The 9th International Conference on e-Business (iNCEB2010)
November 18th – 19th, 2010

Segmentation of University Experts by K-Means Algorithm and Feature Weighting Techniques

Jaruwan Kanjanasupawan
Department of Computer Science
Kasetsart University
Bangkok, Thailand
e-mail: g5064171@ku.ac.th

Anongnart Srivihok
Department of Computer Science
Kasetsart University
Bangkok, Thailand
e-mail: fsciang @ku.ac.th

Abstract— **One type of knowledge banks, named University where domain experts are driving forces in teaching, academic services and research. In big universities, there exist many domains such as agriculture, arts, economics, engineering, medical science, information technology, education, and social science are offered for studying. Searching who is an expert in which area is lengthy. The objective of this study is to propose algorithms for clustering experts from their past studies and research works. Data set were collected from employee profiles of a public university of Thailand. Series of data mining techniques including attribute selection, attribute weighting and two step data clustering by SOM, and K-Means algorithms were used to segment experts into specific groups. These three candidate features weighting techniques include TF-IDF (Term Frequency-Inverse Document Frequency), logarithm weight and augmented weight. It seemed that feature weighting could be used to improve clustering performances. Recommendations for future research are provided.**

Keywords-Expert Clustering; SOM; K-Means; Feature Weight; Expert Extraction

I. INTRODUCTION

Essential knowledge resource of most organization is obtained from experts. The knowledge is important for management, innovation and development of new products or services. Universities are organizations which are able to create and transfer knowledge to society. In the universities, there are several experts who are specialists in different domain. It is a difficult to know who are expert in which knowledge domain. One may guess from the department. For example, both experts from faculty of Economic and faculty of Agriculture who are experts of corn domain. So we ought to cluster experts in the same domain. It is a most basic way.

This experiment introduces clustering techniques for university experts of Thailand by two stage algorithms. SOM (Self- Organizing Maps) was used to determine the best number of clusters. For clustering techniques, we use K-Means algorithm and feature weighting techniques based on TF-IDF (Term Frequency-Inverse Document Frequency), log weight (logarithm weight) and augmented weight. The feature weighting techniques have been use data mining domain such as the clustering dataset and may the better value. In section II, discusses about related work. In section III refers to clustering methods, SOM and K-Means. The Feature weighting techniques demonstrate in section IV. In section V and VI explain the feature selection techniques and clusters evaluators. For section VII, we describe the experiment's processes and results. Last, we conclude the experiments and future works in section VIII.

II. LITERATURES REVIEW

Mahdavi et al. [1] studied clustering techniques on web documents by using K-Means, Harmony, hybrid Harmony K-Means and integrated K-Means in Harmony. The efficient method for clustering was integrated K-Means in Harmony and the iteration efficient method was K-Means. In [2], Pojpongsan and Srivihok demonstrated unsupervised clustering of Food Safety documents by two stage algorithms. SOM was used to determine the number of clusters. For clustering techniques, K-means and Genetic algorithm (GA) were used. The quality of clustering was better for GA while K-means was better for consuming less time. Wongpun and Srivihok [3] experimented attribute selection for classification of bad behaviors of vocational education students. They compared three feature selection techniques such as (1) Correlation-based Feature Selection or CFS (2) Consistency-based Subset Evaluation and (3) Wrapper Subset Evaluation. Then, data set were classified by Naïve Bays, Baysian Belief Network, Ripper and C4.5. Results showed that hybrid classification techniques that was the best efficiency included both genetic search and CFS for *attribute selection* and C4.5 for *classification* algorithm. Lan et al. [4] studied feature weighting techniques for text mining. They offered various techniques (1) TF-IDF (2) logarithm weight, (3) TF.RF (Term Frequency-Relevance Frequency). Data set was sampled from Reuter News Agent. The best performance algorithm was TF.RF and logarithm weight is the second rank.

III. CLUSTERING TECHNIQUES

A. SOM (Self–Organizing Maps)

Self-Organizing Maps (SOM) or Kohonen's map developed in 1982 by Kohonen. SOM is an algorithm for unsupervised learning. The algorithm is mostly used to determine appropriate the number of clusters [5]. SOM is a neural network algorithm that networks have only input layers and output layers [6].

B. K-Means

K-Means is a partitioning method for clustering. It required assigning the number of clusters. The calculation is based on the distances of objects, inter-group should be maximum, and minimum for intra-group [6].

IV. FEATURE WEIGHTING

Feature weighting is used often for data mining and information retrieval tasks. The feature weighting techniques is considered base on terms and document frequencies. We refer to the feature weighting techniques including TF-IDF (Term Frequency-Inverse Document Frequency), logarithm weight and augmented weight.

A. TF-IDF Feature Weighting Techniques

TF-IDF (Term Frequency-Inverse Document Frequency) is a favorable technique that mostly used weighting features. It contemplates base on term frequencies and document frequencies. It calculates as [7]:

$$\text{TF-IDF}_{t,d} = \text{TF}_{t,d} * \text{IDF}_t \quad (1)$$

$TF_{t,d}$ –Term frequencies which occurs in document d.
IDF_t – Scale of terms weight.

IDF_t value define as [7]:

$$\text{IDF}_t = \log \frac{N}{df_t} \quad (2)$$

N –Total number of documents.
df_t –Number of documents which contains term t.

If TF-IDF value is the highest, term t occurs many times within less number of documents. In opposition, term t occurs in most documents.

B. Logarithm Feature Weighting Techniques

Logarithm weight applied from TF-IDF technique. It added logarithm functions that expanded term frequencies. The Formula logarithm weight considered as [4]:

$$\text{Logarithm Weight} = \log(1+\text{TF}_{t,d}) * \text{IDF}_t \quad (3)$$

$TF_{t,d}$ –Term frequencies which occurs in document d.
IDF_t – Scale of terms weight that followed formula (2).

C. Augmented Feature Weighting Techniques

Augment weighting method related TF-IDF technique too. It may consider the length of documents in collection. For weighting computed as [8]:

$$\text{Augmented} = 0.5 + \left(\frac{0.5 * \text{TF}_{t,d}}{\text{Max}_t(\text{TF}_{t,d})}\right) * \text{IDF}_t \quad (4)$$

$TF_{t,d}$ –Term frequencies which occurs in document d.
IDF_t – Scale of terms weight that followed formula (2).

In this case, we do not consider any length documents. We are using 0.5 for weighting because it is a standard value.

V. FEATURE SELECTION

The large features are reasons to consume the resources, and memories. The Feature weighting is a technique to solve this problem. The Hybrid techniques are CFS (Correlation-based Feature Subset Selection) and genetic search.

A. CFS Subset Evaluator

CFS subset evaluator evaluate base on correlation based heuristic and best first search. The efficiency correlation of features determine by maximum merit value. Heuristic calculable will give the highest score of attribute groups which has high relationship and the types of data which have low inter-related relationship in each attribute. It will imply the hypothesis of feature of each group. It can dispel the unrelated attribute. Merit value showed as [3]:

$$\text{Merit}_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad (5)$$

Merits - Group of attribute S including k attributes that chosen from all attributes
r_{cf} - Average value of the groups of chosen attributes from all attribute which have related relationships with the type of data
r_{ff} - Average value of groups of chosen attributes from all attributes which have the inter-related relationship in the same group of chosen attributes

B. Genetic Search

In general CFS subset selection selected attributes based on best first search. The hybrid techniques, CFS subset evaluator and genetic search are the fine technique that referred from Wongpun and Srivihok [3]. If we used other hybrid methods in this case, it cannot indicate any relationship of features. Main processes of genetic search explain as [3, 9]. We calculate fitness from the best Merit values.

VI. CLUSTERING EVALUATORS

The clustering acceptation can estimate by evaluators of clustering. In this work, we are using 3 evaluators as: F-Statistic, R-Squared and Silhouette.

A. F-Statistic

F-Statistic is estimated from statistical analysis to test the ratio between mean square of treatment (summarization distance between groups divided by number of groups to minus 1) and mean square error (summarization of distance within samples divide by number of samples to minus number of groups). F-Statistic should to highest, intergroup are district segment [10].

B. R-Squared

R-Squared (RS) is an estimation index for measurement dissimilarity of groups. The values have ranging between 0 and 1 (0 means high similarity among the clusters, 1 means in opposite). The RS values should be highest because it show farthest distances between of each cluster. RS values can be calculated as follows [10]:

$$RS = 1 - \frac{\sum_{i=1}^{k} \| x_i - \bar{x}_k \|}{\sum_{i=1}^{n} \| x_i - \bar{x} \|} \quad (6)$$

k – Number of clusters
x_i – Data i^{th}
$\bar{x}$ - Mean values of data
n – Number of data
$\bar{x}_k$ - Mean values of cluster

C. Silhouette

Silhouette used to estimate intra-group of clusters. The Silhouette are ranging between -1 to 1 (if the values are close 1, the average distance of clusters are minimum). The average distances within group calculate by (a_i) and the average distances among clusters calculate by (b_i):

$$a_i = \frac{1}{|C_j|} \sum_{r \in c_j} d_p(a_i, x) \quad (7)$$

$$b_i = \frac{1}{|C_j|} \sum_{r \in c_h}^{h \neq j} d_p(a_i, x) \quad (8)$$

C_j – Cluster j^{th}
C_h – Cluster h^{th}
x – Data which interest
d_p – Distance between average distances within group and data x

Silhouette calculates as:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \quad (9)$$

The average value of Silhouette for clustering is the summation of all Silhouettes divided by number of clusters. The Silhouette should be highest that are well clustered [11, 12].

VII. Experiment And Results

The processes of experiment follow from figure I.

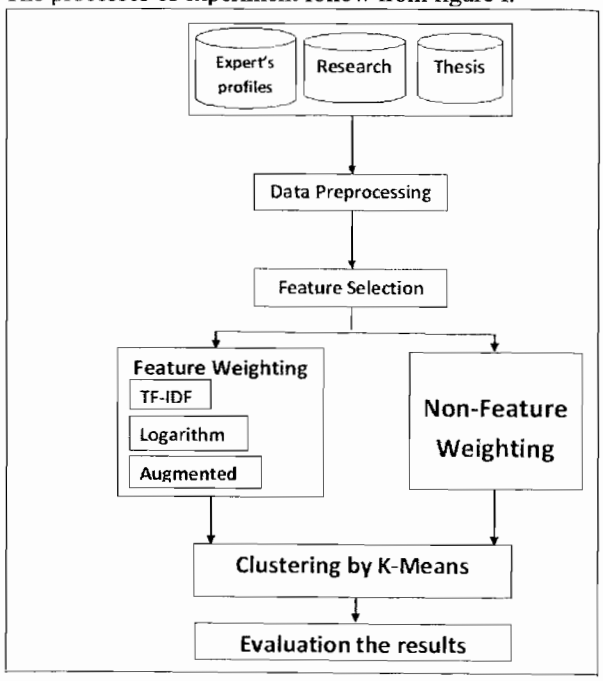

Figure I. Overall the processes of experiment

A. Data & Data Preprocessing

Expert dataset for experiment provided by one public University of Thailand included

1) Expert's profiles: names, date of birth, responsibilities (faculty), study major, degree.

2) Expert's research profile: research grant title, and publication title.

3) *Theses profile*: Thesis title, advisor name, department major.

After combining three profiles, there were 3,194 staff records. First step, keywords which resides in new combined profile were extracted. Keywords are assigned by the experts who owned the research papers or thesis. These keywords were assumed as related to the area of expertise. After keywords extraction, there were 971 keywords, so the beginnings of dataset are 3,194 rows and 971 attributes. The keywords were appraised occurrence of frequencies in each document. It was showing in table I.

B. Feature Selection

Initial data before preprocessing included a large amount of features. It might consume large resources, time and memories. So features selection is a method to decrease the features which are useless. This experiment used CFS subset evaluator and Genetic search to select the attributes. Weka 3.5.8 miner are using for feature selection. The efficient feature selection was 100 population sizes and the features included 258 attributes. So the dataset for next process

included 3,194 rows and 258 attributes. Then, the estimation of term frequencies in each row was showed in table II.

Table I: Dataset's example before preprocessing

Identities	Name	Surname	Project Name	Keyword
12345	Somsri	Suksun	Japanses Parental Time and Time and Budget.	Japan, Child
22456	Somchai	Rukdee	Landscape and Tourism Development Master Plan	Landscape, Tourist
55643	Somsuk	Suksri	Pilot Project of Sufficient Economy for Sustainable Quality of Life of Farming	Pilot, Economic, Farm
...	...	...	...	...

Table II: Dataset's example is estimated for the frequency of keywords

Expert ID	Keywords ID			
	1	2	3	...
1	10	0	3	...
2	1	0	5	...
3	1	2	12	...
...	...	...	...	...

C. *Estimation Number of Clusters*

A problem of cluster, we cannot to estimate the number of clusters. Dataset have not any label in first. We used two stage cluster algorithms in this experiment. First steps, we determine the number of clusters by SOM. The results showing in figure II and III.

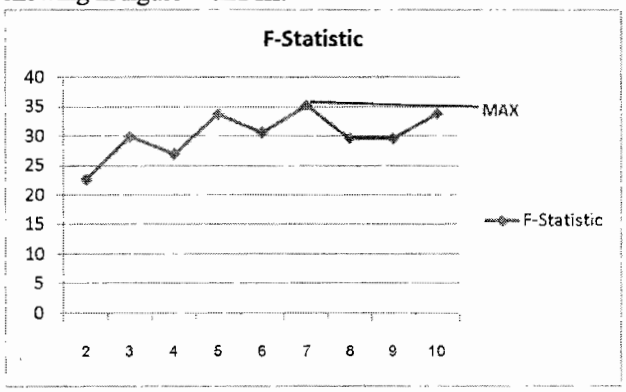

Figure II. F-Statistic for the number of clusters by SOM

From figure II and III, we processed the number of clusters from 2 to 10. The optimization results estimated 7 clusters by best values F-Statistic and R-Squared (RS).

Next question, the results of estimation in (A) may be better than (B). So in table III, we are comparing 7 groups between data with feature selection and data non-feature selection.

The table III evaluated data with feature selection finer than data non-feature selection. Although F-Statistic of data with feature selection is fewer optimization than data non-feature selection, they are held better when calculated Silhouette together.

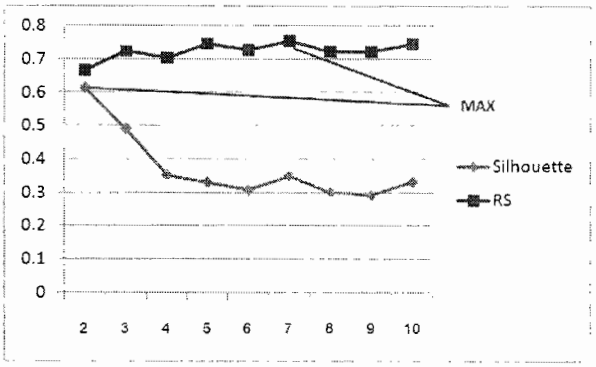

Figure III. RS and Silhouette for the number of clusters by SOM

Table III: Comparison between feature selection and non-feature selection of data set before clustering with SOM into 7 groups.

Number of Groups	Features Selection	F-Statistic	RS	Silhouette
7	No	35.07	**0.93879**	-0.079855
7	Yes	**35.08**	0.75513	**0.35024**

D. *Experts Clustering*

For experts clustering section will explain the results of experts clustering. The clustering processes are using MATLAB 7.0.1. We comparing K-Means with feature weighting techniques such as TF-IDF, logarithm weight and augmented weighting techniques. We running with 7 clusters and data feature selection those determined from C in table IV.

From table IV, K-Means (logarithm weight) was capable clusters. We estimated by F-Statistic and R-Squared while Silhouette was the best with K-Means (TF-IDF).

Table IV: Comparison performances of clustering with K-Means algorithm combined with TF-IDF, logarithm weight and augmented weighting techniques.

Algorithms	Evaluators		
	F-Statistic	RS	Silhouette
K-means	56.586	0.20191	0.83261
K-means(TF-IDF)	56.934	**0.68768**	0.83346
K-means(log)	**79.234**	0.46166	**0.87445**
K-means(augmented)	50.236	0.30611	0.81536

Next step, we will extract the expertise from this and consider the experts in each faculty.

The expert belongs to different groups of expertise. There were 9 domains according to their affiliation as:

(1) Science segment with 432 rows.

(2) Engineering segment with 480 rows.

(3) Agriculture, Natural Resources and Agro-Industry segment with 642 rows.

(4) Business Administration and Economics segment with 231 rows.

(5) Social Sciences segment with 220 rows.

(6) Education segment with 441 rows.

(7) Veterinary Medicine segment with 170 rows.

(8) Research and Development Institute 258 rows.

(9) Offices in the University with 320 rows.

Thus, the total included 3,194 rows, the percentage were depicted in table V

Table V: Percentage of the experts in 9 domains

Grouping by Affiliation	Percents
Science	13.53%
Engineering	15.03%
Agriculture, Natural Resources and Agro-Industry	20.1%
Business Administration and Economics	7.23%
Social Sciences	6.89%
Education	13.81%
Veterinary Medicine	5.32%
Research and Development Institute	8.08%
Offices in the University	10.02%

Table VI: The most frequency keywords in seven clusters of experts by using K-Means algorithm with logarithm weighting features.

	Keyword ranking based on frequency				
Clusters	**Rank1**	**Rank 2**	**Rank3**	**Rank 4**	**Rank 5**
C1	เศรษฐ-ศาสตร์	Econo-mic	ทุน	เงิน	Finance
C2	ป่า	Generate	ปลา	ไข่	ท่องเที่ยว
C3	Rice	ร้อน	คาร์บอน	ไดออกไซด์	สตาร์ช
C4	Corn	หวาน	Baby	Sweet	Rice
C5	สอน	Student	Ad-min	ช่วยสอน	เรียน
C6	Econo-mic	สอน	เศรษฐ-ศาสตร์	Genetic	Indus-trial
C7	ไฟฟ้า	Indus-trial	ไร้สาย	Fire	Safe

Footnote: C^{th} means cluster th.

In table VI presented top five keywords that had most frequencies in each cluster.

Cluster 1 (C1), keywords include "เศรษฐศาสตร์" (Economic), "ทุน" (capital), "เงิน" (money), and "finance". This cluster clearly represents Economics and Finance.

Cluster 2 (C2), keywords include "ป่า" (Forest), generate, "ปลา" (fish), "ไข่" (egg), and "ท่องเที่ยว" (tourism). This cluster represents Forestry or Agriculture.

Cluster 3 (C3), keywords include "Rice", "ร้อน" (hot), "คาร์บอน" (Carbon), "ไดออกไซด์" (dioxide), and "สตาร์ท" (starch). In this cluster, it might represent Agriculture or Environmental Science.

Cluster 4 (C4), the word "Corn", "หวาน" (sweet), baby, "sweet", and "rice". In this cluster, it might represent Agronomy.

Cluster 5 (C5), the word "สอน" (Teaching), student, Admin, "ช่วยสอน" (teaching assistance), and "เรียน" (learning). This cluster clearly represents Education.

Cluster 6 (C6), the word "Economic", "สอน" (Teaching), "เศรษฐศาสตร์" (economics), "Genetics", and "Industrial". In this cluster, it might represent Economics, Education, Science and Engineering.

Cluster 7 (C7), the word "ไฟฟ้า" (Electricity), Industrial, "ไร้สาย" (wires), "fire", and "safe". In this cluster, it should represent Engineering.

From the clustering results in table 6 & 7, it seemed that Cluster 6(C6) was the largest one (2,671 experts) with mixed domain of knowledge in Economics, Education, Science and Engineering. Cluster 2 was the second largest (249 experts), expertise in Forestry or Agriculture. Third rank was Cluster 7 (107 experts), expertise in Engineering. Cluster 3 (63 experts) included expertise in Agriculture or Environmental Science. Cluster 1 (50 experts) represented Economics and Finance. Cluster 5 (43 experts) represented Education. Lastly, the smallest cluster, Cluster 4 (11 experts) represented Agronomy.

Table VII: The number of members of K-Means (logarithm weight) techniques with 7 groups

Cluster	Number of experts in each cluster
C1	50
C2	249
C3	63
C4	11
C5	43
C6	2,671
C7	107

Some word in Thai and English had the same mean because some the project's name was Thai or English any kind.

VIII. Conclusions and Future Work

Experts of one public university in Thailand were clustered by using two stage algorithms: SOM and K-Means algorithms. SOM was used to estimate the number of clusters then data were clustered by K-Means algorithm with feature weighting by three techniques: TF-IDF, logarithm weight and augmented weight.

The dataset were preprocessing, they consisted of 3194 rows and 971 attributes. Since, the dataset had a large number of attributes, there were reasons to consume large amount of resources. Feature selection was a technique to decrease the features numbers. In this study, a hybrid techniques, CFS subset evaluator and Genetic search were used to remove some non significant feature. After feature selection, dataset features were decreased to 258 attributes.

Since, dataset had no label and the initial number of clusters was unknown. SOM was applied to determine the number of clusters which was 7 clusters. Then data set were clustered by K -Means algorithm coupling with four options: (1) only K-Means (2) TF-IDF, (3) logarithm weight and (4) augmented weight techniques. The measures were F-statistics, R squared and Silhouette. The algorithms performing best results were K-Means with logarithm weighting technique.

Based on the clustering results, data set were divided into seven clusters. After rating keywords frequency in each cluster, it can conclude area of experts by Cluster. Cluster1 represents Economics and Finance, Cluster2: Forestry/Agriculture, Cluster3: Agriculture/Environmental Science. Cluster4: Agriculture-Agronomy, Cluster5: Education, Cluster6: Economics, Education and Science. Cluster7: Engineering. It seemed that smaller cluster can be a good candidate for domain clustering.

For future work, it will be fruitful to set an experiment by adding another algorithm to clustering algorithm aside from K-Means such as using ontology for calculating feature weights.

REFERENCES

[1] M. Madahvi, M. H. Chehreghani, H. Abolhassani and R. Forsati, "Novel meta-heuristic algorithms for clustering web documents," in Applied Mathematics and Computation vol. 201, pp. 441-451, 2008.

[2] W. Pojpongsan and A. Srivihok, "Unsupervised Data Clustering using Hybrid Genetic Algorithms," in The 5th International Conference on e-Business (NCEB2006), 2006.

[3] S. Wongpun and A. Srivihok, "Comparison of Attribute Selection Techniques and Algorithms in Classifying Bad Behaviors of Vocational Education Students," in Proceeding of Second IEEE International Conference on Digital Ecosystems and Technologies, 2008.

[4] M. Lan, C. L. Tan, H. B. Low and S. Y. Sung, "A Comprehensive Comparative Study on Term Weighting Schemes for Text Categorization with Support Vector Machines," in International World Wide Web Conference Special interest tracks and posters of the 14th international conference on World Wide Web, pp. 1032-1033, 2005.

[5] K.J. Cios, W. Pedrycz, R.W. Swiniarski and A. Kurgan, "Data Mining: A Knowledge Discovery Approach," in Springer Science + Business Media, LLC, 2007.

[6] P. Tan, M. Steinbach and V. Kumar, "Introduction to Data Mining," in Peason Education, Inc, 2006.

[7] C. D. Manning, P. Raghavan and H. Schutze, "Introduction to Information Retrieval," download June 2008:http://nlp.stanford.edu/IRbook/pdf/irbookonlinereading.pdf.

[8] Y. Gong and X. Liu, "Generic text summarization using relevance measure and Latent Semantic Analysis," in Annual ACM Conference on Research and Development in Information Retrieval, pp. 19-25, 2001.

[9] M. A. Hall and L. A. Smith, "Feature Subset Selection: A Correlation Based Filter Approach," in Neural Information Processing and Intelligent Information Systems, pp. 855-858, 1997.

[10] R. Matignon, "Data mining using SAS Enterprise miner," in Wiley – Interscience Press, 2007.

[11] R. Tibshirani, G. Walther and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," in Journal of the Royal Statistical Society, Series B (J. Royal Statist. Soc. B), pp. 411-423, 2001.

[12] N. Fanizzi, C. d' Amato and F. Esposoto, "Randomized Metric Induction and Evolutionary ConceptualClustering for Semantic Knowledge Bases," in ACM Conference on Information and Knowledge Management (CIKM), pp. 51-60, November 2007.