Introduction
- Why?
- Overview

Citation Analysis for Document Recommendation
- Previous Approaches
- Direction Aware Recommendation

A High Performance Computing Problem
- A specialization of SpMV
- Ordering and Partitioning

Result Diversification

Other Features

Final Thoughts
- Conclusion
- Future Works
**Erik’s tale:** Once upon a time there was a survey paper

The Jimmy John’s scheduling problem

<table>
<thead>
<tr>
<th>scheduling</th>
<th>partitioning</th>
<th>mapping</th>
<th>pipeline</th>
<th>workflow</th>
<th>data flow</th>
<th>task graph</th>
<th>linear chain</th>
<th>sequences (tree)</th>
<th>(serial parallel)</th>
</tr>
</thead>
</table>

But also...

“Scheduling problems in parallel query optimization”

“Bringing skeletons out of the closet: A pragmatic manifesto for skeletal parallel programming”

After 6 months, unknown papers were still uncovered.

Develop software to make the search easier!

---

[Umit V. Catalyurek](http://bmi.osu.edu/hpc)

Ohio State University, Biomedical Informatics

HPC Lab [http://bmi.osu.edu/hpc](http://bmi.osu.edu/hpc)

theadvisor: [http://theadvisor.osu.edu/](http://theadvisor.osu.edu/)
Erik’s tale: Once upon a time there was a survey paper

The Jimmy John’s scheduling problem

- scheduling
- partitioning
- mapping
- pipeline
- workflow
- data flow
- task graph
- linear chain
- sequences (tree)
- (serial parallel)

But also...

“Scheduling problems in parallel query optimization”
“Bringing skeletons out of the closet: A pragmatic manifesto for skeletal parallel programming”
Erik’s tale: Once upon a time there was a survey paper

The Jimmy John’s scheduling problem

<table>
<thead>
<tr>
<th>Scheduling</th>
<th>Partitioning</th>
<th>Mapping</th>
<th>Pipeline</th>
<th>Workflow</th>
<th>Data Flow</th>
<th>Task Graph</th>
<th>Linear</th>
<th>Chain</th>
<th>Sequences</th>
<th>(Tree)</th>
<th>(Serial Parallel)</th>
</tr>
</thead>
</table>

But also...

“Scheduling problems in parallel query optimization”
“Bringing skeletons out of the closet: A pragmatic manifesto for skeletal parallel programming”

After 6 months, unknown papers were still uncovered
Erik’s tale: Once upon a time there was a survey paper

The Jimmy John’s scheduling problem

<table>
<thead>
<tr>
<th>scheduling</th>
<th>partitioning</th>
<th>mapping</th>
<th>pipeline</th>
<th>workflow</th>
<th>data flow</th>
<th>task graph</th>
<th>linear chain</th>
<th>sequences (tree)</th>
<th>(serial parallel)</th>
</tr>
</thead>
</table>

But also...

“Scheduling problems in parallel query optimization”
“Bringing skeletons out of the closet: A pragmatic manifesto for skeletal parallel programming”

After 6 months, unknown papers were still uncovered

Develop software to make the search easier!
Design Goals

**Personalized**
The user should be able to make a query that describes precisely what she is looking for.

**Conceptual**
The system should free of linguistic problems. Ambiguity and synonymy should be taken into accounts.

**Exploratory**
Different perspective should be available. The system should enhance the user’s search.

**Easy to use**
The user should not need to know anything about data mining or algorithms.

Umit V. Catalyurek
Ohio State University, Biomedical Informatics
HPC Lab http://bmi.osu.edu/hpc

theadvisor: http://theadvisor.osu.edu/
The Academic Web Service Ecosystem

DBLP
List of CS papers with clean reference and disambiguated names.
The Academic Web Service Ecosystem

**DBLP**
List of CS papers with clean reference and disambiguated names.

**CiteSeer, {Ref,Ack,Collab}Seer**
Automatically crawled papers in CS. Give PDFs. Contain citation information, full text. Compute similarity.
The Academic Web Service Ecosystem

**DBLP**
List of CS papers with clean reference and disambiguated names.

**Citeseer, \{Ref,Ack,Collab\}Seer**
Automatically crawled papers in CS. Give PDFs. Contain citation information, full text. Compute similarity.

**CiteUlike**
Social paper tagging application. Find paper from researchers with similar interest.
The Academic Web Service Ecosystem

**DBLP**
List of CS papers with clean reference and disambiguated names.

**CiteSeer, \{Ref, Ack, Collab\} Seer**
Automatically crawled papers in CS. Give PDFs. Contain citation information, full text. Compute similarity.

**CiteULike**
Social paper tagging application. Find paper from researchers with similar interest.

**ArnetMiner**
Academic network analysis.
# The Academic Web Service Ecosystem

<table>
<thead>
<tr>
<th>Service</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DBLP</strong></td>
<td>List of CS papers with clean reference and disambiguated names.</td>
</tr>
<tr>
<td><strong>CiteSeer, {Ref,Ack,Collab}Seer</strong></td>
<td>Automatically crawled papers in CS. Give PDFs. Contain citation information, full text. Compute similarity.</td>
</tr>
<tr>
<td><strong>CiteULike</strong></td>
<td>Social paper tagging application. Find paper from researchers with similar interest.</td>
</tr>
<tr>
<td><strong>ArnetMiner</strong></td>
<td>Academic network analysis.</td>
</tr>
<tr>
<td><strong>Mendeley</strong></td>
<td>Application for managing references. Database of reference.</td>
</tr>
</tbody>
</table>

---

Ümit V. Çatalyurek  
Ohio State University, Biomedical Informatics  
HPC Lab [http://bmi.osu.edu/hpc](http://bmi.osu.edu/hpc)  
theadvisor: [http://theadvisor.osu.edu/](http://theadvisor.osu.edu/)
<table>
<thead>
<tr>
<th>Service</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DBLP</strong></td>
<td>List of CS papers with clean reference and disambiguated names.</td>
</tr>
<tr>
<td><strong>CiteSeer, {Ref,Ack,Collab}Seer</strong></td>
<td>Automatically crawled papers in CS. Give PDFs. Contain citation information, full text. Compute similarity.</td>
</tr>
<tr>
<td><strong>Mendeley</strong></td>
<td>Application for managing references. Database of reference.</td>
</tr>
<tr>
<td><strong>Google Scholar</strong></td>
<td>Keyword-based search engine (with citation informations).</td>
</tr>
<tr>
<td><strong>CiteULike</strong></td>
<td>Social paper tagging application. Find paper from researchers with similar interest.</td>
</tr>
<tr>
<td><strong>ArnetMiner</strong></td>
<td>Academic network analysis.</td>
</tr>
<tr>
<td>Service</td>
<td>Description</td>
</tr>
<tr>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td><strong>DBLP</strong></td>
<td>List of CS papers with clean reference and disambiguated names.</td>
</tr>
<tr>
<td><strong>Citeseer, {Ref, Ack, Collab} Seer</strong></td>
<td>Automatically crawled papers in CS. Give PDFs. Contain citation information, full text. Compute similarity.</td>
</tr>
<tr>
<td><strong>Mendeley</strong></td>
<td>Application for managing references. Database of reference.</td>
</tr>
<tr>
<td><strong>CiteUlike</strong></td>
<td>Social paper tagging application. Find paper from researchers with similar interest.</td>
</tr>
<tr>
<td><strong>ArnetMiner</strong></td>
<td>Academic network analysis.</td>
</tr>
<tr>
<td><strong>Google Scholar</strong></td>
<td>Keyword-based search engine (with citation informations).</td>
</tr>
<tr>
<td><strong>Microsoft Academic Search</strong></td>
<td>Keyword-based search engine and Academic network analysis.</td>
</tr>
</tbody>
</table>

---

Ümit V. Çatalyürek
Ohio State University, Biomedical Informatics
HPC Lab [http://bmi.osu.edu/hpc](http://bmi.osu.edu/hpc)
theadvisor: [http://theadvisor.osu.edu/](http://theadvisor.osu.edu/)
<table>
<thead>
<tr>
<th><strong>DBLP</strong></th>
<th><strong>Mendeley</strong></th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th><strong>Citeseer, {Ref,Ack,Collab}Seer</strong></th>
<th><strong>Google Scholar</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatically crawled papers in CS. Give PDFs. Contain citation information, full text. Compute similarity.</td>
<td>Keyword-based search engine (with citation informations).</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>CiteUlike</strong></th>
<th><strong>Microsoft Academic Search</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Social paper tagging application. Find paper from researchers with similar interest.</td>
<td>Keyword-based search engine and Academic network analysis.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>ArnetMiner</strong></th>
<th><strong>IEEE, ACM, Elsevier, JSTOR, ...</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic network analysis.</td>
<td>Publishers or digital libraries with complete text and references. Some suggestions.</td>
</tr>
</tbody>
</table>
A Use Case

1. Put your references in a BibTeX or RIS file

@inproceedings{Kucuktunc12,
  author = {Kucuktunc, Onur and title = {A Large-Scale Sent. booktitle = {Proc. 5th Int' year = {2012},
  }},
}

@inproceedings{weber,
  author = {Weber, In title = {The demographics booktitle = {Proc. 33rd Int pages = {523--530}, year = {2010},
  }},
}

@article{Aral11,
  author = {Aral, Sinan and Ve

2. Select the file or simply drag-and-drop, adjust the parameter, then submit!

Let's get started!

1. Select a BibTeX (*.bib), RIS (*.ris) or EndNote (*.xml) file:

Choose File No file chosen

optional Have a bbl file as well?

2. I want papers to be more

traditional recent

× I authorize the use of my activity for research purposes.

Get recommendations →

3. Get citation, venue, and reviewer recommendations, give feedback...

- [ ] Papers

- Marina Drosou, Evaggelia Pitoura: Search result diversification. [bib] [good SIGMOD Record, 2010.


- Cong Yu, Laks V. S. Lakshmanan, Sihen Recommendation Diversification Using International Conference on Data Engine
System Overview

Architecture

A web-server as a front end. A cluster in the back-end. New instances are dynamically created as the load varies.

Functional

Paper Mapper

Recommendation Engine

Paper Mapper

parameters \( \{k,d,\kappa\} \)

paper IDs

Venue Rec.

Reviewer Rec.

Diversification Engine

Visualization

Relevance Feedback

venues

reviewers

papers
1 Introduction
   - Why?
   - Overview

2 Citation Analysis for Document Recommendation
   - Previous Approaches
   - Direction Aware Recommendation

3 A High Performance Computing Problem
   - A specialization of SpMV
   - Ordering and Partitioning

4 Result Diversification

5 Other Features

6 Final Thoughts
   - Conclusion
   - Future Works
Using the Citation Graph

Hypothesis
If two papers are related or treat the same subject, then they will be close to each other in the citation graph (and reciprocal)

Benefits
- No linguistic $\Rightarrow$ no synonymy, no ambiguity
- Automatically crowd source by researchers

Drawbacks
- Difficult to gather the data (But thanks Citeseer)
- Relies on researcher already having made similar connections
- **Bibliographic coupling** [Kessler63]: papers having similar references are related.

- **Cocitation** [Small73]: papers which are cited by the same papers are related.

- **CCIDF** [Lawrence99]: cocitations weighted with inverse frequencies.
Local Approaches

- **Bibliographic coupling** [Kessler63]: papers having similar references are related.
- **Cocitation** [Small73]: papers which are cited by the same papers are related.
- **CCIDF** [Lawrence99]: cocitations weighted with inverse frequencies.

**Problem:** Considers only level-2 papers based on level-1 information.
Global Approaches

Graph distance-based

- Katz: number of paths between two papers [Strohman07]

Random walk with restarts (RWR) based

- ArticleRank [Li09] (PageRank [Brin98] extension)
- PaperRank [Gori06] (Personalized PageRank [Haveliwala02] extension)

RWR treats the citations and references in the same way
Global Approaches

Graph distance-based
- Katz: number of paths between two papers [Strohman07]

Random walk with restarts (RWR) based
- ArticleRank [Li09] (PageRank [Brin98] extension)
- PaperRank [Gori06] (Personalized PageRank [Haveliwala02] extension)

RWR treats the citations and references in the same way

This is not exploratory!
Let $G = (V, E)$ be the citation graph.

PageRank [Brin98]

Let $\pi_i(u) = (1 - d) \frac{1}{|V|} + d \sum_{v \in N(u)} \frac{\pi_{i-1}(v)}{\delta(v)}$

with $d \in (0 : 1)$ is the damping factor. It converges to a stable distribution.

source: wikipedia
Let $G = (V, E)$ be the citation graph.

**PageRank** [Brin98]

$$\pi_i(u) = (1 - d) \frac{1}{|V|} + d \sum_{v \in N(u)} \frac{\pi_{i-1}(v)}{\delta(v)}$$

with $d \in (0 : 1)$ is the damping factor. It converges to a stable distribution. **But it is not personalized.**

source: wikipedia
Let $G = (V, E)$ be the citation graph

**PageRank [Brin98]**

$$\pi_i(u) = \left(1 - d\right)\frac{1}{|V|} + d \sum_{v \in N(u)} \frac{\pi_{i-1}(v)}{\delta(v)}$$

with $d \in (0 : 1)$ is the damping factor. It converges to a stable distribution. **But it is not personalized.**

**Personalized PageRank [Jeh03]**

$$\pi_i(u) = \left(1 - d\right)p^*(u) + d \sum_{v \in N(u)} \frac{\pi_{i-1}(v)}{\delta(v)}$$

with $\sum p^*(u) = 1$. 

source: wikipedia
Let $G = (V, E)$ be the citation graph

**PageRank [Brin98]**

$$\pi_i(u) = (1 - d) \frac{1}{|V|} + d \sum_{v \in N(u)} \frac{\pi_{i-1}(v)}{\delta(v)}$$

with $d \in (0 : 1)$ is the damping factor. It converges to a stable distribution. **But it is not personalized.**

**Personalized PageRank [Jeh03]**

$$\pi_i(u) = (1 - d)p^*(u) + d \sum_{v \in N(u)} \frac{\pi_{i-1}(v)}{\delta(v)}$$

with $\sum p^*(u) = 1$. **But it is not exploratory.**
Direction Awareness

Time exploration

What if we are interested in searching papers per years. Recent papers? Traditional papers?

Let $M$ be a set of known relevant papers.

Direction Aware Random Walk with Restart

$$\pi_i(u) = (1 - d)p^*(u) + d((1 - \kappa) \sum_{v \in N^+(u)} \frac{\pi_{i-1}(v)}{\delta^-(v)} + \kappa \sum_{v \in N^-(u)} \frac{\pi_{i-1}(v)}{\delta^+(v)})$$

- $d \in (0 : 1)$ is the damping factor.
- $\kappa \in (0 : 1)$.
- $p^*(u) = \frac{1}{|M|}$, if $u \in M$, $p^*(u) = 0$, otherwise.
Exploring in Depth

The diagram illustrates the relationship between the average shortest distance and two parameters, \( \kappa \) and \( \sigma \). The color gradient represents the average shortest distance, with warmer colors indicating higher distances.

---

Umit V. Catalyurek
Ohio State University, Biomedical Informatics
HPC Lab http://bmi.osu.edu/hpc

TheAdvisor: http://theadvisor.osu.edu/
Citation Analysis::Direction Awareness 14 / 40
Exploring in Time

The diagram illustrates a heatmap with axes labeled for average publication year on the right and a parameter $\kappa$ on the x-axis. The y-axis represents another parameter $\sigma$. The color gradient indicates changes over years, possibly showing trends or patterns in the data.
The recovery test

Let’s hide some references from a paper and see if an algorithm can find them.

Results of the experiments with mean average precision (MAP@50) and 95% confidence intervals.

<table>
<thead>
<tr>
<th></th>
<th>hide random</th>
<th>hide recent</th>
<th>hide earlier</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>interval</td>
<td>mean</td>
</tr>
<tr>
<td>DARWR</td>
<td>48.00</td>
<td>46.80 49.20</td>
<td>42.22</td>
</tr>
<tr>
<td>P.R.</td>
<td>56.56</td>
<td>55.31 57.80</td>
<td>38.75</td>
</tr>
<tr>
<td>Katz_β</td>
<td>46.33</td>
<td>45.16 47.50</td>
<td>34.56</td>
</tr>
<tr>
<td>Cocit</td>
<td>44.60</td>
<td>43.39 45.80</td>
<td>14.22</td>
</tr>
<tr>
<td>Cocoup</td>
<td>17.28</td>
<td>16.36 18.20</td>
<td>17.56</td>
</tr>
<tr>
<td>CCIDF</td>
<td>18.05</td>
<td>17.11 19.00</td>
<td>18.97</td>
</tr>
</tbody>
</table>
Outline

1. Introduction
   - Why?
   - Overview

2. Citation Analysis for Document Recommendation
   - Previous Approaches
   - Direction Aware Recommendation

3. A High Performance Computing Problem
   - A specialization of SpMV
   - Ordering and Partitioning

4. Result Diversification

5. Other Features

6. Final Thoughts
   - Conclusion
   - Future Works
A Sparse Matrix-Vector Multiplication (SpMV)

Rewriting DARWR

\[
\pi_i(u) = (1 - d)p^*(u) + d \left( (1 - \kappa) \sum_{\nu \in N^+(u)} \frac{\pi_{i-1}(\nu)}{\delta^-(\nu)} + \kappa \sum_{\nu \in N^-(u)} \frac{\pi_{i-1}(\nu)}{\delta^+(\nu)} \right)
\]

\[
\pi_i(u) = (1 - d)p^*(u) + \sum_{\nu \in N^+(u)} \frac{d(1 - \kappa)}{\delta^-(\nu)} \pi_{i-1}(\nu) + \sum_{\nu \in N^-(u)} \frac{d\kappa}{\delta^+(\nu)} \pi_{i-1}(\nu)
\]

\[
\pi_i = (1 - d)p^* + A\pi_{i-1}
\]
A Sparse Matrix-Vector Multiplication (SpMV)

Rewriting \( \text{DaRWR} \)

\[
\pi_i(u) = (1 - d)p^*(u) + d \left( (1 - \kappa) \sum_{v \in N^+(u)} \frac{\pi_{i-1}(v)}{\delta^-(v)} + \kappa \sum_{v \in N^-(u)} \frac{\pi_{i-1}(v)}{\delta^+(v)} \right)
\]

\[
\pi_i(u) = (1 - d)p^*(u) + \sum_{v \in N^+(u)} \frac{d(1 - \kappa)}{\delta^-(v)} \pi_{i-1}(v) + \sum_{v \in N^-(u)} \frac{d\kappa}{\delta^+(v)} \pi_{i-1}(v)
\]

\[
\pi_i = (1 - d)p^* + A\pi_{i-1}
\]

CCS Full

- Traverse \( A \) column per column.
- Skip columns where \( \pi_{i-1}(v) = 0 \).
- Per edge: 2 non-zeros (2 indices, 2 values)
Rewriting DaRWR

\[
\pi_i(u) = (1 - d)p^*(u) + \sum_{v \in N^+(u)} \frac{d(1 - \kappa)}{\delta^-(v)} \pi_{i-1}(v) + \sum_{v \in N^-(u)} \frac{d\kappa}{\delta^+(v)} \pi_{i-1}(v)
\]

\[
\pi_i = (1 - d)p^* + B^- \frac{d(1 - \kappa)}{\delta^-} \pi_{i-1} + B^+ \frac{d\kappa}{\delta^+} \pi_{i-1}
\]
A Sparse Matrix-Vector Multiplication (SpMV)

Rewriting DaRWR

\[ \pi_i(u) = (1 - d)p^*(u) + \sum_{v \in N^+(u)} \frac{d(1 - \kappa)}{\delta^-(v)} \pi_{i-1}(v) + \sum_{v \in N^-(u)} \frac{d\kappa}{\delta^+(v)} \pi_{i-1}(v) \]

\[ \pi_i = (1 - d)p^* + B^- \frac{d(1 - \kappa)}{\delta^-} \pi_{i-1} \]

CCS Half

- pre-compute: \( \frac{(1-d)\kappa}{\delta^-} \pi_{i-1} \) and \( \frac{(1-d)(1-\kappa)}{\delta^+} \pi_{i-1} \)
- \( B^- \) and \( B^+ \) are 0/1 and symmetric
- Traverse the matrix twice (\( B^- \) and \( B^+ \))
- Skip columns where \( \pi_{i-1}(v) = 0 \)
- Per edge: 1 non-zeros (1 index, 0 values)
A Sparse Matrix-Vector Multiplication (SpMV)

Rewriting DaRWR

\[
\pi_i(u) = (1 - d)p^*(u) + \sum_{v \in N^+(u)} \frac{d(1 - \kappa)}{\delta^-(v)} \pi_{i-1}(v) + \sum_{v \in N^-(u)} \frac{d\kappa}{\delta^+(v)} \pi_{i-1}(v)
\]

\[
\pi_i = (1 - d)p^* + B^- \frac{d(1 - \kappa)}{\delta^-} \pi_{i-1} + B^+ \frac{d\kappa}{\delta^+} \pi_{i-1}
\]

COO Half

- pre-compute: \(\frac{(1-d)\kappa}{\delta^-} \pi_{i-1}\) and \(\frac{(1-d)(1-\kappa)}{\delta^+} \pi_{i-1}\)
- \(B^-\) and \(B^+\) are 0/1 and symmetric
- Traverse the matrix once (\(B^-\) and \(B^+\))
- Arbitrary order. Don’t skip anything.
- Per edge: 1 non-zeros (2 indices, 0 values)
Number of updates

iteration

# updates

- CRS-Full
- CRS-Half
- COO-Half

Umit V. Catalyurek
Ohio State University, Biomedical Informatics
HPC Lab http://bmi.osu.edu/hpc

theadvisor: http://theadvisor.osu.edu/
A HPC computing problem::SpMV

20 / 40
Runtimes

![Runtimes Graph]

- **CRS-Full**
- **CRS-Full (RCM)**
- **CRS-Full (AMD)**
- **CRS-Full (SB)**
- **CRS-Half**
- **CRS-Half (RCM)**
- **CRS-Half (AMD)**
- **CRS-Half (SB)**
- **COO-Half**
- **COO-Half (RCM)**
- **COO-Half (AMD)**
- **COO-Half (SB)**
- **Hybrid**
- **Hybrid (RCM)**
- **Hybrid (AMD)**
- **Hybrid (SB)**

**execution time (s)**

**#partitions**

---

[Ohio State University, Biomedical Informatics HPC Lab](http://bmi.osu.edu/hpc)
Locality

SpMV is sensitive to non-zero locality.
Ordering

Locality
SpMV is sensitive to non-zero locality.

Reverse Cuthill-McKee [Cuthill, McKee, 69]
Order with respect to a Breadth First Search ordering. (Do 10 times, pick best)
Ordering

Locality
SpMV is sensitive to non-zero locality.

Reverse Cuthill-McKee [Cuthill, McKee, 69]
Order with respect to a Breadth First Search ordering. (Do 10 times, pick best)

Approximate Minimum Degree [Amestoy et al.,96]
Greedily, add the vertex whose degree is minimum.
SpMV is sensitive to non-zero locality.

Order with respect to a Breadth First Search ordering. (Do 10 times, pick best)

Greedily, add the vertex whose degree is minimum.

Order by connected components. Remove the highest degree vertex. Repeat.
Partitioning
Run times

![Graph showing run times for different partition sizes and data formats.](image-url)
The goal of diversity is to avoid clustered answers.
The goal of diversity is to avoid clustered answers.
The goal of diversity is to avoid clustered answers.

Relevant

Relevant Diverse
**GRASSHOPPER [Zhu07]**

Vertex selection with absorbing random walks. It greedily selects the highest ranked vertex at each step and turns it into a sink for the next steps.
**GRASSHOPPER [Zhu07]**

Vertex selection with absorbing random walks. It greedily selects the highest ranked vertex at each step and turns it into a sink for the next steps.

---

**DIVRANK [Mei10]**

Greedy vertex selection with the vertex reinforced random walk (VRRW) model. It updates the transition matrix at each iteration with respect to the current/cumulative ranks of the nodes to introduce a *rich-gets-richer* mechanism to the ranking.
**GrassHopper [Zhu07]**

Vertex selection with absorbing random walks. It greedily selects the highest ranked vertex at each step and turns it into a sink for the next steps.

**DivRank [Mei10]**

Greedy vertex selection with the vertex reinforced random walk (VRRW) model. It updates the transition matrix at each iteration with respect to the current/cumulative ranks of the nodes to introduce a *rich-gets-richer* mechanism to the ranking.

**Dragon [Li11]**

Combines relevancy and diversity in *goodness* measure. *Dragon* produces near-optimal solutions.
Problems with bicriteria optimization

Here is a distribution of known algorithms
Problems with bicriteria optimization

Would such an algorithm be of interest?
Problems with bicriteria optimization

Would such an algorithm be of interest?
That algorithm is random!
Our Method: Relaxed Local Maxima ($\gamma$-RLM)

**Intuition:** Instead of top-$k$ results, find top-$\gamma k$ results, and incrementally get local maximas until $|S| = k$. Remove the selected vertices from the subgraph for the next local maxima selection.

**Algorithm 1:** Diversify with $\gamma$-RLM

Input: $G' = (V, E')$, $\pi$, $k$, $\gamma$: relaxation parameter
Output: A list of recommendations $S$

$T \leftarrow \text{SORT}(V)$ w.r.t. $\pi_i$ non-increasing
$R \leftarrow T[1 : \gamma k]$

while $|S| < k$ do
   $R' \leftarrow \text{FINDLOCALMAXIMAS}(G, R, \pi)$
   if $|R'| > k - |S|$ then
      $\text{SORT}(R')$ w.r.t. $\pi_i$ non-increasing
      $R' \leftarrow R'[1 : (k - |S|)]$
   $S \leftarrow S \cup R'$
   $R \leftarrow R \setminus R'$
return $S$

First iteration
$R' = \{r_1, \ldots, r_{15}\}$
$S = \{r_1, \ldots, r_{15}\}$

Second iteration
$R' = \{r_{16}, \ldots, r_{20}\}$
$S = \{r_1, \ldots, r_{20}\}$
GrassHopper and GSparse perform worse based on the diversity measures.

DivRank sacrifices direction-awareness to improve diversity.

FEED and Dragon return almost the same result set as top-k.

IL1 and IL2 have very low usefulness.

γ-RLM performs good overall, runs very efficiently, better choice for the service overall.
Results
Outline

1 Introduction
   • Why?
   • Overview

2 Citation Analysis for Document Recommendation
   • Previous Approaches
   • Direction Aware Recommendation

3 A High Performance Computing Problem
   • A specialization of SpMV
   • Ordering and Partitioning

4 Result Diversification

5 Other Features

6 Final Thoughts
   • Conclusion
   • Future Works
Papers can be tagged are relevant or irrelevant.

- **Positive feedback** (+RF): Relevant results are added to $Q$
- **Negative feedback** (-RF): Irrelevant results are removed from the graph

How long does it take to find the first level-3 paper?

![Graph showing the time to find the first level-3 paper with different feedback options]
Papers can be tagged as relevant or irrelevant.

- **Positive feedback (+RF):** Relevant results are added to $Q$.
- **Negative feedback (-RF):** Irrelevant results are removed from the graph.

More exploration!
Citation Graph Visualization

Legend: references relevant recommended top-100

Notes (1)

Pedro C. Driz, Steve Plimpton, Bruce Hendrickson, Robert W. Leander

Load Balancing for the Numerical Solution of the Navier-Stokes Equations.
Gregory Karagiorgos, Petros Katsafados, Andreas Kontarinis, Nikolaos M. Missirlis, Filippo Tzaferis
the advisor can be accessed programmatically. Emit HTTP requests and obtain JSON, XML, PHP encoded replies.
The advisor can be accessed programmatically. Emit HTTP requests and obtain JSON, XML, PHP encoded replies.

Potential Applications

- Interfacing with article editors (e.g., TexShop)
- Recommendation in bibliography manager (e.g., Mendeley)
- Suggesting reviewers to program committees (e.g., EasyChair)
- Suggesting sessions of interest at conferences (e.g., iConference)
Application Programming Interface

The advisor can be accessed programmatically. Emit HTTP requests and obtain JSON, XML, PHP encoded replies.

Potential Applications

- Interfacing with article editors (e.g., TexShop)
- Recommendation in bibliography manager (e.g., Mendeley)
- Suggesting reviewers to program committees (e.g., EasyChair)
- Suggesting sessions of interest at conferences (e.g., iConference)

Easier to use!
# theadvisor: v1 vs v2

<table>
<thead>
<tr>
<th>Citation Graph</th>
<th>v1</th>
<th>v2</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Vertices</td>
<td>1M</td>
<td>11M</td>
</tr>
<tr>
<td>#Edges</td>
<td>6M</td>
<td>33M</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sources</th>
<th>v1</th>
<th>v2</th>
</tr>
</thead>
<tbody>
<tr>
<td>dblp</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>citeseer (oai)</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>citeseer (raw)</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>inria</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>arnetminer</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>arXiv</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PMC</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Last Update</td>
<td>Dec’11</td>
<td>Apr’13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Features</th>
<th>v1</th>
<th>v2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visualization</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>Editable metadata</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>Searchable(sortable results)</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>1 Graph (CS+PMC)</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>better API for exploratory search</td>
<td>×</td>
<td>✓</td>
</tr>
</tbody>
</table>
Outline

1. Introduction
   - Why?
   - Overview

2. Citation Analysis for Document Recommendation
   - Previous Approaches
   - Direction Aware Recommendation

3. A High Performance Computing Problem
   - A specialization of SpMV
   - Ordering and Partitioning

4. Result Diversification

5. Other Features

6. Final Thoughts
   - Conclusion
   - Future Works
<table>
<thead>
<tr>
<th>Design Goals - Are they matched?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Personalized</strong></td>
</tr>
<tr>
<td>The query is expressed in a very precise way.</td>
</tr>
<tr>
<td><strong>Conceptual</strong></td>
</tr>
<tr>
<td>Using the citation graph allows to avoid all linguistic. Though, it may not be enough to find all relevant papers.</td>
</tr>
<tr>
<td><strong>Exploratory</strong></td>
</tr>
<tr>
<td>Direction Awareness (to choose time), Diversification (to see more topics), Visualization (for manual crawling)</td>
</tr>
<tr>
<td><strong>Easy to use</strong></td>
</tr>
<tr>
<td>Efficient (recommendation in less than 2 seconds), web-based.</td>
</tr>
</tbody>
</table>

*Ümit V. Çatalyürek*

Ohio State University, Biomedical Informatics

HPC Lab [http://bmi.osu.edu/hpc](http://bmi.osu.edu/hpc)

*theadvisor*: [http://theadvisor.osu.edu/](http://theadvisor.osu.edu/)
## Design Goals - Are they matched?

### Personalized
The query is expressed in a very precise way.

### Conceptual
Using the citation graph allows to avoid all linguistic. Though, it may not be enough to find all relevant papers.

### Exploratory
Direction Awareness (to choose time), Diversification (to see more topics), Visualization (for manual crawling)

### Easy to use
Efficient (recommendation in less than 2 seconds), web-based.

---

Is it good enough? Tell us!

---

Ümit V. Çatalyurek  
Ohio State University, Biomedical Informatics  
HPC Lab [http://bmi.osu.edu/hpc]  
theadvisor: [http://theadvisor.osu.edu/]

Final Thoughts::Conclusion  38 / 40
Future works

Clustering
Let’s assume for an instant that we have accurate disambiguated tags for every document. We could restrict analysis to some fields. Improve diversification.

Betweenness Centrality
DaRWR provides recommendation around the query set. What about recommending what is between it?

Contextual information
Distinguishing types of papers and citations. Survey, Method, Application...
Thank you

More information

contact: umit@bmi.osu.edu
visit: http://theadvisor.osu.edu
(or http://bmi.osu.edu/hpc/
or http://bmi.osu.edu/~umit)

Research at HPC lab is supported by