Research Content Recommender

MARIO DIAZ DE LA ROSA

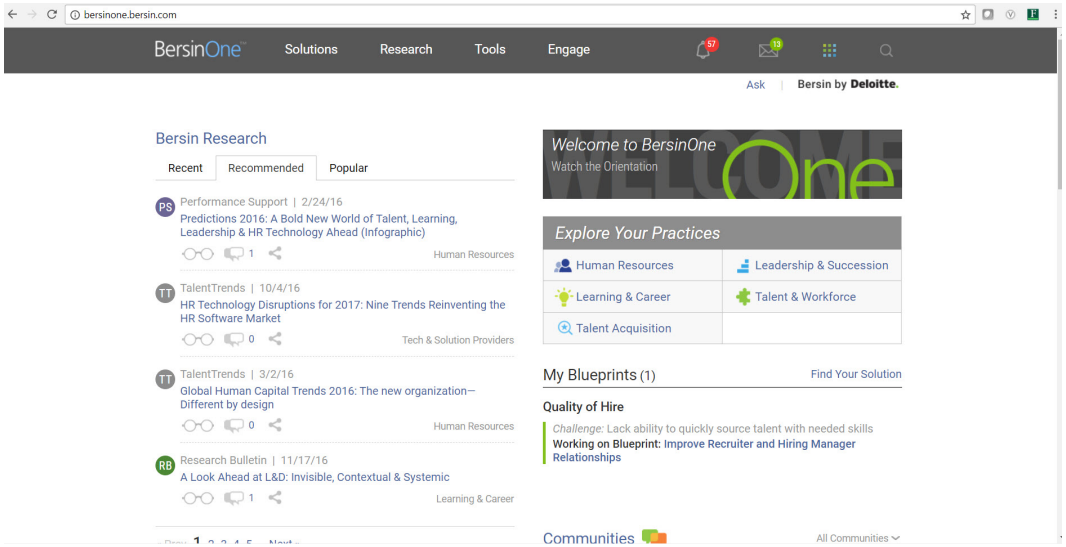
GALVANIZE, SAN FRANCISCO

Bersin by **Deloitte.**

Online research and content service for HR professionals

Provides articles, decision support tools, and online community to help hire, retain, reward best people.

- ~ 45,000 registered users
- ~ 3,000 research items to download



Recommend content tailored to the user's tastes & increase user engagement and retention

- Currently recommend most visited content (~ 2 % accuracy)

GOAL: Harness User Preferences and User Activity to predict favorable content.

Data & User Activity

10 years of User activity data and Content metadata:

- Downloads (~25,000 users)
- Favorites (~1,800 users)
- User's choice of Subject Areas and professional profiles
- Content metadata

Explicit preference measures are minimal.

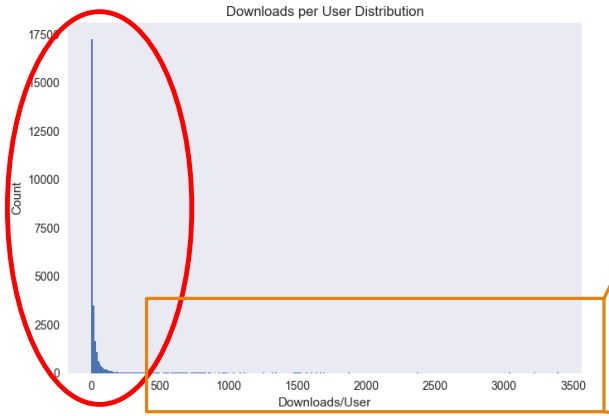
Extract as much signal as possible from implicit preferences

Focus on:

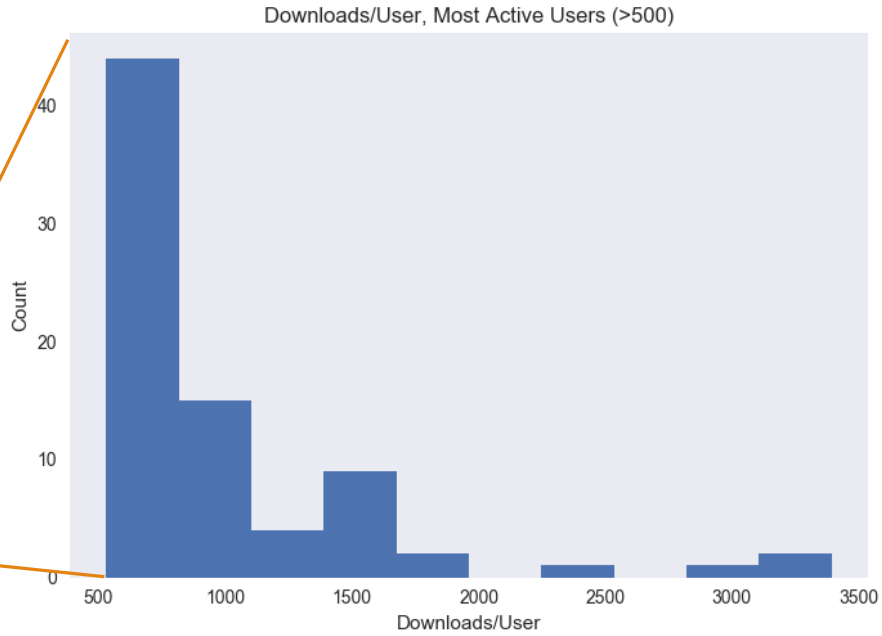
- Download history
- Recommend on User behavior
- Basic User/Item profile features

Data & User Activity

A big cold-start and sparsity problem

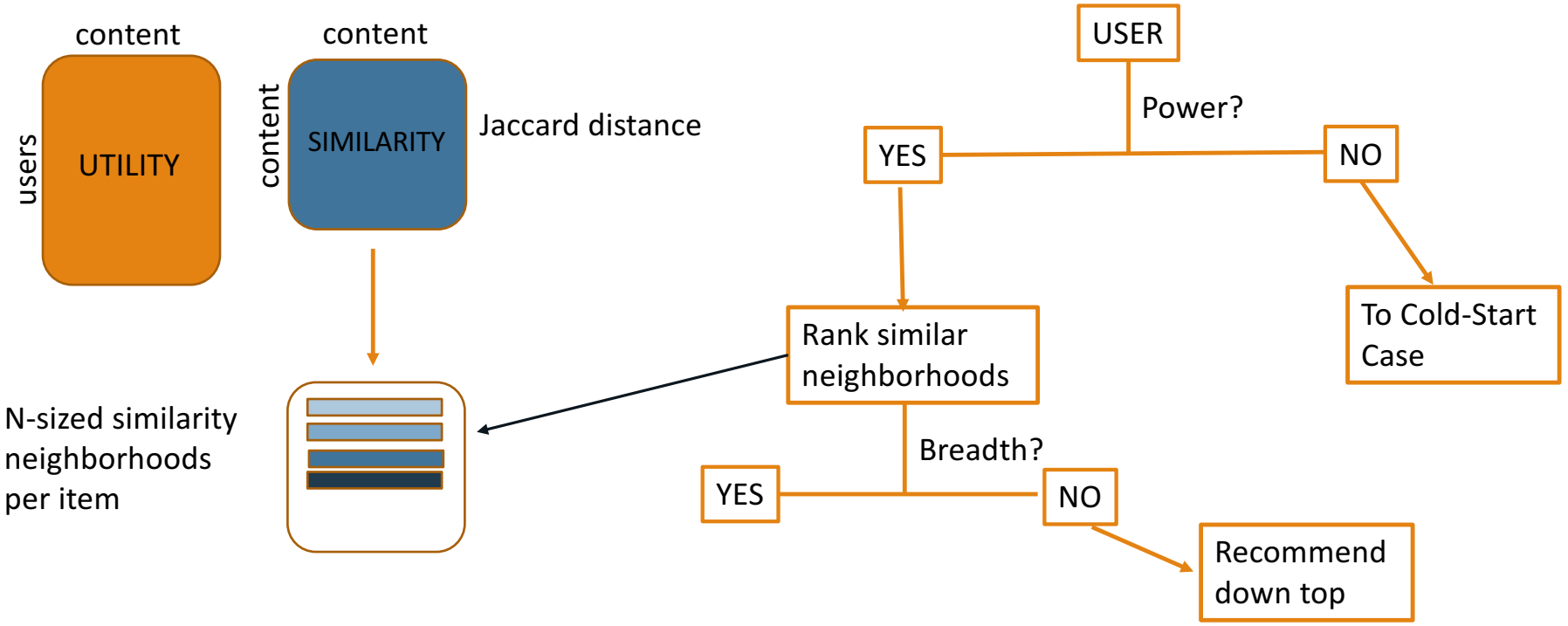


Vast majority of Users download very few documents (1 or 2)



Download activity of top 500 users

Collaborative Filtering



Collaborative Filtering

Parameters:

- Neighborhood size N
- Breadth vs. Depth
- Number of Recommendations

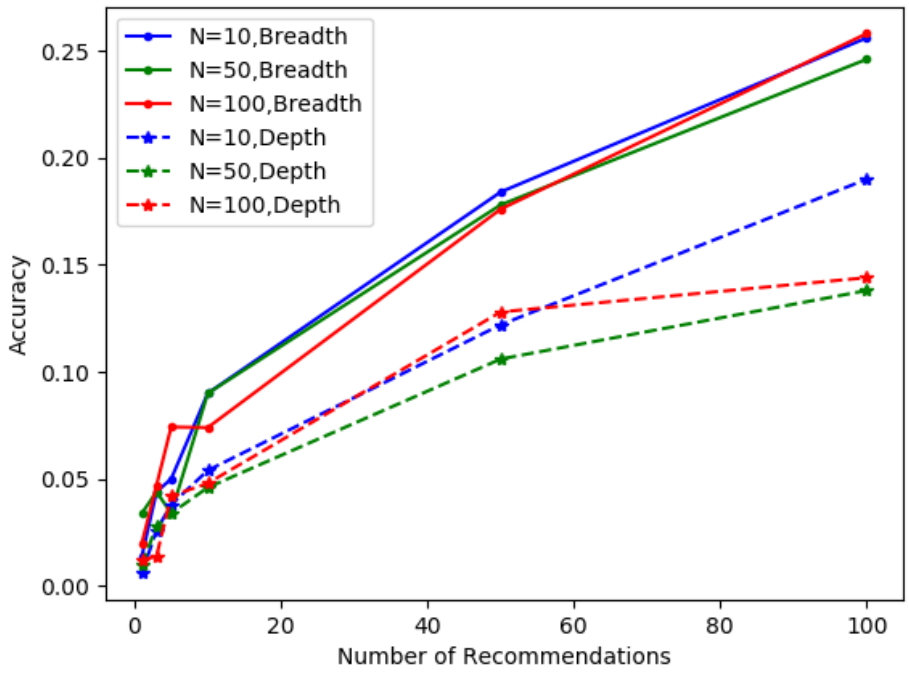
Validation:

- Recommender relies on User behavior
- **Leave-one-out**

Recommendation improves when emphasizing breadth over depth

Use recommender as 1st step to generate initial set, then narrow further with content/profile information

Accuracy = Is item in recommended set?



Conclusions & Recommendations

Users prefer breadth to depth in recommendations

Present A/B testing framework to evaluate recommender performance vs. current baseline

Future:

- More in-depth feature work
- Clustering of users/Items to reduce dimensionality of problem
- Other ML techniques to further narrow down from initial set

Cold-start problem: similarity with active users and nudge

Feature work along with cleaning of the data

Investigate time effect: interest drift, trends

Recommendations on data cleaning/management