Diagnosis at Scale: Detecting the Expertise Level and Knowledge States of Lifelong Professional Learners

Oluwabukola Mayowa Ishola, Gord McCalla
Department of Computer Science
University of Saskatchewan, Saskatoon, Canada
bukola.ishola@usask.ca , mccalla@cs.usask.ca

ABSTRACT
Our research is about identifying gaps in the knowledge of professional software developers, as part of an ongoing project to provide tools to support their lifelong learning needs. We developed metrics that when applied to programmers' online activities in Stack Overflow allowed us to determine the knowledge states of users on specific topics indicating what each user knows they know and their knowledge "gaps", both what they know they don't know and what they don't know they don't know. Further we were able to find patterns that showed that at all levels of expertise there are still "unknown unknowns", and these are particularly dangerous since the software professional is unaware of their weaknesses in these areas.

KEYWORDS: knowledge states, diagnosis, lifelong learning

1. INTRODUCTION
Advanced learning technology research has begun to take on a complex challenge: supporting lifelong learning [1]. Professional learning is an important subset of lifelong learning that is (at least somewhat) more tractable than the full lifelong learning challenge. Professional lifelong learning is an ever more critical issue as the rate at which knowledge is generated in almost every professional discipline continues to accelerate.. Of course, professionals will evolve and develop their skills in the day-to-day practice of their profession, but workplace skills are not exactly the same as professional development because these skills are specific to their job role or even their particular workplace [2]. Professionals can be so overwhelmed with work responsibilities that they are ignorant of important new knowledge that exists.

Our goal in this research is to be able to diagnose the expertise of software professionals. We turned to a categorization of knowledge made by several different people [3,4] In this categorization knowledge can be divided into 4 knowledge states: the things we know we know, the "known knowns" (KK); the things we know we don't know, the "known unknowns" (KU); the things we are not aware we know but we do know, the "unknown knowns" (UK); and, lastly, the things we don't know we don't know, the "unknown unknowns" (UU) [4]. The known unknowns and the unknown unknowns we collectively call the "gaps" in a person's knowledge, and the most worrisome of these are the unknown unknowns, since a person is ignorant of their own ignorance.

2. DIAGNOSIS OF EXPERTISE
The experimental test bed for our research is the well known online programmers' forum called Stack Overflow (SO). We wanted to look for patterns in SO posts that allowed us to diagnose the *expertise* of the SO users. Posts were grouped under their related tags and tags were mapped into appropriate knowledge areas as represented by leaf nodes in the hierarchy shown in figure 1 (akin to that in [8]). Only 5 tags were used in this study: java, python, cplusplus, mysql, and sql. The restriction on choice of tags used in modelling knowledge was employed so as to have enough data about each category. Although, just 5 tags were considered, the total number of posts under each of these tags was large, ranging from a low of 238,487 posts regarding SQL to a high of 708,533 posts regarding java.

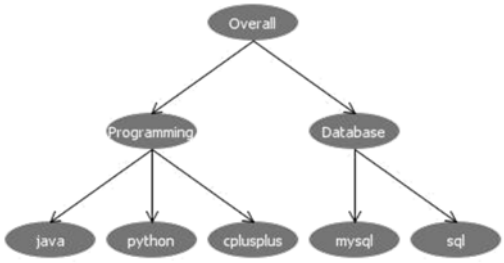

Figure 1. Hierarchical Structure Model Employed In Diagnosis of Expertise

We then determined for each user their expertise level in each of the 5 leaf areas, based on their SO reputation scores in the area. In SO the reputation data fits a power law in which the majority of users have a low reputation score; the higher the score the fewer the number of users. Using the method explored by Jiang [5], we fit a power curve to the actual SO reputation data for each area and computed Xmin and α, where Xmin represents the point where the exponential behavior begins in the dataset and α is the exponential factor. Users below Xmin were considered to be *beginners*. We then divided the remainder of the users into two equal sized chunks, the *intermediate* and *expert* users. Having diagnosed the expertise level of each user in each area, we then inferred their expertise level at the higher level nodes. In making this inference, we took the highest level of expertise of the user on the leaf nodes beneath a non-leaf node and assigned this level to the non-leaf node (recognizing that high expertise in one sub-area transfers to the more generic category, even in the absence of direct evidence). This was done recursively, up the hierarchy.

3. DIAGNOSIS OF KNOWLEDGE STATES
Next we wanted to diagnose the *knowledge states* of each user. Again, we considered only the 5 basic knowledge areas. The "known knowns" were determined by looking at the distinct answers the user has given under each tag that were up-voted. The "known unknowns" were determined by looking at the tags of questions the user has asked. The "unknown unknowns" were determined by looking at the tags of questions that the user has answered where the answer was down voted. At this stage in this work, no metric has been defined for the "unknown knowns"; i.e. the things the user knows but is not aware that they know. To determine the knowledge state of each user on each of the 5 topics represented by the tags we simply count the number of KK, KU, and UU posts for a given tag for a given user and determine the

relative percentage of each. The highest percentage exhibited by the user is diagnosed to be their knowledge state for the topic represented by that specific tag. For instance, a user whose evidence of KK for java is 70%, KU for java is 20% and UU for java is 10%, will be determined to know java, i.e. java is a known known. This process is carried out for all 5 tags, to determine the knowledge state a user exhibited for the topic represented by that tag.

4. RESULTS

In analyzing the data, we computed the average percentage of KK, KU and UU for users in various expertise classes for each of the knowledge areas represented by the 5 tags. For example, considering all users who posted in java whose competency level is '*beginner*', the average percentage for the KK, KU and UU was computed. Aggregate results from the 5 knowledge areas (for all 3 expertise levels) is represented in figure 2 below.

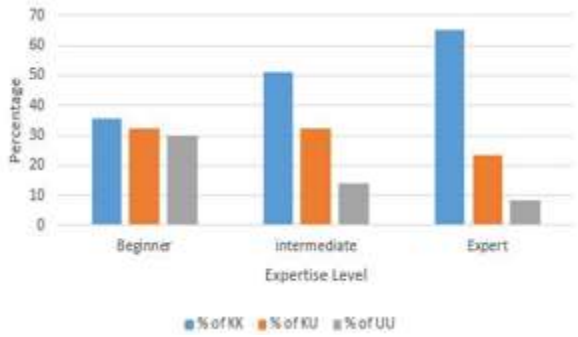

Figure 2. Aggregate Distribution over All Knowledge Areas

Figure 2 shows that as a professional's competency level increases, the proportion of their knowledge that consists of known knowns also increases. This is true for all 5 knowledge areas. This is reasonable, since presumably one measure of a professional's growing capability is that they come to know more (and that they know they know more). Similarly, across all 5 knowledge areas, the proportion of unknown unknowns steadily declines as expertise increases. The overall trend seems to be that the known unknowns continue to constitute about the same proportion of their knowledge when they are of intermediate capability as when they are beginners. Since their known knowns are a higher proportion of their knowledge than when they were beginners, this suggests that at the intermediate stage professionals not only come to know more, but also come to know more about what they don't know. Reassuringly, across all knowledge areas, the proportion of known unknowns decline as a professional of intermediate capability becomes an expert. Again, this suggests that the professional has growing expertise and has acted to reduce his or her known weaknesses. Perhaps the most interesting overall lesson from this analysis is that experts still have a considerable residue of unknown unknowns. The expert himself or herself may indeed find it difficult to believe that the knowledge they have learned and practiced for years is not as comprehensive as they thought. This suggests the need for tools that will enhance the self-awareness of professionals about their knowledge states, especially their unknown unknowns.

5. DISCUSSION

The competency of professionals has been determined in the past mainly by tracking their job performance [6]. This is not sufficient to judge their overall competence in their profession since the job (and the workplace) will likely require only a subset of the skills they need to be fully capable professionals. Moreover, Ley and Kump [7] argued that tasks performed alone is a weak measure in accessing competency of professionals, as competency will at most be judged in comparison to fellow workers rather than with professionals in society at large.

Working in the professional programming domain, our study goes beyond these limitations in several ways. First, we define competence in terms of knowledge states with a particular focus on what is known and unknown to the professional. Further, rather than restricting ourselves to examining job performance for evidence of capability, we look at the actual social interactions of professional programmers as they seek and receive help in a professional forum. Competency is judged in the context of other professionals who are mostly outside their own work places. Our approach also scales to a large number of users (we had access to the data of 888,603 active professionals). The approach also scales temporally: as a discipline evolves new knowledge over time that knowledge will automatically filter into professional interactions, and thus the knowledge states of users on this new knowledge can be readily diagnosed (assuming that the ontology and tag-to-ontology mappings are updated).

To be sure there is much more to be done. We need to confirm the results of this first experiment with further evidence that our diagnoses are accurate. We need to explore other competency and performance metrics that can be mined from SO data. We need to create more refined ontologies that we hope will allow tracking knowledge at a finer grain size. And, ultimately, we wish to create an open user modeling system that can reflect the diagnoses back to the professional user. We believe this approach to "diagnosis at scale" has a promising future in supporting the lifelong learning.needs of professionals.

6. ACKNOWLEDGEMENTS

Thanks to the Natural Sciences and Engineering Research Council of Canada and the U of Saskatchewan for funding this research.

7. REFERENCES

[1] Kay, J., & Kummerfield, B. (2009). Lifelong user modelling goals, issues and challenges. In Proceedings of the Lifelong User Modelling Workshop at UMAP-2009, pp. 27-34).

[2] Bruce, C. S. (1999). Workplace experiences of information literacy. International journal of information management, 19(1), 33-47.

[3] Dunning, D. (2011). 5 The Dunning-Kruger Effect: On Being Ignorant of One's Own Ignorance. Advances in experimental social psychology, 44, 247.

[4] Rumsfeld, D. (2011). Known and unknown: a memoir. Penguin.

[5] Jiang, B. (2013). Head/tail breaks: A new classification scheme for data with a heavy-tailed distribution. The Professional Geographer, 65(3), 482-494.

[6] Ley, T., Ulbrich, A., Scheir, P., Lindstaedt, S. N., Kump, B., & Albert, D. (2008). Modeling competencies for supporting work-integrated learning in knowledge work. Journal of Knowledge Management, 12(6), 31-47.

[7] Ley, T., & Kump, B. (2013). Which User Interactions Predict Levels of Expertise in Work-Integrated Learning. In *Scaling up Learning for Sustained Impact* (pp. 178-190). Springer Berlin Heidelberg.

[8] Ishola, O. M., Shoewu, O., & Olatinwo, S. O. (2013). A Conceptual Design of Analytical Hierarchical Process Model to the Boko Haram Crisis in Nigeria. In Information and Knowledge Management (Vol. 3, No. 3, pp. 1-19).