ADVANCED TOPICS IN INFORMATION RETRIEVAL AND WEB SEARCH

Lecture 9:

Expert Finding

Dr. S M Vahidipour

vahidipour@kashanu.ac.ir

Outline

- **Introduction**
- Approaches
- Evaluation

Knowledge

- Some knowledge is not easy to find
  - Not stored in documents
  - Not stored in databases
  - **It is stored in peoples' minds!**

Definition

- Search scenario:
  - Let’s search for documents that are relevant to topic X.

- Expert finding scenario:
  - Let’s search for documents that are relevant to topic X.

People Expert

Task

- Ranking people based on a topic queried by use

People: Experts

Topic: Subject/Fields

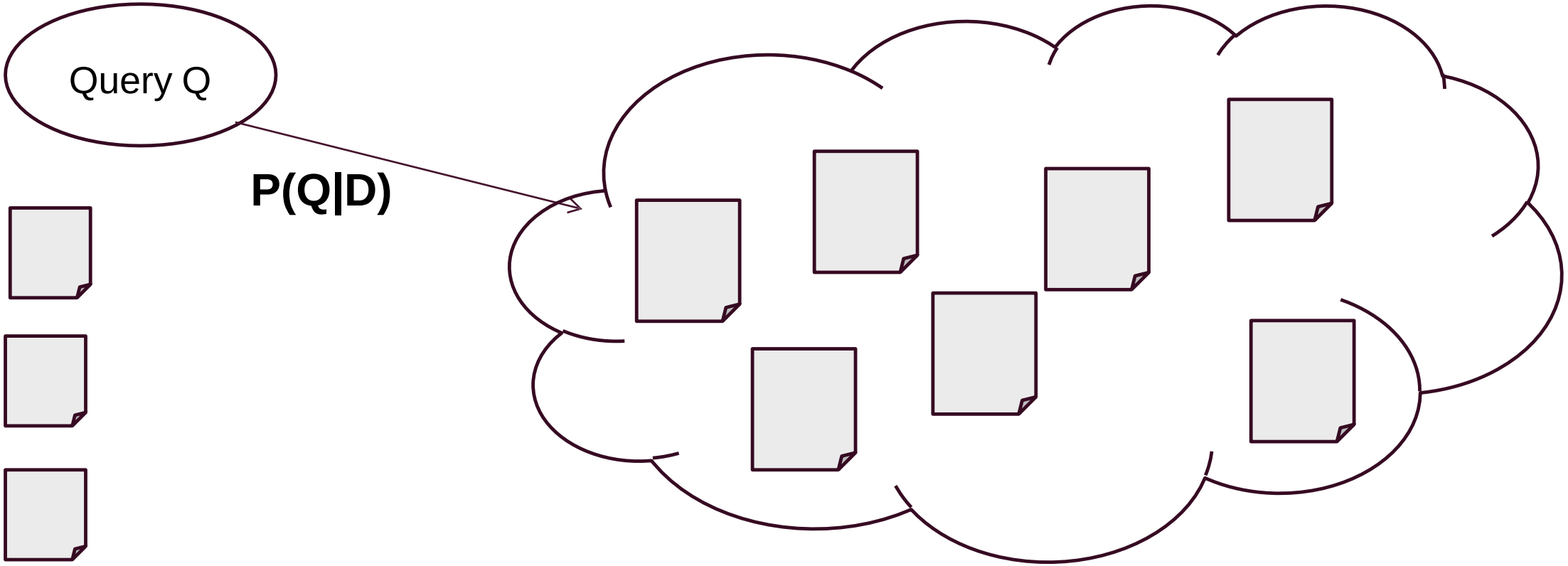
Applications

- Employers: Employees

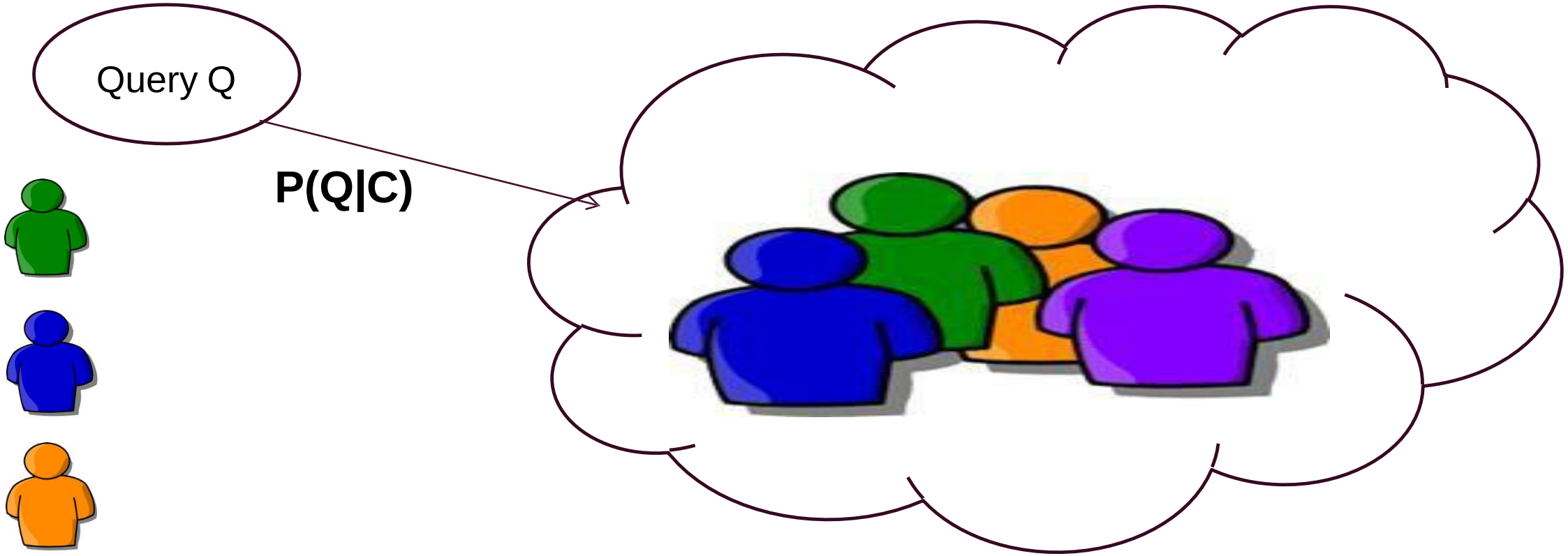
- Conference Committees: Reviewers

- Students: Professors

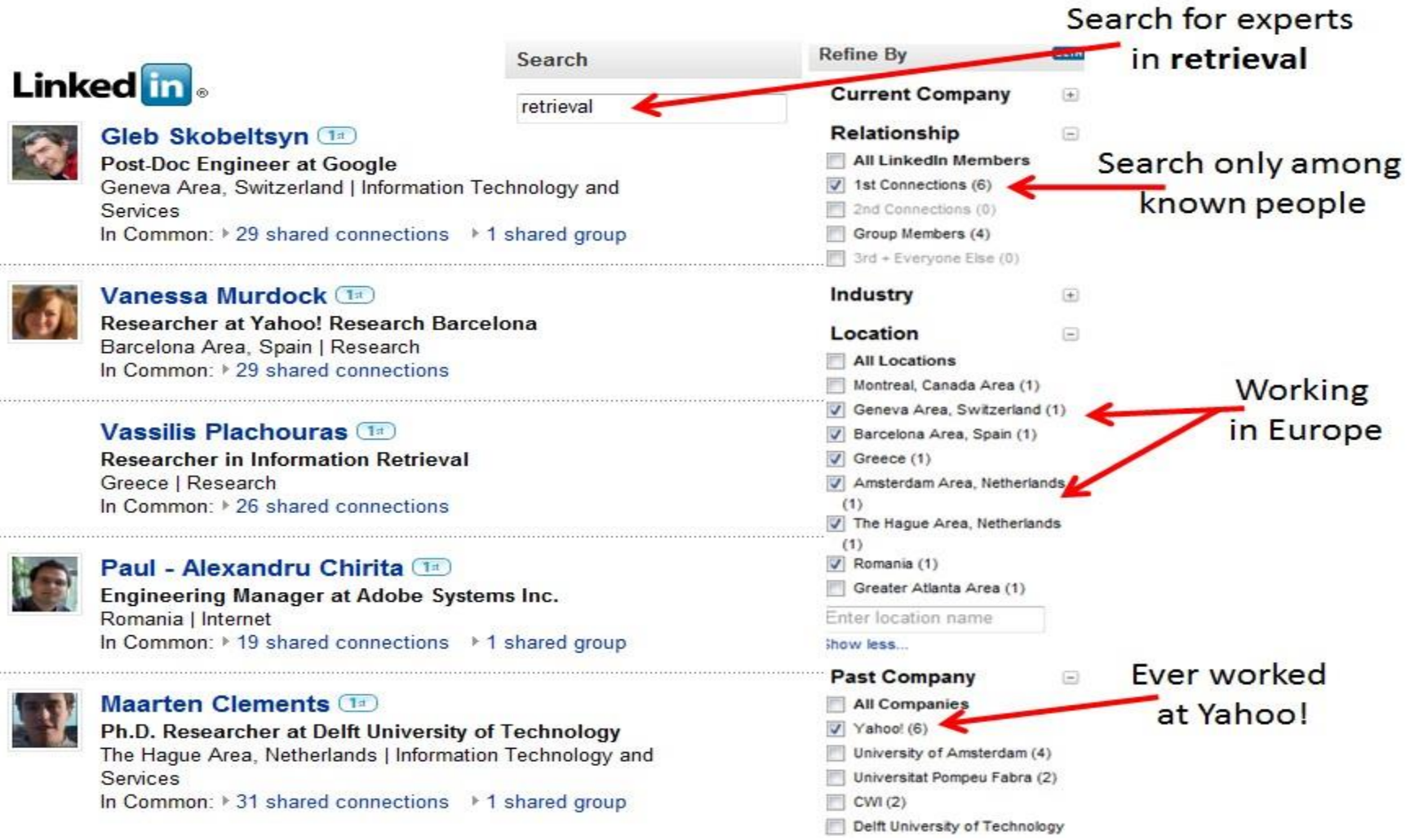
Document Retrieval vs. Expert Finding



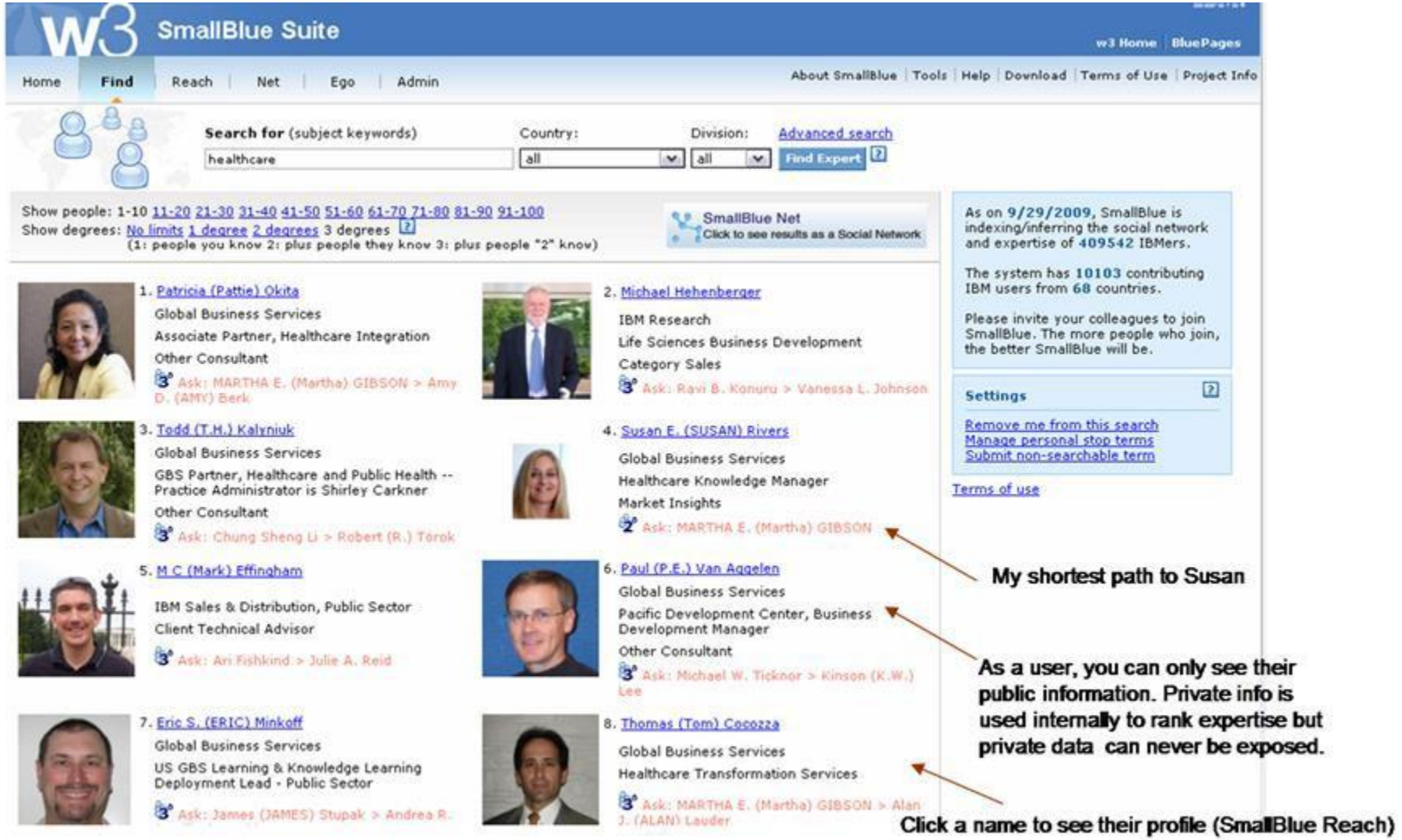
Document Retrieval vs. Expert Finding



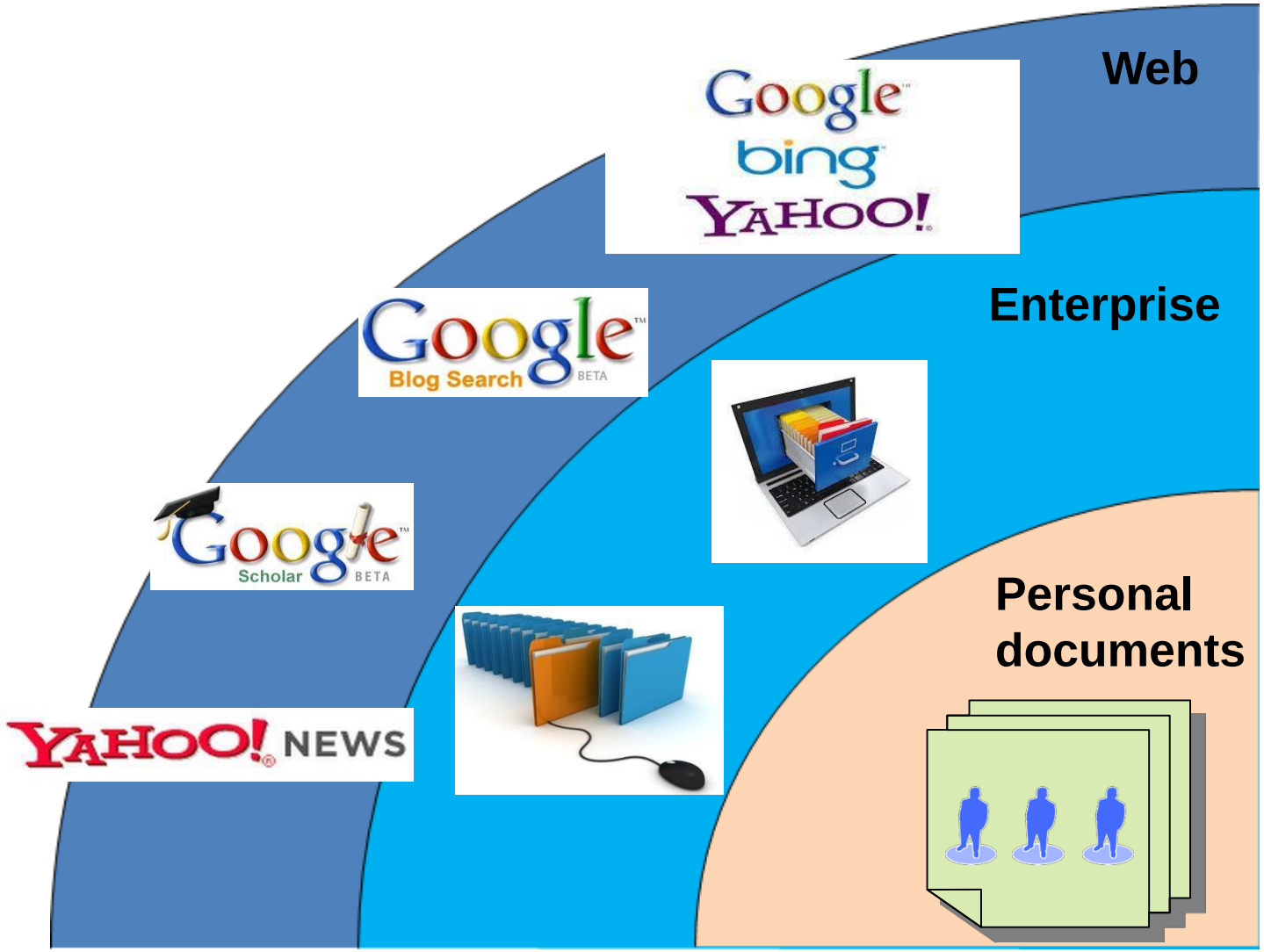
Example


LinkedIn
Search
retrieval
Refine By
Current Company
Relationship
All LinkedIn Members
1st Connections (6)
2nd Connections (0)
Group Members (4)
3rd + Everyone Else (0)
Industry
Location
All Locations
Montreal, Canada Area (1)
Geneva Area, Switzerland (1)
Barcelona Area, Spain (1)
Greece (1)
Amsterdam Area, Netherlands (1)
The Hague Area, Netherlands (1)
Romania (1)
Greater Atlanta Area (1)
Enter location name
Show less...
Past Company
All Companies
Yahoo! (6)
University of Amsterdam (4)
Universitat Pompeu Fabra (2)
CWI (2)
Delft University of Technology
Gleb Skobeltsyn 1st
Post-Doc Engineer at Google
Geneva Area, Switzerland | Information Technology and Services
In Common: 29 shared connections 1 shared group
Vanessa Murdock 1st
Researcher at Yahoo! Research Barcelona
Barcelona Area, Spain | Research
In Common: 29 shared connections
Vassilis Plachouras 1st
Researcher in Information Retrieval
Greece | Research
In Common: 26 shared connections
Paul - Alexandru Chirita 1st
Engineering Manager at Adobe Systems Inc.
Romania | Internet
In Common: 19 shared connections 1 shared group
Maarten Clements 1st
Ph.D. Researcher at Delft University of Technology
The Hague Area, Netherlands | Information Technology and Services
In Common: 31 shared connections 1 shared group
Search for experts in retrieval
Search only among known people
Working in Europe
Ever worked at Yahoo!

Example


SmallBlue Suite
w3 Home | BluePages
Home | Find | Reach | Net | Ego | Admin
About SmallBlue | Tools | Help | Download | Terms of Use | Project Info
Search for (subject keywords)
Country:
Division:
Advanced search
healthcare
all
all
Find Expert
Show people: 1-10 11-20 21-30 31-40 41-50 51-60 61-70 71-80 81-90 91-100
Show degrees: No limits 1 degree 2 degrees 3 degrees
(1: people you know 2: plus people they know 3: plus people "2" know)
SmallBlue Net
Click to see results as a Social Network
As on 9/29/2009, SmallBlue is indexing/inferring the social network and expertise of 409542 IBMers.
The system has 10103 contributing IBM users from 68 countries.
Please invite your colleagues to join SmallBlue. The more people who join, the better SmallBlue will be.
Settings
Remove me from this search
Manage personal stop terms
Submit non-searchable term
Terms of use
1. Patricia (Pattie) Okita
Global Business Services
Associate Partner, Healthcare Integration
Other Consultant
Ask: MARTHA E. (Martha) GIBSON > Amy D. (AMY) Berk
2. Michael Hehenberger
IBM Research
Life Sciences Business Development
Category Sales
Ask: Ravi B. Konuru > Vanessa L. Johnson
3. Todd (T.H.) Kalyniuk
Global Business Services
GBS Partner, Healthcare and Public Health -- Practice Administrator is Shirley Carkner
Other Consultant
Ask: Chung Sheng Li > Robert (R.) Török
4. Susan E. (SUSAN) Rivers
Global Business Services
Healthcare Knowledge Manager
Market Insights
Ask: MARTHA E. (Martha) GIBSON
5. M C (Mark) Effingham
IBM Sales & Distribution, Public Sector
Client Technical Advisor
Ask: Ari Fishkind > Julie A. Reid
6. Paul (P.E.) Van Aggelen
Global Business Services
Pacific Development Center, Business Development Manager
Other Consultant
Ask: Michael W. Ticknor > Kinson (K.W.) Lee
7. Eric S. (ERIC) Minkoff
Global Business Services
US GBS Learning & Knowledge Learning Deployment Lead - Public Sector
Ask: James (JAMES) Stupak > Andrea R.
8. Thomas (Tom) Cocozza
Global Business Services
Healthcare Transformation Services
Ask: MARTHA E. (Martha) GIBSON > Alan J. (ALAN) Lauder
My shortest path to Susan
As a user, you can only see their public information. Private info is used internally to rank expertise but private data can never be exposed.
Click a name to see their profile (SmallBlue Reach)

Expertise Evidence



Example of Documents

- Internal and external websites
- E-mails
- Database records
- Agendas
- Logs
- Blogs
- Wikis
- Address books
- ...

Outline

- Introduction
- **Approaches**
- Evaluation

General Framework

$$P(C|Q) = \frac{P(Q|C)\,P(C)}{P(Q)}$$

$$P(C|Q) \propto \boldsymbol{P(Q|C)}\,P(C)$$

- *P(Q)*
  - Equal for all candidates given a query
- *P(C)*
  - Any priority that can be defined on candidates

Approaches

- Profile-based
  - Building a profile for each candidate
  - Matching it with input queries
- Document-based
  - Using documents to connect queries and candidates
  - Finding relevant documents to the input query
  - Finding the association between documents and candidates

Document-based Approach

- Commonly, co-occurrence information of the person mentions with the query words in the same context is assumed to be evidence of expertise

- In the simplest case, this context is the document itself, so that “all the evidence within the document is descriptive of the candidate’s expertise”

Document-based Expert Finding: Candidate Model

$$P(Q|C) = \prod_{q \in Q} \lambda\, P(q|C) + (1-\lambda) P(q|Corpus)$$

$$P(q|C) = \sum_{d \in D} P(q|d,C).P(d|C)$$

$$P(q|d,C) \propto P(q|d)$$

$$P(d|C) \propto P(C|d).P(d)$$

$$P(Q|C) = \prod_{q \in Q} \lambda\, [\sum_{d \in D} P(q|d).P(C|d).P(d)] + (1-\lambda) P(q|Corpus)$$

Document-based Expert Finding: Document Model

$$P(Q|C) = \sum_{d \in D} P(Q|d, C).P(d|C)$$

$$P(Q|d, C) \propto P(Q|d)$$

$$P(d|C) \propto P(C|d).P(d)$$

$$P(Q|C) = \sum_{d \in D} P(Q|d).P(C|d).P(d)$$

Document-based Expert Finding: Document Model

$$P(Q|C) = \sum_{d \in D} P(Q|d).P(C|d).P(d)$$

$$P(Q|d) = \prod_{q \in Q} \lambda\, P(q|d) + (1-\lambda) P(q|Corpus)$$

$$P(Q|C) = \sum_{d \in D} [\prod_{q \in Q} \lambda\, P(q|d) + (1-\lambda) P(q|Corpus)].P(C|d).P(d)$$

Candidate-Document Association

- *P(C|d)*
  - Frequency-based approach
  - Boolean model

$$P(C|d) = \begin{cases} 1 & \text{if } C \text{ exists in } d \\ 0 & \text{Otherwise} \end{cases}$$

  - Candidates count

$$P(C|d) = \begin{cases} \frac{1}{|C|} & \text{if } C \text{ exists in } d \\ 0 & \text{Otherwise} \end{cases}$$

- *P(d)*
  - Any priority that can be defined on documents

Proximity

- The closer a candidate is to a term the more likely that term is associated with their expertise $P(Q|d,C)$
- Considering the proximity of terms and candidate mentions in the document
  - Terms surrounding candidate mentions form the context of the candidate's expertise
- Defining a window of a fixed size.
  - Small window sizes often lead to high precision but low recall in finding experts
  - Large window sizes lead to high recall but low precision
- It is also possible to consider multiple levels of associations in documents
  - Combining multiple window sizes
  - Exploiting document structure or metadata

Outline

- Introduction
- Approaches
- **Evaluation**

Questions

- Which one is better: candidate model or document model?
- Do we need any proximity in this model? if yes which window size?
- Document-candidate association: frequency-based approach or boolean model?
- Do we need any prior probability for candidates or documents?

Evaluation

- TREC enterprise track
  - TREC 2005: 50 queries
    - Topics: name of working groups on the W3C
    - Experts: members of the working group
  - TREC 2006: 49 queries
    - Topics and experts: assessed manually

- Each person mentioned in documents with name, e-mail, ID number, and abbreviations.

Evaluation

- W3C Corpus
  - 331,037 documents
- Expert List
  - 1092 experts
- Evaluation Metrics
  - Mean Average Precision (MAP)
  - Mean Reciprocal Rank (MRR)

Results: candidate model vs document model

Model	MAP		MRR	
	2005	2006	2005	2006
Candidate Model	0.1888	0.3206	0.4692	0.7264
Document Model	**0.2503**	**0.4660**	**0.6088**	**0.9354**

Further Reading

Expertise Retrieval

By Krisztian Balog, Yi Fang, Maarten de Rijke,

Pavel Serdyukov and Luo Si

Publisher: now

2012

Questions?