Constructing Effective and Efficient Topic-Specific Authority Networks For Expert Finding in Social Media

Reyyan Yeniterzi & Jamie Callan

Social Media for Expert Search

- 72% of the companies use internal social media to find experts within the organization and improve collaboration
  - McKinsey Global Institute survey with >4200 companies

- 56% of the companies use social media for recruiting
  - SHRM 2011 survey on ‘Social Networking Websites and Staffing’
Expert Retrieval Background

- Expert Finding Task
  - TREC Enterprise Track 2005-2008
  - W3C and CSIRO Collections

- State-of-the-art Approaches
  - Profile-based Models [Balog, 2006]
  - Document-based Models [Balog, 2006; Macdonald, 2006]
  - Graph-based Models [Serdyukov, 2008]
  - Learning-based Models [Fang, 2010]

Expert Retrieval in Social Media

- Is writing topic-specific content enough for being considered an expert?
- One also needs to have topic-specific influence over other users
  - authority estimation
  - user authority networks
    - reading, commenting or voting
Outline

- Authority-based approaches
  - PageRank [Brin and Page, 1998]
  - Topic-Sensitive PageRank [Haveliwala, 2002]
  - HITS [Kleinberg, 1999]
- Topic-Candidate Graphs
- Experiments
  - Finding topic-specific expert bloggers
- Conclusion

PageRank (PR) [Brin and Page, 1998]

\[
PR(u) = \frac{1 - d}{|U|} + d \sum_{i \in I_u} \frac{PR(i)}{OL(i)}
\]

- Graph
  - topic-independent
    - all users
    - all user activities over all documents
Topic-Sensitive PageRank (TSPR)  
[Haveliwala, 2002]

- the PageRank graph  
- TSPR Approach
  - PageRank approach +
  - Teleportation is possible only to users that are associated with topic-relevant content

Hyperlink-Induced Topic Search (HITS)  
[Kleinberg, 1999]

- Hub: Sum of authority scores of outgoing edges  
- Authority: Sum of hub scores of incoming edges

- Applied to more topic-specific authority networks
  - to focus the computational effort on relevant nodes
Step 1: Retrieve an initial list of expert candidates, which is called the root set

Step 2: Expand root set into base set, which consists of users who are connected to/from users in the root set
Constructing HITS Graph

- Step 3: Use all users in base set as nodes and all existing interactions among them as edges.

Graph Properties: Nodes & Edges

PageRank Graph | HITS Graph
HITS on users

Topic-Candidate (TC) graphs
Constructing Topic-Candidate Graph

Step 1: Retrieve an initial list of expert candidates, which is called the root set.

Step 2: Expand root set into base set, which consists of users who are connected to/from users in root set due to topic-relevant interactions.
Comparison of Graphs

- PageRank Graph
- HITS Graph
- Topic-Candidate Graph

Experiments

- Finding topic-specific expert bloggers
  - Reading and commenting activity as authority signals
Intra-organizational blog collection from a large multinational IT firm

Access logs
- cover 44 of the 56 months of the collection

Dataset

Number of Posts: 165,414
Number of Comments: 783,356
Number of Employees: >100,000
Number of Posters: 20,354
Number of Commenters: 42,169
Number of Readers: 92,360

Evaluation Data

40 work related topics
- Selected from the access logs of company search engine
- Created by the company employees

Candidate Pools
- Top 10 candidates retrieved from content-based approaches

Assessments – (The collection is not public)
- Performed by author Yeniterzi
- 4-point scale
  - not an expert, some expertise, an expert, very expert
Authority Networks

Content-based Experiments

<table>
<thead>
<tr>
<th>Method</th>
<th>NDCG @1</th>
<th>NDCG @3</th>
<th>NDCG @10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profile [Balog, 2006]</td>
<td>.7000</td>
<td>.6689</td>
<td>.6494</td>
</tr>
<tr>
<td>Votes [MacDonald, 2006]</td>
<td>.3667</td>
<td>.4090</td>
<td>.4140</td>
</tr>
<tr>
<td>ReciprocalRank [MacDonald, 2006]</td>
<td>.7083</td>
<td>.7003</td>
<td>.7281</td>
</tr>
<tr>
<td>CombSUM [MacDonald, 2006]</td>
<td>.6417</td>
<td>.6334</td>
<td>.6168</td>
</tr>
<tr>
<td>CombMNZ [MacDonald, 2006]</td>
<td>.5333</td>
<td>.5295</td>
<td>.5124</td>
</tr>
<tr>
<td>IRW [Serdyukov, 2008]</td>
<td>.5167</td>
<td>.5189</td>
<td>.5159</td>
</tr>
</tbody>
</table>
Authority-based Re-ranking

\[ \text{final} = \text{content}^\alpha \text{reading}^\beta \text{commenting}^\gamma \]

where
\[ \alpha + \beta + \gamma = 1 \]

- Parameter optimization
- 5-fold cross validation

PageRank on Three Types of Graph

- MRR (VE) improvement is statistically significant with p < 0.05
- MAP (VE) improvement is statistically significant with p < 0.10
PageRank on Three Types of Graph

- **Ave. # unassessed candidates introduced**
  - NDCG@1
  - NDCG@10
  - MAP (VE)
  - MRR (VE)

- **Content Baseline**
- **PR Graph**
- **HITS Graph**
- **TC Graph**

MRR (VE) improvement is statistically significant with $p < 0.05$
MAP (VE) improvement is statistically significant with $p < 0.10$

TSPR on Three Types of Graph

- **Ave. # unassessed candidates introduced**
  - NDCG@1
  - NDCG@10
  - MAP (VE)
  - MRR (VE)

- **Content Baseline**
- **PR Graph**
- **HITS Graph**
- **TC Graph**

MRR (VE) improvement is statistically significant with $p < 0.05$
HITS on Three Types of Graph

Graph Size and Running Time Analysis

<table>
<thead>
<tr>
<th>Graph</th>
<th>Average # Nodes</th>
<th>Average # Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>C</td>
</tr>
<tr>
<td>PR</td>
<td>92K</td>
<td>43K</td>
</tr>
<tr>
<td>HITS</td>
<td>57K</td>
<td>14K</td>
</tr>
<tr>
<td>TC</td>
<td>7K</td>
<td>1K</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Approach</th>
<th>Graph</th>
<th>Approximate Running Times (in sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R</td>
</tr>
<tr>
<td>PR</td>
<td>PR</td>
<td>1,203</td>
</tr>
<tr>
<td>HITS</td>
<td>HITS</td>
<td>1,116</td>
</tr>
<tr>
<td>TC</td>
<td>TC</td>
<td>4</td>
</tr>
<tr>
<td>TSPR</td>
<td>PR</td>
<td>1,222</td>
</tr>
<tr>
<td></td>
<td>HITS</td>
<td>1,248</td>
</tr>
<tr>
<td></td>
<td>TC</td>
<td>2</td>
</tr>
<tr>
<td>HITS</td>
<td>PR</td>
<td>478</td>
</tr>
<tr>
<td></td>
<td>HITS</td>
<td>344</td>
</tr>
<tr>
<td></td>
<td>TC</td>
<td>3</td>
</tr>
</tbody>
</table>
Conclusion

- Topic-Candidate graphs
- Statistically significant improvements @ MRR (p<0.05) with PageRank and TSPR approaches
  - Effectiveness
    - 4% @ NDCG@1
    - 8% @ MAP(VE)
    - 17% @ MRR(VE)
  - Efficiency
    - Reading: 20 min to 2 sec
    - Commenting: 1 min to 0.4 sec

Thank you