Bogers, T., Thoonen, W., & van den Bosch, A. (2006). Expertise classification: Collaborative classification vs. automatic extraction. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-20. doi: 10.7152/acro.v17i1.12488

EXPERTISE CLASSIFICATION: COLLABORATIVE CLASSIFICATION VS. AUTOMATIC EXTRACTION

Toine Bogers <a.m.bogers@uvt.nl>
Willem Thoonen <w.w.t.thoonen@uvt.nl>
Antal van den Bosch <antal.vdnBosch@uvt.nl>
ILK / Language and Information Science
Tilburg University, P.O. Box 90153
NL-5000 LE Tilburg, The Netherlands

1. Introduction

Social classification is the process in which a community of users categorizes the resources in that community for their own use. Given enough users and categorization, this will lead to any given resource being represented by a set of labels or descriptors shared throughout the community (Mathes, 2004). Social classification has become an extremely popular way of structuring online communities in recent years. Well-known examples of such communities are the bookmarking websites Furl (http://www.furl.net/) and del.icio.us (http://del.icio.us/), and Flickr (http://www.flickr.com/) where users can post their own photos and tag them.

Social classification, however, is not limited to tagging resources: another possibility is to tag people, examples of which are Consumating (http://www.consumating.com/), a collaborative tag-based personals website, and Kevo (http://www.kevo.com/), a website that lets users tag and contribute media and information on celebrities.

Another application of people tagging is *expertise classification*, an emerging subfield of social classification. Here, members of a group or community are classified and ranked based on the expertise they possess on a particular topic. Expertise classification is essentially comprised of two different components: *expertise tagging* and *expert ranking*. Expertise tagging focuses on describing one person at a time by assigning tags that capture that person's topical expertise, such as ‘speech recognition' or ‘small-world networks'.

Expert ranking, on the other hand, is more task-specific and focuses on ranking the members of a group or community based on their expertise. These rankings are dependent on the topic of an

ISSN: 2324-9773

Bogers, T., Thoonen, W., & van den Bosch, A. (2006). Expertise classification: Collaborative classification vs. automatic extraction. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-20. doi: 10.7152/acro.v17i1.12488

information request, such as, for instance, a query submitted to a search engine. Methods are developed to combine the information about individual members' expertise (tags), to provide on-the-fly query-driven rankings of community members.

Expertise classification can be done in two principal ways. The simplest option follows the principle of social bookmarking websites: members are asked to supply tags that describe their own expertise and to rank the other community members with regard to a specific request for information. Alternatively, automatic expertise classification ideally extracts expertise terms automatically from a user's documents and e-mails by looking for terms that are representative for that user. These terms are then matched on the information request to produce an expert ranking of all community members. In this paper we describe such an automatic method of expertise classification and evaluate it using human expertise classification judgments. In the next section we will describe some of the related work on expertise classification, after which we will describe our automatic method of expertise classification and our evaluation of them in sections 3 and 4. Sections 5.1 and 5.1 describe our findings on expertise tagging and expert rankings, followed by discussion and our conclusions in section 6 and recommendations for future work in section 7.

2. Experts & expertise

Before we can start to extract expertise automatically, we need to get a clear picture of what expertise exactly is and what constitutes an expert. Decades of psychological research into the field focused mainly on quantifiable skills such as playing chess, which can be readily measured and subjected to laboratory experiments (Ross, 2006). Expertise in softer, non-competitive areas such as knowledge organization and dissemination in universities and commercial organizations has been researched using interviews, questionnaires and social network analyses.

Many definitions of expertise have been proposed over the years, each highlighting different individual and social aspects. Individually speaking, expertise is often defined in terms of superior analytical and creative abilities, and the ability to process and apply new information faster than non-experts can (Salthouse, 1991). From a social viewpoint, experts are often regarded as such by other people in the community partly because they meet certain criteria such as specific certifications or diplomas (Sternberg, 1994).

ISSN: 2324-9773

Bogers, T., Thoonen, W., & van den Bosch, A. (2006). Expertise classification: Collaborative classification vs. automatic extraction. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-20. doi: 10.7152/acro.v17i1.12488

In more recent years, automatic methods for expertise classification have been proposed due to the increasing popularity and practical importance of searching for and finding experts in real organizations. Some of the first attempts at expertise classification were reported by Campbell et al. (2003), who experimented with the same kind of identification task. They implemented simple keyword matching in conjunction with link analysis to adequately identify experts in a corpus of e-mail messages sent between people in the same company. TREC 2005 marked the introduction of the `Expert Search Task', aimed at solving the problem of identifying employees who are the experts on a certain topic or in a certain situation (TREC, 2005).

A more recent commercial venture that leverages expertise in social networks is Illumio (http://www.illumio.com/), a software agent that extracts the particulars of a user's expertise and social network by mining e-mail messages, documents, and other data on a user's computer. Using a reverse auction algorithm, it goes from most expert to least, seeking to find an individual willing to answer the question that is being asked.

In our 2006 papers (Bogers & Van den Bosch, 2006, 2006b), we describe a new automatic method called *authoritative re-ranking* that performs expertise tagging and expert ranking. In the first step of the automatic expertise classification phase we extract the expertise terms for each workgroup author, i.e. automatic expertise tagging. We combine these expertise terms and their associated weights to rank all workgroup members on their expertise on any possible query topic. These expert rankings are then used in the second step to improve the information retrieval (IR) process, which none of the aforementioned approaches have done. In this paper, we evaluate the expertise classification component of our approach (described in sections 3 and 4) by contrasting it with a manual collaborative classification approach to expertise.

3. Automatic extraction

Authoritative re-ranking was designed to improve the IR process within workgroups and scientific communities. Research has shown that colleagues are one of the first and preferred sources of information (Procter, 1998; Adar, 1999; Hertzum, 2006). For this reason, we developed a method of modeling the expertise in workgroups and producing rankings of the group members based on their expertise on the topic of a specific query. The second component of authoritative re-ranking reshuffles the original search results using those expert rankings.

ISSN: 2324-9773

Bogers, T., Thoonen, W., & van den Bosch, A. (2006). Expertise classification: Collaborative classification vs. automatic extraction. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-20. doi: 10.7152/acro.v17i1.12488

3.1 Automatic expertise tagging

In the first step of the automatic expertise classification phase we extract the expertise terms for each workgroup author, i.e. automatic expertise tagging. We assume that we can estimate the expertise of each member of the workgroup from the aggregated content of his or her publications[1]. Our re-ranking approach was designed to be used on top of a basic TF·IDF vector space model of information retrieval.

In our experiments, we used the formulas for document weights and query weights as determined by Chisholm (1999). In addition, we incorporated some tried and tested low-level NLP-techniques into our baseline system, such as stop word filtering and stemming. This resulted in a single list of all the informative terms in the collection of all workgroup publications. We then estimated how well each term or phrase pointed to each member by calculating the author-term co-occurrence weights using the Information Gain metric (Zheng & Srihari, 2003). Sorting these lists yields the expertise tags/terms for an author. We calculated these lists for each workgroup member, which resulted in a matrix of term-author expertise weights. Table 1 shows a small part of this matrix.

Table 1. A small part of the author-term expertise weights matrix

term	author A	author B	author C	author D
generalization performance	0.01590	0.00313	0.00019	0.00012
machine learning	0.00400	0.00390	0.00169	0.00400
maximum entropy	0.01587	0.00254	0.00011	0.00009
named entity recognition	0.01592	0.00089	0.00019	0.00015
search results	0.02372	0.00393	0.00015	0.00012
semantic role labelling	0.00858	0.00149	0.00779	0.00018

[1] See section 4.1 for evidence that supports this assumption.

ISSN: 2324-9773

Bogers, T., Thoonen, W., & van den Bosch, A. (2006). Expertise classification: Collaborative classification vs. automatic extraction. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-20. doi: 10.7152/acro.v17i1.12488

3.2 Automatic expert ranking

Ranking the members of a group or community based on their expertise is completely dependent on the topic of an information request, such as a query submitted. Therefore, we developed a way of combining the weighted expertise tags derived in the previous step to rank all workgroup members on their expertise on any possible query topic.

Calculating the expert scores is based on the straightforward assumption that if terms characteristic for author *A* occur in query *Q*, *A* is likely to be more of an expert on *Q*. We extract the most important query terms from each query as described in section 3.1 and look up the corresponding author-term expertise weights in the matrix. We then calculate a weighted average for each author using the query terms' TF·IDF values as the weighting factor. TF·IDF is a popular IR metric for determining the informativeness of a term for a particular text or query and is calculated by multiplying the term frequency of that term in the text with its inverse document frequency.

3.3 Re-ranking of search results

The second component of authoritative re-ranking reshuffles the original search results using those expert rankings. This re-ranking is based on the premise that the documents authored by the experts on a specific query topic are more likely to be relevant to that query. After re-ranking, documents written by experts receive a higher weight while the influence of documents of non-experts is downplayed. See Bogers & Van den Bosch (2006, 2006b) for more information about the re-ranking process. Re-ranking in combination with the automatic expertise classification successfully improved the performance of a baseline IR system with statistically significant gains in R-precision ranging from 1.5% to over 34%.

3.4 Earlier evaluation

Investigating the merits of authoritative re-ranking required testing our approach on test collections that (a) contain information about the authors of each document, and (b) are a realistic representation of a community, such as a workgroup or a scientific community. We used two well-known test collections, CACM and CISI, that both represent scientific communities. CACM is a reference collection composed of all the 3204 article abstracts published in the

ISSN: 2324-9773

Bogers, T., Thoonen, W., & van den Bosch, A. (2006). Expertise classification: Collaborative classification vs. automatic extraction. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-20. doi: 10.7152/acro.v17i1.12488

Communications of the ACM journal from 1958 to 1979, and CISI is made up of 1460 document abstracts selected from a previous collection assembled at ISI (Van Rijsbergen, 1979).

We know of no publicly available IR test collections that represent the body of work published by a workgroup operating in a single institution, which prompted us to create our own: the ILK test collection[2]. ILK contains 166 document titles and abstracts of publications of current and ex-members of the ILK workgroup[3]. The topics of the papers are in the area of machine learning for language engineering and linguistics with subtopics ranging from speech synthesis, morphological analysis, and text analysis to information extraction, text categorization, and information retrieval. We asked the current group members to provide us with queries and the corresponding binary relevance assignments for each document, which resulted in 80 natural language queries. An example of such a query is "can rule induction be used for feature construction in learning language processing tasks?".

4. Collaborative classification

The preliminary evaluation of authoritative re-ranking focused on the evaluation of the combined system. The expertise classification component was evaluated implicitly: if the final re-ranking step produced significant improvements, then the expertise classification step was assumed to be good as well. In this section we describe how we evaluated our automatic expertise tagging and expert ranking components more directly by contrasting them with social classification approach to expertise.

In order to evaluate our automatic approach we need the community members of one or more of our test collections to provide us with expertise tags and expert rankings. The CISI and CACM collections are unsuitable because of their scale and age, so we created a expertise classification questionnaire tailored to the current ILK workgroup members (Thoonen, 2006).

[2] Publicly available at http://ilk.uvt.nl/apropos/

[3] The Induction of Linguistic Knowledge (ILK) workgroup is part of the Department of Language and Information Science of the Faculty of Arts of Tilburg University. It focuses mainly on machine learning for language engineering and linguistics.

ISSN: 2324-9773

Bogers, T., Thoonen, W., & van den Bosch, A. (2006). Expertise classification: Collaborative classification vs. automatic extraction. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-20. doi: 10.7152/acro.v17i1.12488

The ILK workgroup consists of 19 members, 12 of which were included in the questionnaire by extracting their expertise automatically[4]. Ten of the members participated in the questionnaire leading to a response rate of 86.6%. The questionnaire consisted of two main parts, focusing on expertise tagging and expert ranking of all 12 originally selected members. We describe and motivate the questionnaire in the next two subsections; the results are discussed in sections 5.1 and 5.2.

4.1 Expertise tagging

We gave the participants in the questionnaire two tasks that focused on expertise tagging. In the first task they were asked to provide at least 5 keywords or terms that they feel describe their own expertise, such as `information retrieval', `POS tagging' or `speech recognition'.

In the second task, we presented the participants with two lists of automatically extracted expertise terms that were sorted on the expertise weights for each participant separately. The first list was extracted using the optimized settings for our authoritative re-ranking approach as described in Thoonen (2006).. This list—from now on referred to as the *optimal* list—contained 1884 different terms and expertise weights.

However, upon closer inspection of this list, it appeared to contain many single word terms that, combined, are representative of an author's work, but not very informative in terms of expertise, such as ‘data set' and ‘experiments'. We therefore re-ran our authoritative re-ranking experiments with stricter settings[5] that produced fewer terms that humans would consider ‘noisy' and fewer terms in general. This list—from now on referred to as the *strict* list—contained 273 different terms and expertise weights. Finally, to further reduce noise, we also filtered both lists of terms semi-automatically by removing names, URLs and non-English terms.

In the end, we presented the participants with the top 20 terms from each list and asked to rate each of the, in total, 40 expertise tags on how well the term represented the participant’s expertise. Rating was on a 5-point Likert scale from 1 (*very poor*) to 5 (*very well*). We had to restrict the rating process to the two lists of top 20 terms because of time restrictions; we did not

[4] Excluded were members without any publications and ILK's scientific programmers.

[5] For instance, we increased the thresholds that filtered out words that did not occur enough times. See \cite{Thoonen:2006} for details on the optimal settings and these stricter settings.

ISSN: 2324-9773

Bogers, T., Thoonen, W., & van den Bosch, A. (2006). Expertise classification: Collaborative classification vs. automatic extraction. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-20. doi: 10.7152/acro.v17i1.12488

wish to exceed a maximum length of 30 minutes for the questionnaire in order to maximize the response rate and the quality of the responses. For the same reason, we were not able to have participants collaboratively tag each other's expertise by asking them to provide expertise tags for their colleagues. This would have required 30 (10 + 20) tagging tasks instead of 3. Furthermore, we also did not ask participants to tag each other's expertise, as we believe it is more difficult for participants to assign specific expertise tags to their colleagues than to themselves.

We also asked our participants to rate on a 5-point Likert scale (from 1 (*completely disagree*) to 5 (*completely agree*)) whether or not they believe that scientific publications are a good source for identifying an author's expertise. In essence, this partly addresses our assumption made in section 3.1 about publications representing their author's expertise. This assertion was received with an average rating of 4.4 (with 5 being the maximum), indicating that the ILK members also believe that publications are a good representation of a person's academic expertise.

4.2 Collaborative expert ranking

In the expert ranking part of the questionnaire we presented the participants with 10 of the 80 natural language queries. For each of the queries, we asked them to rank the members of the ILK workgroup, including themselves, on their expertise on the query topic, i.e. which colleagues would they turn to with this question and in what order. Workgroup members with no expertise on the topic were to be left blank. Again, we could not ask our participants to do all 80 queries because of time restrictions.

By directly asking the participants to rank the experts in order of expertise, we obtain ranked lists, providing us with the possibility to evaluate the results at a higher level of granularity than, for instance, the W3C corpus used in the TREC Enterprise track (TREC, 2005), which uses only binary relevance judgments: either a person is an expert or not.

The expert ranking part resulted in 10 expert rankings for each of the 10 queries, so we still needed to combine these rankings into one single ranking for each query. The collection of expert rankings for a single query can be seen as a collection of votes for each group member. In creating this final ranking, we wanted to take into account both the vote counts and positions. We used a variation of the Mean Reciprocal Rank (MRR) measure—used to evaluate Question Answering systems—called Normalized Reciprocal Rank (NRR). Using our NRR metric we first

ISSN: 2324-9773

Bogers, T., Thoonen, W., & van den Bosch, A. (2006). Expertise classification: Collaborative classification vs. automatic extraction. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-20. doi: 10.7152/acro.v17i1.12488

calculate the sum of the reciprocal rank[6] (SRR) of all the votes for each author as shown in (1) and then normalize that vector of author SRR scores. This yields the NRR scores for each author.

(1) $$SRR = \sum_{i}^{n} \frac{1}{rank(i)}$$

Table 2 shows a small example of possible votes for three group members. The SSR for author A would be calculated as follows: (3 x $^1/_1$) + (0 x $^1/_2$) + (1 x $^1/_3$) = 3 + 0 + $^1/3$ = 3.333. After normalization over the three scores, NRR_A would be equal to 0.710.

Table 2. A toy example of voting for experts and the corresponding NRR scores

rank	**author A**	**author B**	**author C**
1	3	1	1
2	0	3	1
3	1	0	2
SRR	3.333	2.5	2.167
NRR	0.710	0.532	0.461

We chose to normalize the sum instead of calculating the mean (as in MRR), because MRR does not distinguish between an author with 1 first-place vote and another author with 4 first-place votes. The second author should be ranked higher, but the MRR metric does not take this into account. We normalized the reciprocal rank sum to obtain a convenient value between 0 and 1.

For each query separately, we calculated the NRR scores for all authors and sorted these scores to get the human expert ranking. Members with no votes were sorted alphabetically and added to the end of the ranking. This way we obtain a gold standard of expertise, which is the closest approximation we can make of a collaborative tagging community in our workgroup. In general, IR and natural language processing tasks are preferably evaluated against human performance since man is supposed to be the yardstick of machine intelligence. McDonald (2001) also provides evidence for this: he found that people are relatively good at making judgments about

[6] The reciprocal rank is the reciprocal of the rank of a particular vote. For example, the reciprocal rank of a third-place vote is $^1/_3$.

ISSN: 2324-9773

Bogers, T., Thoonen, W., & van den Bosch, A. (2006). Expertise classification: Collaborative classification vs. automatic extraction. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-20. doi: 10.7152/acro.v17i1.12488

other people's expertise. We therefore consider these expert rankings to be the gold standard by which we should evaluate our automatic approach. Yet another way of looking at constructing rankings from these votes is from the IR point of view. Each vote can be regarded as a relevance judgment and by pooling these judgments we have the relevance information needed to evaluate our automatic approach.

5. Results

The next two subsections present the results of the evaluation of the automatic expertise classification tasks expertise tagging and expert ranking.

5.1 Expertise tagging

The first expertise tagging task in the questionnaire required the participants to provide us with their own expertise keywords and terms. The 10 participants entered a total of 69 terms with an average of 6.9 terms (st. dev = 3.29, range 4-13). The total tag set contained 53 unique terms with 8 terms occurring more than once. These 8 terms are listed in Table 3 and clearly match the overall research focus of the ILK workgroup.

Table 3. A list of expertise tags that occur more than once

term	**frequency count**
machine learning	7
information extraction	4
memory-based learning	3
text-to-speech	2
speech synthesis	2
prosody	2
natural language processing	2
computation linguistics	2

Of these 53 unique tags, 35 (66.0%) were multiword terms with 27 bigrams (such as “machine learning”), 7 trigrams (e.g. “named entity recognition”) and 1 4-gram (“machine learning of

ISSN: 2324-9773

Bogers, T., Thoonen, W., & van den Bosch, A. (2006). Expertise classification: Collaborative classification vs. automatic extraction. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-20. doi: 10.7152/acro.v17i1.12488

language"). With one exception ("prosody"), the only terms that were used by more than one participant were multiword terms. This seems to suggest that humans favor multiword terms for describing their expertise. Many of these terms appear to be higher level descriptive terms that rarely occur in the author's papers themselves, but, when combined, describe the topic of the papers quite well.

In the second expertise tagging task, participants were asked to rate their own top 20 expertise terms from the two lists of automatically extracted terms, corresponding to the optimal and strict settings. Table 4 below shows the average ratings for each author and the global average of all authors combined.

Table 4. The average term ratings of each ILK participant for the optimal and strict terms

ILK member	optimal avg. term rating	st.dev.	strict avg. term rating	st.dev.
1	2.05	1.47	1.7	1.22
2	2.8	1.7	2.8	1.64
3	4	0.79	4	0.72
4	3.6	1	3.25	1.41
5	3.6	1.31	3.5	1.64
6	4.1	0.97	4.2	0.83
7	3.1	1.44	3.65	1.35
8	3	1.12	3.75	0.96
9	3.2	1.44	3.8	1.1
10	2.6	1.73	3.85	1.27
average	3.21	1.30	3.45	1.21

Although the small size of the ILK workgroup and imposed time restrictions make it difficult to draw any definite conclusions, the strict terms appear to be rated slightly higher on average (with a slightly lower standard deviation) by the participants than the terms that were optimal for the computer.

ISSN: 2324-9773

Bogers, T., Thoonen, W., & van den Bosch, A. (2006). Expertise classification: Collaborative classification vs. automatic extraction. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-20. doi: 10.7152/acro.v17i1.12488

If we restrict ourselves to the terms that were rated as 'good' (with a rating of 4 or 5), then 56.5% of the strict terms were rated as good, as opposed to 47% of the optimal terms. All in all, the strict settings seem to have been rated slightly better than the optimal settings.

In addition to these comparisons, we also directly contrasted the human-provided terms with the two automatic term lists by analyzing the occurrence of the human terms in the optimal and strict lists. We first looked for exact matches between human terms and automatically extracted terms. On average, only 22% of the gold standard terms occurred exactly in the optimal list and around 28% occurred in the strict list. This means that slightly more of the strict terms matched the human terms.

If we compare the participant-provided expertise tags with the automatically extracted ones, they are much more often single word terms. The optimal list contained 98.5% single word terms and the strict list 74%. Furthermore, many of the single word terms are not always very descriptive by themselves, such as "data", "performance" and "experiment". This is also evident when we compare the average ratings of single word terms, bigrams, and trigrams in the strict list: the average trigram rating was 4, the average bigram rating was 3.86, and the average rating for single word terms was 3.29. Humans appear to have a clear preference for using more general bigrams and trigrams, as evident from their own terms and the slight preference for the strict list, which contained more bigrams and trigrams.

5.2 Expert ranking

We presented the participants with 10 natural language queries in the expert ranking part of the questionnaire and asked them to rank the ILK members, including themselves, on their expertise on the query topic. For each query, we then calculated the gold standard ranking using NRR as described in section 4.2. Table 5 shows the gold standard NRR scores for each author-query combination.

ISSN: 2324-9773

Bogers, T., Thoonen, W., & van den Bosch, A. (2006). Expertise classification: Collaborative classification vs. automatic extraction. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-20. doi: 10.7152/acro.v17i1.12488

Table 5. The expert scores (calculated using NRR) as determined by the ILK participants themselves

HUMAN	**query**										
author	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**average**
A-01	-	-	-	-	0.02	**0.55**	-	-	-	-	0.29
A-02	**0.88**	**0.33**	**0.42**	**0.76**	**0.78**	**0.61**	0.2	**0.51**	**0.39**	**0.44**	0.53
A-03	0,02	0,16	-	**0,47**	-	-	-	0.03	-	0.02	0.03
A-04	0.2	-	-	-	0.15	0.04	-	0.01	-	0.02	0.08
A-05	**0.29**	**0.92**	0.06	0.24	**0.54**	**0.38**	0.09	**0.86**	0.08	0.21	0.37
A-06	-	-	-	-	-	0.06	-	-	-	-	0.06
A-07	-	-	**0.89**	-	-	-	**0.93**	0.02	**0.91**	-	0.69
A-08	0.05	0.09	0.15	**0.37**	-	-	**0.27**	-	0.07	-	0.17
A-09	-	0.04	0.05	0.04	-	-	0.06	0.01	-	-	0.04
A-10	-	0.06	-	-	-	0.02	0.04	-	0.06	-	0.05
A-11	**0.29**	-	-	-	0.21	0.21	-	0.02	-	**0.87**	0.32
A-12	0.09	-	-	0.03	0.17	**0.36**	-	0.03	-	0.02	0.12
average	0,26	0,27	0,31	0,32	0,31	0,28	0,23	0,19	0,30	0,26	

The data in Table 5 shows that on average only 2 to 3 ILK members were singled out as experts, showing that the participants have a clear sense of which members are experts on which topics. For all but one query there is only one expert whose NRR score is more than one standard deviation higher than the average rating. This means that in 9 out of 10 cases the participants picked out a clear no. 1 expert and that, in general, there seems to be much agreement between the different participants when assigning experts.

Figure 1 shows an example graph of ILK members referring to themselves as experts. This graph was drawn for query 9 “How to detect miscommunications in human-machine dialogues using machine learning?” Author A-07 has the highest NRR score on this query and is clearly also the designated go-to expert if we look at the high number of incoming vertices in the graph.

ISSN: 2324-9773

Bogers, T., Thoonen, W., & van den Bosch, A. (2006). Expertise classification: Collaborative classification vs. automatic extraction. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-20. doi: 10.7152/acro.v17i1.12488

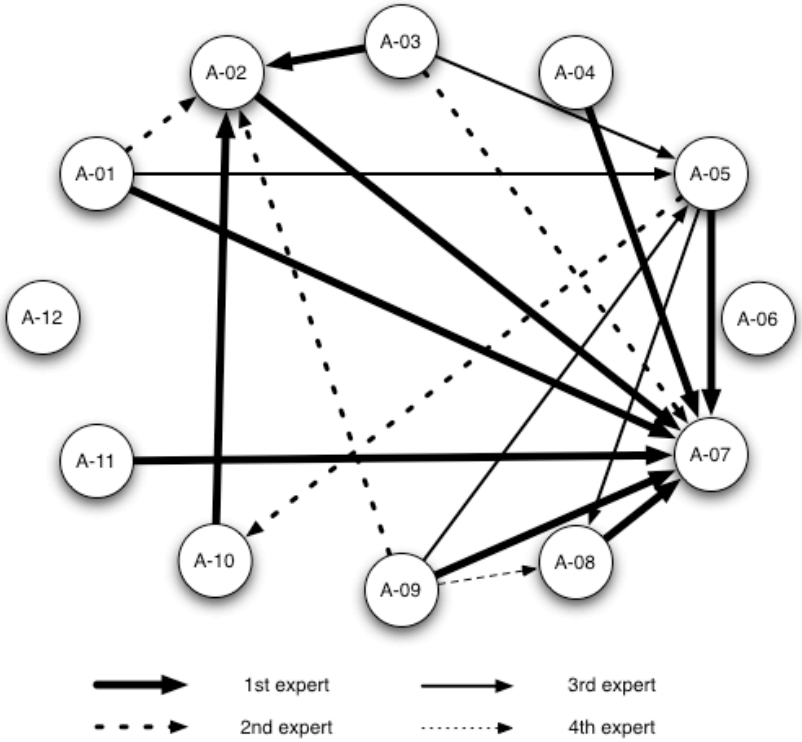

Figure 1. A social graph that displays the voting process for the different experts

Table 6. The expert scores as calculated using the strict settings

STRICT	**query**										
author	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**average**
A-01	**0.11**	0.02	0.07	0.02	**0.26**	**0.57**	0.01	**0.25**	0.08	0.01	0.14
A-02	**0.23**	**0.21**	**0.25**	**0.42**	**0.32**	0.14	**0.11**	**0.35**	**0.19**	**0.20**	0.24
A-03	0.05	0.02	0.07	**0.24**	0.05	0.05	0.02	0.04	0.08	0.02	0.06
A-04	**0.15**	0.02	0.06	**0.13**	**0.16**	0.04	**0.12**	**0.15**	0.07	0.02	0.09
A-05	**0.16**	**0.10**	**0.33**	**0.19**	**0.29**	**0.31**	**0.20**	**0.31**	**0.53**	**0.18**	0.26
A-06	0.04	0.02	0.08	0.02	0.04	0.13	0.01	0.04	0.09	0.01	0.05
A-07	0.05	0.02	**0.31**	0.09	0.05	0.05	**0.12**	0.04	0.08	0.02	0.08
A-08	0.04	0.02	0.13	0.02	0.04	0.03	**0.19**	0.04	0.05	0.01	0.06
A-09	0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.00	0.01
A-10	0.06	0.03	0.09	0.04	0.06	0.12	0.02	0.05	0.10	0.02	0.06
A-11	**0.22**	0.03	**0.20**	0.04	0.06	**0.27**	0.02	0.07	**0.16**	**0.74**	0.18
A-12	**0.13**	0.05	0.11	**0.16**	0.08	0.14	0.03	0.08	0.11	0.04	0.09
average	0.10	0.05	0.14	0.11	0.12	0.16	0.07	0.12	0.13	0.11	

ISSN: 2324-9773

Bogers, T., Thoonen, W., & van den Bosch, A. (2006). Expertise classification: Collaborative classification vs. automatic extraction. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-20. doi: 10.7152/acro.v17i1.12488

Table 7. The performance of the three different automatic systems on each query, calculated using MSE against the gold standard ranking

query	baseline	optimal	strict
1	5.14	2.43	4.00
2	16.17	23.83	12.33
3	17.17	11.60	10.67
4	20.00	12.50	19.67
5	5.33	2.33	4.50
6	12.13	12.00	8.25
7	19.67	11.67	12.00
8	4.63	17.63	10.63
9	8.40	8.80	23.40
10	7.00	12.33	2.00
average MSE	11.56	11.51	10.75

6. Discussion & conclusions

In this paper we described the evaluation of automatic methods of expertise extraction using collaborative tagging techniques. We evaluated both the expertise tagging component, where tags that describe a person's expertise are extracted or elicited, as well as the expert ranking phase, where group members are ranked with regard to a certain query topic.

From our expertise tagging experiments and evaluation, it seems that people demonstrate a clear preference for a small number of higher level descriptive bigrams and trigrams, while the automatic methods employ a much larger set of mostly single word terms, that are related to an author's work but are not directly descriptors. A possible explanation for this is that humans and computers have different limits to how many terms they can actively consider when making (expertise) judgments. Miller (1956) estimated the limits of working memory and showed that people can contemplate only five to nine items at a time. Miller noted that according to this theory, it should be possible to effectively increase short-term memory for low-information-content items by mentally recoding them into a smaller number of high-information-content items. By packing related sets of expertise terms into higher level multiword descriptors, humans

ISSN: 2324-9773

Bogers, T., Thoonen, W., & van den Bosch, A. (2006). Expertise classification: Collaborative classification vs. automatic extraction. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-20. doi: 10.7152/acro.v17i1.12488

Table 6 contains the expert scores for each author-query combination as produced by the authoritative re-ranking approach using the strict settings respectively. We only show the results of the strict settings because they performed best on the evaluation.

The computer assigns an expert score that is higher than average to around 4 members for most queries.

We also created a baseline expert ranking to test our expert ranking algorithm against. The principle behind the baseline is that the position in the ranking should be proportionate to the number of publications. People who have (co-)authored more publications have a better chance to have become an expert (and on multiple topics). The author with the highest number of publications is ranked first, the second most productive author is ranked second, and so on. Because of this, the baseline ranking is the same for every query. We compared the baseline ranking and the two automatic rankings to the gold standard ranking using the Mean Squared Error metric described in (2).

(2)
$$MSE = \frac{\sum_i^n (i - r_i)^2}{n}$$

In this formula, i is the gold standard rank of expert i and r_i is the rank of that expert in one of the automatic rankings. The lower the MSE value, the better the match with 0 being the lowest value, signifying two identical rankings. Table 7 below shows the MSE scores for the different rankings when compared to the gold standard ranking.

The scores in Table 7 show that both automatic systems beat the baseline, albeit barely in the case of the optimal settings.

ISSN: 2324-9773

Bogers, T., Thoonen, W., & van den Bosch, A. (2006). Expertise classification: Collaborative classification vs. automatic extraction. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-20. doi: 10.7152/acro.v17i1.12488

may use a much smaller number of expertise chunks (6.9 on average) than the much larger set of related terms they represent.

These findings suggest that any systems that attempt any kind of automatic resource tagging should take into consideration that humans have a very different view of what are descriptive terms and that they typically use a small number of tags. This is also the case with Flickr where over 80% of the bookmarks are assigned less than five tags (Keller, 2006). Higher level taxonomy terms might be preferred: the questionnaire participants rated the terms from the strict list higher and these terms were more similar to their own provided keywords.

The experiments with creating and evaluating expert rankings compared a baseline approach to two different automatic expert ranking approaches. The baseline expert ranking that focused only on publication count was a fairly strong baseline. Veteran group members are often seen as experts by their group members because of their years of experience and usually have some degree of expertise on many of the group's research subjects. Both our automatic expert ranking approaches beat the baseline, albeit just barely in the case of the system of which the settings were optimized for the optimal re-ranking performance.

The expert rankings constructed based on the strict settings produced the best results when compared to the gold standard. This seems to suggest that forcing the computer to use more 'human' expertise terms when calculating the expert scores brings the computer performance closer to the human rankings.

One possible and likely explanation is that, apparently, successfully re-ranking search results requires a different kind of expert ranking as its input than when the expert ranking itself is the desired end result. At any rate, more research needs to be done on this issue. Another possible influence on the evaluation is the way we constructed the gold standard expert ranking: using another metric than NRR might lead to a slightly different ranking.

One of the problems of our evaluation approach is the size of the data set: 10 participants is barely enough. Arguably, expertise estimations become more reliable as the number of participants increases. However, it is a realistic and typical situation that workgroups are composed of around 10 to 50 people, each with specific interests and not all with enough (co-)authored publications for a reliable expertise extraction, At the same time, according to

ISSN: 2324-9773

Bogers, T., Thoonen, W., & van den Bosch, A. (2006). Expertise classification: Collaborative classification vs. automatic extraction. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-20. doi: 10.7152/acro.v17i1.12488

McDonald (2001), people are usually in good agreement when judging each other's expertise and in our analysis of the expert ranking by the participants we also found that people have a clear preference for two to three experts with always one go-to expert being assigned in 90% of the cases. Therefore, the small sample size need not influence the results to such a large extent.

Another issue is that novice workgroup members can never have the same knowledge of the expertise of their fellow group members. This too, however, is a realistic situation and one of the situations where an automatic expert ranking system would actually be very useful.

7. Future work

We are in the process of making these results available as a complete test collection so that other researchers can test out their own expertise extraction methods[7].

One issue we would like to investigate is whether it is possible to create a hybrid expert ranking system that directly uses the user-provided expertise tags to calculate expert scores and generate the expert ranking. We are interested to see whether this produces a better expert ranking and how well the workgroup search results can be re-ranked using this hybrid expert ranking.

It should also be possible to cluster related computer-extracted terms and link them to higher level descriptive terms. One possibility of doing this would be using a scientific taxonomy such as ACM's topic hierarchy and grouping the most representative words together for each ACM topic. It would be interesting to see if such taxonomy clusters could be used to increase the quality of the predicted expertise tags and the quality of the expert ranking.

Another interesting follow-up experiment would be to have participants perform collaborative expertise tagging in a follow-up questionnaire by letting them tag not only their own expertise but that of their colleagues.

Finally, we would also like to investigate whether it is helpful to use the references in each publication to determine the expertise or authority of the group members. Authors with a large number of referenced papers are more likely to be seen as experts.

[7] This updated test collection will be made available at http://ilk.uvt.nl/apropos/

ISSN: 2324-9773

Bogers, T., Thoonen, W., & van den Bosch, A. (2006). Expertise classification: Collaborative classification vs. automatic extraction. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-20. doi: 10.7152/acro.v17i1.12488

References

Adar, E., Kargar, D., and Stein, L. (1999). Haystack: Per–user Information Environments. *CIKM '99: Proceedings of the Eighth International Conference on Information and Knowledge Management*: 413–422.

Bogers, T., and Van den Bosch, A. (2006). Authoritative re-ranking in fusing authorship-based subcollection search results. In F. de Jong and W. Kraaij (eds.), *Proceedings of the Sixth Belgian-Dutch Information Retrieval Workshop (DIR-2006)*: 49–55. Enschede: Neslia Paniculata.

Bogers, T., and Van den Bosch, A. (2006). Authoritative re-ranking of search results. *Proceedings of the 28th European Conference on Information Retrieval (ECIR 2006)*, vol. 3936 of *Lecture Notes on Computer Science*: 519–522, Berlin: Springer Verlag.

Campbell, C., et al. (2003). Expertise identification using email communications. *CIKM '03: Proceedings of the Twelfth International Conference on Information and Knowledge Management*: 528–531.

Chisholm, E., and Kolga, T. (1999). New term weighting formulas for the vector space method in information retrieval. Technical report ORNL/TM-13756, Computer Science and Mathematics Division, Oak Ridge National Laboratory.

Hertzum, M., and Pejtersen A. M. (2006) The information-seeking practices of engineers: Searching for documents as well as for people. *Information Processing and Management*, 36(5): 761–778.

Keller, P. (2006). Delicious statistics. Retrieved August 30, 2006, from http://www.pui.ch/phred/archives/2005/12/delicious-statistics.html

Mathes, A. (2004). Folksonomies: Cooperative classification and communication through shared metadata. Technical report, University of Illinois Urbana-Champaign.

McDonald, D. W. (2001). Evaluating expertise recommendations. *Proceedings of the ACM 2001 International Conference on Supporting Group Work (GROUP'01)*: 214–223.

ISSN: 2324-9773

Bogers, T., Thoonen, W., & van den Bosch, A. (2006). Expertise classification: Collaborative classification vs. automatic extraction. 17th Annual ASIS&T SIG/CR Classification Research Workshop, 1-20. doi: 10.7152/acro.v17i1.12488

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63: 81-97.

Procter, R., et al. (1998). Genres in support of collaborative information retrieval in the virtual library. *Interacting with Computers* 10(2): 157–172.

Ross, P. E. (2006). The expert mind. *Scientific American*, August 2006.

Salthouse, T. (1991). Expertise as the circumvention of human processing limitations. In *Toward a general theory of expertise* (pp. 286-300). Cambridge: Cambridge University Press.

Sternberg, R. (1994). Cognitive conceptions of expertise. In *Expertise in context* (pp. 149-162). Cambridge, MA: MIT Press.

Thoonen, W. (2006). *Expertise in Werkgroepen*. Master's thesis, Tilburg University, Tilburg.

TREC (2005). TREC Enterprise Track. Retrieved October 2005 from http://www.ins.cwi.nl/projects/trec-ent/wiki/index.php/Main Page.

Van Rijsbergen, C. (1979). *Information retrieval*, 2nd ed. Retrieved from http://www.dcs.gla.ac.uk/Keith/Preface.html.

Zheng, Z., and Srihari, R. (2003). Optimally combining positive and negative features for text categorization. *ICML 2003 Workshop for Learning from Imbalanced Datasets II*.

ISSN: 2324-9773