Estimating Entity Importance via Counting Set Covers

Aristides Gionis (Yahoo! Research)
Theodoros Lappas (Boston University)
Evimaria Terzi (Boston University)

BOSTON UNIVERSITY

Given a set of entities, which are the most important entities to show to the user?

evimaria@cs.bu.edu

Review Management Systems

- Amazon: products with > 30,000 reviews
- Yelp: restaurants with > 3000 reviews

Given a set of reviews about a product, which reviews should be shown to the user?

evimaria@cs.bu.edu

Expertise Management Systems

- Odesk: > 120 million experts
- LinkedIn: > 1 million experts
- Guru: > hundreds of thousands of experts

Given a set of experts, which ones should be selected to perform a task?

evimaria@cs.bu.edu

Existing Paradigms

Given a set of entities, which are the most important entities to show to the user?

Ranking

Coverage

evimaria@cs.bu.edu

Rank Reviews by Helpfulness

Most Helpful Customer Reviews

1,313 of 1,333 people found the following review helpful:

★★★★☆ **Solid ultracompact camera**, March 8, 2008

By Garrett Lowenthal (San Francisco, CA) - See all my reviews
VINE™ VOICE

638 of 659 people found the following review helpful:

★★★★★ **A terrific pocket camera**, March 9, 2008

By Julie Neal (Sanibel Island, Fla.) - See all my reviews
TOP 100 REVIEWER VINE™ VOICE REAL NAME

216 of 222 people found the following review helpful:

★★★★★ **Perfect for me.**, March 10, 2008

By AZ Desert Rat "movie buff" - See all my reviews
VINE™ VOICE

103 of 107 people found the following review helpful:

★★★★★ **Amazon, Amazon, reviewers y'all, tell me which CanonSD is the fairest of all?**, March 24, 2008

By Anjana Nigam (Minneapolis, MN) - See all my reviews
VINE™ VOICE TOP 100 REVIEWER REAL NAME™

40 of 40 people found the following review helpful:

★★★★★ **perfect ultra compact model**, April 2, 2008

By Mark Twain "me" - See all my reviews

This review is from: Canon PowerShot SD1100IS 8MP Digital Camera with 3x Optical Image Stabilized Zoom (Brown) (Electronics)

Canon PowerShot SD1100IS 8MP Digital Camera with 3x Optical Image Stabilized Zoom (Brown)

Rank Experts by Experience

Alice

algorithms(60)

Bob

algorithms(4), python(30)

Cynthia

graphics(4), java (5)

David

graphics (2)

Eleanor

graphics (4), java (3), python (50)

evimaria@cs.bu.edu

Ranking

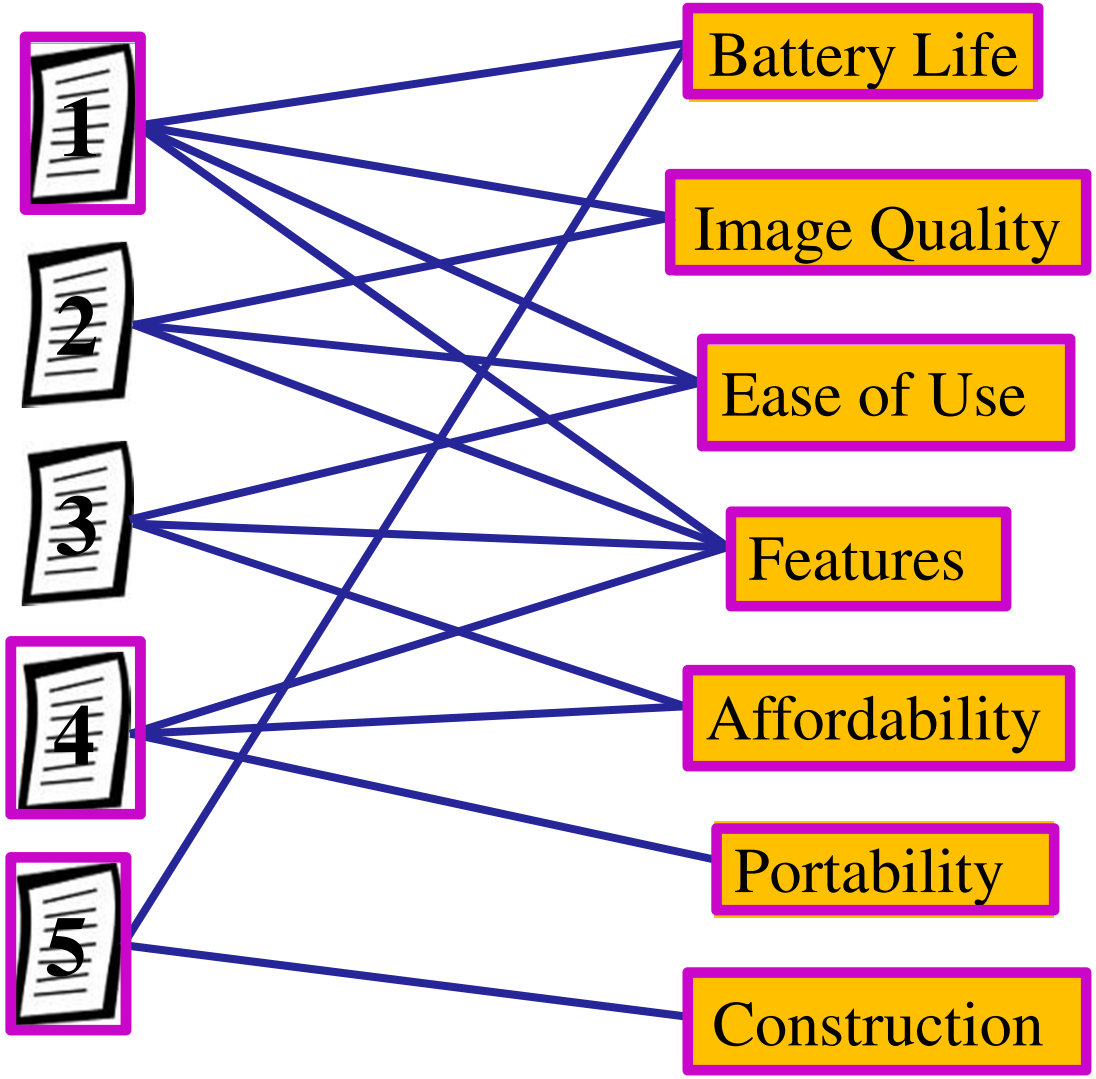
- Democratic and seniority respecting
  - Users vote for helpfulness
  - Experience means expertise

BUT

- Early reviews/ Early experts
- Mainstream reviews/ Mainstream skills
- Lacking **aspect** /**viewpoint** and **skill** **coverage**

evimaria@cs.bu.edu

Coverage of Reviews



Coverage of Experts

- T = {algorithms, java, graphics, python}

Alice	Bob	Cynthia	David	Eleanor
algorithms	algorithms,python	graphics, java	graphics	graphics,java,python

evimaria@cs.bu.edu

Coverage

- Guarantees coverage of viewpoints/skills
- Meritocracy: entities are judged by their marginal contribution

BUT

- Binary importance assignment to entities
- Many equally good subsets

evimaria@cs.bu.edu

Entity Ranking via Coverage

Evaluate the importance of individual entities based on the number of good set covers they participate in

evimaria@cs.bu.edu

Formally

- Universe: $U = \{u_1,\ldots,u_n\}$
- Entities: **C** $= \{E_1,\ldots,E_m\}$, with E_i subset of U
- **Set Cover**: **S** subset of **C** such that $\bigcup_{E \in S} E = U$

Task: for every E compute its **cover score**

$$R(E) = \sum_{S \in L_{SC}} \delta(E,S)\, w(S)$$

Uniform

Threshold

Cardinality-based

evimaria@cs.bu.edu

Complexity of computing cover scores

- Computing one (any) set cover is trivial
- Computing the minimum set cover is NP-hard
- For the cover scores we need to go over ***all*** (exponentially many) set covers.

evimaria@cs.bu.edu

Complexity of Computing Cover Scores

- Computing cover score for each entity is #P-hard
- Cover scores of entities can be approximated efficiently
- **Key Idea:** Counting instead of enumeration

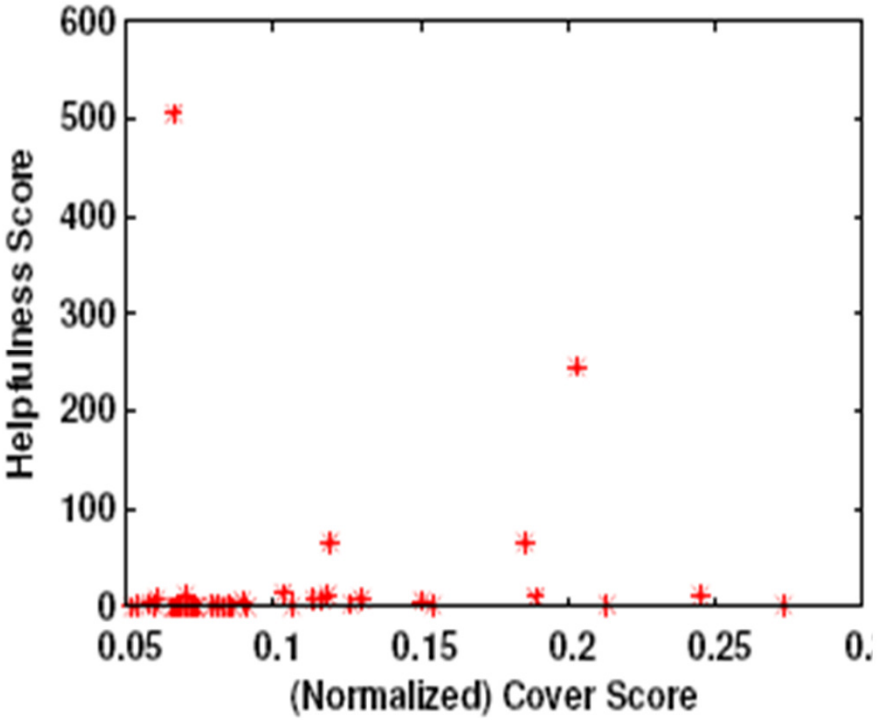
evimaria@cs.bu.edu

Computational Challenges

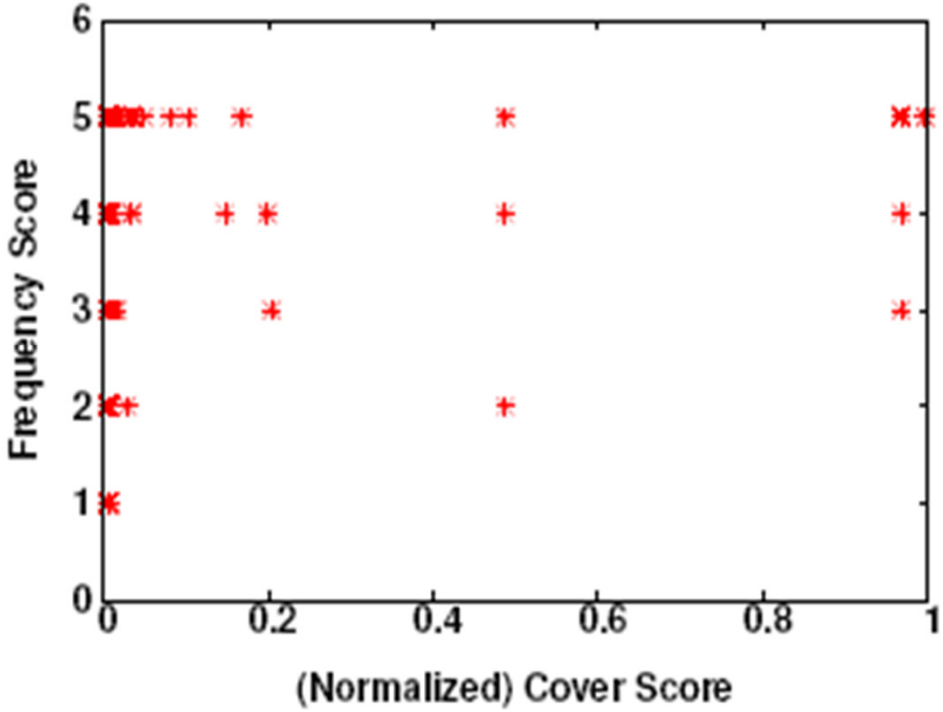
- Naïve Monte Carlo Counting needs exponential number of samples
- Adapt ImportanceSampling for counting satisfying assignments of DNF formulas
  - Compute the cover scores of all entities simultaneously
  - Compression of entities into super-entities
  - Decompose the problem into almost independent components

evimaria@cs.bu.edu

Experimental Results



GPS-reviews dataset

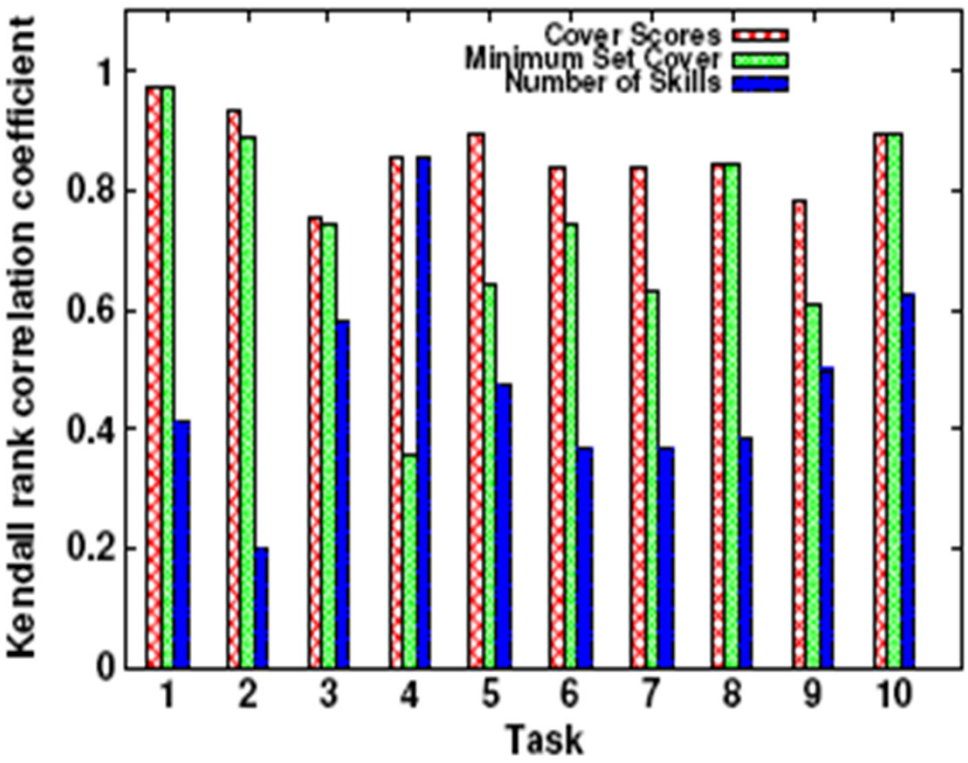


Guru experts dataset

evimaria@cs.bu.edu

User study

- Kendall –τ distance between human rankings and rankings obtained by cover scores, number of skills and minimum set covers.



evimaria@cs.bu.edu

Conclusions and Future Work

- This paper: Ranking via Coverage
- Future work: Coverage via Ranking
  - Select set covers that consist of important entities, (e.g., entities which participate in many set covers)

evimaria@cs.bu.edu