

Automatic people tagging for expertise profiling in the enterprise

Pavel Serdyukov *

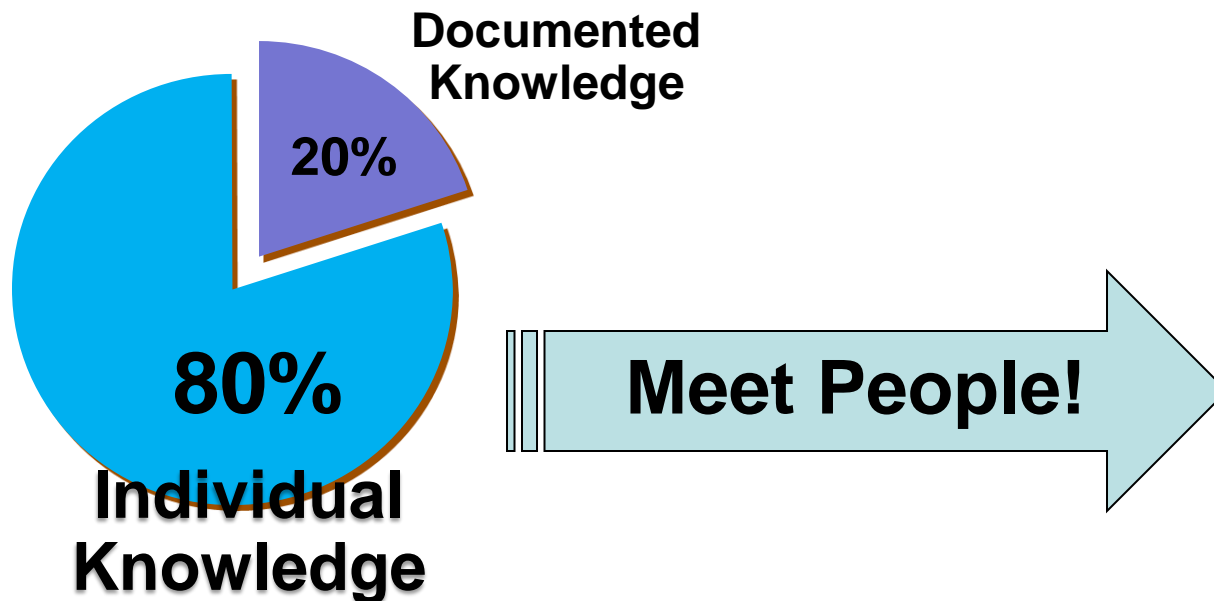
(Yandex, Moscow, Russia)

Mike Taylor, Vishwa Vinay, Matthew Richardson, Ryen White
(Microsoft Research, Cambridge / Redmond)

*** Work was done while visiting MSR Cambridge**

The need for experts

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




What people do **without** special expert finding tools?

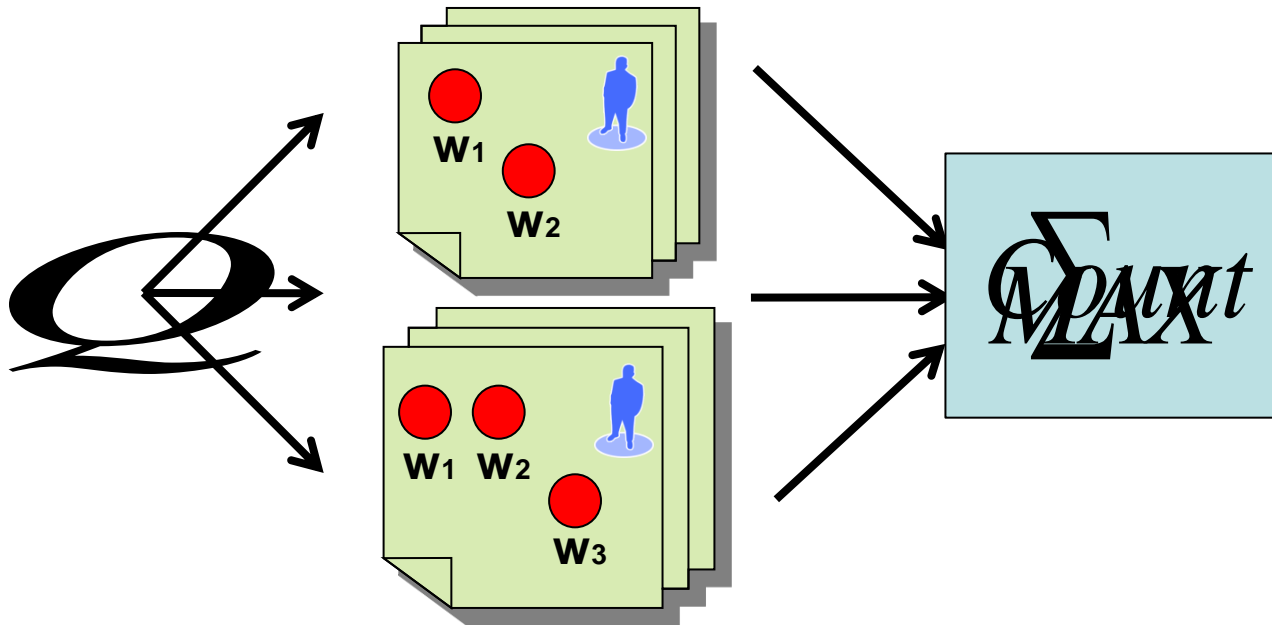
- Independent survey of 170 UK companies
- 50% want to be able to locate expertise
- Only 9% have tools for expert finding
- To find experts:
 - 71% “ask around”
 - 46% use the company directory
 - 34% use the company intranet
 - 30% send a company-wide email

Expert finding

- **Task definition:**
 - Given a short query
 - Rank employees **judged as experts higher than non-experts**
 - Very similar to document retrieval, but...
 - Finding **relevant people**, not **documents**
- Existed as a part of **TREC Enterprise track** for 4 years (2005 - 2008):
 - Community developed nice datasets
 - Lots of papers published
 - **And almost no industrial research!**

Traditional approach

- 1st step: Rank all documents with 
- 2nd step: Aggregate document scores



Typical expert finding output

Query: “csharp programming”

Hard to estimate relevance...
So, why should I click?

People Matches

-   Ali Pezeshk
SDET
US-WLX-SDET
-   Weiyang Chen
Content Publishing
WIN Web Services &
Content
-   Sunghwa Jin
SDET
US-Distributed Application
Server

[View more people »](#)

[Sharp C](#)

Get the best deal. Sharp C at Shopzilla.
www.shopzilla.co.uk

[Sharp Standard Televisions Sale](#)

Save on Sharp Standard Televisions. Compare 100s of Sites
www.shop.com

[Sharp C](#)

Sharp C
www.bizrate.co.uk

[C Standard at Amazon.co.uk](#)

Millions of titles, new and used. Free UK Delivery on Amazon Orders
Amazon.co.uk/books

Compare to snippets/ads!
3 sources of evidence:
Title, URL, Description

Problem

delicious **do not trust** the plain list of names,

Andrew Kennedy ([Microsoft Research Cambridge](#))

research.microsoft.com/en-us/um/people/akenn/

2

language

ocaml

science

types

uk

- So, we |
 - Expert
 - Concise
 - Sentence-free,



build personal summaries:

David Elsweller
@delsweil Nuremberg
IITC

content-bearing

.

- Let's generate



@arjenpdevries/ir-db-research



@giovannaroda/ir-al

Information Retrieval people and conferences (& NLProc, IE, ML, RecSys)



@leobard/pim-researchers

who is who of not forgetting that important tweet - Personal Information Management researchers

People like to tag each other

Farrell, S., Lau, T., Nusser, S., Wilcox, E., and Muller, M. 2007. **Socially augmenting employee profiles with people-tagging.** UIST '07.

Welcome, Eric [[Sign Out](#)]

- My History
- My Profile
- Zeigeist
- Advanced Search
- Report Bugs
- About Fringe

Your tags for Edward:

- java [x]
- mobile [x]
- PIM [x]
- web2.0 [x]

+

Enter tags one at a time. Spaces are replaced with '+'. [Learn more](#)

Tagged by 8 people

application-design design
handhelds java **mobile** NW
sanjose **web2.0** XML

Has tagged 32 people

applications AJAX business
calendars collaboration dev-team
engineer java **mobile** office
PIM strategy telecom tel

Edward Forelli
Research and Development
Senior Software Engineer

Phone: +1. 555-1212 (ext 2391)
E-mail: eforelli@fringe.com
IM: ■ Edward Forelli (I am active)

Local Time: 11:43 AM EST (Wednesday)
Work Location: [NW Research Center \(site info\)](#)
Business Address: 1454 Technology Drive, San Jose, CA 95051
Office: B2-234
Department: [Mobile Device and Software](#)
Notes Mail: Edward Forelli/San Jose/Fringe
Mobile Phone: 1.415.555-1212
Assistant: n/a
Second Life: fringe-kid

Groups and Communities

- [★ mobile-web](#), [Open Source Developers](#), [Tennis NW](#), [web2.0](#), [Web](#)

Weblog: [Edward's Musings](#)

Bookmarks: [Edward's Dogear](#)

What is Web2.0?
A nice overview of companies and definitions to common terms surrounding web2.0 | March 15, 2007

(4) Stephanie Wells, Julie Vallero, Gary Hugo, Michael Muligan

Microsoft IM-an-Expert

Q&A system that finds experts to answer specific questions and mediates the dialog between **an expert** and **the answer seeker**

Stephanie asks **IM-an-Expert** to find an expert

Contact List

 IM an Expert Available

Conversation: Stephanie and Tom

Stephanie: How do you add a calendar drop-down selection in an Excel field?

IM an Expert: I am searching for answerers. Please be patient.

IM an Expert: **Tom** is willing to help. The two of you are now in a conversation.

Stephanie: Thanks Tom!

IM an Expert: No problem

Stephanie: bye

IM an Expert: Please rate the answer you received on a scale from one (not helpful) to five (very helpful)

Stephanie: 5

IM an Expert: You have rated this answer as very helpful. I have passed along the rating to the answer. Please close this window.

IM-an-Expert finds **Tom** and asks to help **Stephanie**

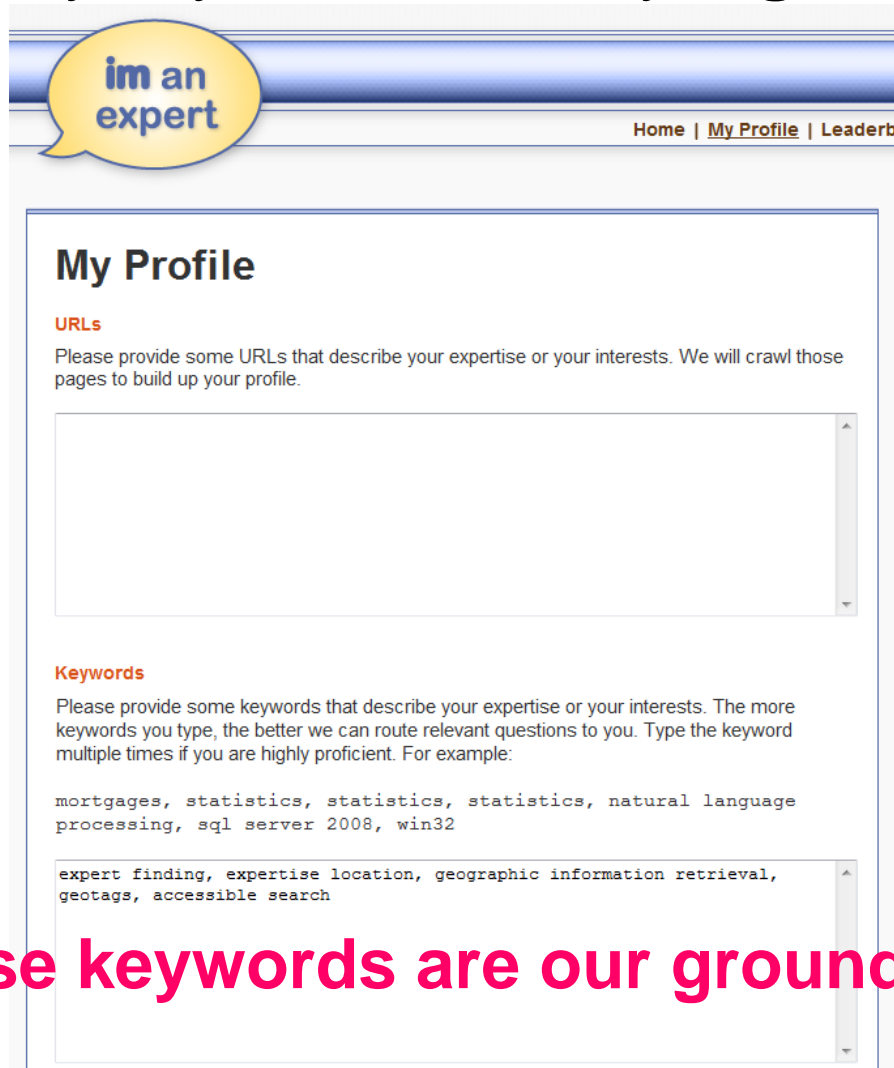
IM an Expert: Sorry for the interruption. Can you help **Stephanie** with the following question?

How do you add a calendar drop-down selection in an Excel field?

Type **yes** to accept question. Close window or type **no** to reject question.

How to make yourself found?

Candidate experts in IM-an-Expert describe their expertise by keywords, so they **tag** themselves



The screenshot shows the 'im an expert' website interface. At the top left is a yellow speech bubble logo with the text 'im an expert'. To the right of the logo is a navigation bar with links for 'Home', 'My Profile', and 'Leaderb'. Below the navigation bar is a section titled 'My Profile'. Under this title, there are two main sections: 'URLs' and 'Keywords'. The 'URLs' section has a text input field with a vertical scrollbar. The 'Keywords' section has a text input field with a vertical scrollbar. The text in the 'Keywords' field is: 'mortgages, statistics, statistics, statistics, natural language processing, sql server 2008, win32' and 'expert finding, expertise location, geographic information retrieval, geotags, accessible search'.

im an expert

Home | [My Profile](#) | Leaderb

My Profile

URLs

Please provide some URLs that describe your expertise or your interests. We will crawl those pages to build up your profile.

Keywords

Please provide some keywords that describe your expertise or your interests. The more keywords you type, the better we can route relevant questions to you. Type the keyword multiple times if you are highly proficient. For example:

mortgages, statistics, statistics, statistics, natural language processing, sql server 2008, win32

expert finding, expertise location, geographic information retrieval, geotags, accessible search

These keywords are our ground truth!

Our task

- **Predict** those **tags** that person specified in personal profile...
- ... **using various expertise evidence sources** related to the person
- Non-unique tags from our training data are our **controlled vocabulary**:
 - So, the task is as well **to recommend tags** for newcomers
 - And actually for any person in Microsoft
- So, let's **rank known tags w.r.t. each person in the enterprise**

- **1167** profiles
 - Alias + Location
 - Gathered
- **Tag stats:**
 - **4450** unique tags
 - **1275** tags
 - **5.5** non-unique tags
 - **1.47** words per profile




es

employee

Expertise evidence sources:

3:45 AM

[Redacted]



Peter Bailey
PRINCIPAL LEAD RESEARCHER
US-SEARCH SDE

Principal applied researcher & dev lead, Whole Page Relevance incubation team, Bing core search

+1 (425) 7064843 X64843
CITY CENTER/19537 City Center Plaza, 19537

pbailey@microsoft.com

[Add as colleague](#) [More information](#)

Overview Organization **Content** Tags and Notes Colleagues Memberships My SharePoint Sites

Sites
Blog

























Documents
Shared Documents
SIGIR2010

Pictures
Shared Pictures

SharePoint Documents

<< My Site Re: se... Partne... DTP Po... DTP An... Team S... DTP Po... DTP Po... >>

Go to Peter Bailey

Type	Name	Last Modified	Location	Properties
	[Redacted]	8/24/2010 10:58 AM	Shared Documents	
	[Redacted]	7/22/2010 10:41 AM	Shared Documents	
	[Redacted]	7/21/2010 10:52 AM	Shared Documents	
	[Redacted]	7/8/2010 6:50 PM	IIIx2010	
	[Redacted]	7/8/2010 6:48 PM	IIIx2010	
	[Redacted]	7/7/2010 7:00 PM	Shared Documents	
	[Redacted]	6/30/2010 2:30 PM	Shared Documents	
	copyright form	6/22/2010 1:30 PM	IIIx2010	
	[Redacted]	6/16/2010 7:12 PM	IIIx2010	
	[Redacted]	6/16/2010 7:09 PM	IIIx2010	
	[Redacted]	6/16/2010 6:46 PM	IIIx2010	
	[Redacted]	6/16/2010 2:17 PM	Shared Documents	

Expertise evidence streams: Click-through sources

- Personal queries to Sharepoint
 - 6 months of queries to Sharepoint (January 2010 – June 2010)
 - **67 unique queries** on average per person
- Clicked documents
 - **433 clicks** on average per person
 - **47 clicked documents** on average per person
- Queries with clicks on authored docs
 - **24 clicks** on average per person
 - **12 unique queries** on average

Streams and features

- Each source contributes streams
 - Authored/related/web/clicked docs:
 - Filenames, titles, snippets, body content
 - Body contents are crawled only for **authored** and **related**
 - Queries, lists:
 - Just query strings / names
- For each stream and each tag we calculate:
 - Binary (**1** if stream contains tag, **0** - otherwise)
 - Language model based score:
$$P(tag | \theta) = \prod_{w \in tag} \{(1 - \lambda) p(w | \theta) + \lambda p(w | Global)\}$$
 - Sum of scores of all records (e.g. **titles** or **queries**) in each stream **are our features**

Importance of deviation

- It's important not only to be “rich” in tag
- But “richer” on average!
- So, transformed features as:

$$X_{tag}^{employeeX} = X_{tag}^{employeeX} - \bar{X}_{tag}, \bar{X}_{tag} = \frac{1}{|training|} \sum_{employeeY \in training} X_{tag}^{employeeY}$$

Additional features

- **Popularity-based priors:**
 - Profile frequency
 - Frequency as query in Sharepoint
- **Quality of tag:**
 - Frequency in Enterprise data (*IDF*)
 - Probability of words in the tag based on Web corpus
 - Using Bing Web N-Gram service *
- **Phrase length:**
 - In words
 - In characters

* <http://research.microsoft.com/en-us/collaboration/focus/cs/bingiton.aspx>

Ranking

- 1167 profiles:
 - **700 (~60%)** as training set, **300 (~25%)** as test set
 - **167 (~15%)** as validation set (to tune parameters)
- In average: ~ **5.8 tags** per person
 - **4098 positive examples**
 - **$\sim 1270 \times 700 = \sim 900,000$** negative examples?
 - Too imbalanced...
 - Too slow to learn...
 - Sampled negatives randomly, tested on validation set:
 - **$\sim 60,000$** was enough to reach maximum AP
- Learned **Logistic Regression** model

Measures

- We rank **tags** by their classification scores
- Measures:
 - **Precision** at ranks *1, 5, 10* (**P@1/5/10**)
 - Average Precision at rank *100* (**AP**)
 - **Success** at rank *5* (**S@5**)

Individual feature performance

Stream	Feature performance			“No expertise evidence” performance		
	P@5	P@10	P@20	P@5	P@10	S@5
<i>ProfileFrequency</i> (baseline)	0.066	0.046	0.253	-	-	-
<i>AuthoredContentLM</i> (baseline)	0.044	0.030	0.180	0.081	0.057	0.310
<i>ListNamesBinary</i>	0.122	0.086	0.437	0.125	0.087	0.437
<i>AuthoredFileNamesBinary</i>	0.071	0.058	0.3	0.093	0.066	0.367
<i>AuthoredTitlesLM</i>	0.072	0.053	0.283	0.085	0.059	0.313
<i>PersonalQueriesBinary</i>	0.055	0.040	0.210	0.059	0.041	0.220
<i>QueriesToAuthoredBinary</i>	0.059	0.038	0.230	0.069	0.043	0.307
<i>RelatedSummaryLM</i>	0.035	0.024	0.163	0.059	0.041	0.257
<i>ClickedTitlesBinary</i>	0.021	0.018	0.087	0.033	0.025	0.133
<i>WebTitlesLM</i>	0.023	0.012	0.093	0.023	0.013	0.097
ALL features	0.171	0.124	0.543	-	-	-

Feature group importance

Stream	P@1	P@5	P@10	AP	S@5
<i>ALL</i>	0.266	0.171	0.122	0.124	0.543
- <i>ProfileFrequency</i>	0.240	0.138	0.102	0.110	0.460
- <i>List</i>	0.146	0.096	0.073	0.074	0.370
- <i>Authored</i>	0.130	0.078	0.065	0.063	0.300
- <i>PersonalQueries</i>	0.090	0.057	0.046	0.047	0.250
- <i>Related</i>	0.060	0.053	0.044	0.030	0.220
- <i>Clicked</i>	0.033	0.046	0.039	0.025	0.180
- <i>Web</i>	0.010	0.034	0.032	0.019	0.143
- <i>QueriesToAuthored</i>	0	0.025	0.023	0.007	0.123

Removing features by feature groups (evidence sources)

Click-through evidence importance

Clickthrough = {PersonalQueries, QueriesToAuth, ClickedDocs}

Stream	P@1	P@5	P@10	AP	S@5
<i>ALL</i>	0.266	0.171	0.122	0.124	0.543
<i>ALL - Click-through</i>	0.266	0.160	0.112	0.117	0.513
Typical enterprise: <i>ALL - Lists - ProfileFrequency</i>	0.146	0.096	0.073	0.074	0.37
Typical enterprise - <i>Click-through</i>	0.093	0.078	0.056	0.050	0.310
Click-through (only)	0.09	0.061	0.047	0.048	0.247

Error analysis (I)

- Some tags **are not predictable** with Enterprise data:
 - Work non-related relevant tags:
“ice cream”, “traveling”, “cooking”, “dancing”, “cricket”, “camping”, “judaism”
 - Tags which are not likely to be used in documents and/or too general:
“design patterns”, “customer satisfaction”, “public speaking”, “best practices”

Error analysis (II)

- Alternative tags used:
 - Predicted: **csharp, e-learning, t-sql**
 - Relevant: **c#, elearning, transactsql**
- More or less general concept used:
 - Predicted: **sql server 2008**
 - Relevant: **sql server**
- Concept expressed differently:
 - Predicted: **machine learning, web search**
 - Relevant: **data mining, search engines**



Susan Dumais

Predicted

search
 msr
 bing
 information retrieval
 web search

Relevant, but not mentioned

Relevant

search
 web search
 enterprise search
 desktop search
 hci

Relevant, but named differently



Vsevolod Dmitriev

Predicted

russia
 ocs
 exchange

Relevant, but less general concept is mentioned

Relevant

exchange 2003
 exchange 2007
 exchange 2010
 ocs 2007
 outlook
 exchange

Relevant, but not mentioned

Conclusions and Future work

- We've shown the way to solve a novel task of **automatic people tagging**:
 - Treated the problem as learning to combine evidences to rank areas of expertise
- Click-through evidence is important
 - But not decisive, at least, for Microsoft
- Future work should consider:
 - Diversity of recommended tagsets
 - Specificity of tags
 - Query dependent tagsets
 - Uncontrolled vocabulary