FIU FLORIDA INTERNATIONAL UNIVERSITY

ELVES research laboratory

dpSmart: a Flexible Group based Recommendation Framework for Digital Repository Systems

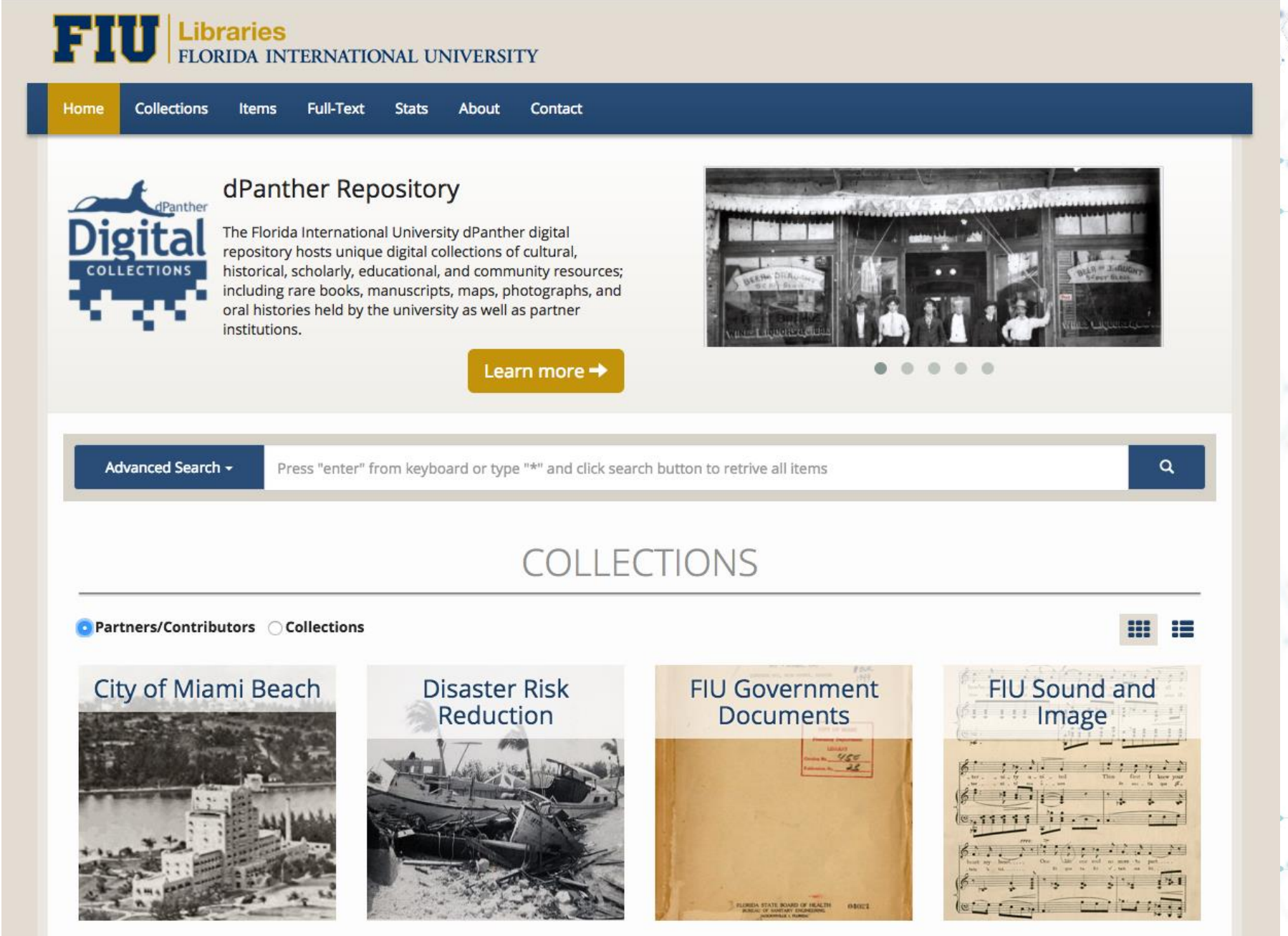
Boyuan Guan, Liting Hu, Pinchao Liu, Hailu Xu, Zhaohui Fu, Qingyang Wang†

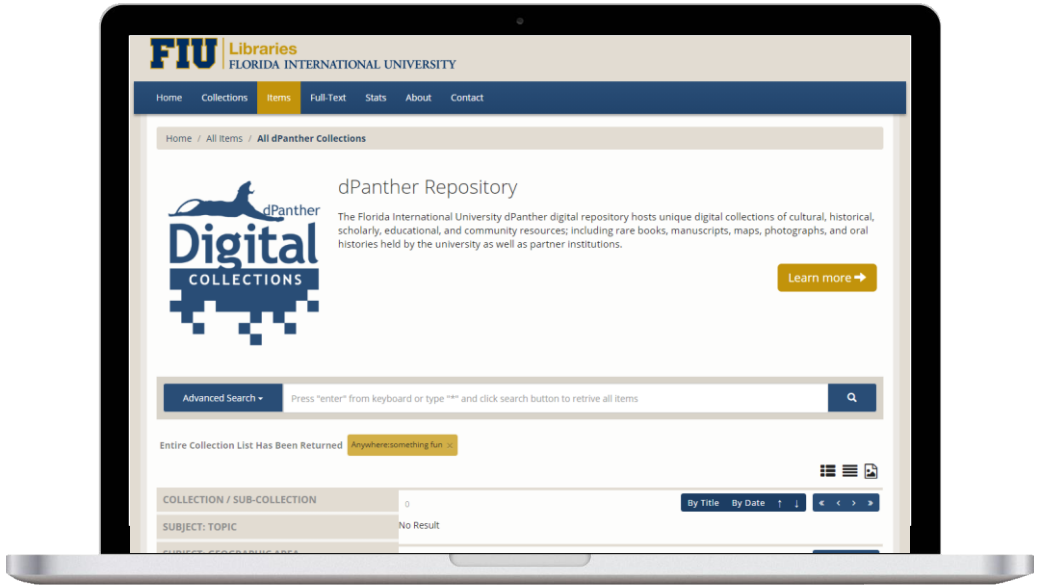
Florida International University
Louisiana State University†

OUTLINE

1. Introduction
2. Related Work
3. System Design
4. Evaluation
5. Conclusion & Questions?

INTRODUCTION


FIU Libraries
FLORIDA INTERNATIONAL UNIVERSITY
Home
Collections
Items
Full-Text
Stats
About
Contact
dPanther Digital COLLECTIONS
dPanther Repository
The Florida International University dPanther digital repository hosts unique digital collections of cultural, historical, scholarly, educational, and community resources; including rare books, manuscripts, maps, photographs, and oral histories held by the university as well as partner institutions.
Learn more
Advanced Search
Press "enter" from keyboard or type "*" and click search button to retrive all items
COLLECTIONS
Partners/Contributors
Collections
City of Miami Beach
Disaster Risk Reduction
FIU Government Documents
FIU Sound and Image

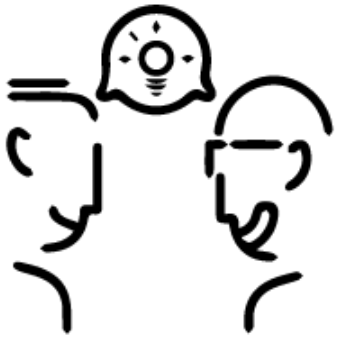


MAIN PROBLEMS

- **Limited Discoverability**
  - 12,376 (24.9%) out of the total 49,737
- **Lack of assistance to explore the system**
  - 700 clients out of 1,127 with a specific record when conducting search
- **Drop off visits**
  - 978,814 visits (58.8%) out of total **1,664,813** visits

General Problems for Digital Library Domain

What do the users WANT to know?
What do the users NEED to know?



Get to know the users & their region information

Awareness about the content

Awareness about the system condition (usages, statistic, user interactions)

Data Mining !! → **Recommendation System**

Challenges

Limited user information

Multiple Recommenders available

Side Effect (Noise)

Impact on the system performance

Intelligent Recommendation System (**dpsmart**)

Building User vectors from web log data only

Allowing multiple recommenders work collaboratively

Grouping Users from stereotyping model

Optimizing the algorithm with multi-process programming

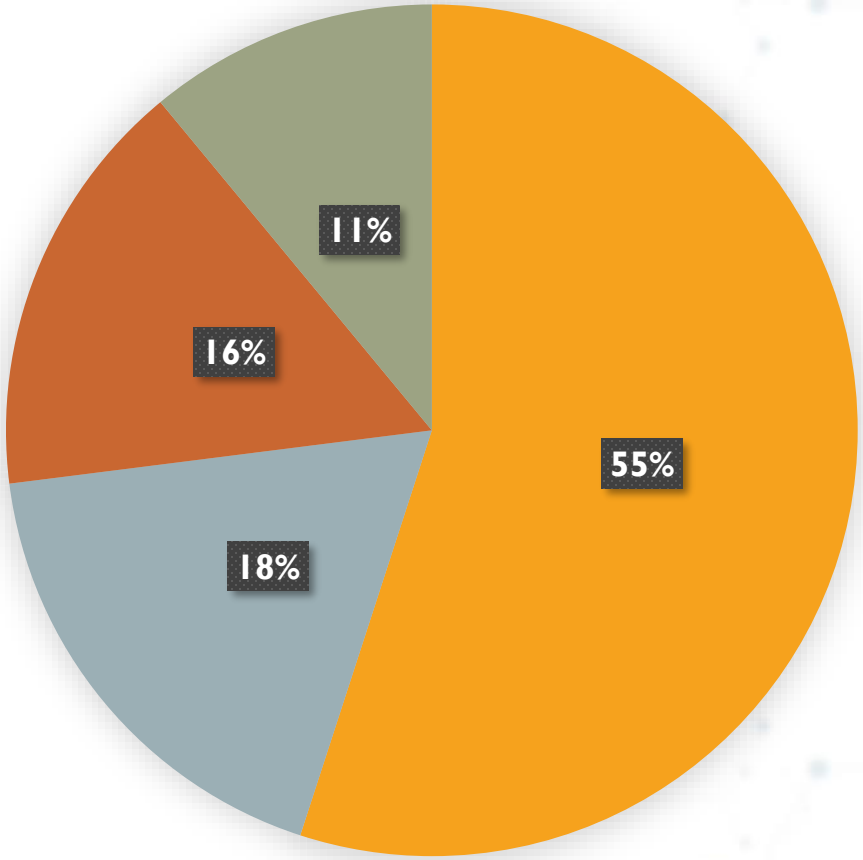
THE TECHNICAL CONTRIBUTIONS

- **Implementing a stereotype based recommendation system in real world system**
- **Avoiding the intensive labor works and automating the process from the log data extraction to model training.**
- **Avoiding the need of personal identifiable information and reducing the noise recommendation.**
- **Improving the system performing by using multi-process programming.**

RELATED WORKS

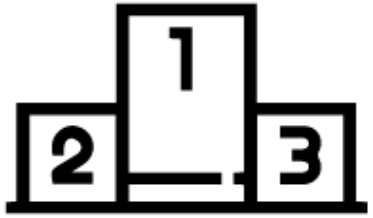
RECOMMENDERS

- **200** research articles were published
- **62** methods proposed



ADDITIONAL METHODS

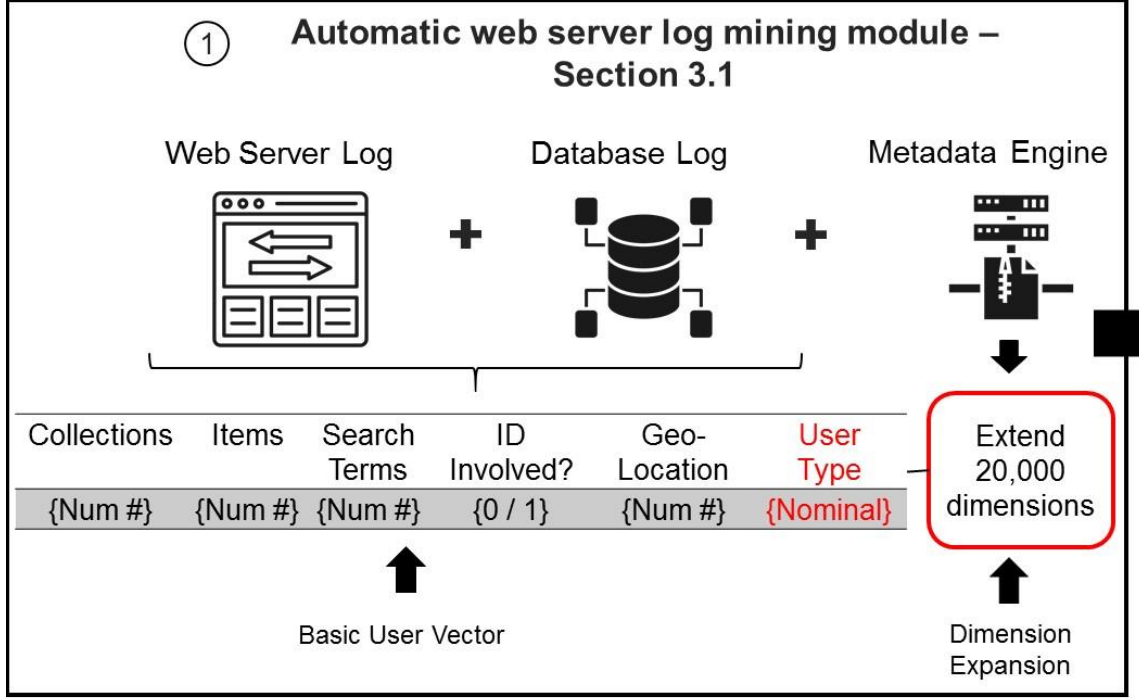
GLOBAL RELEVANCE

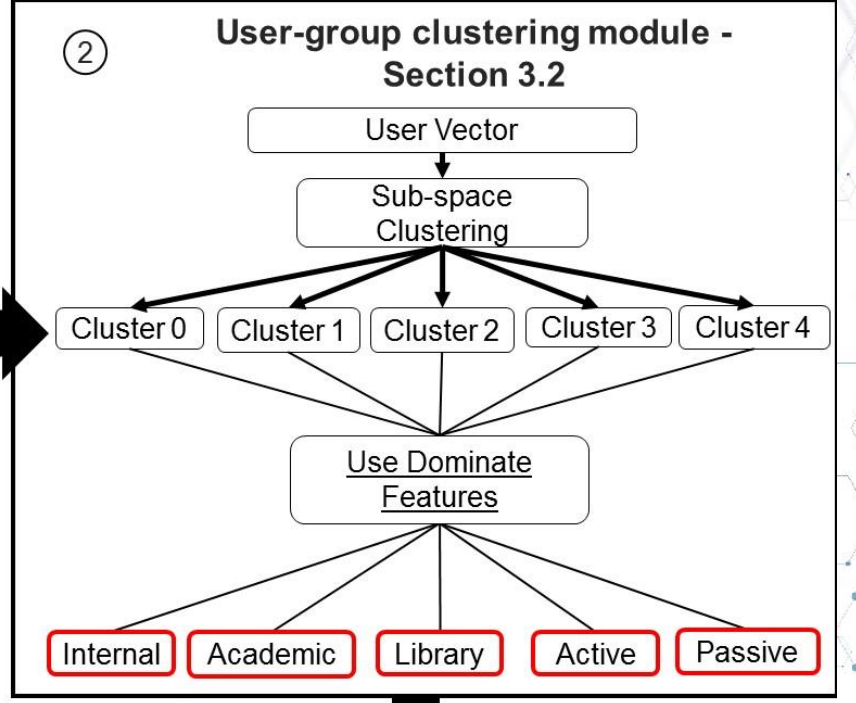


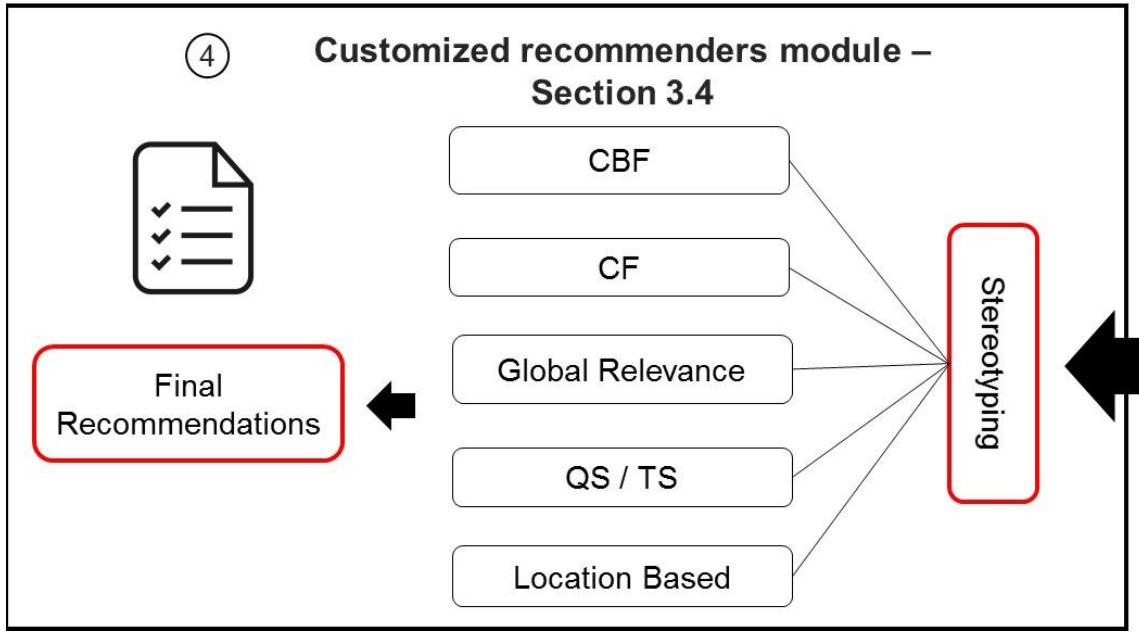
LOCATION BASED

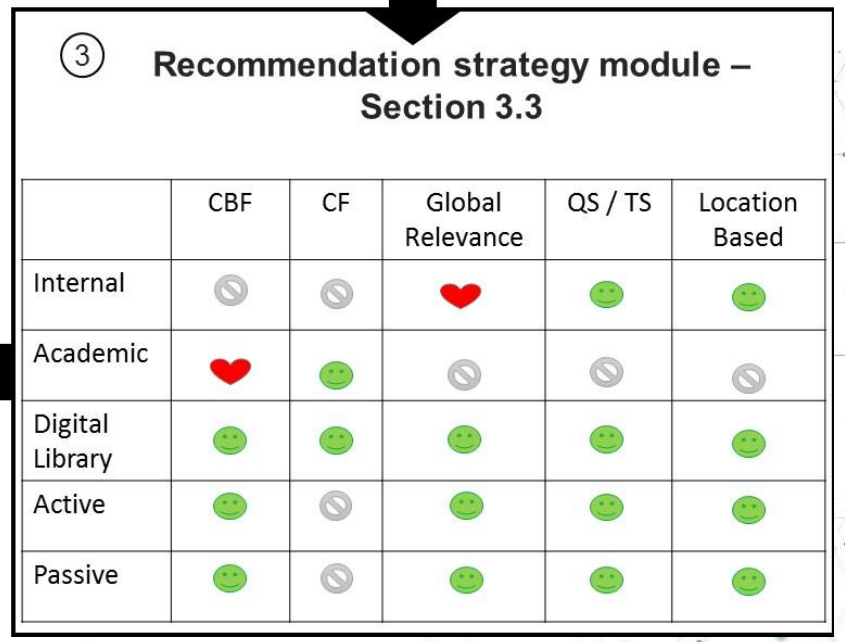
QUERY/TERM SUGGESTION (QS / TS)

SYSTEM DESIGN





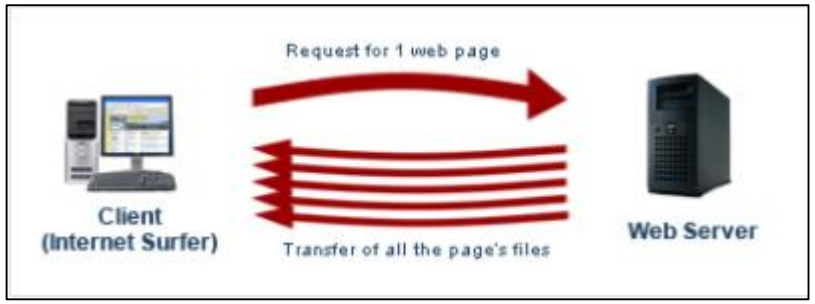




	CBF	CF	Global Relevance	QS / TS	Location Based
Internal	⊘	⊘	❤	☺	☺
Academic	❤	☺	⊘	⊘	⊘
Digital Library	☺	☺	☺	☺	☺
Active	☺	⊘	☺	☺	☺
Passive	☺	⊘	☺	☺	☺

AUTOMATIC WEB SERVER LOG MINING MODULE

Web Server Log



Database Log

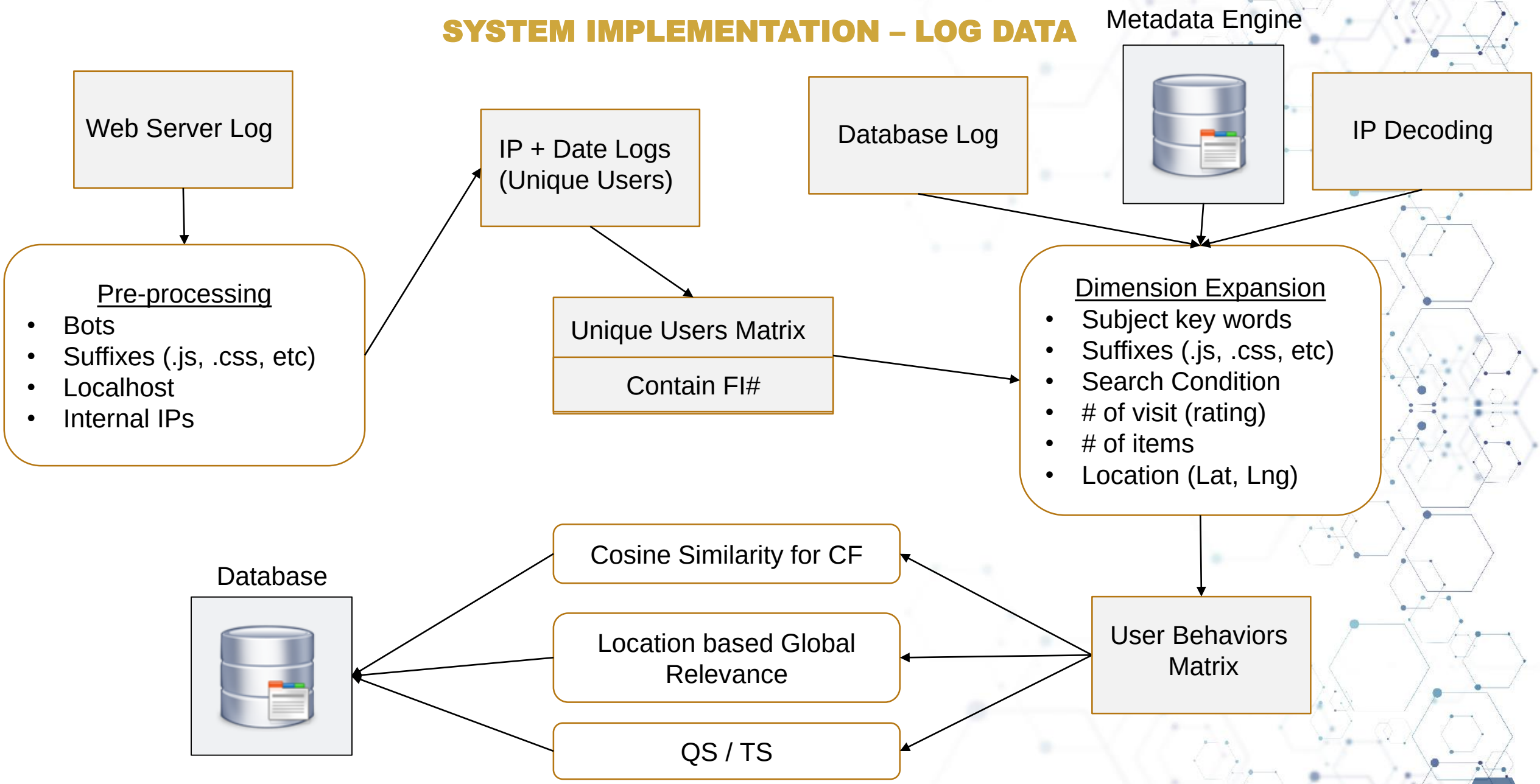
Metadata Engine

Users ➔

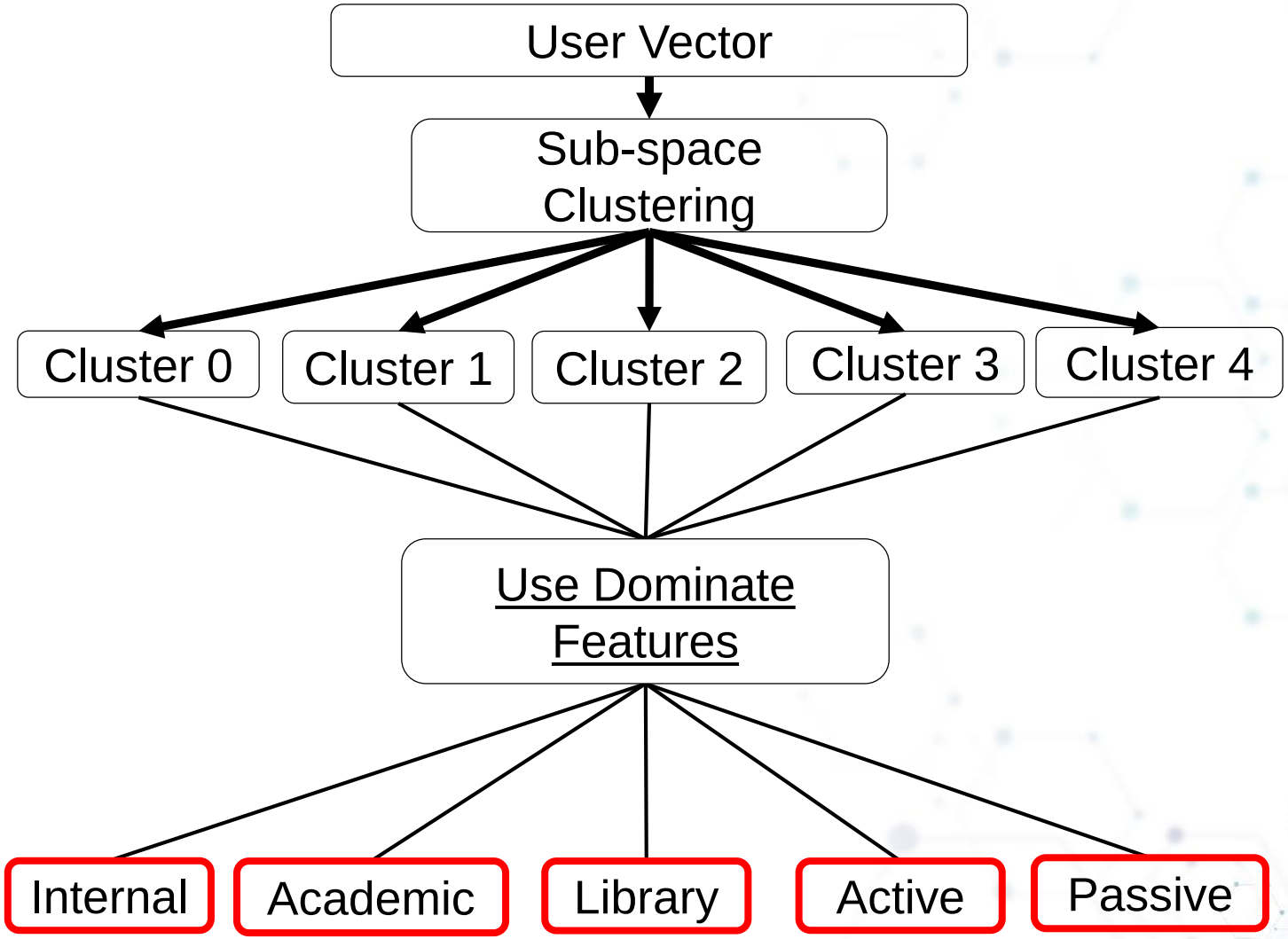
Collections	Items	Search Terms	ID Involved?	Geo-Location	User Type
{number}	{number}	{number}	{True / False}	{Number}	{Nominal}

By linking the records to the metadata content will extend 20,000 dimensions

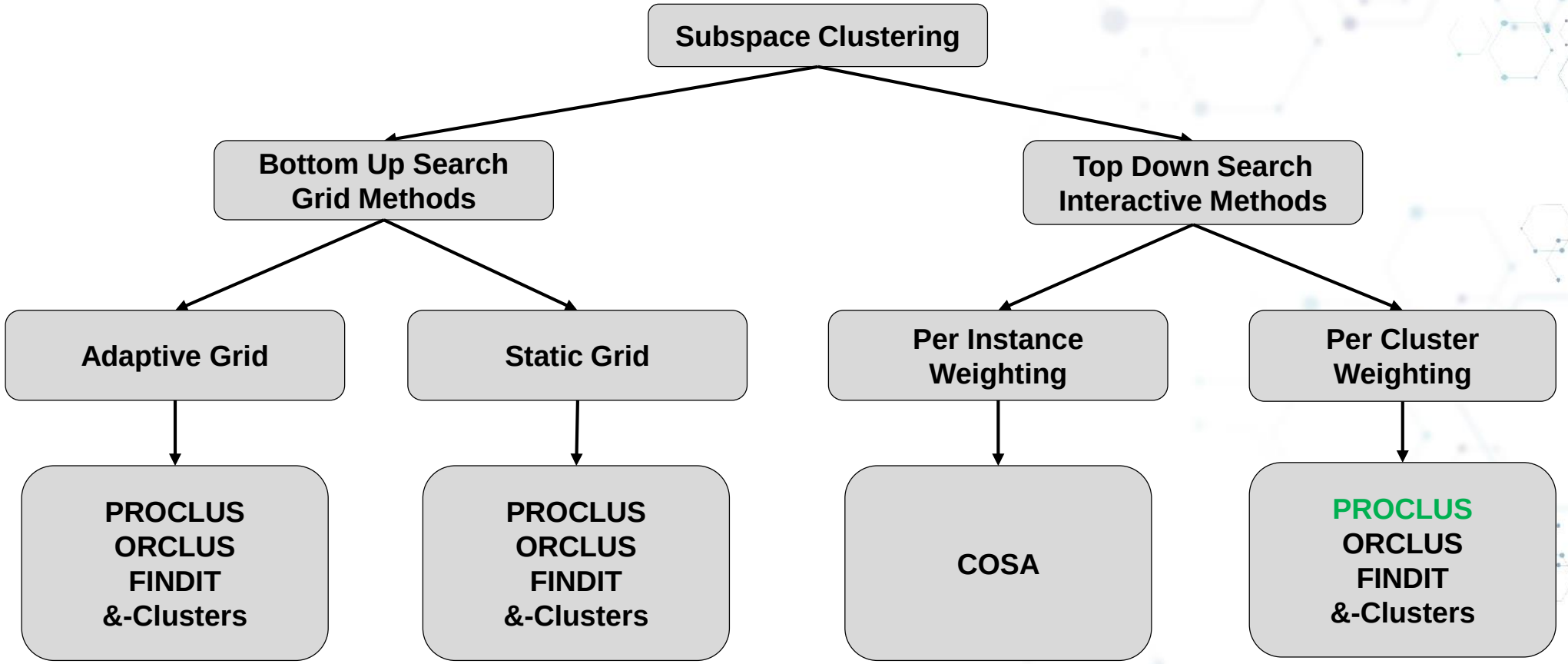
SYSTEM IMPLEMENTATION – LOG DATA



User-group clustering module - Section 3.2

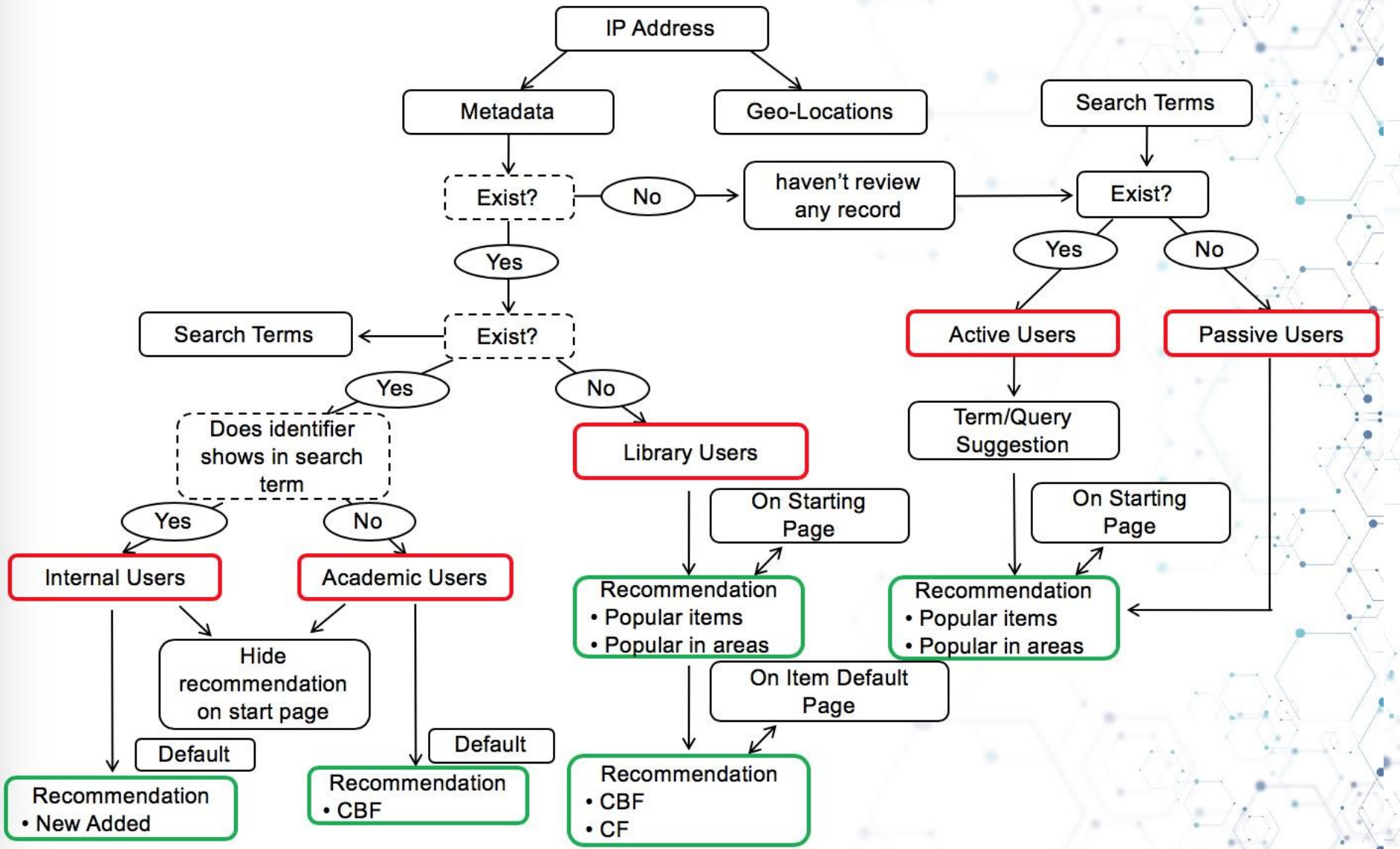


Subspace clustering



WEKA The University of Waikato + Opensubspace v3.31

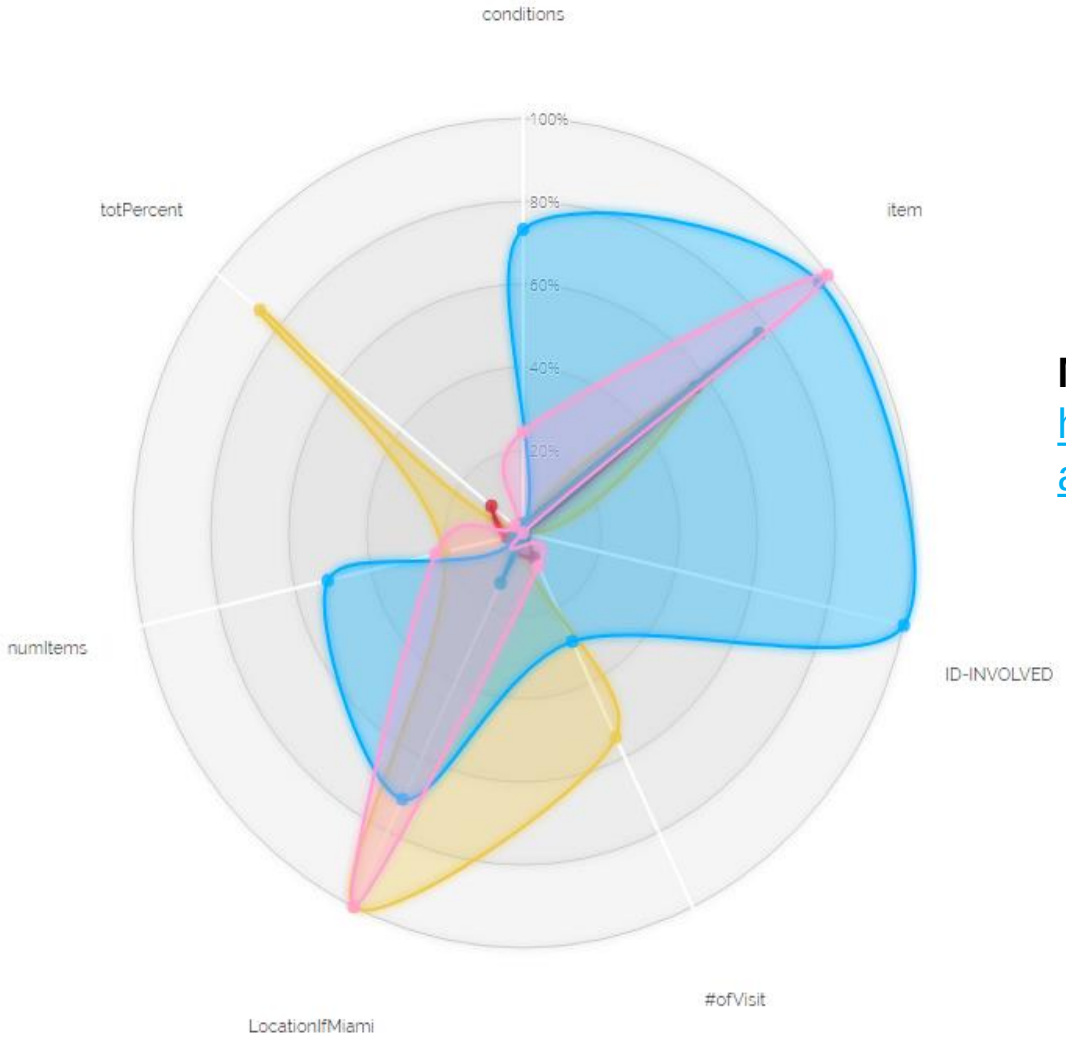
USE DOMINATE FEATUERES TO MARK USER GROUPS



SAMPLE USER RECOGNITION BY USING DOMINATE FEATURES

Basic User Vector: searches,item,idInvolved,numVisits,LocationIfMiami,numItems

- Cluster0:21238 (Active Web Users)
- Cluster1:2565 (Passive Web Users)
- Cluster2:97 (Digital Library Users)
- Cluster3:66 (Internal Users)
- Cluster4:667 (Acadamic Users)



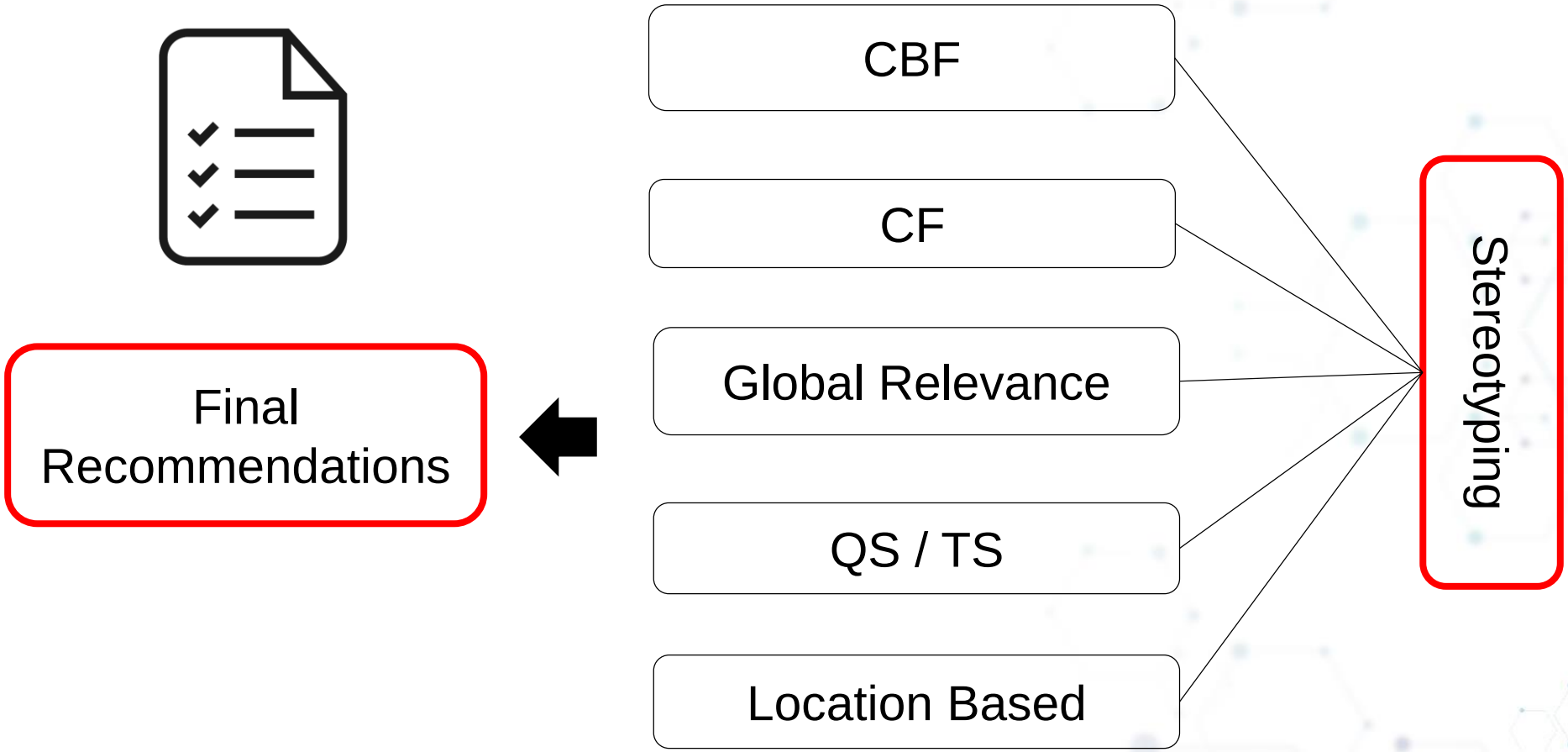
More available here:
http://dpanther.fiu.edu/dpanther/dpsmart/usercluster/

RECOMMENDATION STRATEGY

	CBF	CF	Global Relevance	QS / TS	Location Based
Internal Users	⊘	⊘	❤	☺	☺
Academic Users	❤	☺	⊘	⊘	⊘
Digital Library Users	☺	☺	☺	☺	☺
Active Users	☺	⊘	☺	☺	☺
Passive Users	☺	⊘	☺	☺	☺

4

Customized recommenders module – Section 3.4



CUSTOMIZED RECOMMENDERS

1.Content based filter: calculate the item-similarity among one specific user to find a threshold to provide CB recommendation

$$SW_{(a,b)} = \frac{|\sum_{n=1}^{\infty}((A_n \cap B_n) * W_n)|}{|\sum_{n=1}^{\infty}((A_n \cup B_n) * W_n)|} \quad (2)$$

- Where SW is the weighted Jaccard Similarity
- a, b are any two items in dPanther system
- A,B are any pair metadata field from n selection metadata data fields between item a and b
- w is the weight assigned for a specific pair of one metadata field

2.Collaborative filtering: Calculate user similarity for current user to find if there is a similar user group. Calculation based on Cosine Similarity

$$S_{(i,j)} = \frac{\sum_{u \in U}(R_{(u,i)} - \overline{R}_u)(R_{(u,j)} - \overline{R}_u)}{\sqrt[2]{\sum_{u \in U}((R_{(u,i)} - \overline{R}_u)^2)}\sqrt[2]{\sum_{u \in U}((R_{(u,j)} - \overline{R}_u)^2)}} \quad (4)$$

- Where S is the cosine-based similarity
- U stands for the group of users who rate both item i and item j
- $R_{(u,i)} or R_{(u,j)}$ means the score for such the item by the user
- $\overline{R}_u$ is the average of the u-th userâĂŹs ratings

Very Slow!!!

CUSTOMIZED RECOMMENDERS

3. Global Relevance (GR): Frequency of subject key works appearance in high occurred words plus the location information

$$GR_u \propto N_i \quad where \quad L_i \in L_u \tag{6}$$

- Where GR is the Global Reference Score
- u is the current user
- i is the item in dPanther system
- N is the number of hits of item i
- L is the location information

4. Term Suggestion (TS) / Query Suggestion (QS): The similarity of the term to the metadata is generated by using the Jaccard index as suggestion by:

$$S_{(a,b)} = \frac{|DS_a \cap DS_b|}{|DS_a \cup DS_b|} \tag{7}$$

- Where S is the Similarity Score between the search term and the item subject keywords
- a is the input search term
- b is the item in the system
- DS is the data set consist of the key works

Very Slow!!!

EVALUATION

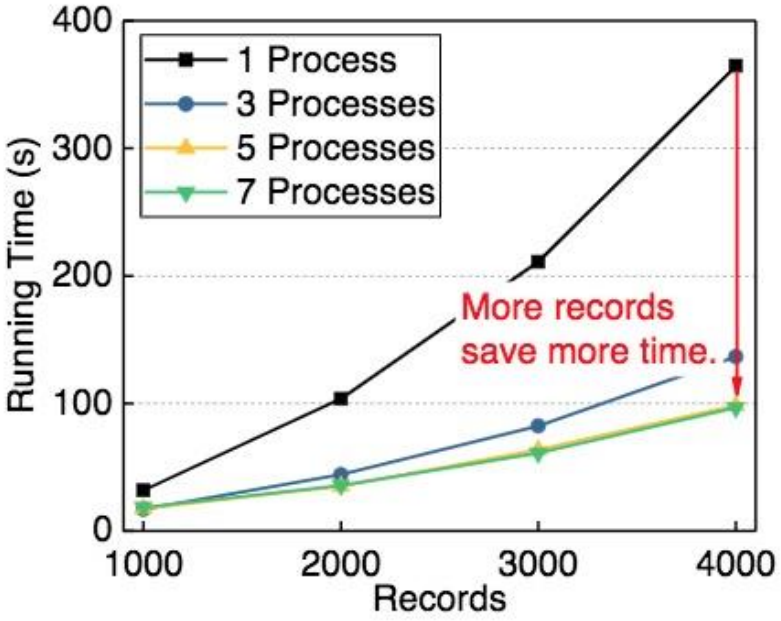
PURPOSE OF EVALUATION

- Whether the multiple-process programming algorithm improves the hosting Digital Repository Systems performance?
- When dpsmart integrated into the dPanther, what are the benefits of dpSmart regarding the **page views, bounce rate & drop-off rate**?
- How are the dPanther usability like **Avg. Time on Page, Avg. Page loading, Avg. Direction Time, Avg. Server Reponses Time, and Avg. Page Overload Time** improved by integrating dpSmart?

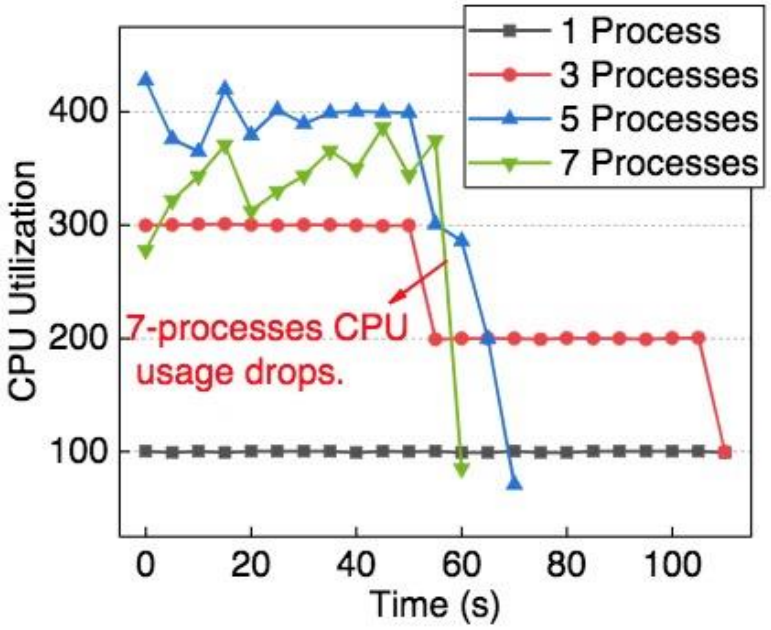
SYSTEM PERFORMANCE EVALUATION

Windows Machine
4 Cores of CPU
32 GB Memory

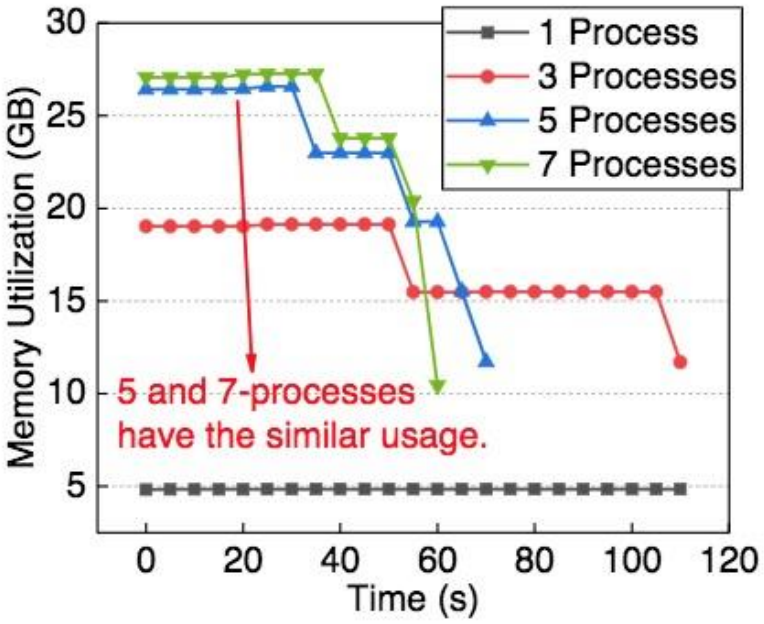
Randomly Selected 4,000 records to run the CBF with 1-process, 3-processes, 5-processes, and 7-processes respectively



(a) The Running Time Comparison.



(b) The CPU Usage.



(c) The Memory Usage.

SYSTEM PERFORMANCE EVALUATION

Windows Machine
4 Cores of CPU
32 GB Memory

Randomly Selected 4,000 records to run the CBF with 1-process, 3-processes, 5-processes, and 7-processes respectively

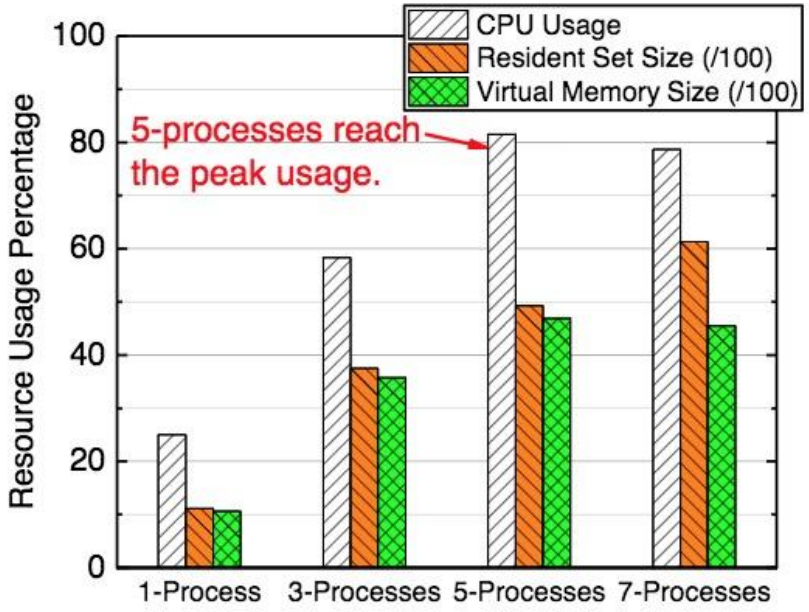


Fig. 4: The Average Usage of CPU, Memory and Virtual Memory by using 1-process, 3-processes, 5-processes, and 7-processes for the task of running 4000 records.

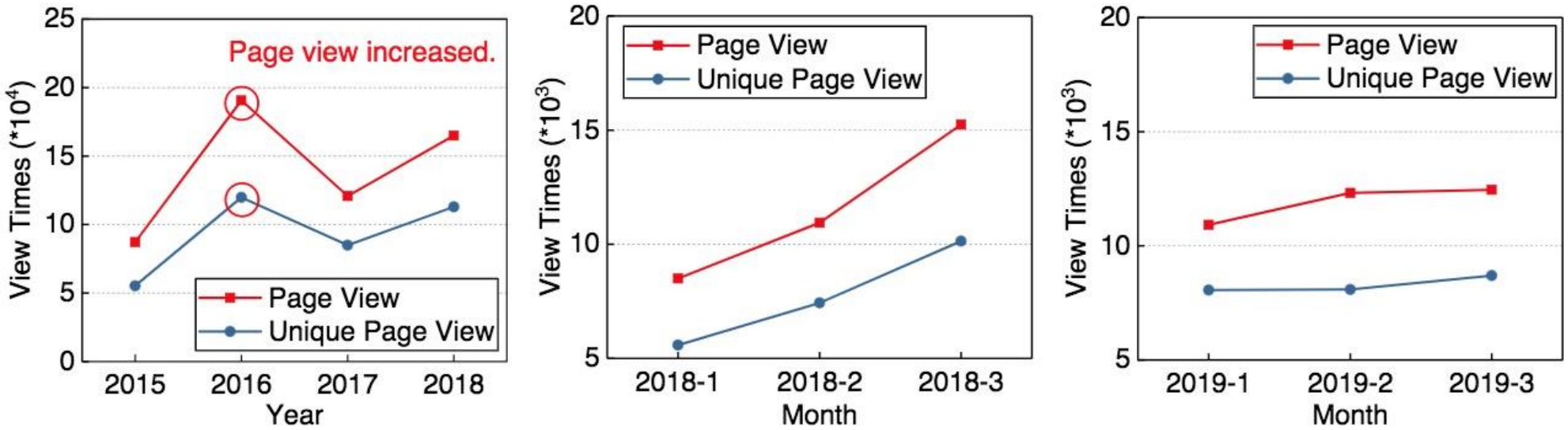
SYSTEM IMPACT EVALUATION

Experiment Time Line:

January, 2016
First module, customized recommender, implement

2018: Model rebuilding multiple times

January 2019: Entire framework implement



(a) The Page View Stats for the year 2015-2018. (b) The Page View Stats from Jan. to Mar., 2018. (c) The Page View Stats from Jan. to Mar., 2019

Fig. 5: The Page View Stats for the year from 2015 to 2018, from January to March in year 2018, and from January to March in year 2019.

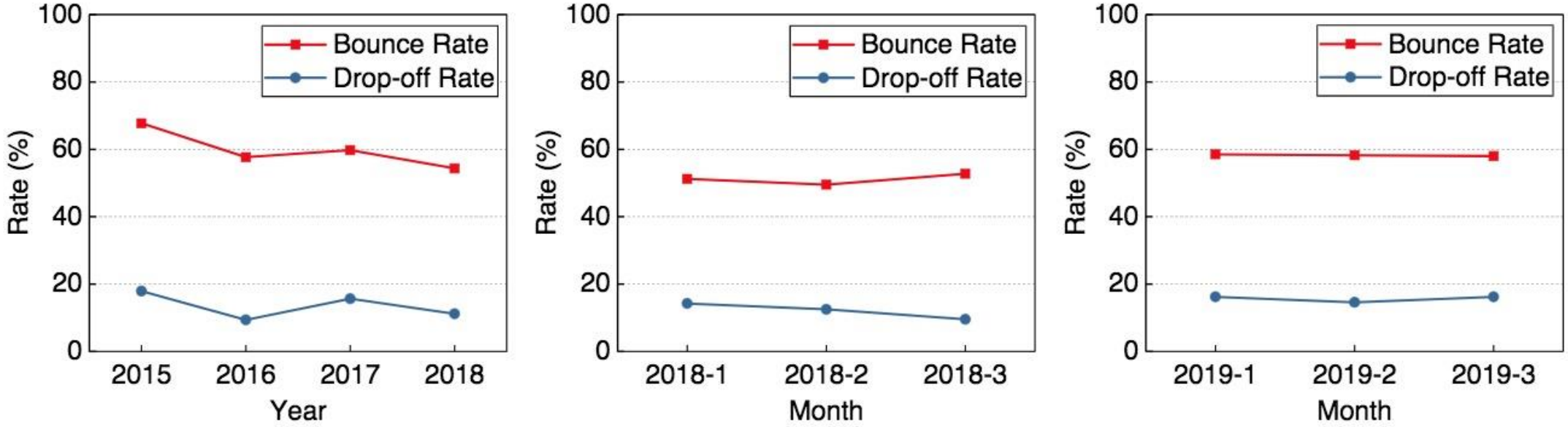
SYSTEM IMPACT EVALUATION

Experiment Time Line:

January, 2016
First module, customized recommender, implement

2018: Model rebuilding multiple times

January 2019: Entire framework implement



(a) The Bounce & Drop-off Rate from the Year 2015-2018.

(b) The Bounce & Drop-off Rate from Jan. to Mar., 2018.

(c) The Bounce & Drop-off Rate from Jan. to Mar., 2019

Fig. 6: The Bounce & Drop-off Rate for the year from 2015 to 2018, from January to March in year 2018, and from January to March in year 2019.

SYSTEM IMPACT EVALUATION

Experiment Time Line:

January, 2016
First module, customized recommender, implement

2018: Model rebuilding multiple times

January 2019: Entire framework implement

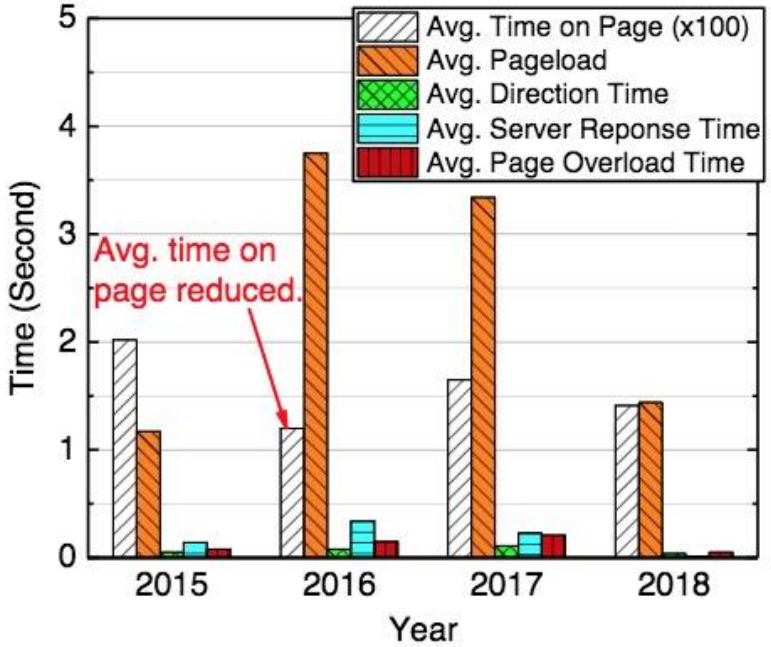


Fig. 7: The system usability statistics from 2015 to 2018.

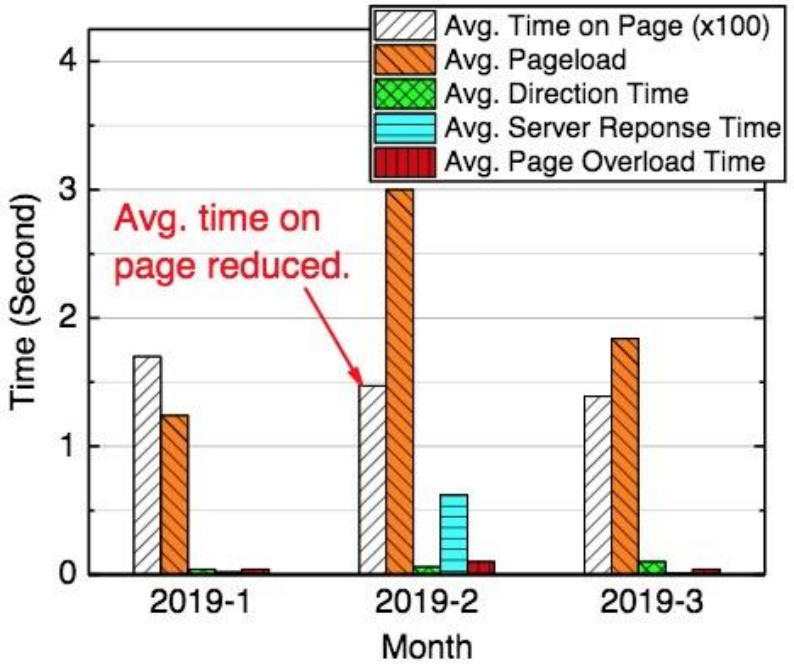


Fig. 8: The system usability statistics from Jan. to Mar., 2019.

CONCLUSION

1. Design **a flexible framework** that allow digital library can build their recommendation system purely from log data and metadata.
2. **Facilitate multiple popular recommenders** and implement it into a real-world Digital Repository System:dPanther (http://dpanther.fiu.edu).
3. Minimize the side effect (noisy recommendation) by **applying customizable group-based recommendation strategy**
4. The experimental evaluation shows that by applying the multi-process programming, the **model building time can be significantly reduced**.
5. The system usage statistics also indicate that during the evaluation time from January 2019 to February 2019, the Page Views have increased compared to 2018, **demonstrating the effectiveness of our proposed framework**.

?

QUESTIONS