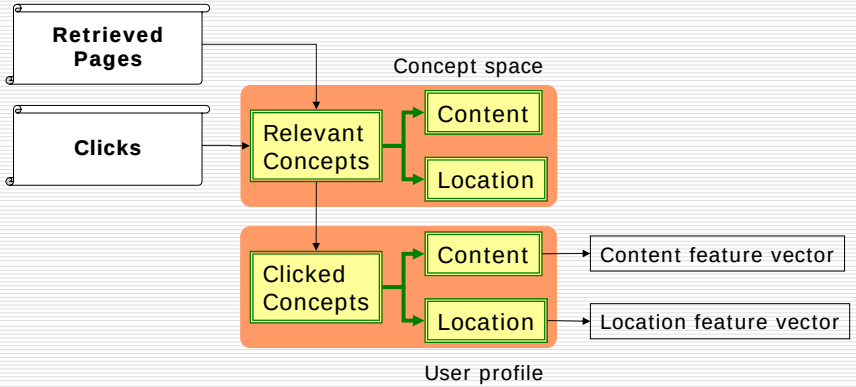
STATISTICS OF THE LOCATION ONTOLOGY

No. of Countries	7	Total No. of Nodes	16899
No. of Regions	190	Country-Region Edges	190
No. of Provinces	6699	Region-Province Edges	1959
No. of Towns	10003	Province-City Edges	14897

- Prebuilt location hierarchy
- A concept that matches a node is a location concept

User Behaviors

- User behaviors are described by the concepts they clicked
- Content feature vector || Location feature vector



Is a concept either 100% content or 100% location?

Hong Kong ⇒ ~100% location
Programming ⇒ ~100% content
Java ⇒ half-half ???
HKUST ⇒ 80-20 ???
What about ``Books'', ``Physics'', ... ?

Measuring Content and Location Richness

- How much content and location is a query associated to?
- A concept is location oriented if it is associated with a large number of different locations
- A concept is content oriented if it is associated with a large number of different concepts
- A concept may be both content and location oriented with different degrees of richness
- Content entropy: $H_C(q) = -\sum_{i=1}^{k} p(c_i) \log p(c_i)$
- Location entropy: $H_L(q) = -\sum_{i=1}^{m} p(l_i) \log p(l_i)$

Measuring Content and Location Interests

- Clicked content entropy: $H_{\overline{C}}(q,u) = -\sum_{i=1}^{t} p(\overline{c}_{iu}) \log p(\overline{c}_{iu})$
- Clicked location entropy: $H_{\overline{L}}(q,u) = -\sum_{i=1}^{v} p(\overline{l}_{iu}) \log p(\overline{l}_{iu})$
- Given a concept, is a user interested in the content and/or the location aspects of the query? Consider ``Java'', ``apple'', etc.
  - Did the user click on a large number of various locations?
  - Did the user click on a large number of various concepts?

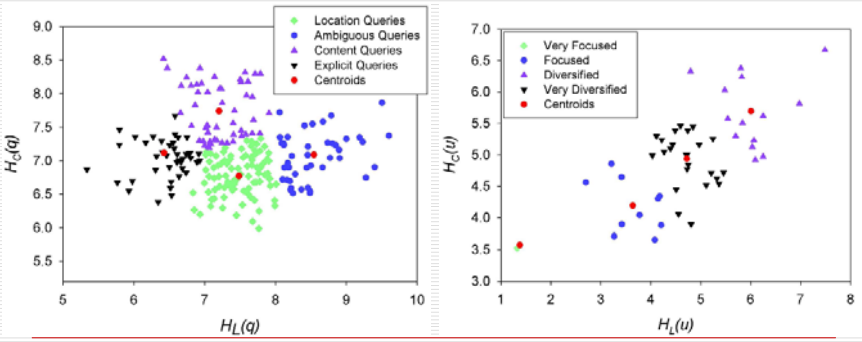
Query Classes

- Four combinations of content and location entropies:
  - low/low, high/low, low/high and high/high
  - Explicit, content, location, and ambiguous queries
  - Note: Beijing is not entirely location-oriented and Manchester is rich in content as well !!!

Explicit	$H_C(q)$	$H_L(q)$	**Location**	$H_C(q)$	$H_L(q)$
Canon	6.6921	5.9792	Beijing	6.6492	8.0116
IBM	6.8683	5.3383	Campus Life	6.7888	7.8522
Sony	6.6698	5.7683	Overseas Study	6.8080	7.8934
Content	$H_C(q)$	$H_L(q)$	**Ambiguous**	$H_C(q)$	$H_L(q)$
Disney Movie	8.1204	6.8074	Manchester	8.3160	7.5705
Dual Core	8.1538	6.9552	Apartment	8.2124	7.5031
Programming	8.3827	6.4718	Shopping	8.0739	7.2339

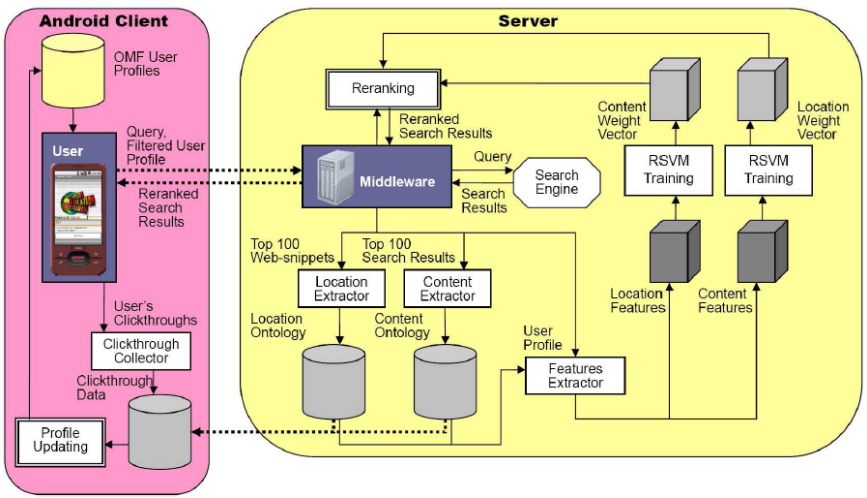
Query and User Classes

- Users can be grouped based on their clicked content and location entropies (50 users and 250 queries)
  - Very focused, focused, diversified and very diversified



Mobility and User Locations

- Searching on desktop:
  - Capture user's interests on locations, not his current location
- Searching on mobile:
  - Capture user's interests around his current location
  - When you are at AsiaWorld Expo, you want to find events and restaurants at or around it
  - But … can we be sure that this is always the case? When you are at the Kowloon Station, you may just want to find information about AsiaWorld Expo or the Airport, not anything around Kowloon Station !!!
- Combination of a user's locations and location interests
  - User had searched and browsed pages about AsiaWorld Expo
  - But then would this be too restricted?



Summary

- The employment of both content and location preferences enhances search precision
- Location-based personalization: If a user is known to be interested in Japan, pages known to be associated with Japan will be ranked higher for his queries even if a query has no indication about Japan (e.g., music)
- Group-based personalization
  - Clicks will not be diluted by naive users
- Group-based recommendation
  - A focused user knows what he/she is doing on the query, and hence his/her clicks (endorsement) benefit other users more

Research Problems

- Better integration of online and mobile activities for better profiling of user interests
  - What indicates what?
  - Selecting the profile concepts to support an engagement
- Consideration of other high-level concepts:
  - Person names, time, actions, goals, plans, events and transactions
- Community-based concept extraction
  - Noise elimination and user segmentation
- Privacy issues
  - Approximate user profiles
- Collaborative filtering
  - User ⇔ Query ⇔ Concepts

Thanks !!!

Q / A