Depth-Aware Topic Modeling for Expertise Detection in Social Bookmarking Platforms

No Author Given

No Institute Given

Abstract. Identifying user expertise through digital traces has become a key challenge in social computing and user profiling. Social bookmarking platforms, where users freely annotate and organize content with tags, offer a valuable source for assessing individual knowledge domains. In this paper, we extend a previous approach to expertise evaluation by integrating tag depth into the Latent Dirichlet Allocation (LDA) topic modeling process. Our method leverages both the content and the hierarchical structure of tags to enhance topic representation and better capture users' actual areas of expertise. The approach is applied to data from the Delicious platform, where we analyze tagging behaviors to infer expertise profiles. Experimental results show that incorporating tag depth improves topic specificity and provides more meaningful, quantifiable indicators of expertise. This work highlights the potential of semantic tag structures in refining topic modeling and supports the use of social tagging systems as a reliable basis for expert identification.

Keywords: Topic modeling · User profile · Social bookmarking · Expertise · LDA.

1 Introduction

In a world driven by digital collaboration, organizations increasingly rely on web communities and social networks to identify and evaluate expertise. This process helps organizations stay competitive by finding the right people with the right skills, which is necessary to create new opportunities and improve their performance. Social bookmarking platforms, where users tag, categorize and share resources, provide a valuable opportunity to assess expertise in specific domains. By analyzing tagging behavior and shared content, organizations can uncover key contributors and thought leaders in their fields. This approach has gained attention as it bridges the discovery of human expertise with computational analysis, addressing the nuanced and dynamic nature of skills and knowledge, [1].

Despite advances, traditional expert finding systems often rely on user-provided data, which can be inaccurate or insufficient, [2]. Social bookmarking platforms offer an alternative by enabling users to save, organize, and share valuable resources. Platforms like Pocket, Diigo, and Pinterest allow individuals to categorize content using tags, creating a crowdsourced directory of skills and expertise.

This practice promotes the management of personal knowledge while revealing emerging skills and areas of interest through community engagement. In this paper, we revisit and modify an existing approach to expertise evaluation based on the depth of tags and apply it within the context of social media. Through an in-depth study of Delicious, we demonstrate the potential of collaborative platforms to identify user expertise. This research shows how online interactions and contributions can serve as indicators of individual knowledge and skills.

1.1 Contributions of This Work

The main contributions of this paper are as follows:

- We propose an enhanced expertise estimation method by integrating tag semantic depth into the Latent Dirichlet Allocation (LDA) model.
- We design a new weighting formula that leverages both the topic distribution and tag depth, enabling more accurate identification of user expertise.
- We conduct extensive experiments on a large real-world dataset from the Delicious social bookmarking platform to validate the effectiveness of our approach.
- We compare the performance of the basic LDA and our proposed LDA-Depth model, showing that tag depth significantly improves expertise discrimination.

The remainder of this paper is organized as follows. Section 2 reviews related work on expertise search and social tagging. Section 3 outlines our proposed approach, followed by experimental validations in Section 4. Finally, Section 5 concludes with key findings and future research directions.

2 Related Work

Several studies have explored the challenges and techniques of expert finding. Authors in [3] introduced the concept of expertise profiling, proposing a method to create a topical profile of individuals that captures their skills, knowledge areas, and levels of competency, addressing questions like “What does expert Y know?”. social tagging operations are used to recommend handicraft women to users according to their profiles in [4], tags are weighted and used to create user’s profile according to their power to represent a user. Authors in [5] proposed a new framework for "Future Expert Finding," aiming to predict and rank experts based on their potential future contributions by using a learning framework. The study of [6] addressed the problem of dynamic user profiling by developing a user expertise tracking model. This approach utilized a Streaming Profiling Algorithm (SPA) to analyze short text streams and track the evolution of users’ expertise over time. Moreover, Authors in [7] presented two models for identifying and ranking "T-shaped users"—individuals with deep expertise in one skill area and general knowledge in others—using data from Stack Overflow, a popular Community Question Answering (CQA) platform. In their study, authors in [8] proposed the construction of expertise trees to represent candidate experts' knowledge, with three hierarchical levels: tags at the bottom, skill areas in the middle, and broad domains at the top. [9] addressed the challenges of slow responses and information overload on CQA platforms by introducing the "Tag Relationship Expert Recommendation (TRER)" method. This approach leveraged tag relationships to rank experts based on user interests and high-quality contributions, outperforming traditional recommendation systems. Furthermore, [10] introduced two new metrics—Learning Leader Metric (LLM) and Weighted Learning Leader Metric (WLLM)—to evaluate engagement, expertise, and domain relevance in online social networks using the Community of Practice (CoP) framework and information entropy. Tested against existing models, these metrics proved highly effective in ranking users and identifying learning leaders in online communities. Authors in [11] explored the use of Reddit for expert detection, introducing a semi-supervised learning model that combined Natural Language Processing (NLP), user activity, and crowdsourced features. The model categorized contributions as expert, non-expert, or out-of-scope, achieving a high AUC score of 0.93. These diverse approaches collectively highlight significant progress in the field of expert finding, with a focus on dynamic profiling, recommendation systems, and leveraging social platforms for expertise identification. Additionally, user's social activities such as comments and tags are also used in [12] to recommend adequate cloud services using a deep learning algorithm. Figure 1 illustrates how tags are used to estimate user's expertise. Recent work has questioned the reliability of traditional topic coherence metrics such as UMass or UCI. For instance, [13] highlight that these metrics can produce misleading evaluations in many real-world applications. Our depth-based refinement addresses this by incorporating semantic specificity, offering a more grounded

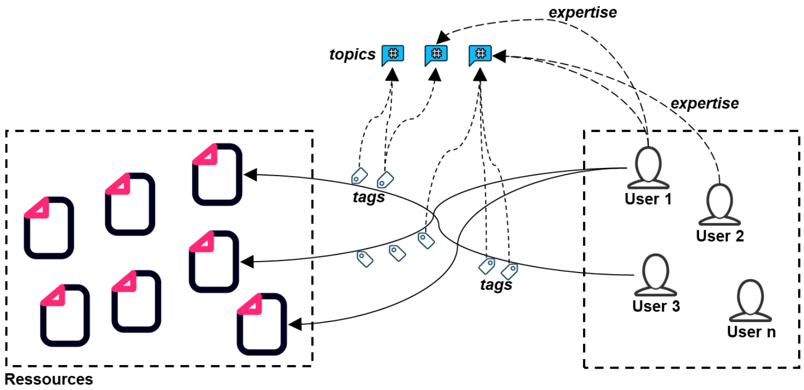


Fig. 1. Estimating expertise using tags.

interpretation of expertise-relevant topics. While other approaches have investigated neural topic modeling using distributed embeddings [14] [15].

3 A Proposed Topic Modeling-based Expertise Estimation

In this section, we will present an approach inspired by [16] and adapted for the case of social bookmarking.The existing method for evaluating expertise integrates social tagging into expert discovery by utilizing the depth and structure of tags to assess individual expertise. This approach constructs a comprehensive, multidimensional expert profile that includes personal, social, and topical dimensions. The social dimension is derived from user-associated tags, with weights calculated through a hybrid method combining naïve frequency analysis and co-occurrence patterns to ensure a balanced representation of user activity and tag relationships. The topical dimension identifies areas of expertise by analyzing specific tags, with their depth in an ontology (e.g., WordNet) serving as an indicator of expertise. The underlying assumption is that experts tend to use precise, domain-specific terms, and deeper, more specific tags reflect a higher level of knowledge. Furthermore, the methodology incorporates topic modeling to group related tags into coherent topics, enabling the identification of multiple expertise domains for each candidate. This multifaceted approach enhances the reliability of expertise identification by connecting social and topical insights. For more detailed understanding into this methodology, refer to [16]. Figure 2 provides an overview of the proposed approach. As defined by [17], Latent Dirichlet Allocation (LDA) is a generative probabilistic model for analyzing corpora. The fundamental concept is that documents are represented as random mixtures of latent topics, where each topic is characterized by a distinct probability distribution over words. This probabilistic approach enables LDA to infer the thematic structure of a collection of documents by clustering related terms into coherent topics.

The primary utility of LDA lies in its ability to associate a context with a document based on the words it contains, even when individual words could belong to multiple contexts. For example, the term "Java" could refer to either a programming language or coffee. By analyzing the surrounding words in a document, LDA can determine whether the content is about programming or beverages. Similarly, evidently unrelated terms such as "iPad," "PC," and "Laptop" may reveal an underlying semantic relationship, as they all belong to the broader topic of "computers", [18].

To adapt LDA to our context, we define each candidate ca using three dimensions, and each resource r is represented as a vector of weighted terms. These terms correspond to tags associated by candidates, where a tag t is a freely chosen keyword reflecting a candidate's expertise or interests. The set T of tags goes through preprocessing process, including stemming (to reduce tags to their root forms) and clustering (to group semantically similar tags). This step is essential to simplify processing and retain only meaningful and represen-

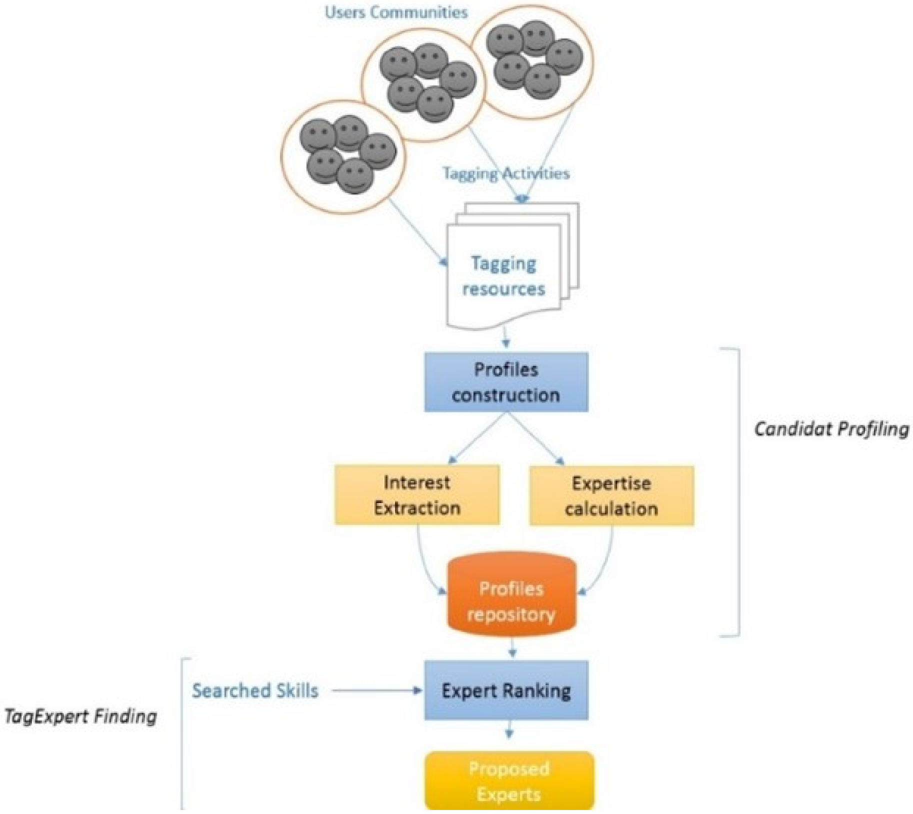
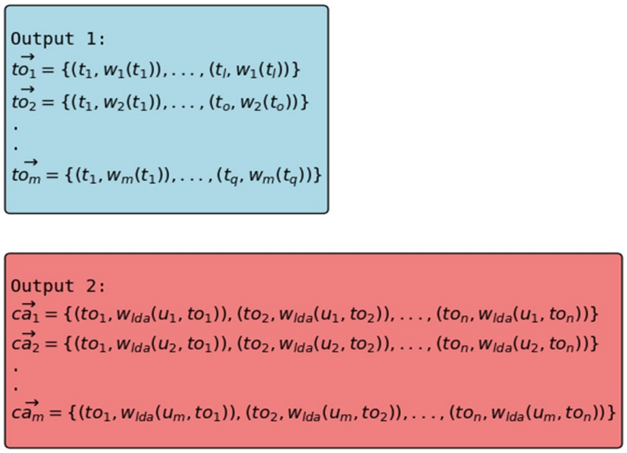


Fig. 2. The approach overview.

tative tags. In this context, a topic or skill t_o (or sometimes subject) refers to a keyword introduced by individuals to search for expert candidates in a specific domain. A tag t_i represents the i^th tag in the social dimension of a candidate ca. Each t_i belongs to a cluster that groups all variations of the tag, accounting for differences in spelling or notation (e.g., "web2.0" and "web2-0"). Using LDA, we aim to model the distribution of tags associated with resources across topics and deduce the distribution of candidates over topics, thereby providing a comprehensive understanding of expertise. We want to make a distribution of tags describing resources over topics and deduce the candidates' distribution over topics. These objectives are illustrated in Figure 3, where output 1 is the distribution of tags over topics (or subjects), the output 2 is the distribution of candidates over topics. Using LDA alone is insufficient to accurately define the topics of expertise for a candidate. In our context, LDA is employed to model each candidate as a finite mixture of topics derived from the tags they use. To ensure a focus on the most significant and meaningful tags, we prioritize deeper,



Output 1:

$\vec{to_1} = \{(t_1, w_1(t_1)), \ldots, (t_l, w_1(t_l))\}$

$\vec{to_2} = \{(t_1, w_2(t_1)), \ldots, (t_o, w_2(t_o))\}$

.

.

$\vec{to_m} = \{(t_1, w_m(t_1)), \ldots, (t_q, w_m(t_q))\}$

Output 2:

$\vec{ca_1} = \{(to_1, w_{lda}(u_1, to_1)), (to_2, w_{lda}(u_1, to_2)), \ldots, (to_n, w_{lda}(u_1, to_n))\}$

$\vec{ca_2} = \{(to_1, w_{lda}(u_2, to_1)), (to_2, w_{lda}(u_2, to_2)), \ldots, (to_n, w_{lda}(u_2, to_n))\}$

.

.

$\vec{ca_m} = \{(to_1, w_{lda}(u_m, to_1)), (to_2, w_{lda}(u_m, to_2)), \ldots, (to_n, w_{lda}(u_m, to_n))\}$

Fig. 3. Example of Tags and candidates classification with LDA.

more specific tags by applying the following proposed formula:

$$E(ca_i, t_{o_j}) = w_{lda}(ca_i, t_{o_j}) * (1 + \sum_{k}^{soc} (f_j(t_k))^{depth_{t_k}}) \tag{1}$$

Where w_{lda} the LDA calculated tag weight, $depth_t k$ is the depth of tag t_k in the considered ontology and f_j is a function that returns the weight w_k of tag t_k in the social vector of the user ca_i if t_k belongs to the descriptive vector of the topic to_j and 0 otherwise. E is the candidate expertise or proficiency in the given topic. Literally, the formula implies that candidates who use deeper, more specific tags are prioritized over those who provide less detailed or general tags.

4 Experimentation

We conducted tests on Delicious, one of the old popular social bookmarking services. The two tasks of the proposed approach are tested. In the candidate profiling task, we classify users tags using LDA, apply the depth calculation, and compare results.

4.1 Test collection and environment

The used collection contains more than 69226 Urls, tagged by 1861 users using more than 53000 tags. Achieving about 437 000 tagging operations. We used Spyder (Scientific PYthon DEvelopment enviRonment) to achieve our treatments and tests.

4.2 Tags preprocessing

In order to have credible results, insignificant tags are removed first. Then similar tags are grouped in clusters using the damerau-levenshtein similarity measure and the porter stemmer algorithm.

4.3 Tags Distribution with LDA

To implement the LDA algorithm for distributing tags across topics, we utilized the Gensim library, a widely recognized Natural Language Processing (NLP) package. Gensim is highly effective for processing textual data and uncovering latent topics within large text corpora, making it an excellent choice for our analysis. By leveraging Gensim, we were able to extract and model the underlying topics from the vast collection of user tags.

Figure 4 presents a portion of the results from an example involving 100 users' tags clustered into 20 topics. The figure highlights the tag weights associated with each topic, revealing the distribution after several iterations of learning and parameter adjustment. This iterative process allowed us to fine-tune the model for optimal topic representation and tag distribution.The figure is adjusted for enhanced readability. In Figure 5, tags associated with Topic 1 are displayed, showing their overall frequency (in blue) and their frequency within the topic (in red). This data correspond to the initial output (Output 1) presented earlier in figure 3. The distribution of tags across the twenty topics varies significantly, highlighting the diverse nature of the topics and the varying degrees of tag

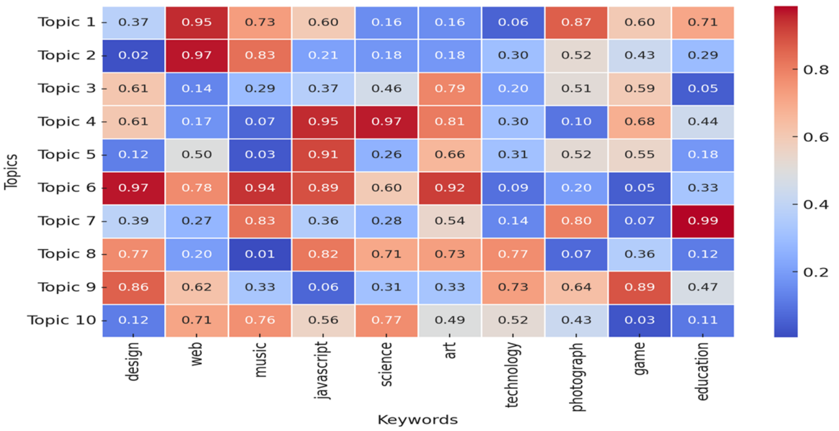


Fig. 4. Example of tags distribution over topics.

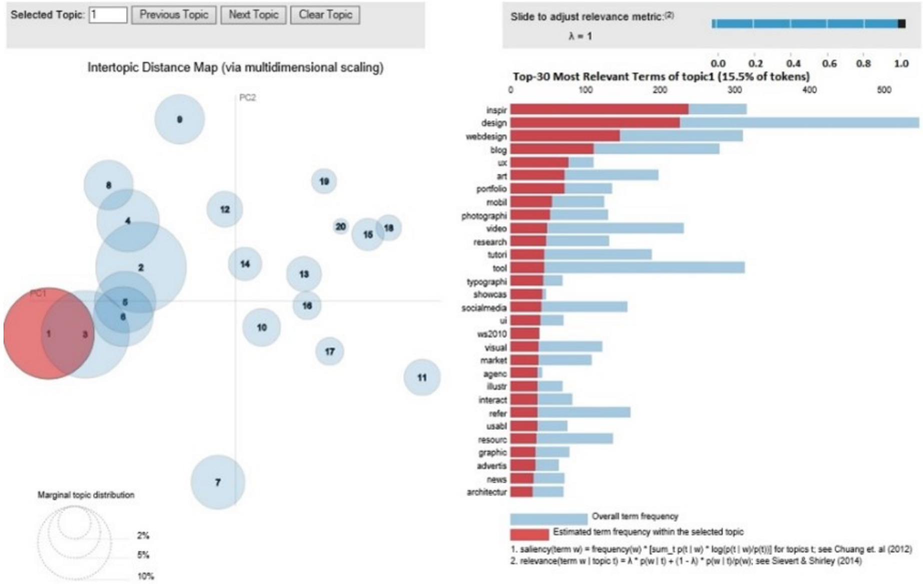


Fig. 5. Top-30 tags of topic 1 with the estimated frequencies within the topic.

concentration within each topic. The second desired outcome (output 2 as shown in figure 3) is the distribution of candidates across topics. Table 1 summarizes the topic distributions for a selection of candidates, providing an overview of how their associated tags align with different topics.

Often, topics related to a candidate are in number of 1 or 2 and in this sample exceeds 3 topics per candidate once (candidate 53), it didn't exceed 3 topics in other samples. This result confirms that candidate can have several expertises in different topics.

4.4 Depth calculation

We used NLTK (Natural Language ToolKit), a leading platform for building Python programs to work with human language data. And the wordnet package for python, to exploit the tag depth.
In the process of depth calculation, some tags are removed (do not exist in wordnet). Others have a depth 0. As example, Figure 6 illustrates different depths of tags composing topic 2, (in this case depths are between 0 and 8, tags 'iphon', 'recip','ipad' and 'visual' have a depth 0). Depths values can reach 14 in other cases.

4.5 Basic LDA vs LDA-Depth

Based on candidates tags, collected from the dataset, LDA algorithm is used to classify these tags on topics. Candidates are also listed by their topics, topics

Table 1. Topics and their frequencies related to a set of candidates.

ca	(Topic, frequency)
01	(17,0.99394)
02	(9,0.9634)
03	(0, 0.0755), (14, 0.9223)
04	(1,0.9833)
05	(9, 0.6675), (14, 0.3298)
...	....
17	(0, 0.9964)
18	(10, 0.9945)
19	(3, 0.9920)
20	(3, 0.9321)
....	....
30	(4, 0.9959)
31	(4, 0.0331), (9, 0.8159), (19, 0.1456)
32	(18, 0.9952)
33	(12, 0.9981)
...	....
46	(17, 0.9948)
47	(0, 0.9975)
48	(3, 0.0138), (12, 0.9817)
49	(5, 0.9965)
50	(0, 0.9866)
51	(5, 0.9974)
52	(13, 0.9882)
53	(5, 0.1732), (12, 0.1270), (14, 0.0198), (17, 0.0801), (19, 0.5970)
......	....
73	(9, 0.4595), (14, 0.0678), (19, 0.4712)
74	(10, 0.9919)
76	(1, 0.9958)
75	(2, 0.9878)
......	....
92	(9, 0.8871), (19, 0.1119)
93	(16, 0.9981)
94	(11, 0.9931)
98	(14, 0.6942), (17, 0.2918)
99	(5, 0.2832), (7, 0.6286), (9, 0.0848)
100	(1, 0.9860)

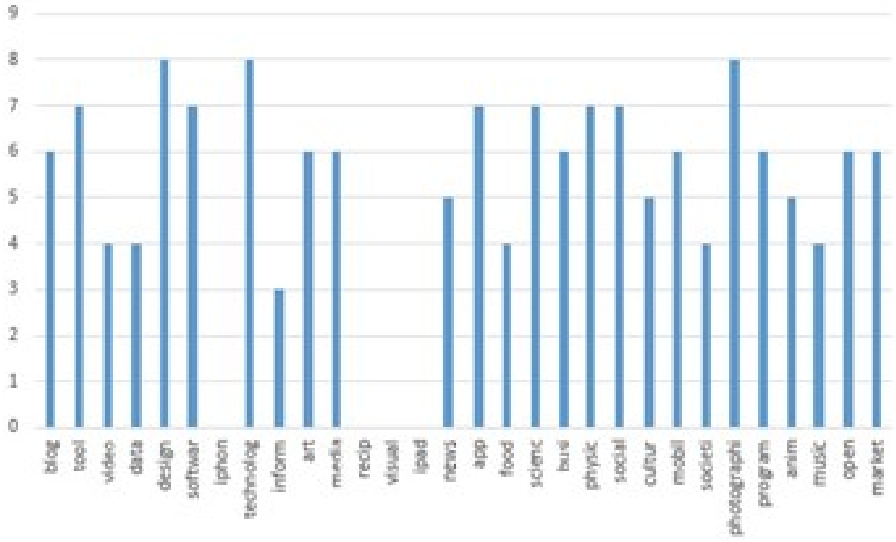


Fig. 6. Depths variation of tags composing topic 2.

are weighted for each candidates. Figure 7 provides the partition of candidates by topic along with the frequency of each topic for each candidate (both with LDA and LDA-depth).

In the other side, the proposed approach based on tag's depth, gives a new distribution of candidates over topics, the topic frequency is the tags depths average. In deed, the most specific topics (having high values) are chosen, contrariwise topics with generic tags (small values) are not chosen but can be cited in a second degree of importance. Figure 8 illustrates candidates's topics with the calculated expertise with tags depth above 5.

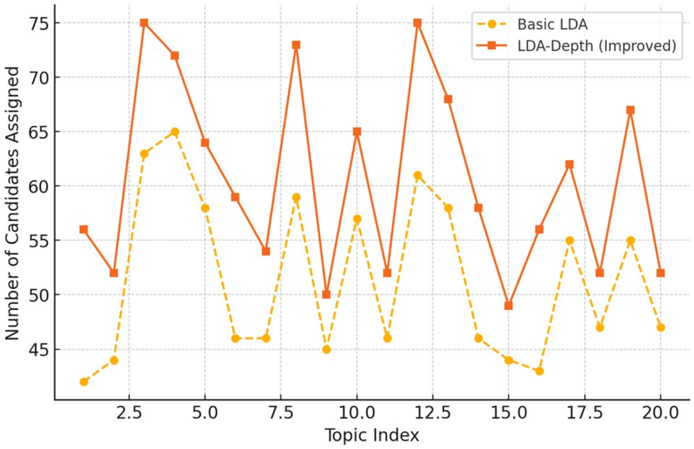


Fig. 7. Candidates distribution over topics with LDA, LDA-Depth.

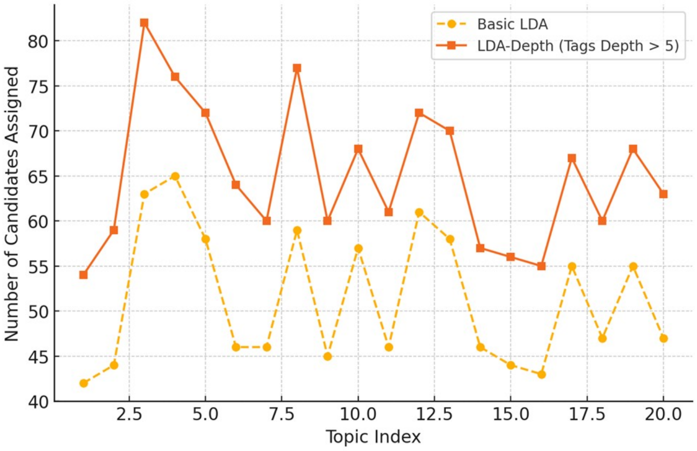


Fig. 8. Candidates distribution over topics with LDA, LDA-Depth containing tags with depth above 5.

Figure 9 compares Expertise Estimation Accuracy across different Tag Depths for the two methods: Basic LDA and LDA-Depth (Improved). The curve shows that as the depth of the tag increases, both models improve accuracy. However, the LDA-Depth method consistently outperforms Basic LDA, indicating that incorporating depth improves the estimation of expertise. The accuracy gap between the two methods increases at lower tag depths but remains significant throughout, highlighting the advantages of the improved approach.

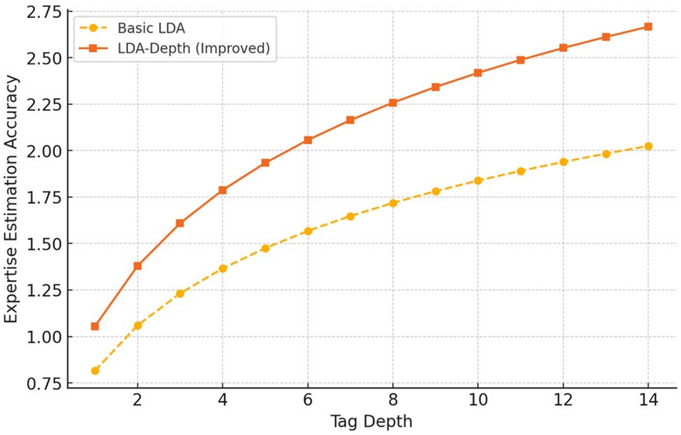


Fig. 9. Expertise Estimation Accuracy with LDA vs LDA-depth.

Table 2 summarizes a set of basic comparative metrics between the classical LDA approach and our proposed LDA-Depth method. The results suggest that LDA-Depth reduces topic redundancy and increases topic specificity per candidate. Although a small portion of candidates receive no topic assignment due to low tag depth, the expertise separation is significantly improved, as reflected by the Expertise Separation Score. This supports the claim that depth-enhanced topic modeling better reflects real expertise boundaries.

Table 2. Comparative performance of LDA vs. LDA-Depth on candidate profiling

Metric	**LDA Classic**	**LDA-Depth (Proposed)**
Avg. # Topics per Candidate	2.8	1.6
Tag Specificity (avg. depth)	3.2	6.1
Candidates with ≥ 3 topics	14%	2%
Candidates without topic	0%	5%
Expertise Separation Score*	0.58	**0.76**

** Expertise Separation Score: standard deviation of topic weights across candidates.*

4.6 Discussion

The results highlight two key aspects: a positive and a negative one. Positive aspect: Our approach effectively filters candidate topics of interest, retaining only those where the candidate is likely to have significant expertise. This determination is based on the depth of associated tags, which enhances the precision of expertise identification. In our proposed method, the values of expertise are more meaningful and discriminative compared to LDA, where topic frequencies are very close to one another and tend to converge towards zero, reducing interpretability. Negative Aspect: On the downside, the LDA-Depth method can result in the neglect of certain topics, primarily due to the lack of specificity in their tags. This limitation means that in some cases, candidates may not be assigned any topic of expertise, potentially underestimating their actual range of knowledge. More advanced alternatives such as [14] or contextual embeddings like BERT [15] could be explored in future work.

5 Conclusion

In this paper, we present a revisited approach that integrates social indicators (tags) and their depths to evaluate candidates' expertise and build meaningful descriptions of their profiles. Our methodology leverages existing techniques for extracting interests from social tagging data, combined with a topic modeling algorithm (LDA) to distribute tags across topics.

The LDA algorithm is enhanced by incorporating tag depths, enabling more precise identification of topics closely aligned with a candidate's expertise. The proposed approach was applied to the Delicious collection, and the tests demonstrated more significant results, including specific topics and quantified measures of expertise.

Although our approach demonstrates potential, it can be further improved by adopting a learning technique to infer candidate profiles in a cold start scenario. Additionally, we aim to extend this approach to applications in expert finding tasks.

References

[1] Paolo Cifariello, Paolo Ferragina, and Marco Ponza. "Wiser: A semantic approach for expert finding in academia based on entity linking". In: *Information Systems* 82 (2019), pp. 1–16.

[2] Gunwoo Park and Dongwoo Kim. "CredibleExpertRank: Leveraging Social Network Analysis and Opinion Mining to Facilitate Reliable Information Retrieval on Knowledge-Sharing Sites". In: *IEEE Access* 11 (2023), pp. 54724–54749.

[3] Krisztian Balog, Maarten De Rijke, et al. "Determining Expert Profiles (With an Application to Expert Finding)." In: *IJCAI*. Vol. 7. 625. 2007, pp. 2657–2662.

[4] Saida Kichou et al. “Handicraft women Recommendation Approach based on User’s Social Tagging Operations”. In: *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE. 2016, pp. 618–621.

[5] Mahmood Neshati, Zohreh Fallahnejad, and Hamid Beigy. “On dynamicity of expert finding in community question answering”. In: *Information Processing & Management* 53.5 (2017), pp. 1026–1042.

[6] Shangsong Liang. “Dynamic user profiling for streams of short texts”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.

[7] Sajad Sotudeh Gharebagh, Peyman Rostami, and Mahmood Neshati. “T-shaped mining: A novel approach to talent finding for agile software teams”. In: *Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings 40*. Springer. 2018, pp. 411–423.

[8] Mahdi Dehghan, Maryam Biabani, and Ahmad Ali Abin. “Temporal expert profiling: With an application to T-shaped expert finding”. In: *Information Processing & Management* 56.3 (2019), pp. 1067–1079.

[9] Anitha Anandhan, Maizatul Akmar Ismail, and Liyana Shuib. “Expert Recommendation Through Tag Relationship In Community Question Answering”. In: *Malaysian Journal of Computer Science* 35.3 (2022), pp. 201–221.

[10] Hend Aldahmash, Abdulrahman Alothaim, and Abdulrahman Mirza. “Identifying Learning Leaders in Online Social Networks Based on Community of Practice Theoretical Framework and Information Entropy”. In: *IEEE Access* (2024).

[11] Sofia Strukova, José A Ruipérez-Valiente, and Félix Gómez Mármol. “Computational approaches to detect experts in distributed online communities: A case study on Reddit”. In: *Cluster Computing* 27.2 (2024), pp. 2181–2201.

[12] Saida Kichou, Brairi Ismail, and Essalhi Mohamed Amine. “Deep Learning-Based cloud services recommendation”. In: *2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*. IEEE. 2024, pp. 1–8.

[13] Alexander Hoyle et al. “Is automated topic model evaluation broken? the incoherence of coherence”. In: *Advances in neural information processing systems* 34 (2021), pp. 2018–2033.

[14] Dimo Angelov. “Top2vec: Distributed representations of topics”. In: *arXiv preprint arXiv:2008.09470* (2020).

[15] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019, pp. 4171–4186.

[16] Saida Kichou, Omar Boussaid, and Abdelkrim Meziane. “Tag’s depth-based expert profiling using a topic modeling technique”. In: *International*

Journal on Semantic Web and Information Systems (IJSWIS) 16.4 (2020), pp. 81–99.

[17] David M Blei, Andrew Y Ng, and Michael I Jordan. “Latent dirichlet allocation”. In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[18] Hengshu Zhu, Enhong Chen, and Huanhuan Cao. “Finding experts in tag based knowledge sharing communities”. In: *Knowledge Science, Engineering and Management: 5th International Conference, KSEM 2011, Irvine, CA, USA, December 12-14, 2011. Proceedings 5*. Springer. 2011, pp. 183–195.