Collaborative Filtering Item Recommendation Methods based on Matrix Factorization and Clustering Approaches

Noor Ifada [1)], Mochammad Kautsar Sophan [2)], Moh. Nurun Fitriantama [3)], Sri Wahyuni [4)]
[1,2,3] Informatics Department, University of Trunojoyo Madura, Bangkalan, Indonesia
[4] Electrical Engineering Department, University of Trunojoyo Madura, Bangkalan, Indonesia
Email: noor.ifada@trunojoyo.ac.id, kautsar@trunojoyo.ac.id, ryan.cwok@gmail.com, s.wahyuni@trunojoyo.ac.id

***Abstract*—This paper studies various collaborative filtering item recommendation methods based on matrix factorization and clustering approaches. We develop six methods that are modified based on three matrix factorization approaches, i.e., Non-Negative Matrix Factorization (NMF), Singular Value Decomposition (SVD), and Principal Component Analysis (PCA); and two clustering techniques, i.e., K-Means and Fuzzy C-Means. The framework of method development consists of three main phases: the matrix factorization for generating the latent factors; the users clustering for grouping similar users; and the user-based collaborative filtering for generating the item recommendation. The recommendation performances are evaluated in terms of F1-Score and Normalized Discounted Cumulative Gain (NDCG) metrics, at various top-N. Using the MovieLens rating dataset, experiment results show that the combination of PCA and K-Means outperforms the other five methods, where the global averages of outperformance are 82.05% in terms of F1-Score and 52.10% in terms of NDCG. It is also worthwhile to note that the combination of K-Means clustering with any matrix factorization techniques is superior compared to that of the Fuzzy C-Means. Additionally, the advantage of implementing the matrix factorization approach depends on which clustering technique used in the method.**

Keywords—clustering, collaborative filtering, item recommendation, matrix factorization

I. Introduction

Collaborative Filtering is widely studied and implemented as the learning model in Recommendation Systems [1]. Methods that implemented such a model generate the list of top-N item recommendations to a target user based on his/her previous ratings along with other users that have considered-similar preferences [1, 2].

The traditional collaborative filtering item recommendation methods are commonly known to suffer from problems that are impacting their performances [2-6]. Researchers have proposed to enhance the methods by implementing the matrix factorization approaches [7-11] or clustering techniques [5, 12-18]. Three latent factor approaches are widely known and used in various areas, namely, Non-Negative Matrix Factorization (NMF), Singular Value Decomposition (SVD), and Principal Component Analysis (PCA) [2]. Meanwhile, K-Means and Fuzzy C-Means are the two popular clustering techniques implemented in Recommendation Systems [19].

Recent studies have shown that combining the matrix factorization approach with the clustering technique is beneficial in enhancing the performance of Collaborative Filtering methods, e.g., SVD and K-Means [20], NMF and K-Means [21], PCA and K-Means [22]. However, to the best of our knowledge, there has not been any in-depth work that investigates and compares the performances of collaborative filtering item recommendation methods based on various combinations of matrix factorization and clustering approaches.

This paper comprehensively studies various collaborative filtering item recommendation methods based on matrix factorization and clustering approaches. We develop and compare the performance of six methods that are modified based on three matrix factorization approaches, i.e., NMF, SVD, and PCA approaches; and two clustering techniques, i.e., K-Means and Fuzzy C-Means. For the empirical analysis, the experiments are conducted on the real-world MovieLens rating dataset.

The contributions of this paper are (1) presenting the performance comparisons of various collaborative filtering item recommendation methods based on matrix factorization and clustering approaches, and (2) providing recommendations on which combination of matrix factorization approach and clustering technique best used in the method.

The rest of this paper is systemized as follows. Section II details the developed collaborative filtering item recommendation methods based on matrix factorization and clustering approaches. Section III presents the empirical analysis and Section IV points out the conclusion and future work of this paper.

II. Collaborative Filtering Item Recommendation Method based on Matrix Factorization and Clustering Approaches

The input of the collaborative filtering item recommendation method based on matrix factorization and clustering approaches is rating data. Assume that $U = \{u_1, u_2, u_3, \dots, u_m\}$ is the set of m users and $I = \{i_1, i_2, i_3, \dots, i_n\}$ is the set of n items. The rating data is represented as a matrix $R \in \mathbb{R}^{m \times n}$ where $r_{u,i}$ is the rating given by a user u towards item i.

The development of the method consists of three main phases: (1) matrix factorization, (2) users clustering, and (3) user-based collaborative filtering. The framework is outlined in Fig. 1.

A. Matrix Factorization

Matrix factorization is a low-rank F approach that decomposes a matrix $S \in \mathbb{R}^{m \times n}$ as $P \in \mathbb{R}^{m \times F}$ and $Q \in \mathbb{R}^{n \times F}$ latent factor matrices such that:

$$S \approx PQ^T \tag{1}$$

We implement the matrix factorization approach to the rating matrix, resulting in users and items latent factor matrices. This paper uses three well-known variations of the matrix factorization approaches:

- NMF: a matrix factorization approach that decomposes a non-negative matrix into two non-negative latent factor matrices [23, 24].
- SVD: a matrix factorization approach that decomposes a matrix into two latent factor matrices of orthonormal singular vectors and a singular value diagonal matrix that works as the controller [11].
- PCA: a matrix factorization approach that decomposes a matrix into two latent factor matrices by using the eigenvalue decomposition of the data covariance matrix to get the principal components [25].

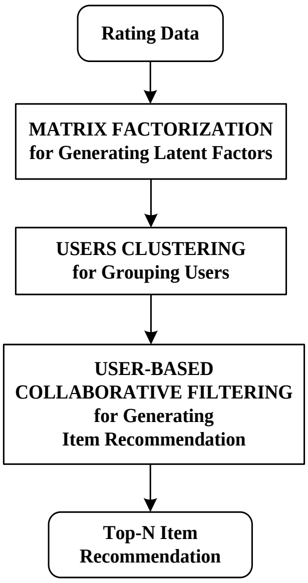

Fig. 1. Framework of Collaborative Filtering Item Recommendation Methods based on Matrix Factorization and Clustering Approaches

```
Algorithm: K-Means Clustering
Process:
1. Randomly initialize $C$ centroids
2. Calculate the distance between user and centroid points.
   Assign each user to its closest centroid
3. Update each centroid by taking the mean of user points
   assigned to the associated cluster
4. Repeat Steps 2 and 3 until the algorithm converges
5. Output $C$ clusters
```

Fig. 2. Algorithm of K-Means clustering

```
Algorithm: Fuzzy C-Means Clustering
Process:
1. Randomly initialize $C$ centroids: $C = \lfloor \mu_{ij} \rfloor$ where $1 = \sum_{k=1}^{C} \mu_{ij}$
2. Calculate the center of each cluster $k$ on attribute $j$
3. Update $C = \lfloor \mu_{ij} \rfloor$
4. Repeat Step 2 until the algorithm converges
5. Output $C$ clusters
```

Fig. 3. Algorithm of Fuzzy C-Means clustering

B. Users Clustering

Clustering is a technique that groups data such that the considered-similar data are joint in the same cluster. We implement the clustering technique to group users based on the users' latent factor matrix. This paper uses two well-known variations of the clustering approaches:

- K-Means clustering: a clustering technique that groups the data based on their distance to the centroid of each group [13, 19]. That is the minimum the distance in the group the higher the similarity between users. The algorithm of K-Means clustering is shown in Fig. 2.
- Fuzzy C-Means clustering: a clustering technique that allows data to belong to more than one cluster [19, 26]. The algorithm of Fuzzy C-Means clustering as shown in Fig. 3.

C. User-based Collaborative Filtering

User-based collaborative filtering is an approach for generating the top-N item recommendation. At this stage, the list of item recommendations for a target user u is generated based on the ratings given by users that belong to the same cluster as the target user. This approach consists of four steps:

1) User similarity

The similarity between target user u and v within the same clusters can be calculated using the Pearson coefficient correlation:

$$Sim(u, v) := \frac{\sum_{i \in I_u \cap I_v} s_{ui} \cdot s_{vi}}{\sqrt{\sum_{i \in I_u \cap I_v} s_{ui}{}^2} \cdot \sqrt{\sum_{i \in I_u \cap I_v} s_{vi}{}^2}} \tag{2}$$

where I_u is the list of items that have been rated by target user u. Meanwhile, s_{ui} is the mean-centered rating of target user u towards item i calculated as:

$$s_{ui} = r_{ui} - \mu_u \quad \forall u \in \{1 \dots m\} \tag{3}$$

where μ_u is the mean rating of user u:

$$\mu_u = \frac{\sum_{i \in I_u} r_{ui}}{|I_u|} \quad \forall u \in \{1 \dots m\} \tag{4}$$

2) User neighborhood

User neighborhood, $X_u(j)$, is the top-k most similar list of users to the target user u, who have rated item j such that $|X_u(j)| \leq k$:

$$X_u(j) := \underset{v \in U_j}{\overset{k}{\operatorname{argmax}}}\ sim(u, v) \tag{5}$$

where U_j is the list of users who have rated item j, and k is the size of users neighborhood.

3) Rating prediction

In the user-based collaborative filtering approach, rating prediction $\hat{r}_{uj}$ of target user u towards item j is calculated as:

$$\hat{r}_{uj} := \mu_u + \frac{\sum_{v \in X_u(j)} Sim(u,v) \cdot s_{vj}}{\sum_{v \in X_u(j)} |Sim(u,v)|} \quad (6)$$

4) Top-N item recommendation

The list of top-N item recommendation for a target user u, $Top_u(N)$, is generated by descending order the list of rating predictions:

$$Top_{\mathrm{u}}(N) := \underset{i \in \hat{I}_u}{\overset{N}{\operatorname{argmax}}}\ \hat{r}_{uj} \quad (7)$$

where $\hat{I}_u$ is the list of items that have not been rated by target user u. In this case, $\hat{I}_u = I - I_u$ whereas $I_u \cap \hat{I}_u = \emptyset$.

III. Empirical Analysis

We conduct experiments to empirically analyze and compare the performances of the developed collaborative filtering item recommendation methods based on matrix factorization and clustering approaches.

A. Experiment Setup

The experiments in this paper are conducted using the MovieLens dataset available from the https://grouplens.org/datasets/movielens/. The dataset consists of 943 users, 1682 movies, and 100000 rating data where the rating range is 1 – 5. Note that we implement the same experimental design as those commonly used in the collaborative filtering item recommendation based researches [12, 14, 15, 27].

To evaluate the method performances, we use the F1-Score and Normalized Discounted Cumulative Gain (NDCG) metrics. The formulation of the top-N recommendation metric scores for a target user u are as follows:

$$F1\text{-}Score_u(N) := \frac{2 \cdot Precision_u(N) \cdot Recall_u(N)}{Precision_u(N) + Recall_u(N)} \quad (8)$$

$$NDCG_u(N) := \frac{DCG_u(N)}{IDCG(N)} \quad (9)$$

where

$$Precision_u(N) := 100 \cdot \frac{|Top_u(N) \cap H_u|}{N} \quad (10)$$

$$Recall_u(N) := 100 \cdot \frac{|Top_u(N) \cap H_u|}{|H_u|} \quad (11)$$

$$DCG_u(N) := \sum_{x=1}^{N} \frac{1}{\log_2(1+x)} \cdot \mathbb{I}(Top_u(x) \in H_u) \quad (12)$$

$$IDCG(N) := \sum_{x=1}^{N} \frac{1}{\log_2(1+x)} \quad (13)$$

where H_u is the ground-truth and $\mathbb{I}(\cdot)$ function results in 1 or 0 to indicate that the condition is either true or false.

B. Methods

We develop six variations of methods based on the three matrix factorization approaches, i.e., NMF, SVD, and PCA; and the two clustering techniques, i.e., K-Means and Fuzzy C-Means used in this paper. Table I shows the labelling of method variations.

TABLE I. THE VARIATION OF METHODS

		Clustering	
		K-Means	***Fuzzy C-Means***
Matrix Factoriza-tion	**NMF**	NMF-KM	NMF-FCM
	PCA	PCA-KM	PCA-FCM
	SVD	SVD-KM	SVD-FCM

To achieve the best performance of each method, we empirically adjust the subsequent parameters: the rank of latent factor model (F), number of clusters (C), and size of users neighborhood (k). As a result, the resulted list of adjustments are as follows:

- NMF-FCM: $F = 512$, $C = 620$, and $k = 7$
- NMF-KM: $F = 4$, $C = 400$, and $k = 5$
- PCA-FCM: $F = 16$, $C = 190$, and $k = 5$
- PCA-KM: $F = 16$, $C = 330$, and $k = 5$
- SVD-FCM: $F = 64$, $C = 490$, and $k = 5$
- SVD-KM: $F = 2$, $C = 380$, and $k = 5$

C. Results and Discussion

Fig. 4 and Fig. 5 respectively show the performance comparisons of the six methods in terms of F1-Score and NDCG, at a various range of top-N. We observe three remarks based on the outcomes.

First, PCA-KM always achieves the best performance compared to the other five methods, in terms of both F1-Score and NDCG metrics, at most top-N. It is only at top-1 that NMF-KM beats PCA-KM in terms of NDCG. Table II and Table III respectively list the global averages of outperformance of PCA-KM, i.e., 82.05% in terms of F1-Score and 52.10% in terms of NDCG. On the contrary, the worst performance is achieved by PCA-FCM at nearly any top-N. It is one time that NMF-FCM performs the worst in terms of F1-Score at top-1. These findings advocate that a collaborative filtering item recommendation method based on matrix factorization and clustering approaches is best implemented by using PCA and K-Means, but certainly not PCA and FCM.

Second, the combination of K-Means clustering with any matrix factorization techniques is superior compared to that of Fuzzy C-Means. This result opposes the outperformance of Fuzzy C-Means compared to K-Means [12], i.e., in the case where no matrix factorization approach is implemented.

The advantage of implementing the matrix factorization approach depends on the choice of the clustering technique used. It is recommended to implement the NMF matrix factorization when the Fuzzy C-Means clustering is used. Meanwhile, a method that uses the K-Means clustering is best coupled with the PCA matrix factorization. This outcome shows that the outperformance of a matrix factorization approach is not an absolute case [25, 28].

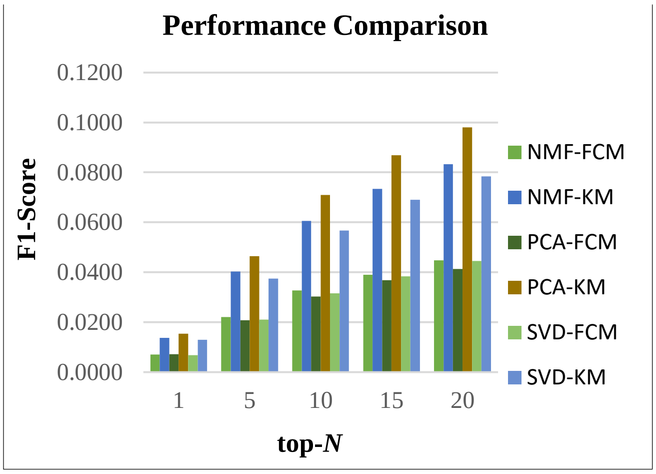

Fig. 4. The F1-Score results

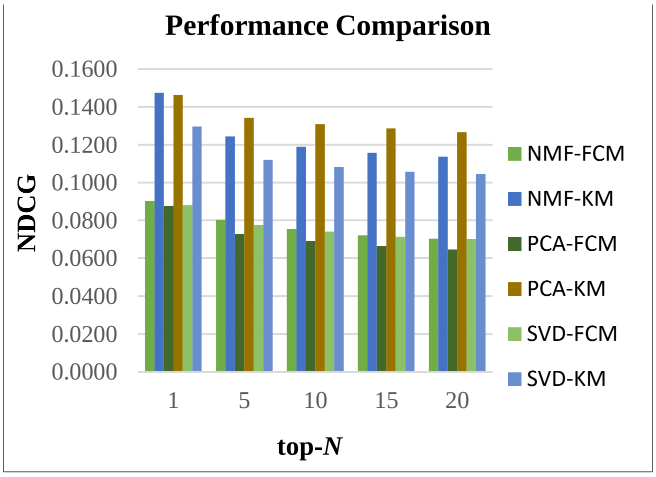

Fig. 5. The NDCG results

TABLE II. THE OUTPERFORMANCE OF PCA-KM IN TERMS OF F1-SCORE

	Outperformance of PCA-KM (%)				
	top-1	*top-5*	*top-10*	*top-15*	*top-20*
NMF-FCM	117.60	110.33	117.08	122.46	119.05
NMF-KM	11.88	15.41	17.06	18.28	17.71
PCA-FCM	112.45	124.16	134.31	136.16	137.52
SVD-FCM	127.60	121.02	125.06	126.67	120.59
SVD-KM	18.55	24.15	25.18	25.79	25.17
AVERAGE	**77.62**	**79.02**	**83.74**	**85.87**	**84.01**
GLOBAL AVERAGE	**82.05**				

TABLE III. THE OUTPERFORMANCE OF PCA-KM IN TERMS OF NDCG

	Outperformance of PCA-KM (%)				
	top-1	*top-5*	*top-10*	*top-15*	*top-20*
NMF-FCM	62.03	66.89	73.33	78.44	79.76
NMF-KM	-0.82	7.82	9.96	11.18	11.37
PCA-FCM	66.68	84.07	89.42	93.28	96.03
SVD-FCM	66.21	72.92	76.63	80.17	80.54
SVD-KM	12.83	19.80	21.03	21.66	21.40
AVERAGE	**41.38**	**50.30**	**54.07**	**56.95**	**57.82**
GLOBAL AVERAGE	**52.10**				

IV. CONCLUSION AND FUTURE WORK

This paper develops and compares the performances of six Collaborative Filtering item recommendation methods based on three matrix factorization (i.e., NMF, SVD, and PCA) approaches and two clustering techniques (i.e., K-Means and Fuzzy C-Means). The recommendation performances are evaluated in terms of F1-Score and NDCG metrics, at various top-*N*.

Experiment results show that PCA-KM outperforms the other five methods, where the global averages of outperformance are 82.05% in terms of F1-Score and 52.10% in terms of NDCG. It is worthwhile to note that the combination of K-Means clustering with any matrix factorization techniques is superior compared to that of Fuzzy C-Means. Additionally, the advantage of implementing the matrix factorization approach depends on the choice of the clustering technique used.

In the future, we are planning to modify the framework of our method such that the users' similarity matrix is used as the input of the matrix factorization process. We also want to further enhance the method by implementing items, instead of the users, latent factor matrix in the current framework.

REFERENCES

[1] J. A. Konstan and J. Riedl, "Recommender systems: from algorithms to user experience," *User Modeling and User-Adapted Interaction*, vol. 22, pp. 101-123, 2012.

[2] C. C. Aggarwal, *Recommender Systems: The Textbook*. Switzerland: Springer International Publishing, 2016.

[3] G. Guo, J. Zhang, and D. Thalmann, "Merging trust in collaborative filtering to alleviate data sparsity and cold start," *Knowledge-Based Systems*, vol. 57, pp. 57-68, 2014.

[4] N. Ifada and R. Nayak, "How Relevant is the Irrelevant Data: Leveraging the Tagging Data for a Learning-to-Rank Model," in Proceeding of *The 19th ACM International Conference on Web Search and Data Mining*, San Francisco, California, US, pp. 23-32, 2016.

[5] N. P. Kumar and Z. Fan, "Hybrid User-Item Based Collaborative Filtering," *Procedia Computer Science*, vol. 60, pp. 1453-1461, 2015.

[6] S. Loh, F. Lorenzi, R. Granada, D. Lichtnow, L. K. Wives, and J. P. de Oliveira, "Identifying Similar Users by their Scientific Publications to Reduce Cold Start in Recommender Systems," in Proceeding of *5th International Conference on Web Information Systems and Technologies (WEBIST 2009)*, pp. 593-600, 2009.

[7] Y. Koren, "Factorization Meets The Neighborhood: A Multifaceted Collaborative Filtering Model," in Proceeding of *The 14th ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, pp. 426-434, 2008.

[8] Y. Koren, R. Bell, and C. Volinsky, "Matrix Factorization Techniques for Recommender Systems," *Computer,* vol. 42, pp. 30-37, 2009.

[9] Y. Koren and J. Sill, "OrdRec: An Ordinal Model for Predicting Personalized Item Rating Distributions," in Proceeding of *The 5th ACM conference on Recommender Systems*, Chicago, Illinois, USA, pp. 117-124, 2011.

[10] D. Bokde, S. Girase, and D. Mukhopadhyay, "Matrix factorization model in collaborative filtering algorithms: A survey," *Procedia Computer Science,* vol. 49, pp. 136-146, 2015.

[11] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Application of dimensionality reduction in recommender system - A case study," Minnesota University Minneapolis Department of Computer Science, Minneapolis2000.

[12] H. Koohi and K. Kiani, "User based Collaborative Filtering using fuzzy C-means," *Measurement,* vol. 91, pp. 134-139, 2016.

[13] P. Phorasim and L. Yu, "Movies recommendation system using collaborative filtering and k-means," *International Journal of Advanced Computer Research,* vol. 7, pp. 52-59, 2017.

[14] N. Sun, Y. Zhuang, and S. Ni, "A Trust-Based Collaborative Filtering Algorithm Using a User Preference Clustering," *Management Science and Engineering,* vol. 11, pp. 9-19, 2017.

[15] J. Zhang, Y. Lin, M. Lin, and J. Liu, "An effective collaborative filtering algorithm based on user preference clustering," *Applied Intelligence,* vol. 45, pp. 230-240, 2016.

[16] G. Guo, J. Zhang, and N. Yorke-Smith, "Leveraging multiviews of trust and similarity to enhance clustering-based recommender systems," *Knowledge-Based Systems,* vol. 74, pp. 14-27, 2015.

[17] N. Ifada, E. H. Prasetyo, and Mula'ab, "Employing sparsity removal approach and Fuzzy C-Means clustering technique on a movie recommendation system," in Proceeding of *The 2018 International Conference on Computer Engineering, Network and Intelligent Multimedia (CENIM)*, Surabaya, Indonesia, pp. 157-162, 2018.

[18] H. Koohi and K. Kiani, "A new method to find neighbor users that improves the performance of collaborative filtering," *Expert Systems with Applications,* vol. 83, pp. 30-39, 2017.

[19] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*, Third ed. Waltham, USA: Morgan Kaufmann, 2012.

[20] M. Abbas, M. U. Riaz, A. Rauf, M. T. Khan, and S. Khalid, "Context-aware Youtube recommender system," in Proceeding of *The 2017 International Conference on Information and Communication Technologies (ICICT)*, pp. 161-164, 2017.

[21] C. Zhang, H. Wang, S. Yang, and Y. Gao, "Incremental nonnegative matrix factorization based on matrix sketching and k-means clustering," in Proceeding of *The 2016 International Conference on Intelligent Data Engineering and Automated Learning* pp. 426-435, 2016.

[22] H. Zarzour, F. Maazouzi, M. Soltani, and C. Chemam, "An improved collaborative filtering recommendation algorithm for big data," in Proceeding of *IFIP International Conference on Computational Intelligence and Its Applications* pp. 660-668, 2018.

[23] J. Kim, Y. He, and H. Park, "Algorithms for Nonnegative Matrix and Tensor Factorizations: A Unified View based on Block Coordinate Descent Framework," *Journal of Global Optimization,* vol. 58, pp. 285-319, 2014.

[24] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13*, ed: MIT Press 2001, pp. 556–562.

[25] L. Zhao, G. Zhuang, and X. Xu, "Facial expression recognition based on PCA and NMF," in Proceeding of *The 7th World Congress on Intelligent Control and Automation*, pp. 6826-6829, 2008.

[26] S. Wei, N. Ye, S. Zhang, X. Huang, and J. Zhu, "Collaborative filtering recommendation algorithm based on item clustering and global similarity," in Proceeding of *The 5th International Conference on Business Intelligence and Financial Engineering (BIFE)*, Lanzhou, China, pp. 69-72, 2012.

[27] N. Ifada, D. R. M. Alim, and M. K. Sophan, "NMF-based DCG Optimization for Collaborative Ranking on Recommendation Systems," in Proceeding of *The 2nd International Conference on Machine Learning and Machine Intelligence (MLMI 2019*, Jakarta, 2019.

[28] Z. Sharifi, M. Rezghi, and M. Nasiri, "A new algorithm for solving data sparsity problem based-on Non negative matrix factorization in recommender systems," in Proceeding of *The 4th International Conference on Computer and Knowledge Engineering (ICCKE)*, pp. 56-61, 2014.