

Personalized Expertise Search at LinkedIn

Viet Ha-Thuc*, Ganesh Venkataraman*, Mario Rodriguez*, Shakti Sinha*, Senthil Sundaram* and Lin Guo*

*Email: vhathuc,ghvenkat,mrodriguez,ssinha,ssundaram and lguo@linkedin.com

LinkedIn, 2029 Steirlin Ct, Mountain View, CA, USA

Abstract—LinkedIn is the largest professional network with more than 350 million members. As the member base increases, searching for experts becomes more and more challenging. In this paper, we propose an approach to address the problem of personalized expertise search on LinkedIn, particularly for exploratory search queries containing *skills*. In the offline phase, we introduce a collaborative filtering approach based on matrix factorization. Our approach estimates expertise scores for both the skills that members list on their profiles as well as the skills they are likely to have but do not explicitly list. In the online phase (at query time) we use expertise scores on these skills as a feature in combination with other features to rank the results. To learn the personalized ranking function, we propose a heuristic to extract training data from search logs while handling position and sample selection biases. We tested our models on two products - LinkedIn homepage and LinkedIn recruiter. A/B tests showed significant improvements in click through rates - 31% for CTR@1 for recruiter (18% for homepage) as well as downstream messages sent from search - 37% for recruiter (20% for homepage). As of writing this paper, these models serve nearly all live traffic for skills search on LinkedIn homepage as well as LinkedIn recruiter.

Keywords—personalized search; expertise scores; learning to rank

I. INTRODUCTION

The LinkedIn platform serves the professional networking needs of 350MM+ members worldwide. Members visit the site for various reasons including recruiting candidates, finding jobs, connecting with other people, reading professional content, etc. About 62% of the company revenue is from Talent Solutions¹, which is a product helping recruiters and corporations around the world find the right talent. Thus, the problem of finding candidates with expertise in some certain areas is extremely important to LinkedIn.

LinkedIn allows members to add *skills* to their profile. Typical example of skills for a software engineer would be - “Algorithms”, “Data Mining”, “Python”, etc. On LinkedIn, there are about 40 thousand standardized skills. Members can also *endorse* skills of other members in their network. Thus, skills are an integral part of members’ profiles to help them showcase their professional expertise (see Figure 1). In this paper, we focus on the problem of personalized expert ranking for queries containing one or several skills.

The area of expert finding was initiated in the field of knowledge management [1]. It became a very active research

Skills & Endorsements

Top Skills



Figure 1. Example of Skill Section with Endorsements

area since the appearance of the Enterprise Track in Text Retrieval Conference (TREC). Over the last decade, many research papers have been published in this area. However, the task of expertise search on LinkedIn has unique challenges.

- 1) **Scale:** With a member base more than 350 million and about 40 thousand skills, it is by far bigger than the benchmark data sets in the literature. For instance, TREC data set contains about 1100 candidates and 50 topics.
- 2) **Sparsity:** Members may not list all the skills they have. The system should estimate expertise scores for the skills that members have but do not explicitly list on their profiles. For example, if a member had listed “hadoop” and “java” explicitly, one could infer expertise scores for “mapreduce” as well.
- 3) **Personalization:** Expertise search on LinkedIn has a strong recruiting intent. Searchers not only care about candidates’ expertise but also many personalized aspects. These personalized aspects may include the social connection distance from the candidate, geographic location, etc.

¹<http://blog.linkedin.com/2015/04/30/linkedin-q1-2015-earnings/>

This paper describes our approach to addressing above mentioned challenges. We propose a way to estimate expertise scores offline *at scale*. Our techniques, which involve collaborative filtering and matrix factorization *infer* expertise scores for skills that members have, including the ones that they do not even explicitly list in their profiles. When a searcher enters a query containing one or several skills, our ranking system uses candidates' expertise scores on these skills as a feature in combination with other features. Our feature set contains personalized (location, social connection distance) as well as non-personalized (text matching) features. The ranking function is learned by leveraging learning-to-rank techniques. We detail practical issues faced when deploying learn-to-rank models in personalized ranking at LinkedIn scale. In particular, we detail our attempts to address position and sample selection biases while extracting training data for learning *personalized ranking* functions from search logs.

The techniques detailed in this work are used to rank a significant percentage of queries within LinkedIn. A/B tests were done on the LinkedIn homepage as well as LinkedIn recruiter search, a paid product. On the homepage search, our proposed approach improves CTR@1 by 18%, CTR@10 by 11% and number of emails sent from search page by 20%. On the premium recruiter search, our proposed approach improves CTR@1, CTR@10 and number of messages sent from search page by 31%, 18% and 37%, respectively. Note that while our examples are shown for skills related to software engineering/data mining, our approach is generic for any skills listed or inferred from our members profile. Our results are validated across different domains including (but not limited to) sales, marketing, engineering management etc. Currently, these models serve nearly **all live traffic** for skills search on LinkedIn homepage as well as LinkedIn recruiter.

The rest of the paper is organized as follows. Section 2 reviews background on expert finding and learning to rank techniques for search ranking. Section 3 presents how we estimate skill expertise scores. Section 4 details how we learn a search ranking function by using the expertise scores in combination with other personalized and non-personalized features. We discuss experimental results in Section 5. Finally, concluding remarks can be found in Section 6.

II. BACKGROUND

A. Expertise scores

Expertise scores yield an ordered list of members for every skill. In other words, it maps the tuple (member, skill) into a score. The closest match to published literature is in the area of expert finding in the field of knowledge management [1]. This became a very active research area since the appearance of TREC 2005 and 2006 [2], [3]. In the track, the task of finding experts is defined as follows:

given a collection of crawled documents including email, homepages, etc., a list of candidates and a list of topics, the task is to rank the candidates given each topic. Almost all of the systems submitted to TREC Enterprise Tracks in 2005 and 2006 and subsequent research ([4], [5], [6], [7]) view this task as a text retrieval problem. They fall into one of the two approaches: candidate-based approach which ranks candidates by textual similarity between the topics and candidates' profiles and document-based approach which first retrieves documents relevant to the topics, then discover candidates mentioned in the documents and estimate their associated scores. Besides these approaches, there is also research on combining multiple evidence associating candidates to topics [8] [9] [10]. Focusing on expert finding on social graphs, Zhang et al. [11], Rode et al. [12] propose a two-step process: (i) using language models or heuristic approaches to compute an initial expertise score for each candidate, and (ii) using graph-based algorithms to propagate scores computed in the first step and re-rank experts.

Compared to the previous research, our work on skill expertise scores is different in the following aspects. First, our system leverages the unique data source of LinkedIn including member profiles containing different sections, such as work experience, education, skills, projects, social endorsements and the high-level signals derived from the profile, such as seniority and popularity. The second unique aspect is scale. Our system estimates expertise of more than 350 million members on about 40 thousand skills, which is by far bigger than the benchmark datasets provided by TREC and others. For instance, TREC dataset contains about 1100 candidates and 50 topics. Finally, previous research typically ranks experts for one topic at a time and does not take into account topic co-occurrence patterns. Our collaborative filtering technique is able to infer expertise scores for the skills that members do not even list in their profiles based the skills they do and relationships amongst skills.

B. Learning to Rank

The second phase in our work that combines skill expertise scores with other signals into a final search ranking is related to the area of learning to rank for search. Learning to rank has been a key problem for information retrieval systems and Web search engines in particular since these systems typically use many features for ranking and that makes it very challenging to manually tune the ranking functions. There has been a lot of research published in the area of learning to rank, typically falling into one of the three following categories. Pointwise approaches view ranking as traditional binary classification or regression problems. Pairwise approaches take input as a set of pairs of documents in the form of one document is more relevant than the other with respect to a specific query. These approaches then learn a ranking functions minimizing the number of

incorrectly ordered pairs. The state-of-the-art for learning to rank is composed of listwise approaches. These approaches typically view the entire ranked list of documents as a learning instance while optimizing some objective function defined over all of the documents, such as mean average precision (MAP) or normalized discounted accumulative gain (NDCG) [13]. We refer the readers who are interested in more details of learning to rank to [14] [15] [16] [17] for more comprehensive reviews.

A key element of learning to rank is collecting ground truth data. Traditionally, ground truth data is labeled by either professional editors or crowd sourcing [14]. Given a pair of query and document, human judges give either a binary judgment (relevant or not-relevant) or graded categories, such as perfect, excellent, good, fair or bad. However, the issues with this approach are (i) It is expensive and not scalable and (ii) It is very hard for the judges to judge the relevance on behalf of some other user, making it challenging to apply the approach for personalized ranking. For these reasons, some previous research proposes to extract labeled data using click logs [18] [19] [20] [21]. Nonetheless, collecting training data from click logs presents its own challenges. We will have to find a way of handling position bias and sample selection bias. User eyes tend to scan from top results to bottom and some of the bottom results may not even be viewed. Thus, marking unclicked results that appear low in the original ranking may result in unfair penalization. Moreover, user clicks generally occur on the top ranked results and it quickly tapers down. So, the documents in the training data are a very biased sample of all results. Radlinski and Joachims [22] propose an approach called FairPair to collect labeled data from click logs and avoid the position bias. Nevertheless, the labeled data is mainly for pairwise learning to rank algorithms, not particularly suitable with the listwise ones. Second, it still suffers the sample selection bias.

III. SKILL EXPERTISE SCORES

The skill expertise algorithm infers skill expertise scores, which can be thought of as the probability that a member is an expert given a skill query containing one of more standardized skills, or $p(\text{expert}|\mathbf{q}, \text{member})$. We discuss what we mean by *expert* in section III-A.

Our approach enables us to score members for multi-skill queries, $\mathbf{q} = (\text{"java"}, \text{"data mining"}, \text{"information retrieval"})$, as well as single-skill queries, $q = (\text{"java"})$. This enables us to rank members given the context of a skill-only query for all members M , where $|M| = m$, given any arbitrary skill combination from the standardized skill set S , where $|S| = s$. We distinguish multi-skill queries from single-skill queries in our notation by bold facing the query vector.

To compute these skill expertise scores for arbitrary queries, we use a 2-step process illustrated in figure 2,

which is composed of: (1) a supervised learning step to compute $p(\text{expert}|\mathbf{q}, \text{member})$ which outputs an initial, sparse, expertise matrix E_i , of dimensions $m \times s$, and (2) a matrix factorization step that factorizes E_i into latent factor matrices, M_k and S_k of dimensions $m \times k$ and $s \times k$ respectively, which when multiplied, allow us to compute a final, denser, expertise matrix E_f . Here, k represents the size of the latent reduced-dimensionality space onto which both, members and skills, are projected, and where $k \ll m$ and $k \ll s$.

Matrix E_i is the interface between the 2 steps in the pipeline. Having this interface explicitly formalized allows for development on both pipeline stages and algorithms to be independent. We now explore the most salient pipeline components in more detail.

A. Preliminary Expertise Learner and Scorer

Expertise is the umbrella term we use to collectively refer to various attributes of members that capture relevant characteristics of those expected to rank high for a given skill query. These attributes are measured using various signals available on the site, many of which rely on complex algorithms for standardization and inference:

- Seniority: we have many ways of measuring seniority, for example, as years of experience from a common reference point (graduation), or as years of experience typically required to reach someones current job title.
- Popularity: which may include measures of number of page views received, or number of endorsements received, as well as more elaborate derived signals such as what can be obtained from running PageRank [23] on the page view graph or the endorsement graph.
- Influence: members have the ability to publish long-form content on the site, and this content is consumed by other members who may choose to further share it with their connections. Members whose content generates a lot of engagement tend to be considered influential.
- Authority: members typically belong to organizations that are hierarchical in nature, and this hierarchy is often reflect on job titles by means of qualifiers, such as *associate*, *senior*, *principal*, *manager*, *director*, or *C-level managers* (e.g. CEO, CFO, CTO).
- Desirability: members are often contacted about job opportunities that may be of interest to them, and some members may be contacted more often than others. This variability in demand from hiring managers or recruiters depicts a measure of demand for a particular set of skills, educational background, or experience.
- Relevance: we also take into account the relevance of a given standardized skill to a member's profile. This signal is derived from the same algorithm that generates recommendations for members to add skills to their profile (see [24]). Note that we differentiate

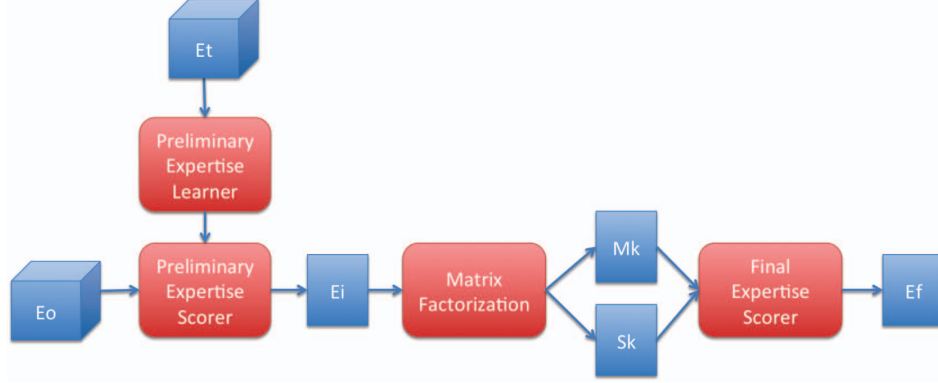


Figure 2. Skill expertise work flow. E_o , of which E_t is a subset, represents a 3-rd order tensor of dimensions $m \times s \times f$ which denotes an f -dimensional feature vector for every member-skill pair. The *Preliminary Expertise Scorer* converts the f -dimensional feature vector into an expertise score, yielding a sparse skill expertise score matrix E_i of dimensions $m \times s$. The *Matrix Factorization* component factorizes E_i into dense factor matrices M_k and S_k , which when multiplied together yield matrix E_f , a denser and final expertise matrix.

between how relevant a given skill is to a profile, vs. the expertise of the profile’s owner at the given skill. For example, the skill *java* is likely relevant to the profile of a recent college graduate out of a Computer Science program, just as it is relevant to the profile of James Gosling, the father of the Java programming language, but both have different levels of expertise in Java.

These attributes become features associated with any given (member, skill) pair, where these pairs are generated from those skills that members list explicitly on their profiles which are above a certain relevance threshold. The feature vectors associated with each (member, skill) pair are consolidated in a 3-rd order tensor of dimensions $m \times s \times f$, which is depicted as E_o in Figure 2, where F is the set of all possible features associated with a (member, skill) pair and $|F| = f$. A subset of E_o is used as training data, denoted E_t and is discussed in the next section.

1) *Training Data*: Our Preliminary Expertise Learner is based on a classifier that learns to distinguish between our positive and negative examples, or more specifically, expert (member, skill) pairs vs. non-expert (member, skill) pairs.

The positive examples are sourced from various lists of experts available online: conference speakers (e.g. Strata), open-source committers (e.g. Apache Software Foundation committers), influential author lists (e.g. LinkedIn influencers), as well as cohorts of members about whom we have an expectation of their expertise (e.g. very senior members, in-demand members, etc). Once we have identified these expert members, and are able to link them with a profile, we extract their explicitly listed skills and use various heuristics to filter their list of explicit skills into a smaller list that may be considered to be their most relevant skills. An example of such heuristic might be to filter out skills with a profile relevance score less than some threshold t . This is done to filter out potential outlier skills added in jest.

The negative examples are similarly generated based on

various strategies. Examples of non-expert (member, skill) pairs may include regular members paired with random skills, or skills that are only mildly relevant to them, as well as members with profiles tagged as *spam* or *fake* (with a certain probability), and skills associated with them.

Finally, the positive and negative examples of (member, skill) pairs are validated using crowd sourcing. The final set is then divided into three subsets: training, testing, and validation sets. Training and testing is discussed in the next section, and validation is discussed in section III-D.

2) *Learning*: We employ a logistic regression model for learning to distinguish between the expert and non-expert (member, skill) pairs using the training data described in the previous section, the primary reason for using a simple linear model being explainability.

The test set is used to evaluate the accuracy of the logistic regression model, as well as to calibrate meta-parameters, such as regularization parameters and cohort mix-proportion in the training set.

3) *Scoring*: At this point in the pipeline, we use the logistic regression model to estimate $p(e|q)$ for those standard skills which members have listed explicitly on their profiles, with the further consideration that those skills need to be highly relevant to the profile in question (so, Karaoke singing for a Director of Search Quality would likely not make the cut).

This gives us a very sparse matrix, denoted as E_i in figure 2, where most of the values in the $m \times s$ matrix are considered to be unknown. In fact, since at the time of this writing, we limit users to listing up to 50 explicit skills on their profiles, and given that we have about $\sim 40,000$ standardized skills, even if all $\sim 350\text{MM}$ members list the maximum number of skills (assuming they are all standard skills), we have a matrix with 99.9% unknown values.

B. Matrix Factorization

In the second stage of the pipeline, we aim to *discover* some of those unknown values in E_i by leveraging collaborative filtering techniques that enable us to uncover insights such as *users who are good at Statistics also tend to be good at Data Mining*.

1) *Normalization*: In order to decouple the factorization algorithm from the particulars of the supervised learning step, prior to factorization, we apply a rank-based inverse normal transformation (see [25]) to those scores in E_i which are known (as well as shift and scale the standard normal scores to be non-negative with an arbitrary mean and standard deviation).

2) *Factorization*: When factorizing the normalized E_i matrix, we sought to find the factors that optimized a loss function similar to equation 3 in [26], with a slight modification:

$$\min_{x_m, y_s} \sum_{m,s} c_{ms} (s_{ms} - x_m^T y_s)^2 + \lambda (\sum_m \|x_m\|^2 + \sum_s \|y_s\|^2) \quad (1)$$

Where the goal is to find a vector $x_m \in \mathbb{R}^k$ for each member m , and a vector $y_s \in \mathbb{R}^k$ for each skill s that will factor member-skill affinities. In other words, skill expertise scores are assumed to be the inner products: $s_{ms} = x_m^T y_s$. We refer to these vectors as the member-factors and skill-factors, respectively. This formulation is different from [26] in that we use the actual normalized expertise score in s_{ms} , rather than simply using an indicator variable denoting whether or not the member has the skill.

Note that in this formulation, similar to [26], we account for *all* (member, skill) pairs, (m, s) , rather than only those which are known. We treat the unknown scores as zeros with low confidence, and the not unknown scores as being high-confidence values. The confidence function is similar to the one used in [26]:

$$c_{m,s} = \begin{cases} \alpha, & \text{if } s_{ms} > 0 \\ 1, & \text{otherwise} \end{cases} \quad (2)$$

The exact values for λ (the regularization parameter), k (the size of the latent reduced-dimensionality space), and α (the high-confidence value for known expertise scores) were determined by cross-validation on the reconstructed matrix.

The reason for this loss function is we know that most members typically only specialize in a very small subset of the entire 40,000-dimensional skill vector space, so *most* of their scores should in fact be zero. This formulation is similar to weighted least squares regression (see [27]), where less weight is given to less precise measurements and more weight is given to more precise measurements when estimating the unknown parameters of the model.

C. Final Expertise Scorer

Having the latent representation of members and skills, we are free to reconstruct the entire expertise matrix. However, we only do so for (member, skill) pairs with high relevance scores between the member's profile and the skill as described in Section III-A

To compute the ranking for a single-skill query, we only need the expertise scores for that skill for members of interest, which is computed by using the inner product of the latent factor vectors $s_{ms} = x_m^T y_s$. To compute the ranking for a multi-skill query, $\mathbf{q} = (s_1, s_2, s_3)$, we can do the same once we have the latent factor representation of the multi-skill query. We define the latent factor representation of the multi-skill query to be a linear combination of the latent factors of each of the skills in the query, $y_{\mathbf{q}} = y_{s_1} + y_{s_2} + y_{s_3}$. This enables us to compute the expertise score for the multi-skill query for a given member m , or $s_{m\mathbf{q}} = x_m^T y_{\mathbf{q}}$.

Rather than projecting the query to the latent factor space at query-time, we can take advantage of the distributive property of vector multiplication over addition so that at query-time, we need only sum the scores for those members satisfying the retrieval requirements of a given query:

$$s_{m\mathbf{q}} = x_m^T \cdot y_{\mathbf{q}} \quad (3)$$

$$s_{m\mathbf{q}} = x_m^T \cdot (y_{s_1} + y_{s_2} + y_{s_3}) \quad (4)$$

$$s_{m\mathbf{q}} = x_m^T \cdot y_{s_1} + x_m^T \cdot y_{s_2} + x_m^T \cdot y_{s_3} \quad (5)$$

To ensure fast retrieval and scoring, the skill expertise scores (e.g. $x_m^T \cdot y_{s_1}, x_m^T \cdot y_{s_2}, x_m^T \cdot y_{s_3}$) are precomputed offline and encoded as payloads in a Lucene [28] inverted index, so that they are readily associated with each of a given member's skills. In this setting, a member ID is a document ID in the inverted index, and the member's skills are terms in that document, and associated with each (doc ID, term ID) pair is a payload-encoded expertise score.

D. Validation on Expertise Scores

For our validation step, we follow the same methodology discussed in [29], using a hold-out validation dataset from our training dataset discussed in section [III-A1], see figure [3] and table [I]. This methodology allows us to use the AUC@K for each of the cohorts to compare different versions of the skill expertise model pipeline.

Each member cohort (e.g. influencer, strata, etc) represents a set of members about whom we have an expectation of how they should rank for a given skill query. For example, given a seed member who is a Strata speaker and given the list of skills in which the seed member is considered an expert at (determined as described in section III-A1), we use his/her top-skill(s) as a skill query for which we generate a ranking of members. After we rank all members with

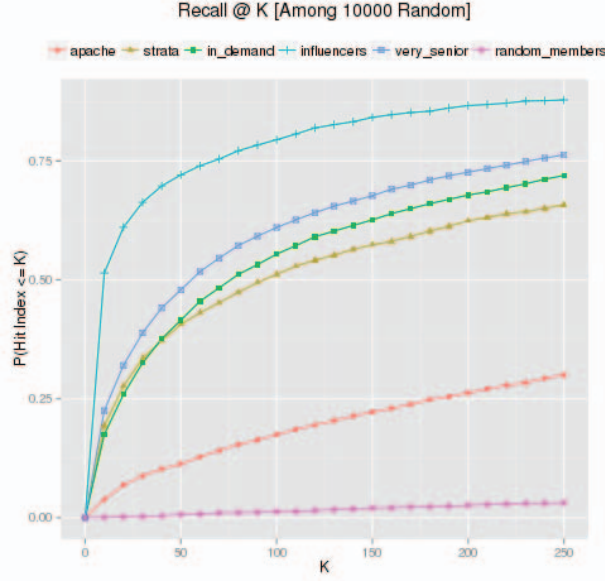


Figure 3. The plot shows the probability that a member of the specified cohort (e.g. Strata speaker) is ranked at a position $i \leq K$, as K varies from 1 to 250. We summarize the performance of the algorithm for a given cohort by using the *area under the curve* (AUC) for a given K , as shown in table [I].

Cohort	AUC@250
Influencer	0.76
Very senior	0.59
In-demand	0.53
Strata	0.50
Apache	0.19
Random	0.02

Table I
AUC @ K=250, SEE FIGURE [3].

regards to the query, we sample 10,000 relevant members, which, together with the seed member, becomes a ranked list of 10,001 for a specific skill query. This methodology is repeated many times over for each of our member cohorts, for multi-skill and single-skill queries, to generate the plot in figure [3]. The plot shows the probability that the seed member is ranked at a given K , as K varies from 1 to 250. We summarize the performance of the algorithm for a given cohort by using the *area under the curve* (AUC) for a given K , as in table [I].

An advantage of this strategy is that it is completely independent of whatever algorithm was used to compute the skill expertise scores.

IV. LEARNING EXPERTISE RANKING FUNCTION

In this section, we present how we leverage learning-to-rank algorithms to learn a ranking function for expertise search. The ranking function combines skill expertise scores described in the previous section as well as other features.

Section 4.1 gives a description on the features in the ranking function. Section 4.2 details the learning algorithm as well as the proposed approach to extract training data from search logs and how to alleviate position and sample selection biases in the log data.

A. Features

In this section, we give an overview of the features used in the ranking function for personalized expertise search. Due to business sensitivity, we cannot give details of how the features are computed. Instead, we give a high-level description. Most of the features are generally divided into the following categories.

Expertise scores As described in Section 3, we estimate an expertise score for every (member, skill) tuple. When a searcher enter a query like “java python senior developer”, we first use a tagger to extract standardized skills mentioned in the query (“java” and “python” in this example). on distributed computing platforms like Hadoop, allows processing very large datasets with complex methods e.g. matrix factorization to produce sophisticated signals like expertise scores. Detailed description of the tagger is beyond the scope of this paper. Interested reader may refer to [30] for more information. For every search result, we compute sum of his or her expertise scores on the skills.

Textual features The most traditional type of features in information retrieval are textual features (e.g. term frequency). These features are computed on different sections of user profiles, such as position titles, position descriptions, specialty, etc.

Geographic features (personalized features) Expertise search on LinkedIn is highly personalized. For instance a simple query like “software developer” from a recruiter will produce very different results when the searcher is in New York City as opposed to Sydney. Location plays an important role in personalizing the results. We created multiple features capturing this.

Social features (personalized features) Another important aspect of personalization is to capture how the results socially relate to the searcher. We leverage a variety of the signals on LinkedIn, such as common friends, companies, groups and schools to generate the features in this category.

B. Training Data

Generally, the objective of learning to rank is to find a ranking function that computes a score representing the relevance between a document and query. In case of personalized ranking in social networks, the ranking function computes relevance score for every triple of (query, document, searcher). As discussed in Section 2, a traditional way of generating labeled data by editors or crowd sourcing is not particularly suitable for personalized relevance at a large scale. Thus, we chose to collect training data using search logs. Process of collecting training data from search

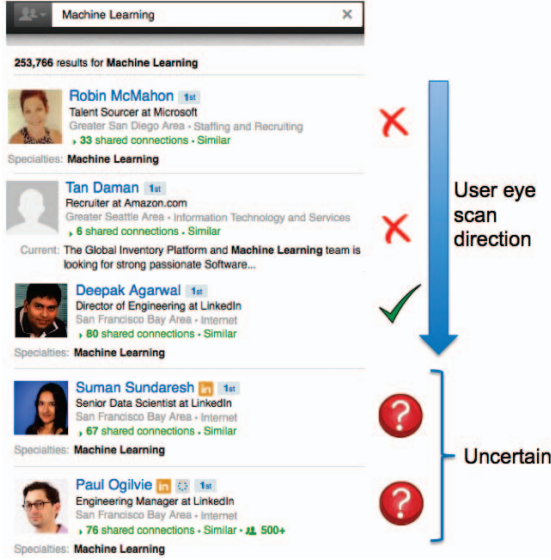


Figure 4. Typically, a searcher’s eyes scan from top results to bottom ones. For example, if the searcher skips results at positions 1 and 2 and clicks on result 3, then we assume that the clicked result is relevant and the skipped results are not. We cannot be certain about the relevance judgement for results 4-5, since the searcher may not have looked at them.

logs, however, presents its own set of challenges. We will have to handle two kinds of biases: position bias and sample selection bias.

Position bias happens when the likelihood of user actions e.g. clicks on search results is not only influenced by the relevance of the results but also their positions in the ranking [31]. A standard way to avoid the position bias is to randomize top- N results, show them to a small percentage of traffic then collect feedback. However, this approach would change the rankings completely after each reload thereby leading to bad search experience. In this work, we use a pseudo-randomization technique: we first apply a raking function to rank all of the results, then re-rank the top- N by a deterministic hash function on member Id, which is independent to result relevance. Thus, the results are stable and the top- N rankings are still orthogonal to all of the features. Given the rankings, typically searcher eyes scan from top results to bottom ones. For example, if the searcher skips results at positions 1 and 2 and interacts with result 3 (see Figure 4), then we assume that result 3 is relevant and the skipped results are not. Also note that results 4 and 5 are also shown to the searcher, but we cannot be certain if he or she looked at them and thinks they are not relevant or stops scanning at result 3. So, we simply ignore all of the results ranked below the last interacted one in the result page. We use *graded relevance* for different searcher actions. For instance, if the searcher sends a message to, clicks or ignores a result, this result has a relevance label of two (perfect), one (good) or zero (bad), respectively.

The results on which we have user feedback are usually

the top ones in the rankings. Thus, the documents in the training data is a biased sample of the overall document space. This can have adverse consequences in training. For instance, let us consider spam feature which indicates if a result is spam or not. Since this is a very important feature and the original ranking function gives a high weight for it, all of the top results are likely to be non-spam. So, when the ranking function is re-trained from the user feedback, the learner (mistakenly) learns that this feature has little discriminative power since all of the training instances have the same value. As a result, it gives a zero weight for this crucial feature. To address this issue, presumably the original ranking is not too bad in the sense that the tail results (very low ranked ones) are generally worse than the top ones, we randomly sample *easy negatives* at the tails of the rankings. These results are still in the set of retrieved documents for the query but very unlikely to get clicked if they were presented. Thus, we consider these easy negatives as bad results and include them into the training set. Introducing easy negatives into the training data reduces the skewness of feature value distributions, particularly for dominant features such as spam. Therefore, it has an effect of diversifying training pools and reducing sample selection bias.

We count every query issue as a unique search. For instance, if the query “machine learning” was issued five times, they are considered as five distinct searches. The reason for this is that even though the skill being searched for is identical across the five searches, they might have had different search contexts when considering other dimensions such as searcher or location etc. Moreover, this keeps the query distribution in the training data the same as in our live search traffic.

C. Model Training

Given a training dataset D , we apply coordinate ascent algorithm, a popular optimization algorithm to search for a solution (feature weights) \mathbf{w}^* that optimizes normalized discounted accumulative gain (NDCG) defined on the graded relevance labels as described above over D .

Without loss of generality, we assume all features are positively correlated with label (if a feature is negatively correlated with label, simply negate it). With this assumption, all of the feature weights are guaranteed to be positive, thus every solution $\mathbf{w} \in R^{+N}$, where N is the number of features, can be mapped to an equivalent one (in terms of ranking) $\lambda \in P^N$ (see Equation [6]), where P^N is a multinomial manifold as described in Equation [7]. The benefit of doing this is that all of the rank-equivalent solutions in R^{+N} is reduced to a single solution in the new space. Thus, the parameter space and as a result the number of local optima are significantly reduced.

$$\lambda_i = \frac{w_i}{\sum_{i=1}^N w_i} \quad (6)$$

$$P^N = \{\theta \in R^N : \forall i \theta_i \geq 0, \sum_{i=0}^N \theta_i = 1\} \quad (7)$$

The coordinate ascent algorithm optimizes the multi-variate non-smooth objective function above by iteratively solving multiple one-dimensional optimization problems. At a time, it searches for optimal weight for a feature while keeping the others fixed and repeats the process. Randomization technique is used during the search process to avoid getting stuck in a local optimum. This algorithm has also shown to be efficient for learning search ranking functions in some other domains [32].

After training different models (ranking functions) resulting from variety of parameter settings, we need to evaluate these models offline before deploying them in production. We use a held-out set from the training data as a test set. The models are used to re-rank the test data and evaluated by NDCG@K where K equals to 10 for LinkedIn homepage search and 25 for recruiter search (these two products will be described in the next section). Those are the numbers of results shown in the first result page of the two products. We also compare coordinate ascent algorithm with other popular LTR algorithms in the literature on the test data. It turns out that coordinate ascent algorithm achieves the best balance between performance and efficiency (scoring time) for our case. Since the paper does not focus on comparing different LTR algorithms (instead it emphasizes on practical challenges faced while deploying LTR in personalized ranking at an industrial scale), we do not show offline evaluation here to save space. Given offline performance, we take a few top models for online A/B tests in production.

V. ONLINE SEARCH EXPERIMENTAL RESULTS

In this section, we verify effectiveness of the ranking functions by using A/B tests on live search traffic. We run the A/B tests on both of LinkedIn homepage search as well as LinkedIn recruiter search. The homepage search (See Figure 5) offers every member the ability to discover people in their networks, subject to visibility rules. The recruiter search (See Figure 6) is an enterprise product where licensed recruiters use to search for candidates in the entire LinkedIn member base. Compared to recruiter search, the user base of homepage search is bigger and much more diverse. Recruiter search users are significantly more active.

Baseline: We compare the machine-learned functions described previously with the legacy ranking functions. The legacy functions for both homepage search and recruiter search use all of the features described in Section 4.1 except expertise scores. Even though the weights of the features were manually tuned, the functions have been running on live traffic for a relatively long time and have been iteratively refined. The weights were tuned over several years and several A/B tests.

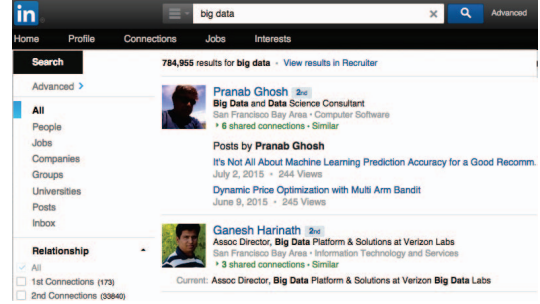


Figure 5. LinkedIn Homepage Search

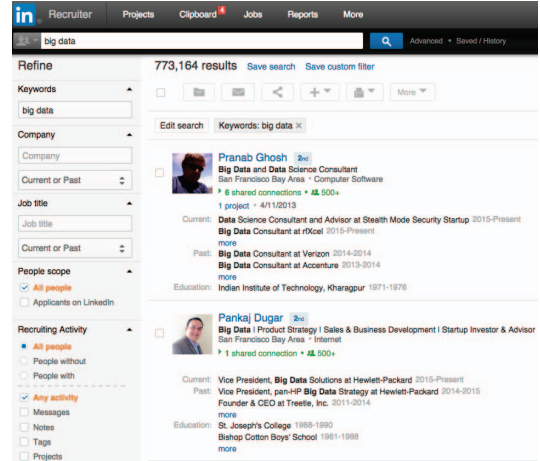


Figure 6. LinkedIn Recruiter Search

We run A/B test on both products for a reasonably long period of time (six weeks) to remove any novelty effect. When a searcher enters a query, we first use a tagger to tag query keywords with labels like *personal name (first and last name)*, *skill*, *location*, *job title*, etc. For detailed description of the tagger, we refer interested readers to [30]. Since we are interested in exploratory expertise search, we focus on queries containing at least one skill and not containing personal names. For searches satisfying this condition, the traffic is randomly split into treatment and control buckets. Searches in the treatment buckets are ranked by the machine-learned functions. Control buckets use legacy ranking functions (our baseline). During six weeks, each bucket on either product contains several hundreds of thousand skill searches.

Based on members' interactions with search results, we compute two sets of metrics. The first one is so called first-order metrics including click-through-rate at the first results (CTR@1), CTR at top ten results (CTR@10) and mean reciprocal rank (MRR). On both products, our final goal is not simply to have searchers click on results and view their profiles, but reach out to those results and eventually hire them. To partially capture this, we use a downstream metric measuring the number messages sent per search. We compare the treatment and control buckets on the metrics.

	Improvement of Treatment over Control
CTR@1	+18%
CTR@10	+11%
MRR	+14%

Table II
FIRST ORDER METRICS ON LINKEDIN HOMEPAGE SEARCH.

	Improvement of Treatment over Control
CTR@1	+31%
CTR@10	+18%
MRR	+22%

Table III
FIRST ORDER METRICS ON LINKEDIN RECRUITER SEARCH.

Table II shows first-order metrics on the homepage search of the baseline and the machine-learned model having the best offline performance. Compared to the baseline, the new ranking function improves 18%, 11% and 14% in terms of CTR@1, CTR@10 and MRR, respectively. Similarly, Table III shows that on LinkedIn recruiter search, the machine-learned ranking function is 31%, 18% and 22% better than the baseline in terms of CTR@1, CTR@10 and MRR.

Regarding to the downstream impacts of the ranking functions, we measure the average number of messages per search that searchers send to results. On LinkedIn, sending a message is typically a way to start a recruiting process. Thus, this metric is particularly important. As shown in Table IV, on homepage search, the new ranking function improves the metric by 20% over the legacy system and on recruiter search, the new ranking function improves 37%. All of the improvements on both first-order and downstream metrics are statistically significant. The significant improvements across all metrics confirm the benefit of the expertise scores and the learning approach. We further conduct additional A/B tests to verify the impact of each. The first A/B test compares the legacy system and a machine-learned model with the same set of features (i.e. without expertise feature). The second test compares machine-learned models with and without the expertise feature. The experimental results show both of the improvements are statistically significant.

Between homepage search and recruiter search, interestingly, the improvements achieved on recruiter search are higher than the ones on homepage search on every metric. This is probably because the users on recruiter search are more active thus they take advantage of the quality improvement more.

	Improvement of Treatment over Control
Homepage Search	+20%
Recruiter Search	+37%

Table IV
THE NUMBER OF DOWNSTREAM MESSAGES PER SEARCH.

VI. CONCLUSIONS

In this paper, we introduce the problem of personalized expertise search and point out technical challenges faced when building such a system at scale. We propose a scalable way to derive expertise scores over a large corpus. Our techniques take into account skill co-occurrence patterns to estimate on both explicit and inferred skills. We leverage state of the art LTR algorithms to learn final ranking functions combining expertise scores with other non-personalized and personalized features. To extract training data, we introduce a deterministic top- N randomization strategy coupled with *easy negatives* alleviating the biases in search logs.

We conducted A/B tests on both LinkedIn homepage search and recruiter search (a premium product) for a reasonably long period of time. The experimental results show substantial improvements over the legacy system. Specifically, on homepage search, the proposed approach increases 18% on CTR@1, 11% on CTR@10, 14% on MRR and 20% on downstream messages sent from search. On recruiter search, the approach achieves even bigger improvements of 31% on CTR@1, 18% on CTR@10, 22% on MRR and 37% on downstream messages. The machine-learned models are serving nearly **all live traffic** for skills search on LinkedIn.

We summarize some of our key findings below:

- For finding experts in large professional networks like LinkedIn, it is crucial to go beyond text matching. Our approach to inferring skill expertise scores yields a significant lift. In terms of normalized feature weights, the expertise scores is amongst the top four most important features in the final ranking function (due to SEO-related issues, we cannot release the actual feature values or their order of importance).
- Collaborative filtering, which has been widely used for recommendation, turns out to also work well for inferring skill expertise scores in professional networks.
- Personalization plays a crucial role in expertise search. For instance, the best geo feature (in terms of normalized feature weights) is also amongst the top four features.
- Re-ranking offline performance on the training data extracted by the deterministic top- N randomization strategy coupled with *easy negatives* is directionally inline with online performance. Between two ranking functions, the one with higher offline re-ranking performance also achieves better results on online A/B test where they are used to rank the whole sets of results. Thus, this strategy is very useful for personalized training data collection and for offline testing to make decision on which models to deploy.
- In terms of system architecture, the two-phase approach (offline and online) is a reasonable choice for production ranking systems. The offline phase runs on distributed computing platforms like Hadoop. It allows

processing very large datasets with complex methods e.g. matrix factorization to produce sophisticated signals like expertise scores. It periodically generates new versions of the signals offline. The online phase then simply consumes the latest versions of the signals in combination with other signals available in the index to generate final rankings in real time.

REFERENCES

- [1] T. H. Davenport and L. Prusak, *Working knowledge: How organizations manage what they know*. Harvard Business Press, 1998.
- [2] N. Craswell, A. P. de Vries, and I. Soboroff, "Overview of the trec 2005 enterprise track," vol. 5, pp. 199–205, 2005.
- [3] I. Soboroff, A. P. de Vries, and N. Craswell, "Overview of the trec 2006 enterprise track," in *Trec*, 2006.
- [4] K. Balog, L. Azzopardi, and M. De Rijke, "Formal models for expert finding in enterprise corpora," in *Proceedings of the 29th ACM SIGIR*, 2006, pp. 43–50.
- [5] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, and L. Si, "Expertise retrieval," *Foundations and Trends in Information Retrieval*, vol. 6, no. 2-3, pp. 127–256, 2012.
- [6] H. Fang and C. Zhai, *Probabilistic models for expert finding*. Springer, 2007.
- [7] D. Petkova and W. B. Croft, "Hierarchical language models for expert finding in enterprise corpora," *International Journal on Artificial Intelligence Tools*, vol. 17, no. 01, pp. 5–18, 2008.
- [8] C. Macdonald, D. Hannah, and I. Ounis, "High quality expertise evidence for expert search," in *Advances in Information Retrieval*. Springer, 2008, pp. 283–295.
- [9] Z. Yang, J. Tang, B. Wang, J. Guo, J. Li, and S. Chen, "Expert2bole: From expert finding to bole search," in *Proceedings of the ACM SIGKDD*, 2009, pp. 1–4.
- [10] K. Hofmann, K. Balog, T. Bogers, and M. de Rijke, "Contextual factors for finding similar experts," *JASIST*, vol. 61, no. 5, pp. 994–1014, 2010.
- [11] J. Zhang, J. Tang, and J. Li, "Expert finding in a social network," in *Advances in Databases: Concepts, Systems and Applications*. Springer, 2007, pp. 1066–1069.
- [12] H. Rode, P. Serdyukov, D. Hiemstra, and H. Zaragoza, "Entity ranking on graphs: Studies on expert finding," 2007.
- [13] Z. Cao, T. Qin, T. Liu, M. Tsai, and H. Li, "Learning to rank: from pairwise approach to listwise approach," pp. 129–136, 2007.
- [14] O. Chapelle and Y. Chang, "Yahoo! learning to rank challenge overview," pp. 1–24, 2011.
- [15] O. Chapelle, Y. Chang, and T. Liu, "Future directions in learning to rank," in *Proceedings of the Yahoo! Learning to Rank Challenge, held at ICML*, 2011, pp. 91–100.
- [16] T. Liu, T. Joachims, H. Li, and C. Zhai, "Introduction to special issue on learning to rank for information retrieval," *Inf. Retr.*, vol. 13, no. 3, pp. 197–200, 2010.
- [17] H. Li, "A short introduction to learning to rank," *IEICE Transactions*, vol. 94-D, no. 10, pp. 1854–1862, 2011.
- [18] T. Joachims, "Optimizing search engines using clickthrough data," pp. 133–142, 2002.
- [19] B. Carterette and R. Jones, "Evaluating search engines by modeling the relationship between relevance and clicks," 2007.
- [20] N. Craswell, O. Zoeter, M. J. Taylor, and B. Ramsey, "An experimental comparison of click position-bias models," in *Proceedings of ACM WSDM*, 2008, pp. 87–94.
- [21] K. Hofmann, A. Schuth, A. Bellogín, and M. de Rijke, "Effects of position bias on click-based recommender evaluation," in *Proceedings of 36th ECIR*, 2014, pp. 624–630.
- [22] F. Radlinski and T. Joachims, "Minimally invasive randomization for collecting unbiased preferences from clickthrough logs," in *Proceedings of the National Conference on Artificial Intelligence*, 2006, p. 1406.
- [23] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Technical Report 1999-66, November 1999.
- [24] M. Bastian, M. Hayes, W. Vaughan, S. Shah, P. Skomoroch, H. J. Kim, S. Uryasev, and C. Lloyd, "Linkedin skills: large-scale topic extraction and inference," in *Proceedings of the 8th ACM RecSys*, 2014, pp. 1–8.
- [25] S. R. Soloman and S. S. Sawilowsky, "Impact of rank-based normalizing transformations on the accuracy of test scores," *Journal of Modern Applied Statistical Methods*, vol. 8, 2009.
- [26] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *Proceedings of the 8th IEEE ICDM*. IEEE Computer Society, 2008, pp. 263–272.
- [27] G. D. Garson, *Weighted Least Squares Regression*. Statistical Associates Publishers, 2013.
- [28] M. McCandless, E. Hatcher, and O. Gospodnetic, *Lucene in Action*. Greenwich, CT, USA: Manning Publications Co., 2010.
- [29] P. Cremonesi, Y. Koren, and R. Turrin, "Performance of recommender algorithms on top-n recommendation tasks," in *Proceedings of the 4th ACM RecSys*. New York, NY, USA: ACM, 2010.
- [30] B. Tan and F. Peng, "Unsupervised query segmentation using generative language models and wikipedia," pp. 347–356, 2008.
- [31] T. Joachims, L. A. Granka, B. Pan, H. Hembrooke, and G. Gay, "Accurately interpreting clickthrough data as implicit feedback," in *Proceedings of the 28th ACM SIGIR*, 2005, pp. 154–161.
- [32] D. Metzler and W. B. Croft, "Linear feature-based models for information retrieval," *Inf. Retr.*, vol. 10, no. 3, pp. 257–274, 2007.