University of Alberta

Database Laboratory

Combining 3 Information Channels for a Better Non-Intrusive Recommender System

Osmar R. Zaïane

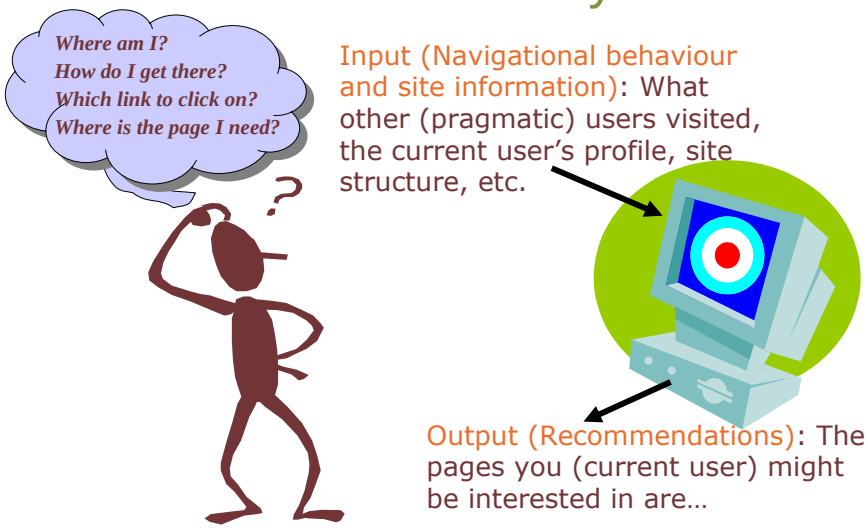
Data Mining Group

University of Alberta - Canada

Web Recommender Systems

- People are finding it increasingly frustrating to locate and access on-line information or resources.
- There are a number of ways to assist on-line users to find the right resources they need.
  - Search Engines
  - The Adaptive Website
  - Visualization Tools
  - Web Recommender Systems (WRSs)
- WRS: anticipating the information needs of on-line users and providing them with recommendations to facilitate and personalize their navigation.

Basic Motivation for a Recommender System



Input (Navigational behaviour and site information): What other (pragmatic) users visited, the current user's profile, site structure, etc.

Output (Recommendations): The pages you (current user) might be interested in are…

Outline

- More on Motivation
- Contributions
- Mission and Mission Identification
- System Framework
- System Evaluation
- Conclusion

Rating-based vs. Activity-based

- Rating-based Web Recommender Systems
  - Recommending users web resources which are highly favored by previous users who have similar preference or taste.
  - Relying heavily on explicit user input to generate the user model, which is either unavailable or considered intrusive.
- Activity-based Web Recommender Systems
  - The user profile is built based on the access log.
  - a web usage recommender system which focuses solely on web server logs has its own problems:
    - Incomplete Information Problem
    - Incorrect Information Problem
    - Persistence Problem

To avoid being intrusive, we focus on activity-based recommendation.

At the heart of Recommender Systems are Collaborative Filtering Algorithms that rely on correlation between individuals

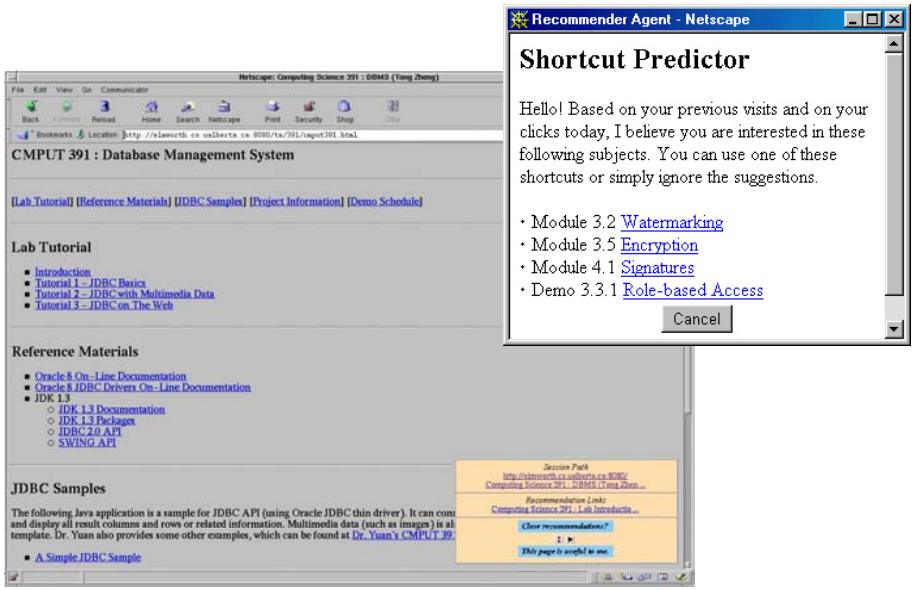
The basic idea of collaborative filtering is people recommending items to one another.

Ratings of Books	1	2	3	4	5	6	7	8
Jane	5	3	3	4	2	1		
Alexander	3	4	2	3	4	5	1	3
Amelia	4	3	1	2	4	2	4	1
Duncan	4	2	1	3	4	1	5	2

Profile

- Jane & Duncan: correlation = .52
- Jane & Alexander: correlation = -.67
- Jane & Amelia: correlation = .23

Recommendations for Jane: Book 7



Recommender with Association Rules and other ideas

- What if we have no ratings?

1. Based on transactions $user_i$ visited $<p_1, p_2,\ldots>$
- If $User_x$ accesses p_a and $<p_a, p_b>$ is frequent itemset in the access logs and $User_x$ never visited p_b then suggest p_b
2. Based on content of visited pages
- Recommend a pages in the content-based cluster (using ranking of nearest neighbours)

Issues with Previous Approaches

- Most consider exclusively web usage data.

  There are other channels to exploit

- Transactions assume information needs are fulfilled sequentially.

  Not true in reality

- Newly added pages are never recommended.

  The new pages may contain the needed data

- Buried and difficult to reach pages are never recommended.

  Defeats the purpose of recommending

- Recommended lists are long and unordered.

  Carefully ranking recommendation is important

Contributions

- A *framework for a hybrid web recommender system*, which combines and makes full use of three different available channels to improve recommendation quality (Access history, Web page content and Page connectivity)
- Propose a novel notion for sub-sessions, *mission*, to capture users' concurrent information needs during on-line navigation (i.e. session).
- Propose a strategy to *take into account newly added or buried web pages* in the candidate list for recommendation.
- Introduce a new on-line navigational *mission-based model*.
- Use linkage information to *rank recommendation* candidates.
- Propose an *evaluation method* for web resource recommendation.

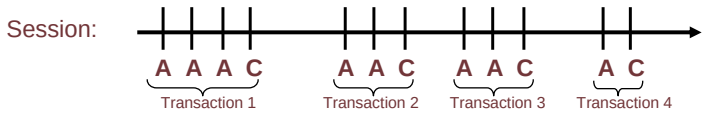
User, Session, and Transaction Identification

- Several pre-processing tasks have to be performed on access log data before applying web mining techniques for pattern discovery.
- Three core pre-processing tasks:
  - User Identification: identifying individual users
  - Visit Session Identification: group all pages clicked by individual users at a given consecutive time into individual visits
  - Transaction Identification: dividing each visit session into transactions, each of which fulfills one information need of the user.
- We use similar pre-processing techniques as in [Cooley et al., 1999] to identify individual users and sessions; propose an improved transaction identification approach.

Transaction Identification

- Two standard approaches
  - *Reference Length* Approach
  - *Maximal Forward Reference* Approach
- Same underlying assumption:

  A visitor may have different information needs during a visit, but all the information needs must be fulfilled **in the sequence**.



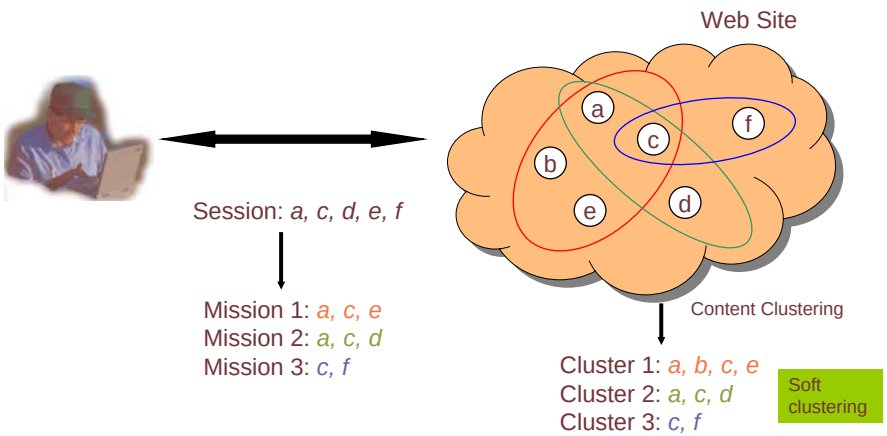
Mission vs. Transaction

- More often than not, we open several browsers to surf a site, looking for different information at the same time. Moreover, we may sometimes interrupt our current goal and start another in the middle, and then return to the original one later on.
- In these scenarios, the transaction identification approaches mistakenly group pages to fulfill users' different information needs into one transaction.
- Because the transaction is the base of any data mining algorithm for pattern discovery, this misclassification would obviously compromise the effect of the data mining task, or even cause it to fail.

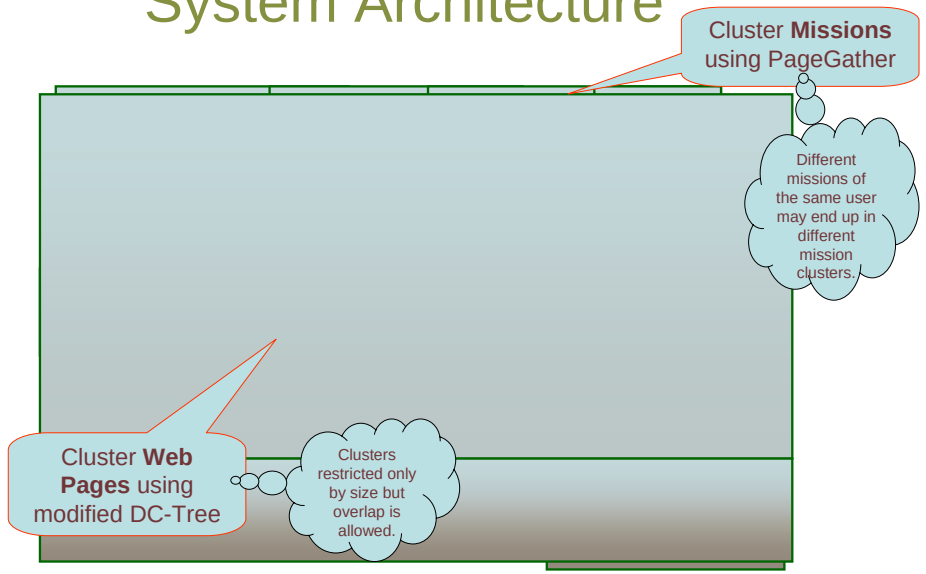
Mission Identification

- Mission Identification – an improved transaction identification approach.
  - Acknowledging that users may visit a website with multiple goals, i.e., different information needs.
  - Making no assumption on the sequence in which these needs are fulfilled.
- *Mission*: a sub-session related to one of these information needs
  - Allowing overlap between missions
  - Representing a concurrent search in the site

How to Identify Missions

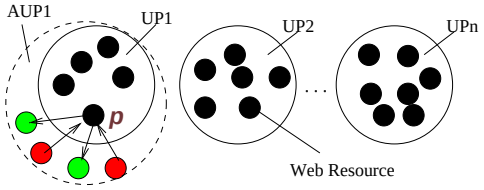


System Architecture



User Profile Improvement

- Providing an opportunity for these rarely visited or newly added pages to be recommended.
- User profile improvement is done in a two-step process.
  - Augmentation



  - Pruning

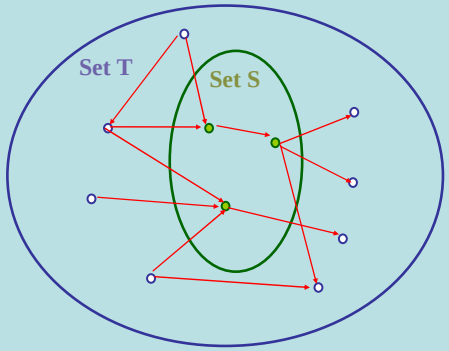
Ranking Recommendation Candidates with Connectivity

- How to present the candidate recommendation items in a suitable order is an important issue.
- We apply the ***HITS* algorithm** on the improved user profile to compute the ***Authority*** and ***Hub*** values of each page.
- The measures of *Hub* allow us to rank the pages within the cluster, and this ranking is used presenting recommendations.
- **Why Hubs and not Authorities?** Authorities would not favor newly added pages. Hubs would give the chance to good new pages to rank high.

Ranking Candidat

- How to present the
a suitable order is
- We apply the ***HITS***
profile to compute
page.
- The measures of *H*
the cluster, and th
recommendations
- **Why Hubs and no**
favor newly added
to good new page

Kleinberg's HITS algorithm (1998)



- Each page p has two values a(p) and h(p) initialized with 1.
- p → q denote "page p has an hyperlink to page q".
- Iteratively update a(p) and h(p)

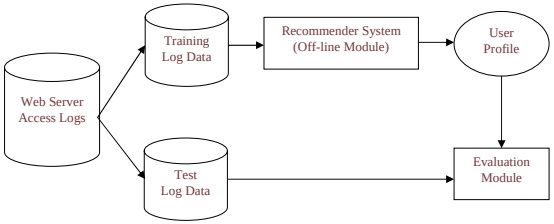
$$a(p) = \sum_{q \to p} h(q) \quad h(p) = \sum_{p \to q} a(q)$$

System Evaluation

- Evaluation Goals:
  - Evaluating the overall performance of our system, and justifying that our work can improve the recommendation quality and usefulness.
  - Justifying that each additional information channel added contributes to this improvement.

Evaluation Methodology

- Methodology: a simulation-based evaluation

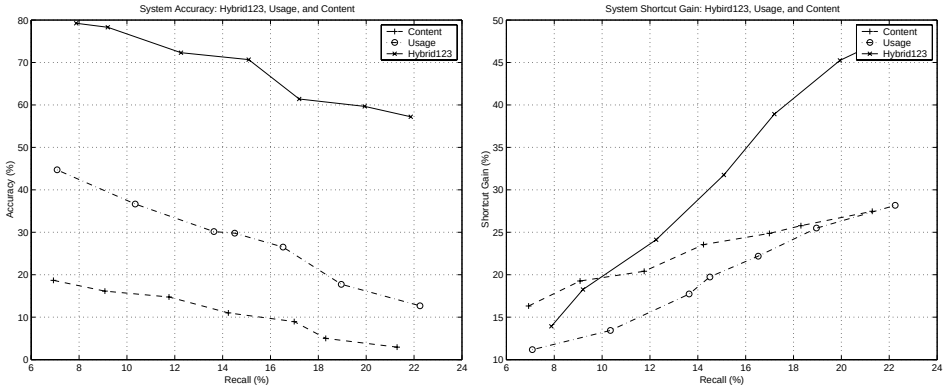


- Dataset: UofA CS Department Web Site

Evaluation Metrics

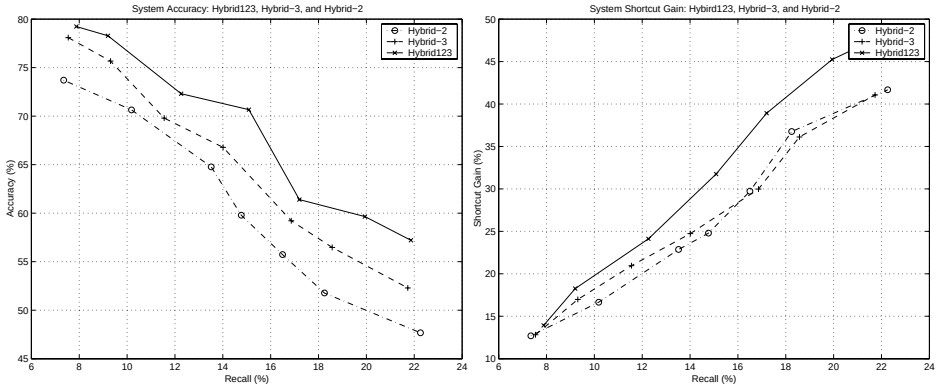
- ***Recommendation Accuracy***: How many recommendations given are correct
- ***Shortcut Gain***: how many clicks the recommendation allows users to jump.
- ***Recommendation Coverage***: Testing the consistency and trend of system performance.

Experiment (1)



A B

Experiment (2)

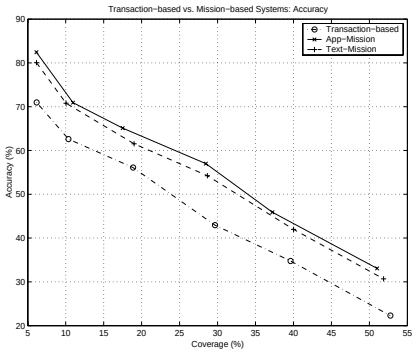


A B

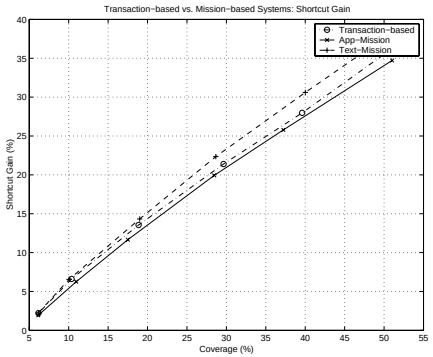
The Case of VIVIDESK

- The dataset generalizes our notion of *mission*, and extends its applicability to the context where content information is not available.
- The notion of mission is proposed to identify concurrent information needs during visit sessions.
- Generally, the mission can be discovered based on the content similarity among pages visited during that session.
- The concept is also suitable and applicable in a context without such information, or to sites containing pages that are not content-rich.
- Our work on VIVIDESK data also highlights the importance to have application related logs rather than just relying on information poor web server access logs.

Experiments on VIVIDESK Data



A



B

Conclusions

- Our experiments show that the combination of usage, content, and structure data in a web recommender system has the potential to improve the quality of the system, as well as to keep the recommendation up-to-date.
- There are various ways to combine these different channels. Our future work will include investigating different methods of combination.
- We introduced the notion of ***mission*** and a strategy to take into account newly added or rarely visited pages,

Questions