Exploiting Submodularity to Tame Information Overload

Khalid El-Arini
Dafna Shahaf
Yisong Yue

Carlos Guestrin
The abundance of books is a distraction

Lucius Annaeus Seneca
4 BC - 65 AD
And it’s getting worse... [Tague et al. 1981]
Search Limitations

**Input**

Google

**Output**

Interaction

New query ➔ Change keywords
Our Approach

Input
Phrase complex information needs

Output
Structured, annotated output

Interaction
New keywords
Learn user’s interests
• Millions of blog posts published every day
• Some stories become disproportionately popular
  • Hard to find information you care about
Our goal: coverage

- Turn down the noise in the blogosphere
  - select a small set of posts that covers the most important stories

January 17, 2009
Our goal: **coverage**

- Turn down the noise in the blogosphere
- Select a small set of posts that **covers** the most **important** stories

Monday, April 9, 2012
Our goal: personalization

- Tailor post selection to user tastes

But, I like sports! I want articles like:

Manchester United fold without a fight as Barcelona claim Champions League

Parker Scores 19 to Lead San Antonio Past Clippers

After personalization based on Zidane’s feedback
Coverage:

- Formalize notion of **covering** the blogosphere
- **Near-optimal solution** for post selection
- Evaluate on **real blog data** and compare against:
  - Google
  - Yahoo! buzz
  - Digg
  - Blogpulse

Personalization:

- Learn a **personalized** coverage function
  - Algorithm for learning user preferences using limited feedback
- Evaluate on **real blog data**
Approach Overview

Blogosphere

Feature Extraction

Coverage Function

Post Selection

Personalization

Monday, April 9, 2012
Document Features

- **Low level**
  - Words, noun phrases, named entities
    - e.g., Obama, China, peanut butter

- **High level**
  - e.g., Topics
  - Topic = probability distribution over words

Inauguration Topic

National Security Topic

Monday, April 9, 2012
Coverage Function

- Some features more important than others
  ⇒
  e.g., weigh by frequency

- A post never covers a feature completely
  ⇒
  use soft notion of coverage, e.g., prob. at least one post in A covers feature $f$
Objective Function for Post Selection

- Want to select a set of posts $\mathcal{A}$ that maximizes

$$F(\mathcal{A}) = \sum_{f \in \mathcal{U}} w_f \text{cover}_\mathcal{A}(f)$$

- Posts shown to user
- Feature set
- Weights on features
- Probability that set $\mathcal{A}$ covers feature $f$

- Maximizing $F(\mathcal{A})$ is NP-hard!

- $F(\mathcal{A})$ is submodular

- Greedy $\Rightarrow (1-1/e)$-approximation
- Lazy greedy (CELF) $\Rightarrow$ very fast, same guarantees

Monday, April 9, 2012
Approach Overview

Blogosphere

Feature Extraction → Coverage Function → Post Selection → Personalization

\[ F(A) = \sum_{f \in U} w_{f \text{cover}}(f) \]

Submodular function optimization
User study:

LDA topics as features

Named entities and common noun phrases as features

We do as well as Yahoo! and Google

Monday, April 9, 2012
User study:

Google performs poorly
We do as well as Yahoo!

Monday, April 9, 2012
User study summary

- Google: good topicality, high redundancy
- Yahoo!: performs well on both, but uses rich features
  - CTR, search trends, user voting, etc.

**TDN performs as well as Yahoo! using only post content**
TDN outline

Coverage:
- Formalize notion of covering the blogosphere
- Near-optimal solution for post selection
- Evaluate on real blog data and compare against:
  - Google
  - Yahoo! Buzz
  - Digg
  - BlogPulse

Personalization:
- Learn a personalized coverage function
  - Algorithm for learning user preferences using limited feedback
- Evaluate on real blog data
Personalization

- People have varied interests

\[\text{Barack Obama} \quad \text{Britney Spears}\]

- **Our Goal**: Learn a personalized coverage function using limited user feedback
Personalize postings

Blogosphere

personalized coverage fn.

personalized post selection

personalization

learn your coverage function

online learning of submodular function problem

Monday, April 9, 2012
Modeling User Preferences

\[ F_\pi(A) = \sum_{f \in U} \pi^*_f \cdot \text{cover}_A(f) \]

- \( \pi^*_f \) represents user preference for feature \( f \)
- Want to learn preference \( \pi^* \) over the features

\( \pi^*_1, \pi^*_2, \pi^*_3, \pi^*_4, \pi^*_5 \) for a politico

\( \pi^*_1, \pi^*_2, \pi^*_3, \pi^*_4, \pi^*_5 \) for a sports fan

Importance of feature in corpus

User preference

Monday, April 9, 2012
User Model

- User scans articles in order
- Stochastically generates feedback (reward)
- Independent of other feedback
- Depends on above articles

\[
\mathbb{E}[r(a) \mid A] = \left(\pi^*\right)^T \Delta(a \mid A)
\]

“Conditional Submodular Independence”

\[
\Delta(a \mid A) = \begin{bmatrix}
F_1(A \cup a) - F_1(A) \\
F_2(A \cup a) - F_2(A) \\
\vdots \\
F_D(A \cup a) - F_D(A)
\end{bmatrix}
\]

\[
\mathbb{E}[r(A)] = \mathbb{E} r(a_1, a_2, \ldots, a_l)
\]
Fitting User’s Feedback

Simple regression approach fits preference vector to expected reward:

\[
\begin{bmatrix}
\Delta(a_1 | A_1) \\
\Delta(a_2 | A_2) \\
\vdots \\
\Delta(a_t | A_t)
\end{bmatrix}
\begin{bmatrix}
\hat{\pi}_1 \\
\vdots \\
\hat{\pi}_d
\end{bmatrix}
= 
\begin{bmatrix}
r_1 \\
r_2 \\
\vdots \\
r_t
\end{bmatrix}
\]

+ some regularization

Submodular advantage of article \(a_2\) wrt each feature

Reward of article \(a_2\)
**Goal:** want to recommend content that user likes
- Exploiting feedback from user, maximizing reward

**However:** user only provides feedback on recommended content
- Explore to collect feedback for new topics
  - Not addressed by [El-Arini, Veda, Shahaf, G. 2009]

**Solution:** algorithm to balance exploration vs exploitation
- Linear Submodular Bandits Problem
Balancing Exploration & Exploitation

- For each slot, maximize trade-off
  - (pick article about **Tennis**)

Estimated coverage gain

\[ \alpha_t \sqrt{\Delta(a|A_t)^T M_{i-1} \Delta(a|A_t)} \]

Uncertainty of estimate

Mean Estimate by Topic

<table>
<thead>
<tr>
<th>( \hat{\pi}_1 )</th>
<th>( \hat{\pi}_2 )</th>
<th>( \hat{\pi}_3 )</th>
<th>( \hat{\pi}_4 )</th>
<th>( \hat{\pi}_5 )</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>( C_1 )</th>
<th>( C_2 )</th>
<th>( C_3 )</th>
<th>( C_4 )</th>
<th>( C_5 )</th>
</tr>
</thead>
</table>

Monday, April 9, 2012
Theorem: For Bandit Alg,

\[(1-1/e) \text{avg}(\pi^*) - \text{avg}(\pi) \to 0\]

i.e., we achieve **no-regret**

Learns a good approximation of the true $\pi^*$

Rate: $d/\sqrt{kt}$, recommending $k$ documents, $T$ rounds, $d$ features

Before any feedback

After 1 day of personalization

After 2 days of personalization

Monday, April 9, 2012
User Study:

27 users in study

Submodular Bandits Wins

Static Weights

Multiplicative Updates (no exploration)

Submodular Bandits Wins

Ties

Losses

Submodular Bandits Wins

Ties

Losses

Monday, April 9, 2012
Problems in High Dimension

- Convergence rate has linear dependency on dimensionality:

\[ \pi^* \]
Hierarchical Bandit Learning [Yue, Hong, G., 2012]

- Assume $\pi^*$ mostly in subspace
  - Dimension $\ell \ll d$
  - E.g., Sports vs Politics

- Coarse to fine bandits
  - Two tiered exploration
  - Significantly fewer examples needed

Original Guarantee:
User Study

~27 users in study

Naïve LSBGreedy

Coarse-to-Fine

Wins

LSBGreedy with Optimal Prior in Full Space

Coarse-to-Fine

Wins

Ties

Losses

Monday, April 9, 2012
Our Approach

Input

Phrase complex information needs

Output

Structured, annotated output

Interaction

Learn user’s interests

what about more complex information needs?
As long as the centuries...unfold, the number of books will grow continually...

as convenient to search for a bit of truth concealed in nature

as to find it hidden away in an immense multitude of bound volumes

-Dennis Diderot, Encyclopédie (1755)

Today: $10^7$ papers in $10^5$ conferences/journals*

How do we cope?

* Thomson Reuters Web of Knowledge
Motivation (1)

Is there an approximation algorithm for the submodular covering problem that doesn’t require an integral-valued objective function?


Any recent papers influenced by this?
Motivation (2)

- It’s 11:30pm Samoa Time. Your “Related Work” section is a bit sparse.

Here are some papers we’ve cited so far.

Anything else?
Recommending Scientific Articles

Model of influence in science

Articles read thus far

Submodular function optimization

Diverse set of recommended articles

Monday, April 9, 2012
Our assumption:

Influence always occurs in the context of concepts
Putting it all together

Maximize $F_Q(A)$ s.t. $|A| \leq k$ (output $k$ papers)

- Submodular maximization problem

\[
F_Q(A) = \sum_{q \in Q} \sum_{c \in C} \gamma_q^{(c)} \text{Influence}_c(q \leftrightarrow A)
\]
But should all users get the same results?
Personalized trust

- Different communities trust different researchers for a given concept

\[ \text{Goal: Estimate personalized trust from limited user input} \]

- e.g., network

Pearl
Kleinberg
Hinton
Specifying trust preferences

- Specifying trust should not be an onerous task
- Assume given (nonexhaustive!) set of trusted papers $B$, e.g.,
  - a BibTeX file of all the researcher’s previous citations
  - a short list of favorite conferences and journals
  - someone else’s citation history!

  a committee member?
  journal editor?
  someone in another field?
  a Turing Award winner?
Personalized Objective

probability of influence between q and at least one paper in A

$$F_{Q|B}(A) = \sum_{q \in Q} \sum_{c \in C} \gamma_q^{(c)} \text{Influence}_c(q \leftarrow A|B)$$

Extra weight in Influence:
Does user trust at least one of authors of d with respect to concept c?
Recommending Scientific Articles

[El-Arini, G. ‘11]

- Articles read thus far
- Model of influence in science
- User’s bibtex file
- Model user’s interests
- Submodular function optimization
- Personalized
- Diverse set of recommended articles

Monday, April 9, 2012
Personalized Recommendations

On Power-Law Relationships of the Internet Topology

Michalis Faloutsos  
U.C. Riverside

Petros Faloutsos  
U. of Toronto

Christos Faloutsos  
Carnegie Mellon Univ.

general recommendations:

personalized recommendations
User Study Evaluation

- 16 PhD students in machine learning

For each:
- Selected a recent publication of participant — the study paper — for which we find related work
- Two variants of our methodology (w/ and w/o trust)
- Three state-of-the-art alternatives:
  - Relational Topic Model (generative model of text and links) [Chang, Blei ‘10]
  - Information Genealogy (uses only document text) [Shaparenko, Joachims ‘07]
  - Google Scholar (based on keywords provided by coauthor)

Double blind study where participant provided with title/author/abstract of one paper at a time, and asked several questions

Monday, April 9, 2012
On average, our approach provides more useful and more must-read papers than comparison techniques.
On average, our approach provides more trustworthy papers than comparison techniques, especially when incorporating participant’s trust preferences.
On average, our approach provides more familiar papers than comparison techniques, especially when incorporating participant’s trust preferences.
In pairwise comparison, our approaches produce more diverse results than the comparison techniques.
Our Approach

Input
Phrase complex information needs

Output
Structured, annotated output

Interaction

Learn user’s interests

what about structured outputs?
Can’t Grasp Credit Crisis? Join the Club

3.19.2008
Input: Pick two articles
(start, goal)

Bridge the gap with a smooth chain of articles
Input: Pick two articles (start, goal)
Output: Bridge the gap with a smooth chain of articles

- Keeping Borrowers Afloat
- A Mortgage Crisis Begins to Spiral
- Investors Grow Wary of Bank's Reliance on Debt
- Markets Can't Wait for Congress to Act
- Bailout Plan Wins Approval

Monday, April 9, 2012
Connecting the Dots [Shahaf, G. '10]

SAY, KID, YOU'RE PRETTY GOOD AT THAT!
HOW'D YOU LIKE TO WORK FOR THE FBI?!
What is a Good Chain?

• What’s wrong with shortest-path?

• Build a graph
  – Node for every article
  – Edges based on similarity
    • Chronological order (DAG)
  – Run BFS
Shortest-path

• A1: Talks Over Ex-Intern's Testimony On Clinton Appear to Bog Down

• A2: Judge Sides with the Government in Microsoft Antitrust Trial

• A3: Who will be the Next Microsoft?
  – trading at a market capitalization...

• A4: Palestinians Planning to Offer Bonds on Euro. Markets

• A5: Clinton Watches as Palestinians Vote to Rescind 1964 Provision

• A6: Contesting the Vote: The Overview; Gore asks Public For Patience; Florida recount
Shortest-path

- **A1:** Talks Over Ex-Intern's Testimony On Clinton Appear to Bog Down
- **A2:** Judge Sides with the Government in Microsoft Antitrust Trial
- **A3:** Who will be the Next Microsoft?– trading at a market capitalization...
- **A4:** Palestinians Planning to Offer Bonds on Euro. Markets
- **A5:** Clinton Watches as Palestinians Vote to Rescind 1964 Provision
- **A6:** Contesting the Vote: The Overview; Gore asks Public For Patience;
Shortest-path

• A1: Talks Over Ex-Intern's Testimony On Clinton Appear to Bog Down

• A2: Judge Sides with the Government in Microsoft + trust Trial

• A3: Who will be the Next Microsoft?
– Trading at a market capitalization…

• A4: Palestinians Planning to Offer Bonds on Euro. Markets

• A5: Clinton Watches as Palestinians Vote to Rescind 1964 Provision

• A6: Contesting the Vote: The Overview; Gore asks Public For Patience;

Stream of consciousness?
- Each transition is strong
- No global theme

Monday, April 9, 2012
More-Coherent Chain

- **B1:** Talks Over Ex-Intern's Testimony On Clinton Appear to Bog Down

- **B2:** Clinton Admits Lewinsky Liaison to Jury

- **B3:** G.O.P. Vote Counter in House Predicts Impeachment of Clinton

- **B4:** Clinton Impeached; He Faces a Senate Trial

- **B5:** Clinton’s Acquittal; Senators Talk About Their Votes

- **B6:** Aides Say Clinton Is Angered As Gore Tries to Break Away

- **B7:** As Election Draws Near, the Race Turns Mean

Contesting the Vote: The Overview; Gore asks Public For Patience;

Monday, April 9, 2012
More-Coherent Chain

- **B1:** Talks Over Ex-Intern's Testimony On Clinton Appear to Bog Down

- **B2:** *Clinton Admits Lewinsky* Liaison to Jury

- **B3:** G.O.P. Vote Counter in House Predicts *Impeachment of Clinton*

- **B4:** *Clinton’s Acquittal*; Senators Talk About Their Votes

- **B5:** Aides Say Clinton Is Angered As *Gore Tries to Break Away*

- **B6:** As *Election Draws Near*, the Race Turns Mean

- **B7:** Contesting the Vote: The Overview; Gore asks Public For Patience;

*Monday, April 9, 2012*
Word Patterns
For Shortest Path Chain

Topic changes every transition (jittery)
Word Patterns
For Coherent Chain

Use this intuition to estimate coherence of chains
(LP-relaxation + randomized rounding)

Topic consistent over transitions
Interaction

Simpson Defense Drops DNA Challenge

Algorithmic ideas from online learning

• MANY black officers say bias Is rampant in LA police force
• Racial split at the end ...

Simpson Verdict

Verdict Blood, glove

Monday, April 9, 2012
Moving Forward: Maps of Info

Can Computers Think? The History and Status of the Debate—Chart 1 of 7 Charts

Selected, structured, annotated relations

query

select important documents

Monday, April 9, 2012
Moving Forward: Maps of Info

- Machines can’t have emotions
- We can imagine artifacts that have feelings
- Deeper understanding → address information overload
- Challenge: build structured view automatically!
- Concept of feeling only applies to living organisms

[Ziff ‘59]
Trains of Thought

• Given a set of documents
• Show important pieces of information
• ... and how they relate
What makes a good map?

1. Coherence
2. Coverage
3. Connectivity
Approach overview

1. Coherence graph G
   - Encodes all coherent chains as graph paths

2. Coverage function f
   - f( ) = ?
   - Find a set of paths that maximize coverage (submodular orienteering)

3. Find high-coverage, high-connectivity paths

Documents D
First Step: Metro Maps of Science

Example query: Reinforcement

Submodular optimization algorithm

Map of Science:
14\% more relevant papers
58\% more fundamental topics

What are the most important topics and representative papers of RL today?

26 earlier-stage grad students

Domain expert evaluation

Scholar + Map

Google Scholar

Relevance
Coverage

Scholar + Map

[Shahaf, G. ’12]
Taming Information Overload

**Input**
- set of query papers, end points of chain
- Phrase complex information needs

**Interaction**
- like/dislike (online learning)
- bibtex file (trust)
- feedback on concepts
- Learn user's interests

**Output**
- efficient algorithms with theoretical guarantees
- smooth chain connecting the dots, metro maps, issue maps
- Structured, annotated output

- multiple user studies ⇒ promising direction for taming challenge of information overload

Monday, April 9, 2012