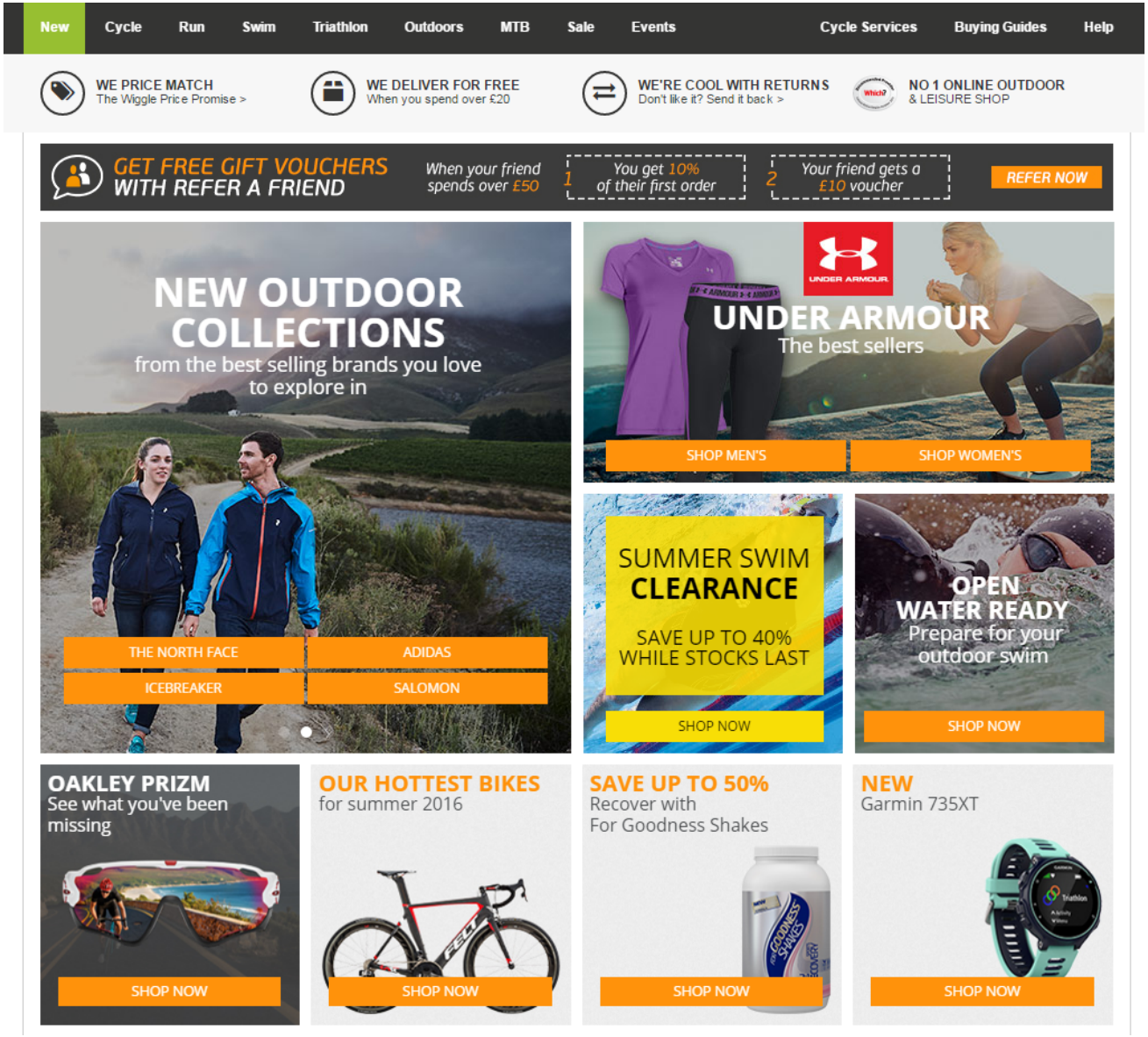
Uncertainty and Diversity in Web Recommendation

James Edwards
j.edwards4@lancaster.ac.uk
Work with David Leslie

Lancaster University

RSS Conference
06 September 2017

Web Recommendation



Diversity

Shop for bikes on Google

Sponsored

B'twin Elops 100 Dutch ...
£139.99
Decathlon UK

"Pendleton Somerby ...
£249.99
Halfords

2016 Scott Plasma ...
£9,699.00
Swift Cycles

"Apollo Claws Kids Bike - 14""
£59.99
Halfords

- Choose a set of elements $A \subseteq \mathcal{A}$
- *Diversity* is variety in the element set.
- Similar elements leads to *redundancy*.
- How to add diversity? How much?

User Preference Uncertainty

- Element value depends on user.
- Different users have different interests.
- What does *this* user want *today*?

Current user interests/preference given by state $x \in \mathcal{X}$.

Provided with topic preference vector $\mathbf{q}$ - a probability mass function on $\mathcal{X}$.

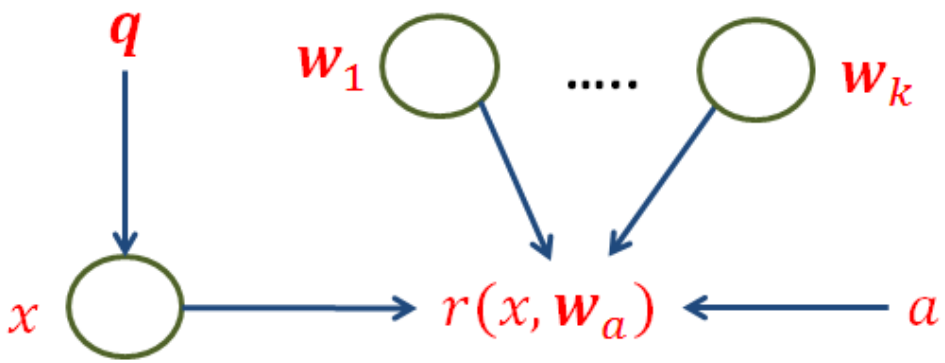
User Preference Uncertainty

- Element value depends on user.
- Different users have different interests.
- What does *this* user want *today*?

Current user interests/preference given by state $x \in \mathcal{X}$.

Provided with topic preference vector $\mathbf{q}$ - a probability mass function on $\mathcal{X}$.

Single Element Click Model



- Each element $a \in \mathcal{A}$ has a weight vector $\mathbf{w} = (w_{a,x})_{x \in \mathcal{X}}$.
- Click probability of single element a is $w_{a,x}$.
- Expected click-through rate (CTR) is $\mathbf{q} \cdot \mathbf{w}_a$.

Example

- $\mathbf{q} = (0.8, 0.2)$.
- $\mathbf{w}_1 = (0.4, 0)$, $\mathbf{w}_2 = (0.3, 0)$, $\mathbf{w}_3 = (0, 0.2)$.
- Independent CTRs: $R_1 = 0.32$, $R_2 = 0.24$, $R_3 = 0.04$.
- User clicks *at most one*.
- Which two should we choose? Depends on interactions.

Probablistic Click Model (PCM)

- Probability that a user is willing to click each element is independent of other elements.

Click probability of a set A of elements is

$$\mathbb{E}\left[1 - \prod_{a \in A}(1 - w_{a,x})\right].$$

Repeating the same elements in the set increases CTR.

Threshold Click Model (TCM)

If topic x and user declines element a would they click element b with $w_{b,x} < w_{a,x}$?

- TCM is a willingness-to-click model.
- Users have click threshold $u \sim U(0, 1)$.
- Given topic x, click probability is $\max_{a \in A} w_{a,x}$.

Click probability for a set A is

$$\mathbb{E}\left[\max_{a \in A} w_{a,x}\right].$$

Example cntd.

- $\mathbf{q} = (0.8, 0.2)$.
- $\mathbf{w}_1 = (0.4, 0), \mathbf{w}_2 = (0.3, 0), \mathbf{w}_3 = (0, 0.2)$.
- Independent CTRs: $R_1 = 0.32$, $R_2 = 0.24$, $R_3 = 0.04$.
- Choose $A = \{1, 2\}$ under PCM but $A = \{1, 3\}$ under TCM.

Set Choosing Methods

- Easy to find CTR of any set but...
- ...combinatorial explosion in number of sets.

Possible Methods

Optimise (OPT): Computationally too slow.

Naive (NAI): Best m elements with largest $\mathbf{q} \cdot \mathbf{w}_a$.

Most Frequent (MFUP): Fix $\tilde{x} = \arg\max q_x$, then select the elements with largest $w_{a,\tilde{x}}$.

Ordered Preference (OUP): Condition on $\tilde{x}_i$ in order of decreasing $q_{\tilde{x}_i}$.

Submodularity

Theorem

Both PCM and TCM are submodular functions of $A \subseteq \mathcal{A}$.

- A greedy sequential method (SEQ) is effective for submodular maximisation problems.
- Choose elements in sequence, conditioning on those already chosen.
- Guarantees solution within $1 - e^{-1} \approx 0.63$ of optimal.

Simulation

Fix $|\mathcal{X}| = 20$, $|\mathcal{A}| = 40$, $|A| = 3$.

Simulate 1000 i.i.d instances. On each:

- $\mathbf{q} \sim \mathit{Dirichlet}(1/20, \ldots, 1/20)$.
- Each $w_{a,x} \overset{i.i.d.}{\sim} 0.1\mathit{Beta}(1, \beta) + 0.9\delta_{0.001}$.
- Select A using each method.
- Calculate CTR under both PCM and TCM.

CTR Results

Set Choosing Method	True Click Model and β Value			
	PCM	PCM	TCM	TCM
	$\beta = 2$	$\beta = 9$	$\beta = 2$	$\beta = 9$
OPT-PCM	0.0%	0.0%	3.4%	6.6%
OPT-TCM	9.2%	20%	0.0%	0.0%
SEQ-PCM	0.0%	0.0%	3.4%	6.6%
SEQ-TCM	8.9%	20%	0.1%	0.1%
NAI	4.2%	0.6%	11%	10%
MFUP	17%	14%	26%	27%
OUP	10%	21%	1.1%	1.4%

Table: Lost CTR as a percentage of optimal CTR.

Diversity/Redundancy Measure

Definition (Overlap)

The overlap of a two element set $A = a_1, a_2$ is

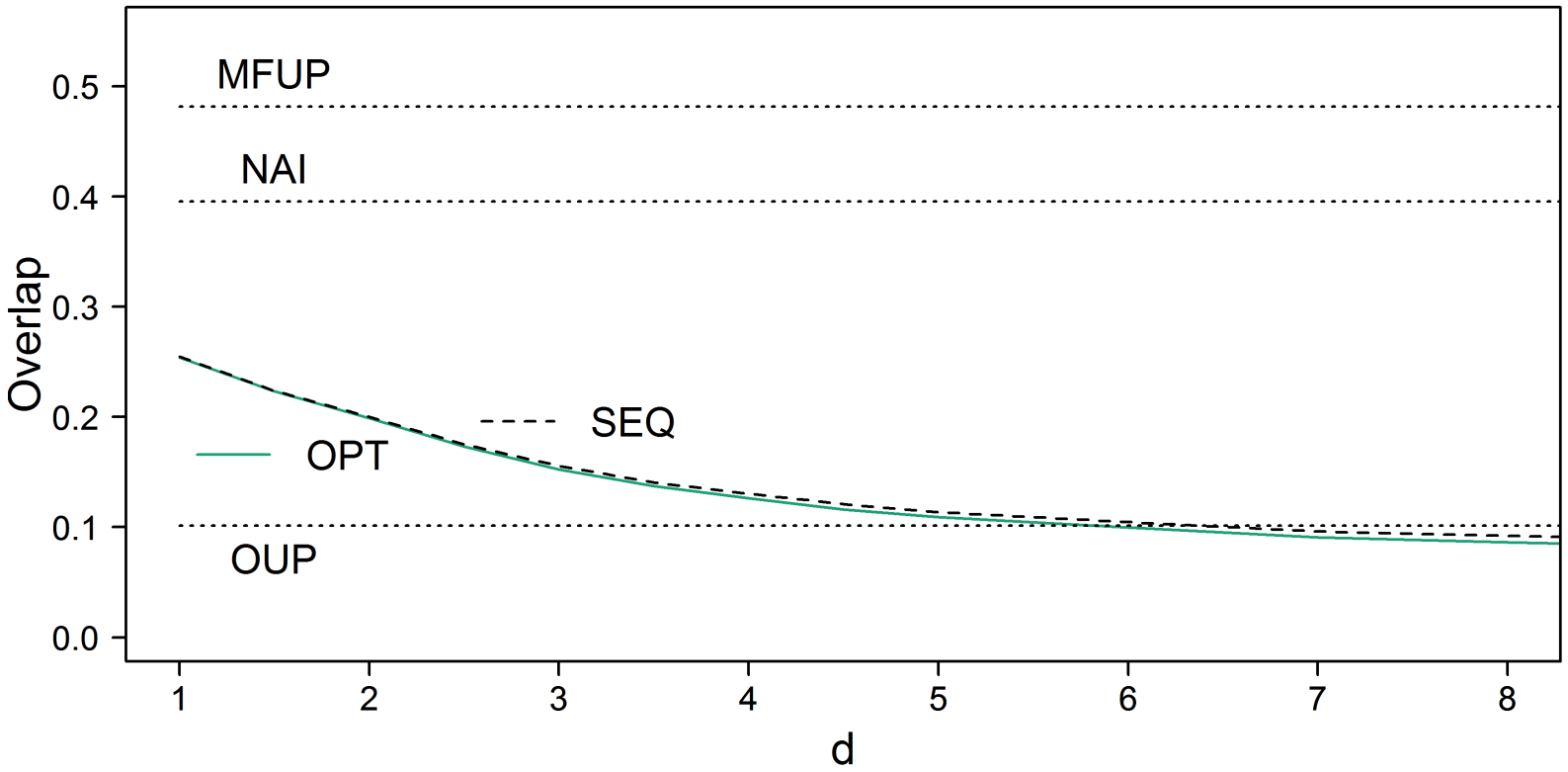
$$overlap(A) = \frac{\sum_{x=1}^{n} \min(w_{1,x}, w_{2,x})}{\min[\sum_{x=1}^{n}(w_{1,x}), \sum_{x=1}^{n}(w_{1,x})]} .$$

For sets of elements larger than 2 the overlap is given by

$$\frac{2}{|A|(|A|-1)} \sum_{a_i, a_j \in A, i<j} overlap(\{a_i, a_j\}).$$

The *diversity* of the set A is given by $1 - overlap(A)$.

Overlap



Summary

- Choosing a set of web elements is different from selecting individual elements.
- Diversity - why, how, and how much?
  - Why? Needed due to user preference uncertainty.
  - How? Maximising CTR with realistic click model creates diversity.
  - How much? Depends on click model and level of uncertainty.
- See Edwards, J. A., & Leslie, D. S. (forthcoming). Diversity as a response to user preference uncertainty. In Statistical Data Science. World Scientific.