Same Places, Same Things, Same People? Mining User Similarity on Social Media

Ido Guy, et al.

IBM Haifa Research Lab

CSCW ’10

2 January, 2015

Jaehwan Lee

Outline

- Introduction
- Experiment
  - Experimental Setup
  - Characterizing Similarity Sources
  - People Recommendation Experiment
- Results
- Conclusion
  - Discussion
  - Future Work

Introduction

- Millions of people use social media applications

- Harvesting similarity information may be useful
  - for information discovery
  - for recommender systems
  - to promote response for advice
  - …

- But How?

Introduction

- Two Relationships Connecting People
  1. Familiarity
     - provides clues to when users may know one another
     - such as
       a. an explicit connection on an SNS
       b. tight collaboration on a wiki page
       c. a public message exchange
  2. Similarity
     - similar behaviors and activities of people who may actually be strangers
     - such as
       a. using the same tags
       b. bookmarking the same web pages
       c. connecting with the same people

Introduction

- Two Relationships Connecting People
  1. Familiarity
     - provides clues to when users may know one another
     - such as
       a. an explicit connection on an SNS
       b. tight collaboration on a wiki page
       c. a public message exchange
  2. **Similarity**
     - similar behaviors and activities of people who may actually be strangers
     - such as
       a. using the same tags
       b. bookmarking the same web pages
       c. connecting with the same people

Introduction - Related Work

- Power of Similarity
  - Accounting for taste: using profile similarity to improve recommender systems [Bonhard et al, '06]
  - study movie recommendations
  - examine how familiarity and similarity affects the decision
    a. profile similarity and rating overlap
    b. familiarity through exposure to the person's profile
  - familiarity did not affect participants' choices, while similarity had a significant influence

Introduction

- Difficulty on evaluation of similarity
  - more challenging than familiarity
  - users easily judge whether they are familiar with someone
  - hard to decide whether someone is similar

- Strategy
  - as similarity in general is hard to evaluate
  - narrowed evaluation to more concrete scenarios
  - such as
    a. "I am interested in reading this person's blogs"
    b. "this person reflects a subset of my expertise".

Introduction – An Overview

- Goal
  - use nine different sources for user similarity

- Hypothesis
  - all sources are useful

- Prerequisite
  1. Similarity relationships are uniquely different from familiarity
  2. Certain types of similarity sources are uniquely different from others

- Verification
  - People recommendation experiment

Experiment – Experimental Setup

- Similarity Sources
  1. a forum system
     - 2,590 forums ,433,000 threads, 45,500 users
  2. a blogging system
     - 16,300 blogs, 144,200 blog entries, 70,000 users, 121,750 comments
  3. a social bookmarking system
     - 359,300 public bookmarks, 552,000 tags, 16,3000 users
  4. a people tagging application
     - 9,300 users who tagged 50,000 other people with 160,000 public tags
  5. three enterprise SNSs
     - 250,000 public connections between 99,000 users
  6. an online communities system
     - 2,800 public communities, 120,500 members

Experiment – Experimental Setup

- Aggregation of different social media
  - use SONAR (SOcial Network ARchitecture)

- 9 Different Sources
  1. *friending* : having the same friend on one of the SNSs
  2. *tagged_by* : being tagged by the same person
  3. *tag_person* : tagging the same person
  4. *tagged_with* ; being tagged with the same tag
  5. *tag_usage* : using the same tag, while tags are collected from the social bookmarking system, the people tagging application and the blogging system
  6. *bookmarks* : bookmarking the same web page
  7. *communities* : being member of the same community
  8. *blogs* : commenting on the same blog entry
  9. *forums* : corresponding on the same forum thread.

Experiment – Experimental Setup

- Three Categories
  1. People
     - sources related to knowing or being known by the same *people*
  2. Things
     - sources related to being interested in the same *things*
  3. Places
     - sources related to being active in the same *places*.

Experiment – Experimental Setup

- Three Categories
  1. People : sources related to knowing or being known by the same *people*
  2. Things : sources related to being interested in the same *things*
  3. Places : sources related to being active in the same *places*

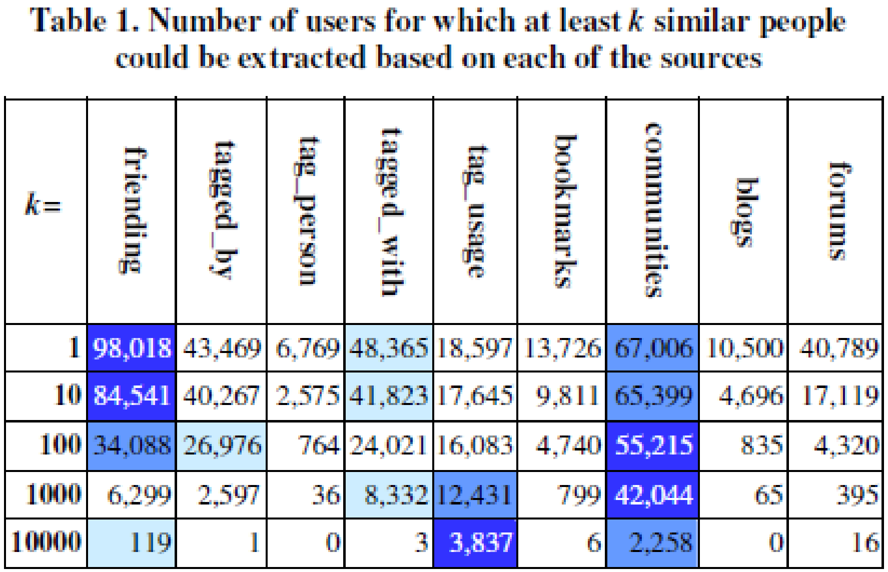


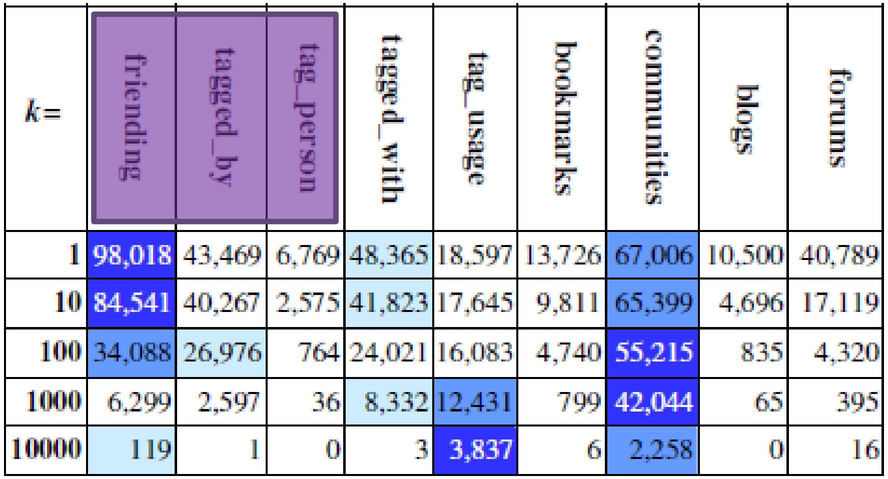
Table 1. Number of users for which at least *k* similar people could be extracted based on each of the sources

k=	friending	tagged_by	tag_person	tagged_with	tag_usage	bookmarks	communities	blogs	forums
1	98,018	43,469	6,769	48,365	18,597	13,726	67,006	10,500	40,789
10	84,541	40,267	2,575	41,823	17,645	9,811	65,399	4,696	17,119
100	34,088	26,976	764	24,021	16,083	4,740	55,215	835	4,320
1000	6,299	2,597	36	8,332	12,431	799	42,044	65	395
10000	119	1	0	3	3,837	6	2,258	0	16

Experiment – Experimental Setup

- Three Categories
  1. People : sources related to knowing or being known by the same *people*
  2. Things : sources related to being interested in the same *things*
  3. Places : sources related to being active in the same *places*

Table 1. Number of users for which at least *k* similar people could be extracted based on each of the sources

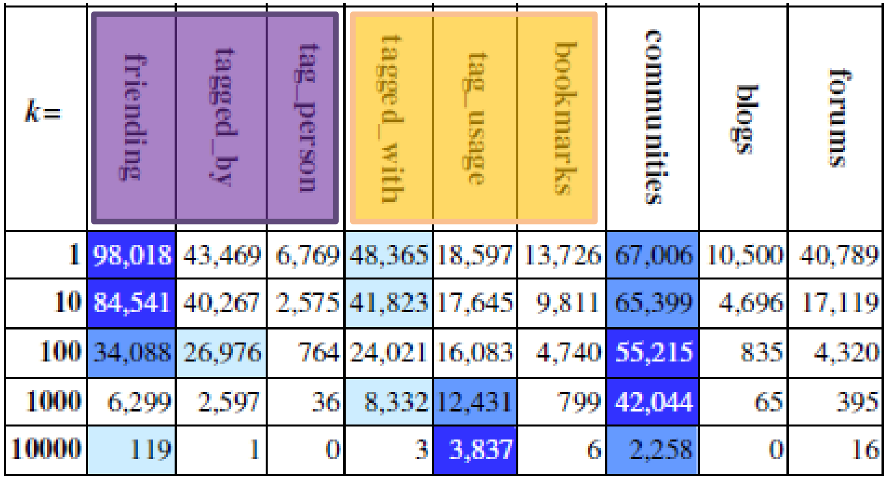


k=	friending	tagged_by	tag_person	tagged_with	tag_usage	bookmarks	communities	blogs	forums
1	98,018	43,469	6,769	48,365	18,597	13,726	67,006	10,500	40,789
10	84,541	40,267	2,575	41,823	17,645	9,811	65,399	4,696	17,119
100	34,088	26,976	764	24,021	16,083	4,740	55,215	835	4,320
1000	6,299	2,597	36	8,332	12,431	799	42,044	65	395
10000	119	1	0	3	3,837	6	2,258	0	16

Experiment – Experimental Setup

- Three Categories
  1. People : sources related to knowing or being known by the same *people*
  2. Things : sources related to being interested in the same *things*
  3. Places : sources related to being active in the same *places*

Table 1. Number of users for which at least *k* similar people could be extracted based on each of the sources

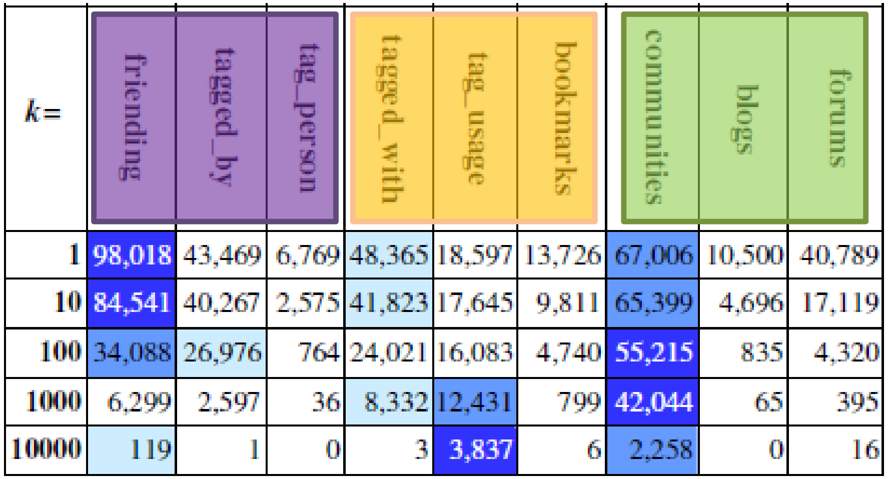


k=	friending	tagged_by	tag_person	tagged_with	tag_usage	bookmarks	communities	blogs	forums
1	98,018	43,469	6,769	48,365	18,597	13,726	67,006	10,500	40,789
10	84,541	40,267	2,575	41,823	17,645	9,811	65,399	4,696	17,119
100	34,088	26,976	764	24,021	16,083	4,740	55,215	835	4,320
1000	6,299	2,597	36	8,332	12,431	799	42,044	65	395
10000	119	1	0	3	3,837	6	2,258	0	16

Experiment – Experimental Setup

- Three Categories
  1. People : sources related to knowing or being known by the same *people*
  2. Things : sources related to being interested in the same *things*
  3. Places : sources related to being active in the same *places*

Table 1. Number of users for which at least *k* similar people could be extracted based on each of the sources



k=	friending	tagged_by	tag_person	tagged_with	tag_usage	bookmarks	communities	blogs	forums
1	98,018	43,469	6,769	48,365	18,597	13,726	67,006	10,500	40,789
10	84,541	40,267	2,575	41,823	17,645	9,811	65,399	4,696	17,119
100	34,088	26,976	764	24,021	16,083	4,740	55,215	835	4,320
1000	6,299	2,597	36	8,332	12,431	799	42,044	65	395
10000	119	1	0	3	3,837	6	2,258	0	16

Experiment – Prerequisite Justification

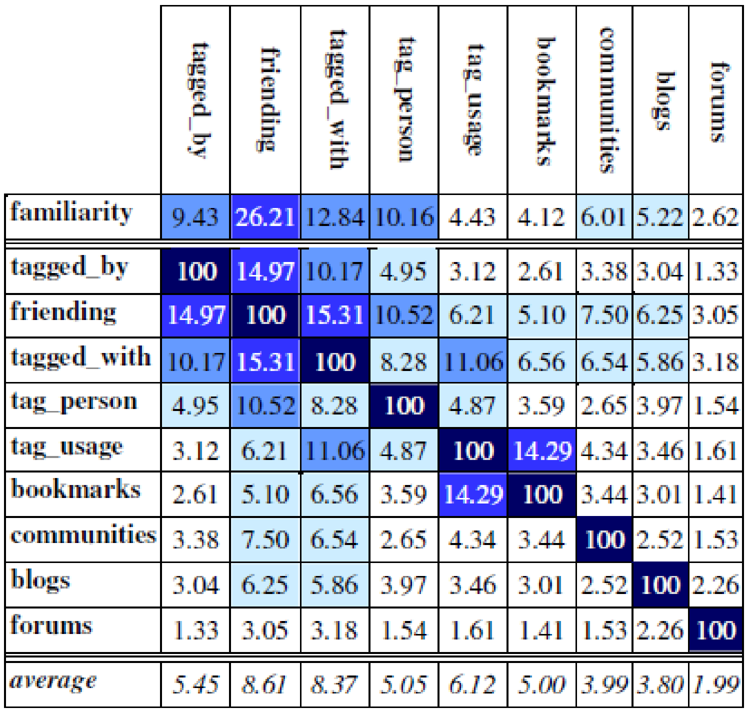
- Population
  - 577 avid users
  - those who make use of all social media applications
  - can be extracted based on each of the 9 sources

- Method
  1. Comparing Similarity to Familiarity
     - aggregate familiarity score
     - match@100 measurement
     - common people between the top 100 individuals
  2. Comparing Similarity Sources
     - match@100 measurement
     - total 36 source-to-source comparisons

Experiment – Prerequisite Justification

- Results
  1. Comparing Similarity to Familiarity
     - highest overlap percentage – 26.2%
     - 9% on average
     - the overlap is not high (p<.001)

Table 2. Mean *Match@100* values for the nine sources

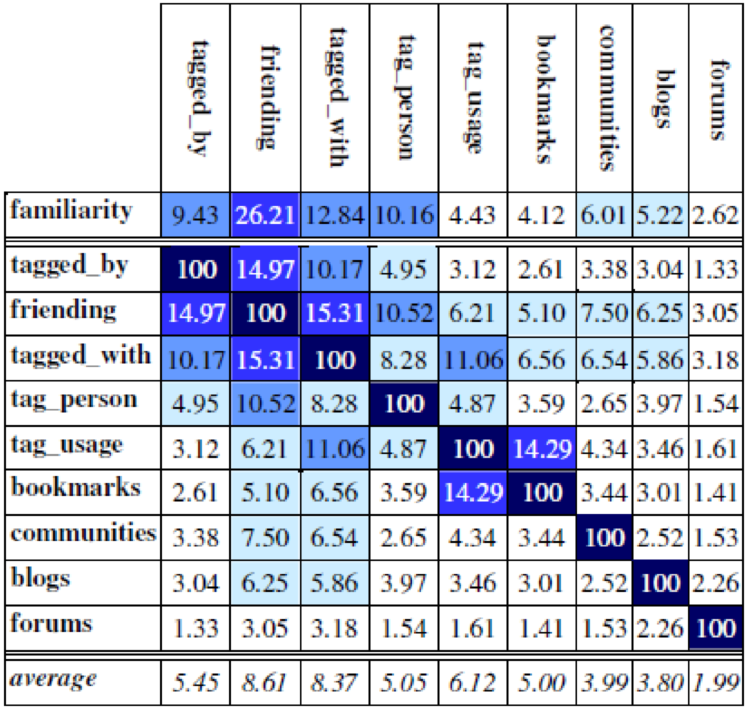


	tagged_by	friending	tagged_with	tag_person	tag_usage	bookmarks	communities	blogs	forums
familiarity	9.43	26.21	12.84	10.16	4.43	4.12	6.01	5.22	2.62
tagged_by	100	14.97	10.17	4.95	3.12	2.61	3.38	3.04	1.33
friending	14.97	100	15.31	10.52	6.21	5.10	7.50	6.25	3.05
tagged_with	10.17	15.31	100	8.28	11.06	6.56	6.54	5.86	3.18
tag_person	4.95	10.52	8.28	100	4.87	3.59	2.65	3.97	1.54
tag_usage	3.12	6.21	11.06	4.87	100	14.29	4.34	3.46	1.61
bookmarks	2.61	5.10	6.56	3.59	14.29	100	3.44	3.01	1.41
communities	3.38	7.50	6.54	2.65	4.34	3.44	100	2.52	1.53
blogs	3.04	6.25	5.86	3.97	3.46	3.01	2.52	100	2.26
forums	1.33	3.05	3.18	1.54	1.61	1.41	1.53	2.26	100
average	*5.45*	*8.61*	*8.37*	*5.05*	*6.12*	*5.00*	*3.99*	*3.80*	*1.99*

Experiment – Prerequisite Justification

- Results
  1. Comparing Similarity to Familiarity
  2. Comparing Similarity Sources
     - no overlap more than 16
     - *places* sources have the lowest
     - tagged_with vs *people*
       - based on tags
       - tags are given by people

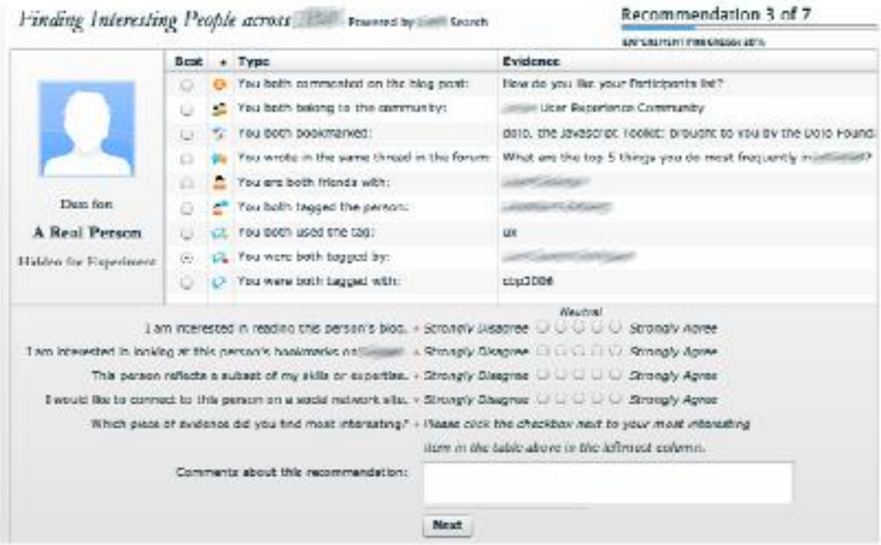
Table 2. Mean *Match@100* values for the nine sources



	tagged_by	friending	tagged_with	tag_person	tag_usage	bookmarks	communities	blogs	forums
familiarity	9.43	26.21	12.84	10.16	4.43	4.12	6.01	5.22	2.62
tagged_by	100	14.97	10.17	4.95	3.12	2.61	3.38	3.04	1.33
friending	14.97	100	15.31	10.52	6.21	5.10	7.50	6.25	3.05
tagged_with	10.17	15.31	100	8.28	11.06	6.56	6.54	5.86	3.18
tag_person	4.95	10.52	8.28	100	4.87	3.59	2.65	3.97	1.54
tag_usage	3.12	6.21	11.06	4.87	100	14.29	4.34	3.46	1.61
bookmarks	2.61	5.10	6.56	3.59	14.29	100	3.44	3.01	1.41
communities	3.38	7.50	6.54	2.65	4.34	3.44	100	2.52	1.53
blogs	3.04	6.25	5.86	3.97	3.46	3.01	2.52	100	2.26
forums	1.33	3.05	3.18	1.54	1.61	1.41	1.53	2.26	100
average	*5.45*	*8.61*	*8.37*	*5.05*	*6.12*	*5.00*	*3.99*	*3.80*	*1.99*

Experiment – People Recommendation

- Method
  1. Give 7 recommended individuals to participants
  2. For each recommendation, up to 9 evidence are given
  3. Participants asked to select the most interesting evidence
  4. Participants asked to rate the similar person



Finding Interesting People across [redacted] Powered by [redacted] Search

Recommendation 3 of 7

Dan fan

A Real Person

Hidden for Experiment

Best		Type	Evidence
○		You both commented on the blog post:	How do you like your Participants list?
○		You both belong to the community:	[redacted] User Experience Community
○		You both bookmarked:	dojo, the Javascript Toolkit: brought to you by the Dojo Found
○		You wrote in the same thread in the forum:	What are the top 5 things you do most frequently in [redacted]?
○		You are both friends with:	[redacted]
○		You both tagged the person:	[redacted]
○		You both used the tag:	ux
◉		You were both tagged by:	[redacted]
○		You were both tagged with:	cLp2008

			Neutral				
I am interested in reading this person's blog. + Strongly Disagree	○	○	○	○	○	Strongly Agree
I am interested in looking at this person's bookmarks on [redacted] + Strongly Disagree	○	○	○	○	○	Strongly Agree
This person reflects a subset of my skills or expertise. + Strongly Disagree	○	○	○	○	○	Strongly Agree
I would like to connect to this person on a social network site. + Strongly Disagree	○	○	○	○	○	Strongly Agree

Which piece of evidence did you find most interesting? + *Please click the checkbox next to your most interesting item in the table above in the leftmost column.*

Comments about this recommendation:

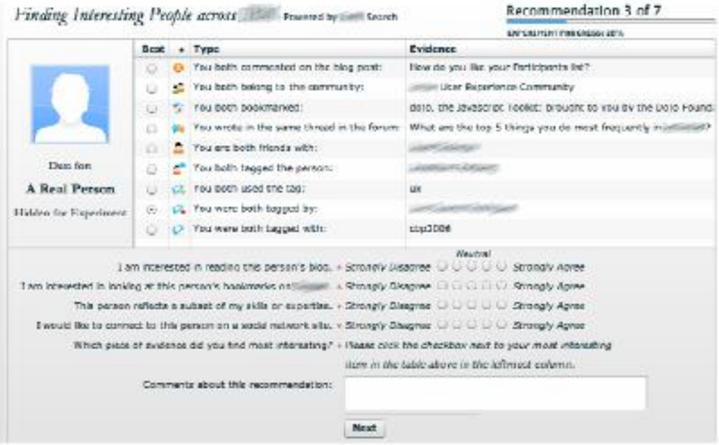
Figure 1. The experimental interface

Experiment – People Recommendation

- Population
  - 300 participants from 557 avid users

- Recommendation Configuration
  - four aggregation (*people, things, places* and *all*)
  - three random single sources
  - total 7, overall 13 different configurations

- Evidence
  - up to 9 evidence total
  - up to 1 for single source
  - up to 3 for three categories



Finding Interesting People across [redacted] Powered by [redacted] Search

Recommendation 3 of 7

[illegible]

Dan fon

A Real Person

Hidden for Experiment

Best	★	Type	Evidence
○		You both commented on the blog post:	How do you like your Participants list?
○		You both belong to the community:	[redacted] User Experience Community
○		You both bookmarked:	dojo, the Javascript Toolkit: brought to you by the Dojo Found
○		You wrote in the same thread in the forum:	What are the top 5 things you do most frequently in [redacted]?
○		You are both friends with:	[redacted]
○		You both tagged the person:	[redacted]
○		You both used the tag:	ux
●		You were both tagged by:	[redacted]
○		You were both tagged with:	ctp2006

Neutral

I am interested in reading this person's blog. * Strongly Disagree ○ ○ ○ ○ ○ Strongly Agree

I am interested in looking at this person's bookmarks on [redacted] * Strongly Disagree ○ ○ ○ ○ ○ Strongly Agree

This person reflects a subset of my skills or expertise. * Strongly Disagree ○ ○ ○ ○ ○ Strongly Agree

I would like to connect to this person on a social network site. * Strongly Disagree ○ ○ ○ ○ ○ Strongly Agree

Which piece of evidence did you find most interesting? * *Please click the checkbox next to your most interesting item in the table above in the leftmost column.*

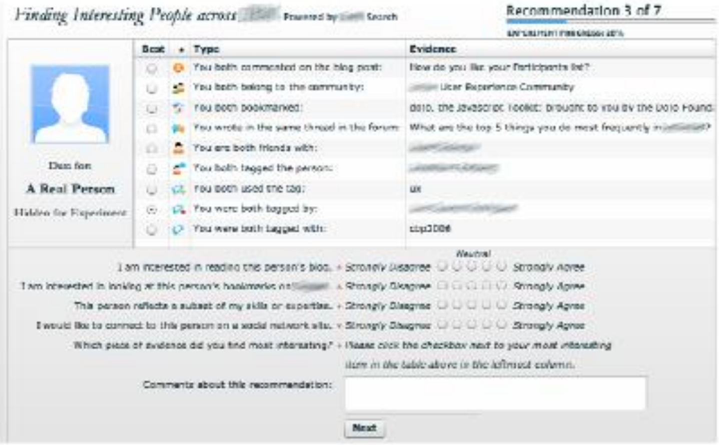
Comments about this recommendation:

Figure 1. The experimental interface

Experiment – People Recommendation

- How to Rate Similarity
  - based on given 4 scenarios
    1. I am interested in reading this person's blog (S1)
    2. I am interested in looking at this person's bookmarks (S2)
    3. This person reflects a subset of my expertise (S3)
    4. I would like to connect to this person on a social network site (S4)
  - 5-point Likert scale (strongly disagree to strongly agree)

- Personal Interests (PI)
  - PI affects S1, S2, S4
  - participants rates PI also



Finding Interesting People across [redacted] Powered by [redacted] Search

Recommendation 3 of 7

[illegible]

Dan for
A Real Person
Hidden for Experiment

Best		Type	Evidence
○		You both commented on the blog post:	How do you like your Participants list?
○		You both belong to the community:	[redacted] User Experience Community
○		You both bookmarked:	dojo, the Javascript Toolkit: brought to you by the Dojo Found
○		You wrote in the same thread in the forum:	What are the top 5 things you do most frequently in [redacted]?
○		You are both friends with:	[redacted]
○		You both tagged the person:	[redacted]
○		You both used the tag:	ux
●		You were both tagged by:	[redacted]
○		You were both tagged with:	cbp2006

		Neutral	
I am interested in reading this person's blog.	Strongly Disagree	○ ○ ○ ○ ○	Strongly Agree
I am interested in looking at this person's bookmarks on [redacted]	Strongly Disagree	○ ○ ○ ○ ○	Strongly Agree
This person reflects a subset of my skills or expertise.	Strongly Disagree	○ ○ ○ ○ ○	Strongly Agree
I would like to connect to this person on a social network site.	Strongly Disagree	○ ○ ○ ○ ○	Strongly Agree

Which piece of evidence did you find most interesting? *Please click the checkbox next to your most interesting item in the table above in the leftmost column.*

Comments about this recommendation:

Figure 1. The experimental interface

Experiment – People Recommendation

- Results
  - One-way ANOVA
    - 13 configurations significantly differ for each scenario (p<.05)
  - Games-Howell post-hoc test
    - marked by '*' yield significantly higher rating than those marked by '+'.

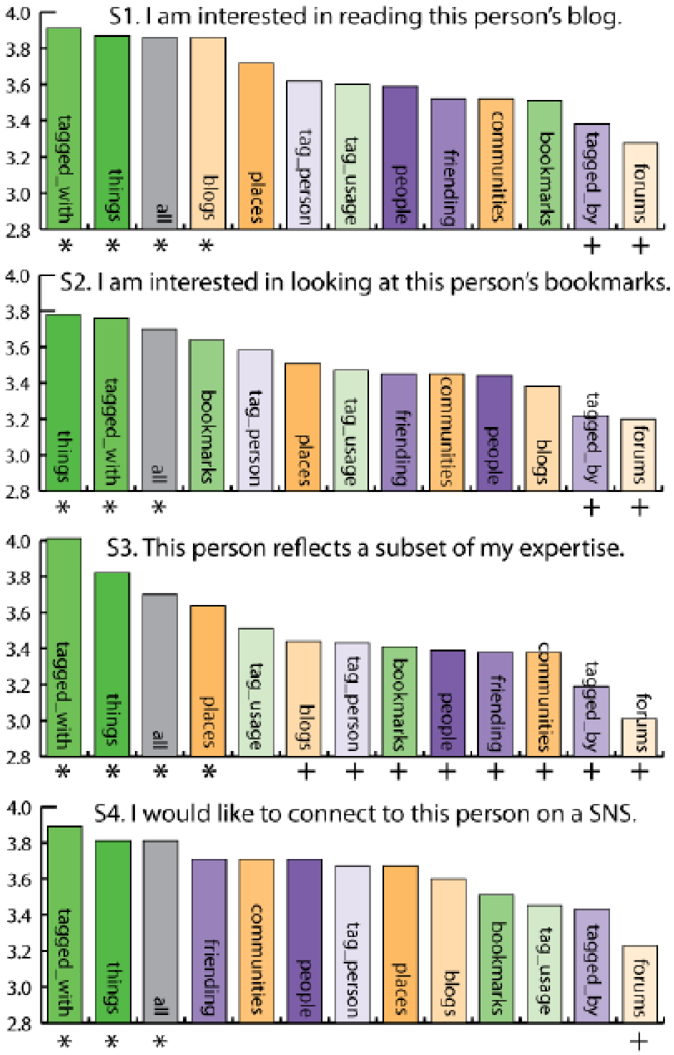
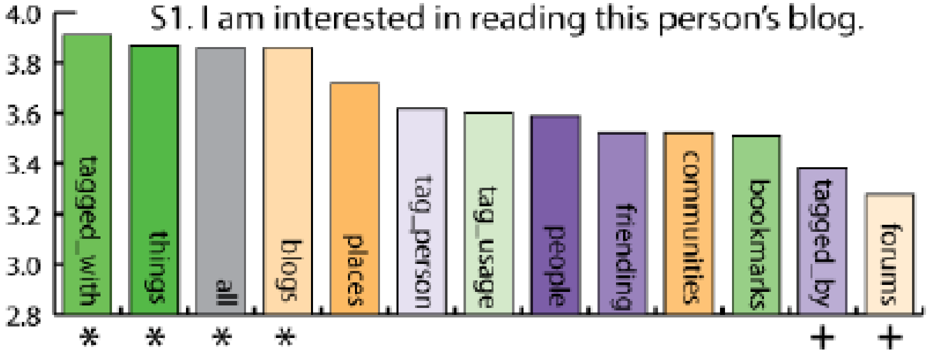


Figure 2. Average rating results for the 13 similarity configurations in each of the four scenarios (S1-S4).

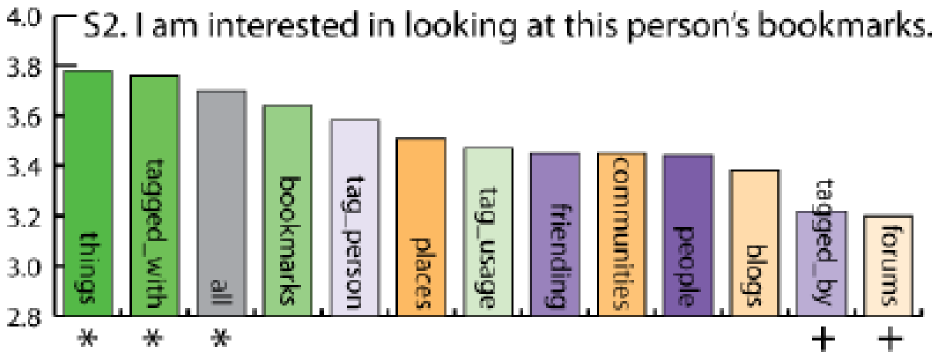
Experiment – People Recommendation

- Result – S1
  - *I am interested in reading this person's blogs*
  - average response 3.68 (SD: 1.08)
  - higher among BLOG-ENJOYERs (average 4.13)
  - three out of top 5 are aggregation



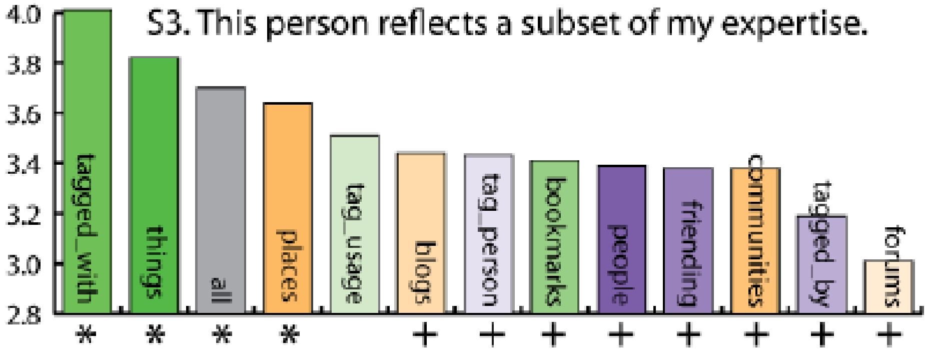
Experiment – People Recommendation

- Result – S2
  - *I am interested in looking at this person's bookmarks*
  - average response 3.54 (SD: 1.09)
  - higher among BOOKMARK-VIEWERs (average 4.05)
  - three out of top 5 are aggregation
  - interesting comment on value of tagged_with
    - "*This person seems to be tagged with my job role. That says a lot. I'd definitely check this person out further [...] even his/her bookmarks*"



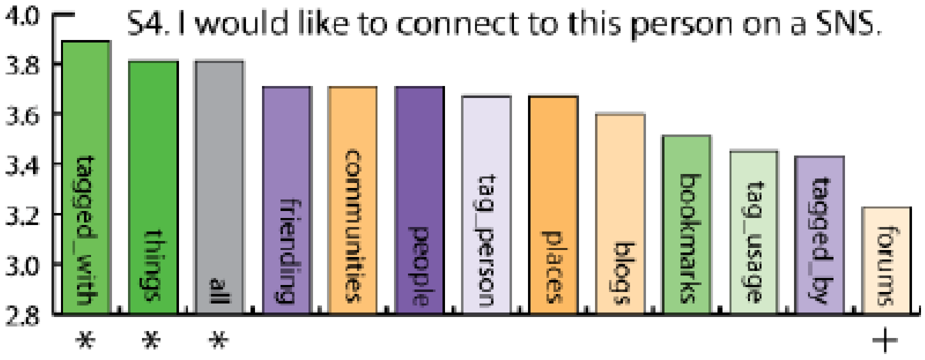
Experiment – People Recommendation

- Result – S3
  - *This person reflects a subset of my expertise*
  - average response 3.47 (SD: 1.1)
  - most diverse from 3.01 for forums to 4.01 for tagged_with
  - positive impact of aggregation
  - tags are good expertise indicators
  - “wisdom of the crowd”
    - tags given by other people are more reflective



Experiment – People Recommendation

- Result – S4
  - *I would like to connect to this person on an SNS*
  - average response 3.67 (SD: 1.14)
  - higher among SNS-LOVERs (average 4.03)
  - friending, people now useful (as expected)
  - tagged_with rules
    - tags given by the crowd are more effective



Experiment – People Recommendation

- Result – Distribution of sources
  - quite different results
    - tagged_with receives only 12% (4th)
  - interesting comments
    - *"I think seeing people that relate to me with outlier information, relevant to me, but things I know less about, is more compelling"*
    - *"…sometimes because it stood out as a 'why is this here?'"*

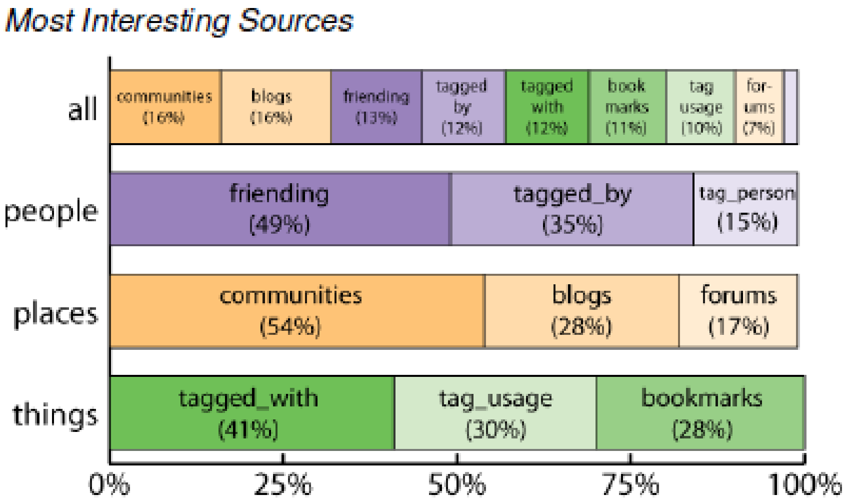


Figure 3. Distributions of sources of most interesting items for the four aggregate configurations

Discussion

- Consistency on Results
  - high rating : tagged_with, things, all
  - low rating : tagged_by, forums
- the “tagged_with”
  - top rated single source for all scenarios
  - combine *things* and *people*
  - compared with tag_person, tagged_by, tag_usage
  - potential power of people tags for mining user similarity
- Value of Aggregation
  - ratings of the aggregates are always higher than the average rating of their sub-sources
  - in some cases higher than any of those alone.

Discussion

- Preference on Diversity
  - users prefer diverse evidence items
  - comments
    - *"People I have different things in common with seem to be more interesting than those where the commonality lies only in one category"*
    - *"It is really the combination of these data points that is interesting"*

- Categories
  - *things* 〉 *places* 〉 *people*
  - exactly opposite to overlaps with familiarity
  - *people* is less effective for similarity detection (even for S4)

- Low rating of forum
  - it says "*We just experienced the same problem - somewhere!*"

Future Work

- Non-anonymized people recommender
  - recommend similar (yet unfamiliar) people to the user
  - based on aggregated sources
- To the Enterprise
  - similarity sources in social media outside the enterprise
- Transitivity
  - beyond similarity between users
  - extended to using similar tags (not only same tags)