Temporal Expertise Profiling

Jan Rybak[1], Krisztian Balog[2], and Kjetil Nørvåg[1]

[1] Norwegian University of Science and Technology, Trondheim, Norway
[2] University of Stavanger, Stavanger, Norway
jan.rybak@idi.ntnu.no, krisztian.balog@uis.no,
kjetil.norvag@idi.ntnu.no

Abstract. We introduce the temporal expertise profiling task: identifying the skills and knowledge of an individual and tracking how they change over time. To be able to capture and distinguish meaningful changes, we propose the concept of a hierarchical expertise profile, where topical areas are organized in a taxonomy. Snapshots of hierarchical profiles are then taken at regular time intervals. Further, we develop methods for detecting and characterizing changes in a person's profile, such as, switching the main field of research or narrowing/broadening the topics of research. Initial results demonstrate the potential of our approach.

1 Introduction

Expertise retrieval refers to the general area of linking humans to knowledge areas, and vice versa [2]. Thanks to the increasing amount of information available online that can be traced and mined for evidence of expertise, there has been a great deal of work in this area within the IR community over the past decade. Specifically, two main expertise retrieval tasks have been investigated: expert finding (*"Who are the experts on topic X?"*) and expert profiling (*"What topics does person Y know about?"*), where the former received considerably more attention than the latter.

In this paper, we focus on the expert profiling task with the ultimate goal of identifying and characterizing changes in expertise of individuals over time. To the best of our knowledge, we are the first to propose this task. To be able to capture and distinguish meaningful changes, we first introduce the concept of a *hierarchical expertise profile*, where topical areas are organized in a taxonomy and expertise is represented as a weighted tree (Section 3.1). *Temporal expertise profile* is then defined as a series of timestamped hierarchical profiles (Section 3.2). Next, we develop methods for detecting and characterizing changes in a person's profile. The core idea of our approach is the identification of so-called *focus nodes*: a single node or small set of nodes that accumulate the majority of the node weights, with respect to a given parent node (Section 4.1). A change occurs if there is a difference in the set of focus nodes between two points in time; the change is then interpreted depending on which level of the topic hierarchy is affected (Section 4.2). We illustrate our approach for a selected person (Section 5).

2 Related Work

Existing work on expert profiling has primarily focused on identifying [5] and ranking [1, 3] topics for a given expert. De Rijke et al. [4] consider hierarchical profiles for

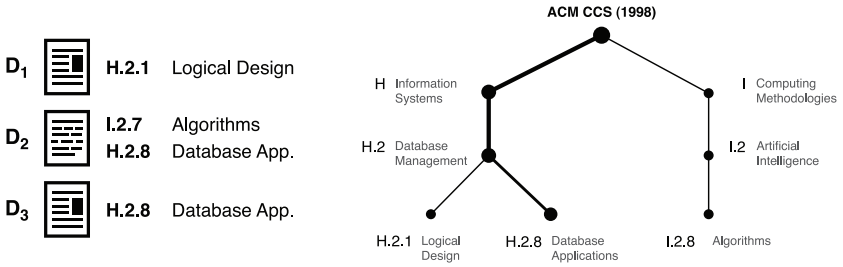

Fig. 1. Example hierarchical expertise profile, constructed from documents shown on the left. Node sizes are set proportional their weight (note that edges are not weighted according to our definition, thickness is only applied here for presentation purposes).

the more general task of entity profiling, however, their work concentrates on evaluation aspects and not on the actual construction of such profiles. Berendsen et al. [3] provide a critical assessment and analysis for the evaluation of expert-profiling systems. Sun et al. [6] present the BibNetMiner system, a system for visualizing bibliographic databases, with a focus on clustering and ranking of conferences. This is subsequently used for author, venue, and research area profiling. None of these works consider the temporal aspects of expertise. Tsatsaronis et al. [7] study the evolution of power graphs for authors over time, based on co-authorship information, the volume of published papers, and impact factors of the respective venues. Albeit they consider temporal aspects, the focus is on classifying authors into 4 predefined types, and not on topical expertise.

3 Temporal Expertise Profile

The purpose of expert profiling is to answer the following question: *"What topics does a person know about?"* In [1] the topical profile of an individual is defined as "a record of the types and areas of skills and knowledge of that individual, together with an identification of levels of 'competency' in each." Based on this definition, we extend the notion of a topical profile to a hierarchical case, where topical areas are not treated as a flat list, but are organized in a taxonomy (where parent-child relationships between topics define a hierarchy.) We represent a person's *hierarchical expertise profile* as a weighted tree, where the weights on the nodes reflect the person's expertise in the given topic. Finally, we define *temporal expertise profiles* as a series of timestamped hierarchical profiles. This allows us to track changes in a person's expertise over time.

3.1 Hierarchical Expertise Profile

The *hierarchical expertise profile* of person a is defined as a weighted tree $T_a = (C, E)$ where tree nodes $C = \{c_1, \ldots, c_n\}$ represent topics and edges E represent hierarchical relationships between topics. We write $e(c_i, c_j)$ to denote that c_j is a sub-topic of c_i. The weights on nodes $\{w_1, \ldots, w_n\}$ indicate the person's expertise on the corresponding topics. We assume that we are given some taxonomy that defines the topics and their hierarchical relationships (such as the ACM Computing Classification System that we will use in our experiments), that is, C and E. Our task, then, is to estimate the node weights $\{w_1, \ldots, w_n\}$.

Following prior work in expertise retrieval [2], we estimate expertise based on a set of documents authored by the person, denoted as D_a. We make the simplifying assumption here that each of these documents d is labeled with one or more leaf-level nodes from the topical taxonomy, d_C. We use the probability $P(c|d)$ to express whether document d belongs to category c. Most documents have a single category assigned to them, but in case there are multiple ones, we distribute the weight evenly across them. Therefore, we set $P(c|d)$ to $1/|d_C|$ if $c \in d_C$ and otherwise set it to 0. It is important to emphasize that we compute direct expertise estimates for leaf nodes only. Formally,

$$w_i = \sum_{d \in D_a} P(c_i|d)P(d). \tag{1}$$

The formula also includes a document prior $P(d)$ which can be used to express the importance of documents; for example, one could assign more importance to articles published at top-tier venues (following the intuition that these might constitute stronger evidence of expertise). However, we leave that to future work and set $P(d) = 1$ for all documents. Note that while we use probabilities in Eq. 1, w_i is not a probability; it is simply a weighted sum of publications that are labeled with a given topic.

For non-leaf nodes we sum up the weights of direct descendants:

$$w_i = \sum_{\{c_j|e(c_i,c_j)\}} w_j. \tag{2}$$

Effectively, weights are calculated in a bottom-up fashion, starting with the leaf nodes and then propagating weights to the upper levels until the root of the tree is reached. An example hierarchical profile is displayed on Figure 1.

3.2 Temporal Expertise Profile

We define the *temporal expertise profile* of a person as a series of hierarchical expertise profiles $\mathcal{T}_a = \{T_a^{t_1}, \ldots, T_a^{t_m}\}$ computed at different points in time, $t_1, \ldots, t_m$. We refer to $T_a^{t_j}$ as the *profile snapshot* taken an t_j. In this work, we assume regular time intervals, but our approach could also be applied to non-regular intervals.

We estimate the weights for leaf nodes $w_i^{t_j}$ for the profile snapshot $T_a^{t_j}$ as a mixture of two components: (1) expertise acquired in the corresponding time period (i.e., in $(t_{i-1}, t_i]$) based on the authored documents, and (2) expertise "carried over" from the past. The first component is the same as in Eq. 1, the only difference being that we restrict ourselves to documents originating from the given time period. As for (2), we use a decay function τ to capture the notion of expertise "fading away" over time.

$$w_i^{t_j} = \lambda \sum_{\substack{d \in D_a \\ d \in (t_{j-1}, t_j]}} P(c_i|d)P(d) + (1-\lambda) \sum_{k=1}^{j-1} \tau(j-k) w_i^{t_k}. \tag{3}$$

This might also be viewed as "smoothing with the past" controlled by parameter λ; for the sake of simplicity, λ is set to 0.5 in our experiments. There are many possibilities for setting the decay function $\tau(t)$, where t denotes the distance in time. We employ linear decay based on time distance with two additional constraints: (1) distances below δ_b are still considered "the present" where there is no decay applied, i.e., $\tau(t) = 1$ if

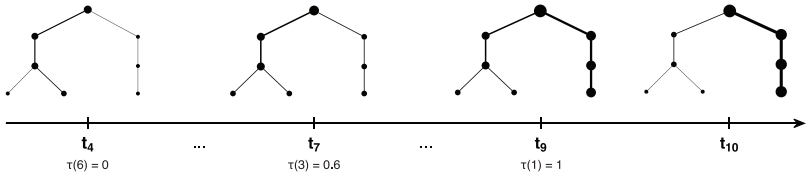

Fig. 2. Example temporal expertise profile. Decay function values $\tau(t)$ are displayed with respect to distances from the rightmost node.

$t \leq \delta_b$; (2) distances beyond δ_e are considered "distant past" that does not have any influence anymore, i.e., $\tau(t) = 0$ if $t \geq \delta_e$. For $\delta_b < t < \delta_e$ a linear decay is applied: $\tau(t) = \frac{\delta_e - t}{\delta_e - \delta_b}$. In our experiments we create profiles at yearly regularity and set $\delta_b = 1$ and $\delta_e = 6$. Figure 2 shows an example of a temporal profile with these settings.

It is important to note that the computations described above are applied only to leaf nodes. The weights for non-leaf nodes are calculated as before, i.e., according to Eq. 2.

4 Detecting Changes

Our general strategy for identifying and characterizing changes in temporal expertise profiles works as follows. First, we pin down a single node or small set of nodes that accumulate the majority of the node weights, with respect to a given parent node c_p, called the *set of focus nodes* $F_p^{t_i}$. This is done for each profile snapshot, where t_i in the superscript indicates the timestamp. Next, we say that a change has occurred if there is a difference in the set of focus nodes between two timestamps t_i and t_j, that is, $F_p^{t_i} \neq F_p^{t_j}$. Finally, we characterize the change based on how exactly $F_p^{t_i}$ and $F_p^{t_j}$ differ; the interpretation depends on the parent node's placement in the topic hierarchy. Specifically, we distinguish between changes depending on whether the top level or lower levels of the topic hierarchy are concerned.

4.1 Identifying Focus

We identify the set of focus nodes with respect to a given parent node as follows. First, we rank nodes by their weight (that is, the node with the highest weight comes first) and set the focus nodes to an empty set. Then, we add nodes iteratively, in a rank-based order, until the weight accumulated in the set, relative to the total weight (that is, the parent node's weight), reaches a threshold. Algorithm 1 details our method.

Algorithm 1 FINDFOCUS identifies the set of focus nodes.

Require: profile snapshot T^{t_i}, parent node c_p, weight threshold $\eta \in [0, 1]$
Ensure: F
1: $F \leftarrow \emptyset, W_F \leftarrow 0$
2: $r \leftarrow sort(\{c_i | e(c_p, c_i)\})$ ▷ r holds indices of nodes sorted by their weight
3: $i \leftarrow 0$
4: **while** $W_F < \eta \cdot w_p$ **do**
5: $F \leftarrow F \bigcup \{c_{r[i]}\}, W_F \leftarrow W_F + w_{r[i]}$
6: $i \leftarrow i + 1$
7: **end while**

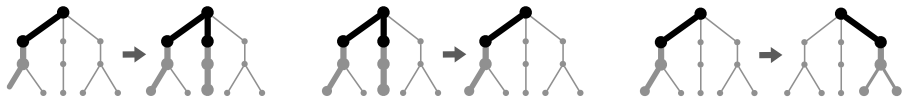

Fig. 3. Changes in the field of research (i.e., top-level nodes). (Left): leaving field, (Middle): moving into field, (Right): switching field.

Fig. 4. Changes in the topic of research (i.e., lower level nodes). (Left): narrowing topics, (Middle): broadening topics, (Right): topic switch.

4.2 Characterizing Changes

We distinguish between three types of changes, depending on how $F_p^{t_i}$ and $F_p^{t_j}$ differ:

- **F-** One of the focus nodes is removed: $|F_p^{t_i}| > 1, \exists c_k : c_k \in F_p^{t_i} \wedge c_k \notin F_p^{t_j}$.
- **F+** New focus node is added: $|F_p^{t_j}| > 1, \exists c_k : c_k \in F_p^{t_j} \wedge c_k \notin F_p^{t_i}$.
- **Fx** Exchanging a single focus node for another: $|F_p^{t_i}| = |F_p^{t_j}| = 1, F_p^{t_i} \neq F_p^{t_j}$.

Changes in the field of research. The top-level nodes of the hierarchy correspond to the *fields of research*. Focus detection performed on the root of the topic tree (as the parent node), therefore, results in the main fields of research of the person. Specific changes are interpreted as follows (see Figure 3 for an illustration).

- **Leaving field** (F-) The person leaves one of multiple main research fields.
- **Moving into field** (F+) The person takes on a new main field of research.
- **Switching field** (Fx) There is a single main field of research and it changes.

Changes in the topics of research. When the parent node used in the change detection method is not the root of the tree, the nodes affected by the changes are at least on level 2 of the hierarchy and correspond to *research topics*. Therefore, changes in the focus set should be interpreted differently; Figure 4 displays some illustrative examples.

- **Narrowing topics** (F-) The focus is distributed between multiple research topics, one of which gets removed.
- **Broadening topics** (F+) A new research topic gets into the focus.
- **Topic switch** (Fx) Research is focused on a single topic and it changes.

5 Results

We use DBLP[3] as our data collection and generated temporal profiles with yearly steps. Each paper was classified according to the 1998 ACM Computing Classification System using an automated approach. Due to space limitations, we display the temporal profile of a single person, Dutch computer scientist Maarten de Rijke. On Figure 5 we can find

[3] `http://www.informatik.uni-trier.de/~ley/db/`

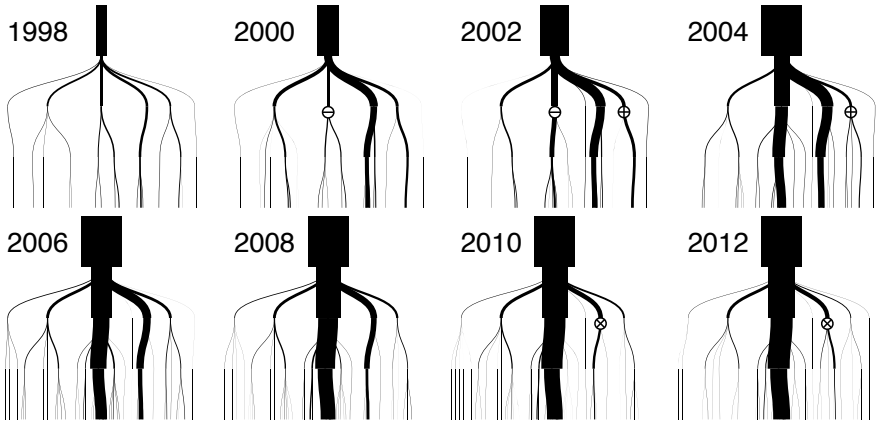

Fig. 5. Temporal expertise profile for a selected person.

examples for all 6 types of change introduced in the previous section: (1) leaving field, 2004 vs. 2008; (2) moving into field, 2002 vs. 2004; (3) switching field, 2002 vs. 2008; (4) narrowing topics, 2000 vs. 2002, denoted with $\ominus$; (5) broadening topics, 2002 vs. 2004, denoted with $\oplus$; (6) topic switch, 2010 vs. 2012, denoted with $\otimes$.

6 Conclusions

We have presented the task of temporal expert profiling and an approach for constructing temporal profiles based on documents labeled with leaf-level categories from a topic taxonomy. Further, we developed methods for identifying and explaining changes in a person's profile. We illustrated our ideas using a collection of computer science papers from DBLP classified according to the ACM taxonomy, but our approach is not limited to this setting; it could be applied, for example, on PubMed using the MeSH concept hierarchy. The evaluation of temporal expertise profiles remains an open question.

Bibliography

[1] K. Balog and M. de Rijke. Determining expert profiles (with an application to expert finding). In *Proceedings of IJCAI '07*, 2007.

[2] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, and L. Si. Expertise retrieval. *Foundations and Trends in Information Retrieval*, 6(2-3), 2012.

[3] R. Berendsen, M. de Rijke, K. Balog, T. Bogers, and A. van den Bosch. On the assessment of expertise profiles. *J. Am. Soc. Inf. Sci. Technol.*, 64(10), Oct. 2013.

[4] M. de Rijke, K. Balog, T. Bogers, and A. van den Bosch. On the evaluation of entity profiles. In *Proceedings of CLEF '10*, 2010.

[5] P. Serdyukov, M. Taylor, V. Vinay, M. Richardson, and R. W. White. Automatic people tagging for expertise profiling in the enterprise. In *Proc. of ECIR '11*, 2011.

[6] Y. Sun, T. Wu, Z. Yin, H. Cheng, J. Han, X. Yin, and P. Zhao. BibNetMiner: mining bibliographic information networks. In *Proceedings of SIGMOD '08*, 2008.

[7] G. Tsatsaronis, I. Varlamis, S. Torge, M. Reimann, K. Nørvåg, M. Schroeder, and M. Zschunke. How to become a group leader? or modeling author types based on graph mining. In *Proceedings of TPDL '11*, 2011.