Building Babel

Large Scale Data Collection in the Cloud
Ian Wesley-Smith
iwsmit@uw.edu
Scholarly Article Recommendation

• Information Overload
  – 50m – 150m articles in existence
Google Scholar

- Recommendation vs Search
  - Serendipity
- Homonymity
- Synonymity

Explicating opinion leadership: Nonpolitical dispositions, information consumption, and civic participation
DV Shah, DA Scheufele - Political Communication, 2006 - Taylor & Francis

Information consumption by Reissner-Nordström black holes
A Strominger, SP Trvceci - Physical Review D, 1993 - APS

Personal information management: from information consumption to curation
S. Whittaker - Annual review of information sciences, 2011 - Wiley Online Library
Netflix/Spotify/Amazon

• User ratings (explicit, implicit)
• Density
  – # user-item interactions >> # items
• Netflix Competition (2006)¹
  – 100m ratings
  – 480k users
  – 17k movies

¹: http://www.netflixprize.com/community/viewtopic.php?id=68
Barriers to Research

• Hard to get datasets
• Difficult to measure effectiveness
  – Judges
  – Citation prediction
Enter Babel

• Provide access to private data sets
• Provide scholarly article recommendations, freely to anyone
  – Feedback data in return
• Evaluate recommenders using usage data
  – With enough traffic could be very fast
Audience

• Publishers
  – Offload expensive research into recommender systems to academia
  – Better recommendations drive more traffic/purchases

• Tool Developers

• Researchers
Requirements

- Fast
- Reliable
- Scalable (lots of data!)
- Easy to use
- Cheap
REST API

curl http://babel-us-east-1.eigenfactor.org/recommendation/aminer/12345
{
   "transaction_id": "46bb84190e9ddfd17700bfafb500ab3c",
   "results": [
      {
         "paper_id": "672",
         "publisher": "aminer"
      },
      {
         "paper_id": "11274",
         "publisher": "aminer"
      }
   ]
}
http://babel.eigenfactor.org
Browser Plugins

[Image of Firefox and Chrome browsers]

Articles

The Eigenfactor™ metrics
CT Bergstrom, JD West - The Journal of ... 2008 - Soc Neuroscience

Quantitative metrics are poor choices for assessing the research output of an individual scholar. Summing impact factors, counting citations, tallying an h-index, or looking at Eigenfactor™ scores (described below)—none of these methods are adequate compared ... Cited by 234 Related articles All 11 versions Cite Save

Assessing citations with the Eigenfactor™ metrics
CT Bergstrom, JD West - Neurology, 2008 - AAN Enterprises

For more than 80 years, researchers and administrators alike have evaluated the prestige and productivity of researchers, institutions, journals, and even nations by counting citations. 1 For the past half-century, the impact factor 2 has been the most prominent of these ... Cited by 91 Related articles All 11 versions Cite Save

Articles

The Eigenfactor™ metrics
CT Bergstrom, JD West - The Journal of ... 2008 - Soc Neuroscience

Quantitative metrics are poor choices for assessing the research output of an individual scholar. Summing impact factors, counting citations, tallying an h-index, or looking at Eigenfactor™ scores (described below)—none of these methods are adequate compared ... Cited by 234 Related articles All 11 versions Cite Save

Assessing citations with the Eigenfactor™ metrics
CT Bergstrom, JD West - Neurology, 2008 - AAN Enterprises

For more than 80 years, researchers and administrators alike have evaluated the prestige and productivity of researchers, institutions, journals, and even nations by counting citations. 1 For the past half-century, the impact factor 2 has been the most prominent of these ... Cited by 91 Related articles All 11 versions Cite Save
Cross-Scale Morphology, Geometry, and Dynamics of Ecosystems

C. S. Holling
Ecological Monographs
Ecological Society of America

Stable URL: www.jstor.org/stable/2937313

Abstract
This paper tests the proposition that a small set of plant, animal, and abiotic processes structure ecosystems across scales in time and space. Earlier studies have suggested that these key structuring processes establish a small number of dominant temporal frequencies that entrain other processes. These frequencies often differ from each other by at least an order of ...

Topics
Boreal forests, Animal physiology, Prairies, Mammals, Terrestrial ecosystems, Landscapes, Birds, Ecosystems, Animals, Species
Babel Architecture
<table>
<thead>
<tr>
<th>Frontend</th>
<th>Application</th>
<th>Package</th>
<th>Deploy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Flask</td>
<td>docker</td>
<td>AWS Elastic Bean Stalk</td>
</tr>
</tbody>
</table>
Swagger UI

GET /recommendation/{publisher}/{paper_id}

Implementation Notes
Generates recommendations for paper_id

Response Class (Status 200)

Model | Model Schema

```
{
    "results": [
        {
            "paper_id": "string",
            "publisher": "string"
        },
        {
            "transaction_id": "string"
        }
    ]
}
```

Response Content Type | application/json

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Description</th>
<th>Parameter Type</th>
<th>Data Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>paper_id</td>
<td>(required)</td>
<td>Publisher assigned identifier of a paper</td>
<td>path</td>
<td>string</td>
</tr>
<tr>
<td>publisher</td>
<td>aminer</td>
<td>Publisher to perform this operation on</td>
<td>path</td>
<td>string</td>
</tr>
</tbody>
</table>
Swagger UI


Request URL
http://babel-us-east-1.eigenfactor.org/recommendation/aminer/12345?limit=5

Response Body
```
{
  "transaction_id": "d403eb04889b9d782f3b14c983bec6",
  "results": [
    {
      "paper_id": "235114",
      "publisher": "aminer"
    },
    {
      "paper_id": "24108",
      "publisher": "aminer"
    },
    {
      "paper_id": "121114",
      "publisher": "aminer"
    },
    {
      "paper_id": "1098220",
      "publisher": "aminer"
    },
    {
      "paper_id": "12586220",
      "publisher": "aminer"
    }
  ]
}
```
<table>
<thead>
<tr>
<th>Frontend</th>
<th>Application</th>
<th>Package</th>
<th>Deploy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Flask</td>
<td>Docker</td>
<td>AWS Elastic Bean Stalk</td>
</tr>
</tbody>
</table>
FROM python:3-onbuild

WORKDIR /var/app

RUN pip3 install virtualenv

RUN virtualenv /var/app

RUN useradd uwsgi -s /bin/false

RUN mkdir /var/log/uwsgi

RUN chown -R uwsgi:uwsgi /var/log/uwsgi

ADD . /var/app

RUN if [-f /var/app/requirements.txt ]; then /var/app/bin/pip install -r /var/app/requirements.txt; fi

ENV BABEL_STAGE beta

EXPOSE 8080

CMD ["python", ".src/application.py"]
Frontend

Application

Package

Deploy

AWS Elastic Bean Stalk
DynamoDB

- AWS NoSQL
  - Key-value store
- Very fast (<10ms)
- Very scalable
  - Specify throughput
- Not too expensive
Issues

• Not all AWS services are created equal
  – Data Pipeline
  – Cloud Search
• Documentation
• SDK/Tooling
• Python & GIL
• Access Keys
Future Directions

- Finish backend
- Expand clients (publishers, tool developers)
- Actually get more recommenders
- Babel 3.0 – simple middleware
  - Automatically logs & add transaction info to outgoing requests