ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

Mexpert: An Algorithm for Finding Cross-disciplinary Experts Using Data Mining Techniques

Chaisiri Sanitphonklang[1] and Nuanwan Soonthornphisaj[2]

ABSTRACT

Multidisciplinary research is the practice of bringing together experts from different disciplines or fields of study to work collaboratively on a common problem, project, or research question.This study proposes an expert finding algorithm, namely Mexpert, to bridge the gap between experts using data mining techniques. Our method has three steps. The first process is data preparation, which involves extracting information from research papers. The first process includes collecting data, extracting keywords, and discovering topics. The second process involves creating a model to group experts with similar knowledge and expertise. The last process involves analyzing the expert profiles using their experience in terms of h-index, average number of publications and total citations, followed by profile ranking. The corpus contains 17,679 papers obtained from SCOPUS. The experimental results reveal that there are four clusters out of 7 fields. We analyzed each cluster obtained from Mexpert and found that most cluster members published multidisciplinary research papers together. These results suggested that our approach can be applied to find a group of experts with different expertise.

Article information:

Keywords: Clustering, K-Means, Elbow Method, Data Mining, Expert Profile

Article history:
Received: August 30, 2023
Revised: October 5, 2023
Accepted: November 24, 2023
Published: December 9, 2023 (Online)

DOI: 10.37936/ecti-cit.2023174.253986

1. INTRODUCTION

Solving complex research problems requires experts from different domains to work together to bring their unique expertise to the table. For example, addressing environmental issues, healthcare challenges, or urban planning may require input from ecologists, economists, medical professionals, engineers, and social scientists. [1]. During the COVID-19 pandemic, there are some examples of multidisciplinary research work [2-3]. For example, cancer prediction is solved using two experts: computer scientists specialising in image processing and gastroenterologists with domain knowledge [4]. The study shows that the collaboration of domain experts and AI experts can enhance prediction performance and reveals that the collaboration of cross-researchers is perfect for overcoming specific problems. Moreover, Zeng et al. state that working in teams with different expertise has improved initiative and efficient research development (Zeng et al. [5]).

Creating a group of multidisciplinary researchers is a challenge for research institutions. According to Scimago Journal Country Rank (SJR), a data source for 27,955 journals worldwide, only 144 are classified as multidisciplinary journals, representing 0.51% of the total number of available research journals worldwide (`www.scimagojr.com/journalrank.php`).

To create multidisciplinary research, a practical methodology for expert finding is needed to bring together experts to conduct a multidisciplinary research project [6-7]. Developing an expert finding algorithm could benefit the research team to solve complex research problems.

This paper aims to apply the clustering technique to discover a group of multidiscipline researchers. According to our previous work [8], we successfully developed a framework to find groups of researchers from 54 departments at Kasetsart University who have the potential to join multidisciplinary research. To clarify, we extend our previous work by exploring new research domains to ensure that the proposed method is appropriate. Expert finding typically involves expert profile creation and ranking [9]. Our framework consists of three processes. The first process is data preparation, which involves extracting

[1,2] The authors are with Department of Computer Science, Faculty of Science, Kasetsart University, Bangkok, Thailand, E-mail: chaisiri.s@ku.th and nuanwan.s@ku.th

information from research papers. This includes paper collecting, keyword extraction, and topic discovery. The second process involves profile clustering to group experts based on their expertise. The last process is to find the ranking to identify the researcher who should join the multidisciplinary research.

The rest of this paper is structured as follows: Section 2 reviews various studies on expert finding; Section 3 presents our proposed method; Section 4 presents the experimental results of Mexpert and Section 5 provides the discussion and conclusions, as well as the directions for future work.

2. RELATED WORK

The following sections present the literature reviews to understand the research background better.

2.1 Finding experts for multidisciplinary research

Over the past few decades, significant advancements have been achieved in this field of research. Nonetheless, finding researchers who can lead research integration and solve complex and challenging problems requires individuals who possess knowledge across various disciplines and can combine their expertise to tackle difficult research questions.

Expert Finding is a tool to find and determine the expertise of individuals in the recruitment industry. Typically, five steps are involved in locating an expert [10]: expertise evidence selection, expert representation, modeling, model evaluation, and user interaction design. Identifying a person's area of expertise depends on the information and knowledge relevant to that person, which can be obtained from databases, archives, and referral networks [11]. Research on commonly developed expert findings can be categorized into two groups. First, search for professionals to join the event and Find experts to answer questions on the weblog.

An expert search application, Fexpert [12], aims to assist users in finding specific expertise by entering research keywords. The system fetches information and displays a list of experts whose expertise matches the entered keywords. The objective is to identify groups of researchers working in the same field but have yet to work together. Their corpus contains research works and a list of keywords.

Shidiq Al-Hakim [13] introduced a technique for small industries to find suitable experts for collaboration in innovation. The method involved using a content portfolio and an expert topic map generated through the TFIDF-VSM technique. This approach produced satisfactory outcomes.

Wei Liang *et al.* [14] presented a method for introducing academic cross-disciplinary collaboration based on big education data and experiences to discover potential research areas and recommend valuable cooperation. This work's limitation was the time it took to process researchers' profile data.

The Topic Professional Level Model (TPLM) identified individuals with relevant knowledge or experience responding to a specific inquiry [15]. TPLM employed topic and professional-level modeling to evaluate a user's expertise in a particular subject by analyzing their textual content and link structure. The TPLMRank algorithm calculates a user's score and identifies expert users. Based on experiments conducted on the Chinese CQA platform-Zhihu dataset, the TPLM-based approach has shown to be more effective than traditional expert-finding methods.

2.2 Data Mining Technique

Data mining [16] typically involves finding valuable, new, potentially beneficial, and understandable patterns in data. This process assists in uncover relationships and global patterns hidden among vast amounts of data in large databases. Still, it is a nontrivial process of identifying valid patterns. A well-known CRISP-DM approach [17] is a flexible framework that can be adapted to various data mining and machine learning projects. This methodology provides a structured approach to data-driven projects, including data understanding and preparation, modeling, evaluation, and deployment.

The objective of the data preparation step is to do data transformation, such as feature selection and feature engineering. The goal is to create a clean, structured dataset for training and testing machine learning models. Keyword extraction plays an essential role in the domain of text analysis. A well-known Python library called RAKE (Rapid Automatic Keyword Extraction) has been applied to delete stop words and separate phrases to extract potential keywords from the text. RAKE algorithm balances short and long keyword phrases, considering the degree values from the word co-occurrence graph. The extracted keywords are essential for the topic modeling.

Latent Dirichlet Allocation (LDA) is a topic modeling algorithm used in natural language processing and machine learning to discover topics within a collection of documents [19]. Note that one of the most critical hyperparameters in LDA is the number of topics (K) the model should discover in the document collection. The choice of K is often based on domain knowledge or subject matter expertise. Selecting an appropriate value for K can significantly impact the quality of topic modeling results.

Several algorithms were proposed to decrease the human bias of topic modeling. For example, the Perplexity-K method proposed by Huang L. *et al.* [20], created the Perplexity-K curve to find the optimal number of topics. Comparative studies were conducted between T-LDA, LDA, and K-Means. The results show that T-LDA outperforms LDA and K-

Table 1: *List of journals in our dataset.*

Journal	Journal Name	The number of Paper
A.	Transactions on Pattern Analysis and Machine Intelligence (IEEE)	4,452
B.	Nature C	1,593
C.	Nature Biotechnology	1,287
D.	Nature Geoscience	1,767
E.	Studies in Mycology	1,106
F.	Veterinary Quarterly	355
G.	Energy and Environmental Science	7,117
	Total	**17,679**

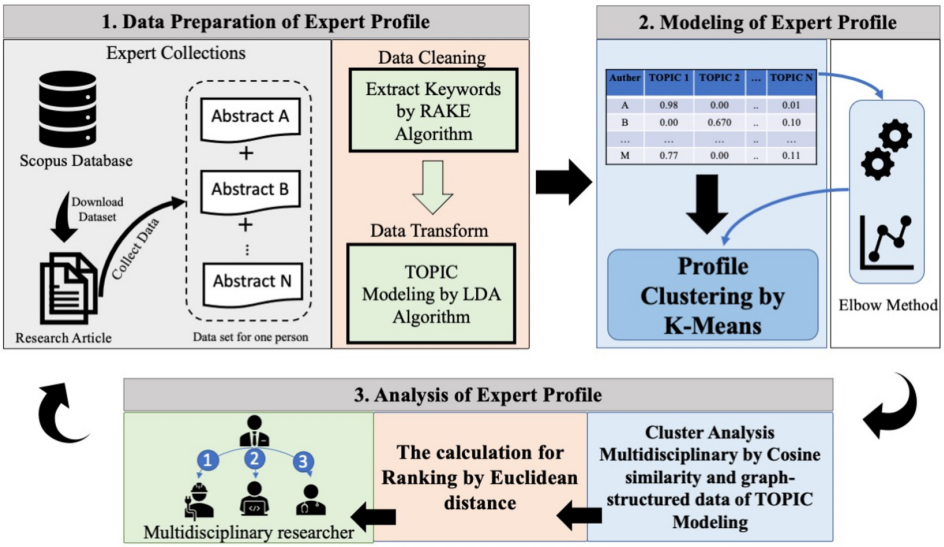


Fig.1: *Mexpert Framework.*

Means regarding topic results, modellingtime, precision, and recall rate. Another research work by Wang and colleagues [21] proposed a method that did not require iteration to assign a number of topics. This method can efficiently identify the optimal number of issues from a piece of information dataset, leading to improved accuracy of the LDA theme model.

The algorithm commonly used for data clustering is K-Means, proposed by Faizan *et al.* [22]. K-Means belongs to the class of centroid-based algorithms and is considered the simplest form of unsupervised learning. However, the limitation of K-Means is that we need to assign the number of clusters (k) in advance. The elbow method can be applied to find the value of k. [23] [24].

3. MEXPERT OF FRAMEWORK

As shown in Fig. 1, the Mexpert framework consists of expert profile preparation, profile modelling, and analysis.

3.1 Data Preparation of Expert Profile

To create the expert profile, we need a collection of research papers from the SCOPUS database. Four steps are performed as follows.

3.1.1 Data collecting

The papers published in 7 journals from 2018 to 2022 are collected. We get 17,679 papers, written by 2,312 researchers. Each paper comprise the author's name, affiliation, abstract, author keywords, index keywords, H-index and citations. Table 1 shows the number of papers from seven journals.

3.1.2 Abstract aggregation for each Expert

The purpose of this step is to aggregate all abstracts written by each researcher. To create a collection of research works for each Expert, we concatenate all abstracts written by each researcher (see Fig. 2).

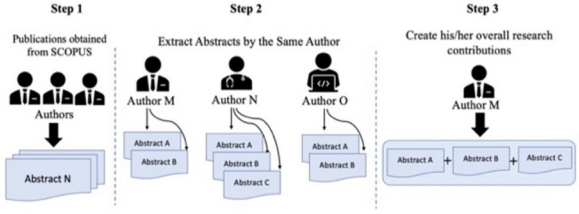


Fig.2: *Abstract Aggregation of each Expert.*

3.1.3 Keyword extraction with RAKE Algorithm

A Python library called RAKE is applied to extract keywords from each of 17,679 abstracts. RAKE works by analyzing the co-occurrence of words and their frequency in the text. Before applying RAKE, we preprocess the collection of abstracts by converting the text to lowercase, removing punctuation and special characters, and tokenizing the text into words or phrases. RAKE breaks down the abstract into individual words or phrases (candidate keywords). It does this by splitting the text based on common stop words (e.g., "and," "the," "in") and punctuation marks. Candidate keywords are sequences of words between these stop words and punctuation. RAKE calculates a score for each candidate keyword based on word frequency and co-occurrence. RAKE normalizes the scores of candidate keywords to account for the different lengths of keywords. Longer keywords tend to have higher scores due to more words, so RAKE divides the score by the number of words in the candidate keyword. The candidate keywords are ranked based on their scores, and the top keywords are selected as the most important keywords or key phrases (See Fig. 3).

```
[('out-of-plane-scattering losses caused', 28.761904761904763),
 ('partial correlated two-port signals simultaneously', 27.166666666666668),
 ('synthesize tunable non-abelian gauge fields', 27.14242424242424),
 ('dual-polarization interferometric fiber optic gyroscope',
  27.138419913419913),
 ('dual-polarization interferometric fiber-optic gyroscopes',
  26.638419913419913),
 ('work introduces real-space building blocks', 25.15),
 ('paired metal strip micro-heaters', 23.666666666666664),
 ('out-of-plane-scattering losses', 22.76190476190476 3),
 ('dual-polarization fiber optic gyroscope', 22.709848484848486),
 ('makes dual-polarization ifog configuration promising', 22.534848484848485),
 ('forward peaked rayleigh scattering light', 22.25),
 ('state-of-art fabrication precision', 22.228571428571428),
 ('achieving single-side radiative quality factors', 21.978571428571428),
 ('giant interferometric fiber optic gyroscopes', 21.570238095238096),
 ('high-sensitive refractive index sensor', 21.506410256410255),
 ('four-state modulation accurately extracts', 21.4),
 ('multiple time-reversal symmetry breaking', 20.60714285714286),
```

Fig.3: *Result of keyword extraction using RAKE library.*

3.1.4 Topic modeling using the LDA Algorithm

Latent Dirichlet Allocation (LDA) is a popular probabilistic model used in the field of natural language processing and text mining for topic modeling. Note that to assign the number of latent topics, we use the Coherence Model [25] to find the optimal number of topics from a given data set. Finally, the hyper-parameter of LDA is set to 40. After setting the hyper-parameter (number of topics), LDA uses an iterative process to refine the assignment of words to topics using probability. The result of LDA is a set of vectors called expert profiles, as shown in Fig. 4.

3.2 Modelling of Expertise Profiles

The objective of this step is to group researchers with similar research backgrounds based on their expert profiles. We applied the K-Mean clustering technique using the following steps.

AUTHOR / TOPIC	1	2	3	....	N
AUTHOR A_1	0.00	0.70	0.30	0.00	0.00
AUTHOR A_2	0.00	0.00	0.70	0.00	0.10
AUTHOR A_3	0.10	0.00	0.00	0.30	0.70
AUTHOR A_4	0.30	0.00	0.00	0.70	0.00
AUTHOR A_5	0.70	0.00	0.00	0.00	0.00
AUTHOR A_6	0.00	0.10	0.00	0.30	0.00
.....	...	...	...	...	...
AUTHOR A_m	0.00	0.00	0.00	0.00	0.70

Fig.4: *Vector of expertise profiles.*

3.2.1 Determining the Number of Clusters for K-Means Algorithm

The elbow method is applied to determine the optimal number of clusters for the K-Means algorithm. The basic idea behind the elbow method is to calculate the within-cluster sum of squares (WCSS) for different values of k (the number of clusters) and look for an "elbow" point in the plot of WCSS against k. The elbow point represents the point at which adding more clusters does not significantly reduce the WCSS and it indicates the optimal number of clusters. The elbow line intersects with the first point of k value, demonstrating the optimal value of k for the clustering algorithm. In this study, we get the optimal value of $k = 4$, as shown in Fig. 5.

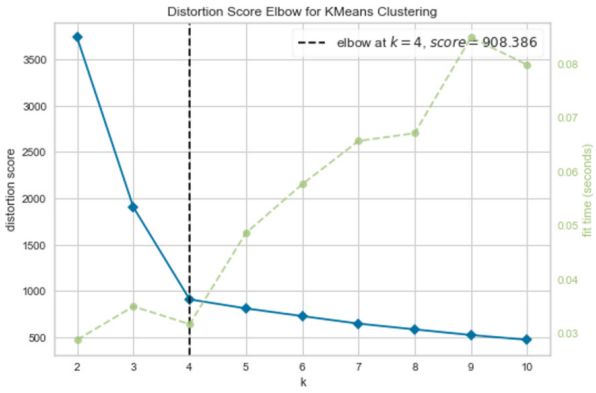


Fig.5: *Elbow Method Visualization.*

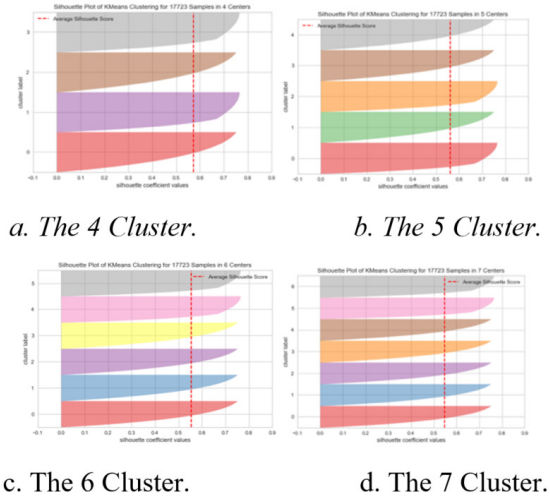
To validate the number of clusters obtained from elbow method, we evaluate the quality of clusters using Silhouette score [26]. This metric quantifies how similar each data point in one cluster is to the data points in the same cluster (cohesion) compared to the nearest neighboring cluster (separation). The Silhouette score ranges from -1 to 1, with higher values indicating better cluster quality. As illustrated in Figure 6, we found that using $k = 4$ provides the best cluster quality.

3.2.2 Expert Profile Clustering using K-Means Algorithm

The K-Means clustering algorithm was applied to group the expert profiles. We found that some ex-

Table 2: *Distribution of research papers in each cluster.*

Cluster No.	Number of researchers from Journal (A-G)							Total No. of experts
	A	B	C	D	E	F	G	
0	0	19	123	0	642	50	0	834
1	0	198	78	8	0	0	1376	1,660
2	1	1544	1864	1766	313	313	6,329	12,130
3	2849	15	131	34	8	7	11	3,055
							Total	17,679



a. The 4 Cluster. *b. The 5 Cluster.*

c. The 6 Cluster. d. The 7 Cluster.

Fig.6: *Silhouette Plot of K-Means clustering data.*

perts worked together in each cluster and published the same paper. Figure 7 depicts the result of clustering by K-Means; there are four different clusters in different colors.

We analyzed all clusters and found that cluster no. 2 and no.3 contain a variety of research papers from 7 journals. Table 2 represents the membership of 4 clusters obtained from the K-Means clustering algorithm.

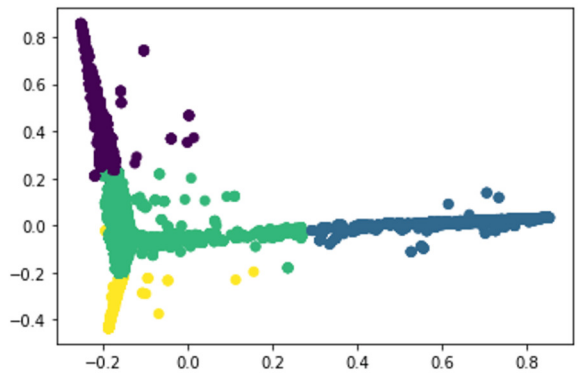


Fig.7: *Visualization of 4 expert profiles clusters.*

3.2.3 Performance Evaluation of K-Means

To evaluate the performance of the K-Means clustering algorithm, we consider the following factors:

a) Within-Cluster Sum of Squares (WCSS)

WCSS measures the total variance within each cluster. The formula for WCSS is the sum of squared distances between data points and their respective cluster centroids. As shown in Table 3, the WCSS value decreases gradually when the number of clusters increases since the reduction of data points in each cluster. Note that for $k = 4$, WCSS is 908.37.

Table 3: *Within-Cluster Sum of Squares (WCSS).*

Number of clusters (K)	The sum of the distances from all points to the Centroid.
K,(2)	3800.30
K,(3)	1980.40
K,(4)	**908.37**
K,(5)	850.76
K,(6)	790.45
K,(7)	654.32
K,(8)	602.45
K,(9)	512.34
K,(10)	498.63

b) Silhouette score

Another approach to evaluate the effectiveness of clustering is through the silhouette score. This metric considers how closely data points are grouped within their respective clusters and the distance between different clusters. It ranges between -1 and 1, where higher values indicate better clustering. Table 4 shows that the highest score is 0.58 when we set $k = 4$.

Table 4: *Silhouette Score.*

Number of clusters (K)	Silhouette Coefficient
K = 2	0.52
K = 3	0.54
K = 4	**0.58**
K = 5	0.56
K = 6	0.55
K = 7	0.53

3.3 Analysis of Expert Profiles

Expert profile analysis consists of 2 steps as follow.

3.3.1 Similarities Calculation of Expert Profiles

1. Given a set of topic vectors, we transform the elements into binary values. (LDA value greater than

0 will be converted to 1) then, we calculate the similarity of expert profiles using cosine similarity (see equation 1). Note that cosine similarity is a metric used to measure how similar two non-zero vectors are in a multi-dimensional space. It is often employed in various fields, including information retrieval, natural language processing, machine learning, and data mining.

$$CoSim(A,B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

Where,
A and B are expert profile of researcher A and B, respectively.

The examples of results are shown in Table 5.

Table 5: *The results of the computation of the similarity of the researchers' profiles.*

Author	Author A_1	Author A_2	Author A_3	Author A_4
Author A_1	1.00	1.00	0.87	0.67
Author A_2	1.00	1.00	0.76	0.67
Author A_3	0.87	0.76	1.00	1.00
...	...	...	...	...
Author A_N	0.67	0.67	1.00	1.00

2.Each cluster's sub-cluster of researchers are discovered using Algorithm1. As shown in Figure 8, the graph is created from expert profiles with the highest similarity value.

Algorithm 1: *Create Sub-cluster*

```
1.  input: LTS is the researchers' profiles.
2.  Output: Adjacency List
3.  Begin
4.  Function :
5.    fun convert (LTS) :
6.      adjList = defaultDict (list)
7.      for n in range (len(LTS)):
8.          for m in range (len (LTS[i]) )
9.              if LTS[n][m] =1.00:
10.                 adjList [i].append (j)
11.     return adjList
12. End
```

3.3.2 Ranking

New feature vectors are created using the h-index, the number of publications, and the total number of

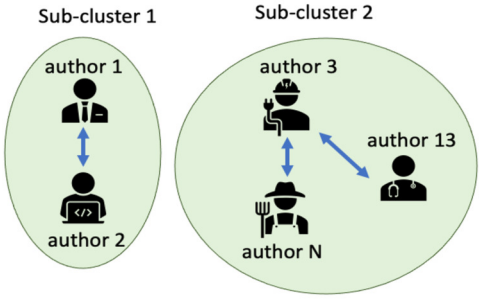


Fig.8: *Graph for Grouping Expert Profiles.*

citations of researchers. The h-index is an author-level metric that measures both the productivity and citation impact of the publications. To reduce the bias, we extract the time period (number of years) that each researcher has experienced in paper publication. Therefore, we get the feature vector as follows.

$$\text{feature vector} = \left(Hindex, \frac{no.\ of\ papers}{no.\ of\ years}, \frac{total\ no.\ of\ citation}{no.\ of\ years}\right)$$

Given a set of feature vectors, the Euclidian distance between researchers of different sub-cluster are calculated (see Fig. 9(4)). The Euclidian distance equation is shown below.

$$d(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}$$

Where,
x_1, y_1 = h-index of researcher, X and Y
x_2, y_2 = the average number of papers per year of researcher, X and Y
x_3, y_3 = the average number of citations per year of researcher, X and Y
The last step is to do an ascending sort based on the distance. The final result is the list of sorted pairs of researchers.

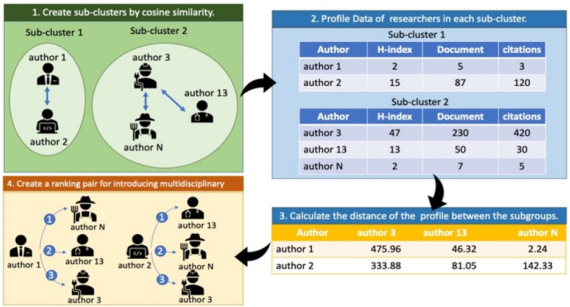


Fig.9: *presents the operating procedure.*

3.3.3 Summary of Mexpert Framework

Mexpert approach to discovering cross-discipline researchers for multidisciplinary research are summarized as follows.

Algorithm: Mexpert Algorithm
Input: *A set of Articles, A*
Output: Cluster of Researcher and profile Ranking

```
Begin

# Abstract concatenation for each Expert

data_set = Read_File (A)
for key, item in data_set :
    abstract = " "
    for paper in item["papers"]:
        abstract += paper["abstract"]
        abstract += " "
    abstract = abstract.strip( )
    dataset_abstract = df.append({Authors, Affiliation, Abstract})

# Keyword extraction with RAKE Algorithm

 dataRake ['Column_text'] = None
 dataRake ['Column_score'] = None
 for index in dataRake.index:
     text = dataRake ['Abstract'][index]
     keyphrases = Rake.run (text)
     dataRake [' Column_score '][index] = keyphrases
     dataRake [' Column_text '][index] = " ".join([keyphrase for keyphrase, score in keyphrases])

# Create a corpus from the output of RAKE.

 texts = [ ]
 for i in dataRake.index:
     tokens = dataRake [' Column_text '][i].split ( )
     for token in tokens:
         int(token)
         tokens.remove(token)
        texts.append (tokens)
 dct = Dictionary (texts)
 dct.Filter out tokens(No_below =N, No_above = M)
 for doc in texts
     corpus = dct.Convert document into BoW (doc)

# Create a set of vectors from expert profiles by LDA Algorithm.

 lda_dataset = LDA(corpus, N)
 for x in corpus
     features_vector =dict(lda_dataset[x]
 lda_vector = DataFrame(features_vector)

# Profile Clustering by K-Means Algorithm
  clusters = KMeans (k, lda_vector)
  write file ('list_Profile')

# Create a graph of Sub-cluster from expert profiles.
  adjList = defaultDict (list_Profile)
 for n in range (len(LTS)):
     for m in range (len (LTS[i]) )
       if LTS[n][m]=1.00:
           adjList [i].append (j)

# Profile Ranking step
# construct a set of feature vectors base on researcher performance

Experience =[ ]
Data_Profile = Read_data(Data_ experience.file)
for H_index, Paper, Citation in Data_Profile:
    H_index = H_index
    avgPaper = Number of Paper / Number of years
    avgCitation = Total number of Citation / Number of Year
    feature = [H_index, avgPaper, avgCitation]
    Experience.append(feature)
# do ascending sort

 for Dis_Expert in Experience:
     P = Dis_Expert[i]
   for Q in Dis_Expert:
       eDistance = Euclidean distance (P,Q)
       E_Distance.append(eDistance)
Expert_Ranking = E_Distance.Sort(reverse=False)
End
```

4. EXPERIMENTAL RESULTS

To validate our proposed method, we need a test set with ground truth. Therefore we retrieve a set of multidisciplinary research paper published in 2022. We randomly select 25 multidisciplinary research papers from 3 journals including *The Innovation, Advanced Science and Science Bulletin*. These journals are categorized as multidisciplinary research journals. After data preparation step, we get 180 researchers. Then we retrieve all published paper written by these researcher. Finally we get 3,177 papers published during 2018-2021.

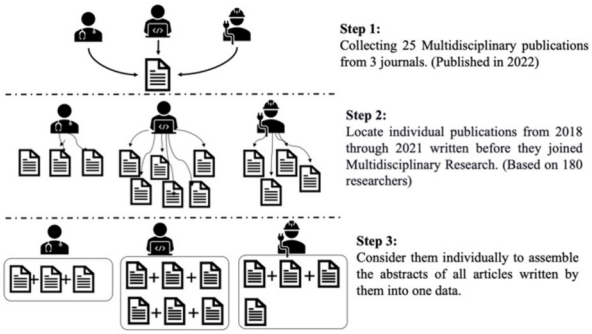


Fig.10: *Test Data Collection.*

Then, we extracted keywords from abstracts of 3,177 papers using the RAKE algorithm. After that, all processes were repeated, including LDA topic modeling and expert profile clustering. Table 7 presents the results of the Coherence method, which found the optimal number of topics for the LDA model. Note that the number of topics in this test set is 5.

Table 6: *The results of the computation of the similarity of the researchers' profiles.*

Author No.	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	0.985	0	0	0	0
2	0.909	0.087	0	0	0
3	0.987	0	0	0	0
...	...	...	...	...	
179	0.912	0	0	0	0
180	0.917	0.052	0	0.029	0

Next, a set of topic vectors from LDA are clustered by the K-Means algorithm. Note that the Elbow method recommends value of $k = 5$ clusters. The visualization of expert profile grouping is shown in Fig. 11.

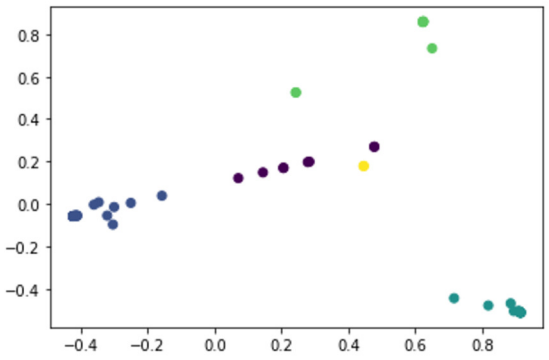


Fig.11: *Visualization of topic vectors from the test set.*

We analyzed each cluster obtained from Mexpert and found that most members of each cluster published the multidisciplinary research paper together. From 5 clusters, there are 94.33% of researchers belonged to the same cluster which shows the effectiveness of our proposed method for grouping researchers with different expertise.

Figure 12 depicts the result from Mexpert and the ground truth (25 multidisciplinary papers). We found that only a few researchers (orange color) are missing from the cluster.

5. CONCLUSIONS

The Mexpert framework is proposed to connect experts from different fields to facilitate collaboration on complex research projects. It is a decision-making tool that suggests a group of researchers from different field to work together using their research profiles.

Fig.12: *The coverage of cross-discipline researchers from 5 clusters obtained from Mexpert.*

The source of research papers are from SCOPUS database. The topic modeling are performed on abstracts after the keyword extraction process. Then the expert profiles are clustered using K-Means algorithm. The results show that experts from different research fields are grouped in the same cluster.

Repeat Steps 1 to 3 using the new data sets of multidisciplinary research researchers. The experimental results show impressive performance since most researchers are in the same cluster, accounting for 94.33%. After completing the clustering from the previous step, the sub-clusters are created using cosine similarity of expert topics vectors. Finally, the ranking step is perform using the Euclidean distance of feature vector which consist of the h-index, average number of paper per year and average of total number of citations.

A sensitive point of Mexpert is to set the number of clusters suitable for K-Means. We suggests to use Elbow method to determine the optimal number of clusters. There may be other methods to explore in the future. In the near future, we plan to extract more interesting features, such as the reference part. Furthermore, more data sets need to be explored as well

ACKNOWLEDGEMENT

We thank the Department of Computer Science, Faculty of Science and Graduate School of Kasetsart University for supporting laboratories and learning resources.

References

[1] B. C. Choi, and A.W.Pak. "Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness," *Clinical and investigative medicine*, vol.29, no. 6, pp. 351- 364, 2006.

[2] E. A. Holmes, R. C. O'Connor, V. H.Perry, I. Tracey, S. Wessely, L.Arseneault and E. Bullmore, "Multidisciplinary research priorities for

the COVID-19 pandemic: a call for action for mental health science," *The Lancet Psychiatry*, vol. 7, no. 6, pp. 547-560, 2020.

[3] S. Choudhury and A. Ghosh, "Ethical considerations of mental health research amidst COVID-19 pandemic: mitigating the challenges," *Indian Journal of Psychological Medicine*, vol. 42, no. 4, pp. 379-381, 2020.

[4] A.J. Grossberg, L. C. Chu, C. R. Deig, E. K. Fishman, W. L. Hwang,, A. Maitra,, and C.R. Thomas Jr. "Multidisciplinary standards of care and recent progress in pancreatic ductal adenocarcinoma," *CA: a cancer journal for clinicians*, vol. 70, no. 5, pp. 375-403, 2020.

[5] A. Zeng, Y. Fan, Z. Di, Y. Wang and S. Havlin, "Fresh teams are associated with original and multidisciplinary research," *Nature human behaviour*, vol. 5, no. 10, pp. 1314-1322, 2021.

[6] A. O'Hagan, C. E. Buck, A. Daneshkhah, R. Eiser, P. Garthwaite, and D. Jenkinson, "Uncertain judgements : eliciting experts' probabilities," John Wiley & Sons, Chichester. 2006, ch. 10.

[7] M.J. Caley, R.A. O'Leary, Fisher, R., Low-Choy, S., Johnson, S., and K. Mengersen, "What is an expert? A systems perspective on expertise," *International Journal of Ecology and evolution*, vol. 4, no.3, pp. 231-242, 2014.

[8] C. Sanitphonklang and N. Soonthornphisaj, "The Discovery of Experts for Multidisciplinary Research using data mining approach," *Proceeding of 22nd IEEE International Computer Science and Engineering Conference (ICSEC)*, pp. 1-4, 2018.

[9] J. Zhang, J. Tang and J. Li, "Expert Finding in a Social Network," *Proceeding of 12th International Conference on Database Systems for Advanced Applications (DASFAA)*, pp. 1066-1069, 2007.

[10] O. Husain *et al.*, "Expert finding systems: A systematic review," *Applied Sciences Journal*, vol. 9 no. 20, pp. 42-50, 2019.

[11] S. Lin *et al.*, "A survey on expert finding techniques," *Journal of Intelligent Information System*, vol. 49, pp. 255-279, 2017.

[12] A. Srivihok, "Building FExpert : System for Searching Experts in Research University Using K-Means Algorithms," *Proceeding International Conference of IEEE Symposium on Computer and Informatics (ISCI)*, pp. 176-179, 2012.

[13] S. Al Hakim, D.I. Sensuse, I. Budi, I. M. I. Subroto, and A. H. A. M. Siagian, "Expert retrieval based on local journals metadata to drive small-medium industries (SMI) collaboration for product innovation," *International Journal of Social Network Analysis and Mining*, vol. 13, no.1, pp. 68. 2023.

[14] W. Liang, X. Zhou, S. Huang, C. Hu, X. Xu, and Q. Jin. "Modellingof cross-disciplinary collaboration for potential field discovery and recommendation based on scholarly big data," *Future generation computer systems*, vol. 87, pp. 591-600, 2018.

[15] S. Wang, D. Jiang, L. Su, Z. Fan, and X. Liu, "Expert finding in CQA based on topic professional level model," *Proceeding of 3rd Springer International Publishing International Conference, DMBD 2018*, pp.17–22, 2018.

[16] A.K. Pujari, *Data mining techniques*. Universities press, India, 2001.

[17] D.L. Olson and D. Dursun, *Advanced data mining techniques*. USA: Springer Science and Business Media, 2008.

[18] R. Stuart, E. Dave, C. Nick and C. Wendy, "Text Mining : Applications and Theory," John Wiley & Sons. Ltd., United States, 2010, ch. 1.

[19] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 24-36, 2017.

[20] L. Huang, J. Ma and C. Chen, "Topic detection from microblogs using T-LDA and perplexity," *Proceeding of 24th IEEE international Conference on Asia-Pacific software engineering conference workshops (APSECW)*, pp. 71-77, 2017.

[21] H. Wang, J. Wang, Y. Zhang, M. Wang and C. Mao. "Optimization of Topic Recognition Model for News Texts Based on LDA," *International Journal Digital of Information Management*, vol. 17, no. 5, pp. 257-269, 2019.

[22] M. Faizan, M. F. Zuhairi, S. Ismail and S.Sultan. "Applications of clustering techniques in data mining: a comparative study," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, pp. 146-153, 2020.

[23] M. Cui, "Introduction to the k-means clustering algorithm based on the elbow method," *Accounting, Auditing and Finance*, vol. 1, no. 1, pp. 5-8, 2020.

[24] E. Umargono, J.E. Suseno and S.V. Gunawan, "K-means clustering optimization using the elbow method and early centroid determination based on mean and median formula," in *The 2nd International Seminar on Science and Technology (ISSTEC 2019)*, pp. 121-129, 2020.

[25] H. Amoualian, W. Lu, E. Gaussier, G. Balikas, M.R. Amini and M. Clausel, "Topical coherence in LDA-based models through induced segmentation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1799-1809, 2017.

[26] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *International Journal of Computational and Applied Mathematics*, vol. 20, pp. 53- 65, 1987.

Chaisiri Sanitphonklang received the B.S. degrees in Information Technology From Pathumthani University (PTU), Thailand, in 2004, and M.S. degrees in Computer Science from National Institute of Development Administration, Bangkok, Thailand, in 2012. He is currently pursuing a PhD in Computer Science in the Faculty of Science. Kasetsart University, Thailand. He research interests include Machine Learning, Data Mining and Internet of Things (IoT).

Nuanwan Soonthornphisaj received the B.Sc. degrees in Computer Science From Thammasat University, Thailand, in 1993, the M.S. degrees in Computer Science From Asian Institute of Technology, Thailand, in 1997, and Ph.D. degree in Computer Engineering From Chulalongkorn University, Thailand, in 2002. She is currently an Associate Professor with the Kasetsart University, Thailand. Her research interests include Machine Learning and Data Mining.