PERSONALIZED NEWS RECOMMENDATION SYSTEM

GITHUB LINK: github.com/VishrutMehta/PersonalizedNewsRecommendationEngine/tree/master

Team Members:

Prateek Sachdev
Vidit Gupta
Mayank Natani
Vishrut Mehta

Machine Learning

BACKGROUND

- Recommendation systems are a type of information filtering systems that recommend products that are likely to be of interest to the user.

- Large number of news articles everyday.

- Dataset:
  -Reuters text classification dataset
  -19043 articles (~79.2 MB)

- ML Problems:
  - Clustering
  - Collaborative Filtering
  - Content Based Recommendation

PROBLEMS

- Businesses need to optimize their recommendation engines to handle recommending thousands and often millions of options across a multitude of channels, while reducing churn and enhancing customer experience.

- Need to cluster news articles, find user similarity(collaborative filtering), and recommend.

- Extract useful information from articles and rank them on parameters like:
  - Entities
  - Recency
  - Interest (global and personal)
  - users with similar interest

- Useful in field of computational advertising, web search, movie recommendation.

- Difficult to evaluate the relevancy of result.

DATA INSIGHTS

- Text/ Time series data

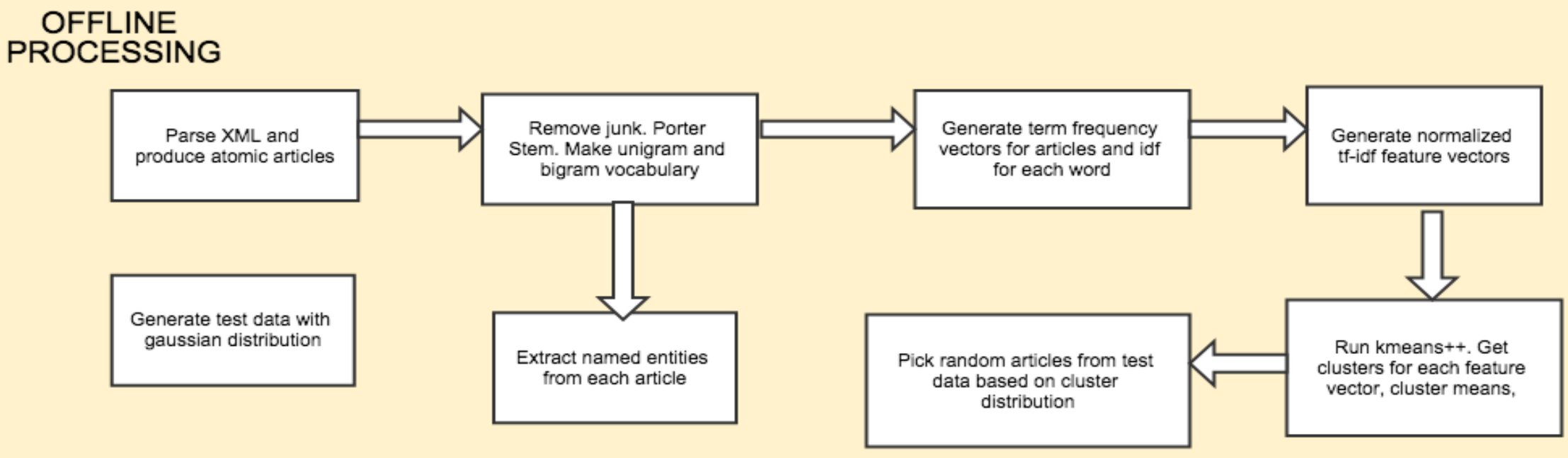
- Dimensions:
  - 5,579 (unigrams) + 9,953 (bigrams) = 15,532
  Unigrams in >= 12 articles out of 19,000
  Bigrams in >= 13 articles out of 19,000

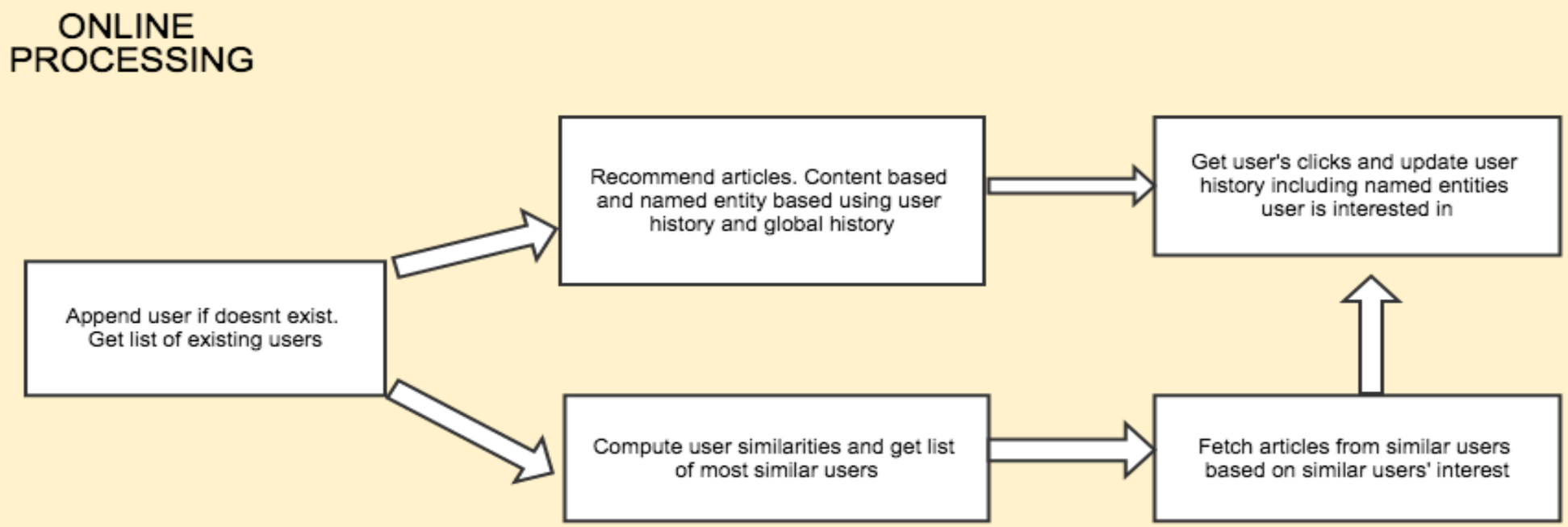
- Clusters:
  15

- Sparsity:
  0.010043

- Entities (unique):
  35,283 in 19,000 articles

SOLUTION


OFFLINE PROCESSING
Parse XML and produce atomic articles
Remove junk. Porter Stem. Make unigram and bigram vocabulary
Generate term frequency vectors for articles and idf for each word
Generate normalized tf-idf feature vectors
Generate test data with gaussian distribution
Extract named entities from each article
Pick random articles from test data based on cluster distribution
Run kmeans++. Get clusters for each feature vector, cluster means,


ONLINE PROCESSING
Append user if doesnt exist. Get list of existing users
Recommend articles. Content based and named entity based using user history and global history
Get user's clicks and update user history including named entities user is interested in
Compute user similarities and get list of most similar users
Fetch articles from similar users based on similar users' interest

OFFLINE PROCESSING

- Once we get normalised feature vectors using TF*IDF, we cluster articles using spherical k-means++

$$IDF(w) = \log\left(\frac{1}{P(w)}\right)$$

$$TF(w \mid d) = \log(1 + N(w \mid d))$$

$$x(w,d) = \frac{TF(w \mid d) * IDF(w)}{\sqrt{\sum_{w' \in d} (TF(w' \mid d) * IDF(w'))^2}}$$

$$x(d)_2 = \sqrt{\sum_{w \in d} x(w,d)^2} = 1$$

$$\partial(n,k,t) = (k == \arg\max_{j=1..K} [\{x(n), m(j,t)\}] = \sum_{d=1}^{D} x(d,n) m(j,d,t)]$$

OFFLINE PROCESSING

$$u(k,t+1)=\frac{\sum_{n=1}^{N}\partial(n,k,t)x(n)}{\sum_{n=1}^{N}\partial(n,k,t)}=Un-normalized\ clusters$$

$$m(k,t+1)=\frac{u(k,t+1)}{\|u(k,t+1)\|_2}=Normalized\ clusters$$

- FFP

$$P(m(0,l+1)=x(n))=\frac{D(x(n))^2}{\sum_{i=1}^{N}D(x(i))^2}$$

$$D(x(n))=D(x(n)\,|\,m(0,1),m(0,2)...m(0,l))=\min_{j=1..l}(\Delta x(n),m(j,0))$$

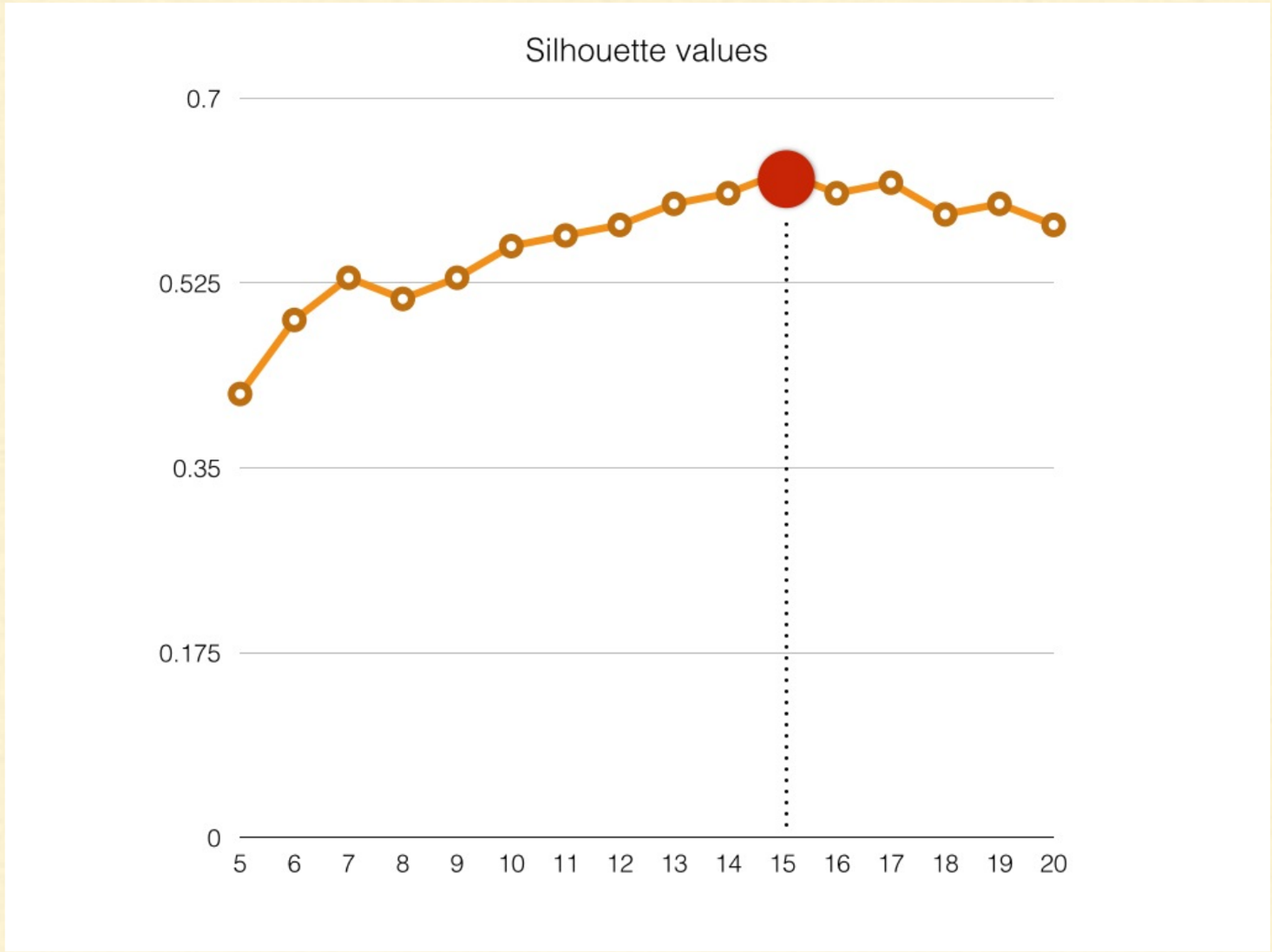
OFFLINE PROCESSING

- We calculate silhouette values to determine natural number of clusters

- Let a(i) be the average dissimilarity of i with all other data within the same cluster.

- Let b(i) be the average dissimilarity of i with all other data of closest cluster i is not assigned to it

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$



OFFLINE PROCESSING

- Named entity recognition

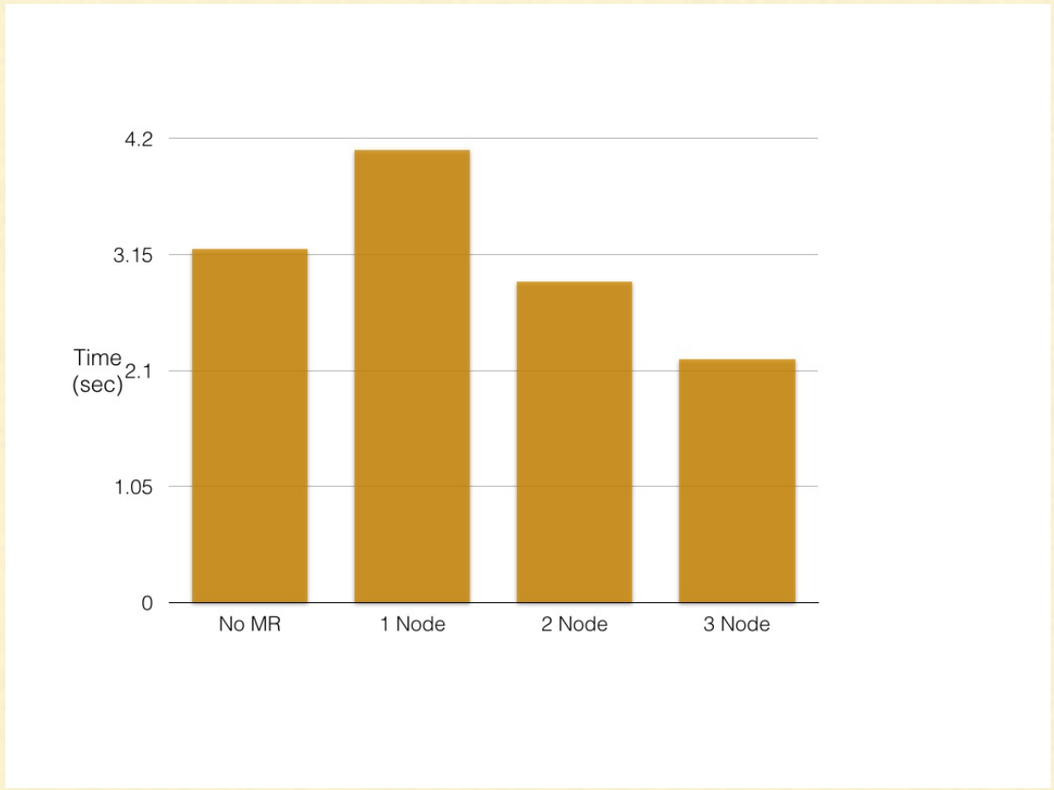
  - Used Stanford's natural language toolkit (nltk) to extract named entities from articles

- Map-Reduce:

- Computed tf-idf using map reduce taking all TF and IDF as input.

  - MAP:
    We take the input as the TF vector and the IDF vector and output the key-value pair - (TF vector, IDF vector).

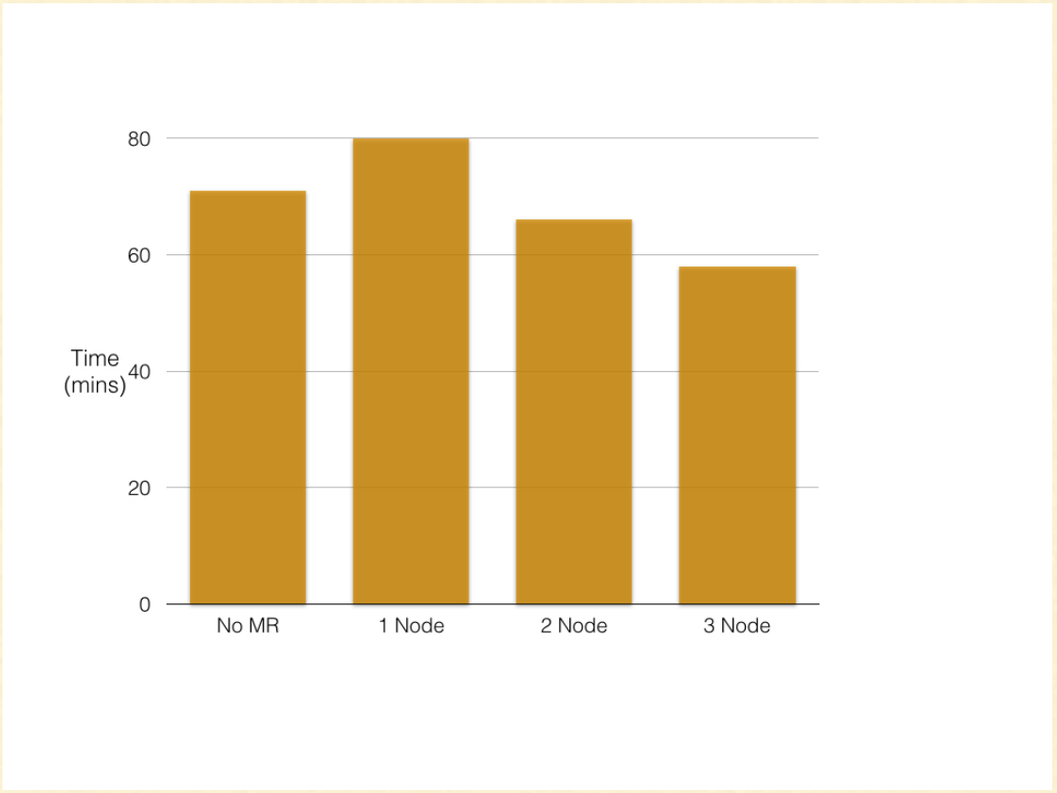
  - REDUCE:
    We multiply the value and give the output as (result vector, 1).



OFFLINE PROCESSING

- Named entity recognition for all articles.

  - MAP
    In map step, we take articles
    as input and output the named entities
    as (key,value) pair as (named_entities, 1).



  - REDUCE
    In reducer, we take these values, and then merge them as (named_entities, total_count).

OFFLINE PROCESSING

- Test data generation

  - Generated gaussian distributions for each cluster vs clicks
  - Picked random article from each cluster for each click in it
  - Saved entities for article
  - Generated random timestamp for each click within a month span

  - For each user i

user(i):

cluster(0) - click(1), click(2),…..click(k1)

cluster(1) - click(1), click(2),…..click(k2)

.

.

.

cluster(j) - click(1), click(2)……click(kj)

where there are j clusters

ONLINE PROCESSING

- Read user's historical data and give interest score to each cluster

$NTS_{i,j} =$ Normalized time stamp for the active user in cluster i

$CS_i =$ Cluster score of cluster i for given user

$\text{num\_clicks}_i =$ Number of clicks in the cluster i

$$NTS_{i,j} = \frac{TS_{i,j}}{\sum_{i=0}^{num\_clusters} \sum_{j=0}^{num\_clicks_i} TS_{i,j}} \qquad CS(i) = \sum_{i=0}^{num\_clicks_i} NTS_{i,j}$$

- Read user's interest in entities along with their scores
- Iterate over randomly selected articles from test data and assign to appropriate cluster using cluster mean vectors
- Extract entities from each of these articles
- Compute entity similarity score for each article with user's entities of interest (dot product)

$$entity\_similarity\_score_i = (user\_entity\_score) \circ (entity\_score_i)$$

ONLINE PROCESSING

- Assign each article a cluster similarity score according to user's interest
  For each article i

$$cluster\_similarity\_score_i = CS_i \text{ for article } i$$

$$final\_score\_user = entity\_similarity\_score + cluster\_similarity\_score$$

- Global Interest

$$entity\_score\_global = \bigcup_{i=1}^{total\_users} entity\_score\_user_i$$

$$TS\_global_i = \bigcup_{j=1}^{total\_users} TS_i^{user_j} \qquad NTS\_global_{i,j} = \frac{TS\_global_{i,j}}{\sum TS\_global_{i,j}}$$

$$final\_score\_global = entity\_score\_global + cluster\_score\_global$$

$$final\_score\_article = w_1 * final\_score\_user + w_2 * final\_score\_global$$

ONLINE PROCESSING

- Calculate user similarity on the basis of user history.

$$normalized\_clicks(c,i) = \frac{clicks(c,i)}{\sqrt{\sum_{c=1}^{num\_clusters} clicks(c,i)^2}}$$

$$clicks(c,i) = number \text{ of clicks in cluster c by user}_i$$

$$user\_similarity(i,j) = normalized\_clicks(i) \circ normalized\_clicks(j)$$

- Rank the users on their similarity and pick top-K most similar users.
- Pick articles read by most similar users with probability proportional to their similarity score.
- Create new user data and periodically merge it with user history.

User Interface

CLUSTER 0

COBANCO INC YEAR NET

OHIO MATTRESS MAY HAVE LOWER 1ST QTR NET

AM INTERNATIONAL INC 2ND QTR JAN 31

BROWN-FORMAN INC 4TH QTR NET

DEAN FOODS SEES STRONG 4TH QTR EARNINGS

Cluster wise articles

News Articles

JAPAN IN LAST DITCH EFFORT TO SAVE CHIP PACT

Japan has launched a last-ditch effort tosalvage its computer micro-chip pact with t
American policy makers setting out its case and instructing its producers to cut out
effort to ward off any catastrophe," Ministry of International Trade and Industry (M
Yamamoto told reporters. "If hasty action is taken in the United States, it will creat
Administration's Economic Policy is expected to meet Thursday to review Japanes
hammered out last year. Under the pact, Tokyo agreed to stop selling cut-price chip
imports of American semiconductors. Washington has accused Japan of reneging o
Asia and by failing to boost American imports, and has threatened to take retaliato
agreement, MITI is asking Japanese chip makers to limit production in the hope tha
the incentive to export. Yamamoto said that Japan will slash output of 256 kilobit d
programmable read only memory chips by 11 pct in the second quarter. This follow

History based recommendations

Similar users also read

TRITON GROUP LTD 4TH QTR JAN 31 NET

Oper shr profit nil vs loss nil Oper net profit 671,000 vs loss 138,000 Sales 104.3 mln vs 70.8 mln Avg shrs 101.2 mln vs 66.8 mln Year Oper shr profit six cts vs profit five cts Oper net profit 6,309,000 vs profit 5,144,000 Sales 349.8 mln vs 303.4 mln Avg shrs 85.0 mln vs 76.3 mln NOTE: Net excludes discontinued operations nil vs gain 196,000 dlrs in quarter and loss 293,000 dlrs vs gain 407,000 dlrs in year. Net excludes tax loss carryforward 1,423,000 dlrs vs reversal of tax credit 625,000 dlrs in quarter and credits 5,437,000 dlrs vs 7,261,000 dlrs in year. Results include U.S. Press Inc from November Three acquisition. Reuter

AAR CORP 3RD QTR FEB 28 NET

Shr 37 cts vs 32 cts Net 3,892,000 vs 2,906,000 Sales 71.8 mln vs 64.5 mln Nine mths Shr 1.08 dlrs vs 91 cts Net 10,946,000 vs 8,206,000 Sales 214.1 mln vs 179.4 mln Avg Shrs 10.5 mln vs 9.1 mln Reuter

Collaborative filtering

DEMO

FUTURE PROSPECTS

- Make a distributed system which can handle more traffic on the website.
- Take into account user's location and language.
- Automating the recommendation system's evaluation.
- Spam filtering of articles.
- Parallelisation of k-means.

THANK YOU !!

QUESTIONS?