Academic Recommendation using Citation Analysis with theadvisor

Erik Saule
joint work with Onur Küçüktunç, Kamer Kaya, Ümit V. Çatalyurek

esaule@bmi.osu.edu

Department of Biomedical Informatics
The Ohio State University

CSTA 2013
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
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) (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!

Erik Saule
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/)
Once upon a time: 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”
Once upon a time: 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”

After 6 months, unknown papers were still uncovered
Once upon a time: 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</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 where 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.
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

<table>
<thead>
<tr>
<th>Service</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBLP</td>
<td>List of CS papers with clean reference and disambiguated names.</td>
</tr>
<tr>
<td>Citeseer, Seer</td>
<td>Automatically crawled papers in CS. Give PDFs. Contain citation information, full text. Compute similarity.</td>
</tr>
<tr>
<td>CiteUlike</td>
<td>Social paper tagging application. Find paper from researchers with similar interest.</td>
</tr>
</tbody>
</table>

---

Erik Saule
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/)
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

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

**Mendeley**
Application for managing references. Database of reference.

**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>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>
</tbody>
</table>

Erik Saule  
Ohio State University, Biomedical Informatics  
HPC Lab [http://bmi.osu.edu/hpc](http://bmi.osu.edu/hpc)
# 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>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><strong>Microsoft Academic Search</strong></td>
<td>Keyword-based search engine and Academic network analysis.</td>
</tr>
</tbody>
</table>

---

Erik Saule  
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>DBLP</td>
<td>List of CS papers with clean reference and disambiguated names.</td>
</tr>
<tr>
<td>CiteSeer, {Ref,Ack,Collab}Seer</td>
<td>Automatically crawled papers in CS. Give PDFs. Contain citation information, full text. Compute similarity.</td>
</tr>
<tr>
<td>Mendeley</td>
<td>Application for managing references. Database of reference.</td>
</tr>
<tr>
<td>CiteUlike</td>
<td>Social paper tagging application. Find paper from researchers with similar interest.</td>
</tr>
<tr>
<td>ArnetMiner</td>
<td>Academic network analysis.</td>
</tr>
<tr>
<td>Google Scholar</td>
<td>Keyword-based search engine (with citation informations).</td>
</tr>
<tr>
<td>Microsoft Academic Search</td>
<td>Keyword-based search engine and Academic network analysis.</td>
</tr>
<tr>
<td>IEEE, ACM, Elsevier, JSTOR, ...</td>
<td>Publishers or digital libraries with complete text and references. Some suggestions.</td>
</tr>
</tbody>
</table>
Put your references in a BibTeX or RIS file

@inproceedings{Kucuktunc12,
  author = {Kucuktunc, Onur and others},
  title = {A Large-Scale Sentiment Analysis System},
  booktitle = {Proc. 5th Int'l Conf. on Web Information Systems Engineering},
  year = {2012},
}

@inproceedings{Weber13,
  author = {Weber, Ina and others},
  title = {The demographics of sentiment analysis},
  booktitle = {Proc. 33rd Int'l Conf. on Research and Development in Information Retrieval},
  pages = {523--530},
  year = {2010},
}

@article{Aral11,
  author = {Aral, Sinan and others},
  title = {The influence of word choice on consumer perceptions of health},
  journal = {Health Psychology},
  year = {2011},
}

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

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

Get recommendations →

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

- Papers
  - Marina Drosou, Evaggelia Pitoura: Search result diversification. [bib] [good]
    SIGMOD Record, 2010.
  - Marina Drosou, Evaggelia Pitoura: Diversity over Continuous Data. [bib] [good]
  - Cong Yu, Laks V. S. Lakshmanan, Sihem Benseddik: Recommendation Diversification Using
    Group Dynamics. [bib] [good]
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
- .bib
- .ris
- .xml

parameters {k,d,κ}

dynamically

Recommendation Engine
- paper IDs

Venue Rec.
- venues

Reviewer Rec.
- reviewers

Diversification Engine
- papers

Visualization

Relevance Feedback

Erik Saule
Ohio State University, Biomedical Informatics
HPC Lab http://bmi.osu.edu/hpc
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
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

Erik Saule
Ohio State University, Biomedical Informatics
HPC Lab http://bmi.osu.edu/hpc
theadvisor: http://theadvisor.osu.edu/
Citation Analysis: 9 / 37
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

![Diagram showing reference and citation edges between nodes representing papers over time from 2001 to 2006.](http://example.com/diagram.png)
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]

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

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]

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

with $d \in (0 : 1)$ is the damping factor. It converges to a stable distribution. But it is not personalized.
Let $G = (V, E)$ be the citation graph

**PageRank [Brin98]**

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

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

**Personalized PageRank [Jeh03]**

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

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

---

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

### PageRank [Brin98]

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

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

### Personalized PageRank [Jeh03]

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

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

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

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

![Graph showing the relationship between $\kappa$ and $\sigma$ with the average shortest distance on the y-axis.](chart)

Erik Saule
Ohio State University, Biomedical Informatics
HPC Lab http://bmi.osu.edu/hpc

Citation Analysis::Direction Awareness 14 / 37
Exploring in Time

Citation Analysis::Direction Awareness
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></th>
<th>hide recent</th>
<th></th>
<th>hide earlier</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>interval</td>
<td>mean</td>
<td>interval</td>
<td>mean</td>
<td>interval</td>
</tr>
<tr>
<td>DaRWR</td>
<td>48.00</td>
<td>46.80 49.20</td>
<td>42.22</td>
<td>40.95 43.50</td>
<td>60.64</td>
<td>59.48 61.80</td>
</tr>
<tr>
<td>P.R.</td>
<td>56.56</td>
<td>55.31 57.80</td>
<td>38.75</td>
<td>37.50 40.00</td>
<td>58.93</td>
<td>57.76 60.10</td>
</tr>
<tr>
<td>Katz$_{\beta}$</td>
<td>46.33</td>
<td>45.16 47.50</td>
<td>34.56</td>
<td>33.42 35.70</td>
<td>44.19</td>
<td>42.97 45.40</td>
</tr>
<tr>
<td>Cocit</td>
<td>44.60</td>
<td>43.39 45.80</td>
<td>14.22</td>
<td>13.25 15.20</td>
<td>55.97</td>
<td>54.64 57.30</td>
</tr>
<tr>
<td>Cocoup</td>
<td>17.28</td>
<td>16.36 18.20</td>
<td>17.56</td>
<td>16.61 18.50</td>
<td>2.93</td>
<td>2.57 3.30</td>
</tr>
<tr>
<td>CCIDF</td>
<td>18.05</td>
<td>17.11 19.00</td>
<td>18.97</td>
<td>17.94 20.00</td>
<td>3.55</td>
<td>3.10 4.00</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
Rewriting DARWR

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

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

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

Rewriting $DA\text{RWR}$

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

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

$$\pi_i = dp^* + A\pi_{i-1}$$

CRS 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

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

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

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

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

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

CRS 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)
Rewriting DARWR

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

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

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

![Graph showing number of updates over iterations for different storage formats: CRS-Full, CRS-Half, and COO-Half. The graph compares the performance and updates count for each format across iterations.](image-url)
Runtimes

![Graph showing runtimes for different storage formats and partition sizes. The graph plots execution time (s) on the y-axis and #partitions on the x-axis. The storage formats include CRS-Full, CRS-Half, COO-Half, and Hybrid. The graph compares the performance of RCM, AMD, and SB algorithms.](image-url)
SpMV is sensitive to non-zero locality.
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.
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.

Slashburn [Kang, Faloutsos,11]
Order by connected components. Remove the highest degree vertex. Repeat.
Partitioning

Erik Saule
Ohio State University, Biomedical Informatics
HPC Lab http://bmi.osu.edu/hpc
theadvisor: http://theadvisor.osu.edu/
A HPC computing problem::Ordering
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

Erik Saule
Ohio State University, Biomedical Informatics
HPC Lab http://bmi.osu.edu/hpc

theadvisor: http://theadvisor.osu.edu/
Result Diversification:: 25 / 37
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.
A Modelization problem

Here is a distribution of known algorithms
A Modelization problem

Would such an algorithm be of interest?

Erik Saule
Ohio State University, Biomedical Informatics
HPC Lab http://bmi.osu.edu/hpc
theadvisor: http://theadvisor.osu.edu/
Result Diversification:: 27 / 37
Would such an algorithm be of interest?
That algorithm is random!

Erik Saule  
Ohio State University, Biomedical Informatics  
HPC Lab http://bmi.osu.edu/hpc
See later talk!
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 as relevant or irrelevant.

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

---

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

---

**More exploration!**
Visualization

theadvisor

Citation Graph Visualization

Load Balancing for the Numerical Solution of the Navier-Stokes Equations.
Gregory Karagiorgos, Petros Katsafados, Andreas Kontarinis, Nikolaos M. Missirlis, Filippos Tzaliris

Notes (1)

Pedro C. Diniz, Steve Plimpton, Bruce Hendrickson, Robert W. Leland
[dblp] [citeeseer] [pdf] [google]
More exploration!
Web service

the **advisor** can be accessed programmatically. Emit HTTP requests and obtain JSON encoded replies.
the **advisor** can be accessed programmatically. Emit HTTP requests and obtain JSON 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)
Web service

The advisor can be accessed programmatically. Emit HTTP requests and obtain JSON 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!
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
### Design Goals - Are they matched?

<table>
<thead>
<tr>
<th>Personalized</th>
<th>The query is expressed in a very precise way.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceptual</td>
<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>Exploratory</td>
<td>Direction Awareness (to choose time), Diversification (to see more topics), Visualization (for manual crawling)</td>
</tr>
<tr>
<td>Easy to use</td>
<td>Efficient (recommendation in less than 2 seconds), web-based.</td>
</tr>
</tbody>
</table>

Erik Saule
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>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>

Is it good enough? Tell us!

Erik Saule  
Ohio State University, Biomedical Informatics  
HPC Lab http://bmi.osu.edu/hpc  
theadvisor: http://theadvisor.osu.edu/
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: esaule@bmi.osu.edu
visit: http://theadvisor.osu.edu
(or http://bmi.osu.edu/hpc/
or http://bmi.osu.edu/~esaule)

Research at HPC lab is supported by

[Logos of various sponsors]