



Diversified Recommendation on Graphs: Pitfalls, Measures, and Algorithms

Onur Küçüktunç^{1,2} Erik Saule¹ Kamer Kaya¹ Ümit V. Çatalyürek^{1,3}

¹Dept. Biomedical Informatics ²Dept. of Computer Science and Engineering ³Dept. of Electrical and Computer Engineering **The Ohio State University**

WWW 2013, May 13-17, 2013, Rio de Janeiro, Brazil.

Outline

- Problem definition
 - Motivation
 - Result diversification algorithms
- How to measure diversity
 - Classical relevance and diversity measures
 - Bicriteria optimization?!
 - Combined measures
- Best Coverage method
 - Complexity, submodularity
 - A greedy solution, relaxation
- Experiments

Problem definition

Let G = (V, E) be an undirected graph. Given a set of m seed nodes $\mathcal{Q} = \{q_1, \ldots, q_m\}$ s.t. $\mathcal{Q} \subseteq V$, and a parameter k, return top-k items which are relevant to the ones in \mathcal{Q} , but diverse among themselves, covering different aspects of the query.



	Online shopping	Academic	Social
= (V, E)	product co-purchasing	paper-to-paper collaboration citations network	friendship network
$\subseteq V$	one productprevious purchasespage visit history	 paper/field of interest set of references researcher himself/herself 	user himself/herselfset of people
$C \subset V$	product recommendations "you might also like…"	references for related work new collaborators	friend recommendations "you might also know…"

G

0

 \mathcal{R}

Problem definition

Let G = (V, E) be an undirected graph. Given a set of m seed nodes $\mathcal{Q} = \{q_1, \ldots, q_m\}$ s.t. $\mathcal{Q} \subseteq V$, and a parameter k, return top-k items which are relevant to the ones in \mathcal{Q} , but diverse among themselves, covering different aspects of the query.



- We assume that the graph itself is the only information we have, and **no categories or intents are available**
 - no comparisons to intent-aware algorithms [Agrawal09,Welch11,etc.]
 - but we will compare against intent-aware measures
- Relevance scores are obtained with **Personalized PageRank (PPR)** [Haveliwala02] $p^*(v) = \begin{cases} 1/m, & \text{if } v \in Q\\ 0, & \text{otherwise.} \end{cases}$

Result diversification algorithms

- GrassHopper [Zhu07]
 - ranks the graph k times
 - turns the highest-ranked vertex into a sink node at each iteration



Result diversification algorithms

- GrassHopper [Zhu07]
 - ranks the graph k times
 - turns the highest-ranked vertex into a sink node at each iteration
- DivRank [Mei10]
 - based on vertex-reinforced random walks (VRRW)
 - adjusts the transition matrix based on the number of visits to the vertices (*rich-gets-richer* mechanism)



Result diversification algorithms

- GrassHopper [Zhu07]
 - ranks the graph k times
 - turns the highest-ranked vertex into a sink node at each iteration
- DivRank [Mei10]
 - based on vertex-reinforced random walks (VRRW)
 - adjusts the transition matrix based on the number of visits to the vertices (*rich-gets-richer* mechanism)
- Dragon [Tong11]
 - based on optimizing the goodness measure
 - punishes the score when two neighbors are included in the results

Measuring diversity

Relevance measures

• Normalized relevance • *l*-step graph density

$$rel(S) = \frac{\sum_{v \in S} \pi_v}{\sum_{i=1}^k \hat{\pi}_i}$$

Difference ratio

$$diff(S, \hat{S}) = 1 - \frac{|S \cap \hat{S}|}{|S|}$$

nDCG

nDCG_k =
$$\frac{\pi_{s_1} + \sum_{i=2}^{k} \frac{\pi_{s_i}}{\log_2 i}}{\hat{\pi}_1 + \sum_{i=2}^{k} \frac{\hat{\pi}_i}{\log_2 i}}$$

Diversity measures

$$\operatorname{dens}_{\ell}(S) = \frac{\sum_{u,v \in S, u \neq v} d_{\ell}(u,v)}{|S| \times (|S| - 1)}$$

l-expansion ratio

$$\sigma_{\ell}(S) = \frac{|N_{\ell}(S)|}{n}$$

where

 $N_{\ell}(S) = S \cup \{v \in (V - S) : \exists u \in S, d(u, v) \le \ell\}$

Kucuktunc et al. "Diversified Recommendation on Graphs: Pitfalls, Measures, and Algorithms", WWW 13

Bicriteria optimization measures

- aggregate a relevance and a diversity measure
- [Carbonell98]

$$f_{MMR}(S) = (1 - \lambda) \sum_{v \in S} \pi_v - \lambda \sum_{u \in S} \max_{\substack{v \in S \\ u \neq v}} sim(u, v)$$

[Li11]
$$f_L(S) = \sum \pi_v + \lambda \frac{|N(S)|}{n}$$

 $v \in S$

• [Vieira11]

$$f_{MSD}(S) = (k-1)(1-\lambda)\sum_{v \in S} \pi_v + 2\lambda \sum_{u \in S} \sum_{\substack{v \in S \\ u \neq v}} div(u,v)$$

 max-sum diversification, max-min diversification, k-similar diversification set, etc. [Gollapudi09]

Bicriteria optimization is not the answer

Google bicrite

bicriteria optimization

[PS] 25.1 Introduction 25.2 Bicriteria Problems - School of Computer ... www.cs.cmu.edu/afs/cs/academic/class/15854-f05/www/.../lec25.ps * Dec 7, 2005 - 25.1 Introduction. In this lecture we will consider bicriteria optimization problems - problems in which there are. two optimization functions.

Bicriteria Optimization Problem of Designing an Index Fund - JStor www.istor.org/stable/3009912 *

by Y Tabata - 1995 - Cited by 32 - Related articles Key words: bicriterion optimization, index fund, market portfolio ... So, the problem can be regarded as a bicriteria optimization problem: (1) to minimize the

Bicriteria Optimization of a Queue with a Controlled Input Stream dl.acm.org/citation.cfm?id=1017053 ~

by AB Plunovskiy - 2004 - Cited by 12 - Related articles Bicriteria Optimization of a Queue with a Controlled Input Stream, 2004 Article. Bibliometrics Data Bibliometrics. · Downloads (6 Weeks): 0 · Downloads (12 ...

[PDF] The Smoothed Number of Pareto Optimal Solutions in Bicriteria ..

www.roeglin.org/publications/IPCO07.pdf by R Beier - Cited by 16 - Related articles

of Pareto optimal solutions for general **bicriteria** integer **optimization** problems in the framework of smoothed analysis. Our analysis is based on a semi-random ...

[PS] Finding Representative Systems for Discrete Bicriteria Optimiza...

kluedo.ub.uni-kl.de/files/1655/hamacher_nr95.ps ▼ by HW Hamacher - 2005 - Cited by 1 - Related articles Finding Representative Systems for Discrete. Bicriteria Optimization Problems by Box. Algorithms. Horst W. Hamacher, Christian Roed Pedersen†, Stefan ...

Bi-criteria optimization of structures liable to instability - Springer

Ink.springer.com/content/pdf/10.1007/BF01742506 ▼ by J Pietrzak - 1994 - Cited by 3 - Related articles BI-criteria optimization of structures liable to instability. J. Pietrzak*. Civil Engineering Department, The University of Beira Interior, P-6200 Covilhg, Portugal

[PDF] A Bicriteria-Optimization-Approach-Based Dimensionality-Red...

www.iro.umontreal.ca/~mignotte/Publications/IEEE_GRS11_.pdf

by M Mignotte - Related articles

reduction model based on a bicriteria global optimization ap- proach for the color display of hyperspectral images. The proposed fusion model is derived from ...

Constructing robust crew schedules with bicriteria optimization onlinelibrary.wiley.com/doi/10.1002/mcda.321/pdf -

by M Ehrgott - 2002 - Cited by 121 - Related articles We develop a bicriteria optimization framework to generate Pareto optimal schedules for the domestic airline. A Pareto optimal schedule is one which does not ...

Ned Dimitrov - Probabilistic Bicriteria Optimization

neddimitrov.org/research/probabilistic-bicriteria-optimization.html Probabilistic Bicriteria Optimization. We consider a multiperiod system operation problem with two conficting objectives, mini-mizing cost and risk. Risk stems ...

0

- Objective: diversify top-10 results
- Two query-oblivious algorithms:

– top-% + random

百度-剪贴本

www.jtben.com/document/358582 ▼ Translate this page Nov 13, 2010 – 百度一下, 你就知道. 主要提供网页、音乐、图片、新闻搜索, 同时有帖 吧和WAP搜索功能。显示"BIDU"的股票报价. www.baidu.com/ - 网页快照 ...

Simple Spiderman Cookies - The Sweet Adventures of Sugar Belle www.sweetsugarbelle.com/2012/08/simple-spider-man-cookies/ -Aug 18, 2012 – Not too long ago I took the kiddos to see The Amazing Spider-Man. Once upon a time I would have told you that I am NOT an action movie kind ...

- top-% + greedy- σ_2

CNN.com - Breaking News, U.S., World, Weather, Entertainment ... www.cnn.com/ -

CNN.com delivers the latest breaking news and information on the latest top stories, weather, business, entertainment, politics, and more. For in-depth coverage, ...

Wikipedia

www.wikipedia.org/ -

Wikipedia, the free encyclopedia that anyone can edit.

Bicriteria optimization is not the answer

normalized relevance and 2-step graph density



- evaluating result diversification as a bicriteria optimization problem with
 - a relevance measure that <u>ignores</u> diversity, and
 - a **diversity** measure that <u>ignores</u> **relevancy**.

Kucuktunc et al. "Diversified Recommendation on Graphs: Pitfalls, Measures, and Algorithms", WWW13

A better measure? Combine both

- We need a combined measure that tightly integrates **both** *relevance* and *diversity* aspects of the result set
- goodness [Tong11]

penalize the score when two results share

an edge

$$f_G(S) = 2\sum_{i \in S} \pi_i - d \sum_{i,j \in S} \mathbf{A}(j,i)\pi_j$$

max-sum relevance
$$- (1-d) \sum_{j \in S} \pi_j \sum_{i \in S} p^*(i)$$

- downside: highly dominated by relevance

Proposed measure: *l*-step expanded relevance

- a combined measure of
 - *l*-step expansion ratio (σ_2)
 - relevance scores (π)
- quantifies: relevance of the covered region of the graph

 $\ell\text{-step}$ expanded relevance:

$$\operatorname{exprel}_{\ell}(S) = \sum_{v \in N_{\ell}(S)} \pi_{v}$$

where $N_{\ell}(S)$ is the ℓ -step expansion set of the result set S, and π is the PPR scores of the items in the graph.



 do some sanity check with this new measure

Kucuktunc et al. "Diversified Recommendation on Graphs: Pitfalls, Measures, and Algorithms", WWW 13

Correlations of the measures



14/25

Proposed algorithm: Best Coverage

- Can we use *l*-step expanded relevance as an objective function?
- **Define:** $exprel_{\ell}$ -diversified top-k ranking (DTR ℓ) $S = \underset{\substack{S' \subseteq V \\ |S'| = k}}{\operatorname{argmax}} exprel_{\ell}(S')$

ALGORITHM 1: BestCoverage

Input: k, G, π, ℓ Output: a list of recommendations S $S = \emptyset$ while |S| < k do $v^* \leftarrow \operatorname{argmax}_v g(v, S)$ $S \leftarrow S \cup \{v^*\}$ return S

- **Complexity:** generalization of *weighted maximum coverage problem*
 - NP-hard!
 - but $exprel_l$ is a submodular function (Lemma 4.2)
 - a greedy solution (Algorithm 1) that selects the item with the *highest marginal utility*

 $g(v, S) = \sum_{v' \in N_{\ell}(\{v\}) - N_{\ell}(S)} \pi_{v'}$ at each step is the best possible polynomial time approximation (proof based on [Nemhauser78])

• **Relaxation:** computes BestCoverage on highest ranked vertices to improve runtime

```
ALGORITHM 2: BestCoverage (relaxed)
  Input: k, G, \pi, \ell
  Output: a list of recommendations S
  S = \emptyset
  SORT(V) w.r.t \pi_i non-increasing
  S1 \leftarrow V[1..k'], i.e., top-k' vertices where k' = k\bar{\delta}^{\ell}
  \forall v \in S1, g(v) \leftarrow g(v, \emptyset)
  \forall v \in S1, c(v) \leftarrow \text{UNCOVERED}
  while |S| < k do
       v^* \leftarrow \operatorname{argmax}_{v \in S1} g(v)
       S \leftarrow S \cup \{v^*\}
       S2 \leftarrow N_{\ell}(\{v^*\})
       for each v' \in S2 do
             if c(v') = \text{UNCOVERED} then
                  S3 \leftarrow N_{\ell}(\{v'\})
                  \forall u \in S3, g(u) \leftarrow g(u) - \pi_{v'}
                   c(v') \leftarrow \text{COVERED}
  return S
```

Experiments

• 5 target application areas, 5 graphs from SNAP

Dataset	V	E	$\overline{\delta}$	D	$D_{90\%}$	CC
AMAZON0601	403.3K	$3.3\mathrm{M}$	16.8	21	7.6	0.42
CA-ASTROPH	$18.7\mathrm{K}$	$396.1 \mathrm{K}$	42.2	14	5.1	0.63
CIT-PATENTS	$3.7\mathrm{M}$	$16.5\mathrm{M}$	8.7	22	9.4	0.09
soc-LiveJournal1	$4.8\mathrm{M}$	$68.9\mathrm{M}$	28.4	18	6.5	0.31
WEB-GOOGLE	$875.7\mathrm{K}$	$5.1\mathrm{M}$	11.6	22	8.1	0.60

- Queries generated based on 3 scenario types
 - one random vertex
 - random vertices from one area of interest
 - multiple vertices from multiple areas of interest

Results – relevance



- Methods should trade-off relevance for better diversity
- Normalized relevance of top-k set is always 1
- DRAGON always return results having 70% similar items to top-k, with more than 80% rel score



- *l*-step expansion ratio (σ₂) gives the graph coverage of the result set: <u>better coverage = better diversity</u>
- BestCoverage and DivRank variants, especially
 BC₂ and PDivRank, have the highest coverage

Results – expanded relevance



- combined measure for relevance and diversity
- BestCoverage variants and GrassHopper perform better
- Although PDivRank gives the highest coverage on amazon graph, it fails to cover the relevant parts!

Results – efficiency



- BC₁ always performs better, with a running time less than, DivRank and GrassHopper
- BC₁ (relaxed) offers reasonable diversity, with a very little overhead on top of the PPR computation

Results – intent aware experiments

- evaluation of *intent-oblivious* algorithms against *intent-aware* measures
- two measures
 - group coverage [Li11]
 - S-recall [Zhai03]
- cit-Patent dataset has the categorical information
 - 426 class labels, belong to 36 subtopics

Results – intent aware experiments

- group coverage [Li11]
 - How many different groups are covered by the results?
 - omits the actual intent of the query



- top-k results are not diverse enough
- AllRandom results cover the most number of groups
- PDivRank and BC₂ follows

Results – intent aware experiments

- S-recall [Zhai03], Intent-coverage [Zhu11]
 - percentage of relevant subtopics covered by the result set
 - the intent is given with the classes of the seed nodes



- AllRandom brings irrelevant items from the search space
- top-k results do not have the necessary diversity
- BC₂ variants and BC₁ perform better than DivRank
- BC₁ (relaxed) and DivRank scores similar, but BC₁r much faster

Kucuktunc et al. "Diversified Recommendation on Graphs: Pitfalls, Measures, and Algorithms", WWW13

Conclusions

- Result diversification should not be evaluated as a bicriteria optimization problem with
 - a **relevance** measure that <u>ignores</u> **diversity**, and
 - a diversity measure that ignores relevancy
- *I*-step expanded relevance is a simple measure that combines both relevance and diversity
- BestCoverage, a greedy solution that maximizes exprel_l is a (1-1/e)-approximation of the optimal solution
- BestCoverage variants perform better than others, its relaxation is extremely efficient
- *goodness* in **DRAGON** is dominated by relevancy
- DivRank variants implicitly optimize expansion ratio



Wexner Medical Center

Thank you

- For more information visit
 - <u>http://bmi.osu.edu/hpc</u>
- Research at the HPC Lab is funded by

