Discovering and Ranking Web Services with BASIL:
A Personalized Approach with Biased Focus

James Caverlee, Ling Liu, and Daniel Rocco
College of Computing
Georgia Institute of Technology
Categorization-based Service Discovery

- Find all stock ticker services:
Categorization-based Service Discovery

- Find all stock ticker services:
Categorization-based Service Discovery

- The UDDI approach
- Group services based on common properties
  - All stock ticker services
  - All services offered by New York companies
  - ...
- A user can search on properties or browse the registry to find candidate matches
Personalized Relevance-Based Service Discovery

- Identify services based on their relationships to other services
  - Not supported by today’s registries
- Sample discovery tasks:
  - Find the top-ten services that offer more coverage than the BLAST services at NCBI
  - Which medical literature sites are more specialized than PubMed
  - …
Personalized Relevance-Based Service Discovery
Techniques for Service Discovery and Ranking

- Based on communities
  - Reputation systems
  - PageRank-style (?)
- Schema/Interface matching
  - Find the services with similar inputs, outputs
- Semantic matching
  - Using a markup like OWL
- **Instance/data matching**
  - Use the data that the service provides to better understand the service
  - Use that data to compare across services
Our Solution: BASIL

- **BiAsed Service dIcovery aLgorithm**
- Three key components:
  - Source-Biased Probing
  - Evaluation and ranking of services with Biased Focus
  - Identification of interesting relationships based on bi-lateral evaluation of biased focus
- Focuses on the nature and degree of **topical relevance**
- Avoids significant human intervention or hand-tuned categorization schemes
We focus on one type of web service

- Data-intensive web services
  - Access to huge amounts of data
  - Tools for searching, manipulating, and analyzing data
  - Examples: Amazon, Google, Lifesciences resources like BLAST (genetic sequence search)

- Unlike transactional services (e.g. for purchasing a box of pencils)
Modeling Data-Intensive Web Services

- **Service Summary**
  - Bag-of-words model
  - XML Tags and Text
- \( \text{ActualSummary}(S_i) = \{(t_1,w_1), (t_2,w_2), \ldots, (t_N,w_N)\} \)
Estimating Service Summaries

- Query-based Sampling [Callan ’99]
  - Send a query; retrieve top-m documents; repeat until stopping condition reached
  - EstSummary(PubMed) [only a fraction of all terms in Actual Summary]
  - Over text databases, need ~300 docs for high-quality estimated summaries

- Good at generating overall summaries
- But not necessarily good for comparing summaries (see paper)
  - Intuition: a service with broad coverage (like Google) will have few terms in common with a service with narrow coverage (like PubMed)
Source-Biased Probing

- Bias the estimate of the target towards the source of bias
  - $\text{EstSummary}_{\text{PubMed}}(\text{Google})$ vs. $\text{EstSummary}(\text{Google})$
- Hone in on what Google has in common with PubMed
Source-Biased Probing

```
SourceBiasedProbing(Source σ, Target τ)
For target service τ, initialize ESUMMARY_σ(τ) = ∅.
repeat
  Invoke the probe term selection algorithm
to select a one-term query probe q from the
source of bias ESUMMARY(σ).
  Send the query q to the target service τ.
  Retrieve the top-m documents from τ.
  Update ESUMMARY_σ(τ) with the terms and
    frequencies from the top-m documents.
until Stop probing condition is met.
return ESUMMARY_σ(τ)
```

Figure 1: Source-Biased Probing Algorithm
Probe Selection

- Uniform random selection
  - Prob(selecting term j) = 1 / N’

- Weighted random selection
  - Prob(selecting term j) = w_j / Sum_i(w_i)

- Weight-based selection
  - Select terms that occur the most times in all documents
  - Select terms that occur in the most documents

- Focal term probing
Instead of treating a source as a single collection of candidate probe terms, let’s try to break the source up into rough groups of co-occurring terms.

Cluster terms (not documents)
- $\text{Term}_j = \{(\text{doc}_1, w_{j1}), \ldots, (\text{doc}_M, w_{jM})\}$

Use off-the-shelf clustering algorithm to find $k$ focal term groups
- Simple KMeans, in this case
Probing with Focal Terms (2)

- Use round-robin selection to choose a probe from each focal term group

Table 1: Example Focal Terms for PubMed

|   | care, education, family, management, ...
|---|---------------------------------------------------------------------------------------|
| 2 | brain, gene, protein, nucleotide, ...
| 3 | clinical, noteworthy, taxonomy, ...
| 4 | experimental, molecular, therapy, ...
| 5 | aids, evidence, research, winter, ...
Evaluating and Ranking Services

- **Biased Focus**
  - Captures the topical focus of a target on the source
    - $\text{focus}_{\text{source}}(\text{Target})$
  - Should range from 0 (no focus) to 1 (complete focus)
  - Not a symmetric measure; for example:
    - $\text{focus}_{\text{PubMed}}(\text{Google}) = \text{high}$
    - $\text{focus}_{\text{Google}}(\text{PubMed}) = \text{low}$
Cosine-Based Biased Focus

- **Cosine**
  - normalized inner product
  - Independent of the vector length

\[
\text{Cosine}_{\sigma}(\tau) = \left( \frac{\sum_{k=1}^{N} w_{\sigma k} w_{\tau k}}{\sqrt{\sum_{k=1}^{N} (w_{\sigma k})^2} \cdot \sqrt{\sum_{k=1}^{N} (w_{\tau k})^2}} \right)
\]

Other metrics discussed in the paper
Identifying Interesting Relationships

- Consider two services: A and B
- Evaluate their relationship by understanding the focus of each with respect to the other
  - $\text{focus}_B(A)$ and $\text{focus}_A(B)$
- Relies on a family of lambda-parameters
- Example:
  - Let $\lambda_{\text{high}} = 0.9$
  - if $\text{focus}_B(A) > 0.9$ and $\text{focus}_A(B) > 0.9$, then A and B are lambda-equivalent
- Of course, determining the appropriate lambda is tricky!
Experimental setup

- Two datasets:
  - Newsgroups
    - 780 collections
    - 100-16,000 documents in each
    - 2.5GB total
  - Web collection – ‘in the wild’
    - 50 real web databases
    - 50 docs collected from each
Probing Efficiency

Figure 3: Probing Efficiency for 100 Pairs

ICSOC 2004
SBP Identifies High Quality Documents

Figure 5: Average Document Quality for 100 Pairs
Precision For 10 Source Newsgroups
# Ranking Web Sources

## Table 2: Identifying Web Sources Relevant to PubMed

<table>
<thead>
<tr>
<th>Query Bias</th>
<th>Source Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMA</td>
<td>Open Directory (13)</td>
</tr>
<tr>
<td>WebMD</td>
<td>Google (27)</td>
</tr>
<tr>
<td>Linux Journal</td>
<td>About (11)</td>
</tr>
<tr>
<td>HealthAtoZ</td>
<td>WebMD (2)</td>
</tr>
<tr>
<td>DevGuru</td>
<td>AMA (1)</td>
</tr>
<tr>
<td>FamilyTree Magazine</td>
<td>HealthAtoZ (4)</td>
</tr>
<tr>
<td>Mayo Clinic</td>
<td>Monster (22)</td>
</tr>
<tr>
<td>Novell Support</td>
<td>Mayo Clinic (7)</td>
</tr>
<tr>
<td>Random House</td>
<td>Random House (9)</td>
</tr>
<tr>
<td>January Magazine</td>
<td>BBC News (12)</td>
</tr>
</tbody>
</table>
## Relationships Relative to PubMed

<table>
<thead>
<tr>
<th>Service (S)</th>
<th>URL</th>
<th>Description</th>
<th>$\text{focus}_{PM}(S)$</th>
<th>$\text{focus}_{SS}(PM)$</th>
<th>Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>WebMD</td>
<td><a href="http://www.webmd.com">www.webmd.com</a></td>
<td>Health/Medical</td>
<td>0.23</td>
<td>0.18</td>
<td>$\lambda$-equivalent</td>
</tr>
<tr>
<td>AMA</td>
<td><a href="http://www.ama-assn.org">www.ama-assn.org</a></td>
<td>Health/Medical</td>
<td>0.19</td>
<td>0.16</td>
<td>$\lambda$-equivalent</td>
</tr>
<tr>
<td>HealthAtoZ</td>
<td><a href="http://www.healthatoz.com">www.healthatoz.com</a></td>
<td>Health/Medical</td>
<td>0.18</td>
<td>0.16</td>
<td>$\lambda$-equivalent</td>
</tr>
<tr>
<td>Open Directory</td>
<td>dmoz.org</td>
<td>Web Directory</td>
<td>0.44</td>
<td>0.08</td>
<td>$\lambda$-superset</td>
</tr>
<tr>
<td>Google</td>
<td><a href="http://www.google.com">www.google.com</a></td>
<td>Web Search Engine</td>
<td>0.37</td>
<td>0.10</td>
<td>$\lambda$-superset</td>
</tr>
<tr>
<td>About</td>
<td><a href="http://www.about.com">www.about.com</a></td>
<td>Web Channels</td>
<td>0.25</td>
<td>0.08</td>
<td>$\lambda$-superset</td>
</tr>
<tr>
<td>Monster</td>
<td><a href="http://www.monster.com">www.monster.com</a></td>
<td>Jobs</td>
<td>0.14</td>
<td>0.08</td>
<td>$\lambda$-overlap</td>
</tr>
<tr>
<td>Mayo Clinic</td>
<td><a href="http://www.mayoclinic.com">www.mayoclinic.com</a></td>
<td>Health/Medical</td>
<td>0.12</td>
<td>0.11</td>
<td>$\lambda$-overlap</td>
</tr>
<tr>
<td>Silicon Investor</td>
<td><a href="http://www.siliconinvestor.com">www.siliconinvestor.com</a></td>
<td>Finance</td>
<td>0.03</td>
<td>0.04</td>
<td>$\lambda$-complement</td>
</tr>
<tr>
<td>Usenet Recipes</td>
<td>recipes2.alastra.com</td>
<td>Recipes</td>
<td>0.02</td>
<td>0.03</td>
<td>$\lambda$-complement</td>
</tr>
</tbody>
</table>

More in the paper!
Conclusions

- Introduced techniques to support Personalized Relevance-Based Service Discovery
  - Source-biased probing
    - Focal term probing
  - Source-biased ranking (with biased focus)
  - Identification of relationships
Open issues

- Exploiting structure
  - E.g. for schema matching, use of ontologies, etc.
- More advanced probing techniques
- Fine-grained inter-service analysis
- Better understanding of complex service computations (e.g. correlating input to output)
- Could extend this “personalization” approach to consider other factors as well
Thank You!