Modeling the Dynamics of Personal Expertise

Yi Fang
Department of Computer Engineering
Santa Clara University
Santa Clara, CA 95053, USA
yfang@scu.edu

Archana Godavarthy
Department of Computer Engineering
Santa Clara University
Santa Clara, CA 95053, USA
ngodavarthy@scu.edu

ABSTRACT

Personal expertise or interests often evolve over time. Despite much work on expertise retrieval in the recent years, very little work has studied the dynamics of personal expertise. In this paper, we propose a probabilistic model to characterize how people change or stick with their expertise. Specifically, three factors are taken into consideration in whether an expert will choose a new expertise area: 1) the personality of the expert in exploring new areas; 2) the similarity between the new area and the expert's current areas; 3) the popularity of the new area. These three factors are integrated into a unified generative process. A predictive language model is derived to estimate the distribution of the expert's words in her future publications. In addition, KL divergence is defined on the predictive language model to quantify and forecast the change of expertise. We conduct the experiments on a testbed of academic publications and the initial results demonstrate the effectiveness of the proposed approach.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Expertise retrieval; Expertise profiling; Temporal change

1. INTRODUCTION

People often change their expertise or interests over time. Capturing how personal expertise evolves can better characterize expert profiles and thus facilitate the task of expertise retrieval which is an important field in Information retrieval [2]. Various factors can affect the dynamics of personal expertise. For example, new and emerging technologies may make the existing ones obsolete and consequently people have to adapt their skills and expertise. Moreover, some people may have worked on the similar fields and thus they may be more likely to move to the new area than others who do not have prior background. On the other hand, the change of expertise is highly personal. Some people may always explore new areas and skills regardless of their prior areas while some other people may stay with the same expertise all the time.

We focus on modeling the dynamics of personal expertise. The task is closely related to expertise retrieval which has been extensively studied in the IR community. However, there exists very little work on investigating the dynamic aspect of expertise. Furthermore, to the best of our knowledge, the existing literature contains no study on predicting personal expertise. In this paper, we propose a probabilistic model to characterize how people change or stick with their expertise. Three factors are taken into consideration: personality, similarity, and popularity. The proposed model can be used to predict what are the next expertise areas a given expert will work on, what words the expert is likely to use in her next paper, and whether there will be a big change in her expertise areas. We conduct the experiments on a testbed of academic publications and the initial results demonstrate that our proposed model can achieve much improved predictive performance over the baselines.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for prof t or commercial advantage and that copies bear this notice and the full citation on the f rst page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specif c permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.
Copyright 2014 ACM 978-1-4503-2257-7/14/07 ...$15.00.
http://dx.doi.org/10.1145/2600428.2609521.

2. RELATED WORK

The existing work on expert profiling has largely focused on finding and ranking topics for a given expert [1, 3, 8]. Very little work in the literature has investigated the temporal and dynamic aspects of expertise. Daud [4] proposed a topic modeling approach called Temporal-Author-Topic (TAT) to simultaneously model text, researchers and time of research papers, but their focus was on discovering topically related researchers for different time periods. Hoonlor [5] investigated the overall trends in computer science research without zooming in the individual researchers. To the best of our knowledge, the closest work to ours is the recent work by Rybak et al. [7]. They studied the temporal expertise profiling task by proposing hierarchical expertise profiles in which topical areas were organized in a taxonomy. They further detected changes in a person's profile based on snapshots of hierarchical profiles taken at various time intervals. Our work differs from theirs in a number of important ways. We explicitly model the factors that affect the change of expertise. The analysis and diagnosis of these factors can help us gain a better understanding of expertise dynamics. Consequently, the proposed model can predict the new topical areas the experts are likely to work on.

3. MODEL

In this paper, we introduce the proposed model in the context of academic publications for the sake of presentation convenience, but the model can be easily adapted to other expert profiling scenarios. In this setting, researchers are experts, and their publications indicate their expertise. In the experiments, we use the Keywords in the publications to define the authors' topical expertise areas. All the publications associated with a given area define the area.

To understand the dynamics of personal expertise, we need to look at how an expert's areas evolve over time. We make the Markov assumption by assuming the expert's expertise areas in year $t+1$, denoted by a_{t+1}, only depends on her areas in year t, denoted by a_t, not depends on the years prior to t. This is a reasonable assumption made by many temporal probabilistic models. Therefore, based on the sum rule and product rule of probabilities, the probability that the expert e will work on the area a_{t+1} in year $t+1$, denoted by $P(a_{t+1}|e)$, is

$$P(a_{t+1}|e) = \sum_{a_t} P(a_{t+1}|a_t, e)P(a_t|e) \quad (1)$$

where $P(a_t|e)$ is the probability that expert e's current area in year t is a_t, and $P(a_{t+1}|a_t, e)$ is the probability that she will work on the area a_{t+1} given her current area a_t. The estimation of $P(a_t|e)$ can be based on the relative frequency of a_t in the expert's publications in year t. Specifically, $P(a_t|e) = \frac{N_{a_t,e}}{N_{e,t}}$, where $N_{a_t,e}$ is the number of times that a_t occurs in e's publications in year t, and $N_{e,t}$ is the total number of times that any topical area occurs in e's publications in t.

The estimation of $P(a_{t+1}|a_t, e)$ is the central component of the proposed model which characterizes how the expert e chooses the next area a_{t+1} given the current one a_t. As discussed in Section 1, we consider three factors to estimate $P(a_{t+1}|a_t, e)$: 1) the personality of the expert in exploring a new area (or the conservativeness to stay in the current areas); 2) the similarity between the new area and the expert's current areas; 3) the popularity of the new area. The subsections below present the modeling of individual factors in detail.

3.1 Conservativeness of an Expert

We can model how the expert e chooses the next area a_{t+1} given the current a_t as the following generative process.

1. Choose to stay in the current expertise area set A_t with the probability $P(a_{t+1} \in A_t|a_t, e)$
   - Choose a specific topical area a_{t+1} from A_t with the probability $P(a_{t+1}|a_t, e, a_{t+1} \in A_t)$
2. Choose to change to the new area set $\bar{A}_t$ with the probability $1 - P(a_{t+1} \in A_t|a_t, e)$
   - Choose a specific topical area a_{t+1} from $\bar{A}_t$ with the probability $P(a_{t+1}|a_t, e, a_{t+1} \in \bar{A}_t)$

The above generative process can also be viewed as a two-class mixture model. Thus, $P(a_{t+1}|a_t, e)$ can be decomposed as

$$\begin{aligned} P(a_{t+1}|a_t, e) &= P(a_{t+1} \in A_t|a_t, e)P(a_{t+1}|a_t, e, a_{t+1} \in A_t) \\ &+ P(a_{t+1} \in \bar{A}_t|a_t, e)P(a_{t+1}|a_t, e, a_{t+1} \in \bar{A}_t) \end{aligned}$$

where $P(a_{t+1} \in A_t|a_t, e)$ is used to model the conservativeness of an expert, i.e., the tendency to stick with the current areas. We can look at how frequently the expert changes her areas in the previous years to estimate it, specifically as follows

$$P(a_{t+1} \in A_t|a_t, e) = \frac{1}{T}\sum_{t=1}^{T} J(A_t, A_{t-1}) = \frac{1}{T}\sum_{t=1}^{T} \frac{|A_t \cap A_{t-1}|}{|A_t \cup A_{t-1}|}$$

where $J(A_t, A_{t-1})$ measures the Jaccard similarity between two sets A_t and A_{t-1}. The intuition is that if the expert is conservative, her current area set A_t and the previous area set A_{t-1} will have a big overlap which leads to a high Jaccard similarity. By definition, the Jaccard similarity is always between 0 and 1. We use the average Jaccard similarity over T years to estimate $P(a_{t+1} \in A_t|a_t, e)$. Consequently, the probability that e will choose a different area in year $t+1$ is simply

$$P(a_{t+1} \in \bar{A}_t|a_t, e) = 1 - P(a_{t+1} \in A_t|a_t, e)$$

Let us then consider the specific area the expert will choose in year $t+1$ given the two different cases. If e stays in the current areas, i.e., $a_{t+1} \in A_t$, the probability that she will select a_{t+1} may be proportional to the frequency of a_{t+1} in e's current publications. Specifically,

$$P(a_{t+1}|a_t, e, a_{t+1} \in A_t) = \frac{N_{a_{t+1},e,t}}{N_{e,t}}$$

where $N_{a_{t+1},e,t}$ is the frequency of the area a_{t+1} occurring in e's publications in year t. $N_{e,t}$ is the total number of times that any topical area occurs in e's publications in t.

3.2 Similarity and Popularity of a New Area

If e chooses a new area, i.e., $a_{t+1} \in \bar{A}_t$, we will consider the similarity between the new area and the current area, and the popularity of the new area. Specifically,

$$\begin{aligned} & P(a_{t+1}|a_t, e, a_{t+1} \in \bar{A}_t) \\ = \; & \beta \times Sim(a_{t+1}, a_t) + (1-\beta) \times Pop(a_{t+1}) \end{aligned}$$

where $Sim(a_{t+1}, a_t)$ measures the similarity between two areas a_{t+1} and a_t. The intuition is based on the fact that if a_{t+1} is more similar with the expert's current area a_t, she is more likely to explore the area a_{t+1}. $Pop(a_{t+1})$ measures the popularity of a_{t+1}. The idea is that if an expertise area gets popular or trendy, people are likely to move to the area regardless of their prior background (e.g., a_t). Both $Sim(a_{t+1}, a_t)$ and $Pop(a_{t+1})$ are normalized probabilities and can be calculated as follows. β is a parameter to trade off between the two probabilities and can be estimated by cross validation. In the experiments, we choose $\beta = 0.5$ by treating the two factors equally.

$$Sim(a_{t+1}, a_t) = \frac{cosine(a_{t+1}, a_t)}{\sum_{a_{t+1}} cosine(a_{t+1}, a_t)} \quad (2)$$

$$Pop(a_{t+1}) = \frac{N_{a_{t+1},t}}{N_t} \quad (3)$$

where $cosine(a_{t+1}, a_t)$ calculates the cosine similarity between the two areas. In the experiments, we aggregate all the abstracts that co-occur with each area into one document. Thus, to estimate the similarity between two areas, we can compute the cosine similarity between the two documents associated with the two areas. The popularity

$Pop(a_{t+1})$ of area a_{t+1} is calculated based on the relative frequency of the area appearing in year t.

3.3 Predictive Language Model

In the previous subsections, we obtain the predicted probability $P(a_{t+1}|e)$ for e over topical areas a_{t+1} (in year $t+1$). Based on it, we can further estimate e's probability over words in $t+1$ as follows

$$P(w_{t+1}|e) = \sum_{a_{t+1}} P(w_{t+1}|a_{t+1})P(a_{t+1}|e) \quad (4)$$

where $P(w_{t+1}|a_{t+1})$ is the probability over words given the topic a_{t+1}. Here we have conditional independence assumption of w_{t+1} and e given a_{t+1}. This can be viewed as a generative process in which expert chooses a topic a_{t+1} with probability $P(a_{t+1}|e)$ and then generates a word w_{t+1} from the topic a_{t+1} with probability $P(w_{t+1}|a_{t+1})$. We can use $P(w_{t+1}|e)$ to predict what are the next words the expert e is likely to use, and thus the above equation defines a predictive language model (PLM) for e. This model is especially useful when the topical areas are not directly observed. For example, in the experiments we use the Keywords in ACM's publications to identify experts' topical areas, but many papers do not include the Keywords. e.g., those not published by ACM. Therefore, we evaluate the proposed approach based on the observed words w_{t+1} instead of the areas a_{t+1}. We can still estimate $P(a_{t+1}|e)$ via Eqn. 1 by using the observed areas/keywords of the expert.

3.4 Predictive Expertise Change Detection

Based on the predictive language model $P(w_{t+1}|e)$ in Section 3.3, we can predict to what extent the expert will change her areas for the next year. Specifically, we can achieve this by quantifying the difference between $P(w_{t+1}|e)$ and $P(w_t|e)$ using Kullback-Leibler (KL) divergence [6] as follows

$$KL(P_{t+1}(w|e)||P_t(w|e)) = \sum_w P_{t+1}(w|e) \log \frac{P_{t+1}(w|e)}{P_t(w|e)} \quad (5)$$

where $P_{t+1}(w|e) = P(w_{t+1}|e)$ is the predictive language model for year $t+1$, and $P_t(w|e) = P(w_t|e)$ is the current language model estimated based on the observed words in year t. KL divergence is a natural and well studied "distance" measure between two distributions. We can utilize it to detect and forecast potential changes in expertise, which could improve expertise retrieval and better understand the dynamics of expertise.

4. EXPERIMENTS

4.1 Data Collection

We created a testbed of expert profiles in the IR academic community. Specifically, we retrieved 500 researchers on "information retrieval" from ArnetMiner[1], a publicly available academic researcher website. From ArnetMiner, we could collect the titles, venues, and years of the publications associated with the researchers. Totally, we obtained 25,255 paper titles. Instead of using the full-text publications, we utilized the abstracts of publications as expertise evidence of the authors since abstracts are concise summaries of the publications. We crawled the abstracts from Google Scholar[2]. Each paper published by ACM also includes a set of keywords that authors use to specify the topics of the paper. We treated these keywords as the topical areas of the authors, and all the abstracts co-occurring with a given keyword were aggregated to define the topical area/keyword. We obtained the keywords of the papers from from ACM Digital Library[3]. A total number of 1,726 keywords were extracted for our dataset. Although ACM publications also include Categories and Subject Descriptors devised by ACM Computing Classification System (ACC), we think they may be too coarse and sometimes cannot accurately reflect authors' specific expertise.

We use the data in year t for training and the data in year $t+1$ for testing. In training, only the papers with any keywords observed are used, while all the papers (in year $t+1$) are used in testing. Except in Section 4.2, the time interval between t and $t+1$ is assumed to be 1 year. In Section 4.2, we vary the time interval from 1 year up to 5 years to investigate how the time interval affects the model performance (e.g., the 3-year interval means using year 2008, 2009, and 2010 for training, and 2011, 2012, and 2013 for testing). We use Lucene 4.3 for indexing with *EnglishAnalyzer*, and a standard list of the stop words are removed and Porter stemming is applied.

4.2 PLM vs Baselines in Perplexity

In this experiment, we use perplexity as the evaluation metric for our proposed predictive language model (PLM) and baseline methods. Perplexity is a quantitative measure for comparing language models. The value of perplexity reflects the ability of a language model to generalize to unseen data. A lower perplexity score indicates better predictive performance. We calculate the average perplexity across all the researchers as follows:

$$perplexity(D_{t+1}) = \frac{1}{N_e} \sum_e \exp - \frac{\sum_{w_{t+1,e} \in D_{t+1,e}} \log \left(P(w_{t+1,e}|e)\right)}{|D_{t+1,e}|}$$

where $w_{t+1,e}$ denotes a word observed in e's publications in $t+1$, and $D_{t+1,e}$ is the set that includes all such words for e. D_{t+1} is the union set that includes $D_{t+1,e}$ for all the experts. N_e is the total number of experts of interest. $P(w_{t+1,e}|e)$ is the language model to be evaluated. For the words only appearing in the test set D_{t+1} but not in the training data, the out-of-vocabulary (OOV) words, we assign them the probability $\frac{1}{|V_{D_t}|}$ where $|V_{D_t}|$ is the size of vocabulary of training corpus.

We compare our PLM model with three baselines with different ways to estimate $P(a_{t+1}|e)$. Base 1 does not consider any change in research areas from a_t to a_{t+1}, and thus assumes $P_{base1}(a_{t+1}|e) = P(a_t|e)$ where $P(a_t|e)$ is the relative frequency of a_t (or a_{t+1}) in e's publications in year t. Base 2 assumes the probability of choosing a_{t+1} is proportional to the similarity between a_t and a_{t+1}, i.e., $P_{base2}(a_{t+1}|e) = \sum_{a_t} sim(a_{t+1}, a_t)P(a_t|e)$ where sim is calculated as Eqn. (2). Base 3 only considers the popularity of the area and has $P_{base3}(a_{t+1}|e) = Pop(a_{t+1})$ where $Pop(a_{t+1})$ is defined in Eqn. (3). Table 1 shows the perplexity results for the

[1]http://arnetminer.org/search/1392615977909?q=information$0020retrieval

[2]http://scholar.google.com/

[3]http://dl.acm.org/

Table 1: Perplexities for the baselines and our proposed model (PLM) over various time intervals.

	1-year	2-year	3-year	4-year	5-year
Base 1	4590.7	2850.2	3430.4	4153.4	3428.4
Base 2	4598.2	2816.8	3386.8	4104.9	3351.4
Base 3	4580.9	2927.9	3611.4	4208.0	3662.3
PLM	960.1	809.2	867.2	943.2	1003.6

baselines and our proposed model (PLM) over various time intervals.

From the results, we can see that our proposed PLM model has substantial improvement over the baselines across all the time intervals. The baseline methods yield similar performance with each other. All the models obtain the best results on the 2-year time interval, which may indicate researchers may often shift their research interests every two years.

4.3 Predicted Topical Areas

In this section, we apply the proposed model in Eqn. 1 to calculate the probability $P(a_{t+1}|e)$ that e will choose area a_{t+1} in year 2013. Based on the descending order of this probability, we can predict the top areas for e in 2013. Table 2 lists the top 3 areas for three researchers, along with the researchers' top areas in 2012 calculated based on the frequency of the areas in theirs publications. Again, the areas are extracted and defined by the keywords in ACM publications (see Section 4.1).

From the results, we can see that the proposed model can predict different top areas for 2013 from 2012, while still reserving the areas of strength of the researchers. For example, Dawei Song's top areas in 2012 did not include "Query expansion", but our model gave such a prediction for 2013. From the titles of his publications in DBLP [4], we can verify that at least two of Dawei Song's papers in 2013 are related to the area "Query expansion". On the other hand, our model still kept some of his previous top areas such as "Query suggestions" in the forecast, which was a good prediction since his publications in 2013 indeed had quite a few about this topic. We can also have some similar findings for other two researchers in the table.

4.4 Predicting Large Changes in Expertise Areas

We rank all the researchers based on the KL divergence, defined in Section 3.4, from 2012 to 2013. The researchers with large KL divergence are predicted to have a relatively considerable change in their topical areas in the new year. Figure 1 shows the top 5 researchers with the largest predicted KL divergence from 2012 and 2013. To see whether the results make sense, we dig into the researchers' previous publications, and find out that they all have worked on a quite diverse set of research areas. For example, Kotagiri Ramamohanarao's fields include Machine learning, Data mining, Information retrieval, Logic programming, Distributed systems, Bioinformatics, Medical imaging, and so on. The frequent change of areas would lead to low value in $P(a_{t+1} \in A_t|a_t, e)$ (see Section 3.1, which is based on Jaccard similarity between two area sets A_t and A_{t+1}), and thus would be expected to have low probability to stick with the current areas.

[4]http://www.informatik.uni-trier.de/ ley/pers/hd/s/Song:Dawei

Table 2: Top areas in 2012 and predicted top areas in 2013 for three researchers

Researcher	Keywords in 2012	Predicted Keywords in 2013
Dawei Song	Concept hierarchy Log analysis Query suggestions	Query expansion Search log mining Query suggestions
Yoelle S. Maarek	Community QA Web retrieval Query analysis	Community QA User interaction Query analysis
Clement T. Yu	Coreference resolution Topic model Diversionary comments	Blog retrieval Faceted blog distillation Subjectivity

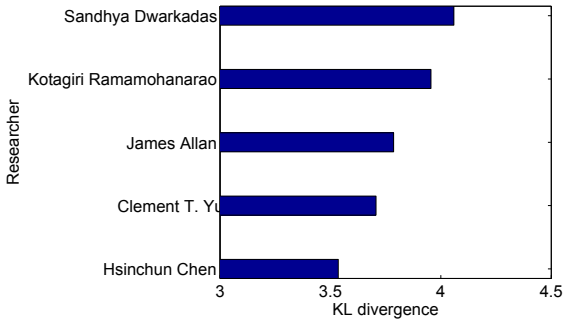

Figure 1: The 5 researchers with the largest predicted KL divergence from 2012 and 2013

5. CONCLUSIONS AND FUTURE WORK

This paper proposes a novel probabilistic model to investigate and predict how personal expertise evolves over time. The model considers the personality of a given expert, the similarity between new areas and the current ones, and also popularity of new areas. The experimental results demonstrated the effectiveness of the proposed model. This work is an initial step towards a promising researcher direction. In the future work, we will explore more factors in modeling expertise dynamics. We will also conduct a more comprehensive set of experiments to evaluate the proposed model on large-scale testbeds.

6. REFERENCES

[1] K. Balog and M. De Rijke. Determining expert profiles (with an application to expert finding). In *IJCAI*, 2007.

[2] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, L. Si, et al. Expertise retrieval. *FnTIR*, 2012.

[3] R. Berendsen, M. Rijke, K. Balog, T. Bogers, and A. Bosch. On the assessment of expertise profiles. *JASIST*, 2013.

[4] A. Daud. Using time topic modeling for semantics-based dynamic research interest finding. *Knowledge-Based Systems*, 2012.

[5] A. Hoonlor, B. K. Szymanski, and M. J. Zaki. Trends in computer science research. *Communications of the ACM*, 2013.

[6] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *SIGIR*, 2001.

[7] J. Rybak, K. Balog, and K. Nørvåg. Temporal expertise profiling. In *ECIR*, 2014.

[8] P. Serdyukov, M. Taylor, V. Vinay, M. Richardson, and R. W. White. Automatic people tagging for expertise profiling in the enterprise. In *ECIR*, 2011.