Web Usage Mining & Personalization in Noisy, Dynamic, and Ambiguous Environments

Olfa Nasraoui
Knowledge Discovery & Web Mining Lab
Dept of Computer Engineering & Computer Sciences
University of Louisville

E-mail: olfa.nasraoui@louisville.edu
URL: http://www.louisville.edu/~o0nasr01

Supported by *US National Science Foundation Career Award IIS-0133948*

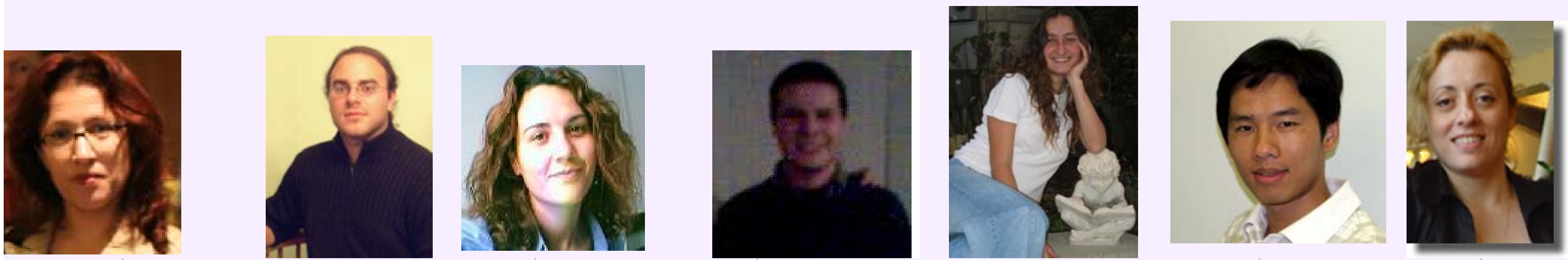
Compressed Vita

- Endowed Chair of E-commerce in the Department of Computer Engineering & Computer Science at the University of Louisville
- Director of the Knowledge Discovery and Web Mining Lab at the University of Louisville.
- Research activities include **Data Mining, Web mining, Web Personalization, and Computational Intelligence (Applications of evolutionary computation and fuzzy set theory)**.
- Served as program co-chair for several conferences & workshops, including WebKDD 2004, 2005, and 2006 workshops on Web Mining and Web Usage Analysis, held in conjunction with ACM SIGKDD International Conferences on Knowledge Discovery and Data Mining (KDD).
- Recipient of US National Science Foundation CAREER Award.

- What I will speak about today is mainly the research products and lessons from a 5-year US National Science Foundation project

My Collaborative Network?

Team: Knowledge Discovery & Web Mining Lab
University of Louisville



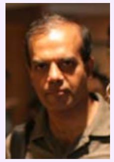
Director: Olfa Nasraoui (speaker)

Current Student Researchers (<u>alphabetically listed</u>):

Jeff Cerwinske, Nurcan Durak, Carlos Rojas, Esin Saka, Zhiyong Zhang, Leyla Zhuhadar

Note: Gender balanced & multicultural ;-)

Past and Present Collaborators



Raghu Krishnapuram, IBM Research

Anupam Joshi, University of Maryland, Baltimore County

Hichem Frigui, University of Louisville

Hyoil Han, Drexel University

Antonio Badia, University of Louisville

Roberta Johnson, University Corporation for Atmospheric Research (UCAR)

Fabio Gonzalez, Nacional University of Colombia

Cesar Cardona, Magnify, Inc.

Elizabeth Leon, Nacional University of Colombia

Jonatan Gomez, Nacional University of Colombia

Introduction

- Information overload: too much information to sift/browse through in order to find desired information
  - Most information on Web is actually irrelevant to a particular user
- This is what motivated interest in techniques for Web personalization
- As they surf a website, users leave a wealth of historic data about what pages they have viewed, choices they have made, etc
- Web Usage Mining: A branch of Web Mining (itself a branch of data mining) that aims to discover interesting patterns from Web usage data (typically Web Log data/clickstreams) (Yan et al. 1996, Cooley et al. 1997, Shahabi, 1997; Zaiane et al. 1998, Spiliopoulou & Faulstich, 1999, Nasraoui et al. 1999, Borges & Levene, 1999, Srivastava et al. 2000, Mobasher et al. 2000; Eirinaki & Vazirgiannis, 2003)

Introduction

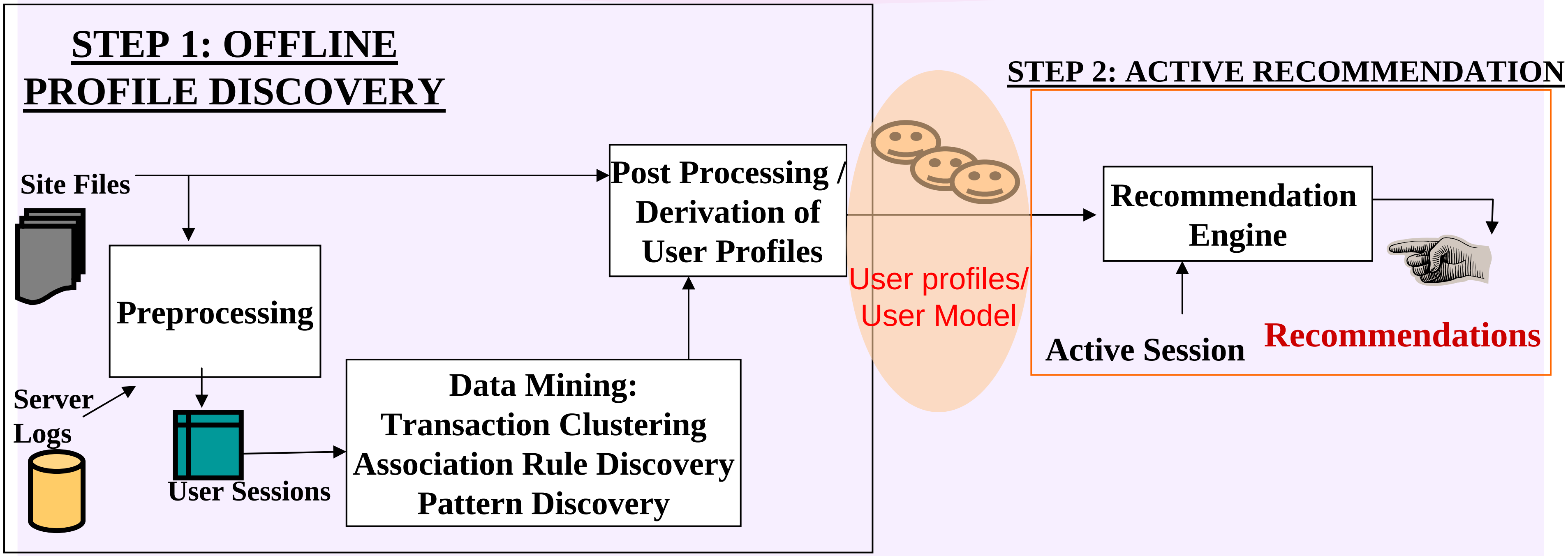
- Web Personalization: Aims to adapt the Website according to the user's activity or interests (Perkowitz & Etzioni, 1997, Breeze et al. 1998, Pazzani, 1999, Schafer et al. 1999, Mulvenna, 2000; Mobasher et al. 2001, Burke. 2002, Joachims, 2002; Adomavicius &. Tuzhilin, 2005)
- Intelligent Web Personalization: often relies on Web Usage Mining (for user modeling)
- Recommender Systems: recommend items of interest to the users depending on their interest (Adomavicius & Tuzhilin, 2005)
  - **Content-based filtering:** recommend items similar to the items liked by current user (Balabanovic & Shoham, 1997)
    - No notion of community of users (specialize only to one user)
  - **Collaborative filtering:** recommend items liked by "similar" users (Konstan et al., 1997; Sarwar et al., 1998; Schafer, 1999)
    - Combine history of a community of users: **explicit** (ratings) or **implicit (clickstreams)**
  - **Hybrids:** combine above (and others)

Focus of our research

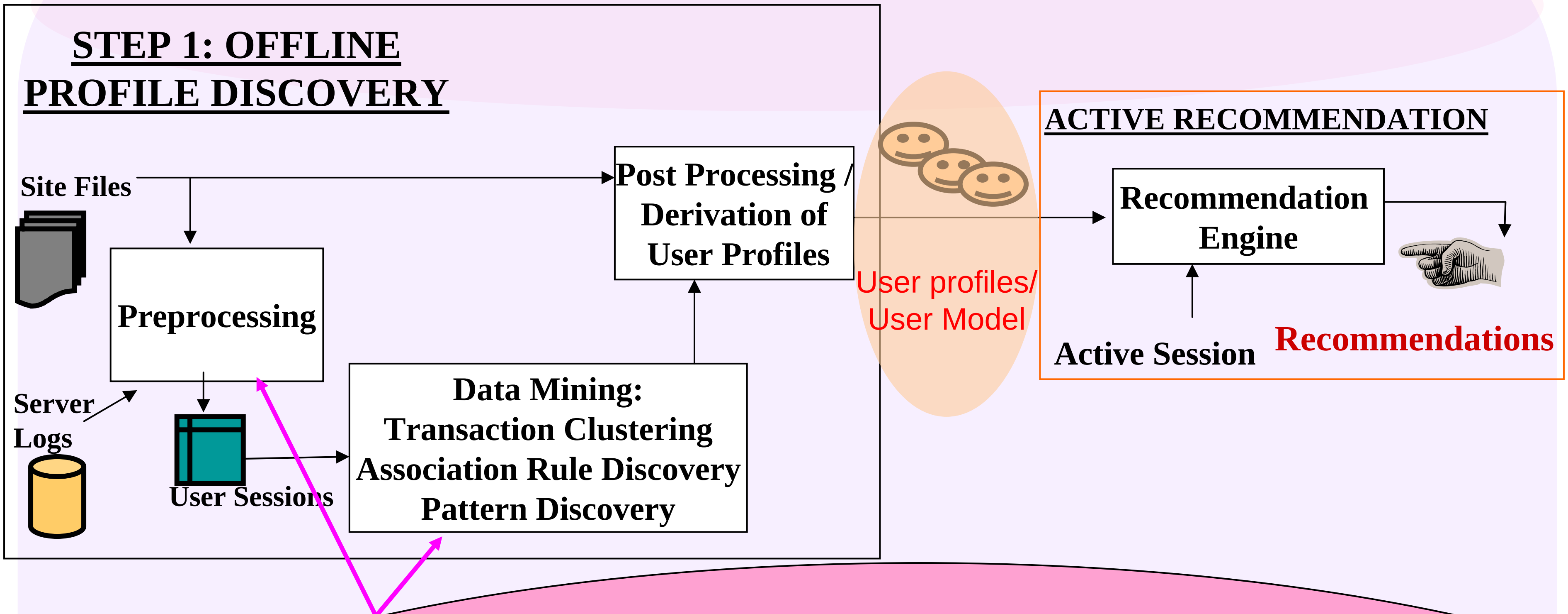
Some Challenges in WUM and Personalization

- **Ambiguity**: the level at which clicks are analyzed (*URL* A, B, or C as basic identifier) is very shallow, almost no meaning
  - Dynamic URLs: meaningless URLs ➔ even more ambiguity
  - Semantic Web Usage Mining: (Oberle et al., 2003)
- **Scalability**: Massive Web Log data that cannot fit in main memory requires techniques that are scalable (stream data mining) (Nasraoui et al.: WebKDD 2003, ICDM 2003)
- **Handling Evolution**: Usage data that changes with time
  - Mining & Validation in dynamic environments: largely unexplored area…except in: (Mitchell et al. 1994; Widmer, 1996; Maloof & Michalski, 2000)
  - In the Web usage domain: (Desikan & Srivastava, 2004; Nasraoui et al.: WebKDD 2003, ICDM 2003, KDD 2005, Computer Networks 2006, CIKM 2006)
- **From Clicks to Concepts**: few efforts exist based on laborious manual construction of concepts, website ontology or taxonomy
  - How to do this automatically? (Berendt et al., 2002; Oberle et al., 2003; Dai & Mobasher, 2002; Eirinaki et al., 2003)
- **Implementing recommender systems** can be slow, costly and a bottle neck especially
  - for researchers who need to perform tests on a variety of websites
  - For website owners that cannot afford expensive or complicated solutions

Different Steps Of our Web Personalization System



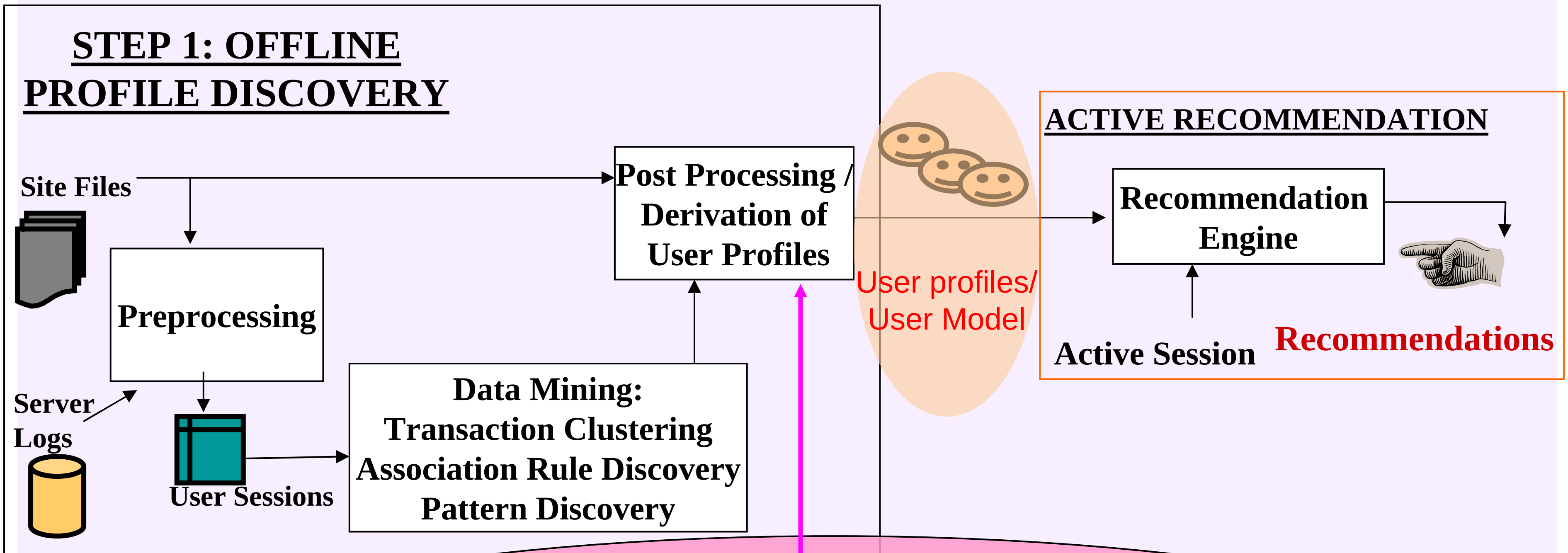
Challenges & Questions in Web Usage Mining



Dealing with Ambiguity: Semantics?

- Implicit taxonomy? (Nasraoui, Krishnapuram, Joshi. 1999)
  - Website hierarchy (can help disambiguation, but limited)
- Explicit taxonomy? (Nasraoui, Soliman, Badia, 2005)
  - From DB associated w/ dynamic URLs
  - Content taxonomy or ontology (can help disambiguation, powerful)
- Concept hierarchy generalization / URL compression / concept abstraction: (Saka & Nasraoui, 2006)
  - How does abstraction affect quality of user models?

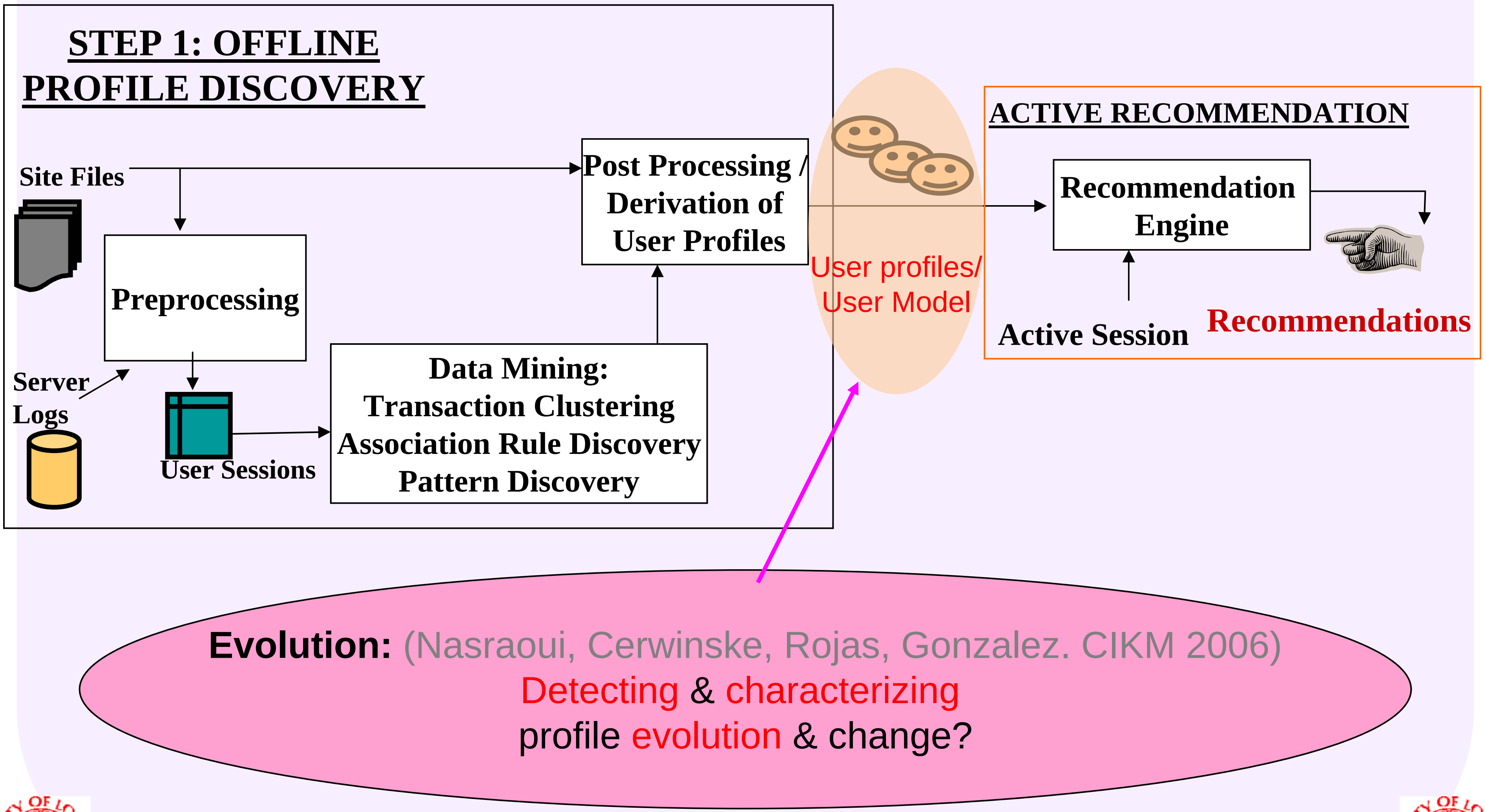
Challenges & Questions in Web Usage Mining



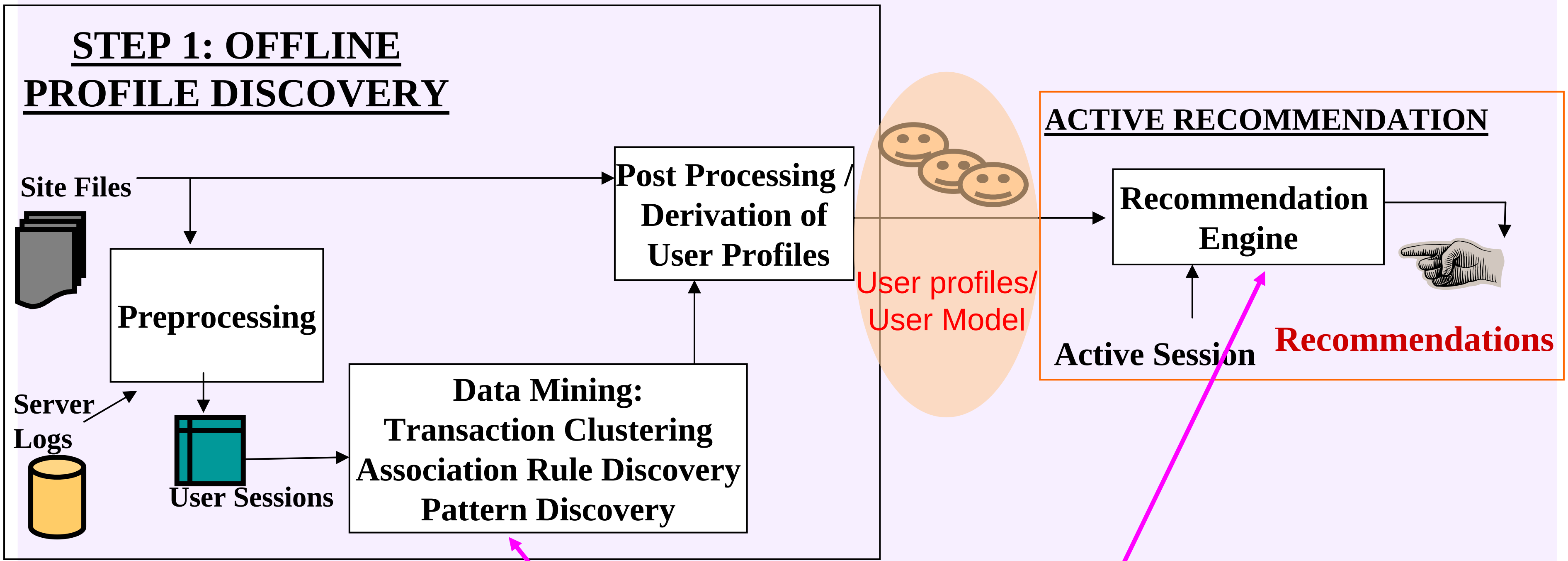
User Profile Post-processing Criteria? (Saka & Nasraoui, 2006)

- Aggregated profiles (frequency average)?
- Robust profiles (discount noise data)?
- How do they really perform?
- How to validate? (Nasraoui & Goswami, SDM 2006)

Challenges & Questions in Web Usage Mining



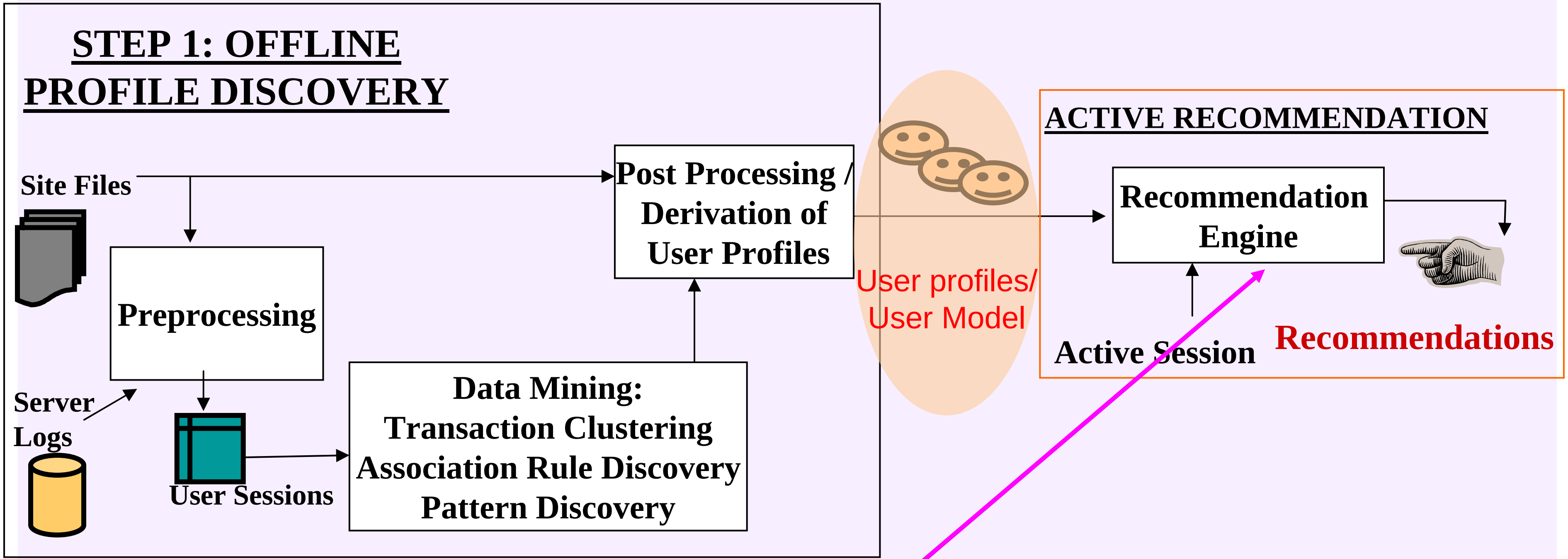
Challenges & Questions in Web Personalization



In case of massive evolving data streams:

- Need stream data mining (Nasraoui et al. ICDM'03, WebKDD 2003)
- Need stream-based recommender systems? (Nasraoui et al. CIKM 2006)
- How do stream-based recommender systems perform under evolution?
- **How to validate** above? (Nasraoui et al. CIKM 2006)

Challenges & Questions in Web Personalization



Implementing Recommender Systems:

- Fast, easy, scalable, cheap, free?
- At least to help support research…
- But Grand advantage: help the little guy…

(Nasraoui, Zhang, Saka, SIGIR-OSIR 2006)

Web Usage Mining:

- **Ambiguity**:
  - Implicit Semantics: website hierarchy
  - Explicit Semantics: DB w/ taxonomy of dynamic URLs
  - What is effect of generalization / URL compression / concept abstraction
- Noise:
- Detecting and characterizing evolution in dynamic environments
- Recommender Systems in dynamic environments
- Fast, Easy, Free Implementation
- Mining Conceptual Web Clickstreams

What's in a click?

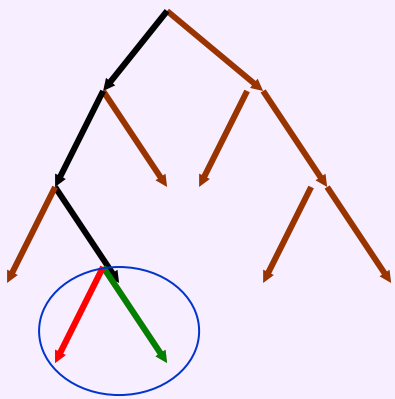
- **Access log:** Record of URLs accessed on Website
- Log entry: access date, time, IP address, URL viewed, …etc.
- Modeling User Sessions: set of clicks, pages, URLs (Cooley et al. 1997)
  - Map URLs on site to indices
  - User session vector *s(i)*: temporally compact sequence of Web accesses by a user (consecutive requests within time threshold: e.g. 45 minutes)
- URLs:
  - Orthogonal? (Traditional approach)
  - Exploit some **implicit concept hierarchy**: website hierarchy (easy to infer from URLs) (Nasraoui, Krishnapuram, Joshi. 1999)
  - **Dynamic URLs**: Exploit some **explicit concept hierarchy**: encoded in Web item database (Nasraoui, Soliman, Badia, 2005)
- How to take above into account?
  - **Integrate into the similarity measure** while clustering

Similarity Measure

(Nasraoui, Krishnapuram, Joshi. 1999)

- Map N_U URLs on site to indices
- User session vector $s^{(i)}$: temporally compact sequence of Web accesses by a user

$$s_j^{(i)} = \begin{cases} 1 & if \quad user \quad accessed \quad j^{th}\, URL \\ 0 & otherwise \end{cases}$$



- If site structure ignored ➔ cosine similarity

$$S_{1,kl} = \frac{\sum_{i=1}^{N_U} s_i^{(k)} s_i^{(l)}}{\sqrt{\sum_{i=1}^{N_U} s_i^{(k)} \sum_{i=1}^{N_U} s_i^{(l)}}}$$

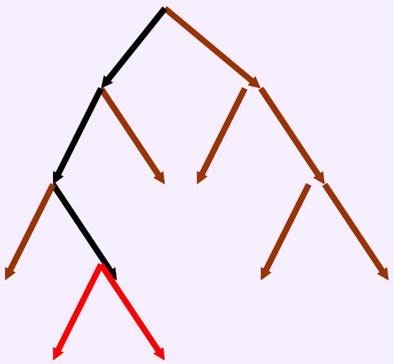
- Taking site structure into account ➔ relate distinct URLs:
  - p_i: path from root to i^{th} URL's node

$$S_u(i,j) = \min\left(1, \frac{|p_i \cap p_j|}{\max\left(1, \max\left(|p_i|, |p_j|\right) - 1\right)}\right)$$

O. Nasraoui and R. Krishnapuram, and A. Joshi. Mining Web Access Logs Using a Relational Clustering Algorithm Based on a Robust Estimator, 8th International World Wide Web Conference, Toronto, pp. 40-41, 1999.

Web Session Similarity Measure: variant of cosine that takes into account item relatedness

Taking site structure into account:



$$S_{1,kl} = \frac{\sum_{i=1}^{N_U} s_i^{(k)} s_i^{(l)}}{\sqrt{\sum_{i=1}^{N_U} s_i^{(k)} \sum_{i=1}^{N_U} s_i^{(l)}}}$$

$$S_{2,kl} = \frac{\sum_{i=1}^{N_U} \sum_{j=1}^{N_U} s_i^{(k)} s_j^{(l)} S_u(i,j)}{\sum_{i=1}^{N_U} s_i^{(k)} \sum_{i=1}^{N_U} s_i^{(l)}}$$

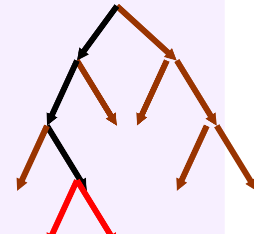
- Final Web Session Similarity = $S_{kl} = \max\left(S_{1,kl}, S_{2,kl}\right)$
- Concept Hierarchies: helpful in many data mining contexts: (E.g. in association rule mining: Srikant &. Agrawal, 1995, in text: Chakrabarti et al., 1997, in Web usage mining: Berendt, 2001, Eirinaki, 2003)

Web Usage Mining:

- **Ambiguity:**
  - Implicit Semantics: website hierarchy
  - Explicit Semantics: DB w/ taxonomy of dynamic URLs
  - What is effect of generalization / URL compression / concept abstraction
- **Noise:**
- **Detecting and characterizing evolution in dynamic environments**
- **Recommender Systems in dynamic environments**
- **Implementation**
- **Mining Conceptual Web Clickstreams**

Role of Similarity Measure: Adding semantics

- **Problem:** Dynamic URLs, such as universal.aspx?id=56:
  - hard to recognize based only on their URL ➔ affects presentation & interpretation of discovered user profiles!
  - hard to relate (among each other) based only on their URL ➔ affects Web usage mining!
- **Solution:** Use available external data that maps dynamic URLs to hierarchically related and more meaningful descriptions
  - ➔ ***Explicit*** **taxonomy: parent item → child item**
  - ➔ transform URL into **regular looking URL**: parent/child/grand-child…etc
  - ➔ handle this URL using previous *implicit* website hierarchy approach: inferred by tokenizing the URL string



- Ultimately, both implicit and explicit taxonomy information are seamlessly incorporated into the data mining algorithm (clustering) via the Web session similarity measure

Olfa Nasraoui, Maha Soliman, and Antonio Badia, Mining Evolving User Profiles and More – A Real Life Case Study, In Proc. Data Mining meets Marketing workshop, New York, NY, 2005.

Mapping Dynamic URLs to Semantic URLs

(Nasraoui, Soliman, Badia, 2005)

- **Problem:** Dynamic URLs, such as **universal.aspx?id=56**, are
  - hard to recognize based only on their URL ➔ affects <u>presentation</u> of profiles!
  - hard to relate (among each other) based only on their URL ➔ affects Web usage <u>mining</u>!.
- **Solution:** We resorted to available external data, provided by the website designers, that maps dynamic URLs to hierarchically related and more meaningful descriptions.

Taxonomy Data Provided by the website designers

menus_id	item_name	item_level	parent_item	sequence	Resource/url
56	Regulations and Laws	1	4939	1	universal.aspx
4939	NST Center®	0		1	Nst

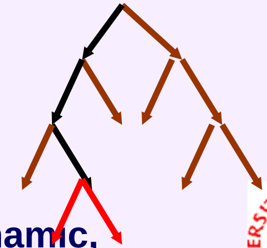
Example: Dynamic URL: **universal.aspx?id=56** ➔
Semantic URL: ***NST Center&reg / Regulations and Laws***

Mapping Dynamic URLs to Semantic URLs (another example)

- universal.aspx?id=**6770** ➔ ?
- *since item #6770 has as parent:* item #56

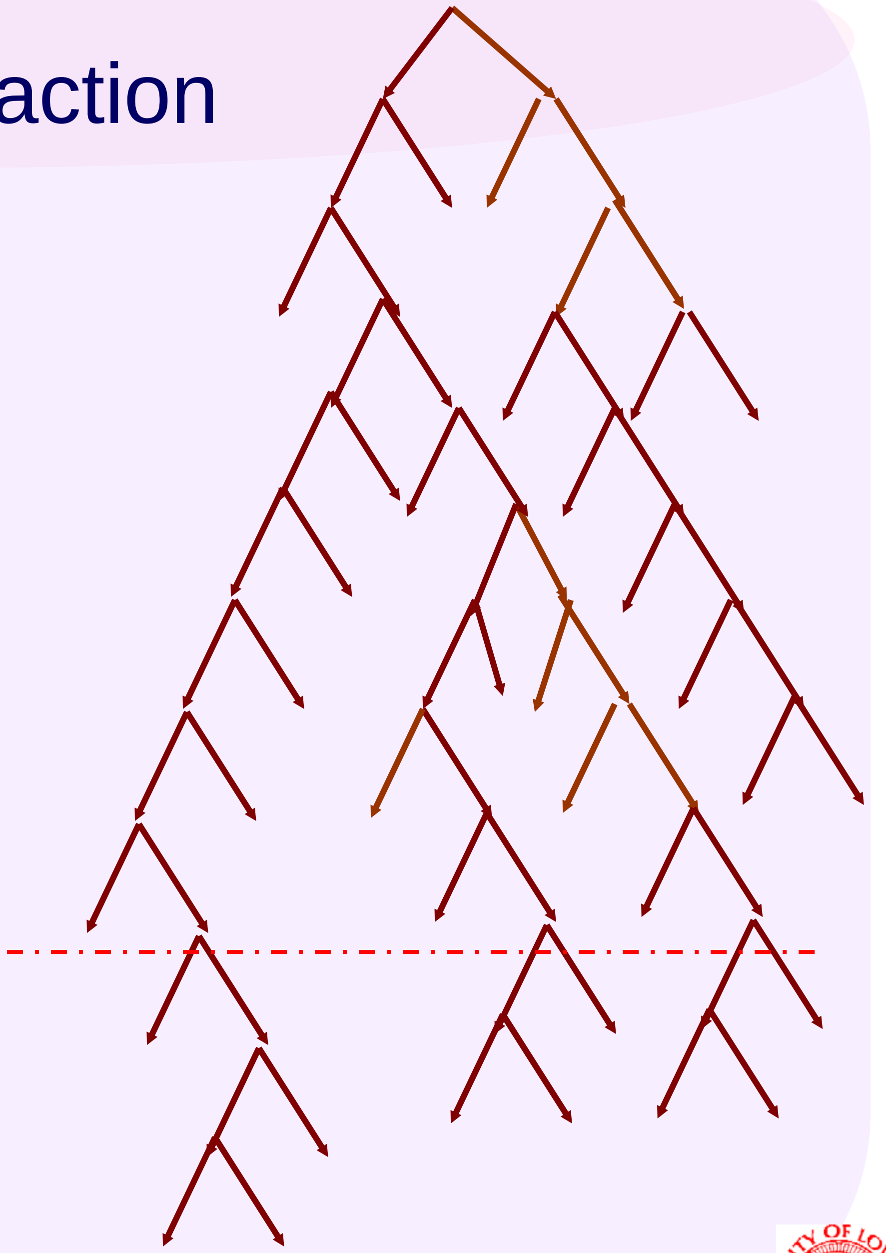
menus_id	item_name	item_level	parent_item	sequence	url
6770	Air Quality and Emission Standards	2	**56**	1	universal.aspx

- Recall: Item 56 = (*NST Center&reg / Regulations and Laws*)
- Hence, universal.aspx?id=**6770** ➔
  - *NST Center&reg / Regulations and Laws / Air Quality and Emission Standards*



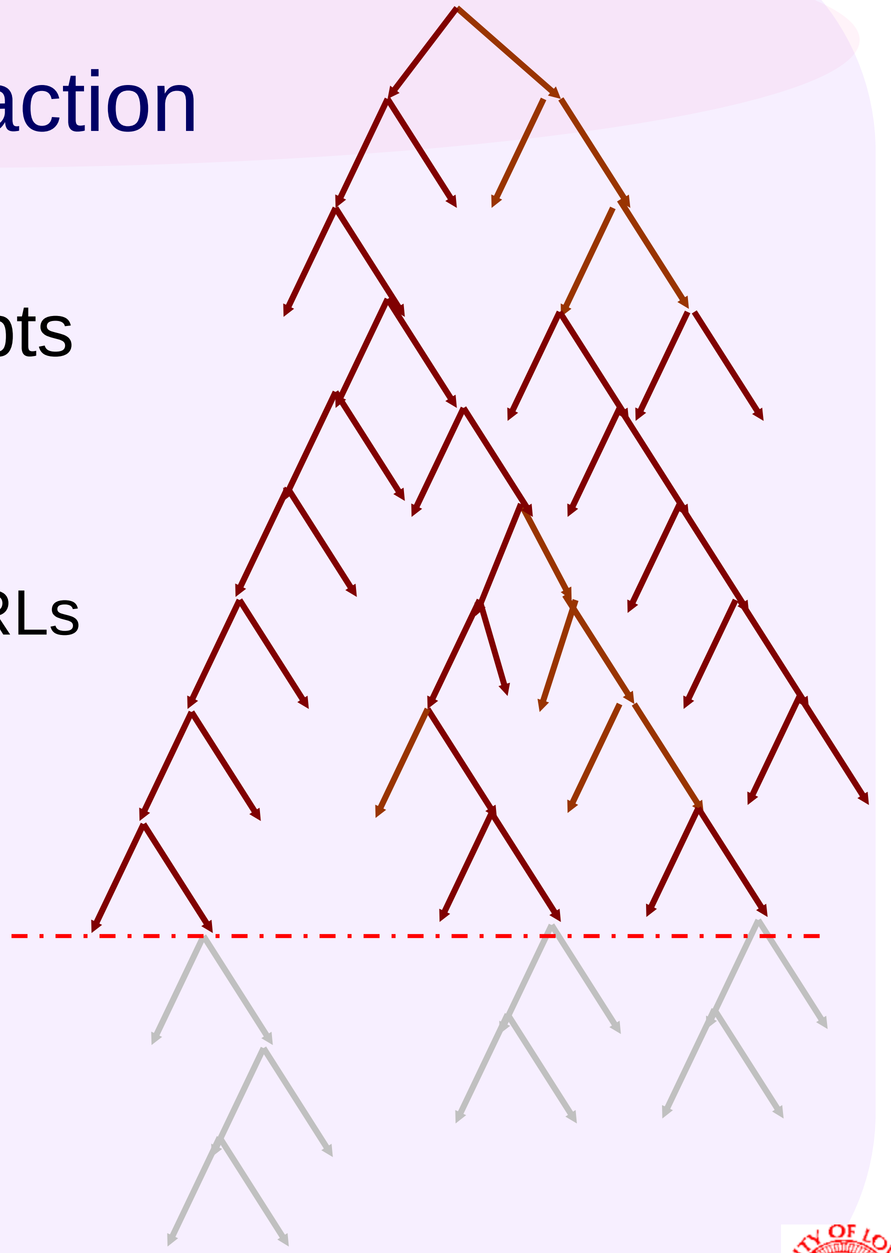
Concept Generalization/Abstraction

- Generalize lower/specific concepts to higher concepts
- Mechanism:
  - IF Sim (URL_i, URL_j) > Threshold THEN merge URLs



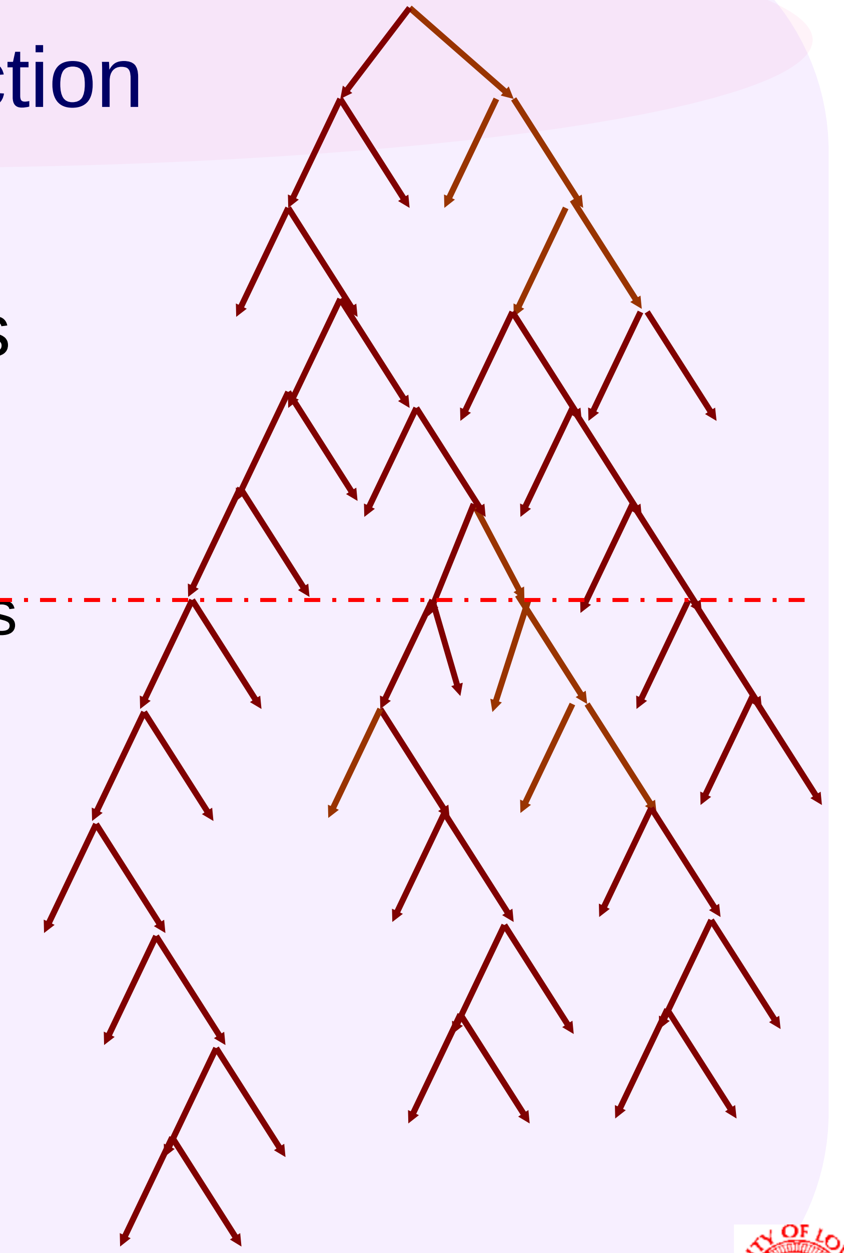
Concept Generalization/Abstraction

- Generalize lower/specific concepts to higher concepts
- Mechanism:
  - IF Sim (URL_i, URL_j) > Threshold THEN merge URLs
- Effects:
  - Helps in disambiguation
  - URL compression
    - Easily reach compression rates in 80% range depending on merging threshold



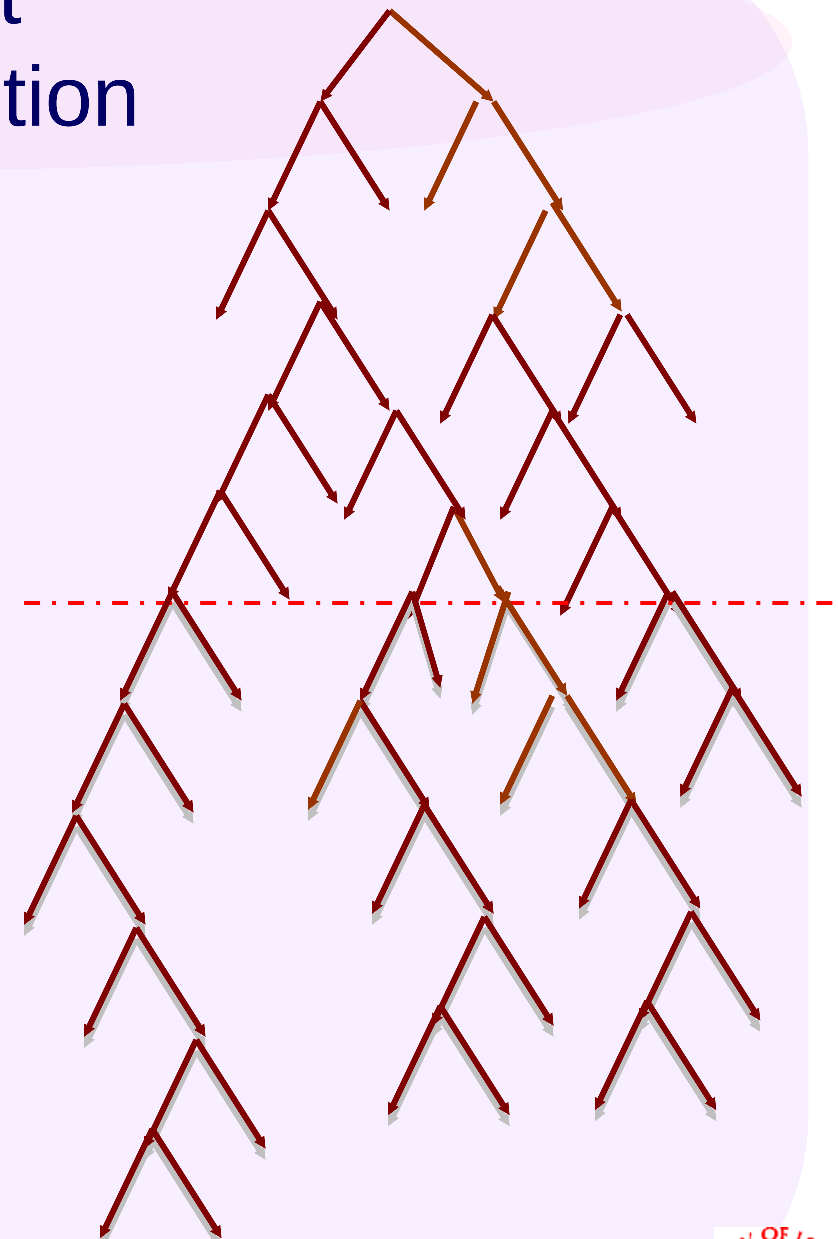
Concept Generalization/Abstraction

- Generalize lower/specific concepts to higher concepts
- Mechanism:
  – IF Sim (URL_i, URL_j) > Threshold THEN merge URLs
  – Effects:
  – Helps in disambiguation
  – URL compression
    - Easily reach compression rates in 90% range depending on merging threshold



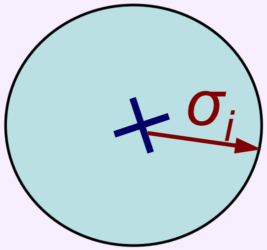
Aggressive Concept Generalization/Abstraction

- Generalize even more lower/specific concepts to higher concepts
- Mechanism:
  - IF Sim (URL_i, URL_j) > Even-bigger-Threshold THEN merge URLs
- More drastic effects:
  - Helps in disambiguation
  - URL compression
    - Easily reach compression rates in 90% range depending on merging threshold



Effect of Compression

- First, the mining + validation methodology:
- Perform Web Usage Mining:
  - Pre-process Web log data (includes URL transformations taking into account implicit or explicit concept hierarchy)
  - Cluster user sessions into optimal number of user profiles using HUNC (Hierarchical Unsupervised Niche Clustering)
    - Localized Error-Tolerant profiles
    - maximize a measure of soft transaction support
    - with dynamically optimized error-tolerance σ

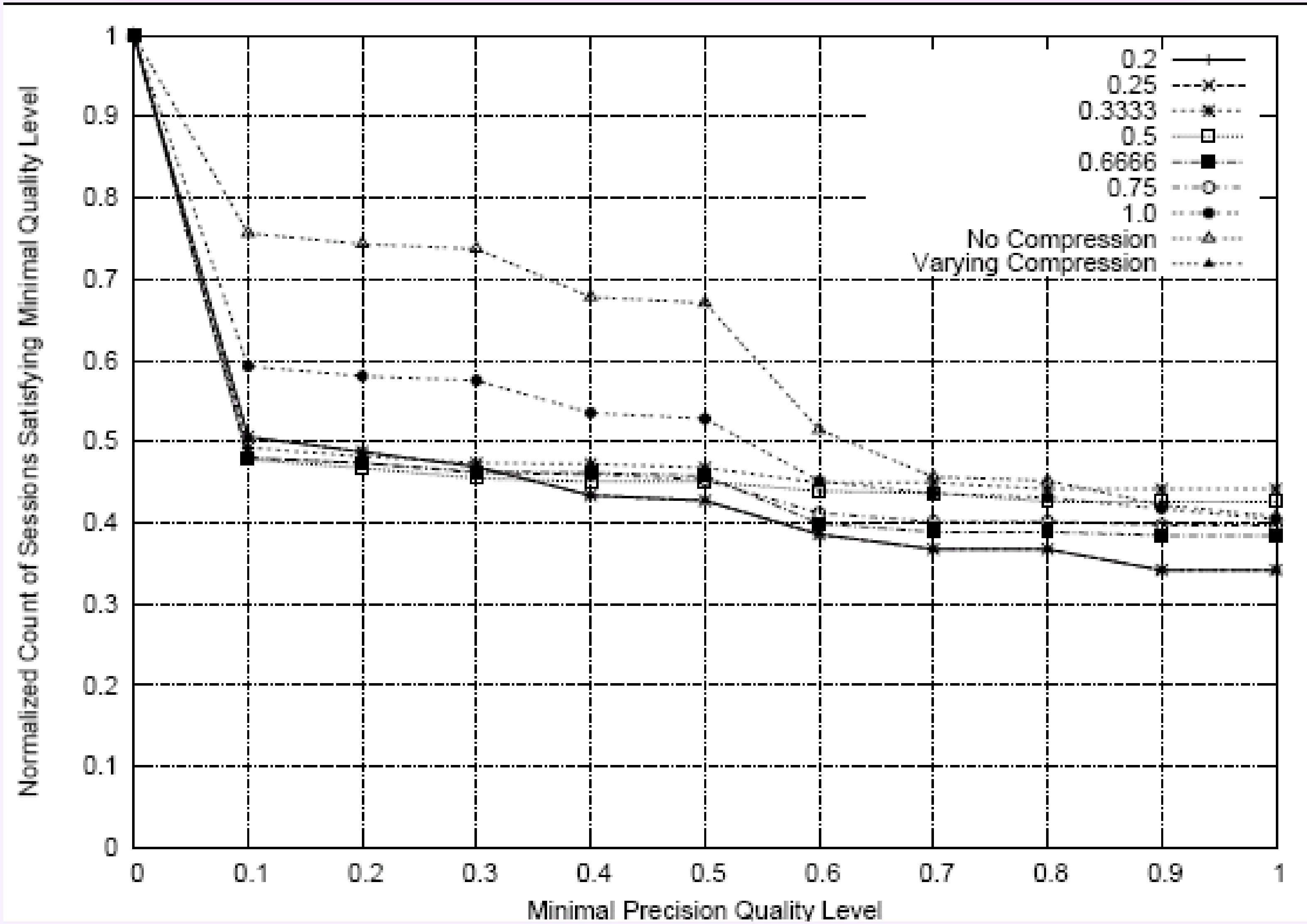


  - Optional Post-processing: (Later…)
    - Frequency Averaging: compute frequency of each URL in each cluster ➔ **profile**
    - Robust Profiles: ignore noisy user sessions when computing the above
- Validate discovered profiles against Web sessions

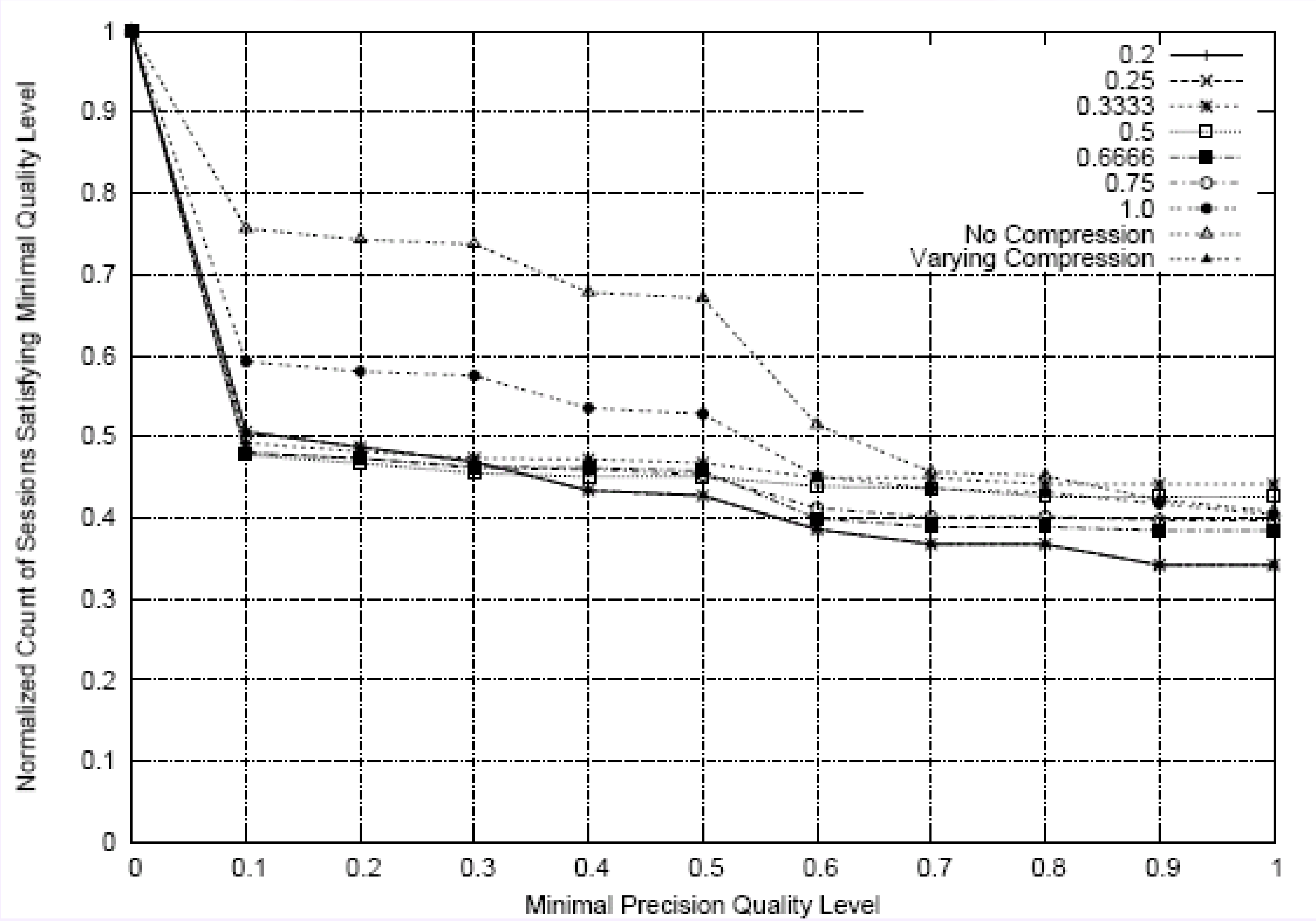
Validation in an Information Retrieval Context (Nasraoui & Goswami, 2005)

- Profiles are patterns that summarize the input transaction data
- Quality of discovered profiles as a summary of the input transactions:
  - Precision (the profile's items are all correct or included in an original input transaction/session, i.e. no extra items)
  - Coverage/recall (a profile's items are complete compared to an transaction or session, i.e. no missed items)
- Interestingness measure: Given $T_i^Q = \{t_j \mid Q_{ij} \geq Q_{\min}\}$, define $Q_i = |T_i^Q| / |T|$
- When $Q_{ij} = Cov_{ij}$, we call Q the *Cumulative Coverage of Transactions*, and it answers the Question
  - Is the data set *completely* summarized/represented by the mined profiles? .
- When $Q_{ij} = Prec_{ij}$, we call Q the *Cumulative Precision of Transactions*, and it answers the Question:
  - Is the data set *faithfully/accurately* summarized/represented by the mined profiles?
- These measures quantify the quality of mined profiles from the point of view of providing an accurate summary of the input data.
- ***Note:*** **Q_i = Probability {Precision ≥ Q_{min}} or Probability {Coverage ≥ Q_{min}}**

Precision Quality



Coverage Quality



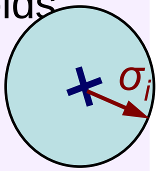
Observations

- Compression decreases Quality (as expected …)
- However, level of compression (or abstraction) is not an important factor
  - What seemed to matter most is "whether" any compression is made or not?
- Compression ➔ distortion of original data (hence reduced quality)
- But let's not forget…
- Compression ➔ reduced sparsity of the session matrix (hence may help clustering results)
- Compression ➔ drastic reduction in # items (hence speed up the mining…)

Handling Noise: Effect of Robustifying the Profiles

(Nasraoui & Krishnapuram, SDM 2002)

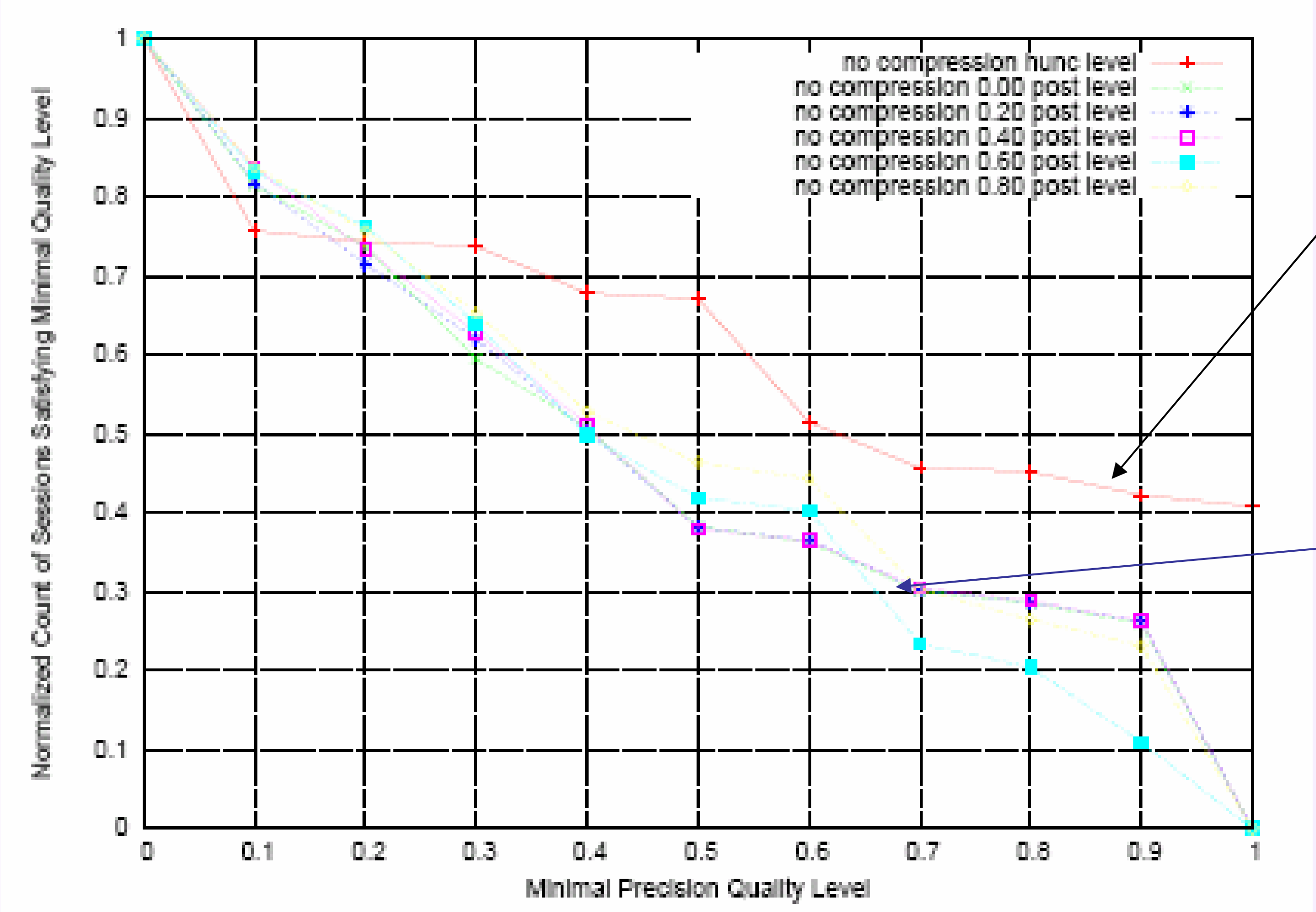
- Perform Web Usage Mining:
  - Pre-process Web log data (includes URL transformations taking into account implicit or explicit concept hierarchy)
  - Cluster user sessions into optimal number of user profiles using HUNC (Hierarchical Unsupervised Niche Clustering)
    - Localized Error-Tolerant profiles
    - maximize a measure of soft transaction support
    - with dynamically optimized error-tolerance σ
  - **Post-process profiles:**
    - Simple Means: Compute (URL-frequency) means/centroids for each cluster



    - Robust Means:
      - Robust weight of a session into a profile (varies between 0 and 1):
        - » $w_{ij} = e(-(1-Sim_{ij})^2 / \sigma_i)$
      - user sessions with $w_{ij} < w_{min}$ are ignored when averaging the URL frequencies in their cluster

- Validate discovered profiles against Web sessions

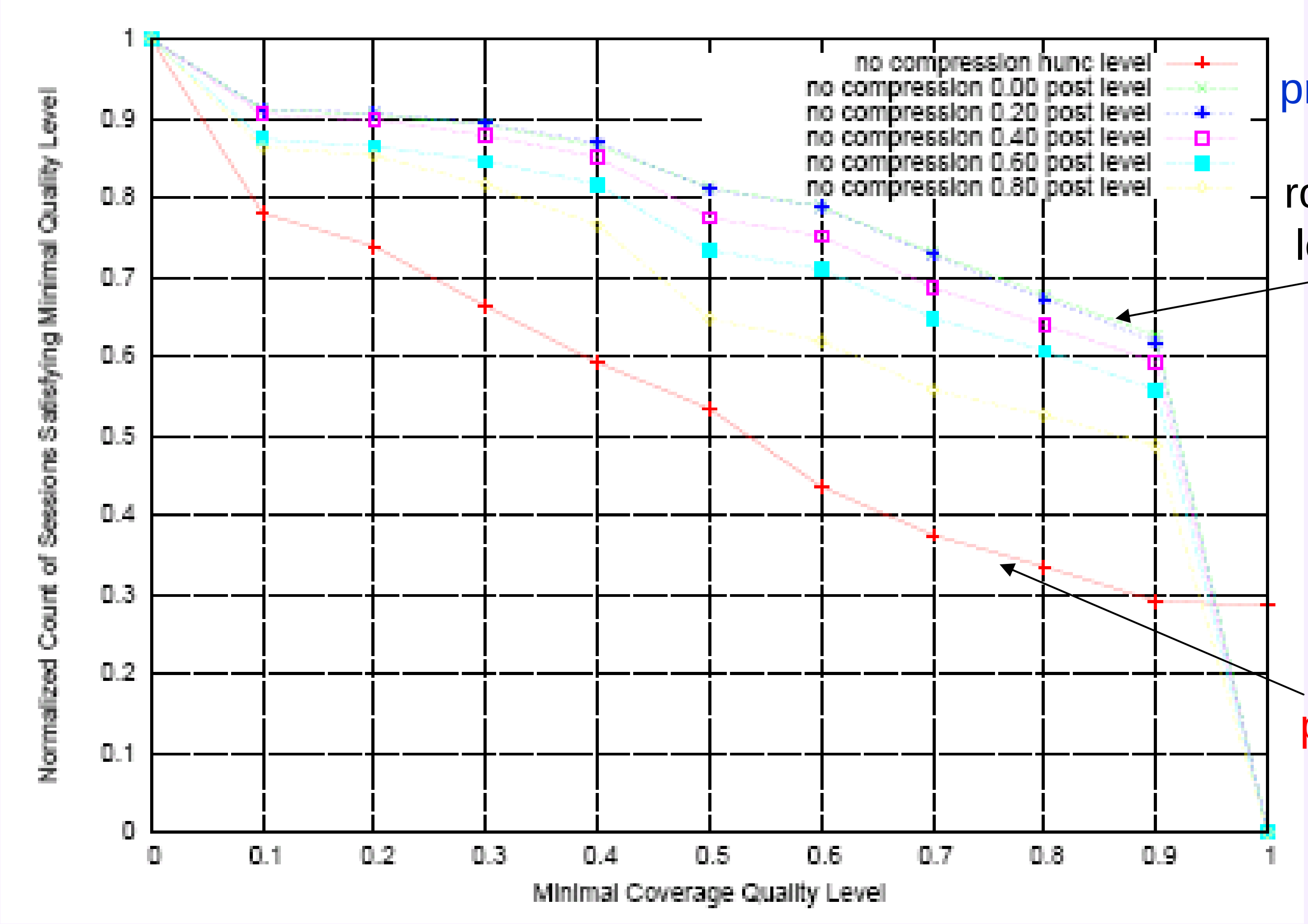
Precision Quality for various robustness levels w_{min}



No post-processing (raw profiles)

Post-processing: various robustness levels

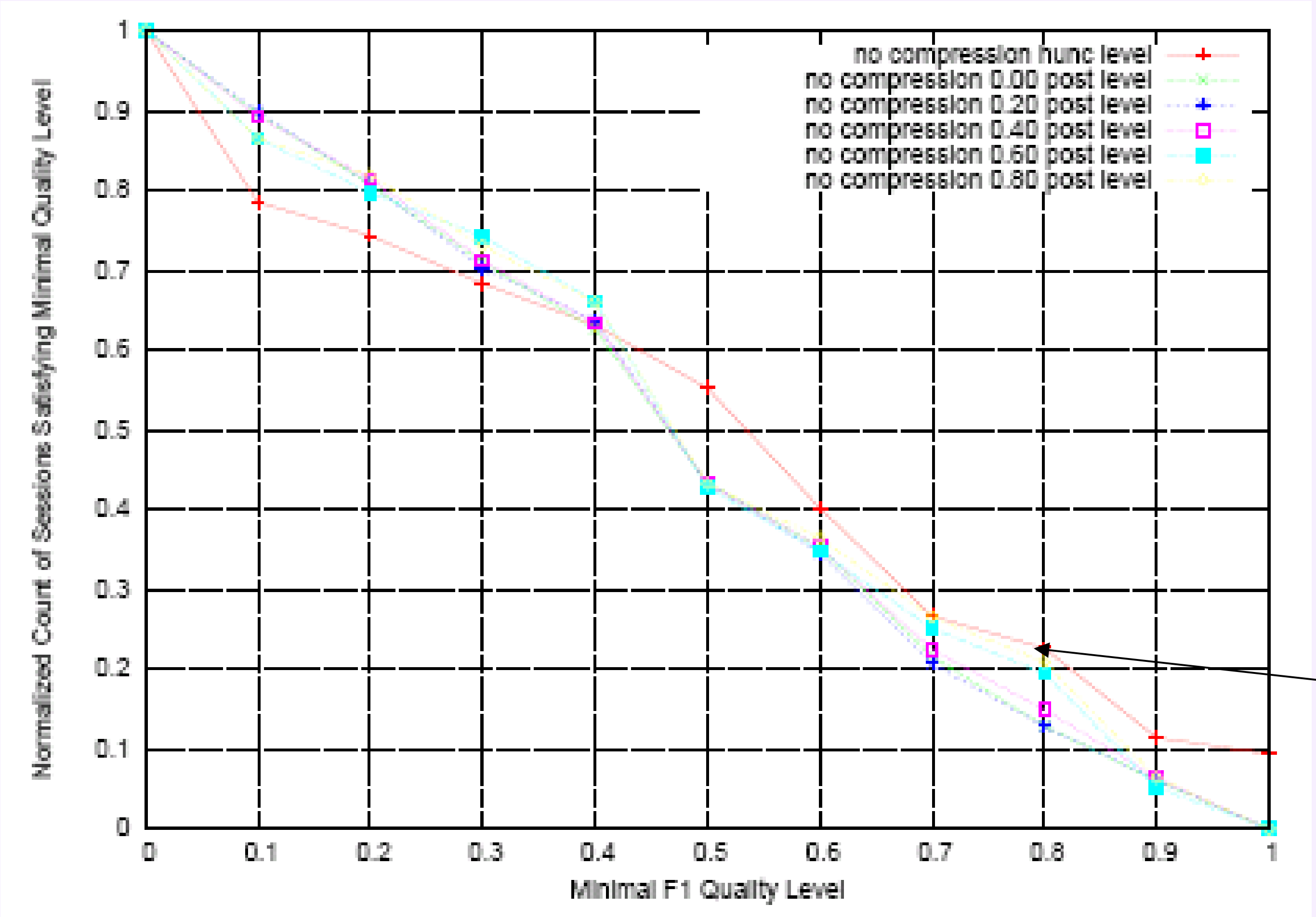
Coverage Quality for various robustness levels w_{min}



Post-processing: Optimal robustness level (0.2)

No post-processing (raw profiles)

F1 Quality for various robustness levels w_{min}



No post-processing (raw profiles)

Observations

- Post-processing decreases Precision
- However, it improves coverage
- Computing the URL frequency means of all sessions in each profile/cluster brings up to the surface some URLs that did not make it through the optimization process resulting in the raw profiles
  - More URLs improve coverage, however, hurt precision

Tracking Evolving Profiles

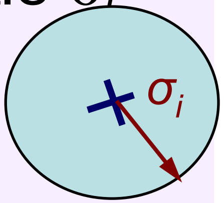
(Nasraoui, Soliman, Badia, 2005)

Mine user sessions in several batches (for each period)

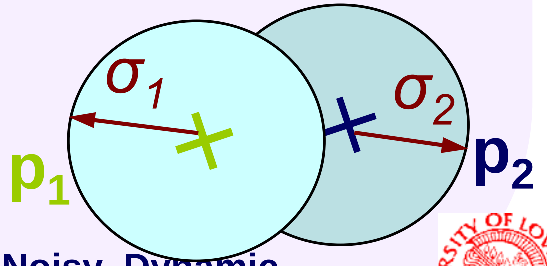
Automated comparison between new profiles and all the old profiles discovered in previous batches.

Each profile $\boldsymbol{p}_i$ is discovered along with an automatically determined measure of scale σ_i

➔ *boundary around each profile*



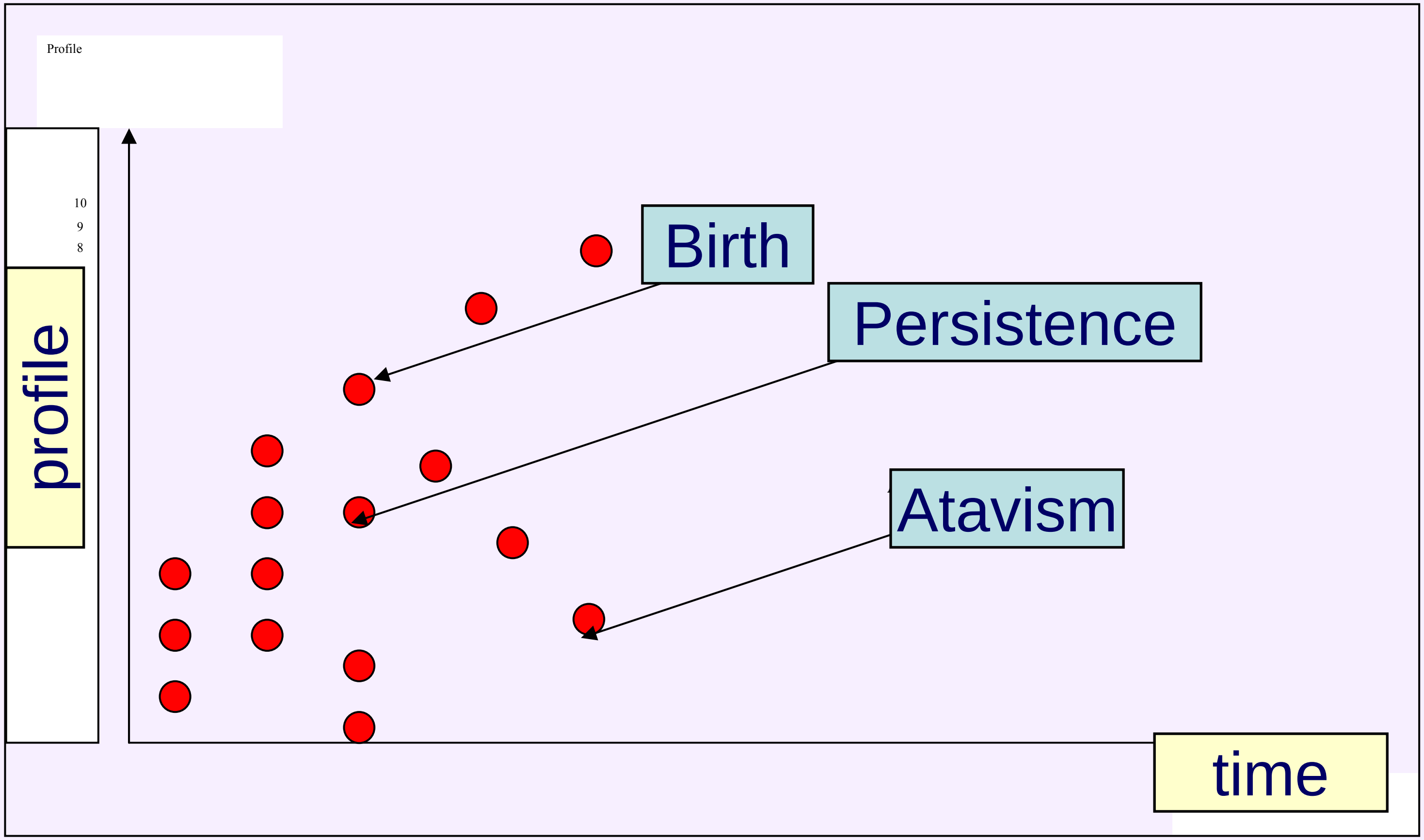
This allows us to automatically determine whether two profiles are compatible based on their distance compared to their respective boundaries



Tracking Evolving Access Patterns

- Four events can be detected from the comparison:
  - **Persistence:** New profiles are compatible with the old profiles.
  - **Birth:** New profiles are incompatible with any previous profile.
  - **Death:** Old profile finds no compatible profile from the new batch.
  - **Atavism:** Old profile that disappears, and then reappears (i.e. via compatibility) again in a later batch

Profile Events



Tracking Evolving Access Patterns: Example of Atavism

Here is one profile in June

The same profile disappears in first 2 weeks of August

This profile reappears again in last 2 weeks of August

	June 04	July 04	Aug 04-Part1	Aug 04-Part2	Sep 04
		/Write_Up_Display.Aspx/Navy Community/Navy/ApprovedInteriorCoatings&Write_Ups_Id/BilgePreservation **/Write_Up_Display.Aspx/Navy Community/Navy/Approved InteriorCoatings&Write_Ups_Id/Ballast Tanks**		**/Write_Up_Display.Aspx/NavyCommunity/Navy/Approved Interior Coatings&Write_Ups_Id/Bilge Preservation** **/Write_Up_Display.Aspx/Write_Ups_Id/Project : MSC Non-Skid** **/Write_Up_Display.Aspx/NavyCommunity/Navy/Approval Of Coatings For Navy Use (Roadmap)**	**Write_Up_Display.Aspx/NavyCommunity/Navy/Approved Interior Coatings&Write_Ups_Id/Bilge Preservation** **/Write_Up_Display.Aspx/NavyCommunity/Navy/Approved Exterior Coatings&Write_Ups_Id/Non-Skid Walk**

Why track Evolving Profiles?

- Form long term evolution patterns for interesting profiles
  - Predict seasonality
  - Support marketing efforts (if marketing campaigns are performed during these periods)
  - Forecast profile re-emergence to improve downstream personalization process via a caching process
    - Frequent atavism ➔ profile should be cached
- Help improve **scalability** of Web usage mining algorithm
  - Process Web usage data in **batches**
  - **Integrate** tracking evolving profiles **within** mining algorithm
  - **Maintain** previously discovered profiles
  - **Eliminate** a majority of the new sessions from analysis (if similar to existing profiles)
  - Focus on typically **smaller data** consisting of sessions from **truly emerging** user profiles

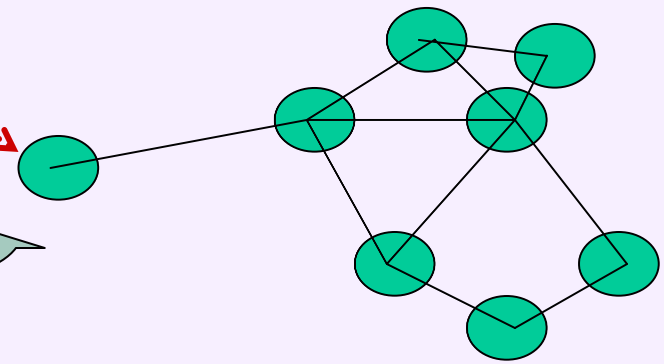
Recommender Systems in Dynamic Usage Environments

- For massive Data streams, must use a stream mining framework
- Furthermore must be able to continuously mine evolving data streams
- TECNO-Streams: Tracking Evolving Clusters in Noisy Streams
  - Inspired by the immune system
  - Immune system: interaction between external agents (antigens) and immune memory (B-cells)
  - Artificial immune system:
    - Antigens = data stream
    - B-cells = cluster/profile stream synopsis = evolving memory
    - B-cells have an age (since their creation)
    - Gradual forgetting of older B-cells
    - B-cells compete to survive by cloning multiple copies of themselves
    - Cloning is proportional to the B-cell stimulation
    - B-cell stimulation: defined as density criterion of data around a profile (this is what is being optimized!)

O. Nasraoui, C. Cardona, C. Rojas, and F. Gonzalez. Mining Evolving User Profiles in Noisy Web Clickstream Data with a Scalable Immune System Clustering Algorithm, in Proc. of WebKDD 2003, Washington DC, Aug. 2003, 71-81.

The Immune Network → Memory

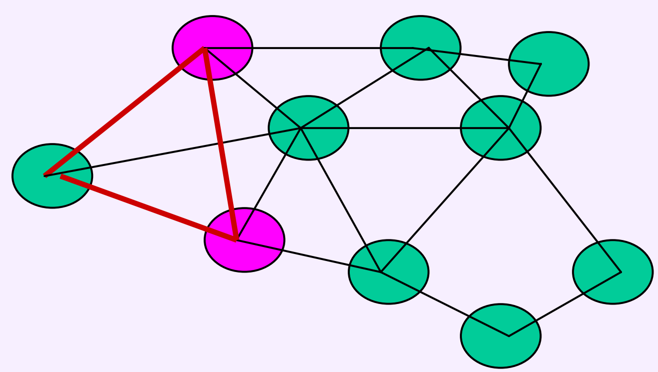
External antigen (RED) stimulates binding B-cell → B-cell (GREEN) clones copies of itself (PINK)



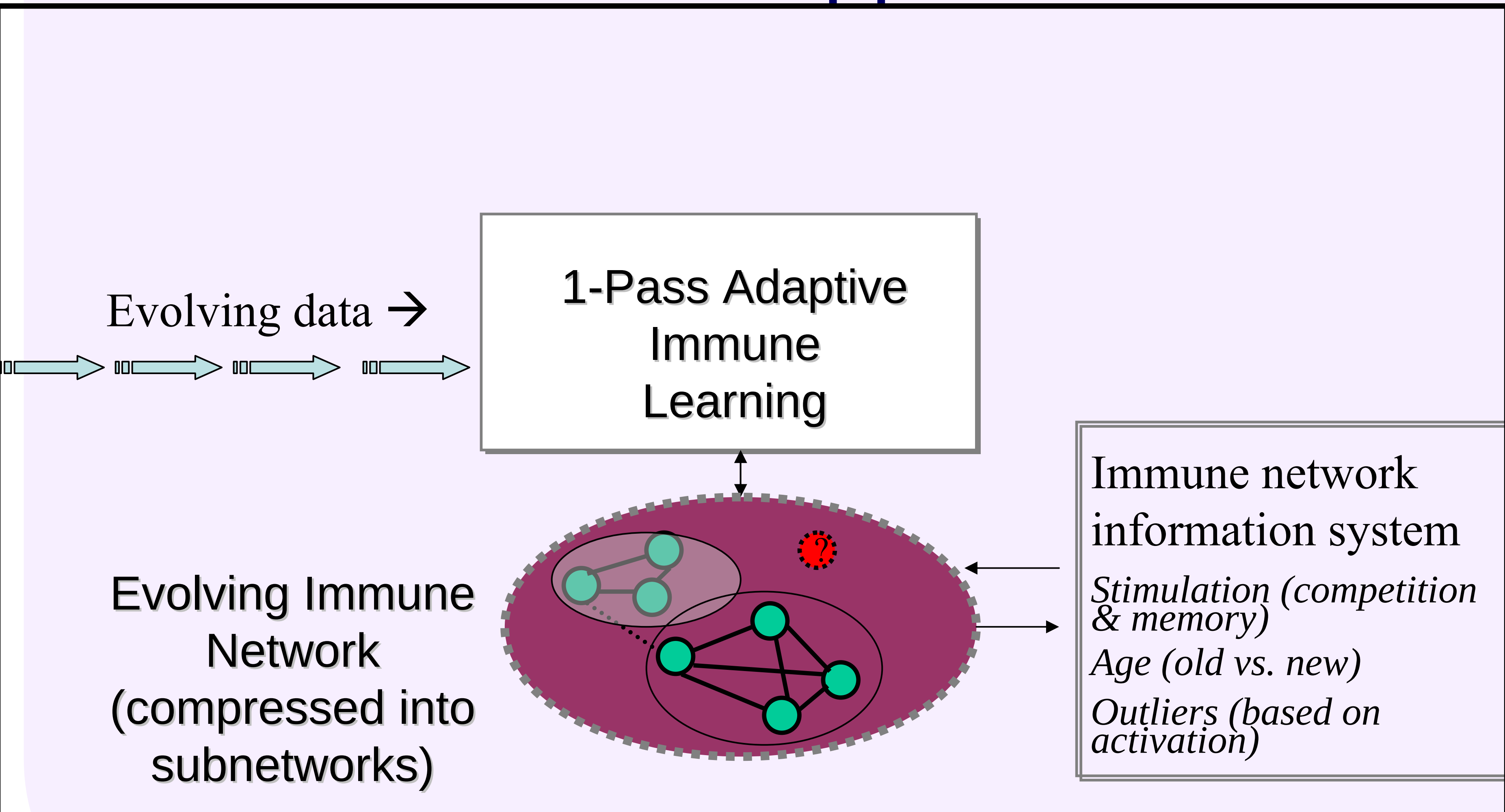
Stimulation breeds Survival

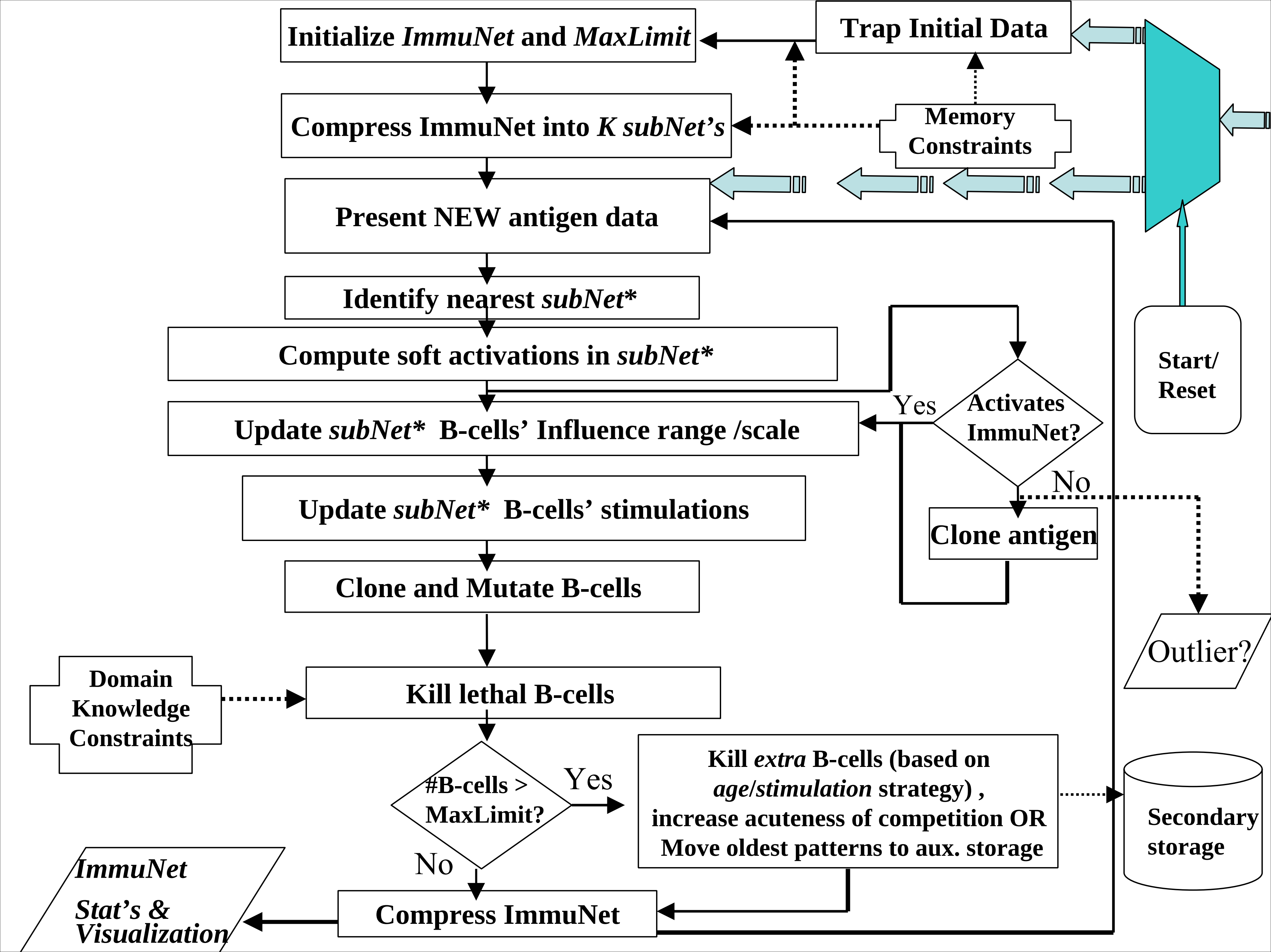
Even after external antigen disappears:

B-cells **co-stimulate** each other → thus sustaining each other → **Memory**!



General Architecture of TECNO-Streams Approach





Initialize ImmuNet and MaxLimit
Trap Initial Data
Compress ImmuNet into K subNet's
Memory Constraints
Present NEW antigen data
Identify nearest subNet*
Compute soft activations in subNet*
Update subNet* B-cells' Influence range /scale
Yes
Activates ImmuNet?
No
Start/ Reset
Update subNet* B-cells' stimulations
Clone antigen
Clone and Mutate B-cells
Outlier?
Domain Knowledge Constraints
Kill lethal B-cells
#B-cells > MaxLimit?
Yes
Kill extra B-cells (based on age/stimulation strategy) , increase acuteness of competition OR Move oldest patterns to aux. storage
Secondary storage
No
ImmuNet Stat's & Visualization
Compress ImmuNet

Adherence to Requirements for Clustering Data Streams (Barbara' 02)

- **Compactness of representation**
  - Network of B-cells: each cell can recognize several antigens
  - B-cells compressed into clusters/sub-networks
- **Fast incremental processing of new data points**
  - New antigen influences only activated sub-network
  - Activated cells updated incrementally
  - Proposed approach learns in **1 pass**.
- **Clear and fast identification of "outliers"**
  - New antigen that does not activate any subnetwork is a potential outlier ➔ create new B-cell to recognize it
  - This new B-cell could grow into a subnetwork (if it is stimulated by a new trend) or die/move to disk (if outlier)

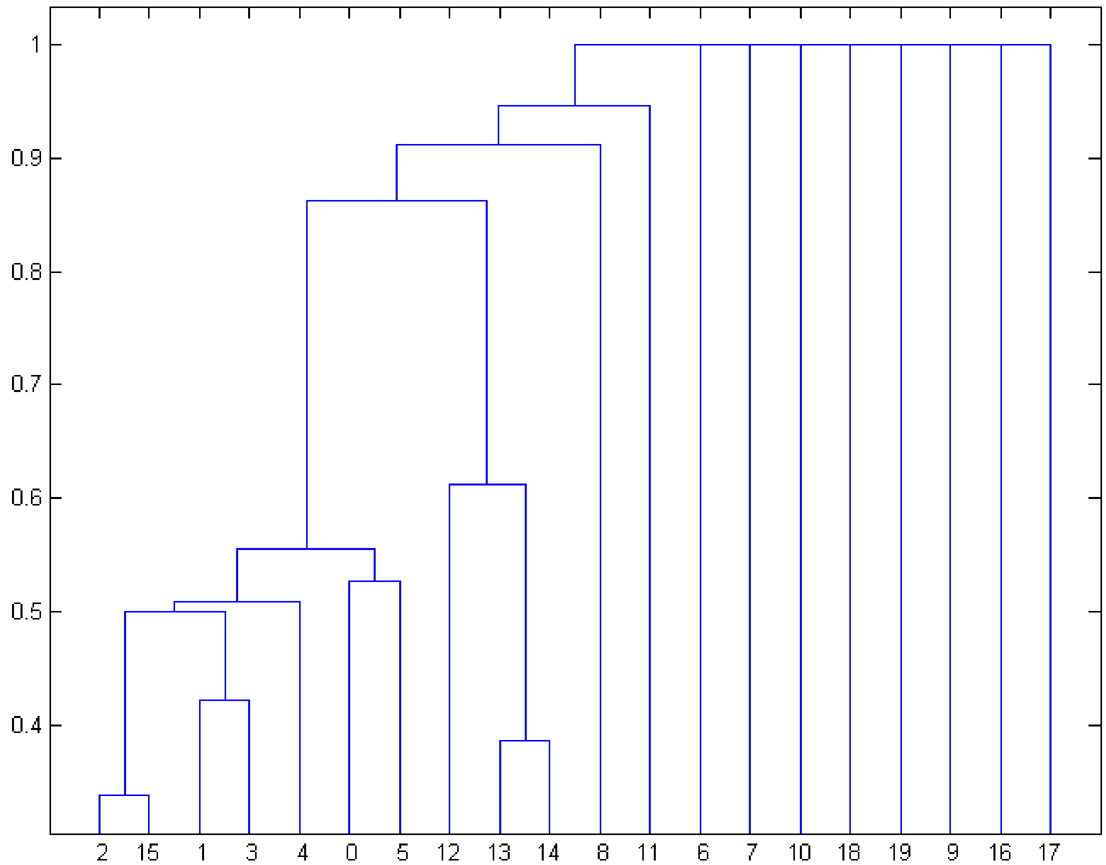
Validation Methodology in Dynamic Environments

- Limit Working Capacity (memory) for Profile Synopsis in TECNO-Streams (or Instance Base for K-NN) to 30 cells/instances
- Perform 1 pass mining + validation
  - First present all combination subset(s) of a real ground-truth session to recommender,
    - Determine closest neighborhood of profiles from TECNO-Stream's synopsis (or instances for KNN)
    - Accumulate URLs in neighborhood
    - Sort and select top N URLs ➔ Recommendations
  - Then Validate against ground-truth/complete session (precision, coverage, F1),
  - Finally present complete session to TECNO-Streams (and K-NN)

Validation Methodology in Dynamic Environments

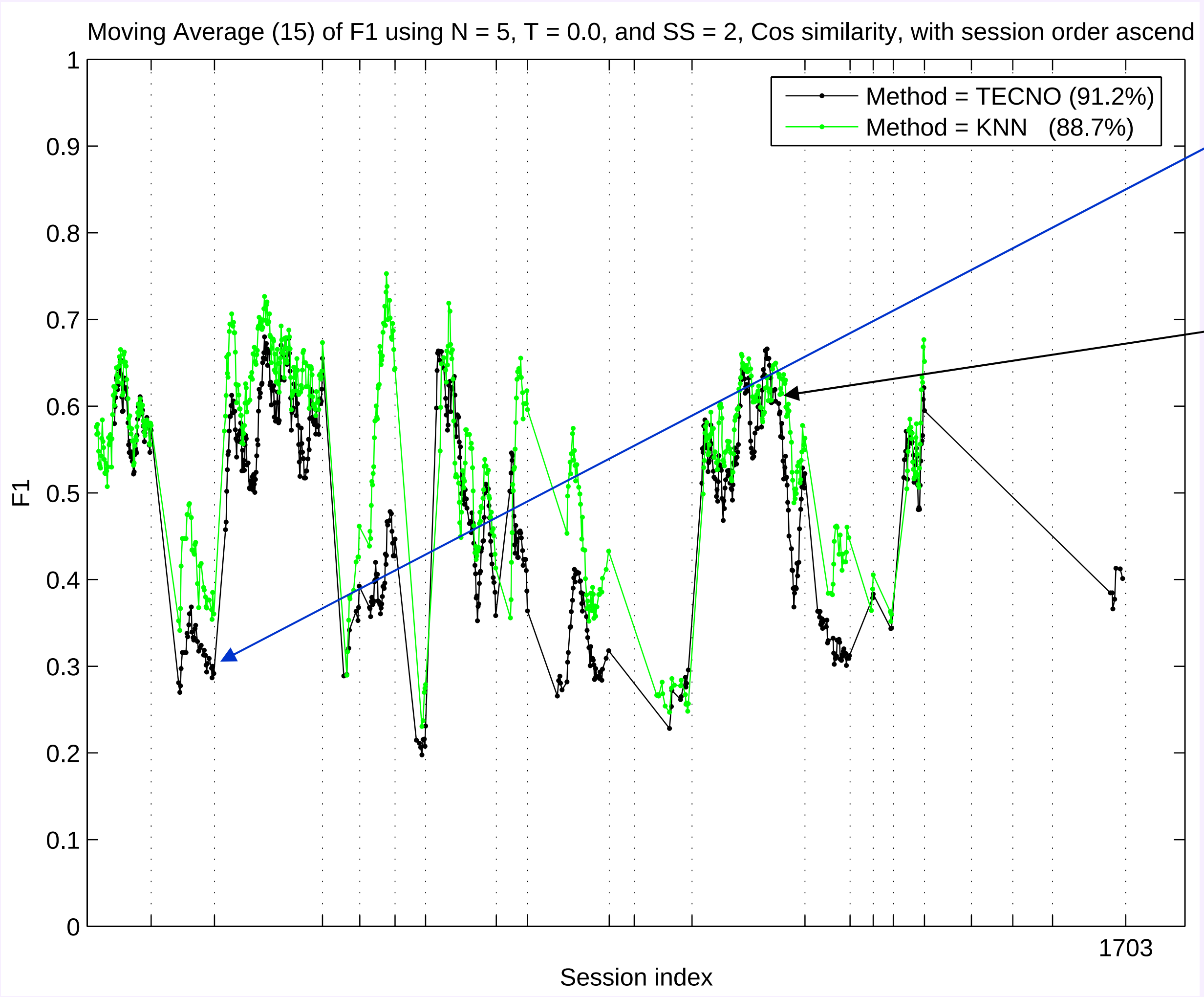
- **Scenario D (Drastic changes):**
  - We partitioned real Web sessions into 20 distinct sets of sessions, each one assigned to one of 20 previously discovered and validated profiles.
  - Then we presented these sessions to the immune clustering + recommendation + validation algorithm one profile at a time. That is, we first present the sessions assigned to ground truth profile/trend 0, then the sessions assigned to profile 1, …, etc.
- **Scenario M (Mild changes):** present Web sessions in chronological order exactly as they were received in real time by the web server
- **Scenario (Repeating Drastic changes):** Same as Scenario D, but presented profiles 1,2,3,4,5,1,2,3,4,5 (Repetition).

Dendogram of the 20 profile (vectors) 1.7K sessions, 343 URLs



Memory capacity limited to 30 nodes in TECNO-Streams' synopsis, 30 KNN-instances

Drastic Changes: F1 versus session number (vertical lines: environment changes),1.7K sessions

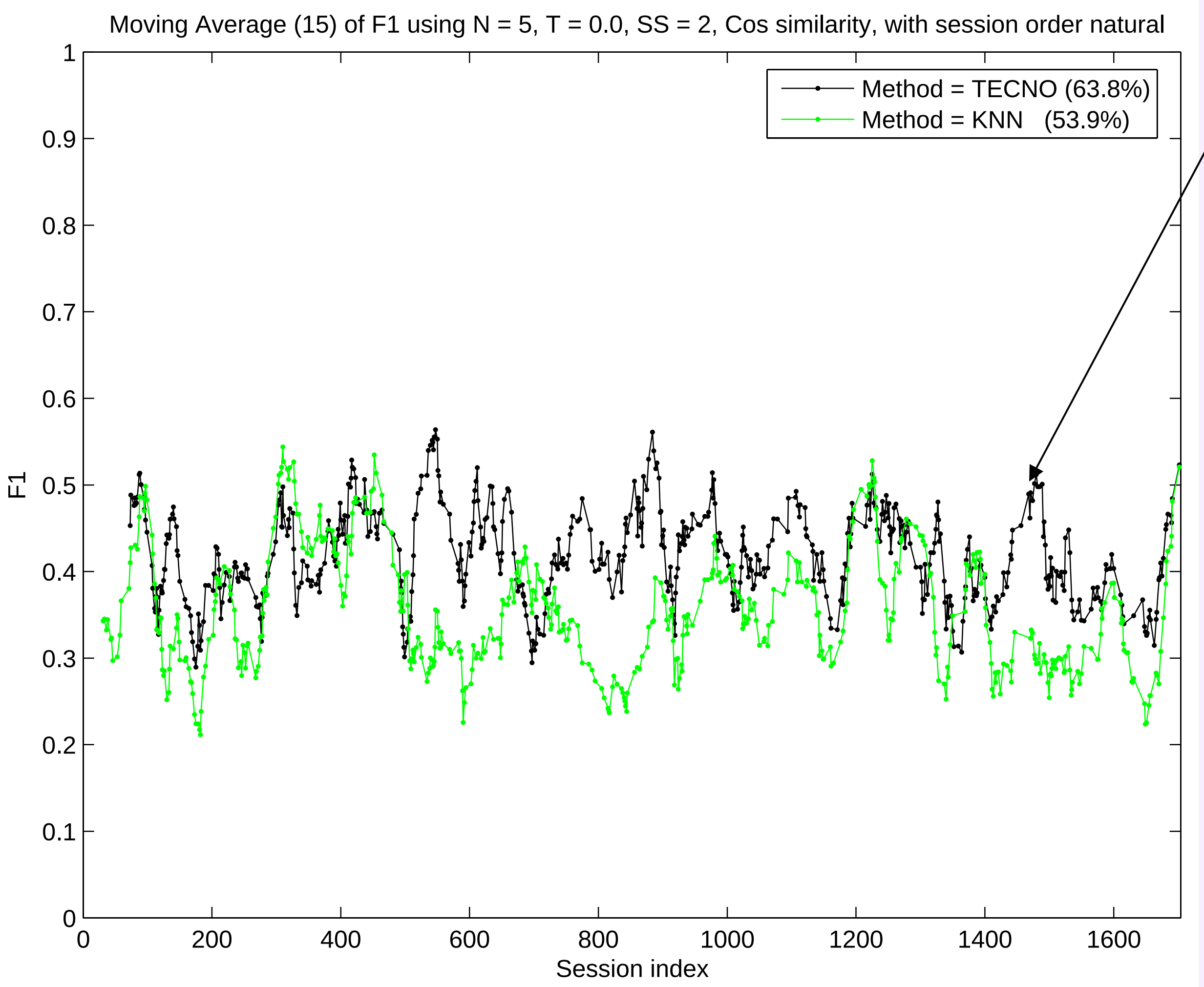


Ramp-up: both deteriorate equally as environment changes

- With sustained environment, KNN climbs higher (intense memorization of immediate past)

- On the other hand TECNO-Streams forms a “compressed” summary via optimization ➔ lossy compression

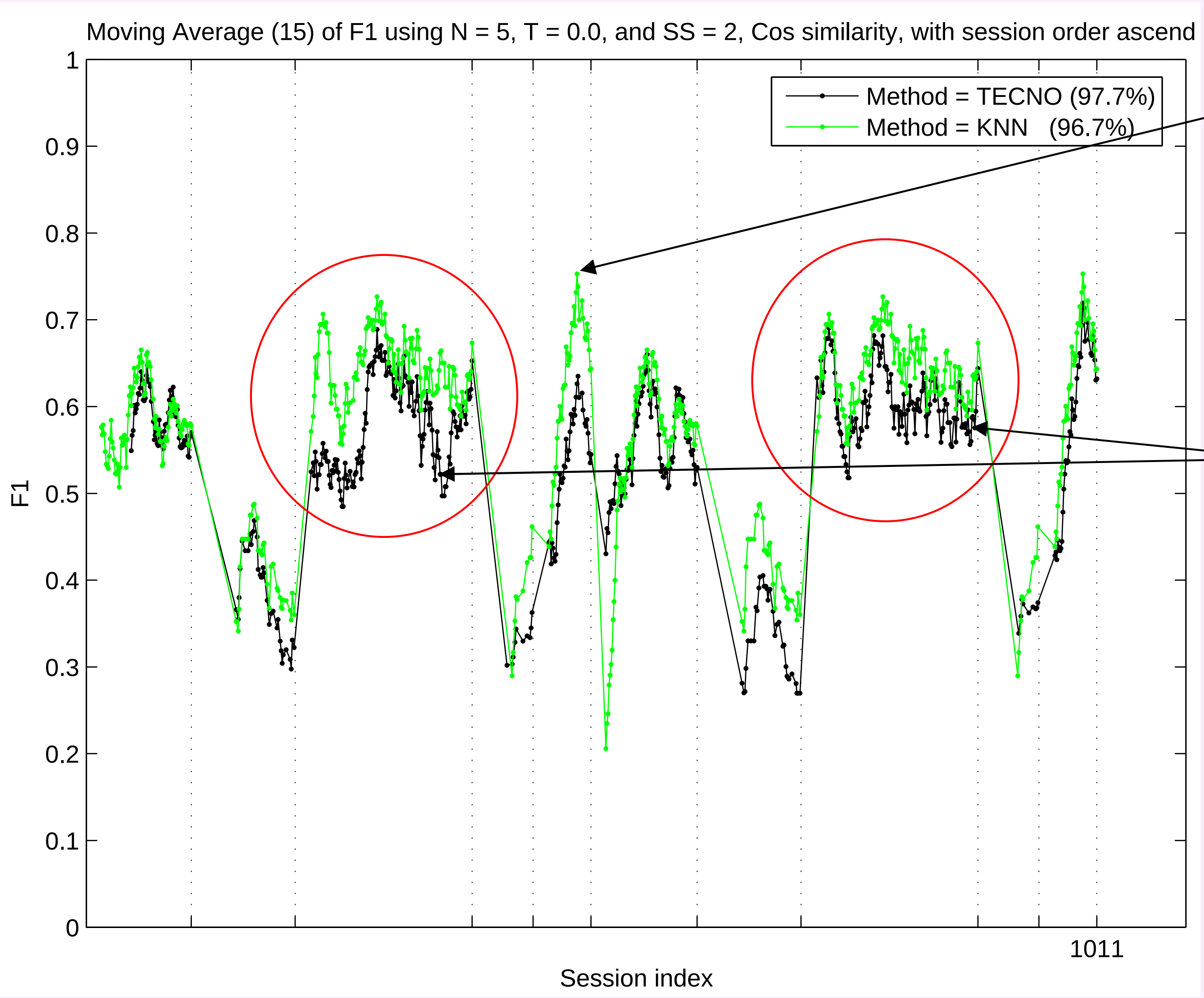
Mild Changes: F1 versus session number, 1.7K sessions



TECNO-Streams higher (noisy, naturally occurring but unexpected fluctuation call for more intelligent optimization?)

The real challenge is that here, ALL 20 usage trends are presented simultaneously as opposed to one at a time (scenario M)!

Repeating Drastic Changes: F1 versus session number (vertical lines: environment changes), 1.7K sessions



KNN higher (same as drastic: intense memorization of immediate past)

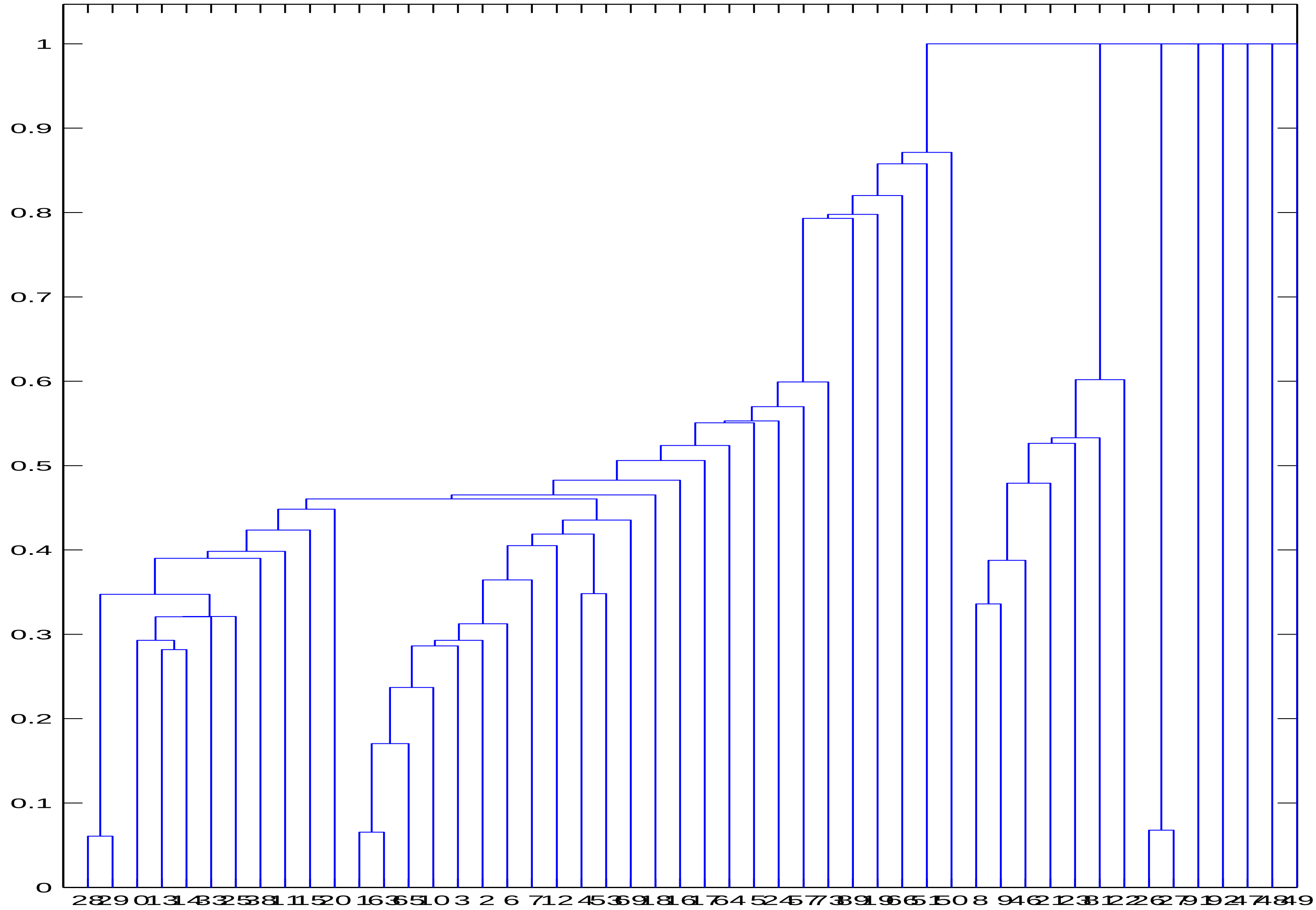
However, the 2nd time that a past environment re-occurs:

-TECNO-Stream's performance improves significantly compared to the 1st time (longer term memory, 2ndary immune response known to be stronger)

- KNN's performance remains identical to the 1st time (deterministic)

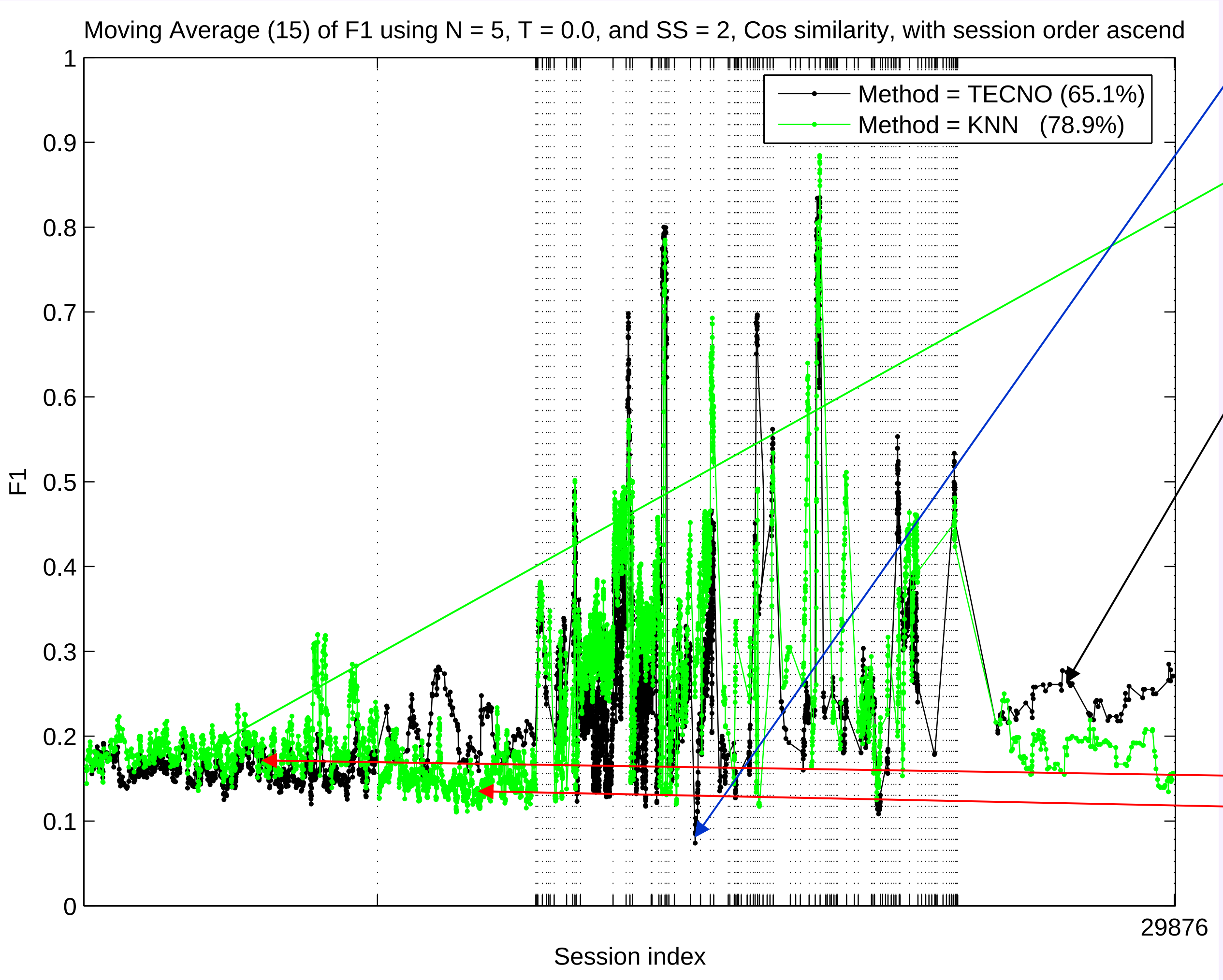
Dendogram of the 93 profile (vectors)

Bigger Data Set (~ 30K sessions, 30K URLs):



Memory capacity limited to 150 nodes in TECNO-Streams' synopsis, 150 KNN-instances

Bigger Data Set (~ 30K sessions, ~18K items): Drastic Changes: F1 versus session number (vertical lines: environment changes)



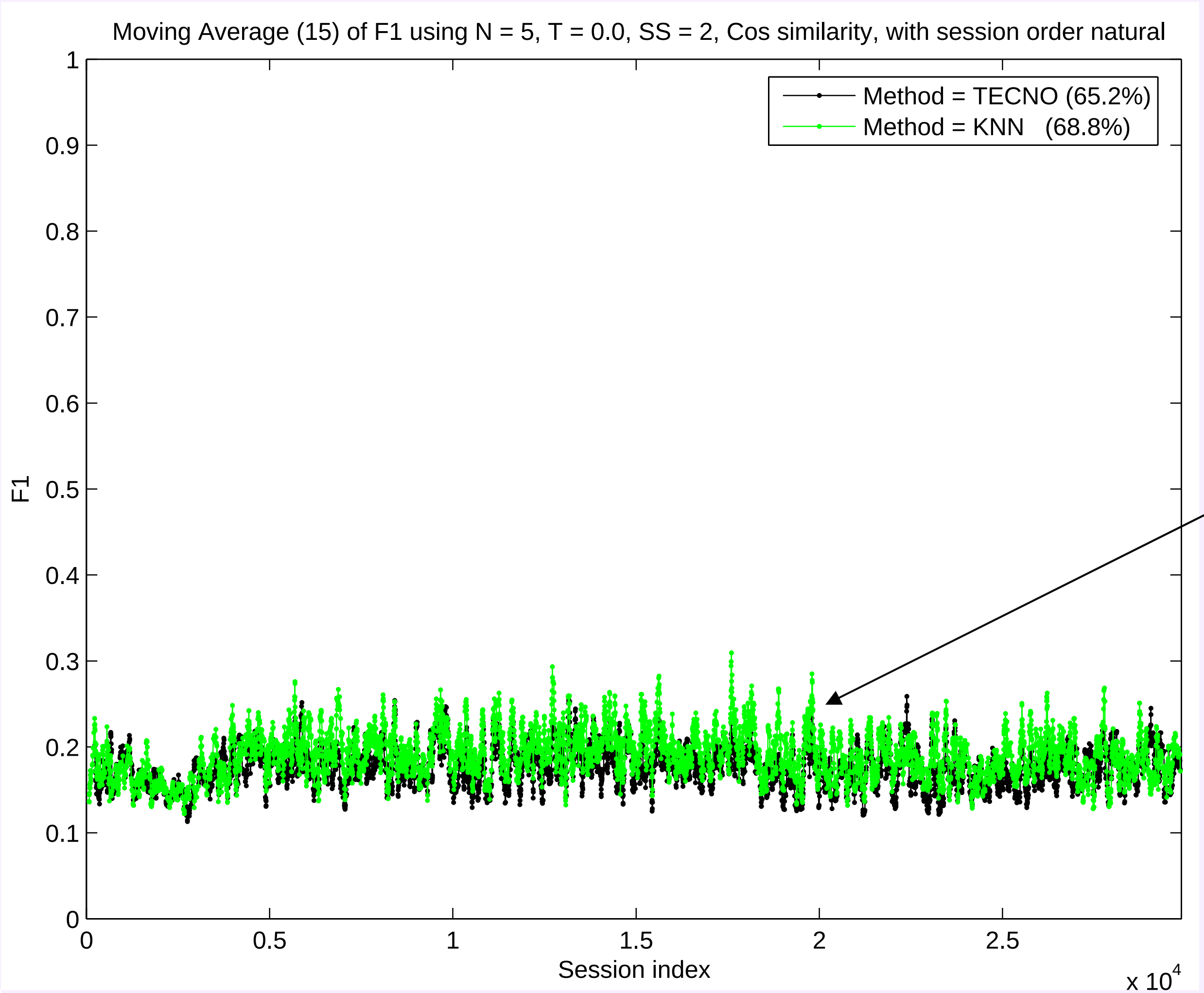
Ramp-up: both deteriorate equally as environment changes

Either one of **KNN** or T**ECNO-Streams** seem to perform better depending on profile

Overall, both recommenders' performances are very poor for "some" usage trends!!! **(Note the dimensionality and sparsity is much higher for the big data!)**

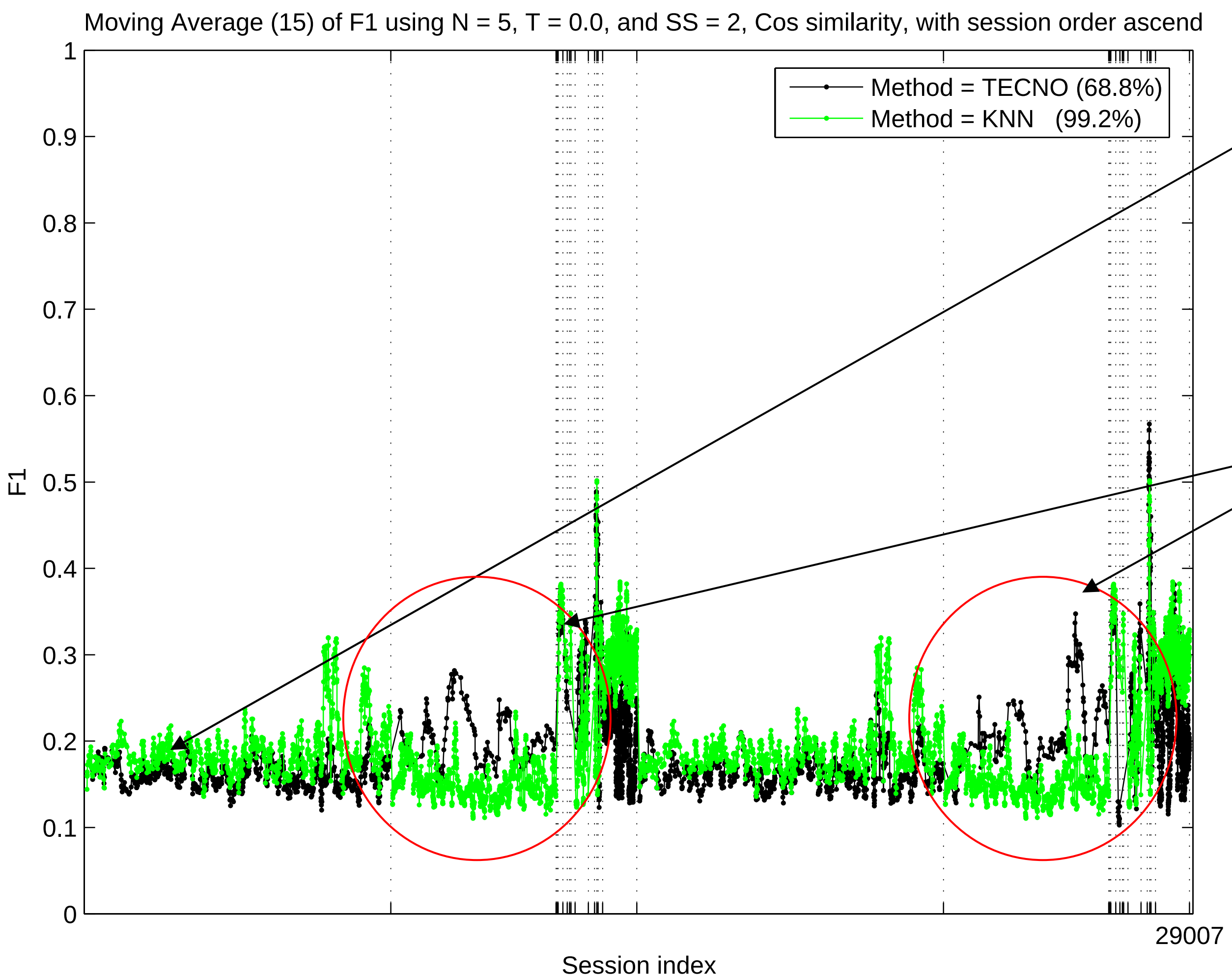
These trends are contaminated by too many noise sessions (close to 50%)!

Bigger Data Set (~ 30K sessions): Mild Changes: F1 versus session number



KNN-Streams
slightly higher ?

Bigger Data Set (~ 30K sessions): Repeating Drastic Changes: F1 versus session number (vertical lines: environment changes)



KNN slightly **higher** (same as drastic: intense memorization of immediate past)

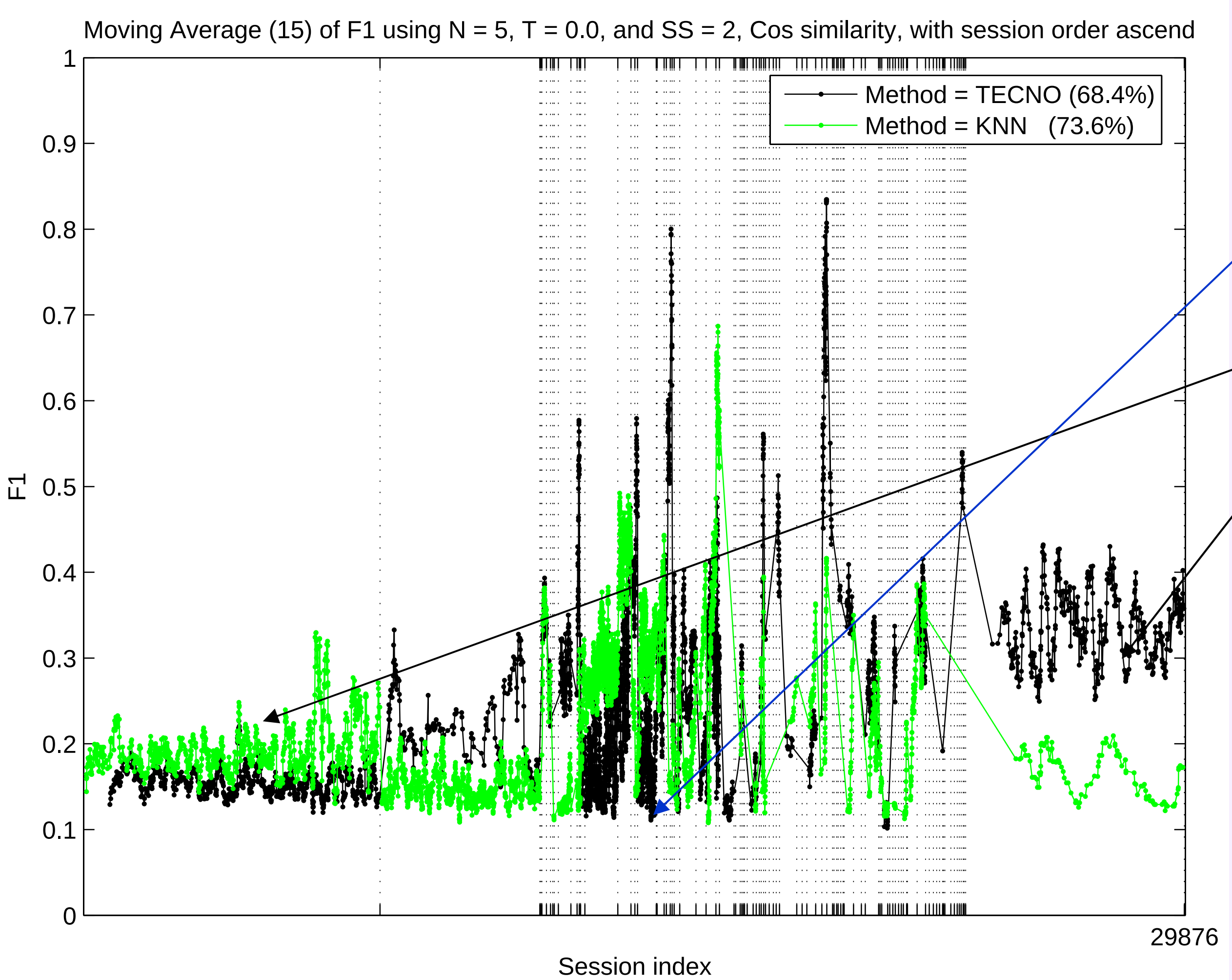
However, the 2nd time that a past environment re-occurs:

-**TECNO-Stream's** performance improves slightly compared to the 1st time (longer term memory, 2ndary immune response known to be stronger)

- KNN's performance remains identical to the 1st time (deterministic)

Memory capacity limited to 500 nodes in TECNO-Streams' synopsis, 500 KNN-instances

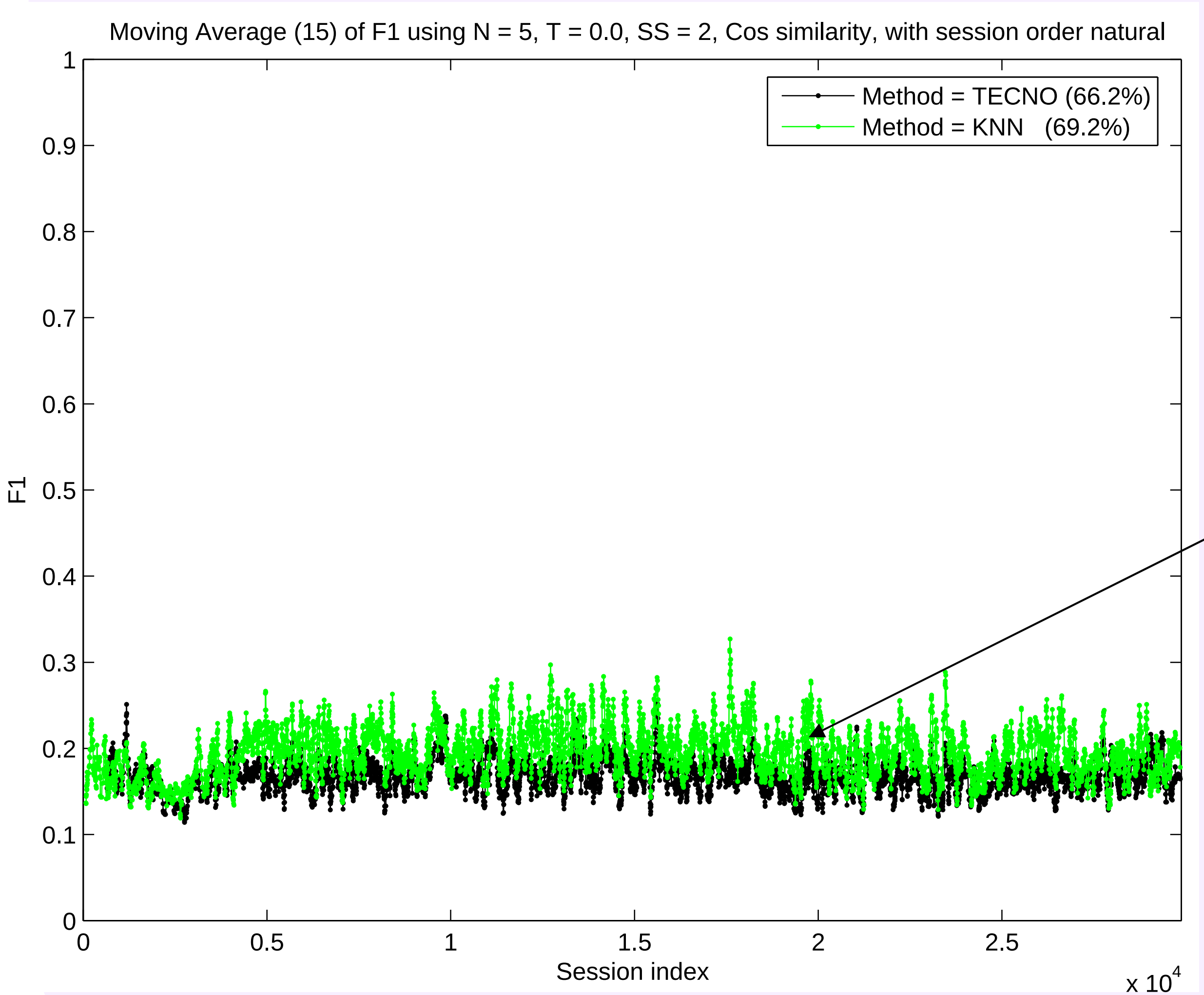
Bigger Data Set (~ 30K sessions): Drastic Changes: F1 versus session number (vertical lines: environment changes)



Ramp-up: both deteriorate equally as environment changes

Either one of **KNN** or **TECNO-Streams** seem to perform better depending on profile

Bigger Data Set (~ 30K sessions): Mild Changes: F1 versus session number



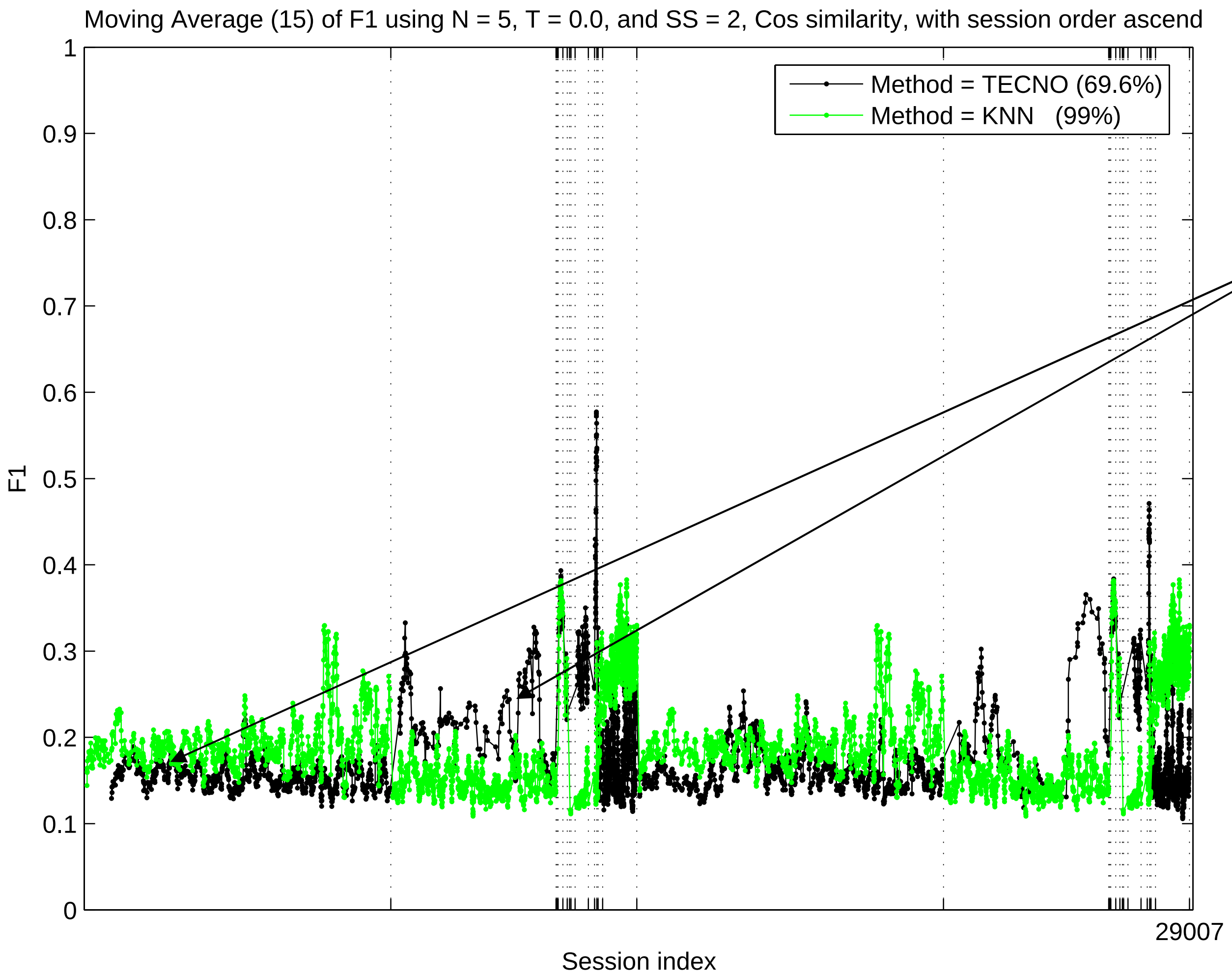
KNN-Streams slightly higher ?

But overall both are poor

Possibly because of extremely high dimensionality (>17000) and sparsity!

which wrecks havoc on Collaborative filtering in streaming environments!!!

Bigger Data Set (~ 30K sessions): Repeating Drastic Changes: F1 versus session number (vertical lines: environment changes)



Either one of **KNN** or T**ECNO-Streams** seem to perform better depending on profile

Personalization: Implementation Issues

- Fast
- Easy
- Scalable
- Cheap?
- ***Free…?***

Summary of Methodology

- Systematic framework for a fast and easy implementation and deployment of a recommendation system
- on one or several affiliated or subject-specific websites
- based on any available combination of *open source tools* that include
  - *crawling,*
  - *indexing, and*
  - *searching capabilities*

Supported Approaches

- Content based filtering (straight forward)
- Collaborative filtering (more complex)
- Hybrids that combine the power of both (2 types):
  - Cascaded (2 options):
    - First collaborative filtering (obtain collaborative recommendations), then content-based filtering (on previous result)
    - First content-based filtering (obtain content-based set of recommendations), then collaborative filtering (on previous result)
  - Parallel/combined:
    - Perform collaborative filtering on original input
    - Perform content-based filtering on original input
    - Then combine resulting recommendations above by weighting, etc.

What for?

- Easily "implement" (existing) recommendation strategies by using a search engine software when it is available,
- Benefit to research and real life applications
  - by taking advantage of search engines' scalable and built-in indexing and query matching features,
  - instead of implementing a strategy from scratch.

Advantages to Expect…

- Multi-Website Integration by Dynamic Linking:
  - dynamic, personalized, and automated linking of partnering or affiliate websites
  - Crawl several websites + connect through common proxy
- Giving Control Back to the User or Community instead of the website/business
  - no need for intervention from websites
- The Open Source Edge
- Tapping into IR Legacy

Search Engine

- 1) Crawling: A crawler retrieves the web pages that are to be included in a searchable collection,
- 2) Parsing: The crawled documents are parsed to extract the terms that they contain,
- 3) Indexing: An inverted index is typically built that maps each parsed term to a set of pages where the term is contained,
- 4) Query matching:
  - Submit input queries in the form of a set of terms to a search engine interface or to a query matching module
  - that compares this query against the existing index,
  - to produce a ranked list of results or web pages.
- Two open source products that enable a fast and free implementation of Web search,
  - Text search engine library: Lucene,
  - Web search engine: Nutch, built on Lucene

Lucene

- D. Cutting and J. Pedersen, Space optimizations for total ranking, RIAO (Computer Assisted IR) 1997
- http://lucene.apache.org/
- high-performance, full-featured text search engine library written in Java,
- can support any application that requires full-text search, especially cross-platform.
- Examples of using Lucene: Inktomi and Wikipedia's search feature
- powerful features through a simple API, include
  - scalable, high-performance indexing,
- available as Open Source software under the Apache License

Lucene's features

- ranked searching
- various query types: phrase, wildcard, proximity, fuzzy, range, and more
- fielded searching (e.g., title, author, contents),
- date-range searching,
- sorting by any field,
- multiple-index searching with merged results,
- allowing simultaneous update and searching
- All the above ➔ Heaven on Earth! for implementing recommender system

Nutch

- http://lucene.apache.org/nutch/: Lucene based Web search
- Adds Web specifics to Lucene: crawler, link-graph database, parsers for HTML and other document formats (pdf, ppt, doc, plain text, etc).
- Document = sequence of Fields .
  - Field values may be stored, indexed, analyzed (to convert to tokens), or vectored.
- Uses Lucene's index: Inverted Index that maps a term → field ID, and a set of document IDs, with the position within each document.
- Given a query, Nutch by default searches URLs, anchors, and content of documents

Proposed Methodology

Two requirements for tweaking a search engine to work like a recommender sys.

1. An index: The source of the recommendations must be indexed in a format that is easy to search.
2. A querying mechanism:
   - the input to the recommendation procedure must be transformable into a *query*
   - Query is expressed in terms of the entities upon which the index is based

Content-based filtering

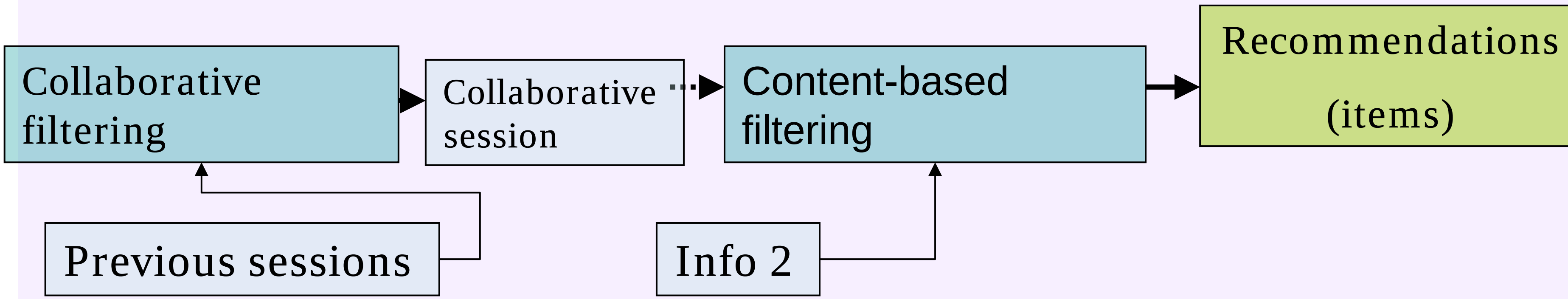
Given a few pages that a user has viewed, the system recommends other pages with content that is similar to the content of the viewed pages

- **Step 1:** Preliminary Crawling and Indexing of website(s) (done offline): to form content of the recommendations, and then forming a reverse index that maps each keyword to a set of pages in which it is contained.
  - Store the most frequent terms in each document as a vector field, that is indexed and used later in retrieval
- **Step 2:** Query Formation and Scoring: transform a new user session into a query that can be submitted to the search engine.
  - Map each URL in user session to a set of content terms (top k frequent terms) using an added package net.nutch.searcher.pageurl.
  - Combine these terms with their frequencies to form a query vector,
  - Submit query to Nutch as a Fielded query (i.e. the query vector is compared to the indexed Web document vector field).
  - Finally, rank results according to cosine similarity with the query vector in the vector space domain
    - modification of the default scoring mechanism of SortComparatorSource in the LuceneQueryOptimizer class (which is part of the package net.nutch.searcher)

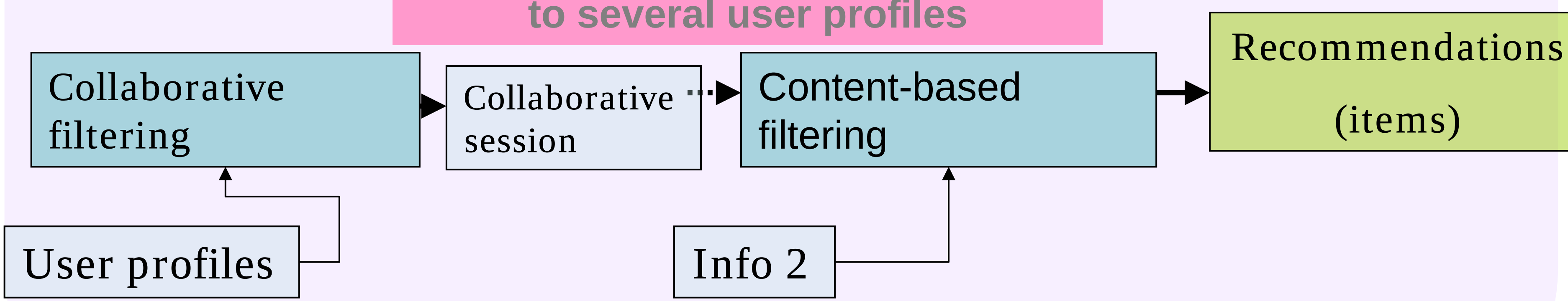
session → URLS → terms → fielded query vector → results (ranked according to cosine similarity (result vector, query vector))

Cascaded Hybrids

Type 1: compares current session to (all) previous sessions



Type 2: compares current session to several user profiles



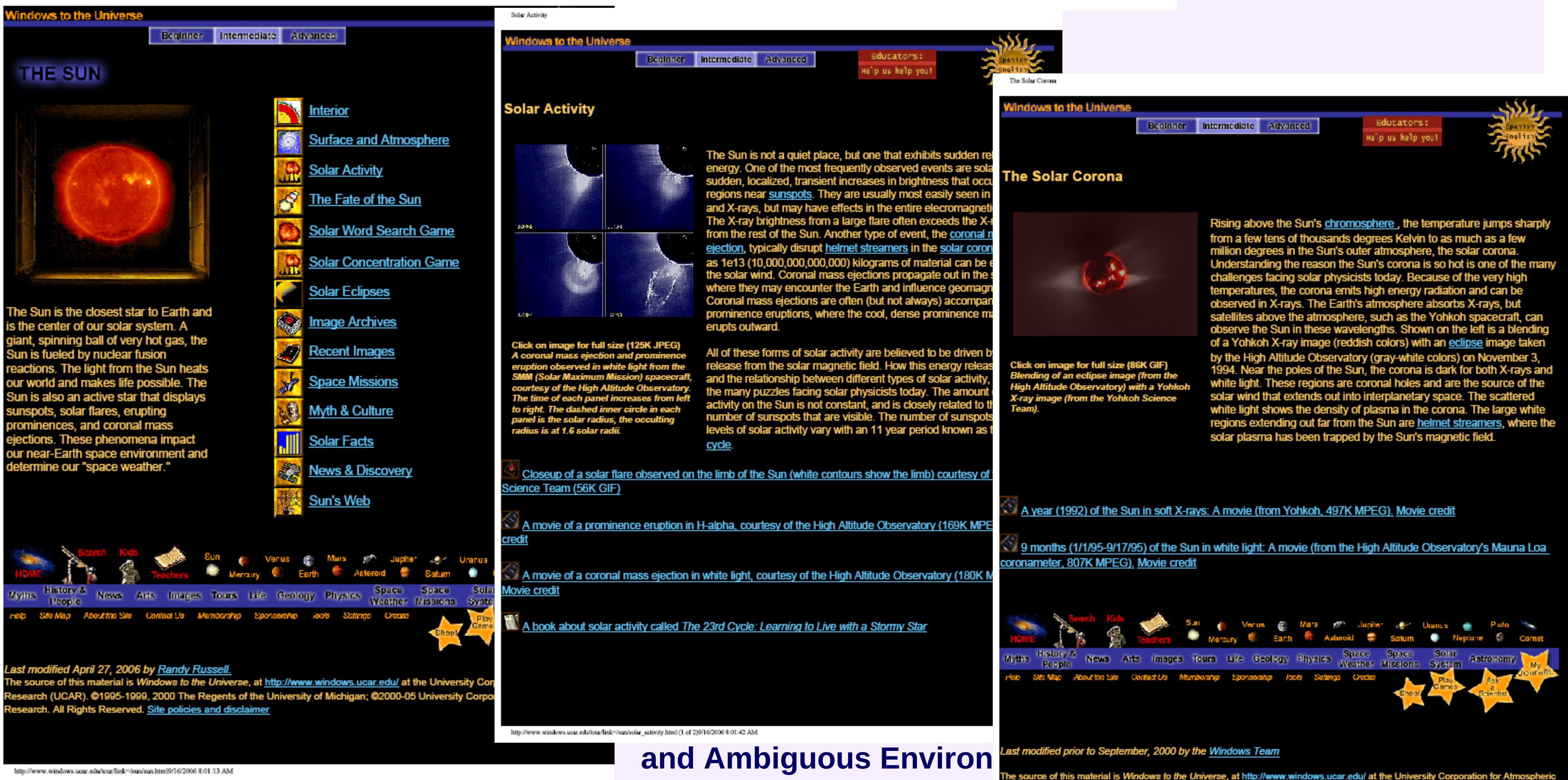
Implementation

- Crawled web pages in following domains:
  - .wikipedia.org
  - .ucar.edu
  - .nasa.gov
  - ➔ (this corresponds to Step 1 of content-based filtering)
- The content was indexed using nutch
- the nutch search engine application was launched to accept queries (in our case transformed user sessions!)
- A proxy was set at one port on our server based on the Open Source SQUID Web proxy software (http://www.squid-cache.org/)
- Additional C code to track each session, convert it to an appropriate query, and submit this query to nutch

Example

Your visit history(3):

- http://www.windows.ucar.edu/tour/link=/sun/sun.html
- http://www.windows.ucar.edu/tour/link=/sun/solar_activity.html
- http://www.windows.ucar.edu/tour/link=/sun/atmosphere/corona.html



and Ambiguous Environ

Content-based Recommendations

Hits **1-10** (out of about 137 total matching pages):

The James Webb Space Telescope
The James Webb Space Telescope + Goddard Space Flight Center + Sciences & Exploration Directorate + Exploration of **...**
http://www.jwst.nasa.gov/glossary.html

UCAR Staff Notes: HIAPER work reaches pivotal stage
UCAR Staff Notes: HIAPER work reaches pivotal stage February 2004 HAO Sunrise NCAR is involved in numerous **...**
http://www.ucar.edu/communications/staffnotes/0402/hao.html (more from www.ucar.edu)

Magnetohydrodynamics - Wikipedia, the free encyclopedia
Magnetohydrodynamics - Wikipedia, the free encyclopedia Your continued donations keep Wikipedia running! Magnetohydrodynamics From Wikipedia, the **...**
http://en.wikipedia.org/wiki/Magnetohydrodynamics (more from en.wikipedia.org)

Imagine the Universe Science
Imagine the Universe Science What's New Site Map NASA Homepage Search: HOME Science Special Exhibit **...**
http://imagine.gsfc.nasa.gov/docs/science/science.html

Talk:Plasma cosmology/Archive 3 - Wikipedia, the free encyclopedia
Talk:Plasma cosmology/Archive 3 - Wikipedia, the free encyclopedia Your continued donations keep Wikipedia running! Talk:Plasma **...**
http://en.wikipedia.org/wiki/Talk:Plasma_cosmology/Archive_3 (more from en.wikipedia.org)

Outstanding Accomplishment Awards

Outstanding Accomplishment Awards The Laurels Pages Awards | Patents | Outstanding Accomplishments | UCAR Awards Database advanced Outstanding Accomplishments Awards ...
http://www.ucar.edu/communications/performance/perfretro.shtml (more from www.ucar.edu)

1998 Eclipse: Enhanced POISE98 Image

1998 Eclipse: Enhanced POISE98 Image 1998 Eclipse: Enhanced POISE98 Image Enhanced, calibra "Intensity" image of the ...
http://mlso.hao.ucar.edu/eclipse98_scarab.html

Publications

Publications Publications Refereed Publications Athay, R. G. and P. G. Judge, 1995: Excitation of O
http://web.hao.ucar.edu/public/asr/asr95/pub.html

NASA - Science Highlights from NASA's Sun-Solar System Connection

NASA - Science Highlights from NASA's Sun-Solar System Connection The nasa.gov site requires tl JavaScripts ...
http://www.nasa.gov/centers/goddard/2004solar.html

SPARTAN 201-3: The Corona

SPARTAN 201-3: The Corona The Corona At the center of ...
http://umbra.nascom.nasa.gov/spartan/the_corona.html

The James Webb Space Telescope

National Aeronautics and Space Administration

+ Goddard Space Flight Center
+ Sciences & Exploration Directorate
+ Exploration of the Universe Division

FIND IT @ NASA :

+ GO

+ Advanced Search

FAST FACTS | FAQ | GLOSSARY | SITE MAP | CONTACT US

HOME
ABOUT JWST
SCIENCE
OBSERVATORY
INSTRUMENTS
NEW TECHNOLOGY
PEOPLE
MULTIMEDIA
FOR SCIENTISTS
FOR PUBLIC
FOR PRESS

Glossary

[A-D] [E-H] [I-M] [N-R] [S-Z]

AACS: Attitude
AAS: American
AAT: Anglo-Au
AGB: Asympto
AGN: Active G
AGU: American
Albedo: Reflec
ALOT: Adaptiv
AMSD: Advanc
AO: Announce

http://www.jwst.nasa.gov/glossary.html (1 of 19)9/16/2006 7:46:34 A

The James Webb Space Telescope

Constellation: A group of stars that, from the vantage point of Earth, appear to make a shape. They are often named after mythological characters, people, animals, and things.

Convection: A transfer of heat energy by circulation through a gas or liquid.

Corona: The Sun's outer atmosphere.

Coronagraph: An instrument used to block out the bright light from the sun or other stars so that nearby faint objects are visible.

Cosmology: The study of the origin, evolution and structure of the universe (the cosmos).

COSPAR: Committee on Space Research

COSTAR: Corrective Optics Space Telescope Axial Replacement (HST)

COTS: Commercial Off-The-Shelf

CPI: Continuous Process Improvement

Crater: A round impression left in a planet or satellite from a meteoroid collision.

• About Staff Notes
• Past Issues
• Favorite Photos
• Communications Home
• UCAR Home
• Search

staff notes monthly

Febru

HAO

NCAR is involved in numerous internationa
Staff Notes Monthly will be highlighting sev

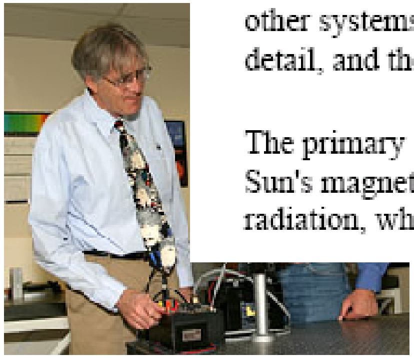
NCAR scientists and engineers, working wi
moving ahead with plans to build a powerfu
the atmosphere by a balloon.

"Throughout this decade, we'll be seeing much more detailed images of the Sun from a number of experiments," Bruce says "Sunrise should be a big part of this research picture."

New insights into magnetic fields

Learning more about the Sun has long been a high scientific priority. The Sun is the main source of light and energy for life on Earth and the principal driver of atmospheric motion. Solar disruptions, such as coronal mass ejections, have profound impacts on our upper atmosphere, touching off geomagnetic storms that affect sensitive communications and other systems on Earth. But scientists need specialized instruments to examine the Sun in detail, and they remain uncertain about the causes of solar variability and disruptions.

The primary goal of the Sunrise project is to investigate the structure and dynamics of the Sun's magnetic field. The magnetic field fuels solar activity and causes variations in radiation, which may be a significant factor in long-term changes in our climate.



Bruce Lites and Kim Streander

At the heart of what's known as the Sunrise project is a lightweight, one-meter telescope that will circle Antarctica for about two weeks at an altitude of approximately 130,000 feet (39,600 meters). Its advanced instrumentation will provide high-resolution images of the Sun's outer surface, or photosphere, enabling scientists to get unprecedented views of small-scale magnetic fields that drive solar variability and profoundly affect Earth's atmosphere.

The international team expects to launch the telescope in late 2007. If the instrument can

nalization in Noisy, Dynamic,
vironments

Magnetohydrodynamics

From Wikipedia, the free encyclopedia

Jump to: navigation, search

Magnetohydrodynamics (MHD) (*magnetofluiddynamics* or *hydromagnetics*) is the academic discipline which studies the dynamics of electrically conducting fluids. Examples of such fluids include plasmas, liquid metals, and salt water. The word *magnetohydrodynamics (MHD)* is derived from *magneto-* meaning magnetic field, and *hydro-* meaning fluid, and -dynamics meaning movement. The field of MHD was initiated by Hannes Alfvén[1], for which he received the Nobel Prize in 1970.

The set of equations which describe MHD are a combination of the Navier-Stokes equations of fluid dynamics and Maxwell's equations of electromagnetism. These differential equations have to be solved simultaneously. This is too complex or impossible to do symbolically in most cases, but there are important classes of analytical solutions (for example, the Solov'ev equilibria). For real-world problems in complex geometries, numeric solutions are found using computers. Because MHD is a *fluid* theory, it cannot treat *kinetic* phenomena, i.e., those in which the existence of discrete particles, or of a non-thermal distribution of their velocities, is important.

Contents

[hide]

Astrophysics

[edit]

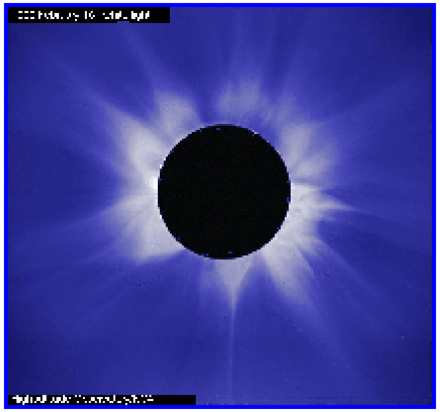
MHD applies quite well to astrophysics since over 99% of the matter content of the Universe is made up of plasma, including stars, the interplanetary medium (space between the planets), the interstellar medium (space between the stars), nebulae and jets. Many astrophysical systems are not in local thermal equilibrium, and therefore require an additional kinematic treatment to describe all the phenomena within the system (see Astrophysical plasma).

Sunspots are caused by the Sun's magnetic fields, as Joseph Larmor theorized in 1919. The solar wind is also governed by MHD. The differential solar rotation may be the long term effect of magnetic drag at the poles of the Sun, an MHD phenomenon due to the Parker spiral shape assumed by the extended magnetic field of the Sun.

Previously, theories describing the creation of the Sun and planets could not explain how the Sun has

The Corona

At the center of our solar system there is a magnetic variable star, our Sun, which drives every cubic centimeter of interplanetary space. The upper atmosphere of the Sun, the *solar corona* extends from the visible disk of the Sun outward, eventually enveloping the earth. The earth, our home planet, is located at a distance of about 200 solar radii from the visible surface of the Sun. The dimension of a solar radius is roughly 700,000 km, approximately twice the distance from the earth to the Moon, and the solar radius is a convenient scale for discussing the solar corona, and the heliosphere, the extension of the solar atmosphere into interplanetary and interstellar space. Astronomers feel comfortable using the solar radius as a measure of length when discussing the corona, the interplanetary medium, and the sizes of other stars.



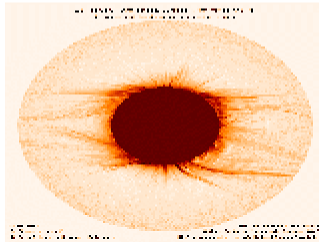
Click image for full-size GIF

Total solar eclipse images of 1980 February (above) and 1988 March (below) taken from sites located in India (1980) and the Philippines (1988) by expeditions from the High Altitude Observatory of Boulder, Colorado. Note that the 1980 image, taken near the *maximum* of the solar activity cycle shows many streamers located at all azimuths around the occulted disk of the Sun. Taken later in the cycle, about a year past the *minimum*, the 1988 image shows several large (bottle-shaped) *helmet streamers* which are restricted to latitudes between N45 and S45. The helmet streamers, which are large scale, dense structures, have measured lifetimes from less than one to more than several solar rotations.

A special telescope, known as the White Light Coronal Camera, was used for both of these observations. Half of the diameter of the dark central image of the moon is equal to a distance of one solar radius.

ynamic,

1998 Eclipse: Enhanced POISE98 Image



Enhanced, calibrated "Intensity" image of the solar corona made at the total solar eclipse on 26 February 1998 in Curacao, Netherlands Antilles, with the HAO Polarimetric Instrument for the Solar Eclipse 1998 (POISE98). This image shows both the plumes at the solar poles and the coronal streamers nearer the Sun's equator at onset of Solar Cycle 23.

The POISE has an aperture of 80 mm, an effective focal length of 1000 mm, and a field of view of 6.5X6.5 solar radii on a Loral CCD with 2034X2034 pixels. The pixel size is 3.1X3.1 arc seconds. The spectral band of the image is set by an Andover filter with a central wavelength of 620 nm and a bandwidth of 10 nm. The CCD camera is a Pixel Vision Spectra Video camera with 16 bit digitization, an electron well depth of 95,000 electrons, and a read noise 20 electrons. The polarization analyzer contains fixed linear and quarter wave polarizers together with a Meadowlark Optics liquid crystal variable retarder. This image is a calibrated "intensity" composite of a series of 0.25, 1.0 and 4 second exposures.

The POISE98 images can be processed numerically to remove the radial gradient due to the rapid outward decrease in the coronal density. This numerical 'flattening' replaces the radially graded optical filter used in the photographic Newkirk camera which was used in previous HAO eclipse expeditions.

Specifically, the processing involved to produce this image was: dividing a r^-5 filtered, 2-D sobel filtered image, by a r^-5 filtered image. This results in great edge enhancements and discernibility of fine structure in the data.

Such fine details improve understanding of a fundamental solar physics question:

> "Why is the solar corona so hot, and how does it get that way?"

Is is known that most of the heating occurs very close to the solar limb, that is, very low in the corona. And, is only at eclipses where one can observe the corona in this region.

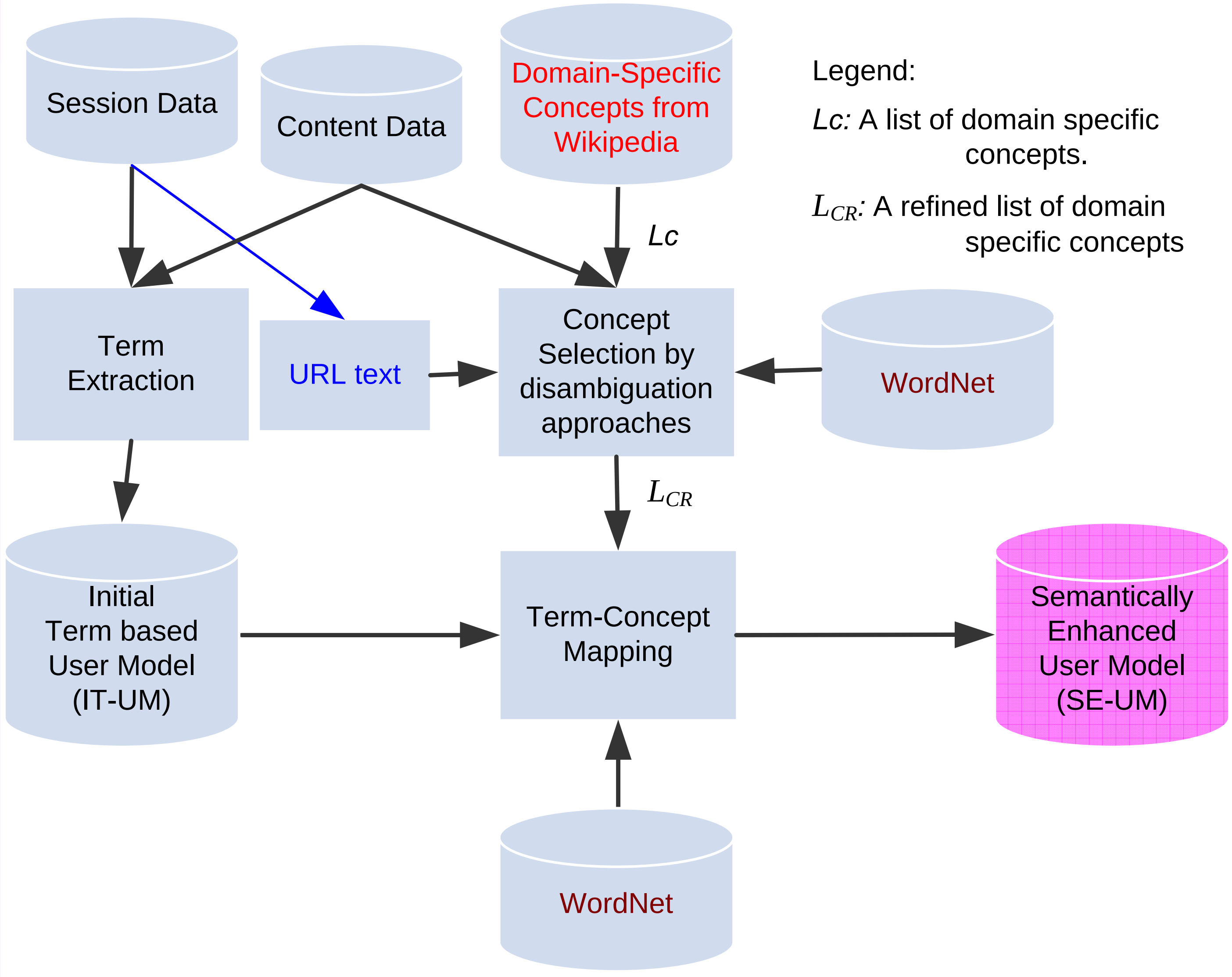
ic,

The HAO expedition team for the '98 eclipse is Alice Lecinski, Kim Streander, David Elmore,

Web Usage Mining:

- **Ambiguity:**
  - Implicit Semantics:
  - Explicit Semantics:
  - What is effect of generalization / URL compression?
- **Noise: Effect of post-processing:**
  - **Robust profiles**
  - **Frequency averaging**
- **Detecting and characterizing evolution in dynamic environments**
- **Recommender Systems in dynamic environments**
- **Recommender Implementation**
- **Mining Conceptual Web Clickstreams**

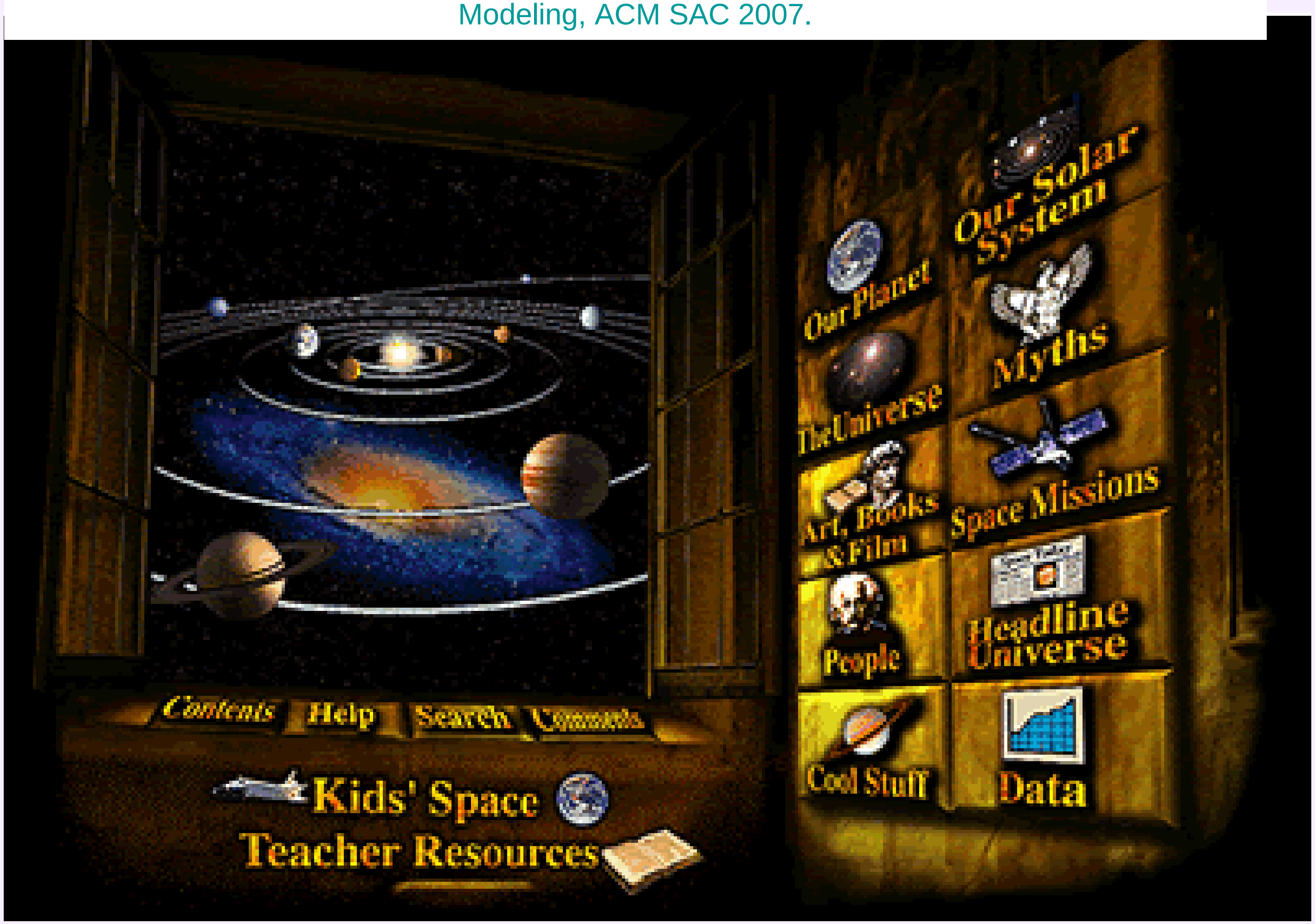
Conceptual User Session Modeling (w/ lead author: Dr. Hyoil Han, Drexel Univ.)



P. Achananuparp, H. Han, O. Nasraoui and R. Johnson, Semantically Enhanced User Modeling, ACM SAC 2007,
also in Tech. report No IST TR-06-1, Drexel University, September 2006.

Windows to the Universe: http://www.windows.ucar.edu (education & outreach website for NASA, NCAR, and other research agencies/groups)

P. Achananuparp, H. Han, O. Nasraoui and R. Johnson, Semantically Enhanced User Modeling, ACM SAC 2007.



Use Wikipedia categories to get large set of Concept terms (specific to physics, astronomy, earth science, etc)

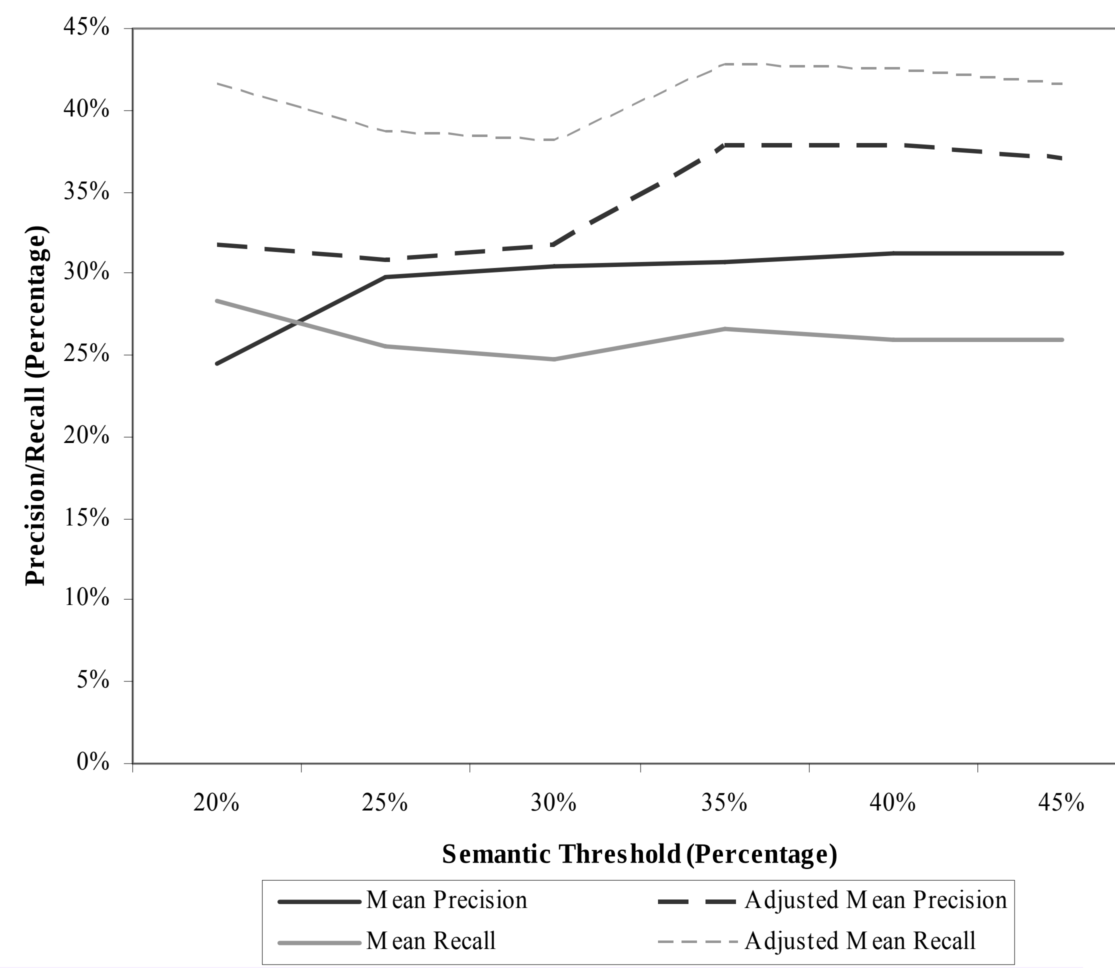
Aberration	Big Bang	Dark Matter
Accretion	Biosphere	Desert
Albedo	Black hole	Diety
Altitude	Blizzard	Doppler Effect
Andromeda	Bursters	Doppler Shift
Antimatter	Cenozoic	Earth
Aphelion	Ceres	Earth's crust
Apogee	Chemical Element	Eclipse
Apoapsis	Chinese Deity	Ecology
Aquatic Mammal	Chromosphere	Ecosystem
Asteroid	Coma	Epoch
Astrobiology	Comet	Equinox
Astrometry	Constellation	Eruption
Astronaut	Corona	Evolution
Astronomer	Coronagraph	Facula
Astronomy	Craters	Filament
Astrophysics	Crustacean	Fireball
Atmosphere	Cryosphere	Flare
Aurora	Culmination	Forecast

Use URLs to prune Wikipedia concepts to those that are relevant to user sessions context (in the usage logs)

Accretion	Fossil	Mars
Astronaut	Fusion	Mercury
Atmosphere	Geosphere	Mineral
Aurora	Granulation	Moon
Biosphere	Halo	Occultation
Black hole	Heliosphere	Phase
Crustacean	Ionosphere	Planet
Culmination	Island	Polarization
Earth	Jupiter	Rain Forest
Eclipse	Light	Satellite
Epoch	Lithosphere	Scientist
Eruption		Star
		Weather
		Zodiac

Map user sessions → term sets (content), → concept sessions

- Find most semantically related concept for each term
  - either the exactly matched concept
  - or a more general concept.
- Use the concept hierarchy in WordNet's taxonomy
  - calculate a path-based measure between term-concept pairs
  - IF Sim < threshold Then unrelated
- Evaluation: Compare automatically extracted concepts in 100 sessions with those assigned by Human evaluator (ground truth) using prevision/recall



P. Achananuparp, H. Han, O. Nasraoui and R. Johnson, Semantically Enhanced User Modeling, ACM SAC 2007.

Summary of Talk: Challenges & Proposed Solutions in Web Usage Mining & Personalization

- Mining Web Clickstreams ➔ User Profiles / User Models
  - Semantics for disambiguation:
    - Implicitly derived (e.g. from website hierarchy)
    - Explicit (e.g. from related Databases that describe a hierarchy of the items/web pages)
    - Content semantics ➔ Conceptual user model
    - Noise ➔ Robust profiles
  - Scalability: how to scale to massive data streams?
    - need to process data in one pass to mine continuously evolving user profiles + work under very stringent constraints
  - Evolution: Track profiles over periods, Define profile evolution events
- Recommender Systems (that use the user profiles/models discovered above)
  - Evolution: Validate continuously mined evolving user profiles against evolution scenarios?
  - Implementation: fast, easy, scalable, cheap, free (use existing open source indexing+ search engine software)

REFERENCES
IN WEB USAGE MINING & PERSONALIZATION

- [1] M. Perkowitz and O. Etzioni. Adaptive web sites: Automatically learning for user access pattern. Proc. 6th int. WWW conference, 1997.
- [2] R. Cooley, B. Mobasher, and J. Srivastava. Web Mining: Information and Pattern discovery on the World Wide Web, Proc. IEEE Intl. Conf. Tools with AI, Newport Beach, CA, pp. 558-567, 1997.
- [3] O. Nasraoui and R. Krishnapuram, and A. Joshi. Mining Web Access Logs Using a Relational Clustering Algorithm Based on a Robust Estimator, 8th International World Wide Web Conference, Toronto, pp. 40-41, 1999.
- [4] O. Nasraoui, R. Krishnapuram, H. Frigui, and A. Joshi. Extracting Web User Profiles Using Relational Competitive Fuzzy Clustering, International Journal on Artificial Intelligence Tools, Vol. 9, No. 4, pp. 509-526, 2000.
- [5] O. Nasraoui, and R. Krishnapuram. A Novel Approach to Unsupervised Robust Clustering using Genetic Niching, Proc. of the 9th IEEE International Conf. on Fuzzy Systems, San Antonio, TX, May 2000, pp. 170-175.
- [6] O. Nasraoui and R. Krishnapuram. A New Evolutionary Approach to Web Usage and Context Sensitive Associations Mining, International Journal on Computational Intelligence and Applications - Special Issue on Internet Intelligent Systems, Vol. 2, No. 3, pp. 339-348, Sep. 2002.
- [7] M. Pazzani and D. Billsus. Learning and revising User Profiles: The identification of Interesting Web Sites, Machine Learning, 27:313—331, 1997.

- [8] Levene, M., Borges, J., and Loizou, G. Zipf's law for Web surfers. *Knowl. Inf. Syst.* 3, 1 (Feb. 2001), 120-129.
- [9] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. Effective personalizaton based on association rule discovery from Web usage data, ACM Workshop on Web information and data management, Atlanta, GA, Nov. 2001.
- [10] J. H. Holland. Adaptation in natural and artificial systems. MIT Press, 1975.
- [13] R. Agrawal and R. Srikant. Fast algorithms for mining association rules, Proceedings of the 20th VLDB Conference, Santiago, Chile, 1994, pp. 487-499.
- [14] G. Linden, B. Smith, and J. York. Amazon.com Recommendations Item-to-item collaborative filtering, IEEE Internet Computing, Vo. 7, No. 1, pp. 76-8
- [15] J. Breese, H. Heckerman, and C. Kadie. Empirical Analysis of Predictive Algorithms for Collaborative Filtering, Proc. 14th Conf. Uncertainty in Artificial Intelligence, pp. 43-52, 1998.
- [16] J.B. Schafer, J. Konstan, and J. Reidel. Recommender Systems in E-Commerce, Proc. ACM Conf. E-commerce, pp. 158-166, 1999.
- [17] J. Srivastava, R. Cooley, M. Deshpande. and P-N Tan, Web usage mining: Discovery and applications of usage patterns from web data, SIGKDD Explorations, Vol. 1, No. 2, Jan 2000, pp. 1-12.

- [18] O. Zaiane, M. Xin, and J. Han. Discovering web access patterns and trends by applying OLAP and data mining technology on web logs, in "Advances in Digital Libraries", 1998, Santa Barbara, CA, pp. 19-29.
- [19] M. Spiliopoulou and L. C. Faulstich. WUM: A Web utilization Miner, in Proceedings of EDBT workshop WebDB98, Valencia, Spain, 1999.
- [20] J. Borges and M. Levene, Data Mining of User Navigation Patterns, in "Web Usage Analysis and User Profiling", Lecture Notes in Computer Science", H. A. Abbass, R. A. Sarker, and C.S. Newton Eds., Springer-Verlag,1999 , pp. 92-111.
- [21] J. R. Quinlan. Induction of Decision Trees. Machine Learning, Vol. 1, pp. 81--106, 1986.
- [22] O. Nasraoui, C. Cardona, C. Rojas, and F. Gonzalez. Mining Evolving User Profiles in Noisy Web Clickstream Data with a Scalable Immune System Clustering Algorithm, in Proc. of WebKDD 2003, Washington DC, Aug. 2003, 71-81.
- [23] G. Adomavicius, A. Tuzhilin. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Trans. Knowl. Data Eng. 17(6): 734-749, 2005.
- [24] M. Pazzani. A Framework for Collaborative, Content-Based and Demographic Filtering, AI Review, 13(5-6):393-408, 1999.
- [25] M. Balabanovic and Y. Shoham. Fab: Content-based, Collaborative Recommendation, Communications of the ACM 40(3): 67-72, March 1997.

- [26] B. Berendt, A. Hotho, and G. Stumme. Towards semantic web mining. In Proc. International Semantic Web Conference (ISWC02), 2002.
- [27] R. Burke. Hybrid recommmender systems: Survey and experiments. In User Modeling and User-Adapted Interaction, 12(4): 331-370,2002.
- [28] D. Oberle, B. Berendt, A. Hotho, and J. Gonzalez. Conceptual User Tracking, in Proc. of the Atlantic Web Intelligence Conference (AWIC) Madrid, Spain, 2003.
- [29] H. Dai and B. Mobasher. Using ontologies to discover domain-level web usage profiles. In Proc. 2nd Semantic Web Mining Workshop at ECML/PKDD-2002.
- [30] M. Eirinaki, H. Lampos, M. Vazirgiannis, I. Varlamis. SEWeP: Using Site Semantics and a Taxonomy to Enhance the Web Personalization Process, in the Proc. of SIGKDD ’03, Washington DC, USA, August 2003.
- [31] P. Van der Putten, J. N. Kok and A. Gupta. Why the Information Explosion Can Be Bad for Data Mining and How Data Fusion Provides a Way Out, In Proc. of the 2nd SIAM International Conference on Data Mining, 2002.
- [32] Miller, G. A. WORDNET: An On-Line Lexical Database, Int. Journal of Lexicography 3-4:235-312, 1990.
- [33] B. Mobasher, H. Dai, T. Luo, Y. Sung, J. Zhu, Integrating Web Usage and Content Mining for More Effective Personalization, in Proc. of the International Conference on E-Commerce and Web Technologies (ECWeb2000), Greenwich, UK, September 2000.

- [34] R. Srikant, R. Agrawal, Mining Generalized Association Rules, in Proc. of 21st VLDB Conf., Zurich, Switzerland, September 1995.
- [35] S. Chakrabarti, B. Dom, R. Agrawal, P. Raghavan, Using taxonomy, discriminants and signatures for navigation in text databases, in Proc. of the 23rd VLDB Conference Athens, Greece, 1997.
- [36] B. Berendt, Understanding Web usage at different levels of abstraction: coarsenin and visualizing sequences, in Proc. of the Mining Log Data Across All Customer TouchPoints Workshop (WEBKDD'01), San Francisco, CA, August 2001
- [37] Desikan P. and Srivastava J., Mining Temporally Evolving Graphs. In Proceedings of "*WebKDD- 2004 workshop on Web Mining and Web Usage Analysis*", B. Mobasher B. Liu, B. Masand, O. Nasraoui, Eds. part of the ACM KDD: Knowledge Discovery and Data Mining Conference, Seattle, WA, 2004.
- [38] Eirinaki M., Vazirgiannis M. Web mining for web personalization. *ACM Transactions On Internet Technology (TOIT)*, 3(1), 1-27, 2003.
- [39] Joachims T., Optimizing search engines using clickthrough data. In *Proc. of the 8t ACM SIGKDD Conference*, 133-142, 2002.
- [40] Nasraoui O., Krishnapuram R., Joshi A., and Kamdar T., Automatic Web User Profiling and Personalization using Robust Fuzzy Relational Clustering, in "*E-Commerce and Intelligent Methods" in the series "Studies in Fuzziness and Soft Computing*", J. Segovia, P. Szczepaniak, and M. Niedzwiedzinski, Ed, Springer-Verlag 2002.

- [41] L. Terveen, W. Hill, B. Amento, D. McDonald, and J. Creter", PHOAKS: A System for Sharing Recommendations", Communications of the ACM, 40(3), 59-62, 1997.
- [42] Balabanovic, M., An Adaptive Web Page Recommendation Service. *First International Conference on Autonomous Agents,* Marina del Rey, CA, 378-385, 1997.
- [43] Konstan J.A., Miller B., Maltz, Herlocker J., Gordon and Riedl J.. GroupLens: Collaborative Filtering for Usenet News. Communications of the ACM, March, p. 77-87, 1997.
- [44] Sarwar, B. M., Konstan, J. A., Borchers, A., Herlocker, J., Miller, B., and Riedl, J. 1998. Using filtering agents to improve prediction quality in the GroupLens research collaborative filtering system. In *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work* , Seattle, Washington, 1998, 345-354
- [45] CDNow.com – http://www.cdnow.com
- [46] T. Yan, M. Jacobsen, H. Garcia-Molina, and U. Dayal. From user access patterns to dynamic hypertext linking. In *Proceedings of the 5th International World Wide Web conference*, Paris, France, 1996.
- [47] C. Shahabi, A. M. Zarkesh, J. Abidi, and V. Shah. Knowledge discovery from users web-page navigation. In *Proceedings of workshop on Research Issues in Data Engineering*, Birmingham, England, 1997.

- [48] O. Nasraoui, C. Rojas, and C. Cardona, A Framework for Mining Evolving Trends in Web Data Streams using Dynamic Learning and Retrospective Validation, in Computer Networks, Special Issue on Web Dynamics, 50(14), Oct., 2006.
- [49] M. D. Mulvenna, S. S. Anand, A. G. Büchner: Personalization on the Net using Web mining: introduction. Commun. ACM 43(8): 122-125 (2000).
- [50] Ganesan, P., Garcia-Molina, H., and Widom, J. 2003. Exploiting hierarchical domain structure to compute similarity. *ACM Trans. Inf. Syst.* 21, 1 (Jan. 2003), 64-93.
- [51] Armstrong, R., Freitag, D., Joachims, T., and Mitchell, T., *WebWatcher: A Learning Apprentice for the World Wide Web*. Proceedings of the 1995 AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, 1995.
- [52] Olfa Nasraoui, Maha Soliman, and Antonio Badia, Mining Evolving User Profiles and More – A Real Life Case Study, In Proc. Data Mining meets Marketing workshop, New York, NY, 2005.
- [53] P. Achananuparp, H. Han, O. Nasraoui and R. Johnson, Semantically Enhanced User Modeling, ACM SAC 2007, Seoul, Korea.
- [54] E. Saka and O. Nasraoui, Effect of Conceptual Abstraction and URL Compression on the Quality of Web Usage Mining, Knowledge Discovery & Web Mining Lab Tech. report No 2006-12-1, University of Louisville, Dec. 2006.
- [55] O. Nasraoui, C. Cardona, C. Rojas, F. González, “TECNO-STREAMS: Tracking Evolving Clusters in Noisy Data Streams with a Scalable Immune System Learning Model”, in Proc. of Third IEEE International Conference on Data Mining (ICDM'03), Melbourne, FL, November 2003, pp. 235-242.
- [56] Nasraoui O., Petenes C., "Combining Web Usage Mining and Fuzzy Inference for Website Personalization", in Proc. of WebKDD 2003 – KDD Workshop on Web mining as a Premise to Effective and Intelligent Web Applications, Washington DC, August 2003, p. 37-48.

- [57] O. Nasraoui, J. Cerwinske, C. Rojas, and F. Gonzalez, Collaborative Filtering in Dynamic Usage Environments, in Proc. Conference on Information and Knowledge Management: CIKM, Arlington, VA, Nov. 2006.
- [58] O. Nasraoui, Z. Zhang, and E. Saka, “Web Recommender System Implementations in Multiple Flavors: Fast and (Care) Free for All”. In Proceedings of the ACM-SIGIR Open Source Information Retrieval workshop, Seattle, WA, Aug. 2006.
- [59] Nasraoui O. and Goswami S., “Mining and Validating Localized Frequent Itemsets with Dynamic Tolerance,” in Proc. SIAM conference on Data Mining, Bethesda, MD, Apr. 2006.
- [60] O. Nasraoui, C. Cardona, and C. Rojas. “Using Retrieval Measures to Assess Similarity in Mining Dynamic Web Clickstreams”. In Proceedings of ACM KDD: Knowledge Discovery and Data Mining Conference, Chicago, IL, 2005, 439-448.
- [61] O. Nasraoui and M. Pavuluri, “Complete this Puzzle: A Connectionist Approach to Accurate Web Recommendations based on a Committee of Predictors”. In Proceedings of “*WebKDD- 2004 workshop on Web Mining and Web Usage Analysis*”, B. Mobasher, B. Liu, B. Masand, O. Nasraoui, Eds. part of the ACM KDD: Knowledge Discovery and Data Mining Conference, Seattle, WA, 2004.
- [62] O. Nasraoui, C. Cardona, and C. Rojas. “Mining of Evolving Web Clickstreams with Explicit Retrieval Similarity Measures”. In Proceedings of “International Web Dynamics Workshop”, International World Wide Web Conference, New York, NY, May. 2004.

- [63] Mitchell T., Caruana R., Freitag D., McDermott, J. and Zabowski D. Experience with a Learning Personal Assistant. Communications of the ACM 37(7), 1994, pp. 81-91.
- [64] Maloof M. and Michalski R. Selecting examples for partial memory learning. Machine Learning, 41(11),2000, pp. 27-52.
- [65] Schlimmer J., and Granger R. Incremental Learning from Noisy Data, Machine Learning, 1(3), 1986, 317-357.
- [66] SchwabI., Pohl W. and KoychevI.Learning to Recommend from Positive Evidence, Proceedings of Intelligent User Interfaces 2000, ACM Press, 241 - 247.
- [67] Widmer G. Tracking Changes through Meta-Learning, Machine Learning 27, 1997, pp. 256-286.
- [68] Widmer G. and Kubat M. Learning in the presence of concept drift and hidden contexts. Machine Learning 23, 1996, pp. 69-101.

Thank You!

- Any questions?