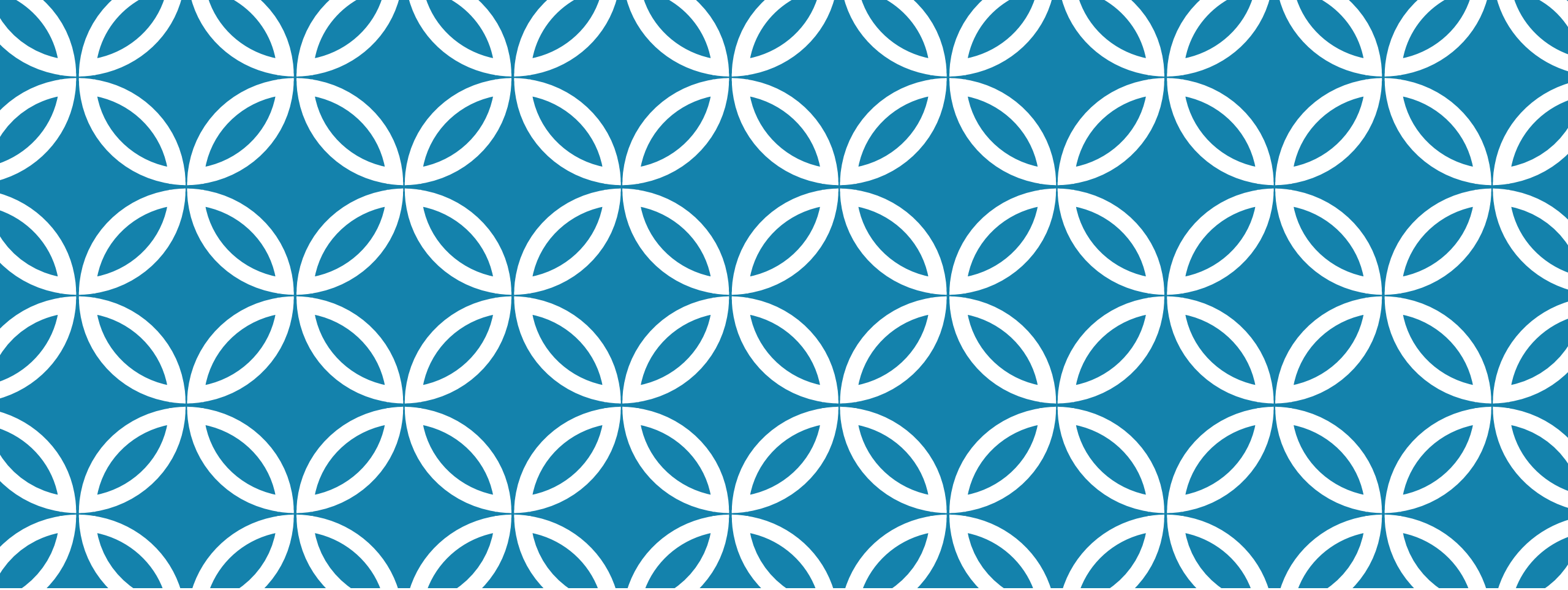


PROJECT PRESENTATION FOR FOLLOWEE RECOMMENDATION IN TWITTER

Noushin salek faramarzi
Nazanin Jafari

INTRODUCTION

Two step process in followee recommendation :

- Topology based approach
- Content based approach

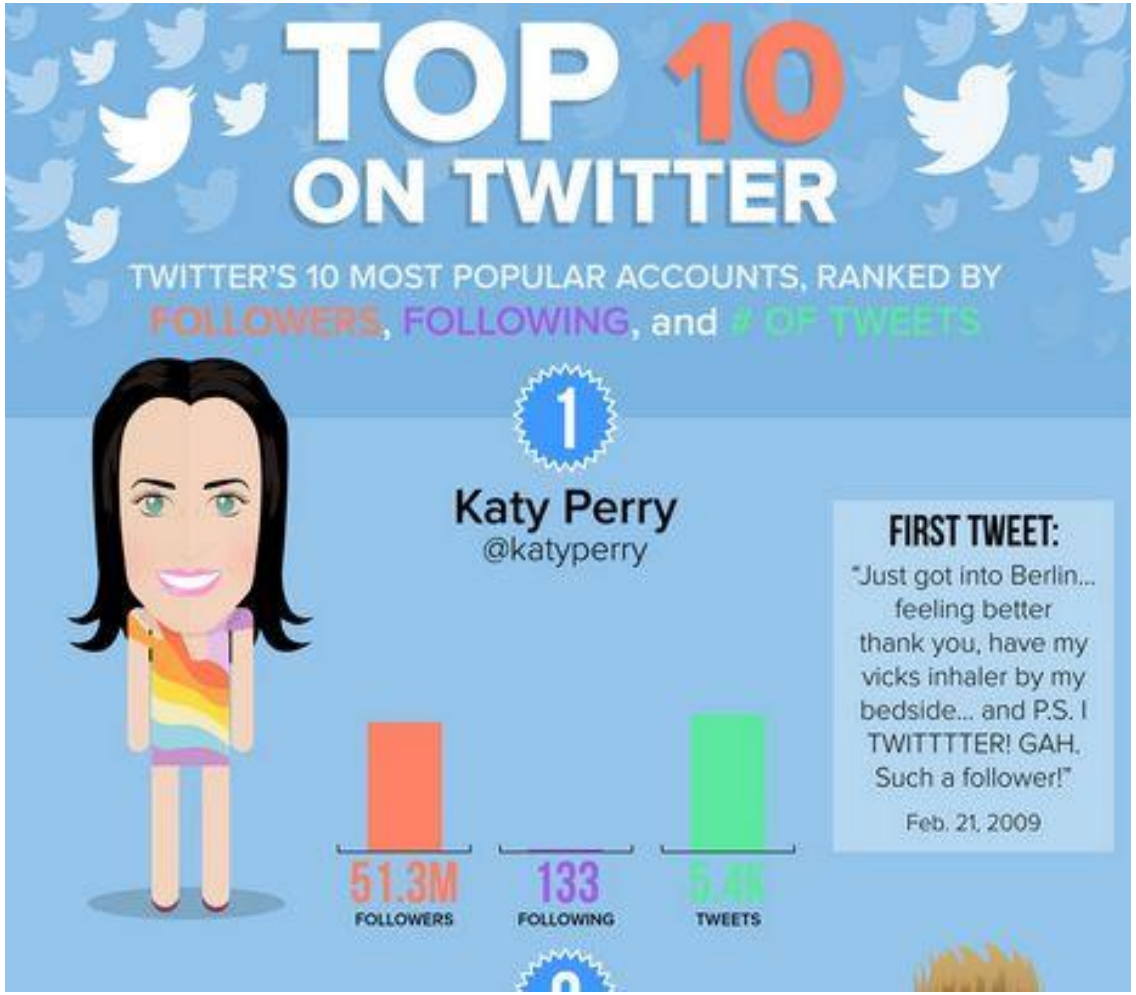
Topology based approach :

- Followee-follower relationship
- Obtaining candidate users to evaluate for content based approach

CONTENT BASED APPROACH

In twitter users are :

- Information sources
- Information seekers
- Friends



CONTENT BASED APPROACH (CONT'D)

Hypothesis :

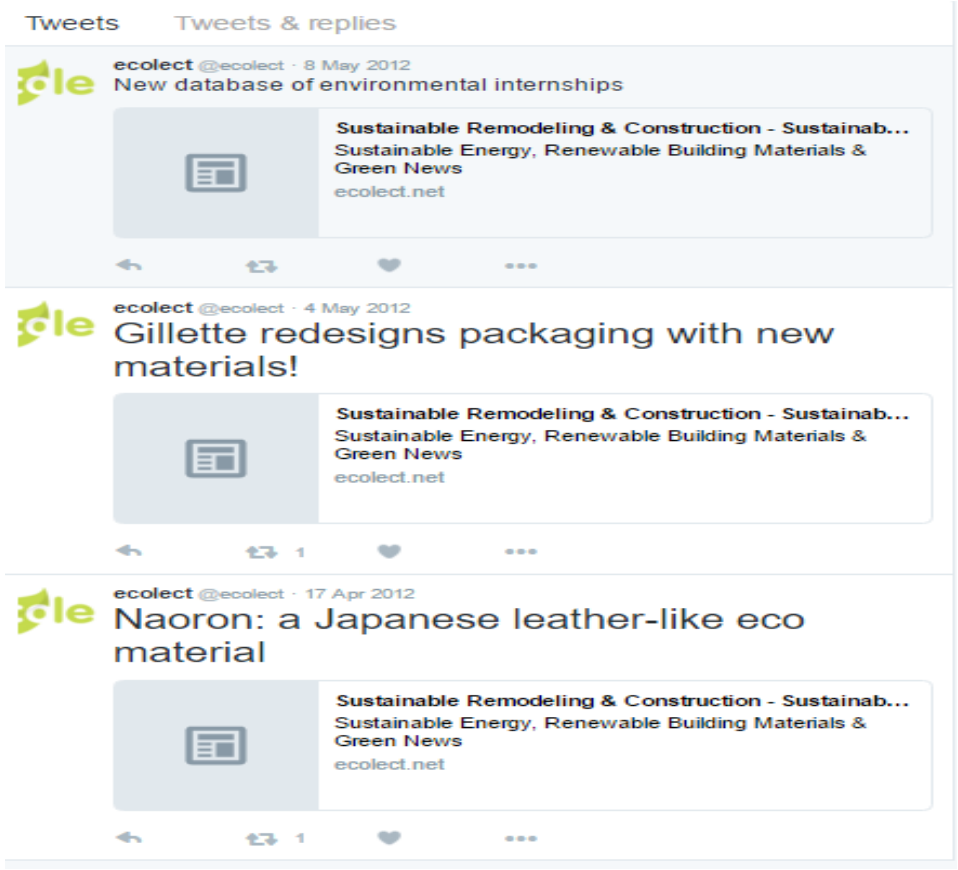
- Interest of target user :
  - relationships in Twitter
  - different content sources

4 PROFILING STRATEGY FOR UNDERSTANDING INTEREST OF TARGET USER

Idea of first profiling strategy :

users are likely to tweet about things that interest them therefore :

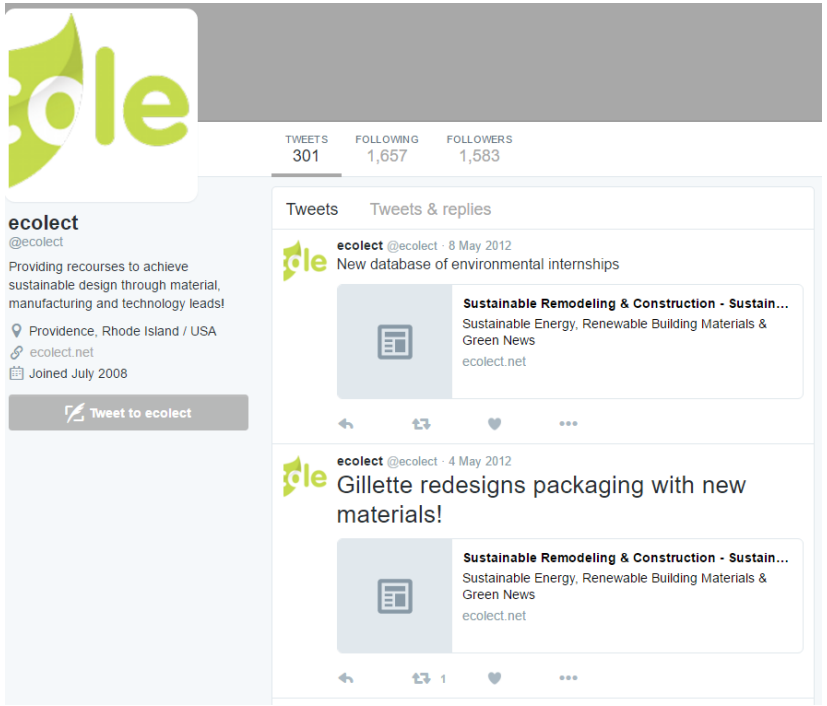
FIRST PROFILING STRATEGY (CONT'D)



- Tweets of a target user are gathered in a set
- And the profile of target user become the aggregation of his/her tweets.
- Similarity of target user and candidate user is :
  - Similarity of profiles of target user and candidate user measured by **cosine similarity**

4 PROFILING STRATEGIES (CONT'D)

Information seekers usually post few tweets themselves, but follow people that actively generate content.



On the other hand information sources actively generate content.

SECOND PROFILING STRATEGY

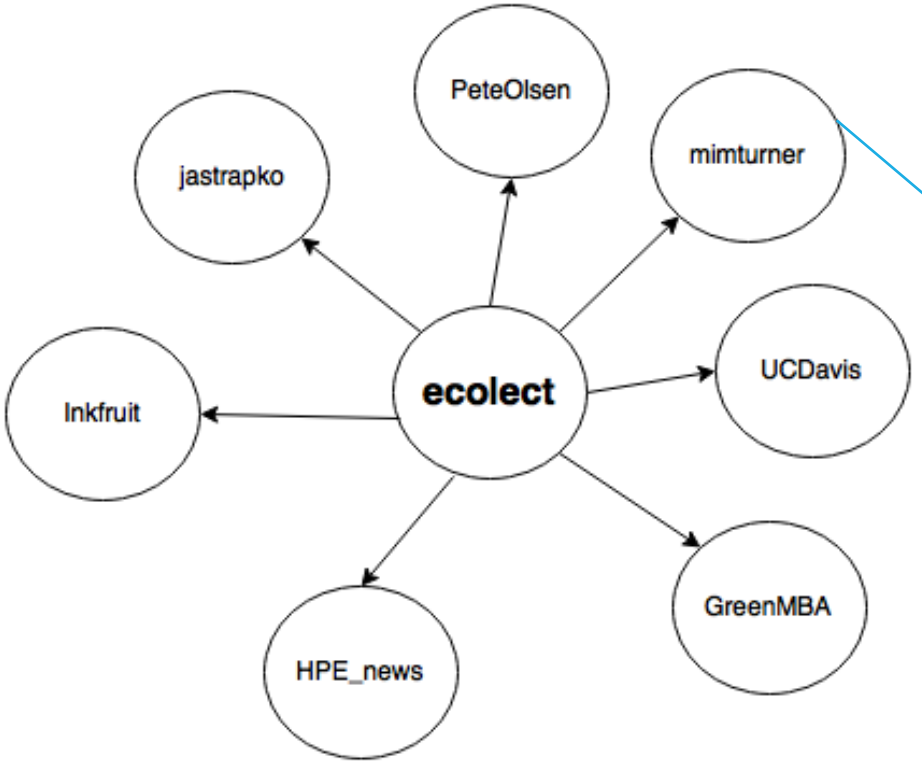
Since followee recommendation is toward information seekers :

- model the interests of a user based on who he/she is following

Second profiling strategy follows above model.

SECOND PROFILING STRATEGY (CONT'D)

The information a user is interested in can be seen as the aggregations of the profiles of his/her followees.



SECOND PROFILING STRATEGY (CONT'D)

For the target user the **profile** models the information he/she likes to read, whereas for the candidate user the profile models the information he/she published.

Similarity of profile of candidate user with the profile of the aggregation of profiles of followees of target user have been measured by cosine similarity.

This profiling strategy aggregates in a **single vector of** the information published by the user followees instead of the user own tweets.

4 PROFILING STRATEGIES (CONT'D)

Second profiling strategy considers all followees as responding to a unique topic of interest, which is not enough.

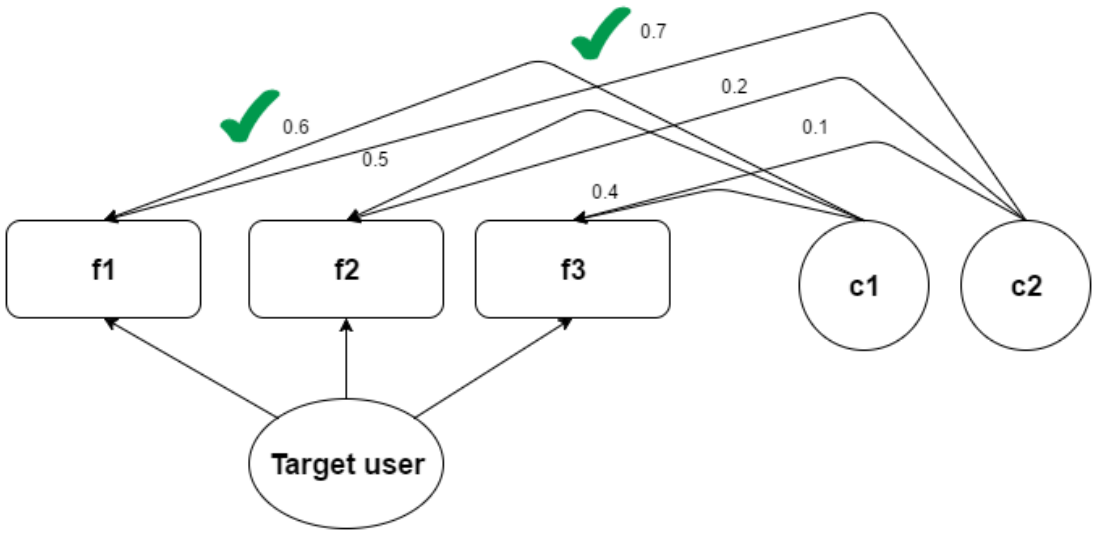
a user may follow celebrities, politicians, sportsmen and other types of users whose information will coexist in the vector representing the target user profile.

THIRD PROFILING STRATEGY

Rather than using a single vector, **multiple vectors** are used, each of which represents a different followee.

This profiling works as follows :

- Similarity of each candidate user with followees calculated and **most similar ones** added to the array list.
- Ranking of candidates are based on highest similarity between each candidate and corresponding followee.



4 PROFILING STRATEGIES (CONT'D)

In a more realistic view of a user information preference:

- users are likely to follow people in **different** interest categories.
- Hence, to assess a more precise description of the user interest a last type of profile groups user followers into meaningful category.

FOURTH PROFILING STRATEGY

The identification of categories, needs to be incrementally discovered.

In this clustering approach

- as soon as the user subscribes to a followee it is assigned to the first cluster or category in the user profile.
- Each subsequent followee is incorporated to either some of the existent categories or to a novel category
  - **depending on its similarity with the current categories.**

FOURTH PROFILING STRATEGY (CONT'D)

clustering algorithm returns a set of categories in which the current followees of the user can be grouped into.

Each time user follows another user, new followee have to incorporate with existing clusters hence centroids need to be defined.

Similarity of new followee and all clustering centroids define which cluster new followee belongs to.

FOURTH PROFILING STRATEGY (CONT'D)

The profile of a user using this strategy is then defined as the set of the centroids of the clusters identified.

Finally similarity of target user with this profile and the candidate user profile have been calculated.

PREPROCESSING DATA

Clearing data from slang ,stemming and lemmatization.

Removing users and their connections to other users whose tweets were less than 10.

DATA SET

Our data set consists of 54,000 followee-follower relationship and their corresponding tweets.

Tweets of users after preprocessing reach more than 3 million.

EVALUATION OF THE HYPOTHESIS

Some of the target user followees are hidden from the recommendation algorithm and then it is verified if they were discovered and suggested as future followees.

Set of followees of each user were partitioned into 70% for training, and 30% for testing.

In order to make the results **less sensitive** to the particular training/testing partitioning:

- **the average and standard deviation of 5 runs for each individual user** are reported,
- each time using a different random partitioning into training and test sets.

EVALUATION METRICS

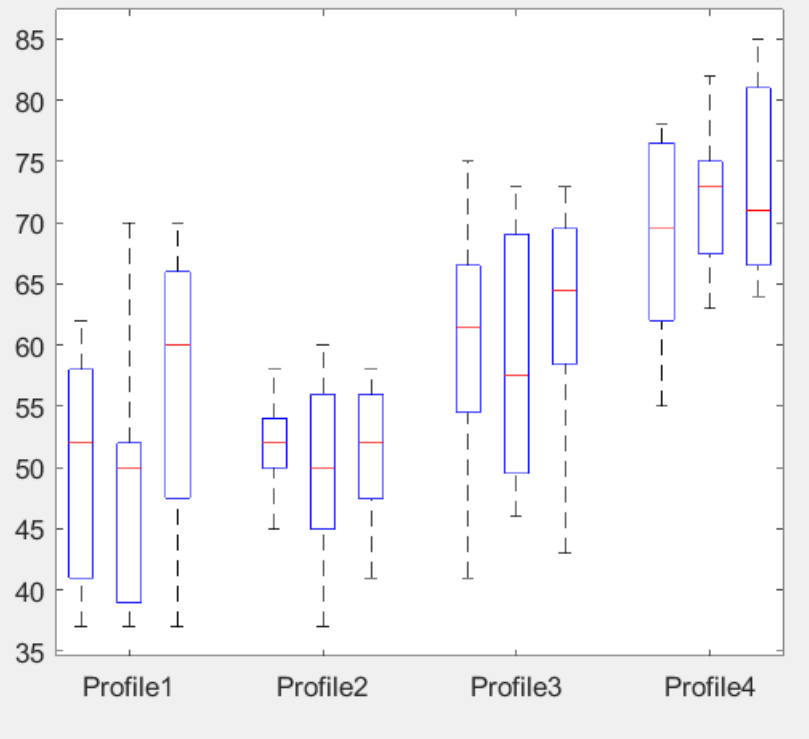
- Hit rate:

- the number of followees in the test set that were also present in the topN recommended followees for a given test user.

$$HR = \frac{number\ of\ hits}{|U_{test}|}$$

EXPERIMENTAL RESULTS

Hit rate results:



Standard deviation and hit rate measures :

	Total STD	*Total Hit Rate*
Profile 0	1.86	51.83
Profile 1	1.85	51.25
Profile 2	1.90	60.54
Profile 3	2.39	71.29

FUTURE WORKS

Implementing 4th profiling strategy with

- single link,
- complete link,
- hit diffusion
- k-means clustering algorithms.

Evaluate which one can have better performance compared to this clustering algorithm.

CONCLUSION

Using first and second profile does not give us high accuracy in both of the hit rate metric and precision metric which is expected logically.

Third and fourth profile strategy had better results leading us to conclude that :

- Information seeker's interest can be obtained by the content published by his/her followee rather than content published by himself.