Algorithmic Problems in Review-Management Systems

References:

- T. Lappas, M. Crovella, E. Terzi: Selecting a Set of Characteristic Reviews. ACM SIGKDD International Conference on Data Mining and Knowledge Discovery, 2012.
- Panayiotis Tsaparas, Alex Ntoulas and Evimaria Terzi: Selecting a comprehensive set of reviews. ACM SIGKDD International Conference on Data Mining and Knowledge Discovery, 2011.

Online-Review Portals

User-generated content

Help customers make informed decisions

The Ecosystem of Review-Management Systems

Users-Customers:

- Read reviews to form opinions

Users-Reviewers:

- Write reviews to express opinions

Users-Merchants

- Receive reviews about their products and services

Problems

Customers

Information Overload

Reviewers

Motivation and Utilization

Merchants

Merchant Feedback

In this lecture

Customers

Information Overload

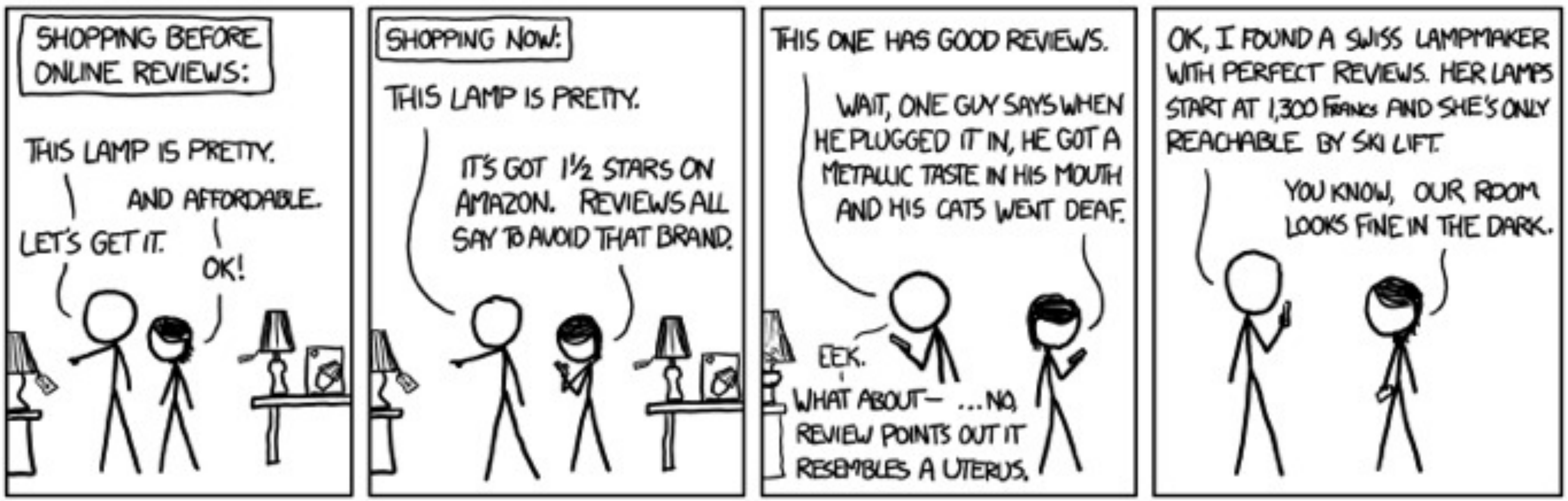
Reviewers

Motivation and Utilization

Merchants

Merchant Feedback

Information overload



Information Overload

Information Overload

Canon PowerShot SD1400IS 14.1 MP Digital Camera with 4x Wide Angle Optical Image Stabilized Zoom and 2.7-Inch LCD (Black)

Buy new: ~~$199.00~~ **$178.95**

65 new from $168.99

14 used from $139.00

Get it by **Monday, Aug 8** if you order in the next **15 hours** and choose one-day shipping.

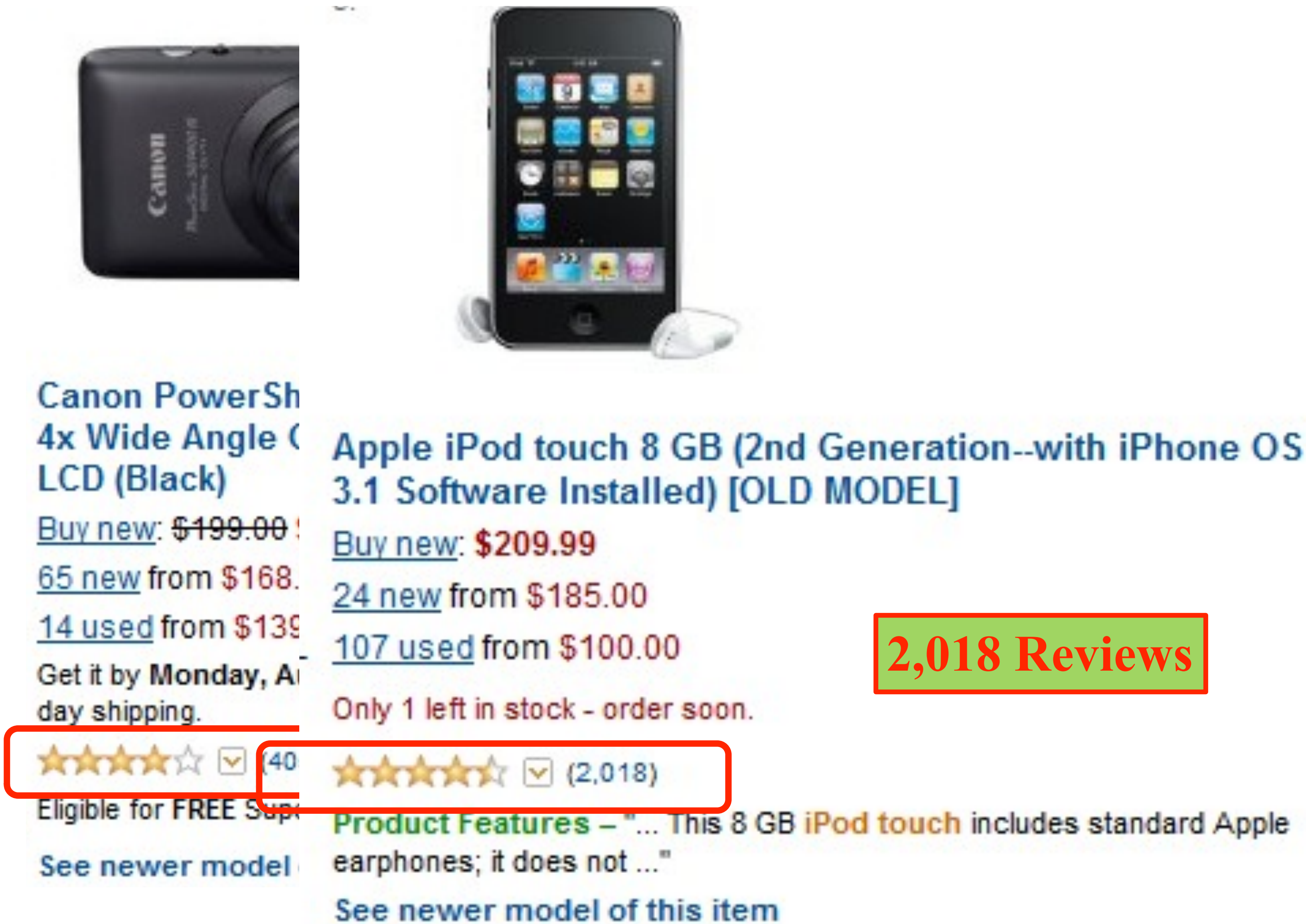
☆☆☆☆☆ (405)

Eligible for **FREE** Super Saver Shipping.

See newer model of this item

405 Reviews

Information Overload



Information Overload

Review 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 by helpfulness

Democratic

- Users vote for ranking

Biased

- Early reviews
- Mainstream reviews
- Lacking **aspect** and **viewpoint** **coverage**

Customer Reviews

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

Average Customer Rating

(938 customer reviews)

5 star:	(647)
4 star:	(208)
3 star:	(34)
2 star:	(14)
1 star:	(35)

Battery life	(3)
Construction quality	(3)
Ease of use	(3)
Features	(3)

› See and rate all 6 attributes.

Create your own review

The Most Helpful Reviews

The most helpful favorable review

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

Solid ultracompact camera

If you need a solid, reliable, and stylish point-and-shoot ultracompact digital camera that produces high-quality images, then the new Canon PowerShot SD1100IS may be right for you.

I am an advanced amateur photographer and own 2 Canon digital cameras (G2 and 20D). Both have served me well over the years but recently I have found myself needing a...

Read the full review ›

Published on March 8, 2008 by Garrett Lowenthal

› See more 5 star, 4 star reviews

Vs.

The most helpful critical review

164 of 183 people found the following review helpful:

The lens error problem is for real....

I got this camera for my daughter (in pink of course) in mid-April. She loves it (size, pictures, etc.) but after less than three months it will only flash "lense error, restart" when it's turned on. Too late to return to Amazon. :(On the bright side, a trip to Canon's website support section got me through to a Repair Request Confirmation. Hopefully, this will just cost...

Read the full review ›

Published on July 12, 2008 by D. Pate

› See more 3 star, 2 star, 1 star reviews

Talk outline

Information Overload -- Coverage

Motivation > Model > Algorithms > Results

Information Overload – Summarization

Motivation > Model > Algorithms > Results

Conclusions

Our goal

Select **a small (size k) set of comprehensive reviews** of

High quality

High attribute coverage

High viewpoint coverage

The Model

The Model

The Model

Item attributes

Battery Life

Image Quality

Ease of Use

Features

Affordability

Portability

Construction

The Model

Reviews

Item attributes

Battery Life

Image Quality

Ease of Use

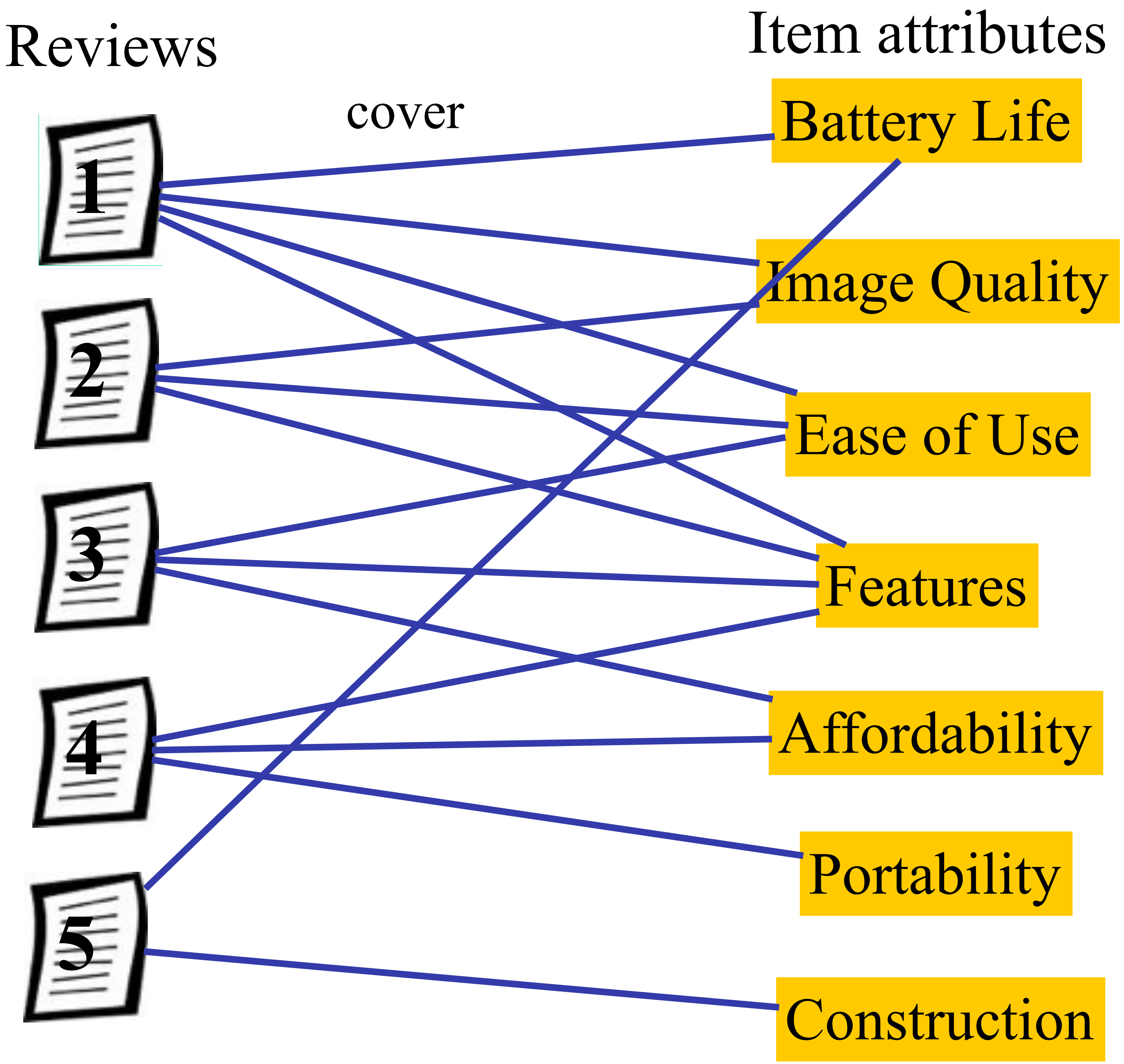
Features

Affordability

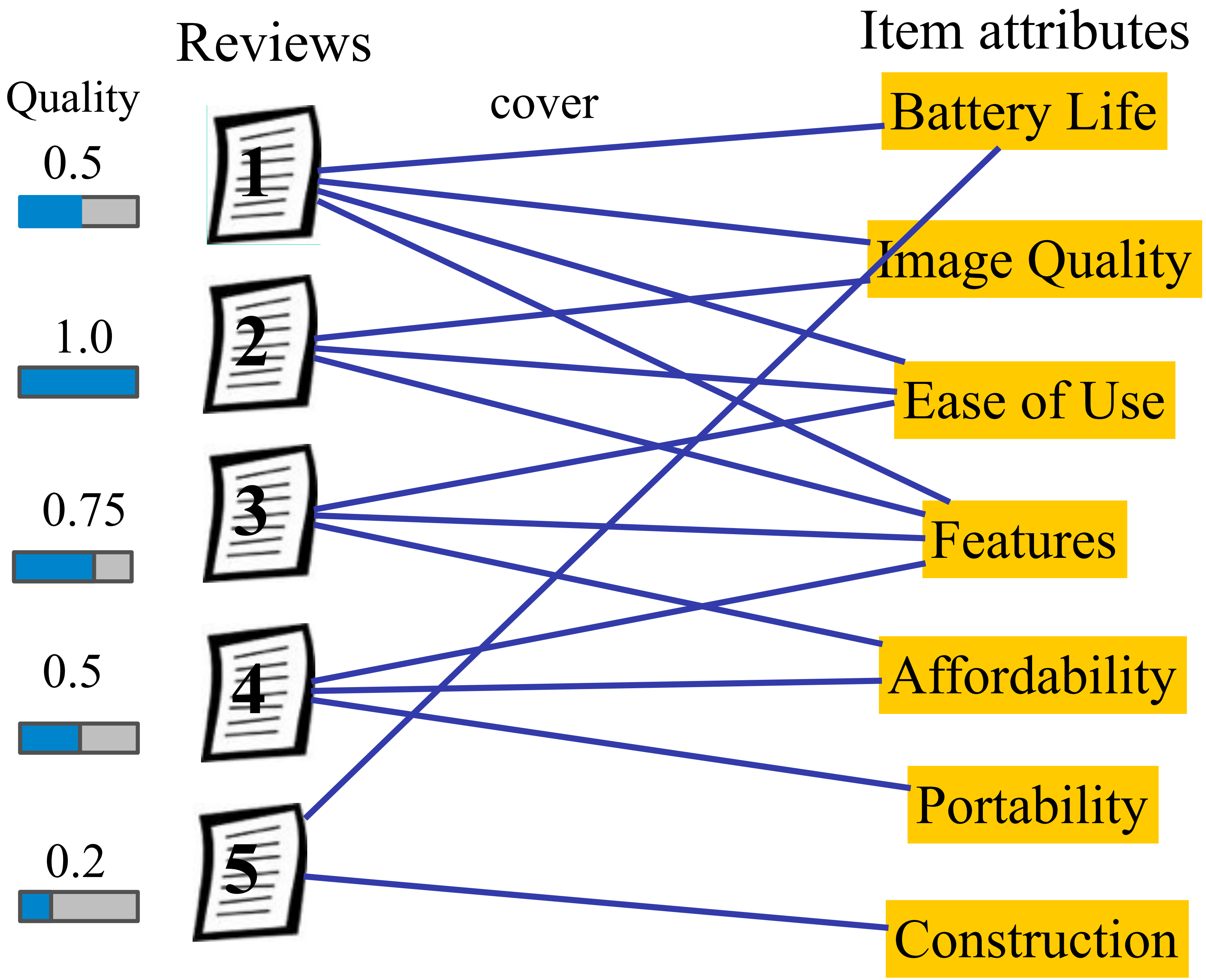
Portability

Construction

The Model

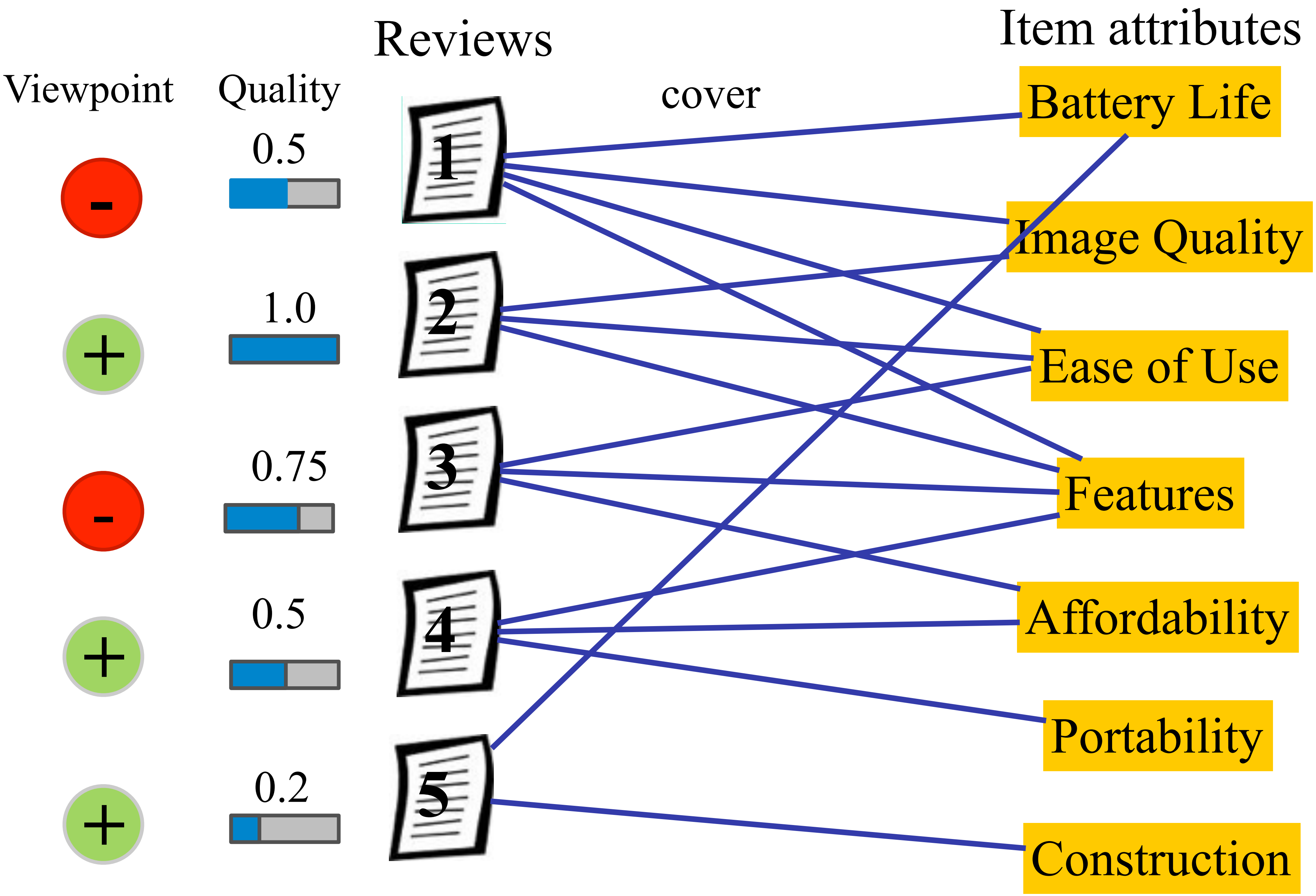

Reviews
Item attributes
cover
1
2
3
4
5
Battery Life
Image Quality
Ease of Use
Features
Affordability
Portability
Construction

The Model



Reviews
Item attributes
Quality
cover
0.5
1.0
0.75
0.5
0.2
1
2
3
4
5
Battery Life
Image Quality
Ease of Use
Features
Affordability
Portability
Construction

The Model


Viewpoint
Quality
Reviews
cover
Item attributes
0.5
1.0
0.75
0.5
0.2
-
+
-
+
+
1
2
3
4
5
Battery Life
Image Quality
Ease of Use
Features
Affordability
Portability
Construction

Our goal

Select **a small (size k) set of comprehensive reviews** of

High quality

High attribute coverage

High viewpoint coverage

General Coverage Problem

How good is a subset of reviews ***S***?

For attribute ***a***:

c(S,a) quantifies how well ***S*** covers ***a***

Coverage Function:

$$F(S) = \sum_{a \in A} c(S, a)$$

General Coverage Problem

Given a collection of reviews select a set of ***k*** reviews ***S*** such that ***F(S)*** is maximized

$$F(S) = \sum_{a \in A} c(S, a)$$

Need to define function ***c(S,a)***

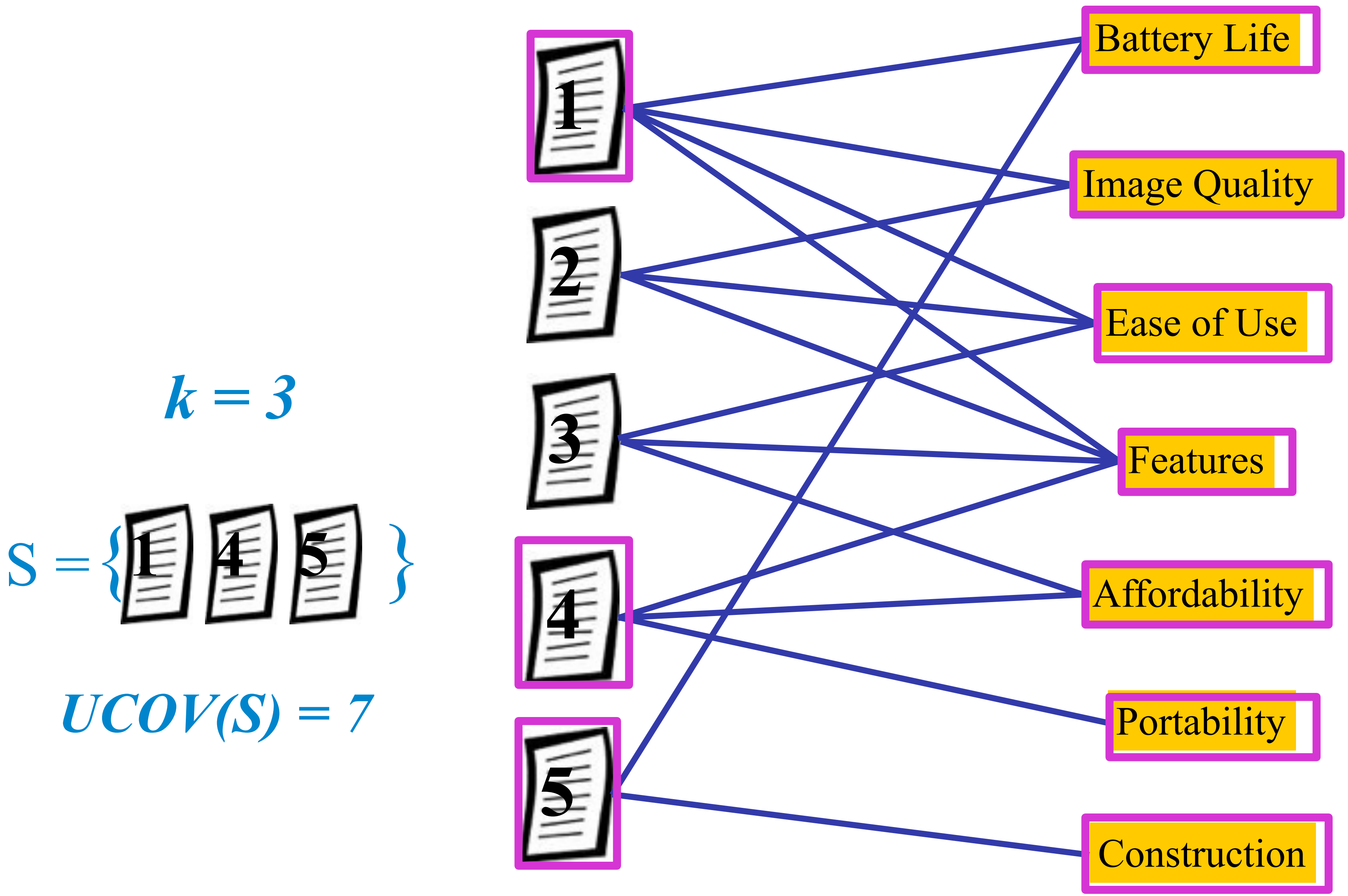
Unit Coverage Problem

$c_u(S,a)=1$ if S covers a

$$\mathrm{UCOV}(S) = \sum_{a \in A} c_u(S,a)$$

Given a collection of reviews select a set of k reviews S such that $UCOV(S)$ is maximized

Unit Coverage



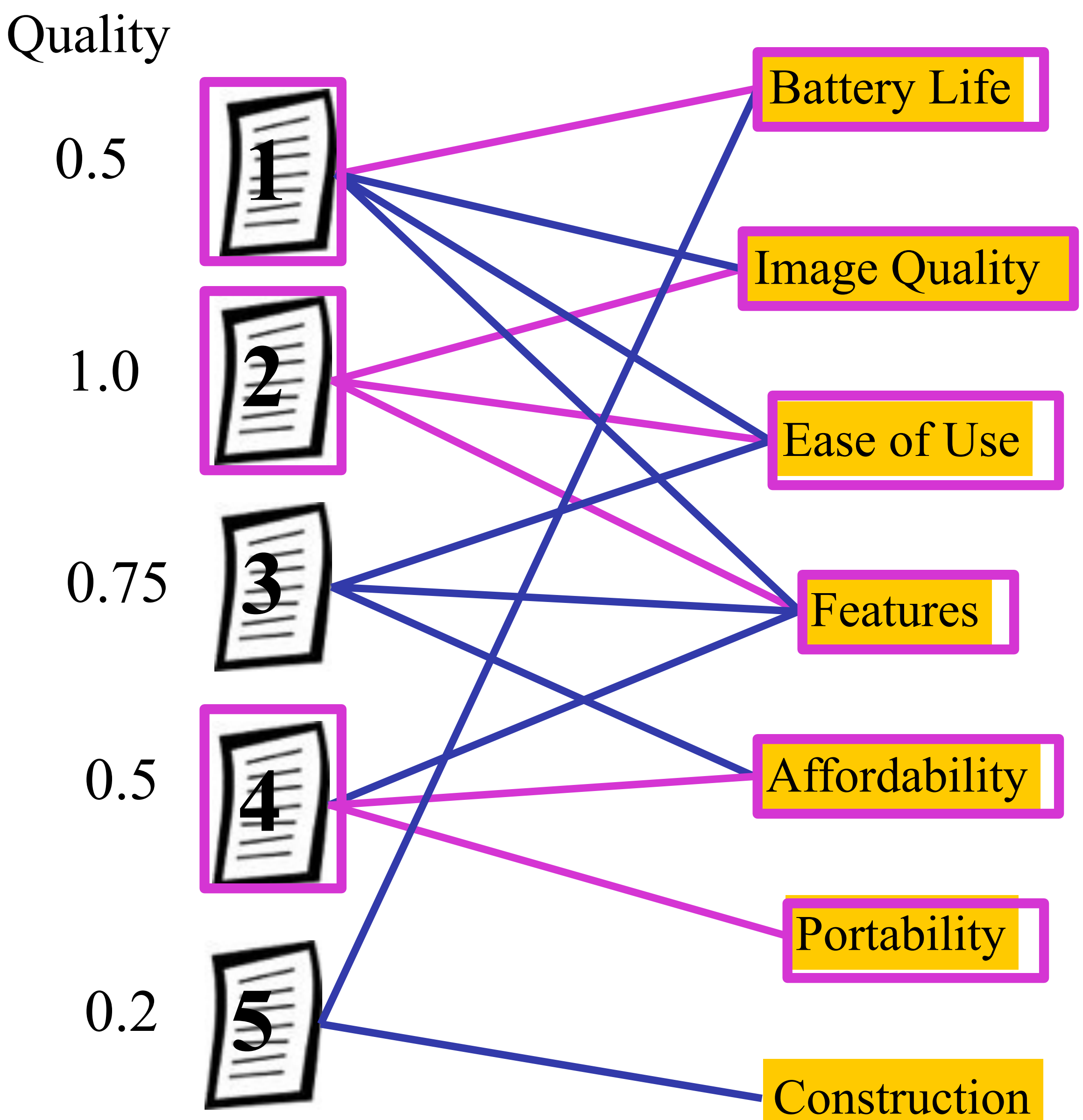
Quality Coverage Problem

$c_q(S,a)$*:*max quality among reviews in ***S*** that cover ***a***

$$\mathrm{QCOV}(S) = \sum_{a \in A} c_q(S,a)$$

Given a collection of reviews select a set of ***k*** reviews ***S*** such that ***QCOV(S)*** is maximized

Quality Coverage



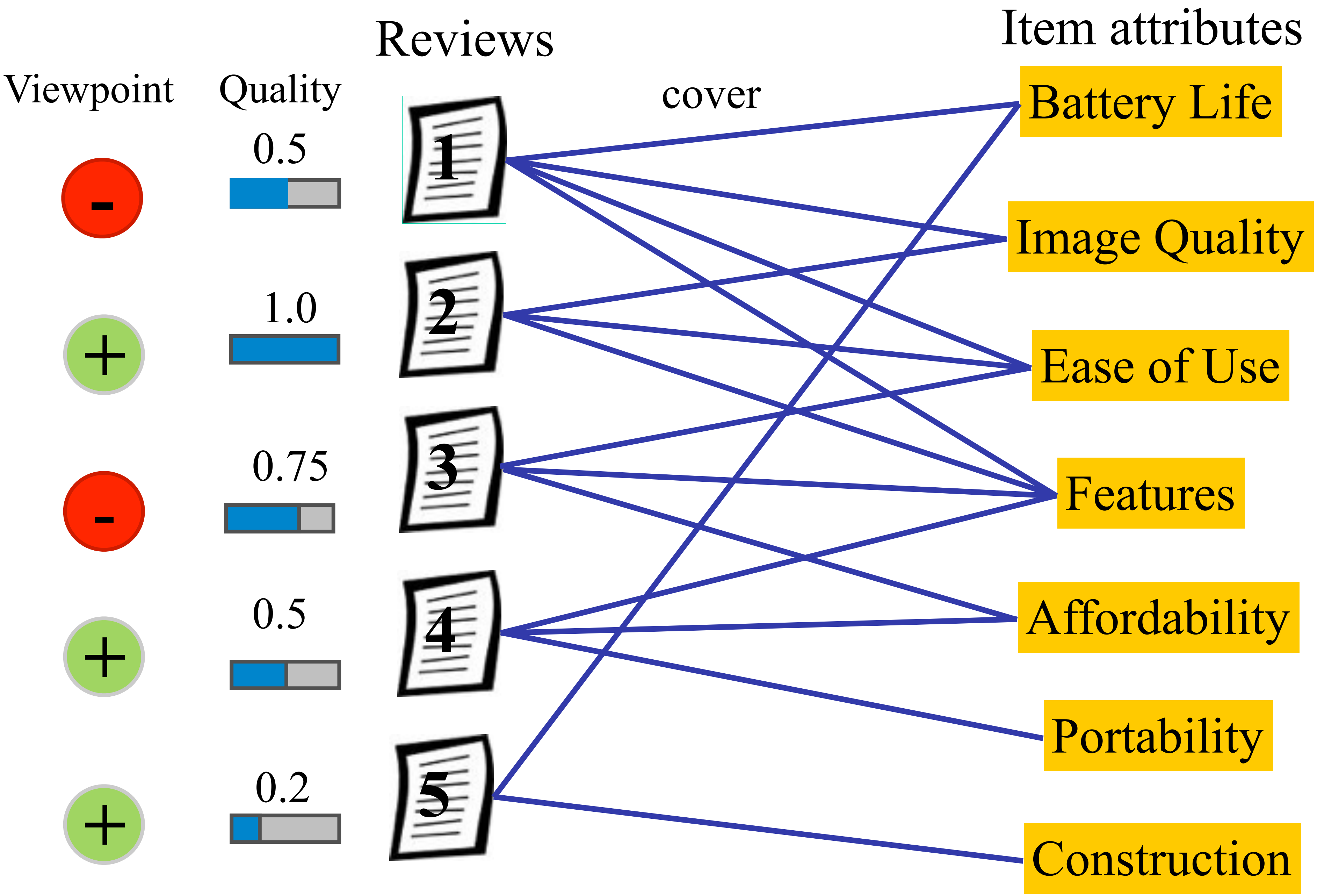
k = 3

$S = \{$ 

$\}$

QCOV(S) = 4.5

The Model



Reviews
Item attributes
Viewpoint
Quality
cover
-
0.5
1
+
1.0
2
-
0.75
3
+
0.5
4
+
0.2
5
Battery Life
Image Quality
Ease of Use
Features
Affordability
Portability
Construction

Our goal

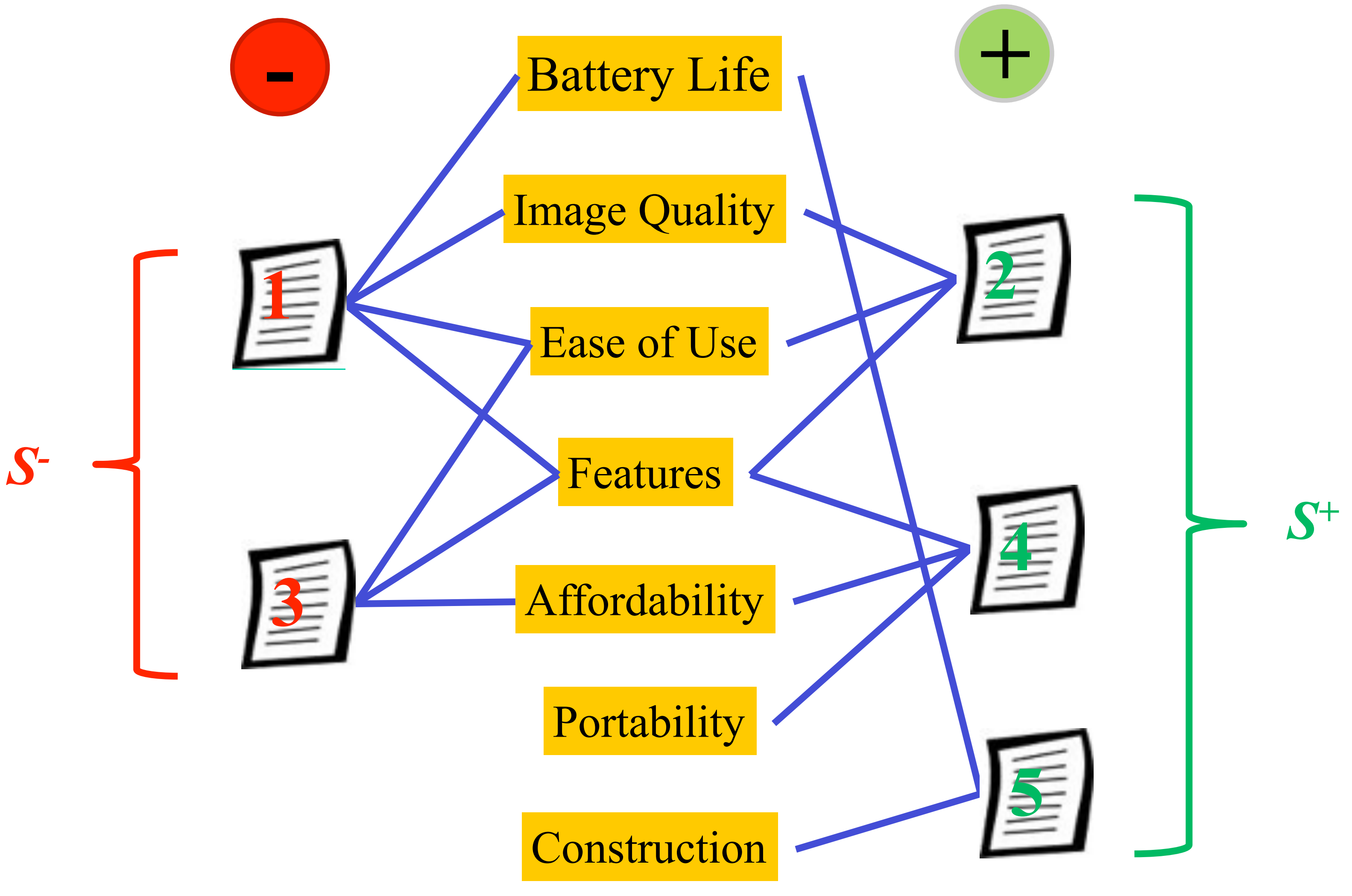
Select **a small (size k) set of comprehensive reviews** of

High quality

High attribute coverage

High viewpoint coverage

Group Coverage



Group Coverage Problem

$$c_g(S,a)=\min\{c(S^+,a),c(S^-,a)\}$$

$$\mathrm{GCOV}(S)=\sum_{a\in A} c_g(S,a)$$

Given a collection of reviews select a set of ***k*** reviews ***S*** such that ***GCOV(S)*** is maximized

Group Unit Coverage Problem

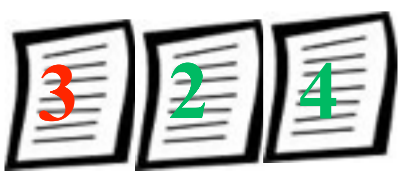
$$c_{gu}(S,a)=\min\{c_u(S^+,a),c_u(S^-,a)\}$$

$$\mathrm{GUCOV}(S)=\sum_{a\in A} c_{gu}(S,a)$$

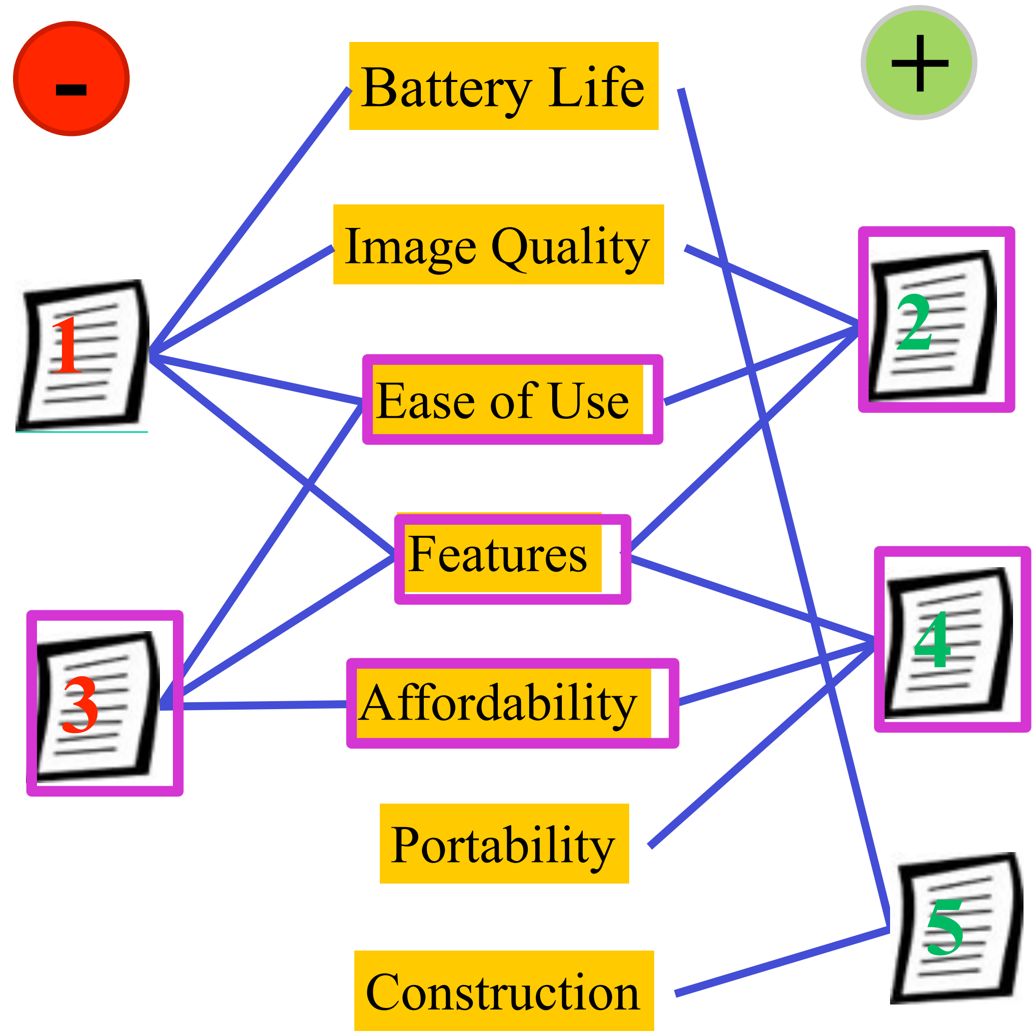
Given a collection of reviews select a set of ***k*** reviews ***S*** such that ***GUCOV(S)*** is maximized

Group Unit Coverage

***k* = 3**

S = {  }

***GUCOV(S)* = 3**



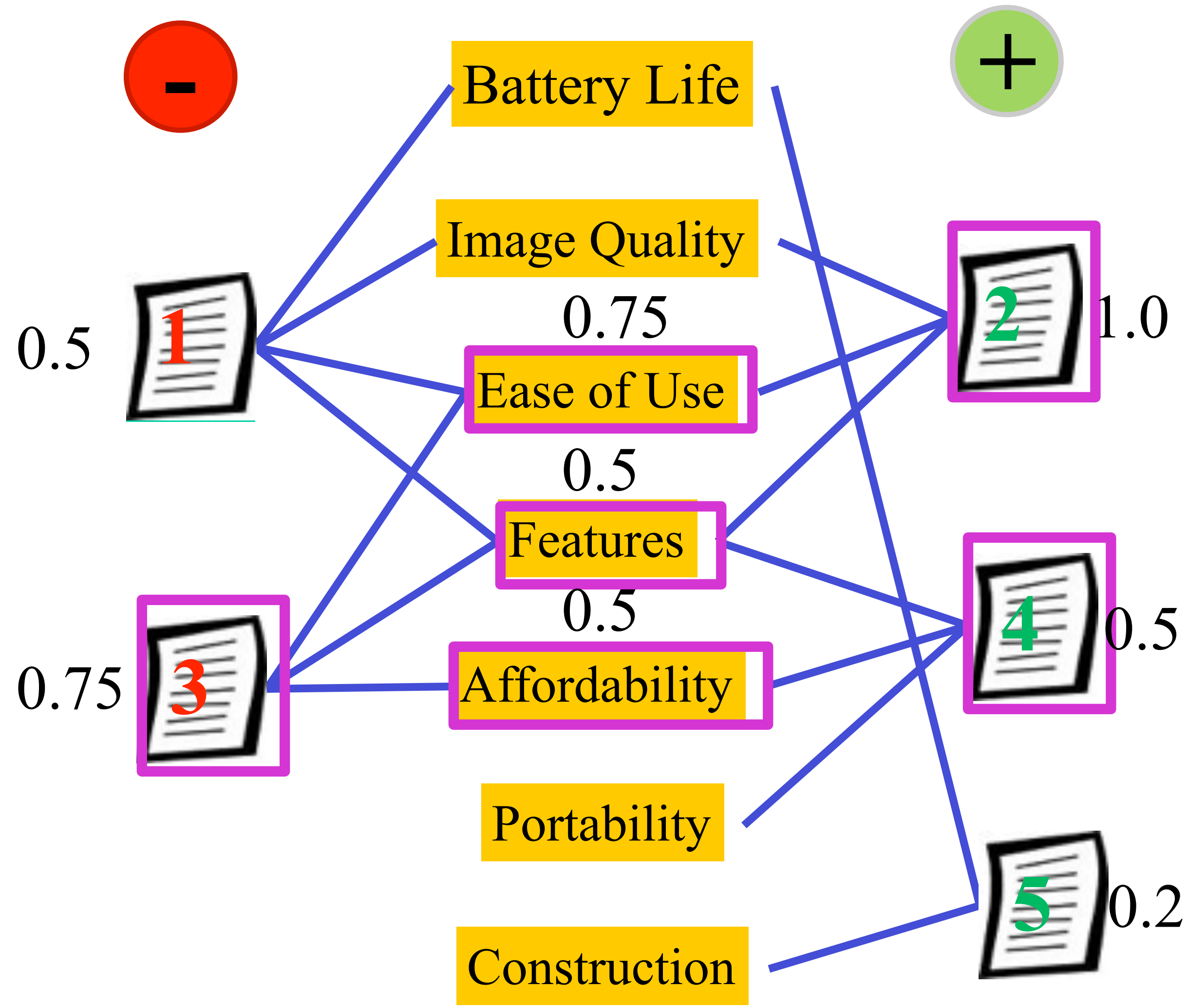
Group Quality Coverage Problem

$$c_{gq}(S,a)=min\{c_q(S^+,a),c_q(S^-,a)\}$$

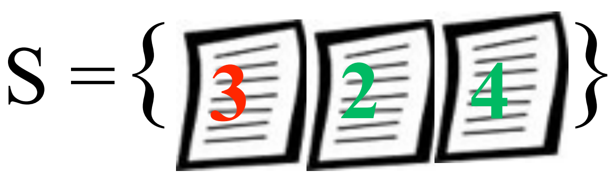
$$\text{GQCOV}(S) = \sum_{a \in A} c_{gq}(S,a)$$

Given a collection of reviews select a set of ***k*** reviews ***S*** such that ***GQCOV(S)*** is maximized

Group Quality Coverage



k* = *3



GQCOV(S) = 1.75

Soft Quality Coverage Problem

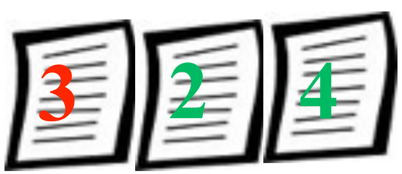
$$c_{sq}(S,a) = c_q(S^+,a) + c_q(S^-,a)$$

$$\mathrm{SQCOV}(S) = \sum_{a \in A} c_{sq}(S,a)$$

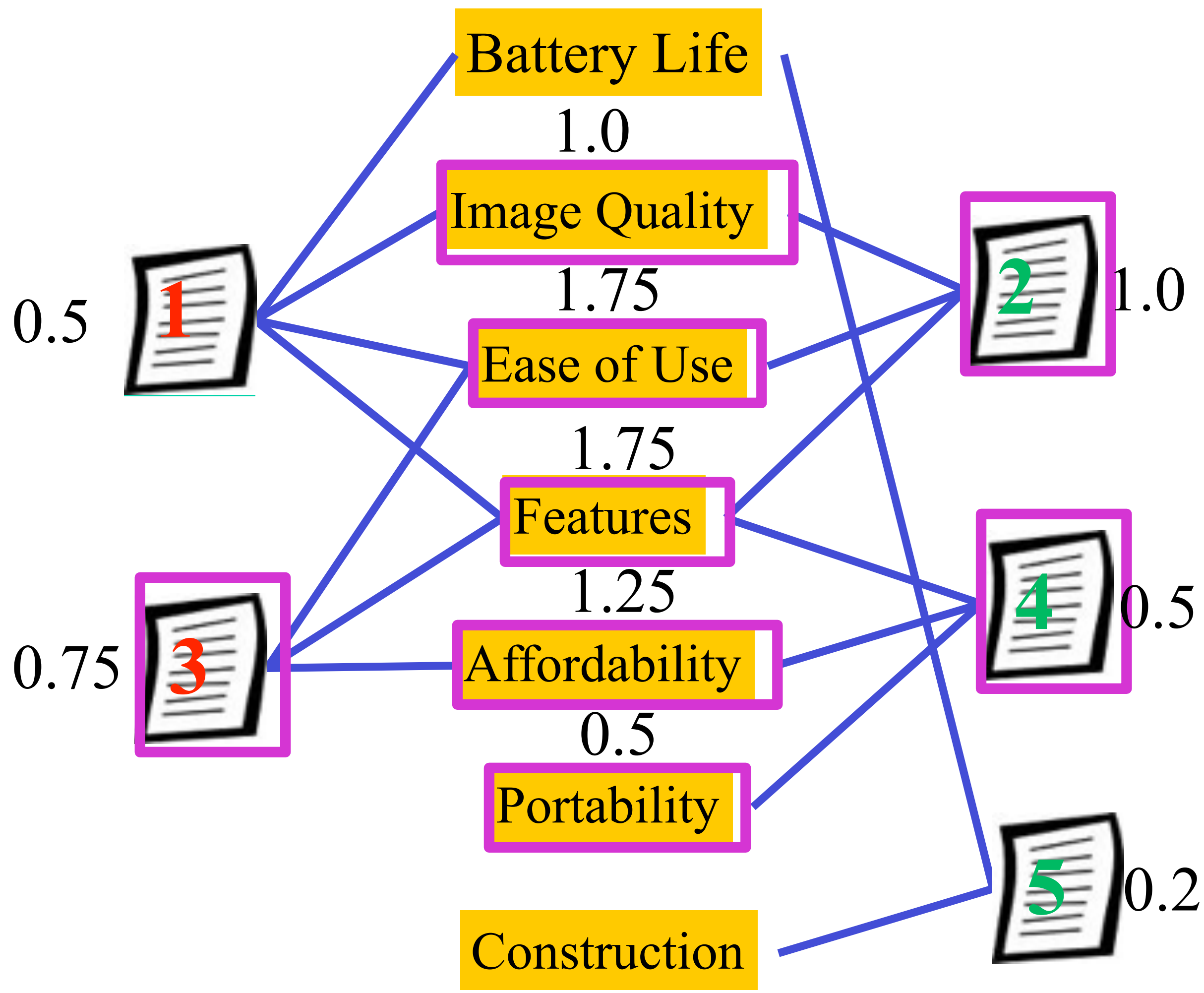
Given a collection of reviews select a set of ***k*** reviews ***S*** such that ***SQCOV(S)*** is maximized

Group Quality Coverage

k = 3

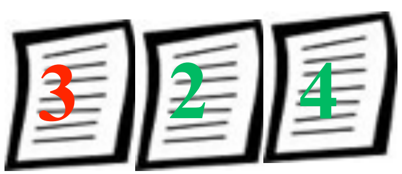
S = {  }

SQCOV(S) = 5.75

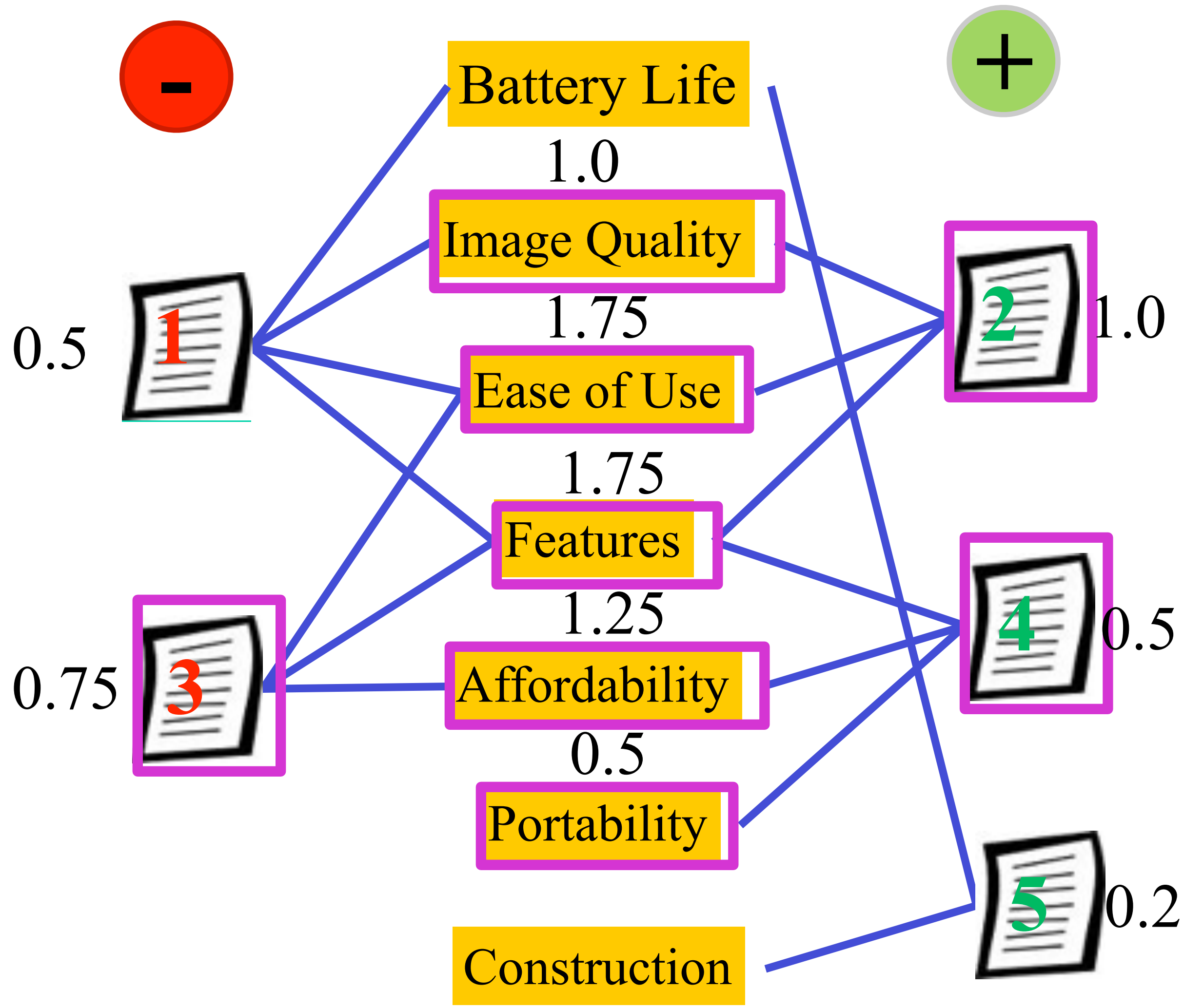


Group Quality Coverage

k = 3

S = {  }

SQCOV(S) = 5.75



Talk outline

Information Overload -- Coverage

Motivation > Model > Algorithms > Results

Information Overload – Summarization

Motivation > Model > Algorithms > Results

Conclusions

Analysis

All versions of the General Coverage problem are NP-hard

The UCOV, QCOV, SQCOV functions are submodular

For $X \subseteq Y$,

$$F(X \cup \{r\}) - F(X) \geq F(Y \cup \{r\}) - F(Y)$$

A simple Greedy algorithm is an (1-1/e) approximation to the optimal

The Greedy algorithm

$S = \emptyset$

While $|S| < k$

for each review r compute

$gain(r) = F(S \cup \{r\}) - F(S)$

$r^* = \text{argmax}_r\ gain(r)$

$S = S \cup \{r^*\}$

Group Coverage

Greedy algorithm does not work

An attribute **cannot** be covered with one review

Bad News: The GUCOV, GQCOV functions are **not** submodular

GreedyPairs: Greedy algorithm on pairs of reviews

The GreedyPairs algorithm

Compute the set P of all pairs of reviews from positive and negative groups

$S = \emptyset$

While $|S| < k$

for each pair p compute

$gain(p) = F(S \cup \{p\}) - F(S)$

$cost(p)$ = reviews in p not in S

$p^* = \text{argmax}_p\ gain(p)/cost(p)$

$S = S \cup \{p^*\}$

Talk outline

Information Overload -- Coverage

Motivation > Model > Algorithms > Results

Information Overload – Summarization

Motivation > Model > Algorithms > Results

Conclusions

Dataset

Data: Bing reviews for Cameras, MP3 Players, Cell Phones

Attributes: Aspect rater tool of Bing

Quality: Helpfulness votes of the corresponding site

Viewpoints: Positive if rating 4 or 5, Negative otherwise

Algorithms: Greedy for UCOV, QCOV, GQCOV, SQCOV

Baselines: Top-Quality, Top-Length

k=5

Quantitative Analysis

Quantitative Analysis

Null-hypothesis ratio: fraction of items for which the results of the algorithm on the measure are close to random (empirical p-value > 0.05 over 1000 random samples)

Quantitative Analysis

Null-hypothesis ratio: fraction of items for which the results of the algorithm on the measure are close to random (empirical p-value > 0.05 over 1000 random samples)

	UCOV	QCOV	GQCOV	SQCOV	Quality
Greedy-UCOV	0.98%	3.43%	70.49%	9.02%	88.24%
Greedy-QCov	6.37%	0.49%	77.87%	11.48%	40.20%
Greedy-GQCOV	61.27%	54.90%	0.00%	50.82%	60.78%
Greedy-SQCov	17.65%	3.43%	9.84%	0.00%	53.43%
Top-Quality	83.33%	51.96%	86.89%	59.02%	1.47%
Top-Length	48.53%	34.80%	61.48%	35.25%	67.65%

Quantitative Analysis

Null-hypothesis ratio: fraction of items for which the results of the algorithm on the measure are close to random (empirical p-value > 0.05 over 1000 random samples)

	UCOV	QCOV	GQCOV	SQCOV	Quality
Greedy-UCOV	0.98%	3.43%	70.49%	9.02%	88.24%
Greedy-QCov	6.37%	0.49%	77.87%	11.48%	40.20%
Greedy-GQCOV	61.27%	54.90%	0.00%	50.82%	60.78%
Greedy-SQCov	17.65%	3.43%	9.84%	0.00%	53.43%
Top-Quality	83.33%	51.96%	86.89%	59.02%	1.47%
Top-Length	48.53%	34.80%	61.48%	35.25%	67.65%

Quantitative Analysis

Null-hypothesis ratio: fraction of items for which the results of the algorithm on the measure are close to random (empirical p-value > 0.05 over 1000 random samples)

	UCOV	QCOV	GQCOV	SQCOV	Quality
Greedy-UCOV	0.98%	3.43%	70.49%	9.02%	88.24%
Greedy-QCov	6.37%	0.49%	77.87%	11.48%	40.20%
Greedy-GQCOV	61.27%	54.90%	0.00%	50.82%	60.78%
Greedy-SQCov	17.65%	3.43%	9.84%	0.00%	53.43%
Top-Quality	83.33%	51.96%	86.89%	59.02%	1.47%
Top-Length	48.53%	34.80%	61.48%	35.25%	67.65%

Quantitative Analysis

Null-hypothesis ratio: fraction of items for which the results of the algorithm on the measure are close to random (empirical p-value > 0.05 over 1000 random samples)

	UCOV	QCOV	GQCOV	SQCOV	Quality
Greedy-UCOV	0.98%	3.43%	70.49%	9.02%	88.24%
Greedy-QCov	6.37%	0.49%	77.87%	11.48%	40.20%
Greedy-GQCOV	61.27%	54.90%	0.00%	50.82%	60.78%
Greedy-SQCov	17.65%	3.43%	9.84%	0.00%	53.43%
Top-Quality	83.33%	51.96%	86.89%	59.02%	1.47%
Top-Length	48.53%	34.80%	61.48%	35.25%	67.65%

Quantitative Analysis

Null-hypothesis ratio: fraction of items for which the results of the algorithm on the measure are close to random (empirical p-value > 0.05 over 1000 random samples)

	UCOV	QCOV	GQCOV	SQCOV	Quality
Greedy-UCOV	0.98%	3.43%	70.49%	9.02%	88.24%
Greedy-QCov	6.37%	0.49%	77.87%	11.48%	40.20%
Greedy-GQCOV	61.27%	54.90%	0.00%	50.82%	60.78%
Greedy-SQCov	17.65%	3.43%	9.84%	0.00%	53.43%
Top-Quality	83.33%	51.96%	86.89%	59.02%	1.47%
Top-Length	48.53%	34.80%	61.48%	35.25%	67.65%

Talk outline

Information Overload -- Coverage

Motivation > Model > Algorithms > Results

Information Overload – Summarization

Motivation > Model > Algorithms > Results

Conclusions

Coverage-based review selection

Holistic

- Provides all aspects of users' opinions

Not statistical

- Ratio of positive and negative reviews (per attribute) is lost

Need for Statistical Summaries

Statistical Summaries

Statistical Summaries

Accurate statistics

- Estimate of the representation of every opinion in the reviewers population

Not narrative

- Users like to read the narrative of reviews

Statistical Review Selection

Talk outline

Information Overload -- Coverage

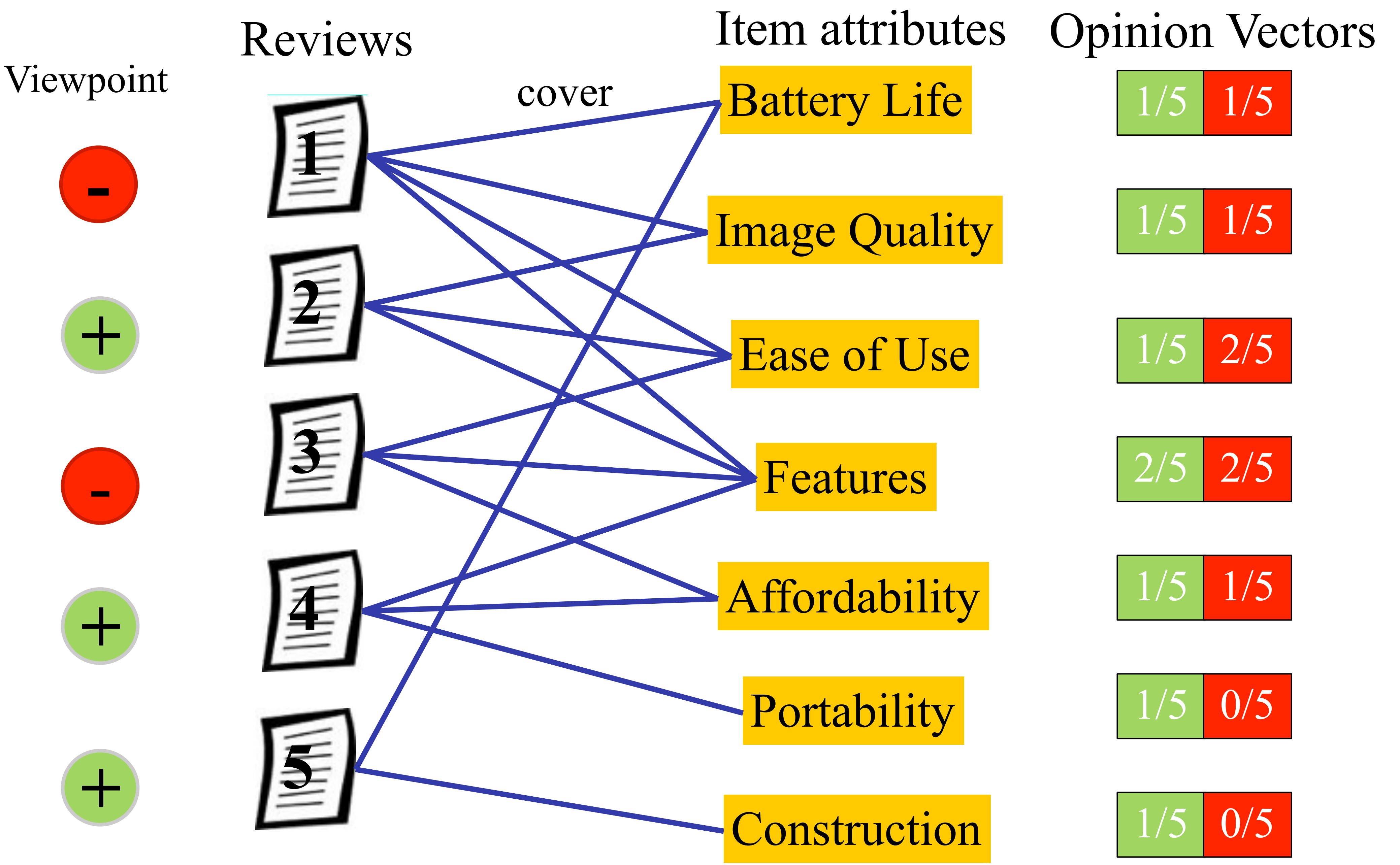
Motivation > Model > Algorithms > Results

Information Overload – Summarization

Motivation > Model > Algorithms > Results

Conclusions

The Model



Reviews
Item attributes
Opinion Vectors
Viewpoint
cover
-
+
-
+
+
1
2
3
4
5
Battery Life
Image Quality
Ease of Use
Features
Affordability
Portability
Construction
1/5 1/5
1/5 1/5
1/5 2/5
2/5 2/5
1/5 1/5
1/5 0/5
1/5 0/5

Our goal

Select a small (size k) set of reviews that approximate the *opinion vector* as well as possible

Statistical Selection

How good is a subset of reviews ***S*** given original review collection ***R***?

For opinionated attribute ***a***:

sc(S,a) quantifies whether ***S*** and ***R*** cover ***a*** similarly

Statistical Coverage Function:

$$F(S) = \sum_{a \in A} sc(S,a)$$

Statistical Selection Problem

Given a collection of reviews ***R*** select a set of ***k*** reviews ***S*** such that ***F(S)*** is minimized

$$F(S) = \sum_{a \in A} sc(S, a)$$

Where: $sc(S,a) = (mean(R,a) - mean(S,a))^2$

$$sc(S,a) = (target\text{-}vector(a) - mean(S,a))^2$$

Talk outline

Information Overload -- Coverage

Motivation > Model > Algorithms > Results

Information Overload – Summarization

Motivation > Model > Algorithms > Results

Conclusions

Analysis

The Statistical Selection problem is NP-hard to approximate for arbitrary target vectors

Several heuristic algorithms: Greedy, Random, Integer-Regression

The Integer-Regression algorithm

For i=1…ℓ

1. [Regression step:] Form a nonnegative real-valued vector x: $F(R_x)$ is small, and the number of nonzero elements of x is not larger than ℓ

Rx ~ target-vector

2. [Integer-transformation step:] Form a nonnegative integer-valued vector s representing k reviews that together approximate x in distribution:

$$L_1\left(\frac{s}{||s||_1}-\frac{x}{||x||_1}\right)$$

is minimized.

Talk outline

Information Overload -- Coverage

Motivation > Model > Algorithms > Results

Information Overload – Summarization

Motivation > Model > Algorithms > Results

Conclusions

Dataset

Data: Amazon reviews for Cameras, MP3 Players, Coffee Makers, Printers, Books, Vacuum Cleaners

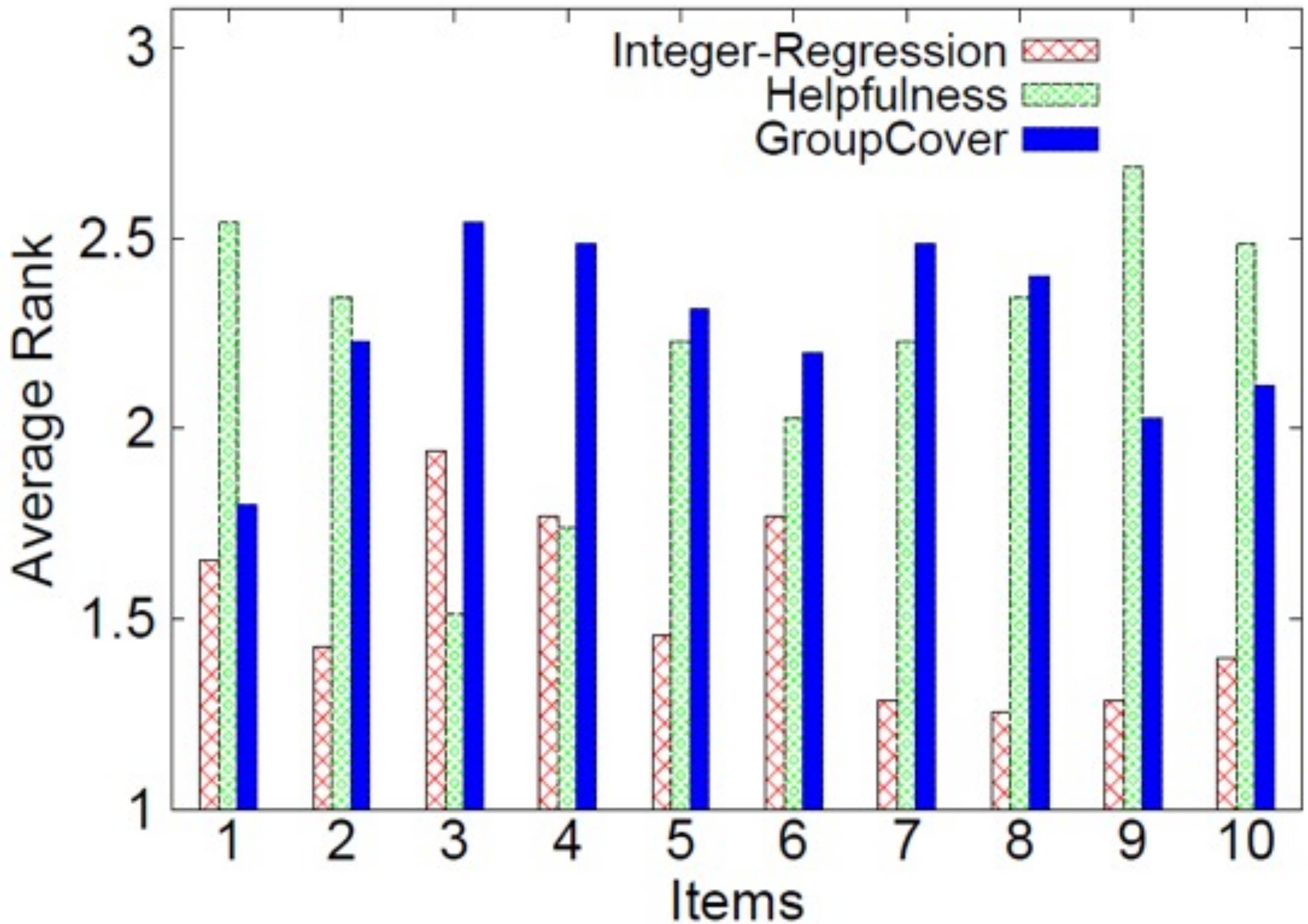
Attributes: Extracted automatically using attribute extractor

Viewpoints: Extracted automatically using attribute extractor

Baselines: Helpfulness, GCoverage

k=5

User Study


Integer-Regression
Helpfulness
GroupCover
Average Rank
3
2.5
2
1.5
1
1
2
3
4
5
6
7
8
9
10
Items

Abundance of Algorithmic Problems

Customers

Information Overload

Discovery of hidden gems

Reviewers

Motivation and Utilization

Merchants

Merchant Feedback