An Automatic Method for Building a Taxonomy of Areas of Expertise

Thi Thu Le[1][a], Tuan-Dung Cao[2][b], Xuan Lam Pham[3], Duc Trung Pham[3] and Toan Luu[4]
[1]Department of Research Methodology, Thuongmai University, Hanoi, Vietnam
[2]School of Information and Communication Technology, Hanoi University of Science and Technology, Hanoi, Vietnam
[3]Department of Computer Science, School of Information Technology in Economics,
National Economics University, Hanoi, Vietnam
[4]Move Digital AG, Zurich, Switzerland

Keywords: Topic Taxonomy Construction, Expert Finding System, Expertise Profile.

Abstract: Although a lot of Expert finding systems have been proposed, there is a need for a comprehensive study on building a knowledge base of areas of expertise. Building an Ontology creates a consistent lexical framework of a domain for representing information, thus processing the data effectively. This study uses the background knowledge of machine learning methods and textual data mining techniques to build adaptive clustering, local embedding, and term ordering modules. By that means, it is possible to construct an Ontology for a domain via representation language and apply it to the Ontology system of expert information. We proposed a new method called TaxoGenDRK (Taxonomy Generator using Database about Research Area and Keyword) based on the method from Chao Zhang et al. (2018)'s research on TaxoGen and an additional module that uses a database of research areas and keywords retrieved from the internet – the data regarded as an uncertain knowledge base for learning about taxonomy. DBLP dataset was used for testing, and the topic was "computer science". The evaluation of the topic taxonomy using TaxogenDRK was implemented via qualitative and quantitative methods, producing a relatively good accuracy compared to other existing studies.

1 INTRODUCTION

For searching systems in general and Expert Finding Systems (EFS) in particular, it is essential to build a knowledge database of the research areas of experts because this will enhance the quality of search and recommendation algorithms (Abramowicz et al., 2011; Husain et al., 2019). An EFS will contain a lot of information about experts in different research fields (Al-Taie et al., 2018). Thus, each area needs a consistent vocabulary framework to display, categorize and process the experts' information effectively. A sufficient knowledge framework will help the systems to find experts, with the keywords as the research areas or related research areas. By that means, the systems can suggest or recommend related researchers. The knowledge base will make the systems work more smartly, producing more accurate and sufficient search results (Lin et al., 2017).

Currently, there are two primary ways to solve the problem: manual and automatic taxonomy building. Manual building is advantageous because it can produce a solid taxonomy but requires experts in the areas as well as considerable time and effort. The knowledge in a specific field is vast and needs deep research to have a good taxonomy. Building automatic taxonomy is supplementary to the manual method, which can use machine learning to mine the data structure to create a quality taxonomy (Liu et al., 2012; Song et al., 2015; Zhang et al., 2018).

Extensive expertise data can be created frequently due to new research orientations and terminologies. The directory can change over time, and most of the data can only be understood by humans, not computers. This gives rise to the appearance of the semantics web and Ontology, where data and knowledge are represented in structures that are comprehensible to computers (Gomez-Perez & Corcho, 2002).

This study uses background knowledge of machine learning methods and textual data mining techniques to build adaptive clustering, local

[a] https://orcid.org/0000-0001-6192-0212
[b] https://orcid.org/0000-0002-3661-9142

Le, T., Cao, T., Pham, L., Pham, T. and Luu, T.
An Automatic Method for Building a Taxonomy of Areas of Expertise.
DOI: 10.5220/0011630500003393
In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023) - Volume 3*, pages 169-176
ISBN: 978-989-758-623-1; ISSN: 2184-433X

embedding, and term ordering modules. By that means, it is possible to construct an Ontology for a domain via representation language and apply it to the Ontology system of expert information. We proposed a new method called TaxoGenDRK (Taxonomy Generator using Database about Research Area and Keyword) based on the method from Zhang et al. (2018)'s research on TaxoGen and an additional module that uses a database of research areas and keywords retrieved from the internet – the data regarded as an uncertain knowledge base for learning about taxonomy.

However, the topic taxonomy for EFS has received scant attention, and there has been little body of research on this topic. The algorithm is used to build a knowledge database on the research areas of experts to serve the EFS. We experiment with the DBLP dataset and aim to address two research questions:

- Question 1. How is the performance of TaxoGenDRK in creating the topic taxonomy in Computer Science?
- Question 2. Is the TaxoGenDRK method suitable for building the knowledge base for EFS?

To realize the research objectives, the study focuses on presenting the theoretical basis and related studies related to the problem, including machine learning and text data processing theories. This is followed by an overview of the solution adopted and a detailed description of the modules used in the method. Afterward, we discuss the tests of the method performed on the input dataset and evaluate its effectiveness. The research results draw conclusions about what has been achieved and the remaining limitations of the method used. In addition, some directions for further research and research recommendations are provided.

2 LITERATURE REVIEW

2.1 Ontology in Expert Finding Systems

Ontologies, which are clear formal descriptions of the concepts in the domain and their relationships (Gruber, 1993), have recently moved from artificial intelligence laboratories to the desktops of subject-matter experts. Ontologies on the Web range from large-scale classification methods for the classification of websites to products for sale and their characteristics (Noy & McGuinness, 2001). Ontologies define a general vocabulary for researchers who need to exchange information within a domain (Çelik et al., 2013). Similarly, an ontology-based method was proposed to find experts (Uddin et al., 2011) in a certain field, primarily a research topic in computer science and engineering.

There are a wide variety of methods that concentrate on EFS, including Fine-Grained (Deng et al., 2008), hybrid topic and language models (Deng et al., 2008), and integrated evidence (Bogers et al., 2008). Additionally, numerous methods involving academic and social networks have been created for locating specialists by the quantity of researchers. However, they are based on text mining and probabilistic approaches for performing expert searches. Based on the collection of information, ontologies used in expert search have been proposed to improve the knowledge base model for academic information (Bukowska et al., 2012).

2.2 Methods for Finding Experts

Numerous research has examined methods for finding experts. According to their primary areas of focus, the existing techniques can be grouped into three major categories.

2.2.1 Content-based Method

Content-based methods have received much attention from different studies. Text REtrieval Conference (TREC) considers the first type of content-based methodology to be an information retrieval task. These methods fall into two categories: profile-centric method and document-centric method (Balog et al., 2012; Petkova & Croft, 2006). All documents or texts related to a candidate are combined into a single personal profile in the profile-centric method, and the ranking score for each candidate is then estimated based on the profile in response to a given query (Balog et al., 2006). However, instead of creating a single expertise profile, the document-centric method analyzes the content of each document separately (Balog et al., 2006; Wu et al., 2009). To take advantage of the benefits of both the profile-centric and document-centric methods, some existing approaches have combined the two to improve expert-finding performance (Petkova & Croft, 2008). By using the term "vectors" with bag-of-words representation, these studies typically focus on matching search results to user queries, which differs from topic-dependent expert finding based on automatically inferred latent topics.

Topic modeling is the second type of content-based method. To overcome the limitations of the earlier topic model, a three-level hierarchical

Bayesian model called Latent Dirichlet allocation (LDA) was proposed (Blei et al., 2003) to change the problems to topic-vector based representation. To extend the effort of the LDA model, an author-topic model was introduced (Rosen-Zvi et al., 2004) to illustrate the connection between the content of documents and the interests of authors by sharing the hyperparameter of all publications by the same authors. As a result, topic models can support in evaluating the contribution of candidates to an inferred topic. They do not, however, analyze the relationship between fields of research that represent authors' knowledge to build a network for more sophisticated identification of experts.

2.2.2 Link Structure-based Method

To describe the direct evidence of candidates' expertise, link structure-based method using PageRank (Ding et al., 2009; Page et al., 1999) and HITS (Kleinberg, 1999) algorithms were applied in analyzing relationships in the scholarly network to find authorized experts. AuthorRank, which is a modification of PageRank with weights among the co-authorship links, was also introduced to solve the connection. Citation graph was used to evaluate the impact of scientific journals and conferences, publications, and academic authors. Several modifications of traditional PageRank for bibliographic networks were presented and achieved a better result than the standard algorithm (Fiala et al., 2008). PageRank with damping factors with weighted algorithms that consider citation, co-authorship topology and co-citation network was proposed to measure author impact and rank them (Ding, 2011; Page et al., 1999). All those efforts aim to improve the accuracy of classical indicators such as impact factor, citation count and H-index by applying different modifications of HITS or PageRank algorithms. However, they are ineffective for identifying the top "experts" without focusing on content features.

2.2.3 Combination of Content-based and Link Structure-based Methods

Some researchers have employed link analysis to further improve the ranking results after using documents or snippet-level content to evaluate topic relevance for each applicant. Text analysis and network analysis were used after gathering all emails particularly related to a topic and examining connections between every pair of individuals, to sort individuals and create "expert graph" (Campbell et al., 2003). Then modified HITS was applied to find ratings for all senders and recipients related to that topic. Candidates' personal profile and publications were used to estimate an initial expert score for each applicant and choose the top applicants to build a subgraph (Zhang et al., 2007). All of them were accomplished by operating through local optimization while ranking among a small subset of applicants. However, a topical factor graph model was suggested to find representative nodes from academic networks on a given topic by utilizing the topic relevance and social ties between linkages (Tang et al., 2009). In this research, we suggest an improved method to identify the relationship between topics for building topic tree-structured hierarchy, which is called topic taxonomy.

2.3 Topic Taxonomy Construction Method

A topic taxonomy is a tree-structured hierarchy in which each node has a group of semantically related terms that each reflect a particular conceptual topic (Shang et al., 2020). Additionally, the topic-subtopic relation should be followed by the parent-child nodes in the hierarchical tree. For example, if a node has children S = c_1, c_2, ..., c_n, then each c_i should be a sub-topic of the node and have the same level of precision as its siblings in S.

It is important to note that a term may be a sub-topic of multiple conceptual topics and hence appear in various nodes. For instance, "search engine" may be a component of both "search engine in information retrieval" and "image matching search engine"; "neural networks" may be a component of both "neural networks in deep learning" and "neural networks for data mining."

The main approaches for building taxonomy were introduced as hyponymy-based methods, clustering-based methods, network clustering-based methods.

3 METHODOLOGY

3.1 Research Process

From an input corpus about a field, we need to create a topic taxonomy for that field. It is a data structure of domain terms, namely a tree-structured hierarchy of topics and terms within the area, in which each node of the tree consists of a set of terms representing a topic, and the child nodes are the sub-topics of the parent node. The hierarchy is shown in the Figure 1. This structure meets the requirements for building a taxonomy of areas of expertise as each topic will usually be described by a group of synonymous terms or terms related to the same content represented by the group's topic.

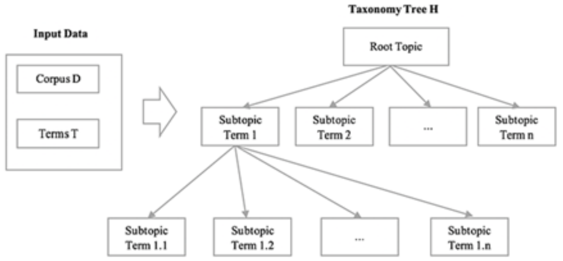

Figure 1: The topic taxonomy problem.

From the above research objectives and directions, this study was implemented in the following stages:

1. A new method called **TaxoGenDRK** (**Taxo**nomy **Gen**erator using **D**atabase about **R**esearch area and **K**eyword) was proposed in an attempt to improve the topic taxonomy method Taxogen. This is a method based on TaxoGen research by Zhang et al. (Zhang et al., 2018), together with an additional module that uses a database of research areas and keywords extracted from the internet – the data regarded as an uncertain knowledge base for learning about taxonomy. This dataset is a data structure about research fields and keywords built from crawled data from Google Scholar[3].
2. An experiment was conducted to build a topic taxonomy, with "Computer Science" as the main topic for testing, using Python language, Google Colab Pro, and the DBLP dataset.
3. An evaluation of the topic taxonomy using TaxogenDRK was conducted from qualitative and quantitative approaches.

The TaxogenDRK method is the focal point of our research.

3.2 Proposed Method: TaxoGenDRK

The overview model of the TaxoGenDRK method is shown in Figure 2 below. The initial input and output are the same as in the TaxoGen method. Two local terms embedding, and adaptive clustering modules are used for term clustering to create the output tree. In addition, the method needs to perform additional data retrieval from experts' research areas and scientific publications before building a dependency graph from the extracted information, and finally, use it to arrange terms based on a topic representation level. The two reused modules will not need to be re-described, but the next section will describe the further developed term ordering module (Figure 3).

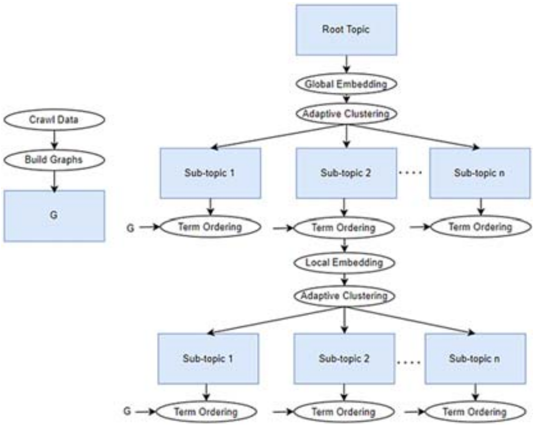

Figure 2: An overview of the TaxoGenDRK method.

The term ordering module includes the tasks shown in Figure 3, starting with the retrieval of data about the experts' research areas and keywords in their scientific publications to the construction of a dataset of research areas and related terms. The data set is in the form of a dependency graph between the research areas and keywords, all of which will be extracted from information pages about experts in their fields of research and scientific publications. Google Scholar was used because it is the leading scientific and expert information page used in the research community. Finally, we combined the classified term clusters to order the terms. The tasks will be described in more detail in the following sections.

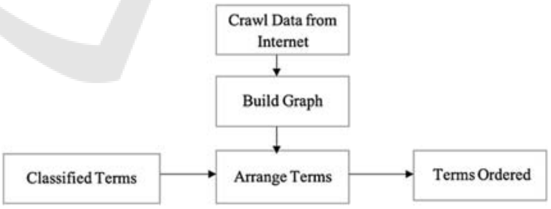

Figure 3: The stages in the term ordering module.

The next step is to build a dependency graph, a set of graphs in which each graph has a node as a research area, the remaining nodes are keywords related to that research area, connected with the research area by a weighted edge, and the weight is calculated by the probability that the keyword is related to the area.

It describes the research fields and the keywords in the field with a weight $w_{ra\text{-}kw}$ representing the probability of the keyword in the field.

[3] https://scholar.google.com/

$$w_{ra-kw} = 1 \; or \; w_{ra-kw} = \frac{1}{len(ra)} \quad (1)$$

In (1), $len(ra)$ is the number of research areas of the expert whose scientific publications are crawled to get the keywords. With two identical keywords extracted from two different publications added to the same topic, the weights will be updated by the higher value between the two keyword weights.

This module has the function of sorting terms according to a level called topic visibility, making it easy to select the top terms of a cluster. In this method, we will use a score to evaluate the topic representation of the node, and each node will be illustrated by the term with the highest topic representation and the top terms with the highest topic representation.

The topic representation for each term for a cluster is calculated using the following formula:

$$score(t, S_k) = cossim(t, S_k) * r(t, S_k) \quad (2)$$

In (2), $cossim(t, S_k)$ is the similarity degree cosine of the embedding vector of term t with the center of cluster S_k, and $r(t, S_k)$ is the degree of representation of the term t for the cluster S_k (as included in the parent node of the cluster S_k during clustering of that parent node).

In the step of ordering terms, the clusters containing the term is the area of research, and the topic representation score of the research area will be recalculated based on the association graph above as follows:

$$newscore(t, S_k) = score(t, S_k) + \sum_{kw \in T} score(kw, S_k) * w_{ra-kw} \quad (3)$$

In (3), T represents the set of terms that are keywords and belong to the dataset of research fields associated with the keywords built above.

3.3 Experiments

3.3.1 Dataset

The test dataset is the DBLP, a commonly used data set in some other topic taxonomy studies. Zhang et al. (2018) published preprocessed DBLP dataset, and this study used the above dataset for testing and evaluation. The above dataset contains a corpus with more than 1.8 million titles of scientific articles in computer science fields.

Because the topic taxonomy requires a large corpus, the effectiveness of the experimental setup is required for its efficiency in optimizing memory space and processing time. The preprocessing steps are used as much as possible as they are performed only once compared to the next repeated steps of the algorithm. In the preprocessing step, the tool NPC (Noun Phrase Chunking) is used to extract noun phrases from scientific article titles to form a term set after removing duplicate words and selecting the most common terms to form a collection of more than thirteen thousand terms. In addition, the next preprocessing steps include indexing the corpus, calculating term frequency on each corpus, learning global embedding vectors, calculating document lengths, and storing the index of documents where the term appears.

3.3.2 Parameters

The study is implemented in Python language version 3.6, and is run on Google Colab Pro. The parameters used for the implementation of the method are as follows: **n_cluster**: 5 (the number of clusters categorized at each level); **max_depth**: 4 (the maximum depth of the taxonomy tree); **filter_threshold**: 0.25 (the threshold for generic term removal in adaptive clustering module); **n_cluster_iter**: 2 (the maximum number of iterations of adaptive clustering); **n_expand**: 100 (the number of extended terms, counting from the central term during local embedding); **n_include**: 10 (the number of terms close to the cluster center chosen to extract the secondary corpus during local embedding); **n_top**: 10 (the number of terms with the highest subject representation selected during the term ordering process).

The study combines the data structure of the research areas and the keywords crawled from Google Scholar. To get the dataset, conduct a query to crawl the data of more than fifty experts in the field of computer science, with a number of leading research citations in several research areas, such as machine learning, computer vision, computer security, and natural language processing. The data structure is as follows: **Number of research fields**: 34 (research fields from the information of experts); **Number of keywords**: 213 (crawled keywords from the publication description pages); **Number of links**: 759 (the link between research areas and keywords); **Number of reliable links**: 113 (link weighted from 0.5 or more); **Average number of links**: 293 (links weighted greater than 0.2 and less than 0.5); **Number of less reliable links**: 353 (link weighted 0.2 or less).

4 RESEARCH RESULT

4.1 Taxonomy Results

The result of building topic classification on the DBLP dataset, with the main topic being "computer science," the topic taxonomy tree is built in 4 levels, and each parent topic is divided into 5 sub-topics. Figure 4 shows a part of the topic taxonomy results, where the sub-topics are represented by the five terms with the highest topic representation and are labeled with the words of the highest topic representation.

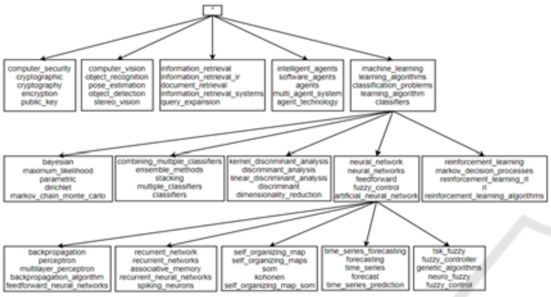

Figure 4: TaxoGenDRK results for classifying Computer Science topics in some clusters.

4.2 Evaluation

4.2.1 Qualitative Evaluation

In the results of the topic taxonomy in the field of "computer science," not all taxonomies give reliable results. Since the number of sub-clusters is fixed at 5, some clusters do not really belong to their parent cluster, although the labels and top terms are still quite semantically clear about the same topic. For instance, the cluster of natural language processing can be seen in Figure 5, which was founded during the clustering process of the topic "information retrieval." Since we fixed the number of clusters of the computer science root node to be 5, the terms on the topic of natural language processing must belong to one of those 5 clusters, and the cluster whose center is close to the term set of the most natural language processing is the information extraction cluster, so it is classified as a child of the information extraction node.

In addition, although the process of adaptive clustering and local embedding helps the clusters to separate clearly, eliminating common terms in the sub-clusters, it is not completely effective for the whole result. In some clusters, after being split, one of the sub-clusters is almost still at the center of the parent cluster, as can be seen in the "search results" cluster at level two in where there is a subcluster that seems to be the center of the parent cluster, and they all represent the same topic. When splitting the above cluster into sub-clusters, there is a cluster located at the center of the parent cluster.

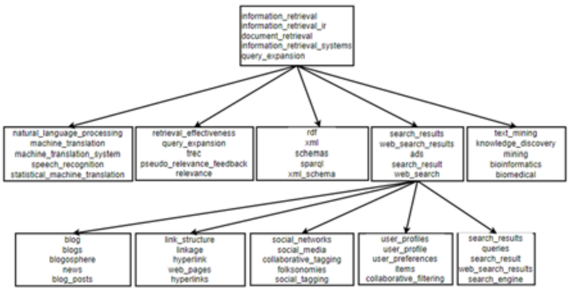

Figure 5: TaxoGenDRK results for the cluster "information retrieval".

4.2.2 Quantitative Evaluation

The quantitative evaluation of a topic taxonomy tree is not an easy task to implement. The researchers made references to related studies, as in (J. Shang, 2020; Zhang et al., 2018), to learn about how to build an assessment of the topic taxonomy tree and to provide quantitative measures to evaluate the results. In general, the related studies all constructed a set of evaluation indicators, then used humans as the main agent to evaluate each selected example, and finally arrived at a quantitative evaluation. From the evaluation indicators, we can compare the results with each other even at a relative level because, in different studies, the main agent to evaluate (humans) are different. The indicators used in this study are: **Relation accuracy**: the accuracy between relationships that measures the accuracy of the relationship between parent and child nodes in a classification tree; **Exclusive sibling**: the semantic separation between sibling nodes in the same parent topic; **Term coherence**: the coherence of terms that measures how well the top terms represent the same topic

In all the above criteria, the data points of the results are given to a team of three engineers working in the field of computer science, who will learn about unknown terms and use their knowledge of known terms to evaluate each result. The majority's result was selected as the final result for each data point. In machine learning taxonomy, the commonly used metrics are Precision, Recall, and F-score. However, in the topic taxonomy problem, each data point will only have True Positive and False Positive, so only the Precision measure will be used for the three evaluation indicators above. The specific formula for the measurements is as follows:

Precision:

$$Precision = \frac{TP}{PP} = \frac{TP}{TP + FP} \tag{4}$$

In (4), FP is the number of False Positives (number of false predictions), TP is the number of True Positives (number of correct predictions), PP is the number of Predicted Positives (total number of predictions), and RP is the number of Real Positives (the total number of correct labels).

The specific results are shown in Table 1. The first column is the name of the methods, and the next columns are respectively the evaluation indicators for the methods, and the blank boxes are the results not evaluated. The methods are: **TaxoGenDRK**: the proposed research method; **TaxoGen**: the method that does not use the term sorting module but only uses the original representative points to generate the top terms; **NoAC**: the method does not use an adaptive clustering module, but uses algorithms to cluster without removing generic terms; **NoLE**: the method does not use local embedding modules but uses global embedding vectors in the clustering steps; **HLDA**: the non-parametric topic hierarchical model tested to build a topic taxonomy tree; **Hclus**: building clustering using global embedding vectors and using only K-means spherical clustering algorithm to build a hierarchical taxonomy tree (Zhang et al., 2018).

Table 1: Results of the topic taxonomy tree evaluation.

Method	Relation Accuracy	Exclusive Sibling	Term Coherence
TaxoGenDRK	0.79	0.76	0.82
TaxoGen	0.77	0.73	0.84
HClus	0.44	0.47	0.62
NoLE	0.64	0.70	-
HLDA	0.27	0.44	-
NoAC	0.56	0.35	-

These results show the effectiveness of the model in building the taxonomy tree. All three modules of adaptive clustering, local embedding, and term sorting provide good indicators, and the results are consistent with the qualitative evaluation in the previous section. The last module is to order the terms, which seems to show only a small effect because it does not affect the process of creating clusters, but when the terms have been hierarchically clustered, it orders them to get the position of the top terms. The results also showed the efficiency of TaxoGenDRK compared to TaxoGen in the evaluation index of the relationships between parent and child nodes as well as the separation between sibling clusters.

5 CONCLUSION

The research process on building an automatic ontology of expertise has brought insights into how to represent knowledge in the current web systems, the approaches, and deployment of machine learning methods, including adaptive clustering module, local embedding module, and term ordering module to solve the problem of building topic taxonomy. The proposed method helped construct a topic taxonomy in the field of “computer science” with a relatively good degree of accuracy compared with the existing studies. Thus, research question 1 has been resolved. The proposed TaxoGenDRK model, however, has given relatively reliable results compared with existing studies on the problem of automatically building topic taxonomies. However, it still has some limitations that must be acknowledged and improved. Some parameters used seem to significantly influence the results, such as the number of clusters in each topic, the threshold to remove generic terms, and so on.

In research question 2, It can be seen that the TaxoGenDRK method is quite suitable for building the knowledge base for EFS. However, the extracted data about the information of research fields and keywords in expert publications are only uncertain knowledge bases. As a result, sometimes, the details learned are not accurate. Further studies can focus on parameters suitable for different steps or different inputs. In addition, the input corpus also dramatically affects the results of classification construction, so it is necessary to do more research on building a generalized input for a topic with complete and relevant contexts for building a taxonomy. Another approach is to create a data set on the research field and keywords in the area with higher reliability. Currently, this study only focuses on the field of computer science with DBLP datasets. Further research can be extended to other fields to build knowledge representations for scientists in diverse research areas.

ACKNOWLEDGEMENTS

The research team would like to thank the Innovation Fund (VinIF), Big Data Research Institute (VinBDI) for financial support to carry out this study.

REFERENCES

Abramowicz, W., Bukowska, E., Dzikowski, J., Filipowska, A., & Kaczmarek, M. (2011). Semantically Enabled Experts Finding System - Ontologies, Reasoning Approach and Web Interface Design. ADBIS.

Al-Taie, M. Z., Kadry, S., & Obasa, A. (2018). Understanding expert finding systems: domains and techniques. *Social Network Analysis and Mining*, *8*(1).

Balog, K., Azzopardi, L., & Rijke, M. d. (2006). *Formal models for expert finding in enterprise corpora* Proceedings of the 29th annual international ACM SIGIR, Seattle, Washington, USA.

Balog, K., Fang, Y., Rijke, M. d., Serdyukov, P., & Si, L. (2012). Expertise Retrieval. *Found. Trends Inf. Retr.*, *6*(2–3), 127–256.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *3*(J. Mach. Learn. Res.), 993–1022.

Bogers, T., Kox, K., & van den Bosch, A. (2008). Using Citation Analysis for Finding Experts in Workgroups.

Bukowska, E., Kaczmarek, M., Stolarski, P., & Abramowicz, W. (2012). Ontology-Based Retrieval of Experts – The Issue of Efficiency and Scalability within the eXtraSpec System. Multidisciplinary Research and Practice for Information Systems, Berlin, Heidelberg.

Campbell, C. S., Maglio, P. P., Cozzi, A., & Dom, B. (2003). *Expertise identification using email communications* Proceedings of the twelfth international conference on Information and knowledge management, New Orleans, LA, USA.

Çelik, D., Karakas, A., Bal, G., Gultunca, C., Elçi, A., Buluz, B., & Alevli, M. C. (2013). Towards an Information Extraction System Based on Ontology to Match Resumes and Jobs. *2013 IEEE 37th Annual Computer Software and Applications Conference Workshops*, 333-338.

Deng, H., King, I., & Lyu, M. R. (2008). Formal Models for Expert Finding on DBLP Bibliography Data. *2008 Eighth IEEE International Conference on Data Mining*, 163-172.

Ding, Y. (2011). Topic-based PageRank on author cocitation networks. *JASIST*, *62*(3), 449-466.

Ding, Y., Yan, E., Frazho, A., & Caverlee, J. (2009). PageRank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, *60*(11), 2229-2243.

Fiala, D., Rousselot, F., & Ježek, K. (2008). PageRank for bibliographic networks. *Scientometrics*, *76*(1), 135-158.

Gomez-Perez, A., & Corcho, O. (2002). Ontology languages for the Semantic Web. *Intelligent Systems, IEEE*, *17*, 54-60.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*, 199-220.

Husain, O., Salim, N., Alias, R. A., Abdelsalam, S., & Hassan, A. (2019). Expert Finding Systems: A Systematic Review. *Applied Sciences*, *9*(20).

J. Shang, X. Z., L. Liu and S. Li. (2020). NetTaxo: Automated Topic Taxonomy Construction from Text-Rich Network. Proceedings of The Web Conference.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM*, *46*, 604-632.

Lin, S., Hong, W., Wang, D., & Li, T. (2017). A survey on expert finding techniques. *J. Intell. Inf. Syst.*, *49*(2), 255–279.

Liu, X., Song, Y., Liu, S., & Wang, H. (2012). Automatic taxonomy construction from keywords. Knowledge Discovery and Data Mining.

Noy, N., & McGuinness, D. (2001). Ontology Development 101: A Guide to Creating Your First Ontology. *Knowledge Systems Laboratory*, *32*.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank Citation Ranking : Bringing Order to the Web. The Web Conference.

Petkova, D., & Croft, W. (2008). Hierarchical Language Models for Expert Finding in Enterprise Corpora. *International Journal on Artificial Intelligence Tools*, *17*, 5-18.

Petkova, D., & Croft, W. B. (2006). Hierarchical Language Models for Expert Finding in Enterprise Corpora. ICTAI'06,

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). *The author-topic model for authors and documents* Proceedings of the 20th conference on Uncertainty in artificial intelligence, Banff, Canada.

Shang, J., Zhang, X., Liu, L., Li, S., & Han, J. (2020). *NetTaxo: Automated Topic Taxonomy Construction from Text-Rich Network* Proceedings of The Web Conference 2020, Taipei, Taiwan.

Song, Y., Liu, S., Liu, X., & Wang, H. (2015). Automatic Taxonomy Construction from Keywords via Scalable Bayesian Rose Trees. *IEEE Transactions on Knowledge and Data Engineering*, *27*(7), 1861-1874.

Tang, J., Sun, J., Wang, C., & Yang, Z. (2009). *Social influence analysis in large-scale networks* Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, Paris, France.

Uddin, M. N., Duong, T. H., Oh, K.-j., & Jo, G.-S. (2011). An Ontology Based Model for Experts Search and Ranking. Intelligent Information and Database Systems, Berlin, Heidelberg.

Wu, H., Pei, Y., & Yu, J. (2009). Hidden Topic Analysis Based Formal Framework for Finding Experts in Metadata Corpus. 2009 Eighth IEEE/ACIS International Conference on Computer and Information Science.

Zhang, C., Tao, F., Chen, X., Shen, J., Jiang, M., Sadler, B. M., Vanni, M. T., & Han, J. (2018). T*axoGen: Unsupervised Topic Taxonomy Construction by Adaptive Term Embedding and Clustering*. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK.

Zhang, J., Tang, J., & Li, J. (2007). Expert Finding in a Social Network. In R. Kotagiri, P. R. Krishna, M. Mohania, & E. Nantajeewarawat, *Advances in Databases: Concepts, Systems and Applications*.