Social Data Exploration

Sihem Amer-Yahia

DR CNRS @ LIG

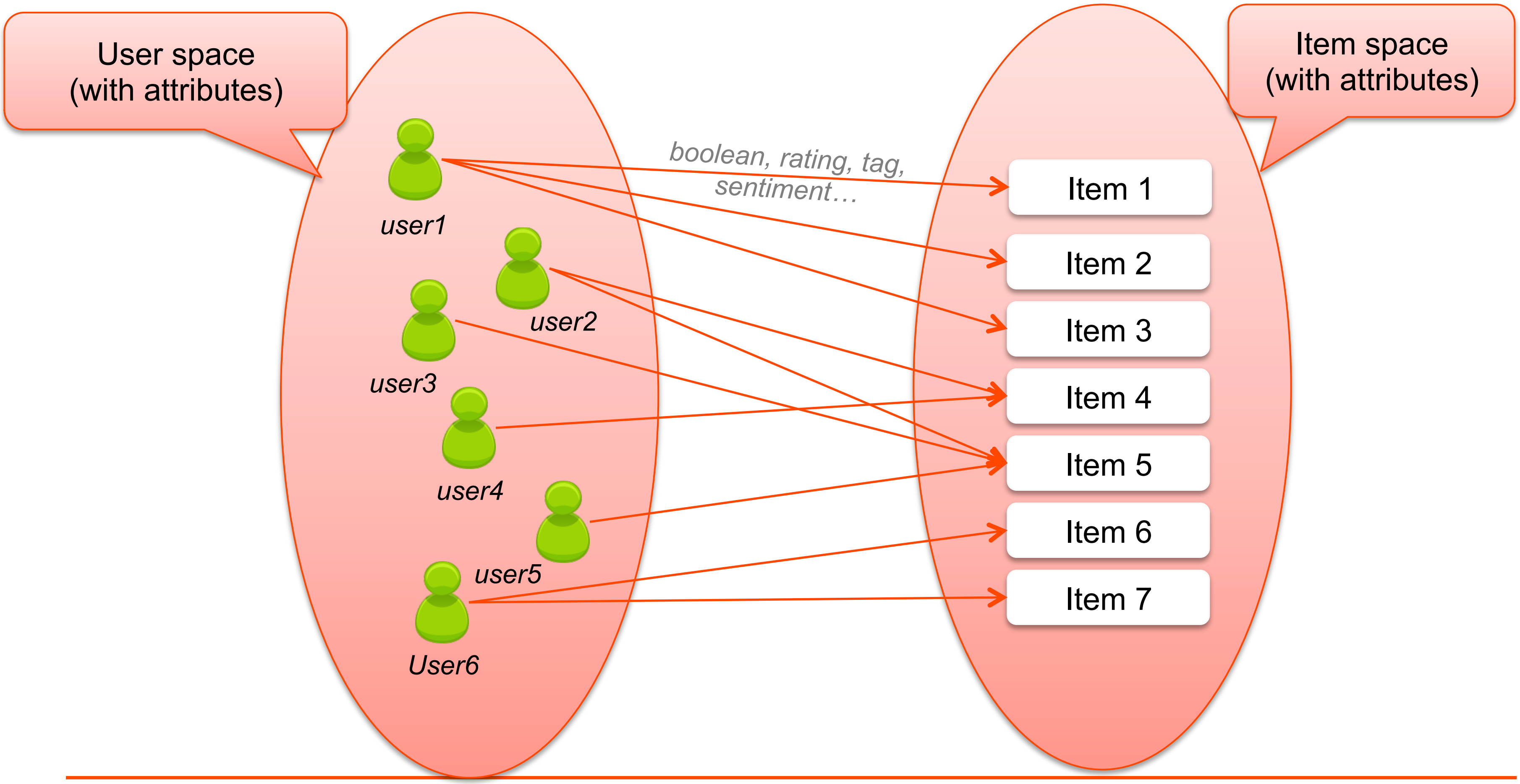
Sihem.Amer-Yahia@imag.fr

Big Data & Optimization Workshop

12ème Séminaire POC

LIP6 Dec 5th, 2014

Collaborative data model



MovieLens instances

ID		Title	Genre	Director		Name	Gender	Location		Rating
1		Titanic	Drama	James Cameron		Amy	Female	New York		8.5
2		Schindler's List	Drama	Steven Speilberg		John	Male	New York		7.0

ID		Title	Genre	Director		Name	Gender	Location		Tags
1		Titanic	Drama	James Cameron		Amy	Female	New York		love, Oscar
2		Schindler's List	Drama	Steven Speilberg		John	Male	New York		history, Oscar

More on MovieLens datasets

http://grouplens.org/datasets/movielens/

MovieLens 100k

100,000 ratings from 1000 users on 1700 movies.

- README.txt
- ml-100k.zip
- Index of unzipped files

MovieLens 1M

1 million ratings from 6000 users on 4000 movies.

- README.txt
- ml-1m.zip

MovieLens 10M

10 million ratings and 100,000 tag applications applied to 10,000 movies by 72,000 users.

- README.html
- ml-10m.zip

Social Data Exploration

- Rating exploration
  - Meaningful Interpretations of Collaborative Ratings
- Tag exploration
  - Who Tags What? An Analysis Framework
- Perspectives

Meaningful Interpretations of Collaborative Ratings

Data Model

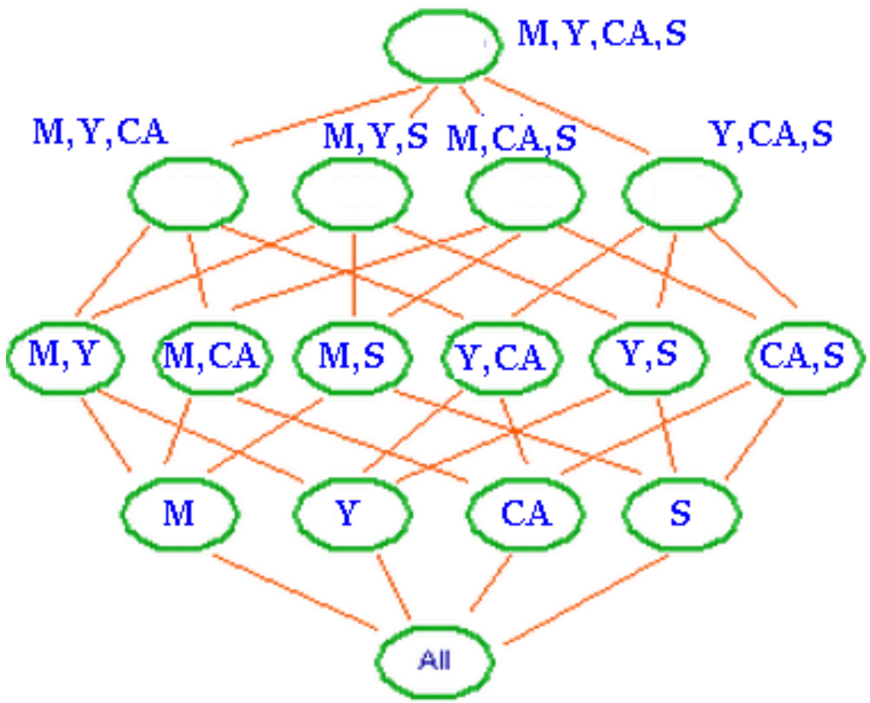
- Collaborative rating site: Set of Items, Set of Users, Ratings
  - Rating tuple: <item attributes, user attributes, rating>

ID		Title	Genre	Director		Name	Gender	Location		Rating
1		Titanic	Drama	James Cameron		Amy	Female	New York		8.5
2		Schindler's List	Drama	Steven Speilberg		John	Male	New York		7.0

- Group: Set of ratings describable by a set of attribute values

- Notion of **group** based on data cube
  - OLAP literature for mining multidimensional data

Exploration Space



Each **node/cuboid** in lattice is a **group**

A = Gender: Male
B = Age: Young
C = Location: CA
D = Occupation: Student

Task
Quickly identify "good" groups in the lattice that help analysts understand ratings effectively

Partial Rating Lattice for a Movie

(M:Male, Y:Young, CA:California, S:Student)

DEM: Meaningful Description Mining

- For an input item covering R_I ratings, return set C of groups, such that: <u>description error</u> $\mathtt{error}(C, R_I)$ is minimized, subject to:
  - $|C| \leq k$;
  - coverage $\mathtt{coverage}(C, R_I) \geq \alpha$

Description Error

Measures how well a group average rating approximates each individual rating belonging to it

$$\begin{aligned}\mathtt{error}(C, R_I) &= \Sigma_{r \in R_I}(E_r) \\ &= \Sigma_{r \in R_I} \mathtt{avg}(|r.s - \mathtt{avg}_{c \in C \wedge r \lessdot c}(c)|)\end{aligned}$$

Coverage: measures percentage of ratings covered by returned groups

- DEM is NP-Hard: proof details in [1]

[1] *MRI: Meaningful Interpretations of Collaborative Ratings, S. Amer-Yahia, Mahashweta Das, Gautam Das and Cong Yu. In the Proceedings of the International Conference on Very Large Databases (PVLDB), 2011.*

DEM: Meaningful Description Mining

- Identify groups of reviewers who consistently share similar ratings on items

DEM: Meaningful Description Mining

THEOREM 1. *The decision version of the problem of meaningful description mining (DEM) is NP-Complete even for boolean databases, where each attribute ia_j in $\mathcal{I}_A$ and each attribute ua_j in $\mathcal{U}_A$ takes either 0 or 1.*

To verify NP-completeness, we reduce the Exact 3-Set Cover problem (EC3) to the decision version of our problem. EC3 is the problem of finding an exact cover for a finite set U, where each of the subsets available for use contain exactly 3 elements. The EC3 problem is proved to be NP-Complete by a reduction from the Three Dimensional Matching problem in computational complexity theory

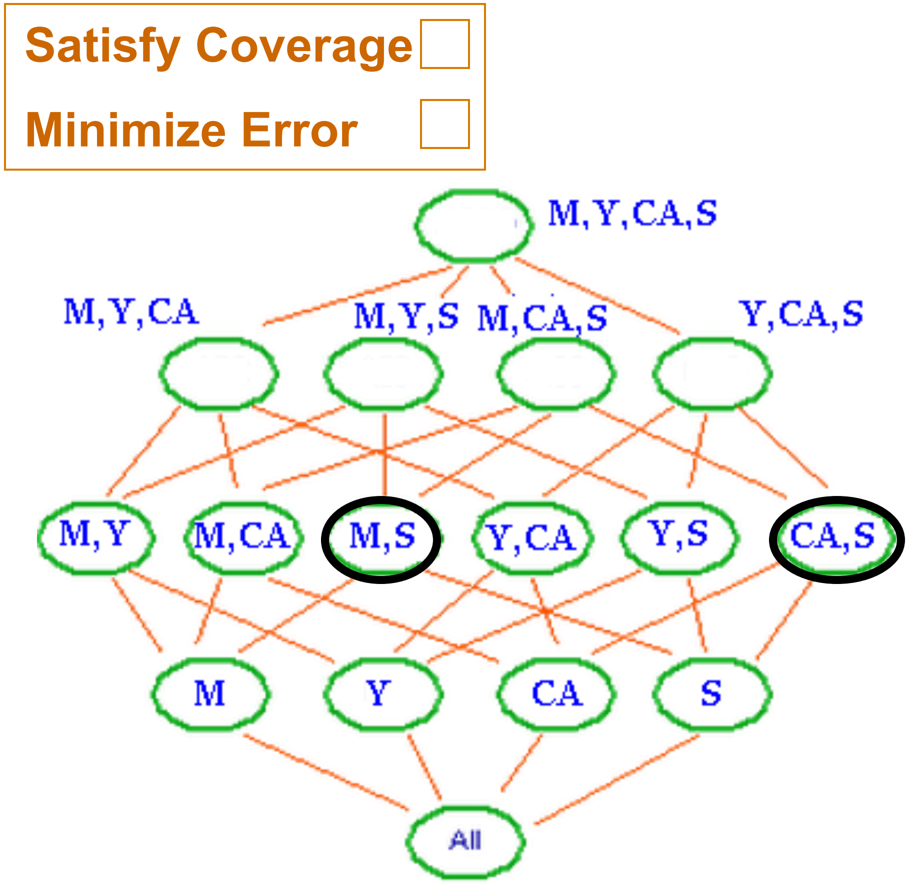
DEM Algorithms

- **Exact Algorithm (E-DEM)**
  - Brute-force enumerating all possible combinations of cuboids in lattice to return the exact (i.e., optimal) set as rating descriptions

- **Random Restart Hill Climbing Algorithm**
  - Often fails to satisfy Coverage constraint; Large number of restarts required
  - Need an algorithm that optimizes both Coverage and Description Error constraints simultaneously

- **Randomized Hill Exploration Algorithm (RHE-DEM)**

RHE-DEM Algorithm

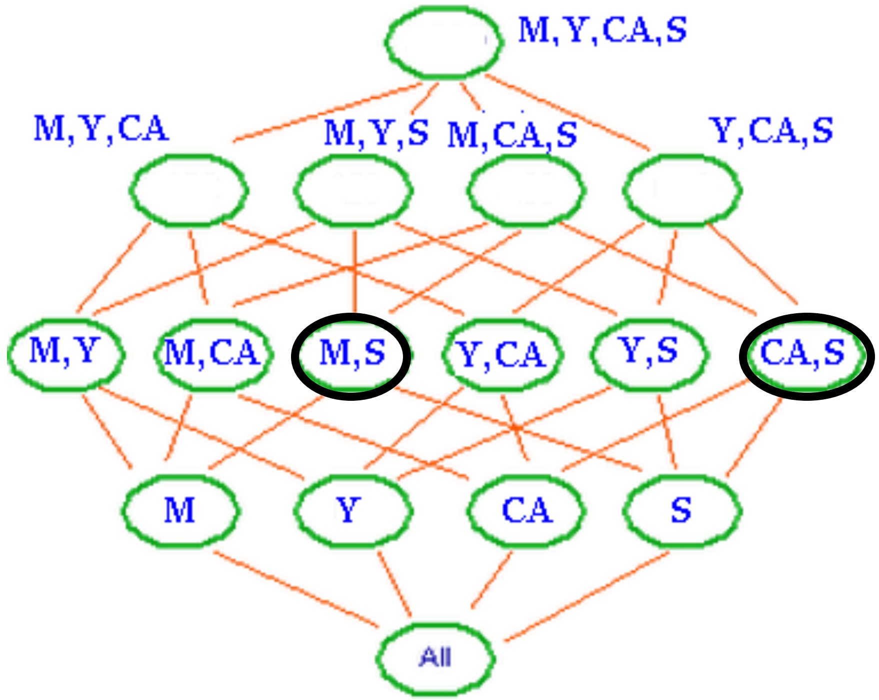


C= {Male, Student}
{California, Student}

RHE-DEM Algorithm

Satisfy Coverage ☐

Minimize Error ☐



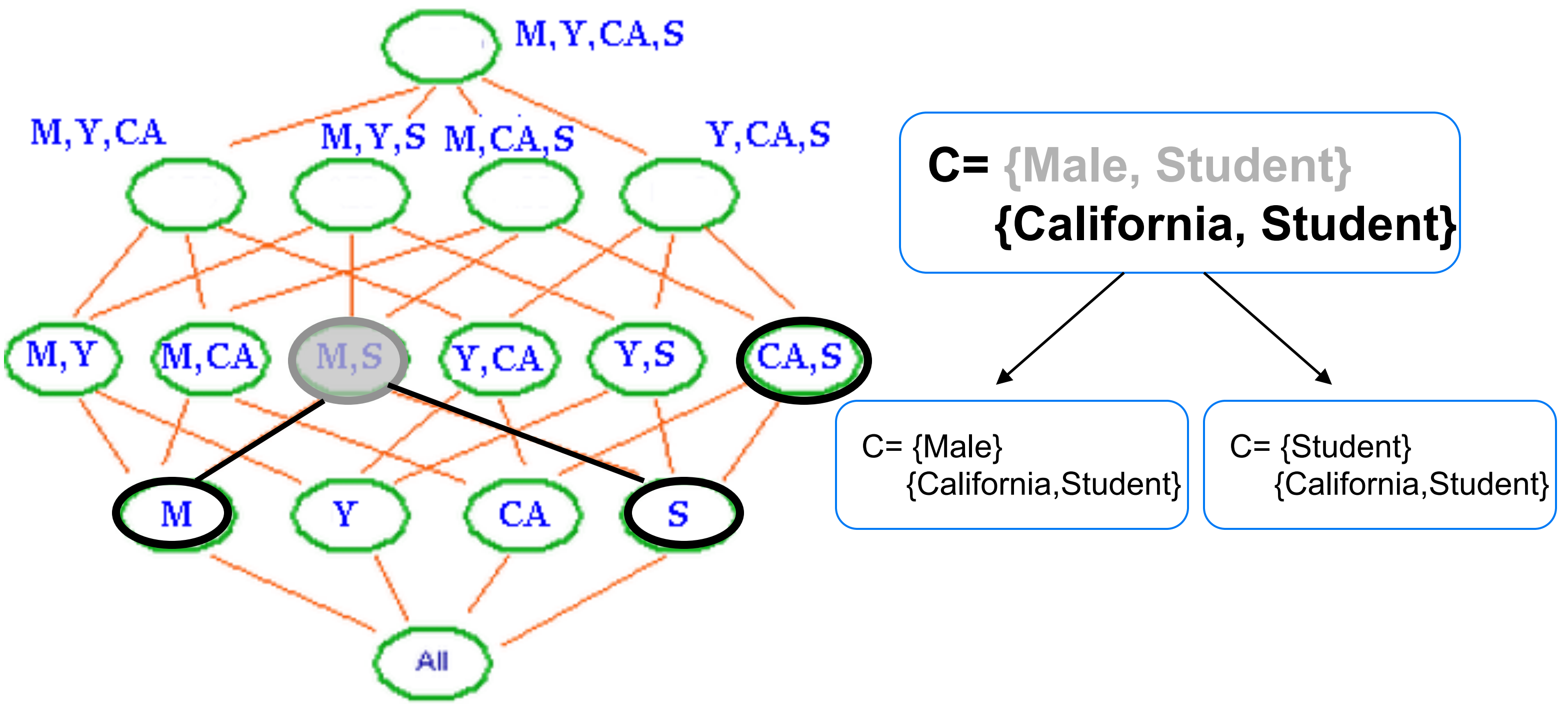
C= {Male, Student}
{California, Student}

Say, C does not satisfy Coverage Constraint

RHE-DEM Algorithm

Satisfy Coverage

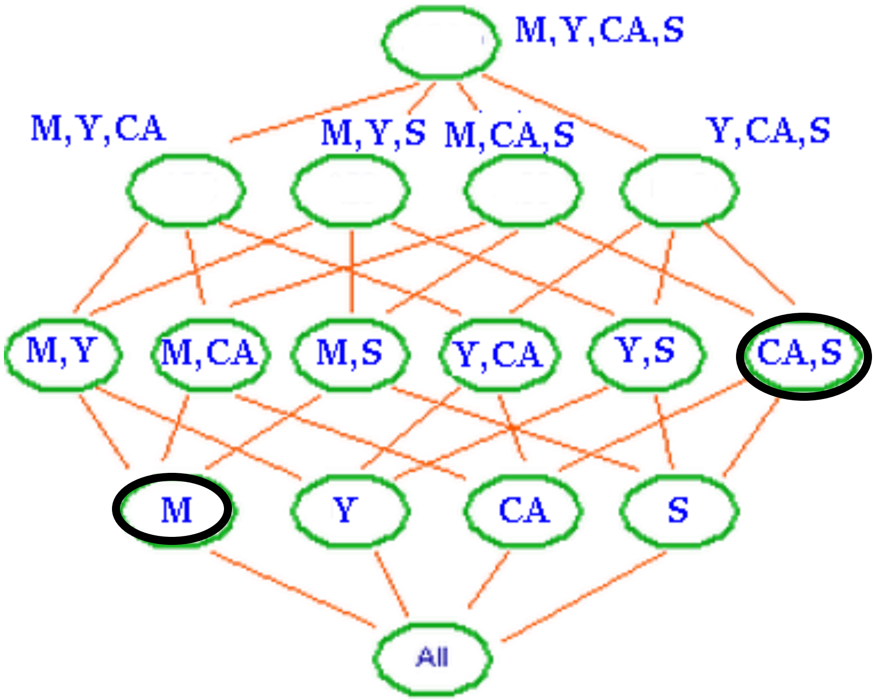
Minimize Error



RHE-DEM Algorithm

Satisfy Coverage √

Minimize Error

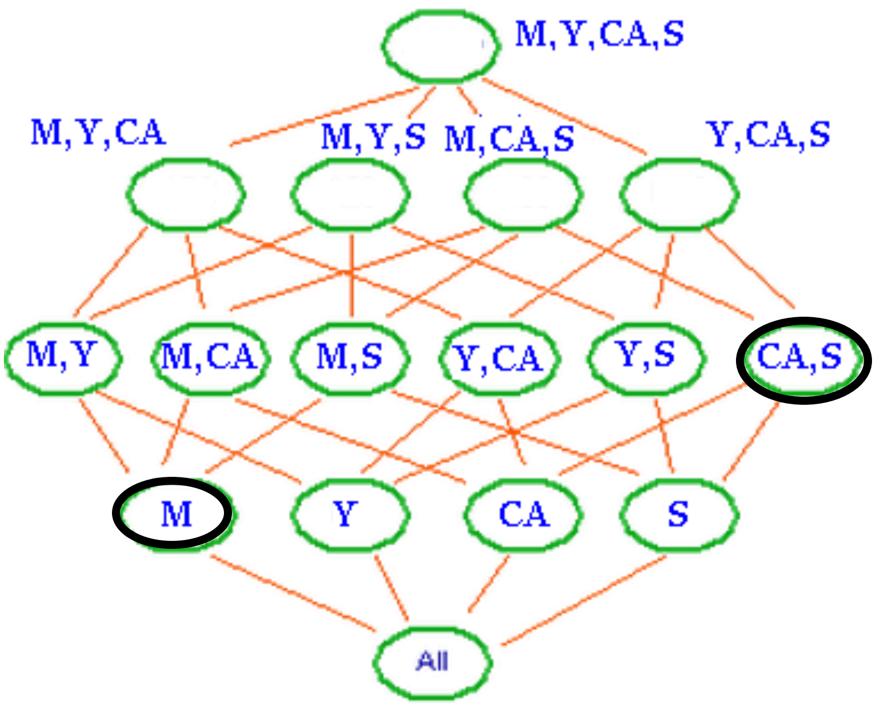


C= {Male}
{California, Student}

Say, C satisfies Coverage Constraint

RHE-DEM Algorithm

Satisfy Coverage	√
Minimize Error	

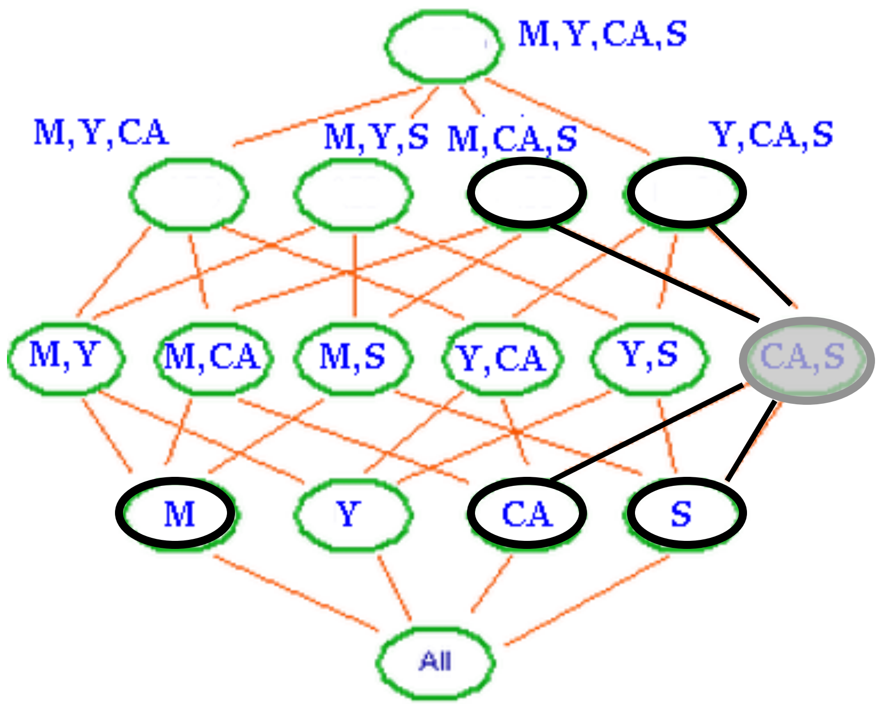


C= {Male}
{California, Student}

RHE-DEM Algorithm

Satisfy Coverage √

Minimize Error



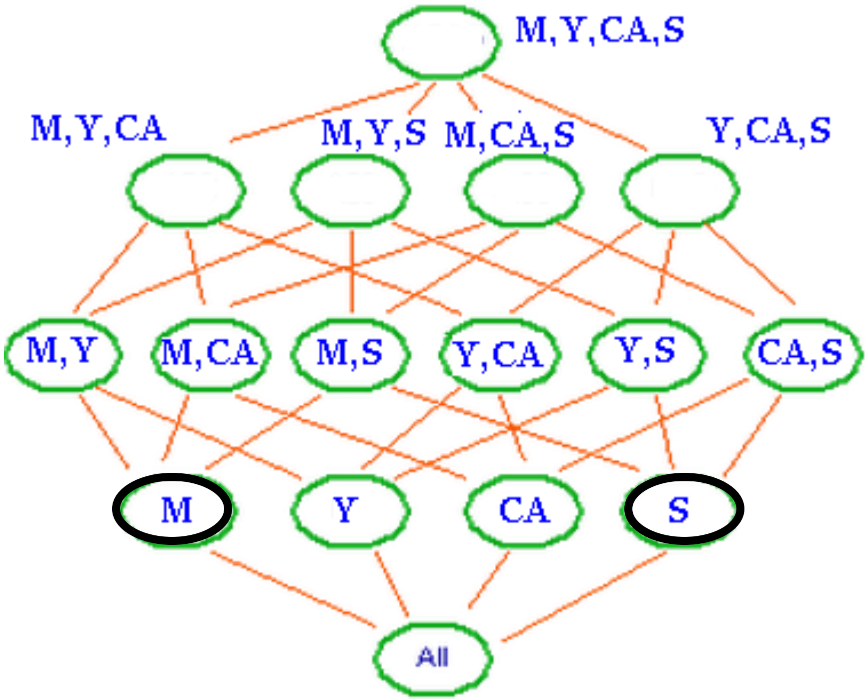
C= {Male}

{California, Student}

RHE-DEM Algorithm

Satisfy Coverage √

Minimize Error √



C= {Male}
{Student}

DIM: Meaningful Difference Mining

- For an input item covering R_I^+ R_I^- ratings, return set C of cuboids, such that:
  - difference balance $\boxed{\texttt{balance}(C, R_I^+, R_I^-)}$ is minimized, subject to:
    - $|C| \leq k$;
    - $\boxed{\texttt{coverage}(C, R_I^+)} \geq \alpha \cap \boxed{\texttt{coverage}(C, R_I^-)} \geq \alpha$

DIM: Meaningful Difference Mining

Difference Balance

Measures whether the positive and negative ratings are "mingled together" (high balance) or "separated apart" (low balance)

$$\mathtt{balance}(C, R_I^+, R_I^-) = m \times \sum_{r_1 \in R_I^+, r_2 \in R_I^-} I_{(r_1, r_2)}$$

where: $m = \frac{1}{|R_I^+| \times |R_I^-|}$, indicator $I_{(r_1,r_2)} = 1$ iff at least one cuboid in C covers r_1, r_2

Coverage

Measures the percentage of +/- ratings covered by returned groups

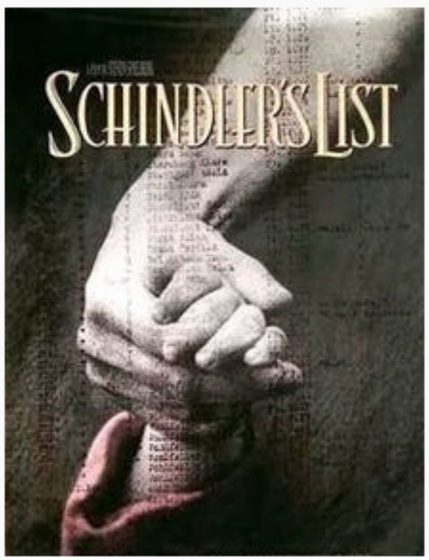
- DIM is NP-Hard: proof details in [1]

[1] S. Amer-Yahia, Mahashweta Das, Gautam Das, Cong Yu: MRI: Meaningful Interpretations of Collaborative Ratings,. In PVLDB 2011.

DIM: Meaningful Difference Mining

- Identify groups of reviewers who consistently disagree on item ratings

Schindler's List



Schindler's List (1993)

R 195 min - Biography | Drama | History - 15 December 1993 (USA)

8.9 Ratings: **8.9**/10 from 329,773 users Metascore: **93**/100

Reviews: 959 user | 95 critic | 23 from Metacritic.com

Teen-aged female reviewers and male middle-aged reviewers have rated this movie inconsistently; their average rating: 7.5

- **Middle-aged male reviewers love this movie, their average rating: 9.1**
- **Teen-aged female reviewers hate this movie, their average rating: 6.2**

Black Swan

Black Swan (2010)

R 108 min - Drama | Mystery | Thriller - 17 December 2010 (USA)

8.3

Ratings: **8.3**/10 from 156,148 users Metascore: **79**/100

Reviews: 892 user | 523 critic | 42 from Metacritic.com

Young female reviewers love this movie, average rating: 9.3

Reviewers from New York love this movie, average rating: 8.7

Young male student reviewers hate this movie, average rating: 6.1

DIM: Meaningful Difference Mining

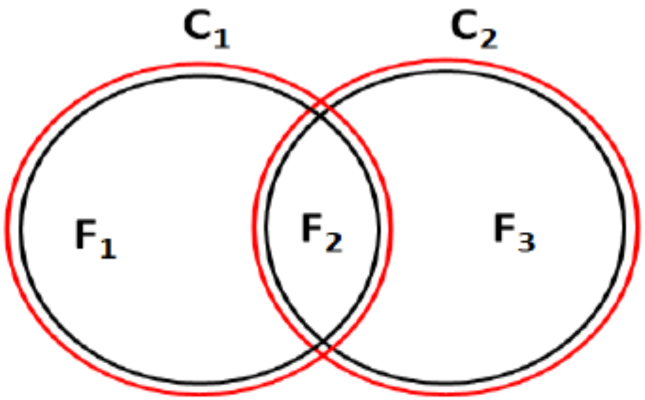
THEOREM 2. *The decision version of the problem of meaningful difference mining (DIM) is NP-Complete even for boolean databases.*

NP-Completeness: reduction of the Exact 3-Set Cover problem (EC3).

DIM Algorithms

- **Exact Algorithm (E-DIM)**
- **Randomized Hill Exploration Algorithm (RHE-DIM)**
  - Unlike DEM “error”, DIM “balance” computation is expensive
    - Quadratic computation scanning all possible positive and negative ratings for each set of cuboids
  - Introduce the concept of **fundamental regions** to aid faster balance computation
    - Each rating tuple is a k-bit vector where a bit is 1 if the tuple is covered by a group
    - A fundamental region is the set of rating tuples that share the same signature
    - Partition space of all ratings and aggregate rating tuples in each region

DIM Algorithms: Fundamental Region



C_1 = {Male, Student}

C_2 = {California, Student}

F	$C_1 C_2$	Count $F(R^+_I), F(R^-_I)$
F_1	1 0	40, 29
F_2	1 1	4, 2
F_3	0 1	2, 2

set of k=2 cuboids having 75 ratings (46+, 33-)

$$\mathtt{balance}(C, R_I^+, R_I^-) \;=\; m \times (\sum_i \mathtt{balance}(C, {R_I^+}_i, {R_I^-}_i) + \sum_{ij} \mathtt{balance}(C, {R_I^+}_{ij}, {R_I^-}_{ij})) \quad (1)$$

balance = $\frac{1}{46\times 33} \times (40\times 29 + 4\times 2 + 2\times 2 + (40\times 2 + 4\times 29) + (4\times 2 + 2\times 2))$

Summary of Rating Exploration

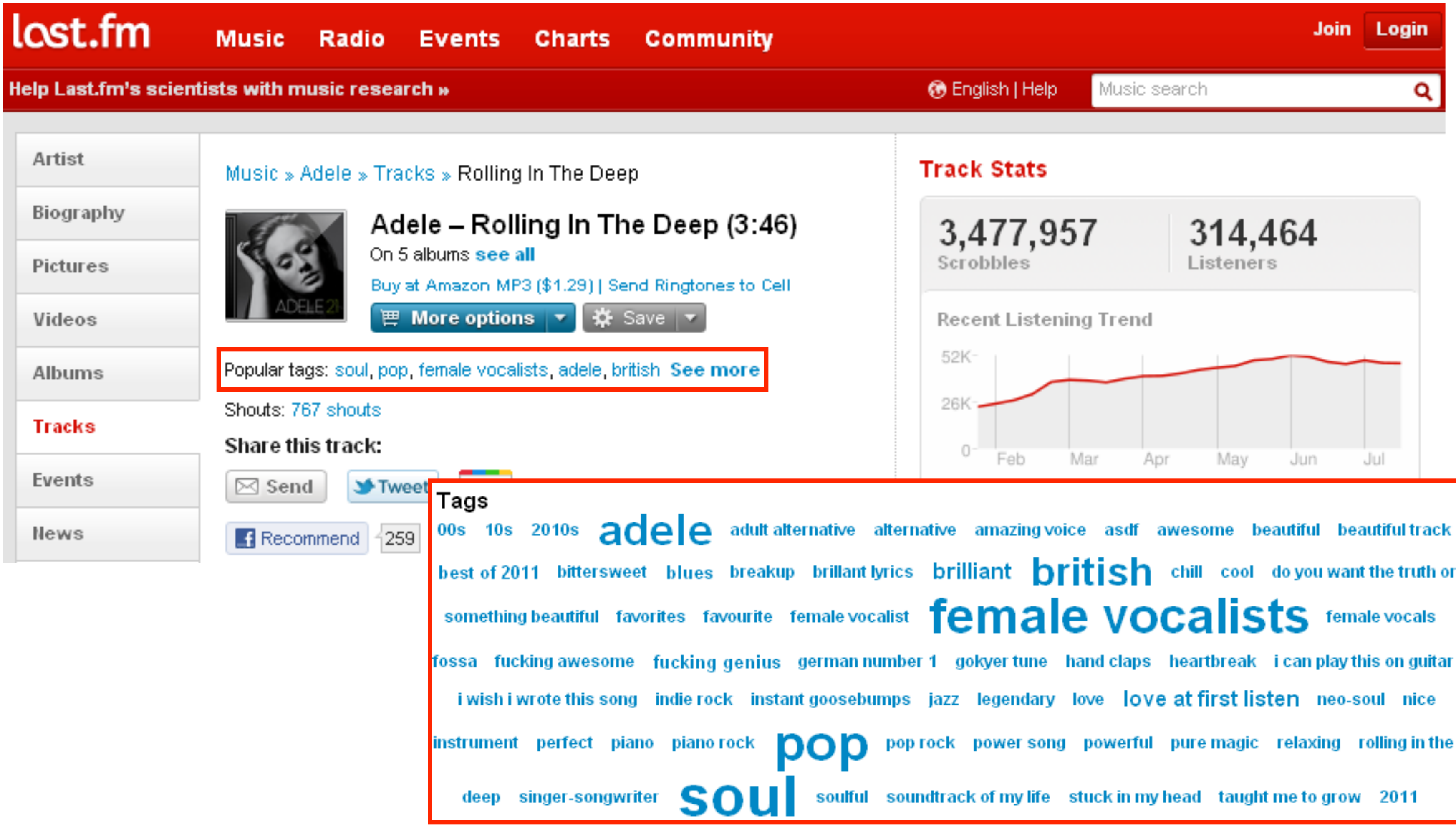
- DEM and DIM are hard problems
  - Leverage the lattice structure to improve coverage
  - Exploit properties of rating function for faster error computation
- Explore other rating aggregation functions
- Explore other constraints: e.g., group size
- Explore other optimization dimensions: group diversity

Social Data Exploration

- Rating exploration
  - MRI: Meaningful Interpretations of Collaborative Ratings
- **Tag exploration**
  - **Who Tags What? An Analysis Framework**
- Perspectives

Collaborative Tagging Site (Amazon)

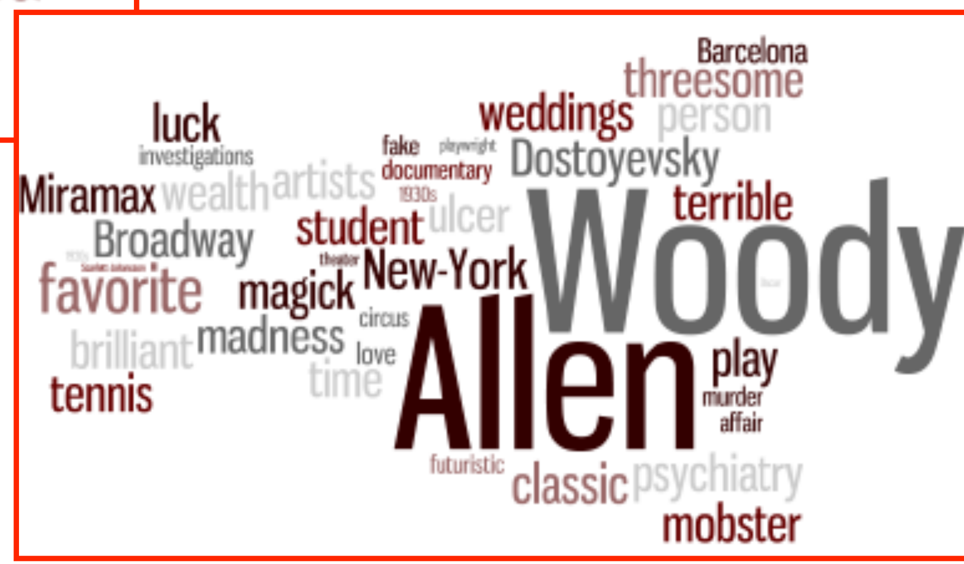
Collaborative Tagging Site (LastFM)



Exploring Collaborative Tagging in MovieLens

Tag Signature for all Users

Woody Allen



Tag Signature for all CA Users

Exploring Collaborative Tagging

- Exploration considers three dimensions
  - **User, Item, Tag**
- and two alternative measures
  - **Similarity, Diversity**

Data Model

- Tagging action tuple: <user attributes, item attributes, tags>

ID		Title	Genre	Director		Name	Gender	Location		Tags
1		Titanic	Drama	James Cameron		Amy	Female	New York		love, Oscar
2		Schindler's List	Drama	Steven Speilberg		John	Male	New York		history, Oscar

Tagging Behavior Dual Mining Problem (TagDM)

DEFINITION 4. **Tagging Behavior Dual Mining (TagDM) Problem.** *Given a triple* $\langle G, C, O \rangle$ *in the TagDM framework where* G *is the input set of tagging actions and* C, O *are the sets of constraints and optimization criteria respectively, the Tagging Behavior Dual Mining problem is to identify a set of tagging action groups,* $G^{opt} = \{g_1, g_2, \ldots\}$ *for* $b \in \{\texttt{users}, \texttt{items}, \texttt{tags}\}$ *and* $m \in \{\texttt{similarity}, \texttt{diversity}\}$, *such that:*

- $\forall g_x \in G^{opt}$, g_x *is user- and/or item-describable;*
- $k_{lo} \leq |G^{opt}| \leq k_{hi}$;
- $Support_G^{G^{opt}} \geq p$;
- $\forall c_i \in C, c_i.F(G^{opt}, b, m) \geq$ *threshold;*
- $\Sigma_{o_j \in O}, o_j.F(G^{opt}, b, m)$ *is maximized.*

Tagging Behavior Dual mining Problem Instance

PROBLEM 1. *Identify a set of tagging action groups,* $G^{opt} = \{g_1, g_2, \ldots\}$, *such that:*

- $\forall g_x \in G^{opt}$, g_x *is user- and/or item-describable;*
- $1 \leq |G^{opt}| \leq k$;
- $Support_G^{G^{opt}} \geq p$;
- $F_1(G^{opt}, \texttt{users}, \texttt{similarity}) \geq q$;
- $F_2(G^{opt}, \texttt{items}, \texttt{diversity}) \geq r$;
- $F_3(G^{opt}, \texttt{tags}, \texttt{similarity})$ *is maximized.*

Problem: Tagging Behavior Dual Mining (TagDM)

- Identify **similar** groups of reviewers who share **similar** tagging behavior for **diverse** set of items

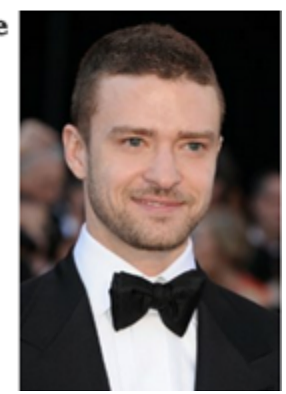
Male Young

Jennifer Aniston

comedy, drama, friendship

Justin Timberlake

drama, friendship



Tagging Behavior Dual Mining Instance

PROBLEM 4. *Identify a set of tagging action groups,* $G^{opt} = \{g_1, g_2, \ldots\}$, *such that:*

- $\forall g_x \in G^{opt}$, g_x *is user- and/or item-describable;*
- $1 \leq |G^{opt}| \leq k$;
- $Support_G^{G^{opt}} \geq p$;
- $F_1(G^{opt}, \mathtt{users}, \mathtt{diversity}) \geq q$;
- $F_2(G^{opt}, \mathtt{items}, \mathtt{similarity}) \geq r$;
- $F_3(G^{opt}, \mathtt{tags}, \mathtt{diversity})$ *is maximized.*

Tagging Behavior Dual Mining Instance

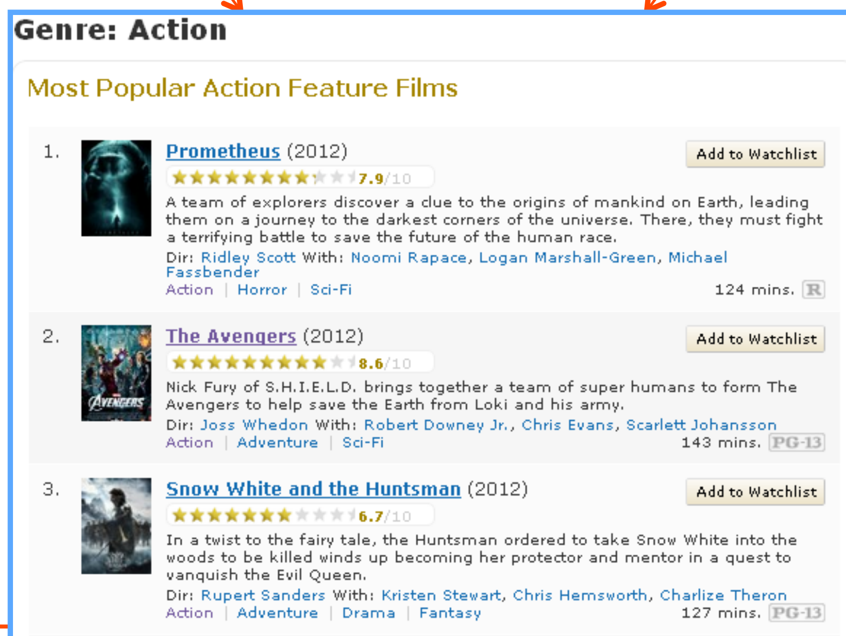
- Identify **diverse** groups of reviewers who share **diverse** tagging behavior for **similar** items

Male Teen

gun, special effects

Female Teen

violence, gory



TagDM is NP-Hard
(proof details in [2])

THEOREM 1. *The decision version of the TagDM problem is NP-Complete.*

PROOF. The membership of decision version of TagDM problem in NP is obvious. To verify NP-Completeness, we reduce Complete Bipartite Subgraph problem (CBS) to our problem and argue that a solution to CBS exists, *if and only if*, a solution our instance of TagDM exists. First, we show that the problem CBS is NP-Complete.

LEMMA 1. *Complete bipartite subgraph problem (CBS) is NP-Complete.*

[2] Mahashweta Das, Saravanan Thirumuruganathan, Sihem AmerYahia, Gautam Das, Cong Yu: Who Tags What? An Analysis Framework, In PVLDB 2012.

Two algorithms

- LSH (Locality Sensitive Hashing) based algorithm to handle TagDM problem instances optimizing similarity
- FDP (Facility Dispersion Problem) based algorithm handles TagDM problem instances optimizing diversity

- Both rely on computing tag signatures for groups
  - Latent Dirichlet Allocation to aggregate tags
  - Comparison between signatures based on cosine

Algorithm: LSH Based

- LSH (Locality Sensitive Hashing) based algorithm handles TagDM problem instances optimizing similarity
- LSH is popular to solve nearest neighbor search problems in high dimensions
- LSH hashes similar input items into same bucket with high probability
  - We hash group tag signature vectors into buckets, and then rank the buckets based on the strength of their (tag) similarity
- **SM-LSH**
  - Returns a set of groups, $\leq$ k having maximum similarity in tagging behavior, measured by comparing distances between group tag signature vectors
- **SM-LSH-Fi:** Handles hard constraints by Filtering result of SM-LSH
- **SM-LSH-Fo:** Handles hard constraints by Folding them to SM-LSH

Algorithm: LSH Based

- Hashing function for **SM-LSH**
  - We use LSH scheme in [3] that employs a family of hashing functions based on cosine similarity

    $$cos(\theta(T_{rep}(g_x), T_{rep}(g_y))) = \frac{|T_{rep}(g_x).T_{rep}(g_y)|}{\sqrt{|T_{rep}(g_x)|.|T_{rep}(g_x)|}}$$

    where $T_{rep}(g)$ is the tag signature vector for group g

  - Probability of finding the optimal result set by SM-LSH is bounded by: (proof details in paper)

    $$P(G^{opt}) \geq 1 - \sum_{x,y \in [1,k]} [1 - \left(\frac{\theta(T_{rep}(g_x), T_{rep}(g_y))}{\pi}\right)^{d'}]$$

    where d' is the dimensionality of hash signatures (buckets)
  - We employ iterative relaxation to tune d' in each iteration (Monte Carlo randomized algorithm) so that post-processing of hash tables yields non-null result set

[3]: M. Charikar. Similarity estimation techniques from rounding algorithms. In STOC, 2002

Algorithm: LSH Based

- **SM-LSH-Fi**: Dealing with constraints by Filtering
  - For each hash table, check for satisfiability of hard constraints in each bucket, and then rank filtered buckets on tagging similarity
    - Often yields null results

- **SM-LSH-Fo**: Dealing with constraints by Folding
  - Fold hard constraints maximizing similarity as soft constraints into SM-LSH
  - Hash similar input tagging action groups (similar with respect to group tag signature vector and user and/or item attributes) into the same bucket with high probability

However, it is non-obvious how LSH hash functions may be inversed to account for dissimilarity while preserving LSH properties

Algorithm: FDP Based

- FDP (Facility Dispersion Problem) based algorithm handles TagDM problem instances optimizing diversity

- FDP problem locates facilities on a network in order to maximize distance between facilities
  - We find tagging groups maximizing diversity (distance) betweeen tag signature vectors

- We intialize a pair of facilities with maximum weight, and then add nodes with maximum distance to those selected, in each subsequent iteration [4]

[4]: S. S. Ravi, D. J. Rosenkrantz, and G. K. Tayi. Facility dispersion problems: Heuristics and special cases. In WADS, 2002

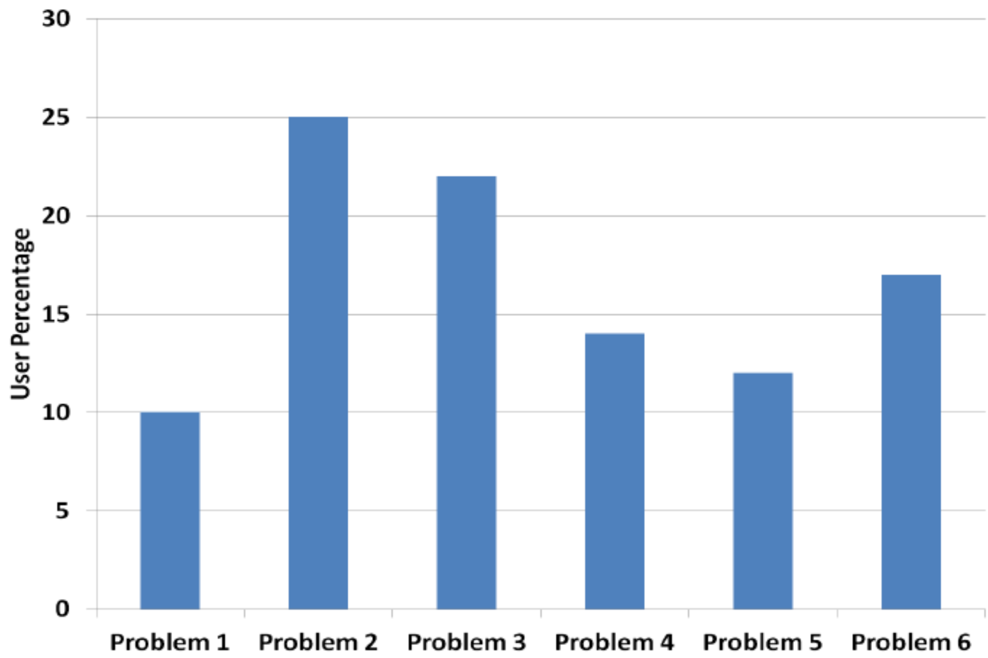
Algorithm: FDP Based

- **DV-FDP**
  - Returns a set of groups, $\leq$ k having maximum diversity in tagging behavior, measured by maximizing average pairwise distance between group tag signature vectors

  - If G^{opt} and G^{app} represent the set of k (k $\geq$ 2) tagging action groups returned by optimal and our approximate DV-FDP algorithm, and tag signature vectors satisfy triangular inequality: (Proof details in paper)

    $G^{opt}/G^{app} \leq 4$
- **DV-FDP-Fi**
  - Handles hard constraints by Filtering result of DV-FDP
- **DV-FDP-Fo**
  - Handles hard constraints by Folding them to DV-FDP

Some anecdotal evidence on analysts' prefs



Users prefer TagDM Problems ***2*** *(find similar user sub-populations who agree most on their tagging behavior for a* ***diverse*** *set of items),* ***3*** *(find* ***diverse*** *user sub-populations who agree most on their tagging behavior for a similar set of items) and* ***6*** *(find similar user sub-populations who* ***disagree*** *most on their tagging behavior for a similar set of items), having diversity as the measure for exactly one of the tagging component: item, user and tag respectively.*

Summary and Perspectives

- The notion of group is central to social data exploration
  - Because it is meaningful to analysts: groups are describable
  - Because group relationships can be explored for efficient space exploration

Perspective 1

- Rating exploration
  - A single dimension was optimized at a time (error or balance)
  - We could formulate a problem that seeks k most uniform (minimize error) and most diverse groups (least overlapping)

Perspective 2

- There is a total of **112** concrete problem instances that our TagDM framework captures!

ID	User	Item	Tag	C	O
1	similarity	similarity	similarity	U,I	T
2	similarity	diversity	similarity	U,I	T
3	diversity	similarity	similarity	U,I	T
4	diversity	similarity	diversity	U,I	T
5	similarity	diversity	diversity	U,I	T
6	similarity	similarity	diversity	U,I	T

Concrete Problem Instantiations. Column C lists the constraint dimensions Column O lists the optimization dimensions.

- And those optimize the tagging dimensions only

Perspective 3

- Social data exploration over time