Enhancing Expert Finding Using Organizational Hierarchies

Maryam Karimzadehgan (U. Illinois Urbana-Champaign)*,
Ryen White (MSR), Matthew Richardson (MSR)

Presented by Ryen White
Microsoft Research

*** MSR Intern, Summer 08**

Motivation for expert finding

- Some questions cannot be answered using a Web search engine
  - Involve tacit / procedural knowledge, internal org topics
- Some solutions:
  - Social connections (ask people, follow referrals)
    - Time-consuming in large organizations
  - Post to forum or mail distribution list
    - May be unanswered, interrupt many, high latency
  - Find one or more candidate experts and present the question to them
    - **Finding these experts is the challenge of expert finding…**

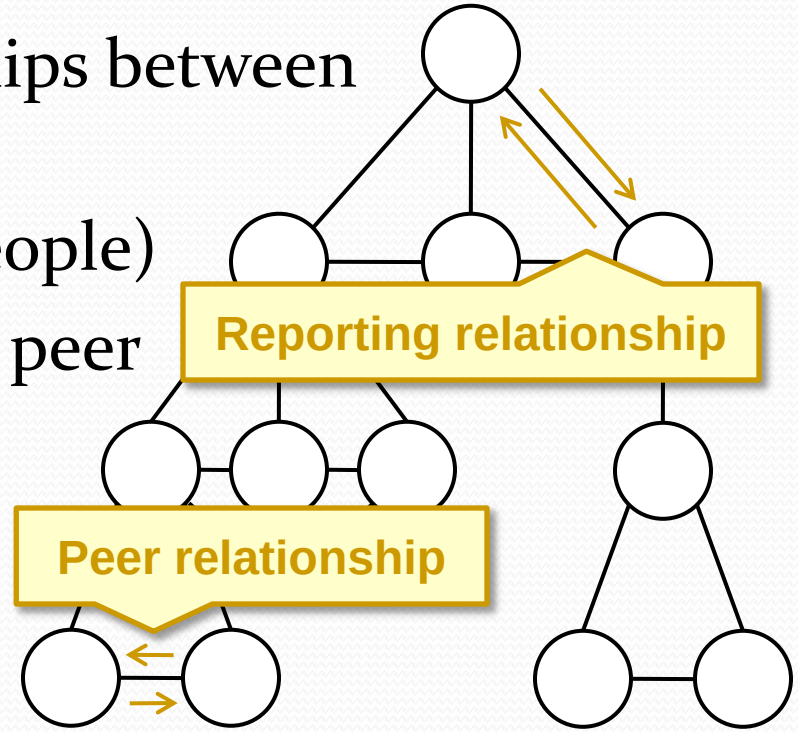
Overview

- Task in expert finding is to find people in an organization with expertise on query topic
- Profiles typically constructed for each member from sources such as email / shared documents

- **What if we don't have a profile for everyone?**
- Can we use organizational hierarchy to help us find experts without profiles and refine others' profiles?
- Propose and evaluate algorithm that considers org. member **and the expertise of his or her neighbors**

Organizational hierarchy

- Depicts managerial relationships between organizational members
- Nodes represent members (people)
- Links represent reporting and peer relationships
- Peers are members with the same direct manager



- **Can we use the hierarchy to improve expert finding by sharing expertise around the hierarchy?**

Does proximity ⇒ shared expertise?

- Before we can use neighbors as a proxy for a member's expertise we must know if their expertise is comparable
- People who work in the same group may have similar interests and expertise because:
  - They work on the same product
  - Their role is probably similar (dev, test, HR, legal, sales)
- Neighbors may be good proxies for those with no profile

- **But we should check to be sure…**

Does proximity ⇒ shared expertise?

- We conducted a study with Microsoft Corporation
- MS employs over 150,000 people, inc. temps/vendors
- By crawling internal email distribution lists we created profiles for 24% of employees via their sent mail
  - Demonstrates the challenge (76% had no profile)
- Selected random question from internal "idunno" list:

  Subject: *Standard clip art catalog or library*

  Body: *Do we have a corporate standard collection of clip art to use in presentations, specs, etc.?*
- Found candidates, asked them to rate own expertise

Does proximity ⇒ shared expertise?

- Asked for self-evaluation 0/1/2 = couldn't answer / some knowledge / could answer
- Emailed immediate neighbors same self-evaluation

Source member rating	Mean neighbor rating	N
0	0.45	46
1	0.86	39
2	1.41	61

- A organizational member's expertise correlates strongly neighbor expertise (caveat: for this particular question)
- Neighbors' expertise may be a good proxy for missing profiles or useful to refine existing profiles

Expert Modeling Techniques

Baseline

- Language-modeling approach
- Build profile based on email associated with person
- Compute probability that this model generates query

Text representation of expertise for j^{th} expert

Number of times word w occurs in e_j

Estimated from all expertise docs, E

$$p(q \mid e_j) = \prod_{w \in q} \frac{c(w, e_j) + \mu p(w \mid E)}{N_{e_j} + \mu}$$

Total number of words in e_j

Dirichlet prior – set empirically

Hierarchy-based algorithm

- Baseline only effective if we have email for all members
  - Since this is unlikely, we propose to use org. hierarchy
- All members scored w/ Baseline (many get zero score)
- Then, their scores are smoothed with neighbors

α weights member versus neighbors

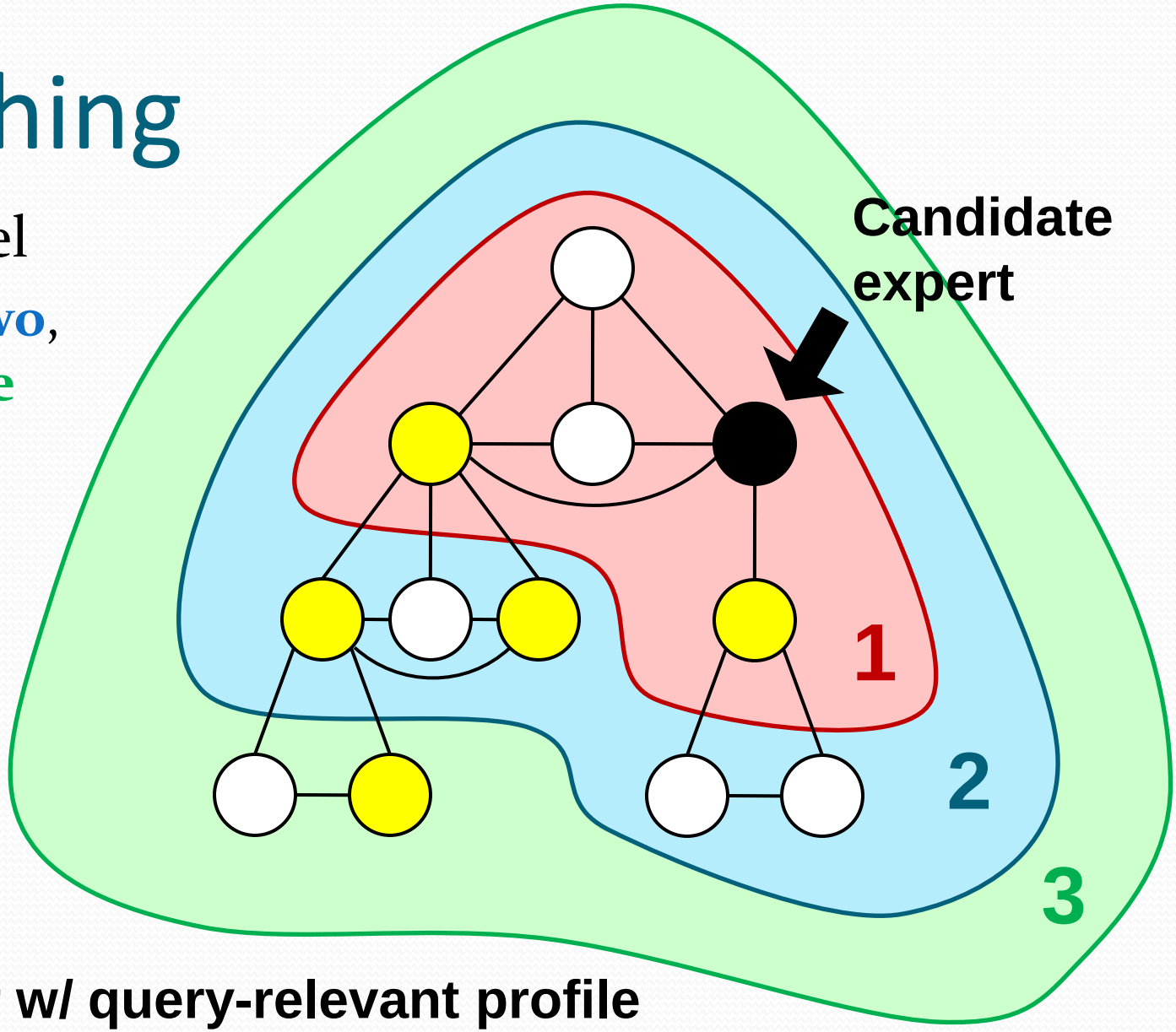
Initial scores using Baseline

$$p_{smooth}(q \mid e_j) = \alpha p(q \mid e_j) + \frac{(1-\alpha)}{N_j} \sum_{i=1}^{N_j} p(q \mid e_i)$$

Number of neighbors of *j*

Smoothing

- Multi-level
  - **One**, **two**, or **three**



= member w/ query-relevant profile

Evaluation

Expert profiling

- Profiles were constructed for organizational members
- Emails sent to internal discussion lists within MS
  - Stemmed text, only used text they wrote (not question)
  - "idunno" list was excluded from this crawl
- Average number of emails per employee = 29
- Median number of emails per employee = 6
- We have outgoing emails for only approximately 36,000 employees (there are ~153,000 employees)
  - We have information for only 24% of all employees

Expert-rating data

- Compare the baseline and hierarchy-based algorithms
- Expert rating data used as ground truth
- Devise and distribute survey with 20 randomly-selected questions from internal "idunno" discussion list
  - **Examples of questions from the list:** *Where can I get technical support for MS SQL Server? Who is the MS representative for college recruiting at UT Austin?*
- Survey was distributed to the 1832 member of the discussion list, 189 respondents rated their expertise as 0/1/2 for each of the 20 questions
  - 0/1/2 = couldn't answer / some knowledge / could answer

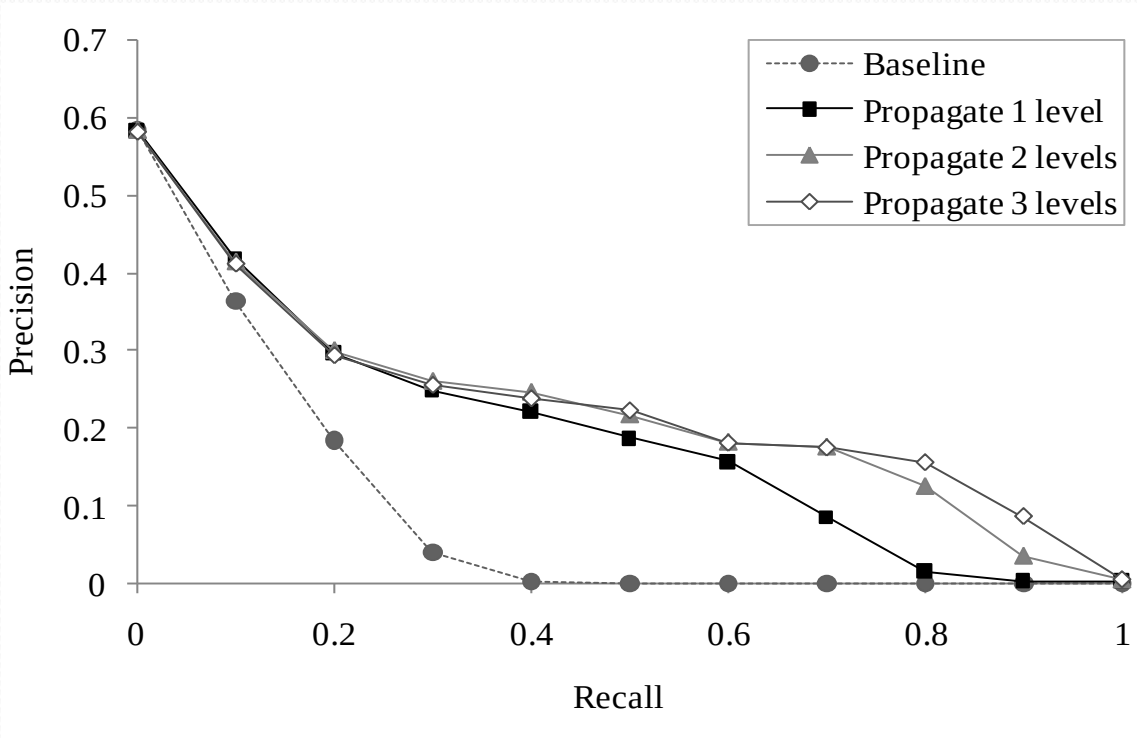
Methodology

- Baseline is sub-part of hierarchy-based algorithm
  - Allowed us to determine the effect of using hierarchy
- Set Dirichlet prior, μ, to 100 and the hierarchy smoothing parameter, α, to 0.9 - both determined empirically via parameter sweeps
- Used subjects of 20 selected questions as test queries
- Expert rating of 2 = relevant, 0/1 = non-relevant
- Generated a ranked list of employees using each alg.
- Computed precision-recall and avg. over all queries

Evaluation Results

Precision-recall

- Ranked all employees for each question
- Kept only those for whom we had ratings (189 total)
- Interpolated-averaged 11-point PR curve

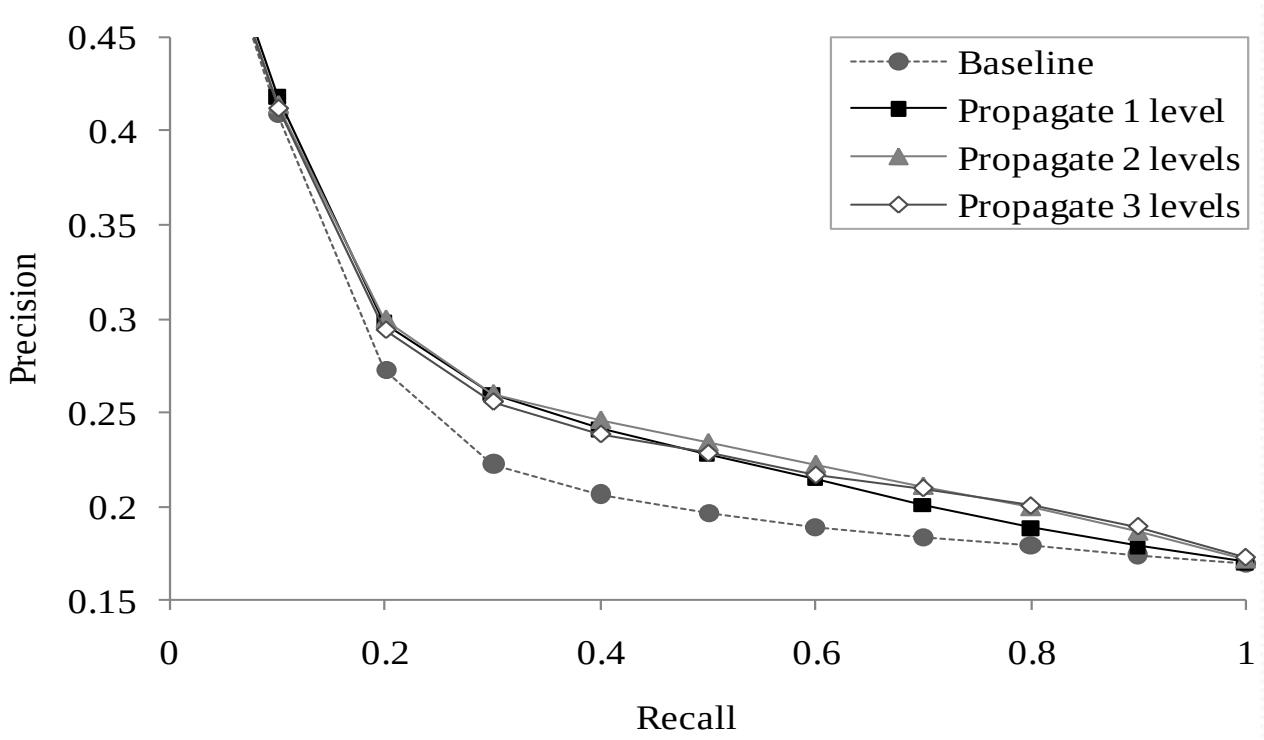


Precision-recall - ranking

- Prior findings could be explained by hierarchy-based algorithm returning more employees
- We used each algorithm to rank all employees
- We kept only those for which we had expert ratings, maintaining their relative rank order.
- We did not ignore rated employees that were not retrieved, but we appended them to the end of the result list in random order
- Computed precision-recall curves for each algorithm, where each point was averaged across 100 runs

Precision-recall - ranking

- Interpolated precision at zero for all alg. is approx. 0.58
- Hierarchy-based algorithm also better at ranking



Further opportunities

- We investigated propagating keywords around the hierarchy rather than scores
  - Keyword performance was significantly worse
    - Perhaps because of low keyword quality or a shortage of information about each employee (only a few emails each)
- Weighting edges between organizational members based on their relationship
  - Peer-to-peer ≠ manager-to-subordinate
- Experiment with other sources
  - Whitepapers, websites, communication patterns

Summary

- **Expertise representation:**
  - Use org. hierarchy to address data sparseness challenge when we lack information for all org. members
- **Expertise modeling:**
  - Hierarchy-based algorithm to share expertise info.
- **Evaluation:**
  - Org. hierarchy and human-evaluated data from Microsoft
- **Outcome:**
  - Org. hierarchy improves expert finding – useful on its own or perhaps as a feature in machine learning (future work)