Social Media-based Expertise Evidence

Arnon Yogev*, Ido Guy**, Inbal Ronen*, Naama Zwerdling*, and Maya Barnea*
*IBM Research-Haifa, Israel, **Yahoo Labs, Haifa, Israel[1]
*{arnony, inbal, naamaz, mayab}@il.ibm.com, **idoguy@acm.org

Abstract. Social media provides a fertile ground for expertise location. The public nature of the data supports expertise inference with little privacy infringement and, in addition, presentation of direct and detailed evidence for an expert's skillfulness in the queried topic. In this work, we study the use of social media for expertise evidence. We conducted two user surveys of enterprise social media users within a large global organization, in which participants were asked to rate anonymous experts based on artificial and real evidence originating from different types of social media data. Our results indicate that the social media data types perceived most convincing as evidence are not necessarily the ones from which expertise can be inferred most precisely or effectively. We describe these results in detail and discuss implications for designers and architects of expertise location systems.

Introduction

Social media is becoming more and more prevalent on the web and also inside enterprises. Enterprise social media platforms, such as Jive[2], Yammer[3] and IBM Connections[4], enable employees to share and collaborate through blogs, wikis, communities, and other social media applications. Employees engage on these platforms by describing their projects and solutions, problems they have

---

[1] Part of the research was conducted while working at IBM Research

[2] https://www.jivesoftware.com/

[3] https://www.yammer.com/

[4] http://www-03.ibm.com/software/products/en/conn

encountered, customers they have met, ideas they want to promote, and more (DiMicco et al., 2008).

While traditionally expertise used to be mined through explicit organizational skill sources such as CVs and project databases, as well as private sources, such as email and local file systems (Becerra-Fernandez, 2000; Reichling at al., 2009; Lin et al., 2009), social media enables expertise identification by implicit means with little privacy infringement. Through their public contributions to social media, employees leave traces of their areas of interest and knowledge, which enables effective expertise mining. The dynamic and up-to-date nature of contributions facilitates the constant update of an employee's expertise, whereas organizational databases are often not updated on a regular basis and thus do not provide accurate information.

Evidence has been shown to be highly important in search and recommendation systems for explaining the reasoning behind retrieved search results or recommended items. For example, search engines highlight matching terms in their returned results and recommender systems add explanations such as "We recommend you this movie because you watched another movie with the same actors". We believe that evidence may play a key role in expertise location systems too, yet the topic has hardly received any attention in the literature thus far. Evidence provides transparency for why a recommendation has been made and therefore increases the confidence of a user in the system (Tintarev and Masthoff, 2007; Sinha and Swearingen, 2002). In particular, for expertise location systems (McDonald et al., 2008; Balog et al., 2006), evidence enables users to understand why a person is suggested as an expert, which can increase their likelihood to approach the recommended individual.

The rich and public nature of social media data allows for providing high quality evidence for expertise, which is otherwise often hidden and inaccessible. As social media evolves, more content types emerge, such as blogs, microblogs, wikis, communities, shared files, and bookmarks. In addition, in social media users may be associated with content not just as its authors, but also as commenters, taggers, 'likers', or members. A previous study has shown that these other types of user-content relationships are highly useful for expertise mining and inference (Guy et al., 2013). It was speculated that with social media, experts often engage not necessarily as the authors, but rather as feedback providers or annotators through commenting, tagging, and liking. Due to the diversity of content types and possible user associations with content, a large variety of expertise evidence types may be exposed through social media contributions. For example, a person can be identified as an expert in a topic because s/he authored blog posts on the topic, commented on wiki pages, uploaded files, answered related questions, participated in related communities, and more.

In this paper, we examine which types of social media-based evidence users perceive as strongest indicators of expertise. As already mentioned, previous research has mined these different contribution types to identify experts (Guy et al., 2013). Different social media activities were weighted according to how much they reflect expertise in a topic. However, activities that provide strong evidence from an analytics perspective are not necessarily the best indicators from a user's perspective. Due to the variety and richness of data sources, there are many potential expertise indicators and a need arises to decide which of them to display in the application's user interface. This is especially true for mobile devices where real estate for display is small. Using this research, user interface designers can more sophistically decide which evidence types to highlight and on which to put less focus.

We examined the perception of evidence through two unique user surveys within a large global enterprise. In each of the surveys, participants rated anonymous experts presented to them along with evidence items. In the first survey, we sought to understand how enterprise social media users perceive potential evidence types with synthetically composed evidence, while in the second survey we examined the users' perception through the presentation of real evidence related to topics extracted from a deployed expertise location system within the studied organization. Participants were asked to rate the experts based on the presented evidence. We then compared results from both surveys.

The rest of this paper is organized as follows. The next section discusses related work. Following, we describe our experimental setup and the two surveys we conducted. We then present our results, which indicate that there is a substantial difference with regards to social media evidence presentation as compared to expertise mining. The paper concludes by summarizing these differences and their implications for expertise location system builders.

Related work

Expertise location, in particular within organizations, has been widely studied (Yiman-Seid et al., 2003; Reichling and Wulf, 2009). With the growth of social media usage within and outside the firewall, social media-based expertise location emerged. Examples outside the firewall exist for social applications, such as question and answering (Adamic et al., 2008), forums (Kardan et al., 2010), microblogs (Xu, 2014), and collaborative tagging (Noll et al., 2009). Chi (2012) provides a good overview of the challenges of such applications. Li et al. (2013) describe a system that returns a ranked list of experts for a given query built over the LinkedIn social graph. Within the enterprise, social media applications such as corporate blogs (Kolari et al., 2008), communities (Zhang et al., 2007), and people tagging (Farrell et al., 2007) were explored. A combined approach was conducted by Serdyukov and Hiemstra (2008) who concluded that expertise derived from sources outside the organization such as Yahoo and Google and their combinations with organizational data are often of higher quality than organizational data only. Varshney et al. (2014) used social media to infer skills within the enterprise by mining enterprise and social data. They developed a classification methodology to predict expertise based on features derived from the digital footprints of employees with the label set coming from the organization's expertise taxonomy. Guy et al. (2013) conducted a comprehensive study of expertise location based on social media data through a user survey in which participants provided self-evaluations for their areas of expertise and interest. The analysis focused on which data sources provided better basis for expertise inference, to most effectively predict the self-evaluation ground truth.

The importance of the display of evidence (often referred to as explanations) has been widely highlighted in the broad domain of recommender systems. Tintarev and Masthoff (2007) defined seven aims of evidence, transparency, scrutability, trust, effectiveness, persuasiveness, efficiency and satisfaction. They compared various academic and commercial systems according to these criteria. Sinha and Swearingen (2002) found that transparency in the form of explanations led users to have more confidence in the recommendations, as they could understand the justification for the system's choices. Herlocker et al. (2000) explained the challenge of generating explanations for collaborative filtering systems. They suggested that the black-box image of recommender systems might be one of the reasons for why they have gained much less acceptance in high-risk domains, such as holiday packages or investment portfolios than in low-risk domains, such as music or movies. Guy et al. (2009) showed the positive effect of evidence on employees' trust in a people recommendation system and their willingness to invite those people into their enterprise social network. Vig et al. (2009) used tag-based explanations to recommend movies. They concluded that tag relevance and tag preference play a key role for justification, effectiveness, and compatibility with the user's mood of recommendations.

Expertise location has been sometimes associated with the recommender system domain, referred to as "expert recommendation" (McDonald et al., 2008; Reichling and Wulf, 2009). Yet, since it is usually triggered by an explicit user query, expertise location more naturally belongs to the search domain. Pu and Chen (2006) investigated different explanation interfaces for recommender systems and compared their contribution to trust formation. They found that what they call organization-based explanations were most effective and increased the user's intention to return to their application as it assisted them in the comparison of products. Organization-based explanations adhere to five principles they defined such as categorization and diversity.

To the best of our knowledge, almost no publication has focused on the presentation of evidence for expertise location. Macdonald et al. (2008) built an expert search system that suggested people with relevant expertise to a topic of interest. Similarly to the approach taken in our paper, they used a document-based method for computing an expertise score. They proposed five techniques to predict the quality of documents (which they call evidence) within a candidate's profile in the expert search task. Those techniques were tested over three TREC topic sets. Balog et al. (2006) showed that the document-based approach outperformed a candidate-based approach for expert identification. Both papers did not address the issue of the presentation of expertise evidence to the user, but rather referred to the identified documents as evidence during score computation.

Experimental setup

Social media applications

Our experiments were based on an enterprise social software application suite, which has been commonly used in the studied organization (a large global enterprise) for over eight years. The suite includes eight types of social media applications:

- A blogging system that allows users to write blog posts, comment on posts, and 'like' posts.
- A social bookmarking system that allows users to share bookmarks of both intranet and internet pages.
- A file sharing system that allows users to upload files and read files uploaded by other users.
- A forum system that allows users to start new topics and comment on their own or other topics.
- A microblogging system that allows users to write short messages (status updates) of up to 500 characters on their profile wall, comment on their own or other messages, and 'like' other messages.
- A wiki system that allows multiple users to co-author wiki pages.
- A tagging system that allows users to tag each other, or to tag content items, such as blog posts or wiki pages.
- A communities system that allows users to create (and own) new communities of interest, or become members of existing communities. A community can include content items such as blogs, bookmarks, files, forums, and wikis.

Each application has its own entity type(s), and users can have different associations to the entities. For example, for the blogging application, the corresponding entity type is a blog post, and a user can be associated with the blog post as an author, a commenter or a liker. We refer to a combination of an entity type and an association (e.g., blog post liker, wiki page author) as an "expertise indicator" or, in short, an "indicator".

Application	Entity Type	Associations	#People
Blogs	Post (386,851)	Author (386,851) Commenter (301,975) Liker (375,544)	55,377
Bookmarks	Bookmark (1,459,510)	Bookmarker (1,459,510)	34,138
Communities	Community (149,152)	Owner (149,152) Member (2,855,139)	402,434
Files	File (384,527)	Author (384,527) Reader (1,382,610)	54,851
Forums	Topic (548,832)	Author (548,832) Commenter (1,391,417)	142,853
Microblogs	Message (1,021,927)	Author (1,021,927) Commenter (280,812) Liker (145,109)	97,891
Tags	Content Tag (357,553)	Tagger (3,226,711)	115,279
	Profile Tag (96,628)	Tagged-With (1,248,972)	186,381
Wikis	Page (1,240,834)	Author (1,240,834)	112,619

Table I. Social media applications, entity types, and associations.

Table I details the applications we experimented with, their related entity types, and the set of associations by which users can relate to the entities. The table also provides an overview of the usage level of the different applications in the studied organization: the numbers in parentheses indicate the amount of social entities and associations of each type. The rightmost column indicates, per application, the number of unique people who are "covered" by it, i.e., the number of employees who are associated with at least one entity. Overall, it can be seen that each application contains a large number of entities and associations, and covers at least 30,000 individuals.

Expertise location system

The social applications mentioned in the previous section form the basis of the Expertise Locator (EL) system used in the studied organization. To the best of our knowledge, the EL system is the first to use social media-based expertise evidence, thus our study's results can potentially be of significant value for it and for similar systems to come. The EL system enables web and mobile users to

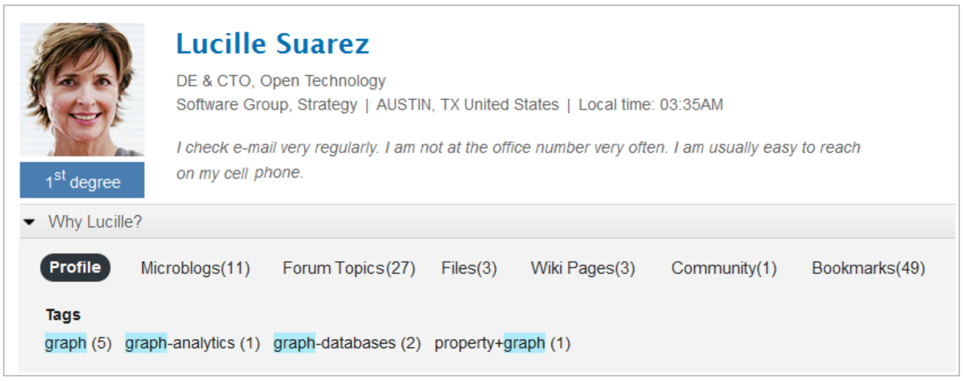

Figure 1. A "graph" expert returned by the Expertise Locator web application

search for experts on a certain topic, using a document-centric approach where each social entity, such as a blog entry, wiki page, or tag, is stored as a document in an index. The experts are retrieved based on their social activity in the context of the provided topic. First, documents related to the queried topic are retrieved. Then, the people associated with these documents (entities) are scored based on weights assigned to combinations of document types and associations to the document, as in Guy et al. (2013). The association of a person with a social media entity in the result is returned as evidence. Figure 1 shows a top expert for the query "graph", as displayed in the EL web application. The EL evidence presentation is divided into different tabs according to the originating social media application. The heading of each tab includes the number of evidence items of the corresponding evidence type; only tabs of evidence types with nonzero items are presented. The tab selected by default is the "Profile" tab. This design choice might not be optimal, since (1) it is not clear that profiles are indeed the strongest evidence and (2) by default the user sees a single type of evidence rather than a mixture. Better understanding the value of the various evidence types can help designers improve expertise presentation for the end user.

Figure 2a shows the same top expert for the query "graph", as well as the evidence, as displayed by the mobile application of the EL system. Since the screen is smaller, the approach taken was showing a summary of the evidence first and upon the user tapping on a specific summary line, showing the corresponding details. Figure 2b illustrates the list of bookmarks, triggered when the user taps "49 related bookmarks" on the screen shown in Figure 2a. The only exception to this is the display of tags, which are presented inline as in Figure 2a and are not attached to any additional content. Users are not exposed to any direct evidence aside from tags, unless they explicitly choose to tap one of the summary lines. In addition, there is no indication of the strongest evidence item (e.g., a specific blog post or file) when a multitude of evidence exits, as in Figure 2a. Overall, the smaller screen of mobile devices makes it even more important to understand social media-based expertise evidence. As we will detail in the following section, the design of our two surveys mimics evidence presentation in

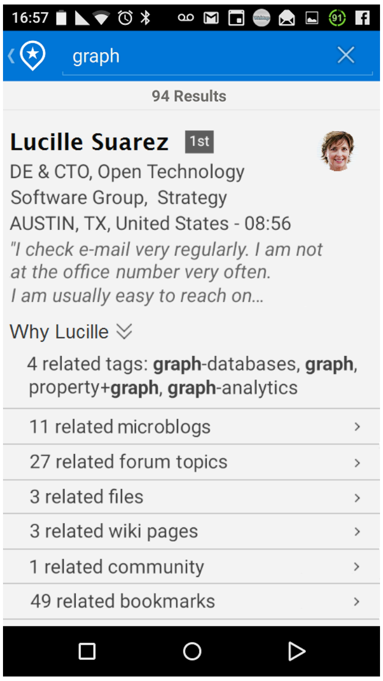

Figure 2a. A "graph" expert as presented in the EL mobile application

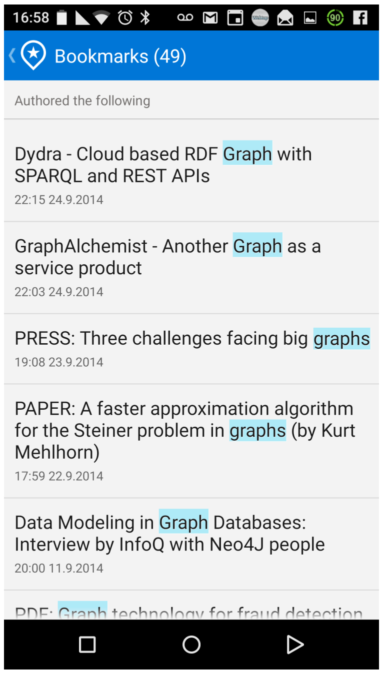

Figure 2b. Bookmark evidence of the "graph" expert in the mobile application

the EL applications: the first survey focuses on evidence summary of multiple types and the second survey focuses on specific evidence types, with real content.

User surveys

In this work, we examine the different expertise indicators and how strongly each indicator supports one's expertise when presented as evidence. To this end, we conducted two separate user surveys, with different sets of participants, where each survey focused on evidence in a different manner. In the first survey, the most active users of the organization's social software application suite were asked to rate experts, basing their decision solely on the type and the count of the expert-related activity. In the second survey, users of the EL system were asked to rate experts, basing their decision on the actual content of the expertise evidence, i.e., specific evidence examples collected from real data of the studied organization's EL system.

User survey I

In the first user survey, we focused on the types and counts of expertise indicators, without referring to their actual content. The goal was to identify the most significant indicators, as perceived by the users, without referring to specific evidence items. Each participant was presented with a list of 10 anonymous experts, whose data was synthetically composed. For each expert it included a list of three related indicators, which explained why they were recommended as experts. Each indicator included the type and count of the entities related to the expert (e.g., “Authored 5 blog posts”, “Read 2 files”). For indicator types, we experimented with all available entity types and associations as detailed in Table I (16 in total). For counts, we opted to experiment with two values - 2 and 5. These values were selected by examining real data from the EL system: we calculated the number of evidence items per indicator in a large collection of results for different topics and found 2 to be the median and 5 to be the 75 percentile. We did not include other count values in the survey, as this would have resulted in investigating too many combinations of indicators and counts.

Figure 3 illustrates an anonymous expert as presented in the survey. The three evidence items are presented on the left. Based on the displayed indicators, the participant was asked to rate each expert’s expertise level on a 5-point Likert scale, where 1 indicated “Not an expert” and 5 indicated “Definitely an expert”. Participants were also asked to select which of the three indicators was the ‘best’ one, i.e., the one they found the most significant. Finally, participants had an option to add a comment about a specific expert and a general comment to express their thoughts and give feedback at the end of the survey.

One of the goals of the survey was to retrieve an average rating score for each indicator. Since each expert’s evidence included 3 different indicators, there was a need to balance the effect of indicators on one another, e.g., avoid a case where ‘blog liker’ received higher average ratings because it often appeared together with ‘wiki author’. To this end, we referred to an ‘evidence configuration’ as a combination of 3 different indicators (out of the 16), where each indicator’s count value was either 2 or 5. An example for an evidence configuration could be: “authored 5 wiki pages, commented on 2 blog posts, read 2 files”. The survey was designed such that each configuration was scored by exactly one participant, thus making sure each indicator appeared alongside all the possible combinations of indicators and enabling a “fair competition” among all indicators and their counts. The total number of configurations was (16·2)·(15·2)·(14·2)/(3!)=4,480.

As each participant was asked to score 10 experts and exactly 4480 configurations had to be scored, we needed 448 participants for the survey. We invited the 3000 most active users of the social software suite to participate in the survey. Invitations were sent via email and 617 participants responded (20.5%). We then selected the first 448 responses so that each evidence configuration was

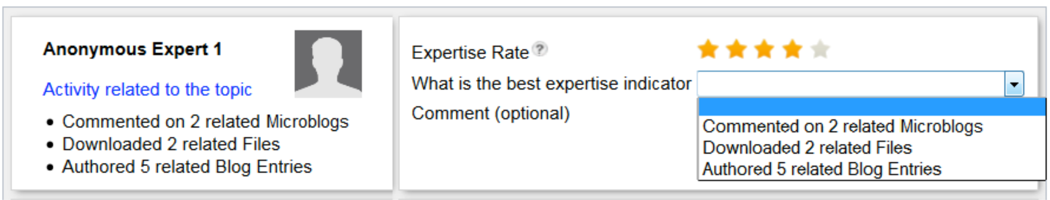

Figure 3. An anonymous expert as presented in Survey I.

covered exactly once. The rest of the responses were discarded, as the number of total responses did not reach the next multiplication of 448.

Many of the participants who submitted the survey chose to add comments (either on a specific expert or general ones) in order to express their thoughts on an expert, an indicator, or the expertise location system in general.

User survey II

The second user survey focused on the content of the indicators. The goal was to present the user with real and detailed evidence extracted from the EL system and compare the results with those of the first survey.

In this survey, participants were presented with a real topic of expertise, which they had previously queried for in the EL system (e.g. “Big Data”, “Java”, “Gamification”). We selected the 30 most popular topics, i.e., those issued as a query to the EL system by the largest number of users. An invitee list was then generated by selecting users who had issued a query on one or more of the 30 topics at least once, in order to increase the likelihood that a participant had knowledge about the topic and could assess the quality of the evidence items. Each participant was asked to rate the expertise level of 5 anonymous experts on a 5-point Likert scale. The evidence list for each expert was comprised of either 2 or 5 evidence items associated with the expert, all of the same type. Each item included a title (e.g., ‘Project Breadcrumbs Overview’) and a link to its page within the social software suite, allowing participants to access the actual blog post, wiki page, community, etc. Tag indicators (*tagger* and *tagged-with*) were an exception, as they have no associated content other than the tag itself. In order to avoid biased decisions, the experts were fully anonymized and their identity was removed both from the survey page and from the corresponding evidence pages to which the links pointed. Figure 4 illustrates two anonymous “banking” experts as presented in the second survey.

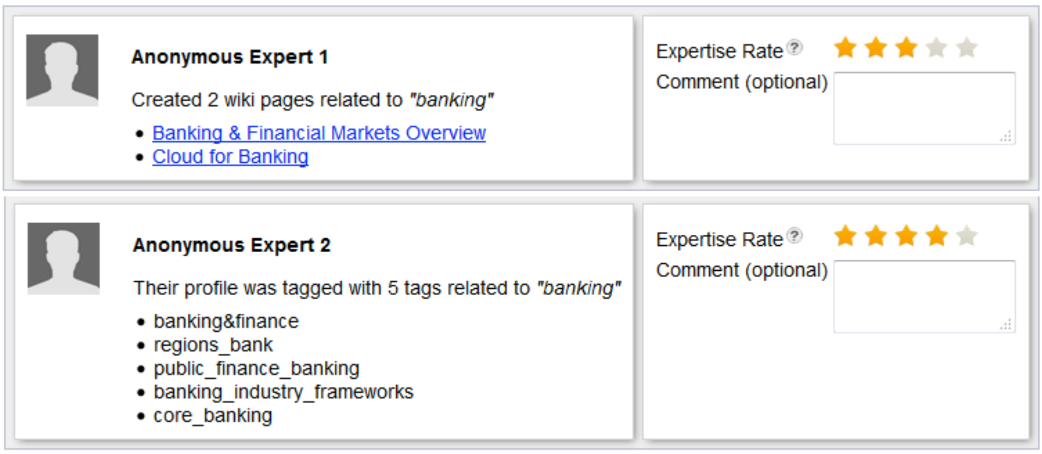

Figure 4. Two anonymous experts as presented in Survey II.

In order to have sufficient information to compute an average score for each of the 32 indicator-count combinations (16 indicator types, 2 possible count values), we used a round-robin method such that each indicator was scored by the same number of participants.

Overall, the survey was sent to 3000 users who had used the EL system to perform one of the 30 queries, out of which 579 users responded (19.3%). Each of the 16 indicator types was rated by 180 participants, half with count 2 and half with count 5. Similarly to the first survey, many of the users added comments, either on a specific expert or general ones.

Experimental results

User survey I

In this section, we analyze the results of the first survey. We examine which indicators were found to be stronger than others, considering both the average scores and the proportion each indicator was voted as "best". Recall that each evidence configuration was rated exactly once, such that each indicator symmetrically appeared with all other indicators. We first analyze the ratings given to each of the 16 indicator types, and at the end of the section we inspect the ratings while also considering the count values.

The first analysis focuses on the average rating per indicator, as shown in Figure 5a. The error bars indicate 95% confidence intervals for the rating averages. We observe a division of the indicators into three groups: the first group, which received the highest scores (indicators ranked 1-6), consists of indicators that show significant contribution of content, such as authorship and community ownership. The second group (ranks 7-9,11) consists of indicators that show light contribution of content, such as commenting and microblogging. In the third group, which received the lowest scores (ranks 10,12-16), we find indicators that show "minimal" activity, without any contribution of content, such as tagging, liking and reading. We now focus on each of these three groups.

The top scored group of indicators consists of 4 authorship indicators (blog, wiki, file, and forum), community owning, and profile tagging. In Guy et al. (2013), it was found that authorship is not a stronger indicator for expertise than commenting and liking, when it comes to social media. Here we see that in terms of evidence, authorship still prevails, with all authorship types (aside from microblogs) leading the list. Even in forums, where authorship is often about asking a question, it is still perceived higher than commenting, which in many cases contains an answer. The two top authorship indicators were blogs, ranked 1st with an average score of 3.75 and wikis, ranked 3rd with 3.46. These two social entities typically contain rich content, as one participant noted: "*Authoring*

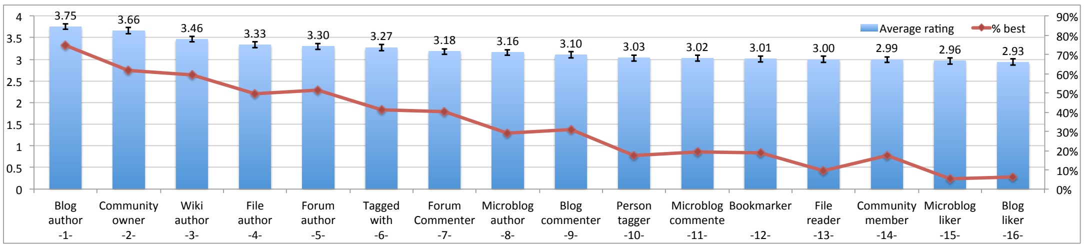

Figure 5a. User survey I: Average ratings (primary y axis) and 'best' proportions (secondary y axis).

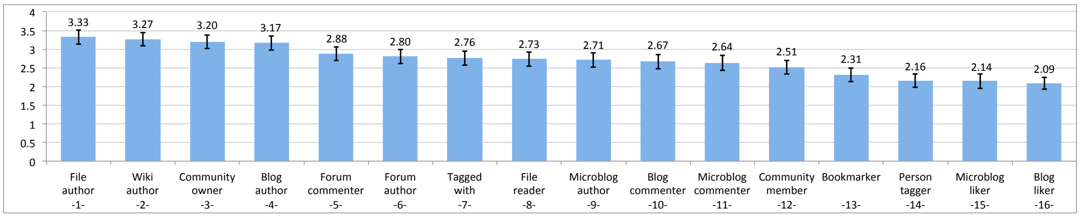

Figure 5b. User survey II: Average ratings.

multiple blogs is a sign of someone who has something worthwhile to say". Community ownership (2nd with 3.55) is also perceived as a strong evidence for expertise, as owning a community related to some topic suggests a strong involvement in it (Ronen et al., 2014). Community ownership was also considered a type of "authorship" by some users, since the owners manage the community. One participant explained: "The *owner is the one who really manages the communities and edits all the relevant content on the communities*." The last indicator in this group (6th with 3.27) is *tagged-with*, which is a kind of an outlier, as a person whose profile was tagged with a topic did not contribute content, but was indicated as related to the topic by other users. One participant commented: "*The tags indicate others recognize and accentuate the expertise shared*." In Guy et al. (2013), people tagging was found to be the most precise data source. It is also the most direct indicator of expertise and reflects the "wisdom of the crowd" rather than the person's own activity. In spite of all that, *tagged-with* is ranked below all authorship forms, except for microblogs.

In the second group, we find indicators that represent light contribution of content. These indicators received medium scores and were ranked between the authorship indicators and the reading\tagging\liking indicators. Leading the list is *forum commenter* with an average score of 3.18. This is the strongest indicator of all commenting evidence, since a forum comment is often an answer to a question, as one participant noted: "*Forum topics are generally discussing issues where an expert is being sought. Adding comments could mean you are offering your expertise on the forum topic or question*." The next commenting indicator is for blogs, which also topped the authorship indicators.

The only authorship indicator included in the second group is *microblog author*. Microblogs had the best combination of precision and recall in Guy et al.

(2013). This concise form of text expression was shown to be very effective for expertise mining in the enterprise. Here we see that despite its effectiveness for expertise inference, it is perceived as particularly weak evidence by users. It is the weakest authorship indicator and is lower than *tagged-with* and *forum commenter*. The reason may be that microblogs are limited to 500 characters and are typically used for sharing short messages, rather than contributing "heavy" authoritative content. The difference between authoring a microblog and authoring a blog was expressed in several comments made by participants, for example: "*Blogs are typically lengthy and inform on a particular area of expertise vs. microblogs which tend to be a quick/short update*".

In the third group, we find indicators that reflect more passive feedback, with no contribution of new content by the potential expert. These indicators received the lowest scores and were mentioned by many participants as weak or even meaningless. This list of indicators includes tagging, reading, or being a member of a community. At the bottom of the list are the two liking indicators - *blog liker* and *microblog liker*. Blogs and microblogs are similar in many characteristics, with the former having richer content. As we have seen this difference leads to a big gap in ranks between the two in terms of authorship and commenting evidence. When it comes to liking, however, there is no such difference and both are at the bottom of the list. One of the more surprising yet very consistent findings of the study by Guy et al. (2013) was that expertise inference from a user's liking activity is as effective as inference from their authorship or commenting activity. It was speculated that with social media, experts may engage in feedback providing as much, and sometimes more frequently, as they do by authoring content. Our survey reveals that expertise location users still do not "buy" liking and other weaker forms of feedback as strong expertise evidence. Our participants' comments imply that liking evidence shows a person's interest in the topic, rather than knowledge or expertise, e.g., "*[...] liking only indicates some level of interest*." Guy et al. (2013) compared inferring expertise and interest from social media data sources and found that in practice liking equally reflects expertise as it reflects interest in a topic.

The selection of 'best' evidence, also drawn in Figure 5a, is generally consistent with the average rating and also supports the partitioning into the three abovementioned groups: substantial content contribution, light content contribution, and no content contribution. The differences among the groups are even more clearly reflected through the proportion of 'best indicator' selection, i.e., the ratio between the number of times an indicator was selected as 'best' and the number of configurations it appeared in. The 'best' proportion of heavy-content indicators is the highest. *Blog author*, the top ranked indicator, was also the evidence that was chosen as best most often, exactly 75% of the configurations it appeared in. Similarly to the average ranking, the light content contribution indicators are placed in the middle. Closing the list, once again, are

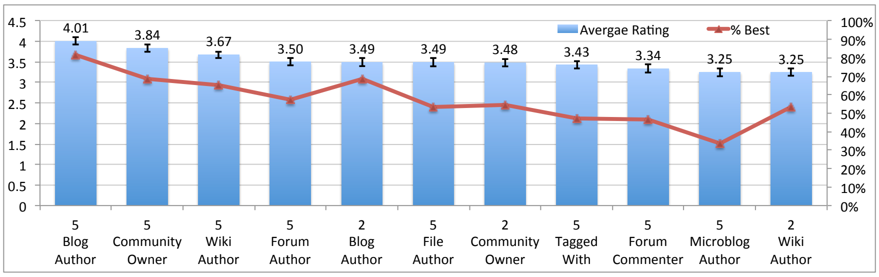

Figure 6. User survey I: Top average ratings and 'best' proportion for type-count combinations.

the two liking indicators, voted as best in only about 5% of their configuration occurrences. One difference between the average rating results and the 'best indicator' results is for *forum author*, which received a lower average score than *file author,* but was selected as 'best' more often. Another difference is for *community member*, which was lower than *file reader* on average rating, but higher in terms of 'best' selection. Despite these differences, the fact that the ranking of indicators based on 'best' selection is generally in agreement with the ranking of indicators based on their average rating, increases our confidence in these results.

We further inspected the comparison among all indicators, considering the number of evidence items (2 or 5) that were presented to the participants. Figure 6 shows the top of the list of indicators when distinguishing between the two counts (2 and 5). Generally, we see that the combination of quantity and quality of the evidence determines its ultimate ranking. It appears that the type (quality) plays a more central role, and stronger types are perceived as better evidence even if their quantity is lower. We sought to identify indicators at the top of the list with a count of 2, which are strong enough to outrank other indicators with a count of 5. It can be seen that blog authorship, the top evidence type, is also high when only including two items, lower only than 5 community ownerships, forum and wiki authorships. Another example is 2 owned communities, which are also lower than 5 authored files, but higher than all other 5's, including tagged-with and microblog authorship. As can also be seen in Figure 6, the selection of 'best' indicator is even more favorable of strong types with a lower count.

As mentioned above, the participants of the first user survey were asked to rank the experts based on evidence type and counts only, without referring to the content itself. Many of them expressed a concern, finding it difficult to evaluate expertise based on this information only, as one participant commented: "*In all cases I would also review the content that [the expert] authored or commented on to try and get a better understanding of their knowledge on the subject…*". This lays the ground for our second survey.

User survey II

In this section, we analyze the results of the second survey. We examine which indicators received better ratings than others and compare the results to those of the first survey. Note that the most significant difference between the two surveys is the content of each indicator. In survey II, participants were presented with 5 experts for the same topic (which they had searched for in the past), and a list of 2 or 5 evidence items per expert. The participants were able to click the title of each item and access its (anonymized) page, allowing them to view the community, read the blog, download the file, etc.

At high level, the results of survey II are similar to survey I: Authorship indicators are at the top along with community ownership. Liking is at the bottom, and commenting (as well as weak authorship indicators – forums and microblogs) are in the middle. There are, however, a few prominent differences between the results of the two surveys, which we discuss in the following paragraphs. One important thing to note is that the average ratings of the two surveys are not comparable – in the first survey participants rated experts according to combinations of evidence types, whereas in the second each expert had evidence of one type. For that reason, we see that survey I's average ratings are generally higher than survey II's, as the participants perceived evidence with combinations of 3 indicators as more convincing than evidence that included single indicators. Therefore, when analyzing the results of both surveys, we compare the rank of each indicator, rather than the rating it received.

As can be seen in Figure 5b, at the top of survey II's ratings list we find the same four indicators as in survey I. Three strong authorship indicators (blog, wiki, file), along with community ownership, are at the top, by a relatively large margin. The order of the four, however, is different between the surveys: *blog author,* the top indicator of survey I, moved down to number 4, whereas wiki and file authorship moved from ranks 3 and 4 to the top 2 spots. Authoring a blog on some topic sounds like a solid indication of a person's expertise. On the other hand, many blogs describing a product or a technology are written for publicity, rather than for providing technical details. They also tend to reflect a personal viewpoint of the author towards a topic, with many socializing aspects, rather than objective authoritative content. This may be the reason why this indicator was ranked lower when the users actually viewed the content of the blog. On the same note, wiki pages often contain technical information (e.g. APIs, step-by-step guidelines, etc.), and so do shared files. Wikis and files typically contain collaboration data used for working together and sharing content, which can explain the two indicators being ranked higher when including the content. One participant explained: "*files generally indicate more 'hands on' expertise*."

Forum commenter, which was ranked 7th in the first survey, moved up to 4th in the second survey, outranking *forum author* and *tagged-with*. Forum is the only type of social application where the commenting association was ranked higher than the authorship association. As mentioned before, commenting on a forum topic often provides an answer to a question posted by the author. Survey II participants could see the content of the entire forum thread and realize that the author is the one seeking for an expert, while the commenter is the knowledgeable individual.

As in survey I, authoring and commenting on microblogs were substantially weaker than other content types, indicating that the intuition of participants in survey I, who perceived microblogs as a way to share short messages rather than contribute content, was reinforced when they saw the actual evidence on survey II. Microblogs, shown to be very effective for expertise mining by Guy et al. (2013), were ranked even lower in survey II than in survey I.

The indicator with the largest rank difference between the two surveys is *file reader*, which moved up 5 spots from 13th in survey I to 8th in survey II. When the survey participants viewed the content of the files, the technical and detailed nature of the files might have compensated for the fact that the expert was not the author, but merely a reader. One participant, who gave a rating of 4 to a "security" expert who read 2 related files, commented: "*Great shared information on the topics. May not be his own, but the topics strike at the heart of security*". Other participants who rated file reading experts seem to have considered the fact that the expert made the effort to read the file, as one participant noted: "*Maybe not a top expert, but has taken the time to read on the topic*". Although ranked below average in the context of expertise mining in Guy et al. (2013), file related indicators were rated high by survey II participants, who perceived writing and reading files as a strong indication of expertise.

Another indicator that moved up is *community member*, ranked 14th in survey I and 12th in survey II. After viewing the communities' content, many participants appreciated the fact that the person is part of the community, in spite the fact that they did not own it. A participant who gave a rating of 5 to a "retail" expert, who is a member of 5 communities, commented: "*This member has great access to 'retail' information that can be very beneficial for the end user!*"

At the bottom of the list are the two 'liking' indicators, which received the lowest ratings on both surveys. Liking a content item is perceived as a very poor evidence of a person's expertise, regardless of the quality of the liked content. Several participants commented on this issue, e.g., "*Liking is not promoting nor is it necessarily social, it may be just someone 'checking a box' cause everyone else is doing it*" or "*Shows interest but not expertise*".

Differences in the results of the two surveys also emerge when examining how the count values affected the average scores. Figure 7 shows the top indicators when distinguishing between the two counts (2 and 5). Counts seem to play a stronger role in the second survey than in the first. The first 8 indicators (out of $16 \cdot 2 = 32$) in survey II are of count 5, compared to only the first 4 in survey I. A suggested explanation for this result is the fact that an expertise evidence item in

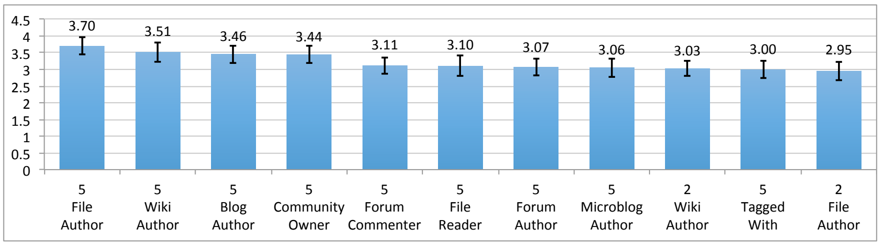

Figure 7. User survey II: Top average ratings for type-count combinations.

survey I was comprised of three indicators, thus the count of each indicator was less significant. On the other hand, an evidence item in survey II included a single indicator, giving more weight to its count.

The results of the two surveys portrait a coherent picture, where heavy content contribution is perceived as a strong indicator of expertise, followed by light and minimal contribution of content. These findings are also consistent with regards to both metrics applied in the first survey (average rating and 'best' selection). The internal order within the three groups, however, was different between the surveys, as a close examination of the content provided participants with additional information and different insights.

Discussion and future work

Our results indicate a substantial difference between social media data used for evidence presentation as compared to expertise mining: for evidence, authorship of "heavy" content is most productive, while more concise indicators, which involve less content, are perceived as weaker evidence. In this section, we summarize the differences, discuss the implications and limitations of our study, and suggest future directions.

In a previous study it was shown, based on self-evaluation ground truth, that when it comes to social media, authorship is not a stronger indicator than content feedback signals, such as commenting, tagging, or liking (Guy et al., 2013). This result, while may seem counter-intuitive, was shown consistent across a variety of social media data sources and for various types of feedback signals. Our own study, by contrast, shows that when it comes to social media-based evidence, users still possess the basic intuition that authorship and management of heavy content are most meaningful. Other forms of feedback are perceived as weaker evidence and the less content they involve, the less convincing they are. Commenting is perceived stronger than tagging, which is typically stronger than reading, liking, or being a member. 'Liking' of content, which was found valuable for expertise mining, was particularly perceived as a weak expertise indicator, even when the number of evidence items was higher.

In the expertise mining study by Guy et al. (2013), profile tagging was identified as the most precise data source for expertise. This was explained by the fact that profile tags are the most direct means for indicating a person's expertise and reflect the perception of a person by others, rather than their own associated content. The profile tagging data source had lower recall, however, since it was not as commonly used as some of the other data sources. The data source with the best combination of precision and recall was microblogs. It was speculated that a person's expertise areas are well propagated through this fast-pace, real-time, concise form of expression. Our experiments for evidence reveal a very different picture: profile tags were perceived weaker evidence than almost all authorship types, such as blogs, wikis, files, and even forums, where authorship often expresses a question rather than an answer. The one authorship type that did not outrank profile tags was microblogs, which had the lowest scores among all authorship indicators. It appears that while profile tags and microblogs serve as good data sources for expert mining and ranking, they are less convincing as evidence items.

These results have various practical implications. Expertise locator designers often need to carefully consider how much and which evidence to present alongside an expert that was found matching the user's query. As demonstrated in our studied organization's EL system, since the user is presented with a list of experts and for each there is a need to show evidence, the real estate for evidence presentation might not be large and require smart selections. This may even be a bigger challenge on mobile devices, which have less user interface space. For example, even if the user interface shows multiple tabs with different evidence types, as in the EL system's web application, the tab order, and particularly the default tab to be presented, play a key role in evidence presentation. Our results indicate that the default choice made in the studied organization to show profile tags may not be ideal; showing authorship of rich-content documents, such as blogs, wikis, or files, may be more desirable.

From a backend/analytics perspective, our results indicate that expertise evidence may require its own calculation. While this may entail further computational cost, our findings suggest that using the scores from the expert ranking algorithms to calculate the top evidence to be presented might not be the best choice: while expertise inference can often be achieved by aggregating many signals such as liked blog posts and "tweets", for evidence it is important to include the most authoritative pieces, if such exist.

We experimented with two different user surveys: the first included synthesized evidence types and counts, without real data, allowing us to artificially compose evidence lists and compare users' perception of them; the second survey delved into presentation of specific evidence item types with actual titles and content pages, originating from a real social media-based expertise location system. The two surveys enabled us to examine users' perception more thoroughly and provided a broader picture. We believe that the results' consistency across the two survey grants more validity to our findings.

Our experiments were conducted in one organization and are naturally influenced by this organization's characteristics and particular way of enterprise social media use. Social media in our organization is used mainly for collaborative work and information sharing, which might explain why employees view authorship of new content as critical. It could be that other organizations, which use such systems mainly for socializing and networking, would see more value in the tagging or liking operations. We hope to see more research on the topic in other environments, which would further validate and extend the results reported here. We note, however, that our study included the most common types of social media applications, each with a large user base of tens of thousands of employees. In addition, both of our surveys were based on responses from hundreds of employees. We therefore believe that our fundamental results regarding authorship, commenting, and liking, and massive versus concise content, are likely to be valid in other organizations as well.

Our evaluation in this paper was based on user surveys. As mentioned, we conducted two different surveys to examine evidence perception from different aspects and validate our results. Yet, both surveys presented anonymous experts and asked participants to assess the level of confidence they have in their expertise based on the presented evidence. We did not use a field study of a live system such as the studied organization's EL system, since it is in early stages of deployment, and its usage level is not broad enough to allow appropriate evaluation. As social media-based expertise locators continue to evolve, future research should apply techniques such as A/B testing, to examine the effect of evidence presentation on real users as they search for experts.

References

Adamic, L.A., Zhang J., Bakshy, E., and Ackerman, M.S. (2008): 'Knowledge sharing and yahoo answers: everyone knows something'. *Proc. WWW '08*, pp. 665-674.

Balog, K., Azzopardi, L., and De Rijke, M. (2006): 'Formal models for expert finding in enterprise corpora'. *Proc. SIGIR '06*, pp. 43-50.

Becerra-Fernandez, I. (2000): 'Facilitating the online search of experts at NASA using expert seeker people-finder'. *Proc. PAKM '00*.

Chi, E. H. (2012): 'Who knows?: searching for expertise on the social web: technical perspective'. *Communications of the ACM*, *55*(4), pp. 110-110.

DiMicco, J., Millen, D. R., Geyer, W., Dugan, C., Brownholtz, B., & Muller, M. (2008): 'Motivations for social networking at work'. *Proc. CSCW '08*, pp. 711-720.

Farrell, S., Lau, T., Nusser, S., Wilcox, E., and Muller, M. (2007): 'Socially augmenting employee profiles with people-tagging'. *Proc. UIST '07,* pp. 91-100.

Guy, I., Avraham, U., Carmel, D., Ur, S., Jacovi, M., and Ronen, I. (2013): 'Mining expertise and interests from social media'. *Proc. WWW '13,* pp. 515-526.

Guy, I., Ronen, I., and Wilcox, E. (2009): 'Do you know?: recommending people to invite into your social network'. In *Proc. IUI '09*, pp. 77-86.

Herlocker, J. L., Konstan, J. A., and Riedl, J. (2000): 'Explaining collaborative filtering recommendations'. *Proc. CSCW '00*, pp. 241-250.

Kardan, A., Garakani, M., and Bahrani, B. (2010): 'A method to automatically construct a user knowledge model in a forum environment'. *Proc. SIGIR '10*, pp. 717-718.

Kolari, P., Finin, T., Lyons, K., and Yesha, Y. (2008): 'Expert search using internal corporate blogs'. *Workshop on Future Challenges in Expertise Retrieval, SIGIR '08*, pp. 2-5.

Li, C. T., & Shan, M. K. (2013): 'X2-Search: Contextual Expert Search in Social Networks'. *Conference on Technologies and Applications of Artificial Intelligence (TAAI),* pp. 176-181

McDonald, D.W. and Ackerman, M.S. (2000): 'Expertise recommender: a flexible recommendation system and architecture'. *Proc. CSCW '00*, pp. 231-240.

Macdonald, C., Hannah, D., and Ounis, I. (2008): 'High quality expertise evidence for expert search'. In *Advances in Information Retrieval*, pp. 283-295.

Noll, M.G., Yeung, A.C., Gibbins, N., Meinel, C., and Shadbolt, N. (2009): 'Telling experts from spammers: expertise ranking in folksonomies'. *Proc. SIGIR '09*, pp. 612-619.

Pu, P., & Chen, L. (2006, January): 'Trust building with explanation interfaces'. *Proceedings of the 11th international conference on Intelligent user interfaces*, pp. 93-100.

Reichling, T. and Wulf, V. (2009): 'Expert recommender systems in practice: evaluating semi-automatic profile generation'. *Proc. CHI '09*, pp. 59-68.

Ronen, I., Guy, I., Kravi, E. and Barnea, M. 'Recommending social media content to community owners'. *Proc. SIGIR '14*, pp. 243-252.

Serdyukov, P. and Hiemstra, D. 2008): 'Being omnipresent to be almighty: the importance of global web evidence for organizational expert finding'. *Workshop on Future Challenges in Expertise Retrieval, SIGIR '08*, pp. 17-24.

Sinha, R., and Swearingen, K. (2002): 'The role of transparency in recommender systems'. In *CHI'02 extended abstracts on Human factors in computing systems*, pp. 830-831.

Tintarev, N., and Masthoff, J. (2007): 'A survey of explanations in recommender systems'. In *23rd International Conference on Data Engineering Workshop,* pp. 801-810.

Varshney, K. R., Chenthamarakshan, V., Fancher, S. W., Wang, J., Fang, D., and Mojsilovic, A. (2014): 'Predicting Employee Expertise for Talent Management in the Enterprise'. *Proc. KDD '14*, pp. 1729-1738.

Vig, J., Sen, S., and Riedl, J. (2009): 'Tagsplanations: explaining recommendations using tags'. In *Proceedings of the 14th international conference on Intelligent user interfaces*, pp. 47-56.

Xu, Z. (2014). '*Expertise Retrieval in Enterprise Microblogs with Enhanced Models and Brokers'* (Doctoral dissertation, The Ohio State University).

Yiman-Seid, D. and Kobsa, A. (2003): 'Expert-finding systems for organizations: Problem and domain analysis and the DEMOIR approach'. *JOCEC 13, 1* (2003), pp. 1-24.

Zhang, J., Ackerman, M.S., and Adamic, L. (2007): 'Expertise networks in online communities: structure and algorithms'. *Proc. WWW '07*, pp. 221-230.