LEARNING BASED APPROACH FOR PERSONALIZED EXPERT DETECTION

A Thesis

by

SANGHITA BANDYOPADHYAY

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Richard Furuta
Co-Chair of Committee,	Frank M. Shipman, III
Committee Member,	Luis Filipe Viera de Castro
Head of Department,	Dilma Da Silva

May 2016

Major Subject: Computer Engineering

ABSTRACT

In recent years, identifying experts has gained significant attention in the research area. The main motivation behind it is to facilitate the process of locating the correct individual capable of answering our queries. There has been a lot of focus on building expert recommendation systems. The main focus of these systems is to effectively build an expert profile in order to facilitate recognition. We argue that definition of an expert is a very subjective term and it has a major dependency on the individual initiating the search. There has also been a lot of research on personalizing search results. The two main methods applied in the design of these techniques are (1) Using explicit feedback (ratings etc.) (2) Using implicit feedback (mouse movements etc.). We propose TAK, a learning-based framework for accurate retrieval of experts based on tacit knowledge of the user placing the request. We focus on defining the tacit knowledge of the user based on implicit features like experience and education to deduce the preference of the user and generate more specific and targeted suggestions. The increasing usage of social media for everyday communication has made it a suitable repository of user specific information. Thus, we base our study on *LinkedIn*, which is a social media application pervasively being used for exchanging information and locating qualified individuals. We use crowd preference knowledge to create a learning-based framework and augment the result with the expert profile created from *LinkedIn* to provide expert recommendations to the user. This enables the user to make an informed decision. A comparative analysis of the results of the proposed method to the method applied by LinkedIn proves that the former provides more popular suggestions to the latter. It further proves that cultivated tacit knowledge with years of experience has an impact on expert selection decision.

TABLE OF CONTENTS

Page

Page

Page

Page

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

With shortening of deadlines and increased demand for on-time task completion, the search for experts has grown manifold. However, due to the humongous amount of information available on the web, the task of locating an expert has become even more complicated. A simple query to www.google.com returns results in the range of a few trillions. Due to the huge amount of information, we tend to miss out on the best source and select the one we view first. It is the same case while looking for an expert. Thus an expert recommendation system which can accurately lead the users to the best experts is very critical.

There has been a lot of research in the field of expert recommendation systems. One of the earliest works on expert recommendation was based on determining topical similarity via semantic analysis of user data [1] and rank them with respect to the query. In order to view experts within a question and answering forum (line AskMe.com),[2] proposed an algorithm to infer expertise by analyzing relevant posts. Social activity is another measure quite frequently used to determine expertise. [1] analyses the ratings given by users to specific items and finds users with similar choices. Consequently, they use unsupervised learning techniques on this data to analyze its characteristics and use it to provide item recommendations.

These techniques utilize the characteristics of the expert profile and score them based on their authority over the query topic. Thus every user will receive the same set of experts, irrespective of the background of the user. Also, it might prevent the user from making a well informed decision while selecting an expert, since he might not be capable of even judging the level of expertise.

Online shopping portals like Amazon.com have been performing personalized recommendations for a while now. They suggest items to users based on items they have purchased in the past. Recently there has been some focus on developing expert systems tailored to the needs of the user. For example, [3] proposed a method to infer user preferences by monitoring their interactions with the system and made

appropriate suggestions. A major concern while providing accurate suggestions is assessing the query submitted by the user. The Google search engine tries to solve this issue by customizing its search results based on the location of the system from which it is being accessed. [4] proposes a method to build user context in order to make sense of the query submitted by the user. The proposed method uses explicit information available via user interactions to capture the user context and make suggestions based on the identified context.

These methods assume every user to have similar responses under similar circumstances. They do not consider the influence of tacit factors on the decision of the user. We argue that a method that combines user profiling with expert recommendation would generate more accurate expert recommendations. [3] presents a method to create user profiles based on their navigation behavior. However, they fail to consider factors like age, gender and background while creating the profile. We argue that tacit features have a great influence over the decision of the user while looking for an expert. Factors like educational background and years of experience in industry will heavily bias the expert selection process.

For example, the criteria for choosing a *JAVA* expert for a person with 20 years of experience in *JAVA* will be completely different from that of a recent graduate. While the former would focus more on specific domain expertise, the later would be more interested in targeted achievements. There has been very little research on studying the influence of such factors over user decisions. Thus, this thesis focuses on analyzing the influence of these factors on users' decision in selecting an expert and develop a framework to test the observations so obtained.

The main intended contributions of this thesis are:

(1) An in-depth analysis of knowledge structure of a user profile in LinkedIn, i.e. the various aspects of knowledge which can be inferred from the information contained within a profile. In order to do the same, we collect user profile data from *LinkedIn* to study the structure of the data provided. To this we apply supervised learning

approaches to extract the most influential and non-overlapping features.

(2) Quantification of the tacit knowledge of the user and the degree of influence it has on the decisions of the user while selecting an expert. We use experience and education information of the user to represent the tacit knowledge of the user.

(3) Develop a learning based framework to predict the user preference and determine its degree of influence on the expert features identified from a user profile in LinkedIn. The features investigated are:

(i) Profile based features (i.e. the features which would be based on the content of the candidate profile)

(ii) Preference based features (i.e. these would capture the relative importance given by a user to different aspects of an expert while making a decision)

(iii) Skill based features (i.e. these would reflect the relevance of the information contained in the user profile with the query term)

Based on the above mentioned features, we propose to calculate an expertise score for each candidate expert, based on which experts will be recommended for the particular user.

To access the accuracy of the expertise scoring strategy based on predicted user preferences, we perform a user study. The results from the same are used to compare the accuracy of the suggestions generated by the proposed method to the method implemented by LinkedIn.

2 RELATED WORK

Expert recommendation has become a major research topic. Previous work can be classified into recognizing expertise, user profiling, personalized recommendation system and expert ranking methods.

2.1 Recognizing Expertise

Identifying people who would qualify as experts has been recognized as a major area of research for a long time. The approaches to recognize experts can be classified into three major categories; content based methods, link analysis methods and collaborative filtering-based approaches. Some of the initial works on analyzing link structure to infer expertise used HITS [5] and PageRank [6]. They constructed graphs based on topical similarity and deciphered expertise based on the analysis of the links. [7] discussed a tool *SmallBlue* which inferred expertise by creating topic specific links connecting people within an organization and inferred expertise based on the characteristics of the link. Expertise has also been inferred using co-authorship and citation information to build the adjacency graph of the users [8] and selecting the user with the highest in-degree as an expert. In the *content based approaches*, one of the earliest proposed methods [1] used description of the topic to find documents pertaining to the user from the web and used a scoring function to rank the experts according to the similarity of the documents to the query term. In recent times, [9] suggested a method to profile group members based on the query term and rank members according to a scoring function applied on the generated profile values. The *collaborative filtering-based approaches* predict the interests of a user from preferences recorded by other users. [8] suggested a method to search possible collaborators. They organized users into clusters based on information obtained from them as well as from the web and then with the help of a collaborative filtering algorithm, they located the nearest neighbor to suggest a possible collaborator. [2]

proposed a hybrid approach, which used both content based search and graph based analysis to locate an expert.

2.2 User Profiling

Before initiating the search for an expert, it is highly essential to properly understand the specific requirements of the concerned user. There has been a considerable amount of research in constructing user profiles to identify their preferences and provide personalized services. [10] used ratings explicitly given by the user along with the items with which the user has interacted, to build the profile. Users' browsing history has also been used with collaborative filtering to infer a user's preference [11].

2.3 Personalized Recommendation System

[12] proposed a personalized PageRank method to enable personalized Web searches. Haveliwala [13] calculated 'topic sensitive' scores to rank each result with respect to the search term. Chang et al. [14] calculated document authority with a method similar to Kleinberg's HITS algorithm [15]. [16] proposed a method which used semantic evidence found in the user profile as well as interest scores, to re-rank expert's search result.

2.4 Expert Ranking Methods

[17] used a linear combination of topical authority and local authority to calculate expertise scores. Another paper which proposed a model to determine community sensitive expertise [18], also suggested a linear combination of document-based model with community-sensitive author-rank, to establish expertise scores.

3 USERS AND TACIT KNOWLEDGE

In the previous section we have seen that there has been a lot of focus on developing personalized systems. However little has been done to observe the influence of the tacit dimension [19] on such systems. We hypothesize that the tacit knowledge of the user, gathered from years of experience, would influence his decision while judging expertise. The main focus of this study is to identify the tacit characteristics which influence an individual's decision while selecting an expert.

3.1 A Profile in LinkedIn

LinkedIn is one of the fastest growing professional network[20]. Thus we choose LinkedIn as the test bed to focus our study. We began our investigation by obtaining a complete understanding of a user profile in LinkedIn and the format in which information is presented within it. Given the fact that it is used pervasively for professional networking, we hypothesized it to be an ideal place to look for experts as well. We used a web crawler to collect 500 LinkedIn profiles. These profiles were obtained by crawling the contact list of an individual LinkedIn user in a single day. From an informal viewing of these profiles, we observed that each profile was divided into predefined sections where users could list their achievements and describe their background. Each of these sections represented a particular dimension of information about the user. These sections varied in terms of frequency of usage across profiles as shown in Table 1. For example, certain people used the education block extensively to list their academic achievements while others used the projects block to list their projects and never used the education block. A LinkedIn profile presents two types of information;

(1) Explicit Blocks: Information blocks like Education and Summary which directly contributed to the technical skills of the user.

Block Type	Block Name	Frequency
Explicit Block	headline	486
	summary	356
	experience	462
	recommendation	360
	certifications	153
	education	424
	skills	342
	additional information	146
	publications	315
	honors	336
	courses	323
	projects	323
	patents	315
	organizations	256
Implicit Block	languages	356
	volunteering	56
	test score	134
	number of connections	478
	contact information	352
	profile image	390

Table 1: Frequency of Information Blocks within LinkedIn

(2) Implicit Blocks: Information blocks like contact information, setting a profile image and providing contact information, from Table 1, which did not contribute directly to the technical skills of the user but had an implicit contribution to the overall

profile information. We created a list of all the available blocks of information and observed their respective frequency of usage. In order to obtain the frequency value, we utilized the pool of 500 LinkedIn profiles that had been collected earlier. By parsing each profile with respect to the explicit and implicit block boundaries, we observed certain frequencies as listed in Table 1. Based on the observed frequencies, we choose to further investigate the textual content within these blocks with a frequency of usage greater than 70%. The observations will be discussed next.

3.1.1 Summary

This section captured the main essence of the user's personality. Given that, it is the first piece of information that any person reads about the user, the textual content used in it indicated how the user would want others to view his profile. Some users preferred to write a line or two about their domain skills while others preferred to write a long list of all of their domain related skills. We observed that the length of the summary did not provide a good measure to judge the contribution of the content to the user profile. Rather, an analysis of the words used in this block gave a better idea about the level of confidence of the user in the domain specified by him.

3.1.2 Experience

LinkedIn provides a predefined format in which users can describe their work profile. We observed that every detail added to this section has three components: (1) Position of Responsibility; (2) Duration; (3) Description of work. Each of these components had a contribution of their own. A combination of these components contributed to knowledge about the general technical background of the individual. Here, the textual content of the third component, i.e. description of work, displayed a level of awareness of the user about his mentioned skill set. For example, if a person had mentioned AJAX in his skill set but did not mention it in the description of his work, it indicated that either the person had never worked in that domain or he was

not confident enough to mention it as a part of his work history.

3.1.3 Education

For this section, LinkedIn provides a predefined format as well. It presents the highest technical degree achieved by the user. In certain cases, we observed that the order in which the degrees were mentioned reflected self-preference, i.e. a user who would be more interested to work in human resource management would mention the relevant degree above all other irrelevant degrees. However, in most of the profiles, it gave a general idea about the educational background of the user.

3.1.4 Project

This section reflected work-related milestones achieved by the user as well as a degree of confidence in the domain. The description (if provided) reflected a willingness of the individual to work in similar domain. Thus, overall it gave an idea about the intra-domain confidence of the user.

3.1.5 Courses

This section provided information about all courses completed by the user. People choose to mention either coursework or courses done apart from requisite subjects. A mention of the courses again indicated willingness of the individual to further work in the related domain. Thus it indicated a degree of domain confidence.

3.1.6 Publication

The textual content of this section indicated a degree of self-interest on the part of the user within the domain mentioned. The general idea behind a successful publication reflected interest of the individual to work on an idea within the field.

3.1.7 Scores

This section provided test scores of the user. It was used either to mention CGPA or score obtained from other competitive tests. Inclusion of this field in the user profile indicated a level of authority and confidence of the user as well as a degree of interest in the mentioned field.

3.1.8 Certifications

Certifications imply additional skills acquired by the user apart from regular course work. Thus we hypothesized that this might also indicate self-interest. But more than interest, it indicated a degree of effort invested by the user. From the user profiles, we observed that it indicated the extra mile tread by the user to acquire skills which in turn would make him better in his domain of interest.

3.1.9 Honors

This section specifically quantified the extra mile traveled by the user to register an achievement in comparison to his peers. A mention of it indicated their proficiency in the respective field in comparison to their cohorts.

In most of the observed profiles, the cumulative information provided by the text extracted from the information blocks mentioned in Table 2, defined a LinkedIn profile and helped a user to judge expertise. As per our earlier argument, the tacit knowledge of the user might influence the relative importance assigned by the user to these blocks of information individually.

3.2 How Users Rank Experts

How does LinkedIn rank search results? Having explored the profile structure within LinkedIn and given the pervasive usage of LinkedIn among software professionals, we proceeded to look for experts within LinkedIn and investigate its search

algorithm. As a preliminary step, we used the search URL shown in Figure 1 provided by LinkedIn to look for experts in JAVA. The URL presented us with 1,000,000 profiles, matched for the search term JAVA. On further investigating the profiles, we found that the search results were ranked based on their proximity with the current user and frequency of usage of the query term within the respective textual content. Do people accept the ranking given by LinkedIn to a set of experts? In order to

Category
Summary
Experience
Education
Project
Courses
Publication
Scores
Certifications
Honors

Table 2: Information Blocks within LinkedIn

find an answer to this, we tried to gather crowd-opinion for judging a set of candidate experts by questioning avid users of LinkedIn. We created a pool of candidate experts using the LinkedIn search URL shown in Figure 1 and accumulated additional information corresponding to each candidate, by mining the textual content contained in their respective profiles.

On initial attempts to collect data from LinkedIn we found that LinkedIn imposed a restriction on the total number of profiles that a user can access in a single day as

well as on the total number of results of a search query that could be viewed overall. We overcame this difficulty by utilizing the increased search limit feature provided by LinkedIn. A total of 1000 LinkedIn profiles were accumulated over a period of 7 days. With this pool of expert candidates ranked in the order as provided by LinkedIn, we proceeded to separate them into sample subsets. To gain insight into crowd perception, we interacted with a few people and registered their opinion on the ranking approach followed by LinkedIn within these subsets of candidate experts.

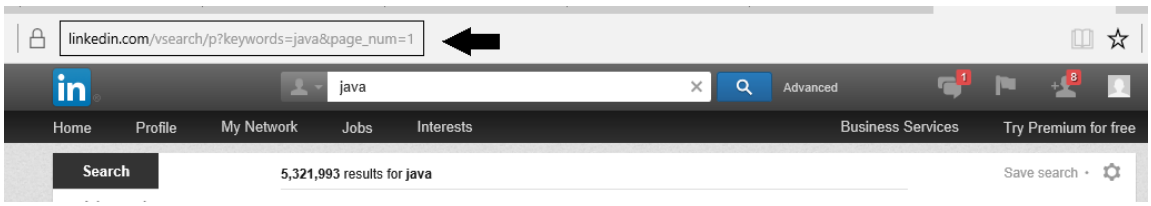


Figure 1: LinkedIn Search URL

Based on earlier work[21], we considered educational background and professional experience to play a powerful role in shaping the tacit knowledge of a user. Thus with reference to our earlier argument, that tacit knowledge of a user can influence his expert selection decision, we considered all the opinions with respect to the academic and professional background of the people we interacted with.

Our aim was to get an agreement from people on the rank given by LinkedIn. We expected most of the users to be in agreement with the rank provided by the

LinkedIn search API. However, we observed substantial digression during the course of interaction. People provided varied opinion on the same set of expert profiles thus proving that frequency of usage of terms was an insufficient condition to decide expert selection criteria. Since few of them considered the rank given by LinkedIn to be correct, we concluded that frequency of usage of terms was one of the criteria considered while judging expertise. Based on our earlier argument that tacit knowledge of the user influenced their expert selection criteria, we proceeded to observe the responses gathered with respect to their academic and professional background. People with similar background had similar opinion about the candidate expert sets. Thus we concluded that users having similar background, had a higher probability of judging experts similarly. This proved that academic and professional background of the user influenced their expert selection criteria.

3.3 Power of Tacit Knowledge

What is Tacit Knowledge? Knowledge which is used by people without being aware of [22] is Tacit Knowledge. Tacit knowledge, as described in [23], is difficult to explain and is gathered through life experiences and impressions. It is not only difficult to explain but also rarely changes with time. It is ingrained within an individuals' sub-consciousness and is used by the user without being aware of it. As defined by Michael Polanyi [19], tacit knowledge is a knowledge that is; *implied or indicated but not actually expressed*, thus gathered from previous experiences and encounters. [21] states that tacit knowledge comprises of experience, thinking, competence, commitment and deed. In our previous observations, we found academic background and professional experience as indicators of tacit knowledge of the user.

How does tacit knowledge influence people's decision? From the above discussed interactions, we observed that academic background and professional experience influenced the expert selection criteria of a user and resulted in a similar ranking pattern. Thus, we proceeded to infer its degree of influence by exploring a solution

to another question, i.e. can we quantify the influence of tacit knowledge? Polanyi argued that informed guesses, hunches and imaginings that are part of exploratory acts are motivated by what he described as 'passions'. According to him, such knowledge might well be aimed at discovering 'truth', but is not necessarily in a form that can be stated in propositional or formal terms. From this statement, we inferred that defining tacit knowledge in formal terms was a difficult task. However, Polanyi proceeded to argue that, tacit knowledge can be reduced to explicit knowledge which is a quantifiable factor[21]. Such explicit knowledge includes documents, record, files and data. Thus, by hypothesizing on the same lines, we planned to base our observation of tacit knowledge by collecting people's opinion on the same. Since frequency of usage of words was observed to be one of the criteria preferred by people to rank experts, we proceeded to identify user's preference towards the textual content within the various blocks of information present in a LinkedIn profile. We limited our discussion to a few specific features extracted from the predefined information blocks from Table 3 within a LinkedIn profile. By conducting informal discussions, we gathered the preference of a few LinkedIn users, about the top five characteristics which they considered to be most important while judging an expert. We asked their opinion about expertise of the candidate profile by observing individual information block as well as the entire profile as a whole. Our aim was to observe any discrepancies created as a result of increased focus on particular information blocks. In several cases, the overall rank was similar to the rank given with respect to certain information blocks from Table 3 whereas in others they did not match. This indicated that in the former situation, the matching information blocks had a positive affect on the rank of an expert while in the later case the information blocks were ignored. One of the users focused more on the content mentioned in the education block and their overall rank was same as the rank given by observing the education block. The same person showed minimal interest in the content of the project block and hence the rank with respect to number of projects differed from the overall rank. People with

similar academic and professional background were observed to have similar opinion while ranking experts. In the next section we present a detailed description of the recorded observations from the pilot studies.

Category
Relevant Skills
Technical Positions
Area of Experience
Years of Experience
Highest Technical Degree Attained
Length of Summary
Number of Certification
Number of Honors
Quality of Text in Explanation
Relevant Courses taken
Number of Projects
Number of Publications
Number of Organizations
Number of Patents

Table 3: Features within Information Blocks

4 QUANTIFYING CROWD KNOWLEDGE

In the previous section we observed that the academic and professional background of the users lead them to associate varying importance to various information blocks within a LinkedIn profile. In this section we present the recorded observations from the pilot studies and provide our interpretations. Based on prior studies[21], we hypothesized that educational background and professional experience might play a powerful role in shaping the tacit knowledge of a user. In the pilot studies mentioned in the previous section, we observed that tacit knowledge resulted in similar ranking pattern among users as well as similar importance to various information blocks.

To further analyze the observed patterns, we defined the profile for each user comprising of three features from Table 4 and defined it as the tacit knowledge vector. We clustered the set of users into separate sets based on their educational background from Table 6 and professional experience from Table 5. We did not use additional information data to cluster the users, since the data obtained in it was not consistent enough to be used as a measure to cluster profiles. Some users had listed their academic achievements in this section and some had listed extra-curricular achievements. While academic achievements contributed in determining expertise of a user, extra-curricular achievements did not contribute in judging technical expertise. Our intention behind collecting data corresponding to additional information was to judge users based on their motivation to invest extra effort and perform better. However, the data presented by the participants digressed from the technical context and hence could not be used. We proceeded by observing the influence of academic and profession background on the relative importance of information block features from Table 3 while judging an expert.

4.1 Educational Background

We analyzed the effect of educational background on the preference towards the individual features identified from a profile in LinkedIn as seen in Table 3. We

observed that users with similar education background displayed similar preferences towards the information content of an expert profile. Educational background was defined based on the highest level of education completed and we created three clusters based on this score as presented in Table 7. Similarity was determined based on the ranking assigned by them to the individual information blocks while rating an expert in LinkedIn. We grouped together information blocks that displayed similar variance across clusters.

4.1.1 Area of Experience, Technical Degree, Years of Experience

Users gave highest preference to these three features, i.e. 3 out of 5 as shown in Figure 2. The overall preference given to each of these features was approximately the same as well. We observed a similar pattern of preference towards these three features. The preference towards each of them reduced with increase in educational background score. By analyzing the text presented corresponding to each of these features (Experience, Education), we were able to infer the professional and academic background of the user. The content related to area of experience was presented in the Experience block and it included the description corresponding to every professional experience of the user in terms of the technical positions held. This section also had the time span for which each of these technical positions were held, which presented the sum total of the years that the user had been in the professional world. Thus, the combined content from these two blocks gave an overview of the professional journey of the user. Similarly, the highest technical degree achieved was mentioned in the Education block and the content gave an idea about the academic journey of the user. A combination of these three features presented the background knowledge of the user. Users in cluster 1 gave greater importance to background while judging an expert and there was a gradual decrease in preference towards background from cluster 2 to 3. Users in cluster 3, gave the least importance to background. Thus, as people gathered a richer professional and educational background they associated

Feature
Experience
Education
Additional Information

Table 4: Tacit Knowledge Vector

Years of Experience	Score
Less than 1	1
1 - 3 years	2
3 - 5 years	3
5 - 10 years	4
More than 10 years	5

Table 5: Scoring System for Years of Experience

Degree Achieved	Score
Bachelor	1
Master	2
Doctorate	3
Post-Doctorate	4

Table 6: Scoring System for Highest Technical Degree Achieved

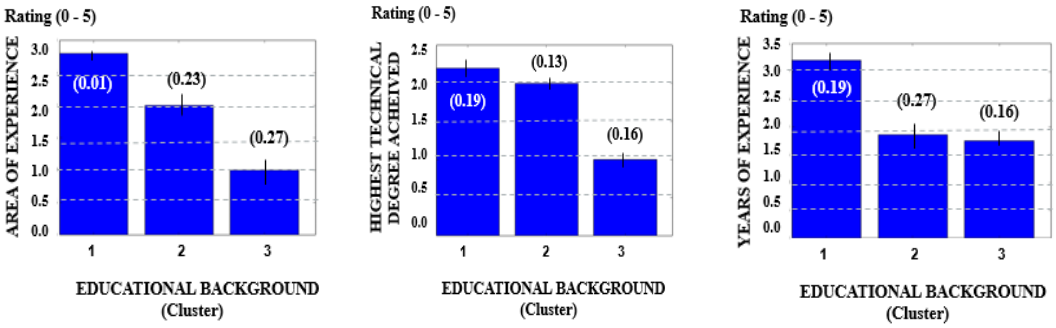
Value	Score	Cluster
Bachelor	1	1
Master	2	2
PhD	3	3
Post-Doctorate	4	3

Table 7: Clustering based on Educational Background

less importance to background of the candidate in determining expertise. The preference towards technical degree reduced considerably from cluster 2 to cluster 3. People in cluster 3, possessed a higher technical degree in comparison to people in cluster 2. Thus, our initial assumption that people in cluster 3 would give greater importance to technical degree was proved wrong. The preference towards years of experience also decreased considerably from cluster 1 to cluster 2. This showed that as people proceeded in academic field, due to investment of more time in the educational field, they considered years of experience in professional field an unimportant factor while judging expertise.

4.1.2 Technical Positions

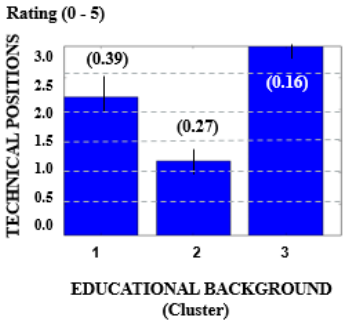
Even though the average preference observed for technical positions was same as area of experience, the pattern of variance was quite different as shown in Figure 3. Users in cluster 1 gave quite high preference to it, which was similar to preference for area of experience. However, there was a substantial increase in preference for technical positions from cluster 2 to cluster 3 unlike area of experience. The information corresponding to technical positions was obtained from the experience section. From an analysis of the content, we concluded that apart from conveying background



(a) Area of Experience (b) Highest Technical Degree Achieved (c) Years of Experience

Figure 2: Area of Experience, Technical Degree, Years of Experience

strength of the user, it also represented a degree of stability of the user. For example, a user with high degree of stability would have 10 years of work experience and 6 positions, but a user with low degree of stability would have 15 positions in 10 years of work experience. Users in cluster 1 preferred an expert who had listed a greater number of technical positions but users in cluster 2 displayed contrasting response. This showed that number of jobs in professional world translated into more knowledge for users with less academic background. Also, users in cluster 2 had the highest probability of being influenced by instability. According to an article published by the Center for Economic and Policy Research, 'How Much Does Employee Turnover Really Cost Your Business?' [24], tenure was the least for the age group 25–29 years. As per our observation, people in cluster 2 would belong to the age group of 25–29 years and thus be influenced by attrition the most. The low preference assigned by these people indicates that instability is considered as a negative trait in experts. As the level of education increased, users gave more importance to positions. This showed that, with increase in academic background people preferred stability and rewarded people with high degree of stability while penalizing people who displayed more instability.

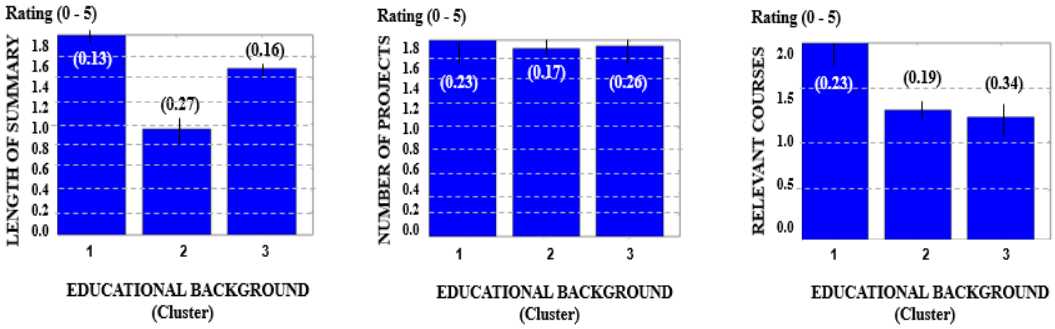


(a) Technical Positions

Figure 3: Technical Positions

4.1.3 Number of Projects, Length of Summary, Relevant Courses

The average score of the three mentioned features was approximately the same i.e. 1.8 as shown in Figure 4. The observed pattern was quite similar for variation of preferences among clusters. The content presented in summary showed the capability of the user to present his knowledge in a comprehensible manner. A concise and to-the-point summary implied a clear knowledge of the domain whereas a summary with rare mention of the skills showed ambiguity and low confidence on the part of the user. Well written project descriptions added value to the clarity of knowledge as well as presented a degree of confidence. Mention of coursework provided proof of knowledge of the user. Thus a combination of these features, presented the level of domain knowledge of the user. From our observation, users in cluster 1 displayed highest preference towards domain knowledge of the expert for judging expertise. Irrespective of the professional experience of the user, their academic background was the lowest, which implied that they had experienced the professional world



(a) Length of Summary (b) Number of Projects (c) Relevant Courses

Figure 4: Number of Projects, Length of Summary, Relevant Courses

as a beginner without having a substantial amount of credibility and would have faced the challenge of establishing their credibility with minimal knowledge of the domain. It was inferred that as a result of facing such a challenge, they valued domain knowledge. People in cluster 2 gave relatively less importance to domain knowledge. These were the individuals who had viewed the professional world with a certain degree of credibility without substantial domain knowledge. Thus, the challenge to establish professional credibility, due to insufficient academic background, might have increased the value for domain knowledge and a higher academic background warranted more credibility in the professional world.

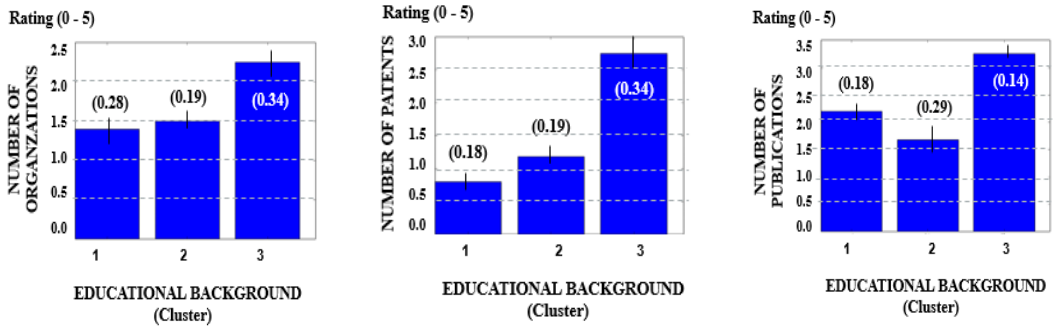
4.1.4 Number of Patents, Publication and Organizations

These features received a high average preference of 3 out of 5 and showed a similar pattern of variance across clusters as shown in Figure 5. The preference for these features gradually increased with increase in quality of academic background. The mention of patents in a user's profile, displayed a sense of willingness to contribute to

the community by investing additional effort and time. The textual content within organizations presented the avant-gard efforts of the user. Similarly, publications also showed extra effort and willingness to work hard. A combination of these features, conveyed the willingness of the user to invest extra time and resources to make things better. The relative preference for this block of information increased significantly with increase in quality of education. With increase in academic background the effort invested to gather better knowledge also increased. As the academic knowledge of the user increased, they associated more importance to investment of additional effort in judging expertise. Thus, we inferred that users who would themselves be investing extra effort to improve their knowledge, would associate more importance to it while judging expertise. However, even though the sense conveyed by the section of publications was similar to the other two, it had a slightly different degree of preference among users in cluster 1 and 2, i.e. there was a decrease in preference towards additional effort to judge expertise. This showed that at earlier phases of academic education, users focused more on publications than patents or organizations.

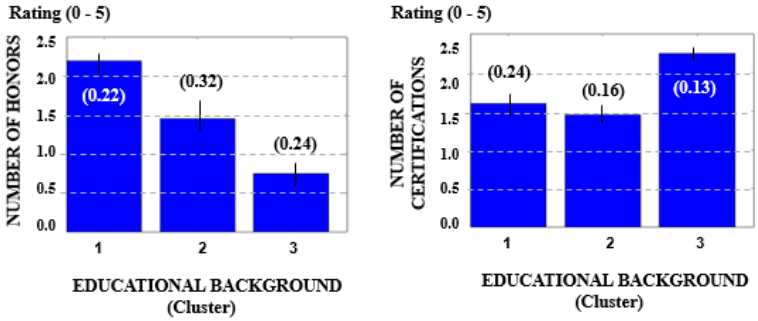
4.1.5 Number of Certification, Number of Honors

Though the pattern of variance observed in these features was inversely related to one another, the average preference to these features was the same, i.e. 2.5 out of 5 as shown in Figure 6. While people in cluster 1 gave highest importance to honors, they preferred certifications the least. Certifications are basic qualifying exams required to demonstrate solid understanding of the respective domain. They are usually not a mandatory requirement in the professional world and thus its presence shows a degree of self-interest of the user to prove the worth of his knowledge. Similarly, the text pertaining to honors section presented the demonstrated ability of the user to perform better among peers. Thus a combination of these features shows self-interest of the user. From our observation we inferred that, people with relatively strong academic background preferred certification the most which demanded a higher quality of domain knowledge. Since, honors did not demand a higher quality of considerably



(a) Number of Organizations (b) Number of Patents (c) Number of Publications

Figure 5: Number of Patents, Number of Publication, Number of Organizations



(a) Number of Honors (b) Number of Certification

Figure 6: Number of Certification, Number of Honors

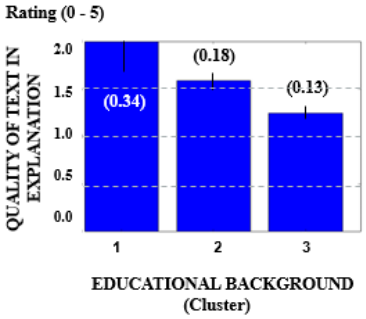
higher than the rest,domain knowledge and reflected the potential of the user in comparison to his peers, it was highly valued by people with lesser academic background. We also note that with increase in quality of education, the importance of comparison with peers decreased.

4.1.6 Quality of Text in Explanation

We had included this field to get the response of people on how much value they would associate with the quality of the text in the profile. Quality of the text implied inclusion of more technically relevant terms without being descriptive. The users in cluster 1 associated maximum importance to this feature and its importance diminished with increase in level of education as shown in Figure 7. This observation was in contrast to the observation for length of summary, which had an increase in importance with increase in degree of education. Thus, we concluded that this would have resulted due to error in judgment on the part of the user, i.e. participants would have considered quality as total information contained in the section rather than actual technical terms. However, we decided to study the observations for further investigation and found that people in cluster 1 gave maximum importance to explanation quality and people in cluster 3 gave least importance to the same. Thus, by taking into account the error in judgment, we inferred that with increase in educational background, the importance for quality of text in description increased.

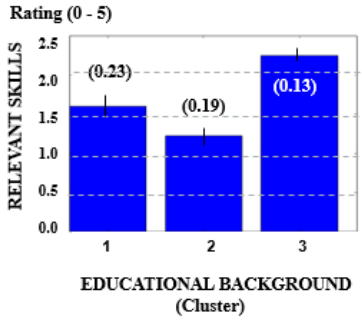
4.1.7 Relevant Skills

People in cluster 3 gave maximum importance to this section and people in cluster 2 gave least importance to it when they were asked to judge proficiency of experts as shown in Figure 8. The information presented in this section included a precise list of domain skills of the user who was being judged for his proficiency. Based on the



(a) Quality of Text in Explanation

Figure 7: Quality of Text in Explanation



(a) Relevant Skills

Figure 8: Relevant Skills

observation we could explain that, people in cluster 1 preferred having plethora of knowledge in various skills instead of limiting themselves to a particular domain. Thus they associated a greater importance to skills. In contrast, people in cluster 2 preferred to remain focused on a particular domain and thus associated relatively less importance to a variety of skills. With further increase in quality of academic background, people associated usefulness with number of skills at disposal. Thus, the importance of skills increased again.

4.2 Professional Background

Value	Score	Cluster
Less than 1 year	1	1
1 to 3 years	2	2
3 to 5 years	3	3
5 to 10 years	4	4
More than 10 years	5	5

Table 8: Clustering based on Professional Background

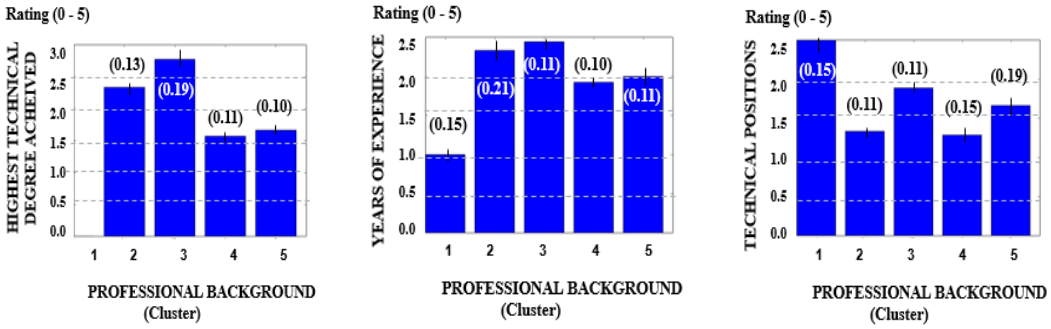
Participants with similar professional background had similar preferences for judging an expert. We defined professional background according to the number of years of experience in the professional world and created five clusters based on this score presented in Table 8. The similarity was determined based on the ranking assigned by them to the 10 information blocks while rating an expert in LinkedIn. Based on the results of the survey, we made several observations which have been elaborated in the section below.

4.2.1 Technical Degree, Years of Experience, Technical Positions

The observed pattern of variance for these three features was approximately constant across the experience clusters and the average preference was 2.5 out of 5 as shown in Figure 9. As mentioned previously, these three features presented the academic and professional background of the user. We observed that the preference to background of expert, followed a Gaussian curve with the peak achieved for people in cluster 3. These were the individuals who had been in the professional world for 3 to 5 years. According to a report from the Council of Graduate Schools, the largest percentage of enrolled students in fall 2007 were between the ages of 25 and 29, which was also the case 10 and 20 years earlier[25]. Given that 21-24 years is the usual age at which people get their undergraduate degree, people in cluster 3 would fall in the age group of 25-29 years. Thus, from this statistic it was clear that, people in cluster 3 would have the highest motivation to focus on improving quality of academic background as compared to others. Thus they had a higher preference to the same as a measure to judge expertise. The only exception to this observation was for cluster 1 in case of technical positions. In contrast to the other observations, the people in this cluster displayed an unusually high preference towards background knowledge. We attributed this to the lack of experience within the professional world, which would lead to perception of technical positions as a mark of achievement.

4.2.2 Area of Experience

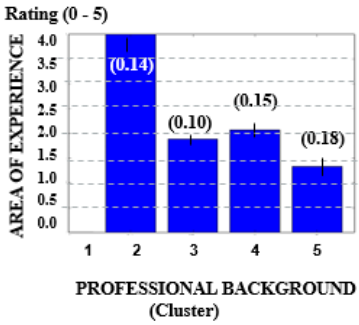
Though the information conveyed corresponding to this feature represented the background of the expert, the pattern of variance was quite dissimilar to the previous features representing the same (Highest Technical Degree Achieved,Years of Experience, Technical positions). Also, the average preference for this feature was considerably higher than the rest, i.e. 4 out of 5 as shown in Figure 10. The content corresponding to this feature showed the professional background of the expert.



(a) Highest Technical Degree Achieved (b) Years of Experience (c) Technical Positions

Figure 9: Technical Degree, Years of Experience, Technical Positions

Along-with this, it also gave a sense of the variety of domains the user had worked on, i.e. it added a dimension of versatility to the professional background of the user. Being new entrants into the professional world, they would consider versatility as a mark of instability leading to less importance to this feature. However, with a basic experience of the professional world and having been presented with the opportunity to work in a plethora of domains, they would opine versatility as a mark of an expert. With further time spent in the professional world and increase in focus on a particular domain, there would be a gradual reduction in importance towards versatility to judge an expert. An unusual observation to this hypothesis was the high preference given by people in cluster 3. There was an observed increase in value for versatility. This can be attributed to the fact that being inclined towards increasing radius of academic and professional domain, they would prefer background with versatility as the mark of an expert.

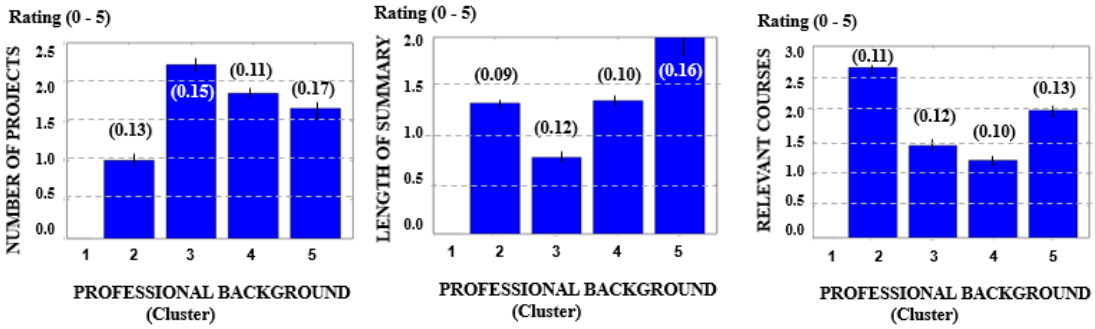


(a) Area of Experience

Figure 10: Area of Experience

4.2.3 Number of Projects, Length of Summary, Relevant Courses

The average preference to these features was approximately constant,i.e. 1.8 out of 5 as shown in Figure 11. This was similar to the observed preference within education clusters. Thus, across clusters,users tended to assign less preference to these features. However, the pattern of variance for these features was inconsistent. Given, that they presented a similar contextual dimension of the user, i.e. the domain knowledge of the expert, we present them together. The users in cluster 1 gave least importance to domain knowledge. Being new to the professional world and gotten the opportunity to experiment with different domains, they would give less importance to domain knowledge while judging an expert. However, with increase in the time spent within the professional world, the value for domain knowledge increases which can be viewed by the increase in preference by people in cluster 2. With further increase in professional background score, the value of domain knowledge is observed to decrease which is depicted by fall in preference score for length of summary and relevant courses. However, there is an increase in preference for projects. This can



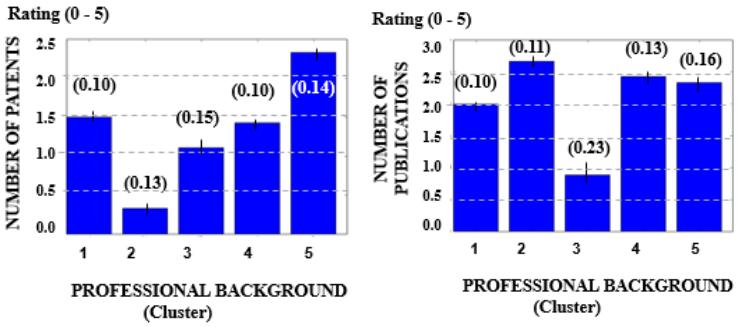
(a) Number of Projects (b) Length of Summary (c) Relevant Courses

Figure 11: Number of Projects, Length of Summary, Relevant Courses

be attributed to the fact that people in cluster 3 with a tendency to further their background quantitatively would consider projects as a mark of expertise. Again, with increase in years spent in the professional world, as preference for summary and courses increases, number of projects decreases. This can be attributed to the fact that, with greater professional experience, the value for quality over quantity increases.

4.2.4 Number of Patents, Number of Publication

The average preference for these features was 3.1 and they showed similar patterns of variance in-between clusters as shown in Figure 12. As described previously, these features presented a degree of additional effort invested by the user. People with minimal experience in the professional world (cluster 1 and 2), showed a high preference towards the degree of perceived effort while judging an expert. This can be attributed to the early motivation and energy in the new entrants and a zeal to succeed which would result in high value for judging expertise by degree of effort.



(a) Number of Patents (b) Number of Publication

Figure 12: Number of Patents, Number of Publication

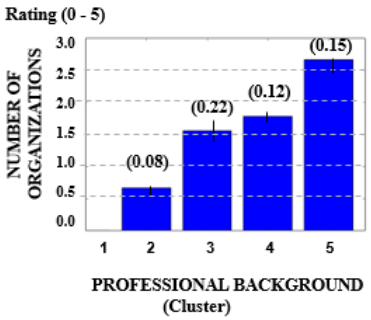
However, people at cluster 3 showed an unusually low preference for additional effort. This might have been a result of indecision between pursuing in professional world or re-entering academic world. Further on, the value for effort steadily increases.

4.2.5 Number of Organizations

The textual content relative to this feature,presented the initiatives taken by the user in order to innovate or contribute to the society. From the observations, we found that the value for this feature steadily increased with increase in quality of professional background as shown in Figure 13. Since, this requires a degree of self-confidence the value of this feature is realized with more time spent in the professional world.

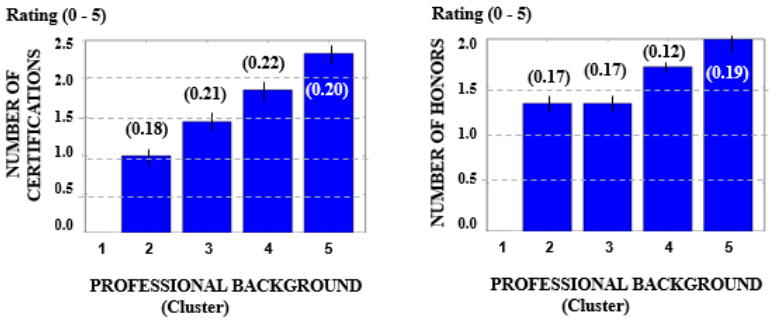
4.2.6 Number of Certification, Number of Honors

The average preference to these features was 2.3 out of 5 as shown in Figure 14. The textual context presented by these features showed a degree of self-interest and



(a) Number of Organizations

Figure 13: Number of Organizations



(a) Number of Certification (b) Number of Honors

Figure 14: Number of Certification, Number of Honors

motivation on the part of the expert. Certifications show the self interest of the individual in improving the quality of his skill. Honors presented a proof of quality of the expert in comparison to his peers. The preference for self-interest increased

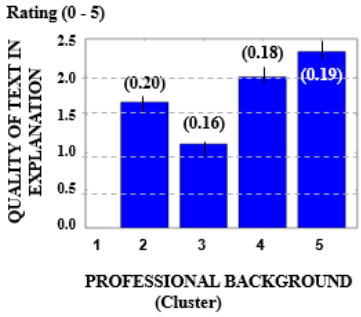
gradually from cluster 1 to cluster 5. This can be explained based on the fact that with more experience in the professional world, people realize the importance of quality of skill.

4.2.7 Quality of Text in Explanation

The average preference for this section was 2.5 out of 5 as shown in Figure 15. The pattern of variance of preference for this feature was similar to the pattern for length summary. Thus this proved our conjecture that, there was an error of judgment on part of the participants and they considered quality of text as greater words in the description.

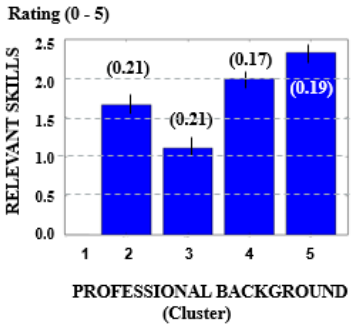
4.2.8 Relevant Skills

People in cluster 5 preferred it the most followed by people in cluster 2 as shown in Figure 16. The information presented in this section included a precise list of domain skills of the user. This observation was similar to the one for quality of text, though the current section received a much higher average preference across clusters than the former. This can be a result of skill being the foremost qualifying criteria to identify domain skills of the user. Based on these observations, we observed that the text in a LinkedIn profile represented 4 separate dimensions of an expert, i.e. background of expert, domain knowledge of expert, willingness to invest effort and self-interest of the expert. Since there was an ambiguity in the observations for quality of text in explanation, we decided to divide the score for quality of text in explanation into seven separate quality features representing quality of the content in individual sections and determine the value as a percentage of the value given to those sections by the participant. The final set of identified features have been presented in Table 9.



(a) Quality of Text in Explanation

Figure 15: Quality of Text in Explanation



(a) Relevant Skills

Figure 16: Relevant Skills

4.3 Extracting Expert Features

To investigate if there was any similarity in the content of the information blocks presented in prevous chapter and test the possibility of further clustering information blocks, we analyzed the textual content of each information block. In order to detect presence of mutual information we used (1) Cosine similarity, (2) Jaccard coefficient, (3) Kullback-Leibler divergence.

4.3.1 Cosine Similarity

Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them [26]. In case of text documents, it measures the differences between them by converting each document into a feature vector comprising of TF-IDF scores of individual words. We calculated the cosine similarity between each pair of blocks and observed that there were discrepancies in the result. For example, it returned a high value of similarity between courses and experience and quite low similarity between education and publication. Since these, results were not in sync with our assumptions, we decided to use a different measure. We also reasoned that, since the two block vectors being considered were multi-dimensional and sparse, thus the measure was not quite reliable.

4.3.2 Jaccard Coefficient

The Jaccard index, also known as the Jaccard similarity coefficient (originally coined coefficient de communauté by Paul Jaccard), is a statistic used for comparing the similarity and diversity of sample sets [20]. It gives a measure of the common terms present in each vector space. Thus we calculated the Jaccard coefficient for each pair of blocks. However, due to unequal size of the feature vector, we again observed certain discrepancies, i.e. when a feature vector with 10 features was com-

pared to one with 1000 features, the result obtained was not quite a reliable measure of similarity between the two.

4.3.3 Kullback-Leibler Divergence

In probability theory and information theory, the Kullback–Leibler divergence (also information divergence, information gain, relative entropy, KLIC, or KL divergence) is a non-symmetric measure of the difference between two probability distributions P and Q [27]. By considering each block as probability distribution of the topics in them, it gave us a measure of similarity between the two clusterings and was not affected by the multidimensional nature of the information. We present the results obtained in Table 11. The mapping of the information blocks has been presented in Table 10.

Original Feature	Modified Feature	Section
Highest technical degree achieved	Education Score	Education
Technical Positions	Position Quality	Experience
Area of Experience	Experience Text Value	Experience
Years of Experience	Experience Rating	Experience
Length of Summary	Summary Value	Summary
Number of Projects	Number of Projects	Projects
Relevant Courses taken	Number of Courses	Courses
Number of Organizations	Number of Organizations	Organizations
Number of Patents	Number of Patents	Patents
Number of Publications	Number of Publications	Publications
Number of Certifications	Number of Certifications	Certifications
Number of Honors	Number of Honors	Honors
Quality of text in explanation	Project Value,Publication Value, Certification Value,Honors Value	Quality
Relevant Skills	Skills	Skills

Table 9: Clustered Profile Features

Information Block	Name
Education	F1
Experience	F2
Projects	F3
Summary	F4
Courses	F5
Certification	F6
Honors	F7
Publication	F8
Scores	F9

Table 10: Mapped Information Blocks

As observed from the f-divergence scores in Table 11, we observed mutual similarity

between (1) Education and Experience. (2) Summary, Project and Courses. (3) Certification and Honors. (4) Publication and Scores. For example, people with similar academic background displayed similar variance in preference towards education and experience. Thus, by incorporating the results from the mutual similarity analysis with the earlier observations, we obtained the set of features for classifying content in an expert profile presented in Table 12.

f-vs-f	F1	F2	F3	F4	F5	F6	F7	F8	F9
F1	1.0	1.0	0.71	0.89	0.71	0.50	0.66	0.31	0.40
F2	1.0	1.0	0.71	0.89	0.71	0.50	0.66	0.31	0.40
F3	0.89	0.89	1.0	0.94	0.93	0.31	0.37	0.54	0.26
F4	0.89	0.89	0.94	1.0	0.89	0.29	0.39	0.34	0.33
F5	0.71	0.71	0.93	0.87	1.0	0.34	0.39	0.45	0.35
F6	0.50	0.50	0.31	0.29	0.34	1.0	0.89	0.46	0.27
F7	0.66	0.66	0.37	0.39	0.39	0.89	1.0	0.23	0.24
F8	0.31	0.52	0.54	0.34	0.45	0.46	0.23	1.0	0.85
F9	0.40	0.50	0.26	0.33	0.35	0.27	0.27	0.86	1.0

Table 11: F-divergence of Information Blocks

Features	Section
Education Score, Position Quality, Experience Text Value, Experience Rating	Education and Experience
Summary Value, Number of Projects, Project Value, Number of Courses	Summary, Project and Courses
Number of Organizations, Number of Patents, Publication Value, Number of Publications	Publication and Scores
Number of Certifications, Number of Honors, Honors Value, Certification Value	Certification and Honors
Skills	Skills

Table 12: Expert Profile Features

5 FEATURES FOR EXPERTISE

Having presented the tacit knowledge vector in Table 4 and the expert profile features in Table 12, in this section we construct the expert feature vector. Based on the expert profile features, we had observed that they can be classified into authority based features and quality based features. For example Number of Projects could be regarded as an authority based feature as it establishes the authority of the candidate expert in the area of projects. Quality of Projects would be regarded as a quality based feature as it represents the textual quality of the information conveyed with respect to this feature. [28] states the difference between an expert and a novice as a progression from a superficial and literal understanding of problems (a qualitative mark of the cognition of novices) to an articulated, conceptual, and principled understanding (a qualitative mark of the cognition of experts). We use this intuition to distinguish expert profiles on the basis of quality of the information content and quantity of authority content. Thus we categorize the identified expert features contained within every information block as (1) Qualitative; (2) Authoritative. We classify the features by recognizing the content represented by them as qualitative or authoritative. A quality feature presents the quality of the information mentioned with respect to the feature. Authority feature is based on the absolute numeric values of the laurels or milestones achieved by the user. Based on the feature categorization achieved in previous section, we define three processing steps to construct the expert profile vector: (1) Extract profile based features; (2) Incorporate preference based features; (3) Filter skill based features; (4) Construct Expert Feature Vector. Each of these processing steps are discussed in detail in the following section.

5.1 Extract Profile Based Features

This class of features captures the aspects of the experts profile which reflect his general domain specific competences. We begin by analyzing the textual information

contained within the profile of a candidate expert with respect to the features identified in previous section presented in Table 12 and classify them as (1) Background based; (2) Domain knowledge based; (3) Effort based; (4) Interest based.

5.1.1 Background Based Features

These are the set of features which are derived from the information presented in the profile of the candidate expert to express his knowledge background. Academia and professional world are the two places, from where a user cultivates his background and hones skills for excelling in future tasks. The features representing this have been extracted to represent the overall quality and authority of the user based on the highest technical degree attained as well as years of domain experience.

Experience Quality ($Q_{\mathbf{experience}}$)

This represents the quality of text used by the user to describe his professional experience. From our user studies we had observed that people tended to differentiate profiles based on the usage of terms in it. While some groups valued the text quality of the section others did not. Quality is determined in terms of the domain defined in his skill set, i.e. greater usage of words in the skill set or words which are closely related to it will result in a higher quality of text. Thus, to determine the quality value we pre-process the text in the experience section by removing stopwords, stemming the words and determining the similarity of the pre-processed set of words with respect to the skillset and assign it a quality value. Suppose the user mentions JAVA as one of the many skills in his skill set and mentions JDBC as an area of work or study in the current section, then it would get a quality value of 0.5. The value 0.5 is owing to the fact that JAVA has a frequency rank 2 in the Wikipedia page for JDBC (after removing stopwords and other high frequency words which do not contribute to the uniqueness of the document). Thus for term t(JDBC) and given skill set S, its quality value is calculated as given below.

$$quality(t) = \max_{i=1}^{n} \frac{1}{rank(w,t)}$$

Here, n represents the total number of words in the block after removing unimportant words, rank(w,t) gives the rank of w (JAVA) in the Wikipedia page for t (JDBC) determined with respect to its frequency of occurrence in the page.

Experience Authority ($A_{\mathbf{score}}, A_{\mathbf{jobs}}$)

This feature identifies the years of experience of the current user (A_{score}) and assigns it a score from 1–5 based on the scoring system in Table 5. It also identifies the number of jobs (A_{jobs}) held by the user in those years. By observing the contextual information conveyed with respect to these two features, we inferred that number of years of experience conveyed a quantitative aspect of professional experience of the user while number of jobs conveyed the value of the person to the organization resulting in a number of promotions or instability of the user resulting in a number of job switches. Overall, both of these features conveyed the background of the user.

Education Authority ($A_{\mathbf{education}}$)

This feature identifies the highest technical degree achieved by the current user ($A_{\text{education}}$) and assigns it a score from 1–4 based on the scoring system in Table 5. The academic degrees held by the user are quite valuable in creating a background profile of the user.

The above set of features are linearly combined using the following formula to generate $F_{\text{background}}$. The value of this feature represents the content quality as well as degree of quantitative excellence of the user's background.

$$F_{\text{background}} = Q_{\text{experience}} + \log_{10}(A_{\text{score}} + A_{\text{jobs}} + A_{\text{education}})$$

5.1.2 Domain Knowledge Based Features

A person with good understanding of his domain, would be able to present it more succinctly than a novice, who would end up presenting it in a more convoluted and lengthy manner with lesser density of domain related terms[28]. They capture the level of confidence of the person with respect to the content in them as well as indicate a degree of awareness of the user about his domain. The set of features in this section represents the domain awareness of the user.

Summary Quality ($Q_{\mathbf{summary}}$)

This feature reflects a degree of command over the mentioned skill set. A person with a greater awareness of the skills listed in his skill set would mention them in his summary in order to make it more noticeable. We evaluate the value of this feature with the quality parameter described above.

Project Quality ($Q_{\mathbf{project}}$)

This feature represents the ability of the user to mention his work in a manner which would be easiest to comprehend. A greater density of text not related to his specified domain, is perceived as a negative trait for the user. It indicates that the user lacks awareness or is not confident enough to present the absolute details.

Project Authority ($A_{\mathbf{project}}$)

This feature presents the measurable milestones achieved by the user in his domain. These are not necessarily any special recognition in the field but represent the quan-tity of work done by the user in that domain thus adding to his credibility within the domain.

Course Authority ($A_{\textbf{course}}$)

The courses taken by the user is a reflection of the domain of interest of the user. People specify only those courses that according to them have helped them furthering there skills. It represents the tools acquired by the user thereby strengthening his credibility within his domain. Thus this feature has a similar intent and strengthens the position of the user within his mentioned domain.

The above set of features are linearly combined using the following formula to generate F_{domain}.

$$F_{\text{domain}} = (Q_{\text{summary}} + Q_{\text{project}}) + \log_{10}(A_{\text{project}} + A_{\text{course}})$$

5.1.3 Effort Based Features

The set of features in this section, represent the achievements of the user in his work as a result of effort invested by him. It is a measure of the sincerity and dedication displayed by the user towards his work. The authoritative measure captures the recognized prowess of the user in his domain and the qualitative measure captures his interest in the work.

Certification Quality ($Q_{\textbf{certification}}$)

A certification is a credential that you earn to show that you have specific skills or knowledge. Thus, it represents the additional effort invested by the user. The quality aspect is determined in terms of its relevance to his skill set which in turn justifies the effort for furthering his specified skills.

Honor Quality ($Q_{\textbf{honor}}$)

This feature represents the achievements of the user by investing additional effort in his domain. Users tend to mention the awards or recognition given to them as a result of their hard work and giving better performance in comparison to his peers.

Certification Authority ($A_{\textbf{certification}}$)

This feature quantitatively measures the additional effort of the user in terms of the number of credentials attained by him.

The above set of features are linearly combined using the following formula to generate F_{effort}.

$$F_{\text{effort}} = (Q_{\text{certification}} + Q_{\text{honor}}) + \log_{10}(A_{\text{certification}} + A_{\text{honor}})$$

5.1.4 Interest Based Features

Apart from effort, the information conveyed through the profiles also reflects the self-interest of the user in furthering his skills. We measure this self-interest in terms of two features mentioned below. Their qualitative aspect reflects an interest in the user in developing his skill set and the authority aspect reflects the measurable milestones achieved by him as a result of being faithful to his interest.

Publication Quality ($Q_{\textbf{publication}}$)

In general terms, publishing is the process of production and dissemination of literature, music, or information i.e. the activity of making information available to the general public. Given the nature of the work, it warrants self-interest on the part of the author. Thus we identify this feature as a reflection of the quality effort invested by the user by self-interest, to make an impact within his domain.

Publication Authority ($A_{\textbf{publication}}$)

This feature reflects the measurable value achieved by the user in terms of the number of documents published.

Score Authority ($A_{\textbf{score}}$)

The section pertaining to score value, contains information about the scores of the user achieved in competitive examinations. This reflects the measurable value achieved by the user in terms of quantitatively defining his dominance within his domain motivated by self-interest.

We refrain from using the quality value of the score since, we observe that all user profiles which mention score values have a above average value for the score. People with below average scores refrain from mentioning it in their profile. The above set of features are linearly combined using the following formula to generate F_{interest}.

$$F_{\text{interest}} = Q_{\text{publication}} + \log_{10}(A_{\text{publication}} + A_{\text{score}})$$

5.2 Apply Filter to generate Preference Based Features

In this step we identify the preference of the current user while evaluating the profile of an expert and incorporate the preference values with the profile based feature values inorder to appropriately adjust their values based on current user's preference. Preference based features represent the preferences of the current user while selecting an expert profile. The knowledge requisite for this feature is obtained from the crowd perception recorded by our pilot studies as presented in Table 13. The tacit knowledge vector of each participant along-with their responses gathered from the pilot studies are incorporated into a learning framework to generate the

preference for an incoming user. The expert profile preference is generated as the output of the learning framework.

Topic	Data Type	Total no of responses
Java	Choice of Expert	2,211
Java	Choice of Feature	3,553

Table 13: Collected Crowd Preference Data

5.2.1 Background Importance

This feature reflects the degree of preference associated by the user to the background of an expert($R_{\text{background}}$). For example if the user considers a high technical degree to be critical in determining expertise then he would focus more on this section while choosing an expert. Based on crowd knowledge, we predict the importance given by current user to education and experience blocks. These values are incorporated into our estimation. We utilize the predicted values of these features to calculate the mean weightage given by the user to the background of an individual while judging his profile for an expert and append the modified value to the expert profile preference. The modified value of $F_{\text{background}}$ is estimated as given below and appended to the expert profile vector.

$$F_{\text{new background}} = F_{\text{background}} * R_{\text{background}}$$

5.2.2 Domain Knowledge Importance

This feature represents the preference associated by the user to domain knowledge while selecting an expert(R_{domain}). For example a user might consider a course in machine learning highly critical for being an expert in data mining. Thus this section would receive a greater weighting for that user. Using crowd knowledge to predict the value of these features, we calculate the mean weighting given by the user to the domain knowledge of an individual while judging his profile for an expert and append the modified value to the expert profile preference. The modified value of F_{domain} is estimated as below and appended to the expert profile vector.

$$F_{\text{new domain}} = F_{\text{domain}} * R_{\text{domain}}$$

5.2.3 Effort Importance

This feature represents the preference associated by the user to efforts invested in cultivating expertise in a particular domain(R_{effort}). We believe certification and honors recognize the effort invested by the user to further his domain knowledge. If a user considers certification as a critical measure to determine expertise in a domain,then this field would obtain greater value for the user. We refer to the predicted values of these features and accordingly calculate the mean weighting given by the user to the effort displayed by an individual in improvising his domain related skills and excel at the task assigned. The modified value is then appended to the expert profile preference. The modified value of F_{effort} is estimated as below and appended to the expert profile vector.

$$F_{\text{new effort}} = F_{\text{effort}} * R_{\text{effort}}$$

5.2.4 Interest Importance

This feature represents the preference associated by the user to self-interest of the candidate(R_{interest}). A user who appreciates extra effort put in by the candidate by self-interest and considers it as a vital attribute of an expert, would assign greater value to this

Topic	Feature Name
Background importance	$F_{\text{new background}}$
Domain knowledge importance	$F_{\text{new domain knowledge}}$
Effort importance	$F_{\text{new effort}}$
Interest importance	$F_{\text{new interest}}$

Table 14: Expert Profile Preference

feature. Again crowd knowledge is used to predict the value of these features and the mean weightage, given by the user to self-interest and self-motivation reflected from the profile of the candidate, is calculated and appended to the expert profile preference. The modified value of F_{interest} is estimated as below and appended to the expert profile vector.

$$F_{\text{new interest}} = F_{\text{interest}} * R_{\text{interest}}$$

5.3 Augment Skill Based Features

In this step we define two features to represent the relevance of the expert profile to the query term and append them to the expert profile vector. By scrutinizing the skill set of the expert, we determine its relevance and closeness to the query term. We discuss both the measure in detail below.

5.3.1 Closeness

This feature quantifies the quality based density of terms found in the candidate expert profile with respect to their relevance to the domain(C_{skill}). It gives us an insight into how closely related are his domain skills with the information provided by him in the information blocks within his profile. Thus, if a person has mentioned JAVA as a skill but has not mentioned JAVA in any of the information blocks, then his profile will receive a low closeness score. In contrast, if a person has mentioned JAVA as a skill and has mentioned having worked as Java Lead, then his profile would get a relatively higher closeness score. In the feature, we consider a relative match instead of an absolute match, i.e. given a skill JAVA if the person mentions JDBC in his information block, then his profile would receive a score less than 1 but not 0. With higher frequency of mention of the term JAVA the value of closeness for this profile would increase. Thus, if we represent the skill set of the individual by a set S, then value of closeness is calculated as :

$$C_{\text{skill}} = \frac{1}{N_p} \sum_{w \in p} (\min_{s \in S}(rank(s, w)))$$

N_p is the total number of words in the profile of the user after pruning stopwords, stemming the terms and removing conjunctions. $rank(s, w)$ indicates the rank of word s (s is a constituent of the set S) in the Wikipedia page for w with respect to its frequency of occurrence within the content.

5.3.2 Relevance

This section quantifies the quality of the text within the user profile with respect to the query term. This feature(C_{query}) aims to evaluate the relevance of the information within the candidate expert profile with the query term. For example, the query term is PYTHON and the person has not mentioned PYTHON anywhere in any of the information blocks, then his profile would receive a low relevance score. On the contrary, given the query term PYTHON if the person mentions having done projects in PYTHON then his profile would receive a high relevance score. An important aspect of this feature is it calculates a degree of relevance with the query term instead of a boolean match i.e. if *Flask* or *Django* have been mentioned in the users' profile, then the user profile would receive a value less than 1 but not 0. Given the query term t, the value of relevance would be calculated as :

$$C_{query} = \frac{1}{N_p}(\sum_{w \in p}((rank(t, w)))$$

N_p is the total number of words in the profile of the user after pruning stopwords, stemming the terms and removing conjunctions. $rank(s, w)$ indicates the rank of word s in the Wikipedia page for w with respect to its frequency of occurrence within the content.

5.4 Expert Feature Vector

Having incorporated the evaluated values in the expert feature vector, we proceed to present the final expert feature vector in Table 15. At each step we modify the values obtained by text analysis of the information provided by them and append it to the expert feature vector. The expert feature vector represents the set of values which would be used to evaluated the expertise of an expert with respect tot he current user. It comprises of $F_{\text{new background}}$, $F_{\text{new domain}}$, $F_{\text{new effort}}$, $F_{\text{new interest}}$, C_{skill} and C_{query} values shown in Table 14. Our motivation behind constructing the expert

feature vector was to boost those information blocks within the expert profile which would be majorly considered by the user while choosing an expert and give less relative importance to less important ones. It represents the measuring stick used by the current user to judge an expert.

Profile based	Background based	Experience Quality, $Q_{experience}$
		Experience Authority, A_{score}, A_{jobs}
		Education Authority, $A_{education}$
	Domain knowledge based	Summary Quality, $Q_{summary}$
		Project Quality, $Q_{project}$
		Project Authority, $A_{project}$
		Course Authority, A_{course}
	Effort based	Certification Quality, $Q_{certification}$
		Honor Quality, Q_{honor}
		Certification Authority, $A_{certification}$
	Interest based	Publication Quality, $Q_{publication}$
		Publication Authority, $A_{publication}$
		Score Authority, A_{score}
Preference based	Background importance, $R_{education}$,$R_{experience}$	
	Domain importance, $R_{summary}$,$R_{project}$, $R_{courses}$	
	Effort importance, $R_{certification}$,R_{honors}	
	Interest importance, $R_{publications}$,R_{scores}	
Skill based	Closeness, C_{skill}	
	Relevance, C_{query}	

Table 15: Expert Feature Vector

6 TAK: TACIT KNOWLEDGE BASED APPROACH

We have identified the preferences of the user and constructed the expert feature vector based on the same. In this section we present in detail our proposed learning framework created to predict the preferences of the current user. The foundation of this framework is based on the knowledge gathered from the user studies presented in previous sections and the expert feature vector hence constructed. Given that this approach bases its predictions on the estimated tacit knowledge of the user, we call it as TAK (Tacit Knowledge based Approach). Based on the pilot studies, we had categorized the profile features into four major characteristics to classify the information presented in an expert's profile. We observed that tacit knowledge of the user, which is developed as a result of his education and experience, leads to difference in relative weighting of the four expert characteristics. Given a query on a particular topic, TAK identifies the preferences of the current user based on the responses collected from the pilot studies and assists in construction of the expert feature vector. In order to rank experts accurately, TAK creates a tacit knowledge vector on the basis of his professional and academic background identified using his LinkedIn profile. The constituents of a user profile and their role in construction of the expert feature vector will be discussed below.

6.1 Revisiting Problem Statement

Given a user John and a query term JAVA, we aim to populate the list of JAVA experts whom John would consider as experts. Prior to presenting our solution to this problem, we state our definition of an expert. We define an expert as the person with the capability to provide the best possible solution to the problem at hand in the most comprehensible way with respect to the current user. Suppose John, who has 10 years of experience in JAVA and has a Master's degree, is looking for an expert in JAVA. We argue that, his tacit knowledge would have been augmented as a result

of 10 years of experience and a Master's degree and this would influence his selection of an expert. Thus based on his tacit knowledge he will give varying importance to different blocks of information available within a candidate expert profile and judge his expertise accordingly.

6.2 Overview of Solution

How to accurately identify the exact expectations of the user from an expert and make accurate suggestions? In order to do the same we developed a method to evaluate the user profile data and identify the preferences of the user from the same. In the following section we discuss the steps involved in predicting a user's preference and structuring the expert profile appropriately in order to identify an expert for the current user.

6.2.1 Create Tacit Knowledge Vector

In order to tackle this problem, we define a profile for John at the onset. To accurately identify the preference of John, it is highly critical to evaluate the tacit knowledge of John from the profile information of the user. John's tacit knowledge vector is constructed by semantic analysis of the data mined from his LinkedIn profile and represented as $\{F_U = \{f_{u1}, f_{u2}, ...f_{um}\}\}$. The first feature, i.e. education, is calculated based on the highest level of education attained by him and it is assigned a value according to predetermined score as discussed earlier. Next, experience is estimated based on the years of experience in the professional world and a value is assigned according to predetermined score. The value for additional information is decided based on the presence of predetermined set of attributes presented in Table 16 and is set to 1 if information is found in either of these sections. Thus, the tacit knowledge vector for John is { Education : 2 , Experience : 5, Additional Information : 0 }.

6.2.2 Get User Preference

Having identified the tacit knowledge of the current user we proceed to predict his expert selection criteria. In Section 4, we presented two pilot studies. We use the

Patents
Publications
Certifications
Honors
Organizations
Scores
Projects

Table 16: User Additional Information

data obtained from the gathered responses to prepare the training set for the supervised learning framework. The tacit knowledge vector is created for each participant based on the professional and academic information pertaining to the people interviewed. Their preferences for each category mentioned in Table 14 are mapped to a four digit number where each digit signifies the preference of the user for each of the profile features mentioned in Table 14. We define it as the response number as shown in Figure 17. The user preferences are obtained from initial interactions where users had recorded their preferences among the provided profile features and ranked their preference on a scale of 1–5. The features from the survey are grouped together as mentioned earlier in Table 15. The value of each digit in the response is assigned a value from 0–5 based on average value of the rank indicated by the participants

in each of the constituent features. Hence, if a participant indicated a rank of 2 for technical positions, area of experience, years of experience and highest technical degree achieved, then the response number would begin with 2. The tacit knowledge vector of every participant along-with the respective response number, forms the training data set for the framework. The profile of the current user evaluated in the previous step, is input to the framework. k-Nearest Neighbor algorithm is used to determine the expert profile preference as mentioned in Table 14, for a new user, i.e. the relative weighting given by the new user to the profile features.

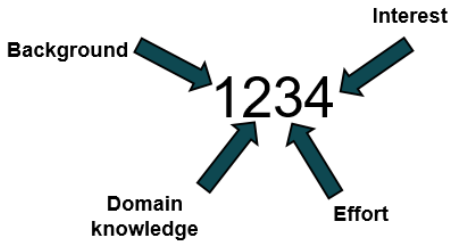


Figure 17: Response Number

6.2.3 Pooling Experts

Who will qualify as an expert for a particular user? The decision of the user will depend on two basic factors:
(1)Credibility of the candidate with respect to the query term.
(2)Credibility of the candidate with respect to the user.
We determine the former measure based on the information retrieved from the candidate's profile which displays his level of authority with respect to the query term. For example if the user mentions 10 certifications in Sun certified courses, then that is an indication of his degree of authority in the related domain, i.e. Java. This

information is obtained by analysing the textual content within the profile of the candidate expert. The later measure is dependent on the user's preference based on his tacit knowledge. For example, we have a user who has a PhD and 2 years of experience and a candidate with PhD and 20 years of experience. Without considering tacit knowledge, for the user the candidate would seem to be a clear choice as an expert, but owing to life experiences, his preference might be different thus making our assumption erroneous. In this section, we measure the value of the former, i.e. the credibility of the candidate with respect to the query term. Based on the query term, we first create a pool of candidate experts from the list returned by LinkedIn search URL shown in Figure 1. The candidates in this list are those who have mentioned the query term within their profile with varying frequency. A LinkedIn query returns on an average 1,000,000 results. However, from our investigation we have found that

Total size of Pool	Average Quality of Content
300	0.7
1200	0.4

Table 17: Information Content Quality

more than 90% of the profiles in the search result do not provide sufficient evidence to verify their relevance since they are out of network connections. On examining the list further, we observed that neither does this list rank the candidates according to their level of expertise nor does it provide any information to the user to take an

informed decision. Thus, we use this list as the candidate expert pool and proceed to define each candidate by a set of features, expert profile vector. The features in the vector are extracted by mining the information contained in each candidate's profile. In order to decide the total size of the candidate pool, we verified a set of 10,000 LinkedIn profiles and observed the maximum number of connections per user to be 1,200 and the average number of connections to be 300. We tested by setting the size of the pool to 1,200 as well as 300 (information quality details mentioned in Table 17). The average quality of the information content across the profiles was observed to be better with 300 as the total size.

$$Q_p = \frac{1}{R_n} \sum_{w \in R} \mathrm{v(w)}$$

R is the set of words which had the query term in their www.wikipedia.org page. R_n is the size of the set of words R. v(w) is the rank of the query term in the www.wikipedia.org page for the word w. Thus, we set the total size of the expert pool to be 300.

6.2.4 Expert Profiling

The profile of an expert constitutes the expert profile vector. Using a web crawler written in Python2.7, we extract the profile page for each candidate profile. BeautifulSoup4[29] is used to parse the contents of the page and assign values to them. The features representing the profile of the expert from Table 12 is classified into two major divisions: (1) Features representing authority. (2) Features representing quality. The former presents the quantifiable aspects of the achievements and credibility of the user while the later presents the qualitative aspect of the information with respect to each individual feature. For example, number of projects is a feature representing authority of the expert with respect to the projects he has been a part of while project value is a feature representing quality of the expert by conveying

the quality of the text used to present the project descriptions. The authority based features are evaluated from their absolute values extracted from the data in the information block. In order to evaluate the content quality we build a dictionary of domain relevant terms for the particular expert from the list of skills mentioned in the skill section in the profile of the expert. Each skill term is appended with their respective content value based on their closeness with the query term. The closeness value of a skill is calculated with respect to the query term. In-order to calculate the closeness, we devised a ranking approach after removing the stopwords and sentence connectors using NTLK corpus. $value(w, q)$ represents the value of word

$$value(w, q) = \frac{N_q}{f(w)}(w \in N_q)$$

w in wikipedia page for q. N_q represents the set of words obtained after removing the stop words and sentence connectors from the set of words obtained from the wikipedia page for q. f(w) is the number of occurrences of term w in the set N_q. For example, the user mentions JDBC in his profile and the query term is Java, then we find the rank of Java in www.wikipedia.org page for JDBC. Java ranks 2 based on its frequency of occurrence (after removing sentence connectors) in the wikipedia page for JDBC. Thus the value assigned to JDBC will be $\frac{1}{2}$ (i.e. $value$(Java,JDBC) = $\frac{1}{2}$). Instead of using TFIDF score to rank the words in the information block, we device a ranking method by removing stopwords and sentence connectors because of the absence of a corpus of documents to compare with. In the absence of the corpus of documents, the TFIDF score obtained was erroneous with more frequent terms like conjunctions receiving higher weigtage.

6.2.5 Expert Profile Vector Creation

Having evaluated the quality and authority based expert profile features, the expert profile vector is created comprising of $F_{\text{new background}}$, $F_{\text{new domain}}$, $F_{\text{new effort}}$, $F_{\text{new interest}}$, C_{skill} and C_{query} values as shown in Table 15.

6.2.6 Scoring Method

In the previous step we created the expert profile vector. The six constituent features are $F_{\text{new background}}$, $F_{\text{new domain}}$, $F_{\text{new interest}}$, $F_{\text{new effort}}$, C_{skill} and C_{query}. The final expertise score is calculated based on their values. Given a candidate e, the expertise score is calculated by,

$$score(e) = (F_{\text{new background}} + F_{\text{new domain}} + F_{\text{new interest}} + F_{\text{new effort}}) * (\frac{(C_{\text{skill}} + C_{\text{query}})}{2})$$

Thus, a score is calculated for each candidate expert and based on the value of the score, the candidates are ranked in ascending order. In this section we have presented the learning framework, TAK, created to predict expert selection preference of the current user and ranked the candidate experts accordingly.

7 RESULTS AND EVALUATION

In the previous section we have presented TAK, which accurately ranks experts with respect to the current user. Our algorithm predicts the tacit knowledge of the user from the data collected from the user's online profile. This tacit knowledge is used to rank candidate experts. In the current section we present an evaluation strategy and a set of metrics to test the efficacy of the approach. We report the results obtained by an experiment conducted to test the accuracy of the method.

7.1 Experimental Setup

The dataset used in the experiment has been described in Table 13, which comprises of the responses gathered as a result of informal interaction with people.

7.1.1 Query

The query designed by us to test the efficacy of our approach, had a list of 10 candidate experts. The participants were instructed to read through a given user profile and rank the given candidate experts from the point of view of the given user. The term used to determine expertise was Java.

7.1.2 Candidates

We created 30 candidate subsets each with the public URL of a LinkedIn user in it and 10 candidate Java experts. The user profiles were obtained from 30 randomly selected profiles with varied proficiency in Java and a publicly available url. We mapped 10 candidate expert profiles to each user profile. The candidate expert profiles were obtained using the search API given by LinkedIn.

7.1.3 Method

The textual content presented in the public LinkedIn profile of the 30 chosen users was analysed in order to create the tacit knowledge vector of the user. This vector was

provided as input for a new user to TAK. Corresponding to the tacit knowledge vector of the user, TAK predicted a response number. The response number comprised of the predicted preference of the user for each of the 15 profile features required to construct the expert profile vector. We proceeded to extract information from the profiles of the 10 candidate experts and incorporated the response number values with it to construct the expert feature vector as presented in Table 15 for each candidate expert. Based on the expert profile vector, the candidates were ranked with the candidate obtaining the highest score securing the highest rank.

7.1.4 Gathering Ground Truth

In order to cross validate the rank generated by our framework, we conducted an AMT (Amazon Mechanical Turks) survey wherein we gathered Turker's opinions on the rankings generated by our method. We designed a HIT(Human Intelligence Test) which was made available to each turker. The qualification for a turker to participate in our survey was set as Master. Masters are an elite group of Workers, who have demonstrated superior performance while completing thousands of HITs across the Mechanical Turk marketplace[30]. Thus this to ensured that the opinion provided by them was highly credible.

7.1.5 The HIT Design

A total of 30 HITs were created for the 30 user profiles with 10 candidate expert profiles per user. At the onset of the survey, each turker was given a brief description of his role in the survey and the steps to be followed. The first page of the form presented a topic to the turker with a brief explanation of the context. We conducted the survey with Java as the query term. Followed by the topic, the participant was presented with a user profile and instructed to carefully observe it. We explained to them the importance of being vigilant in the set of instructions given to them. Next they were given a set of 10 expert profiles and were asked to rate each candidate on a

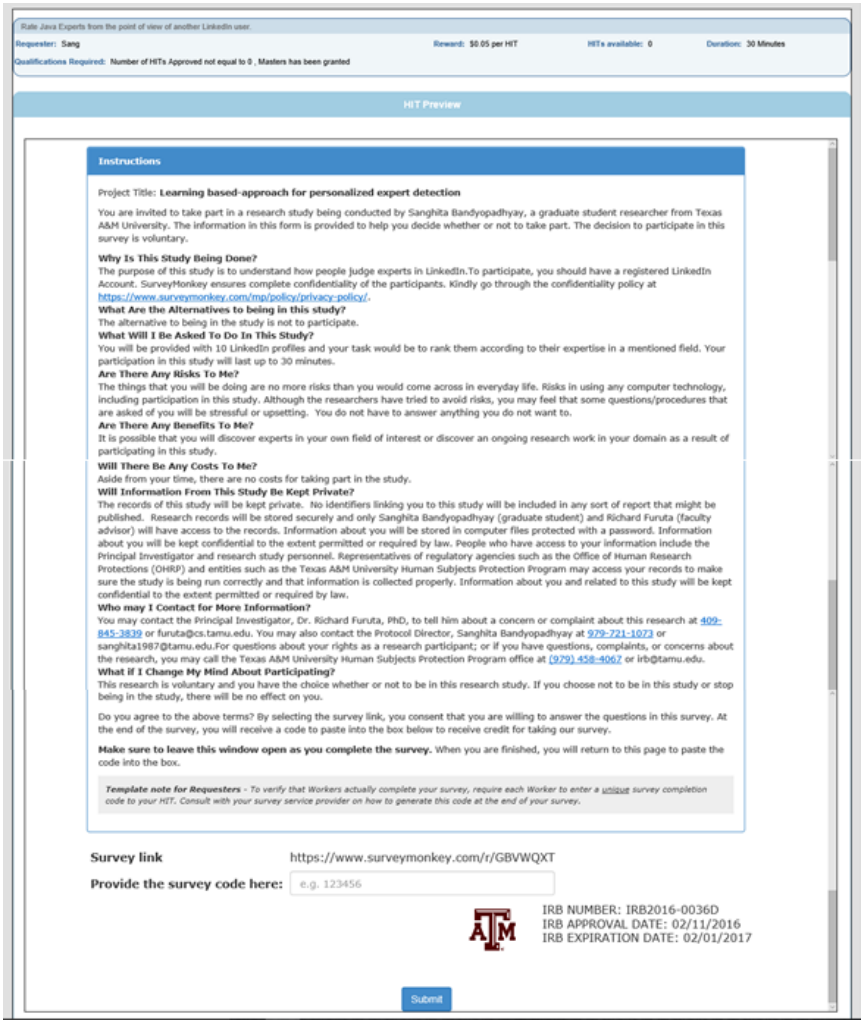


Figure 18: HIT User Consent

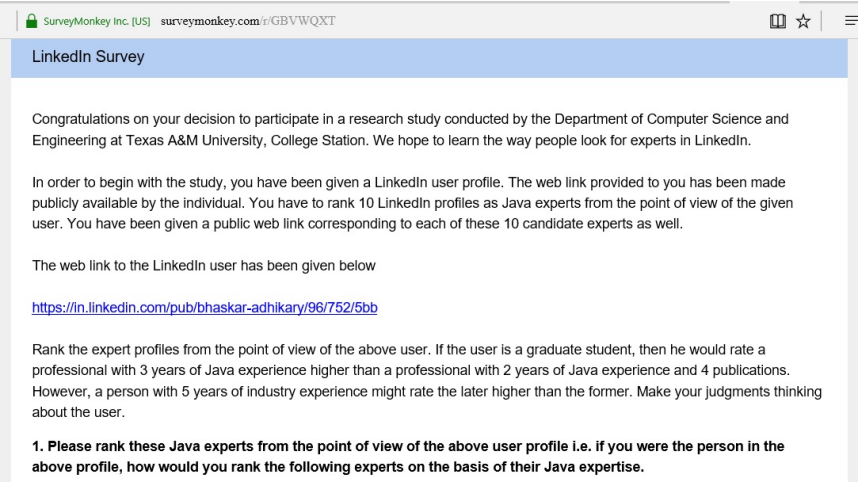


Figure 19: Survey Introduction

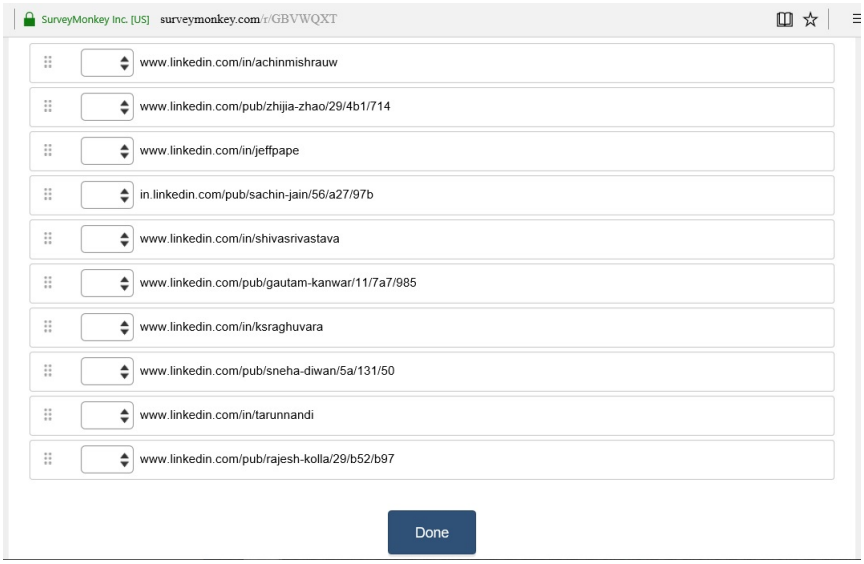


Figure 20: Questionnaire

scale from 1–10 from the point of view of the user whose profile they had just visited. So, the task of the turker was to rank the 10 profiles as the given user would, from 1–10. In order to guarantee credible responses, the qualification of the turker was set to Master. To further ensure reliability of the results, in every set of 10 candidate experts we included one candidate profile with no mention of the query term in their qualifications. This was done to identify responses with low reliability and eliminate them from the final assessment.

7.1.6 Turker Agreement

A total of 1,021 HIT responses were collected. To verify the reliability of the responses, we evaluated the accuracy of the ranks provided from the point of view of each user. The accuracy for a user u is defined as

$$\text{accuracy(u)} = \frac{\text{No. of majority judgments}}{\text{No. of judgments}}$$

Majority judgments was defined as the final cumulative ranking obtained from all the collected responses. Thus, the value of accuracy would be 1 in-case all the participants provide the same ranking from the point of view of the given user and would be 0 in-case none of the rankings match. In Table 18 we provide the value for accuracy calculated per user profile for M-1[1] and M-2[2]. The average accuracy across all users using tacit-knowledge based approach is 0.572 which indicates that around 5 among 9 turkers agreed on the ranking assigned to the set of experts. The average accuracy across all users using the ranking provided by LinkedIn is 0.28 which indicates that around 3 among 9 turkers agreed on the ranking assigned to the set of experts.

[1]ranking given by the TAK
[2]ranking given by LinkedIn

7.1.7 Evaluation Metrics

We tested the accuracy of our framework against the HIT responses. In-order to

User	M-1	M-2	User	M-1	M-2	User	M-1	M-2
1	0.743	0.094	11	0.523	0.369	21	0.560	0.229
2	0.431	0.317	12	0.614	0.297	22	0.623	0.208
3	0.440	0.286	13	0.614	0.371	23	0.614	0.269
4	0.514	0.254	14	0.640	0.351	24	0.571	0.254
5	0.531	0.243	15	0.631	0.251	25	0.566	0.277
6	0.554	0.291	16	0.591	0.200	26	0.631	0.266
7	0.483	0.277	17	0.557	0.294	27	0.551	0.383
8	0.594	0.288	18	0.580	0.226	28	0.537	0.363
9	0.563	0.343	19	0.588	0.271	29	0.523	0.343
10	0.517	0.323	20	0.651	0.237	30	0.620	0.223

Table 18: Accuracy Value per Method

test the accuracy of the framework, we followed [31] which presented three evaluation metrics namely : (1) Recall@k (2) Precision@k (3) NDCG@k. These are quite popular measures for evaluating web search and url rankings.

Recall@k

Measures the fraction of relevant rankings that are retrieved from the responses i.e. the fraction of the ranking generated by the proposed method which is identical to the ranking given by the users. It represents an average of the relevant rankings effectively obtained at each rank among all the registered responses. This indicates

the efficiency of our method in generating rank values which are agreeable to the users. We have considered k to be 10. The value for Recall@10 ranges from 0 to 1. Here a value of 1 indicates that all of the responses are in agreement with the rank generated by the learning based method. It is calculated by the formula given below where e_i is the expert whose rank in the current set is being considered.

$$\text{Recall@10} = \sum_{i=1}^{10} \frac{\text{rank}(e_i, q)}{10}$$

rank(e_i,q) provides the number of relevant ranking values for e_i which are identical to the values generated by the tested method. The sequence in which the experts are being considered is same as the sequence maintained in the response forms(Figure 18, Figure 19, Figure 20).

Precision@k

Measures the fraction of the registered responses that are relevant i.e. the fraction of the users who believe the ranking generated by our method is correct. It represents an average of the relevant rankings effectively obtained at each rank among all the responses till that rank. It is calculated sequentially as per the order maintained in the survey form. An average of all the precision values obtained at each rank provides the precision metric for an individual form. By obtaining the average value from all the survey forms we obtain the final value for precision. This indicates the popularity of the results of our ranking algorithm. It is given by the formula

$$\text{Precision@10} = \sum_{i=1}^{10} \frac{p_i}{10} \ , \begin{cases} p_i = 1, \text{if rank}(e_i, q) >= 8 \\ p_i = 0, \text{otherwise} \end{cases}$$

NDCG@k

Measures the relevance of the ranking generated for the users. A higher value of NDCG indicates greater relevance. If a relevant user is ranked lowly, then it is less likely that they will be chosen as an expert. Thus if a method returns a ranking with a greater value of NDCG, it will rank relevant candidate experts higher. We measure the value of NDCG@10 for each subset of 10 candidate experts in each collected response since we customize the ideal ranking with respect to a particular subset.

$$\text{NDCG@10} = \frac{\text{DCG@10}}{\text{IDCG@10}}$$

$$\text{DCG@10} = \text{rank}(e_1, q) + \sum_{i=2}^{10} \frac{\text{rank}(e_i, q)}{log_2 i}$$

The maximum DCG@10 values across all the responses is considered as the ideal discounted cumulative gain i.e. IDCG@10.

7.2 Results

Having presented the metrics and the data used in the conducted experiment, we report the results in the current section. Table 19 presents the Precision@10, Recall@10 and NDCG@10 values. We have compared these values between the ranking generated by LinkedIn and TAK.

7.2.1 Comparitive Analysis

We compare the ranking generated by LinkedIn with that of our proposed method, TAK. Table 19 shows the Recall@10, Precision@10 and NDCG@10 values of each method in ranking experts. We observe that the proposed method clearly performs

better than LinkedIn. The average observed value of Precision@10 is around 0.5717, which indicates that at our approach provides a 50% probability of providing the desired expert ranking in comparision to the ranking provided by LinkedIn which has a Precision@10 value of 0.2883 indicating a probability of 30% approximately. This result is further strengthened by the Recall@10 value of 0.3145 in comparison with 0.1552 in case of LinkedIn rankings. The value of Recall@10 suggests an acceptance rate of 31% in comparison to 15.5% incase of the ranking given by LinkedIn. This indicates that the ranking strategy implemented by our approach provides more popular results in comparision to the results given by LinkedIn. The popularity of our approach is further strengthened by the NDCG@10 value of 0.6425 in case of TAK and 0.6398 in case of the ranking given by LinkedIn. Though in both cases the value of NDCG@10 does not meet the qualitative requirement of 0.8, our approach performs marginally better, hence providing better quality rankings in comparison to LinkedIn.

Methods	P@10	R@10	NDCG@10
LinkedIn	0.2883	0.1552	0.6398
TKA	**0.5717**	**0.3145**	**0.6425**

Table 19: Results

8 CONCLUSION

8.1 Conclusion

In this work, we present tacit knowledge as an agent influencing user's expert selection decision. We present and evaluate a tacit knowledge based learning method, i.e. TAK, to predict a set of experts for a given user. Previous works to predict experts used the knowledge pertaining to the expert to evaluate relative expertise thus resulting in the same set of experts for different users. My thesis creates a learning based framework, TAK, which evaluates tacit knowledge of an incoming user to predict a set of experts for him. The method utilizes the tacit knowledge gathered from carefully gathered user perceptions. Based on the learnt user background, it effectively differentiates between different users and predicts experts specific to their expected preference. The accuracy of TAK is evaluated by investigating crowd opinion. It is clearly presents the popular results.

8.2 Future Work

Our current method has been tested with only a particular software skill. It's robustness will be further improved by testing it with multiple software skills as well as non-software skills. Instability in the academic as well as professional world will be investigated as another dimension which could influence tacit knowledge of a user.

REFERENCES

[1] T. Yukawa, K. Kasahara, T. Kita, and T. Kato, "An expert recommendation system using concept-based relevance discernment," *13th IEEE International Conference on Tools with Artificial Intelligence*, 2001.

[2] A. Omidvar, M. Garakani, and S. Hamid R., "Context based user ranking in forums for expert finding using wordnet dictionary and social network analysis," *Information Technology and Management*, 2014.

[3] A. Hawalaha and M. Fasli, "Dynamic user profiles for web personalisation," *Expert Systems with Applications*, 2015.

[4] Z. Xu, H. Chen, and J. Yu, "Generating personalized web search using semantic context," *The Scientific World Journal*, 2015.

[5] C. Christopher S., M. Paul P., A. Cozzi, and B. Dom, "Expertise identification using email communications," *CIKM '03*, 2003.

[6] B. Dom, I. Eiron, A. Cozzi, and Y. Zhang, "Graph-based ranking algorithms for e-mail expertise analysis," *DMKD '03*, 2003.

[7] K. Ehrlich, C.-Y. Lin, and V. Griffiths-Fisher., "Searching for experts in the enterprise: Combining text and social network analysis.," *GROUP '07*, 2007.

[8] Y. Nesrine Ben, S. Narjès Bellamine Ben, and H. B. Ghezala., "Community-based collaboration recommendation to support mixed decision-making support.," *Journal of Decision Systems*, 2014.

[9] A. Bozzon, M. Brambilla, S. Ceri, M. Silvestri, and G. Vesc, "Choosing the right crowd:expert finding in social networks," *16th International Conference ACM*, 2013.

[10] E. Zhong, N. Liu, Y. Shi, and S. Rajan, "Building discriminative user profiles for large-scale content recommendation," *KDD '15*, 2015.

[11] K. Sugiyama, K. Hatano, and M. Yoshikawa., "Adaptive web search based on user profile constructed without any effort from users.," *WWW '04*, 2004.

[12] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web.," Stanford InfoLab, Technical Report, 1999.

[13] T. H. Haveliwala, "Topic-sensitive pagerank.," *11th International World Wide Web Conference (WWW2002)*, 2002.

[14] H. Chang, D. Cohn, and M. A. K., "Learning to create customized authority lists.," *17th International Conference on Machine Learning (ICML 2000)*, 2002.

[15] J. M. Kleinberg., "Authoritative sources in a hyperlinked environment.," *Journal of the ACM*, 1999.

[16] A. Sieg, B. Mobasher, and R. Burke., "Ontological user profiles for representing context in web search.," *WI-IATW '07*, 2014.

[17] S. Hashemi, M. Neshati, and H. Beigy, "Expertise retrieval in bibliographic network: A topic dominance learning approach," *22nd ACM international conference*, 2013.

[18] H. Deng, I. King, and M. Lyu, "Enhanced models for expertise retrieval using community-aware strategies," *IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics*, 2012.

[19] M. Polanyi, "The tacit dimension," *Gloucester. 1983. 107 p*, 1967.

[20] C. Smith. (2015). By the numbers: 125+ amazing linkedin statistics, [Online]. Available: http://expandedramblings.com/index.php/by-the-numbers-a-few-important-linkedin-stats/.

[21] R. Lang. (2014). Exploring the relationship between "tacit knowledge" and "explicit knowledge", [Online]. Available: https://villasophiasalon.wordpress.com/2014/01/08/exploring-the-relationship-between-tacit-knowledge-and-explicit-knowledge/.

[22] C. C. Marshall and F. M. Shipman, "Which semantic web?" *HYPERTEXT '03*, 2003.

[23] V. Collinson and T. F. Cook, *Organizational Learning: Improving Learning, Teaching, and Leading in School Systems*. SAGE Publications (CA), 2006.

[24] D. Rosnick. (2012). How much does employee turnover really cost your business? [Online]. Available: http://www.cepr.net/calculators/turnover_calc.html.

[25] D. D. Martin. (2013). Succeed as a nontraditional grad school applicant., [Online]. Available: http://www.usnews.com/education/blogs/graduate-school-road-map/2013/08/16/succeed-as-a-nontraditional-grad-school-applicant.

[26] E. Davoodi, M. Afsharchi, and K. Kianmehr, "A social-network based approach to expert recomendation system.," *Hybrid Artificial Intelligent Systems*, 2012.

[27] C. S. of Mines. (2015). Cosine similarity, [Online]. Available: http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/sphilip/cos.html.

[28] R. Documentation. (2015). Kullback-leibler distance, [Online]. Available: http://rug.mnhn.fr/seewave/HTML/MAN/kl.dist.html.

[29] L. Richardson. (2004-2015). Beautiful soup documentation, [Online]. Available: http://www.crummy.com/software/BeautifulSoup/bs4/doc/.

[30] A. W. Services. (2014). Amazon mechanical turk, api reference(api version 2014-08-15), [Online]. Available: http://docs.aws.amazon.com/AWSMechTurk/latest/AWSMturkAPI/.

[31] Z. Cheng, J. Caverlee, H. Barthwal, and V. Bachani, "Who is the barbecue king of texas?: A geo-spatial approach to finding local experts on twitter.," *SIGIR '14*, 2014.