IJICTR
International Journal of Information & Communication Technology Research

Volume 14- Number 3 – 2022 (37 -47)

Improving Content-Based Recommender System For Clustering Documents Based on Ontology And New Hierarchical Clustering Method

Maryam Hourali
Electronic engineering
Malek-Ashtar University of Technology
Tehran, Iran
maryam_hourali@yahoo.com

Mansoureh Hourali*
Industrial Engineering
Payame Noor University
Tehran, Iran
hourali@pnu.ac.ir

Received: 5 March 2022 – Revised: 12 March 2022 - Accepted: 4 Seotember 2022

Abstract—**Today we live in a period that is known to an area of communication. By increasing the information on the internet, the extra news are published on news agencies websites or other resources, the users are confused more with the problems of finding their desired information and related news. Among these are recommended systems they can automatically finding the news and information of their favorite's users and suggesting to them too. This article attempts to improve the user's interests and user's satisfactions by refining the content based recommendation system to suggest better sources to their users. A clustering approach has been used to carry out this improvement. An attempt has been made to define a cluster threshold for clustering the same news and information in the K-means clustering algorithm. By detecting best resemblance criterion value and using an external knowledge base (ontology), we could generalize words into a set of related words (instead of using them alone). This approach is promoted the accuracy of news clustering and use the provided cluster to find user's favorite news and also could have suggest the news to the user. Since the dataset has an important and influential role in advisory recommended systems, the standard Persian dataset is not provided and not published yet. In this research, an attempted has been made to connect and publish the dataset to finish the effect of this vacuum. The data are collected and crawl 8 periods of days from the Tabnak news agency website. The profile of each volunteers has been created and also saved at the same time as they read the favorite news on that period of time. An analysis shows that the proposed clustering approach provided by the NMI criterion has reached 70.2% on our the dataset. Also, using the suggested clustering recommendation system yield 89.2% performance based on the accuracy criterion, which shows an improvement of 8.5% in a standardized way.**

Keywords: Recommender system; Persian news; Hierarchical clustering; Ontology.

Article type: Research Article

Publisher: ICT Research Institute

*Corresponding Author

[DOI: 10.52547/itrc.14.3.37]

I. Introduction

The ever-growing in news titles and information leads to the confusion of website and news social media users. Companies that use electronic trade need strategies to convince the viewers to turn into loyal customers. The lack of news recommendation systems results in many news titles being recommended to the user without considering their taste which istes their time or draws them away from the website's intentions. On the other hand, the news and advertisement links will fail to reach their desired audience. Furthermore, the display of a large volume of news with no regard for the user's interest renders them unable to read the new titles that interest them. This problem will only result in the dissatisfaction of website users, but it will also reduce the productivity of the news impact.

Despite its many advantages, the internet results in problems mainly for two groups. The first group are content developers and the other group are the users. Since it has become pretty easy to access information on the internet, maintaining the users and gaining their satisfaction has become one of the important goals of content developers.

On the other hand, the presence of various organizations built to satisfy users' needs has confused them in making the best choice. This abundant and growing volume of information on the internet and web pages has made decision-making and selecting the desired information, data, or products difficult for many web users. This is the spark that intrigued the researchers to search for a solution to deal with this essential problem of the new era which is known as "data overflow". Two approaches have been proposed so far to deal with this problem. The first approach is the use of the two concepts of information recovery and information filtering. The main limitation of these two concepts in giving suggestions is that unlike human recommenders (e.g. friends, family, etc.), these two methods are hardly able to distinguish and differentiate high-quality and low-quality items to suggest for a subject or product. This problem resulted in the emergence of a second approach entitled recommender system. These new systems have resolved the issues of the first approach.

A recommender system can provide a user with the best suggestion and prevent them from isting their time to review all items by suggesting suitable items based on the user's desire according to their behavior and their referral information. These systems also save the user from confusion in an abundance of information. The need for recommender systems is highlighted more than ever given the increasing volume of information.

Searching the websites of news social media where many news titles are published for one's desired and favorite titles is difficult and time-consuming, which is a persistent problem in these websites. Recommender systems provide the user with the most suitable news by analyzing their behavior.

Considering that problem of data scatter is a fundamental and essential issue in recommender systems (the respective background has been explained in the "literature review" section), one of the solutions to this problem is using content-based recommender systems to overcome the problem. Considering that we are dealing with text processing, this research seeks to use text clustering to suggest the products (items) in the same cluster to the user due to their close content, similar text, and therefore, the overall similarity between the items. One of the algorithms to do this can be the application of the meaning similarity in content-based systems that helps resolve the problems of recommender systems so that we reduce the recommender systems' errors in the field of text suggestion.

In this study, we seek to introduce content-based recommender systems that provide news website users with the best suggestions considering the characteristics of news and history and reduce the information overflow issues in the age of information.

II. Related work

Recommendation systems have developed rapidly, and various domains have used them, such as movies, music, news, books, restaurants, and other media. In addition, several researchers have developed recommendation systems with many existing approaches, including demographic filtering, content-based filtering, collaborative filtering, and hybrid filtering [1,2].

One of the most prevalent approaches to recommendation systems is collaborative filtering [3–6]. This approach is capable of generating recommendations based on the ratings provided by the users for several items. Collaborative filtering consists of two methods: model-based and memory-based. The first method uses a model built from the ratings to generate recommendations, while the second method utilizes similarity metrics to get the distance between two users/items [4,7].

In recent years, several researchers have proposed collaborative filtering using the similarity metrics approach to increase the accuracy of recommendations. Some of the proposed similarity metrics are Proximity-Significance-Singularity (PSS) [8], Bhattacharyya [9], multi-level collaborative filtering [11], item frequency-based similarity [13], Triangle Multiplying

Jaccard (TMJ) [12], and three impact factors-based similarity [1]. However, these similarity metrics only consider the user rating score to calculate similarities between users. The user rating score is the value given directly by the user in assessing the selected or purchased product. The score ranges from 1 to 5, with a score of 1 indicating that the user really dislikes the selected product and a score of 5 indicating that the user really likes the selected product.

Recently, the development of similarity metrics has considered the user rating score and user

[DOI: 10.52547/itrc.14.3.37]

behavior score. The similarity metrics that have adopted user behavior scores in calculating similarity are User score Probability Collaborative Filtering (UPCF) [13] and User Profile Correlation-based Similarity (UPCSim) [14]. However, adding the user behavior score variable in the similarity calculation causes the computation to be more complex. Consequently, it consumes time with the increasing data.

Several studies [15-17] have utilized clustering methods by reducing large amounts of data to overcome the computational complexity of recommendation systems. The studies applied the partition-based clustering methods by determination of the number of clusters directly. The problem of these studies is determining the number of clusters without measuring the clustering quality to get the optimal number of clusters that affect the results of the recommendations.

One of the well-known recommendation systems is memory-based collaborative filtering that utilizes similarity metrics. Recently, the similarity metrics have taken into account the user rating and user behavior scores. The user behavior score indicates the user preference in each product type (genre). The added user behavior score to the similarity metric results in more complex computation. To reduce the complex computation, we combined the clustering method and user behavior score-based similarity. The clustering method applies k-means clustering by determination of the number of clusters using the Silhouette Coefficient. Whereas the user behavior score-based similarity utilizes User Profile Correlation-based Similarity (UPCSim). The experimental results with the MovieLens 100k dataset showed a faster computation time of 4.16 s. In addition, the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) values decreased by 1.88% and 1.46% compared to the baseline algorithm [18].

(Dhanani and et al. , 2021) proposed a graph clustering based novel Legal Document Recommendation System (LDRS) that forms clusters of referentially similar judgments and within those clusters find semantically relevant judgments. Hence, pairwise similarity scores are computed for each cluster to restrict search space within-cluster only instead of the entire corpus. Thus, the proposed LDRS severely reduces the number of similarity computations that enable large numbers of judgments to be handled. It exploits a highly scalable Louvain approach to cluster judgment citation network, and Doc2Vec to capture the semantic relevance among judgments within a cluster. The efficacy and efficiency of the proposed LDRS are evaluated and analyzed using the large real-life judgments of the Supreme Court of India. The experimental results demonstrated the encouraging performance of proposed LDRS in terms of Accuracy, F1-Scores, MCC Scores, and computational complexity, which validates the applicability for scalable recommender systems. Despite the success demonstrated, the proposed approach is limited by its consideration only for judgments with at least one citation. In a real scenario, there are many judgments without any single citation In the future, consideration of all judgments would be an interesting aspect to enhance the proposed LDRS[19].

To overcome the shortage of described methods, in this paper we attempt to improve the user's interests and user's satisfactions by refining the content-based recommendation system to suggest better sources to their users. A clustering approach has been used to carry out this improvement. An attempt has been made to define a cluster threshold for clustering the same news and information in the K-means clustering algorithm

The Remainder of This Paper is Organized as Follows. Section III presents the proposed method. Section IV concludes the results of the study.

III. THE PROPOSED METHOD

This section discusses the proposed method to recommend news using the proposed hierarchical clustering algorithm used in the present study. Figure 1 demonstrates the stages of the proposed method.

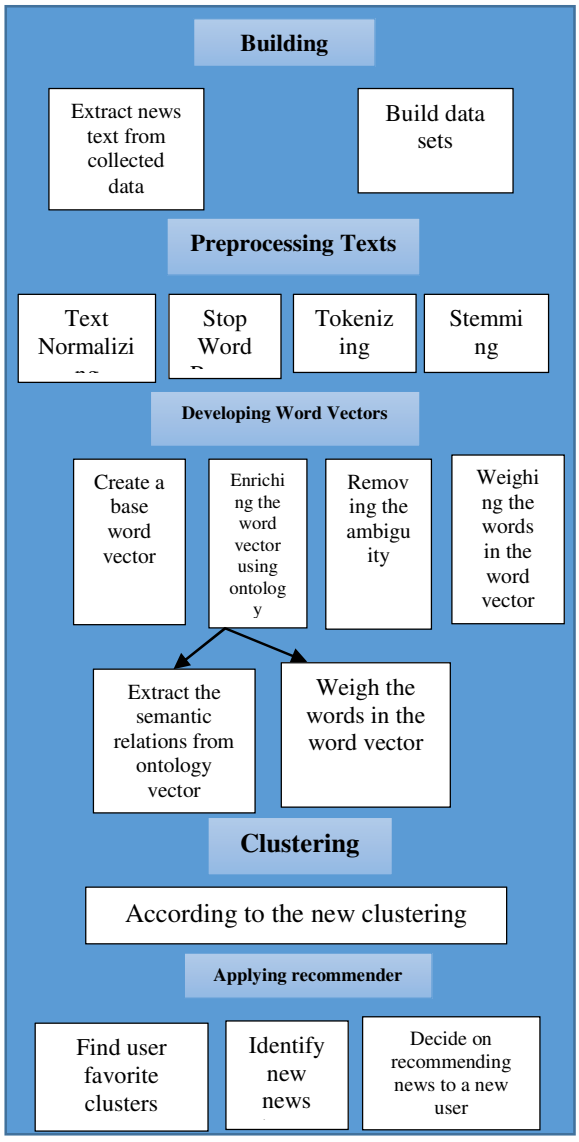

Figure 1. The proposed recommender system.

[DOI: 10.52547/itrc.14.3.37]

A. Building the dataset

The dataset used to develop the recommendations is pretty important in research on recommender systems, to the extent that many of the available studies can be distinguished from one another based on the configuration they have used [4]. No suitable dataset has been developed for the Persian language to be used in recommender systems while many suitable and standard dataset are available for the English language and are even used in the Persian research to evaluate the ability and confirm the efficiency of the researchers proposed methods. One of the shortcomings in Persian and the expansion of the recommender systems in this language is the lack of a proper dataset to evaluate the efficiency of the methods proposed in this field. The present study develops and introduces a new and standard Persian dataset called Persian News Recommender to be used in research on text clustering, recommender systems, text classification, etc. which is among the innovations of the present study and helps develop a standard dataset for researchers in this field. Linguists in this field were used to evaluate the database. Since the agreement reached 0.88 based on the Kappa criterion, it can be concluded that it is appropriate.

Cohen's kappa is calculated with the following formula.

$$\kappa = \frac{p_0 - p_e}{1 - p_e}, \qquad (1)$$

Where P_0 is the overall accuracy of the model and P_e is the measure of the agreement between the model predictions and the actual class values as if happening by chance.

1. Normalizing the texts

After the news texts are extracted from web pages, each news is placed in a file based on the standard mentioned above. Then, the normalization operation is conducted on the texts that increase the readability of texts in the system and leads to less ambiguity in the morphing of the news which is done at the next stage. This also makes the words that are different from the system's perspective (due to the difference in the codes for some of the similar words) uniform which results in a reduced number of frequency calculation errors. Some of the operations performed at this stage include creating half-spaces between the two-part words that are written with no space, creating half-spaces between the two-part words that are written separately, turning letters such as "k" and "y" into their standard Persian equivalents, and converting Arabic numerals to standard Persian numerals[18].

2. Morphing the texts

There are many words in each language that are used abundantly in daily conversations but are useless in text processing and most natural language artificial intelligence processing. Thus, these words are removed in processing to reduce calculation complexities and improve accuracy.

Normalization and morphing of the texts in the present study are conducted using a Python language digestion library and stop-words are removed using the program implemented in Python.

After the normalization, the words in the texts are turned int morphemes considering whether they are made out of one or several parts, which is considered the based for the operations at the following stages of clustering [19].

3. Elimination of stop-words (redundant words)

There are many words in each language that are used abundantly in daily conversations but are useless in text processing and most natural language artificial intelligence processing. Thus, these words are removed in processing to reduce calculation complexities and improve accuracy.

Normalization and morphing of the texts in the present study are conducted using a Python language digestion library and stop-words are removed using the program implemented in Python[18].

B. Developing word vectors

1. Create a base word vector

After the pre-processing phase, the second phase (i.e. clustering) started. The initial goal in this phase is development of word vectors. For this purpose, all of the news documents which are selected to train the model, are integrated and an integrated file is produced. After the development of the integrated words (stop-words are already removed in the pre-processing phase), the unique words in the integrated file are extracted and made the base word vector. Considering the large size of the developed word vector (since it included many words) and the fact that the length of the word vector would increase significantly after being enriched through the extraction of semantic relations between the words from ontology, the volume of calculations would increase significantly which would reduce performance.

The features with the highest effectiveness must be selected to reduce the size of the feature vector. A threshold is used in the present for this purpose, which means only the words with repetition numbers higher than a specific amount areare included in the world vector. Thus, the words that areare repeated fewer times than the threshold value areare deemed unimportant in text processing (clustering) and areare eliminated from the base word vector. After the less-used words areare eliminated from the word vector, unique ids are areattributed to the remaining words.

2. Enriching the word vector using ontology

The FarsNet ontology isis used in the present study to extract the semantic relationships between the words because of its quality, availability, and the fact that it isis supported by its producer group. Four relationships of synonyms, antonyms, hypernyms, and hyponyms are used in the present study. For each word in the base word vector, its semantic relationships are extracted (which needs

[DOI: 10.52547/itrc.14.3.37]

clarifications since each word might have several semantic relationships). After removing the ambiguity, the relationships are added to the word vector and enriched for the clustering operation. This results in the news texts that do not have similar features to educational texts but are similar to their semantic relationship to be correctly clustered.

3. Removing the ambiguity

Considering that for each of the relationships mentioned in II.1 section, a large semantic vector is developed for each of the features, the dimension of the features increases significantly which leads to ambiguity in the clustering operation, higher computational complexity, and lower accuracy. Thus, ambiguities must be removed at the next stage. For this purpose, a threshold is considered for each of these four relationships, so that only a specific number of the extracted relationships play a role in the clustering operation. Considering that these initially extracted relationships are more similar to the main feature in terms of concept, the first extracted relationships to reach the threshold are considered the winner one, and the other is considered loser one.

4.Weighing the words in the word vector

TF-DF method and the 1, 2, and 3 equations are used and the values obtained from equation 3 are normalized to obtain the weight of each of the words in the word vector using the equation 4.

$$tf(t_k, d_i) = \begin{cases} \#(t_k, d_i) & t_k \in Vector\ of\ d_i \\ 0 & t_k \notin Vector\ of\ d_i \end{cases} \quad (2)$$

$$idf(t_k) = log \frac{|D|}{|\{d : t \in d\}| + 1} \quad (3)$$

$$tfidf(t_k, d_i) = tf(t_k, d_i) * idf(t_k) \quad (4)$$

$$w_{ki} = normTFIDF(t_k, d_i) = \frac{tfidf(t_k, d_i)}{\sqrt{\sum_k (tfidf(t_k, d_i))^2}} \quad (5)$$

C. Clustering

Since the present study focuses on content-based recommender systems and these systems are based on the relationships between the item and user and do not include the information of other users, so that their tastes and interests can be used to recommend a product (item) to the given user unlike the participatory recommender system, most of the information required to make recommendations must be obtained from text (news) documents. The present study sought to design and implement an optimal clustering system through texts processing which can provide the user with the documents that would more interest him among the other clusters. A study on related documents also revealed that the cosines similarity between similar news is higher than a threshold. The proposed clustering to cluster news is as follows

After the word vector is created for the documents, the word vector of all educational documents is entered into the total list. The first document of the list is considered as the reference and its cosines distance from the other documents is calculated. As explained regarding the relevant documents mentioned above, the documents with distances lower than a certain amount from the reference documents are removed from the total list and transferred to the "RefList". The RefList contains the documents that are probably in the same cluster. The average of the documents in the Reflist is calculated and saved as the center of the document cluster in the Centroid List. This repeats until the total list is empty. After the total list gets empty, cluster centers will be inside the CentroidList. The CentroidList is the input center of the new hierarchical clustering algorithm. Then, the distance between each document and the cluster center is calculated, and each document is placed in the respective cluster considering the smallest distance to the center. This is repeated until the end condition is satisfied (the end condition is an MSE error of less than 0.0003 or the execution of the loop up to 500 repetitions.).

IV. Results and Analysis (Apply the Recommender)

In this section, the proposed method is analyzed, investigated, and performance evaluated based on the results of its implementation. The conventional methods in this field are used to carry out the evaluation.

A. Evaluation criteria

The two criteria of accuracy and NMI(Normalized Mutual Information) are used to evaluate the proposed clustering. Besides, the criterion of accuracy is used to evaluate the recommender system.

B. Developing the word vector

After pre-processing, the base word vector must be created. At this stage, 15 thresholds are used to extract the features of the words remaining in the documents. The threshold means that the selected features must be repeated more times than the value of the threshold. The reason for using 15 thresholds is to evaluate the clustering accuracy and NMI after the selection of each threshold, and to reach maximum productivity. NMI and accuracy are calculated after each threshold is implemented, the base word vector is developed for each threshold, and the proposed clustering is implemented to obtain the optimal threshold. Table I demonstrates the number of words after the implementation of each threshold.

[DOI: 10.52547/itrc.14.3.37]

TABLE I. NUMBER OF WORD VECTORS IN DIFFERENT THRESHOLDS

Threshed	Number of words Basic words
٢T	٨٧٨٥
٢T	٦٥٠٣
٣T	٥٢٢٦
٤T	٤٤٠٠
٥T	٣٧٩٤
٦T	٣٣٣٨
٧T	٣٠٠٩
٨T	٢٧٣٣
٩T	٢٤٩٤
١٠T	٢٣٠١
١١T	٢١٢٠
١٢T	١٩٨٠
١٣T	١٨٥٧
١٤T	١٧٣٣

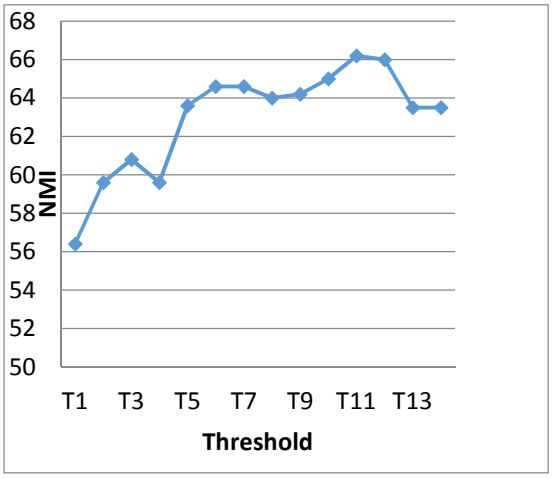

Figure 2. Investigation the effect of word vector threshold on clustering performance

It must be mentioned that the NMI value for each threshold demonstrated in Figure ٢ is obtained from the mean of the 10 repetitions of base word vector operation. It has been used the proposed clustering method to prevent the NMI value obtained for the clustering to be biased by any specific factor. It is observed that decreasing the number of words in the word vector eliminates more ineffective features and only effective features remain in it. This reduces the volume of calculations and improves the process of clustering. On the other hand, Figure 2 demonstrates that increasing the threshold eliminates the effective features which reduce the clustering efficiency and the NMI value. The best threshold is revealed to be T11 in the present study. The following stages of clustering are carried out considering a threshold of T11.

C. The proposed cluster threshold

The previous section mentioned the threshold of the word vector and the impact of various thresholds on clustering. This section will discuss the cluster threshold. The documents that are more similar than a specific amount are placed in the same cluster, which means that if the similarity between two documents passes a specific threshold, they are considered to be related and are placed in the same cluster. The optimal cluster is thus obtained by applying various thresholds.

It must be noted that two identical documents (copied) will have a similarity criterion of 1 and two completely different documents will have a similarity criterion of zero. The similarity criterion will vary in a range of [0, 1].

Cluster thresholds ranging from extremely strict (similarity criterion of 0.9) to extremely loose (similarity criterion of 0.001) are considered. Figure 3 demonstrates the average of 10 repetitions of base word vector operations based on the proposed cluster for each threshold.

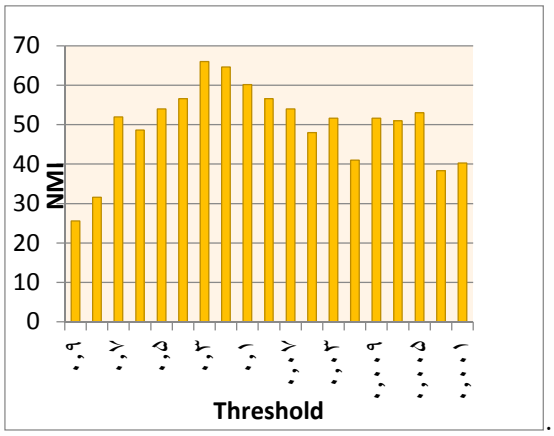

Figure 3. Results of cluster threshold applications

It can be observed from Figure ٣ that the best results are obtained for cluster thresholds between 0.1 and 0.3. Further study of the results has been demonstrated for NMI in Figure 4 and the number of clusters in Figure 5. A larger number of experiments is repeated to investigate the relationship between the number of clusters and the NMI obtained for the cluster threshold (repeated 30 times for each threshold).

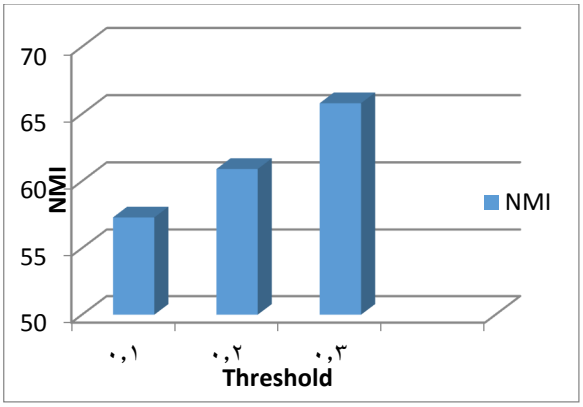

Figure 4. NMI for cluster threshold Figure.

[DOI: 10.52547/itrc.14.3.37]

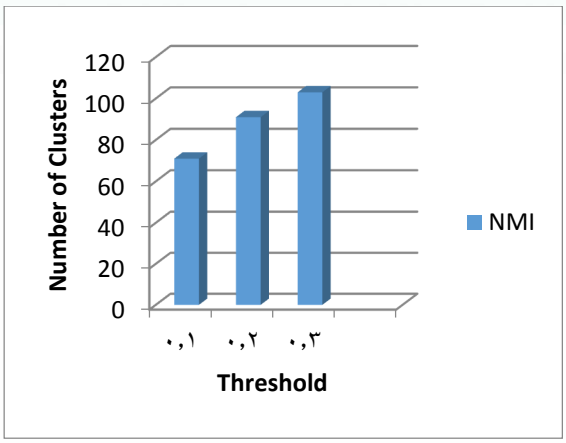

Figure 5. Number of clusters created based on different thresholds

The high cluster accuracy and maximum similarity of the documents in each cluster are significant for better recommendations to users. The cluster threshold is considered 0.3 in the following stages of the work.

D. Selecting the similarity criterion

After similarity criteria are obtained from studies and articles in the field of text analysis, it is discovered that Pearson, Sorens, Jakard, and cosines similarity criteria are the priorities of these studies. This section compares the performance and results of these criteria to select the superior similarity threshold.

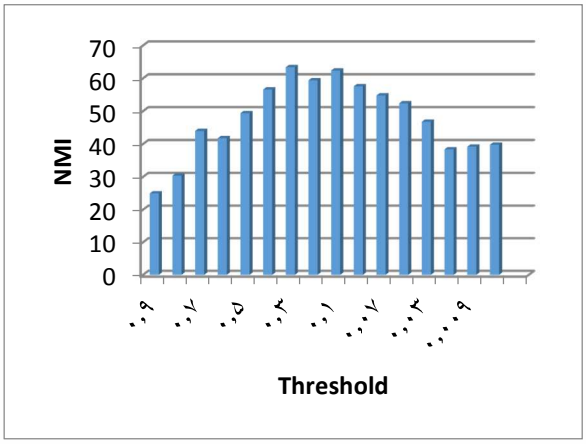

Figure 6. Results from Pearson similarity criterion.

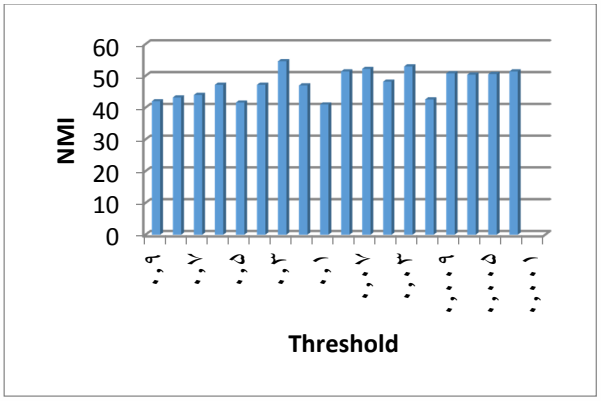

Figure 7. Results from the Soreness similarity criterion

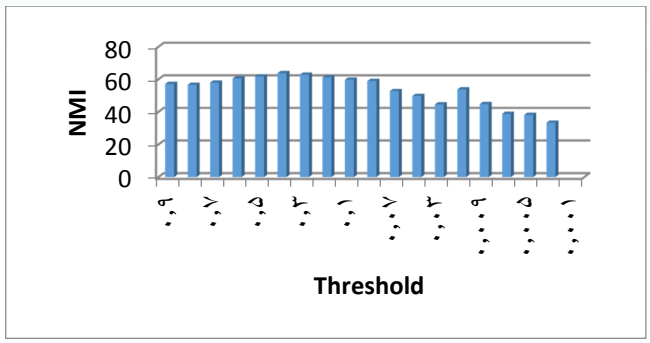

Figure 8. Results from the Jacquard similarity criterion

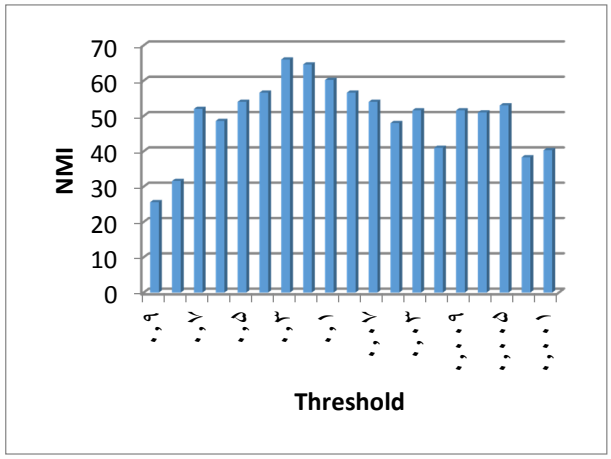

Figure 9. Results from the cosine similarity criterion

The results of the NMI average of 10 repetitions of applying the base word vector to the proposed clustering are obtained in the experiments on the diagrams above for each cluster threshold.

It can be concluded from the charts above that the two Jakard and cosines similarity criteria performed well, and the cosines criterion performed better than the cosines criterion in some cases. Henceforth, all the tests will be conducted using the cosines similarity criterion.

E. Enriching the word vector through ONTOLOGY

At this stage, the base word vector is expanded by ontology so that the semantic relationships in the base word vector are enriched and the final word vector is achieved. The use of ontology is only so that if words that are not among the features in the word vector but are among the semantic relationships of the features in the word vector are identified correctly, and similar documents are identified accordingly.

For this purpose, the four relationships of synonyms, antonyms, hypernyms, and hyponyms are used to enrich the base word vector. The problem is that the semantic relationships of the words are so abundant that the system suffers from ambiguity as the result. Specific combinations such as those mentioned in table II are considered to remove the ambiguity. The first 10 combinations are the those mentioned in reference (18), and three antonym combinations have been added in this dissertation since the mentioned reference did not include antonyms.

[DOI: 10.52547/itrc.14.3.37]

TABLE II. SELECTION OF DIFFERENT COMBINATIONS TO EXTRACT THE RELEVANT RELATIONSHIPS FROM ONTOLOGY

different combinations
synonym2hypernym0hyponym0antonym0
synonym 2hypernym2hyponym2antonym0
synonym 3hypernym0hyponym0antonym0
synonym 3hypernym2hyponym2antonym0
synonym 4hypernym0hyponym0antonym0
synonym 4hypernym2hyponym1antonym0
synonym 5hypernym0hyponym0antonym0
synonym 5hypernym2hyponym0antonym0
synonym 6hypernym0hyponym0antonym0
synonym 6hypernym1hyponym0antonym0
synonym 4hypernym0hyponym0antonym2
synonym 4hypernym2hyponym1antonym2
synonym 6hypernym1hyponym0antonym2

FarsNet v.2.5 ontology is used in word vector enrichment. The question that arises is whether the extracted semantic relationships help improve the cluster or not. Figure 10 demonstrates the impact of threshold on the cluster performance before word vector enrichmen it. The question here is that: considering a threshold for the extracted semantic relationships improve the system? The impact of the threshold on semantic relationships is investigated before extracting the semantic relationships for various thresholds, developing the word vector for documents, and implementing the cluster. According to which extracted three semantic relationships including four synonym, zero hypernyms, and zero hyponyms for defense news, the semantic relationships are obtained from FarsNet ontology with the threshold of T11 for the development of the word vector. The threshold is then implemented on the extracted semantic relationships, and the results are recorded in the mentioned figure. Figure 10 demonstrates that the T1 threshold is suitable for semantic relationships.

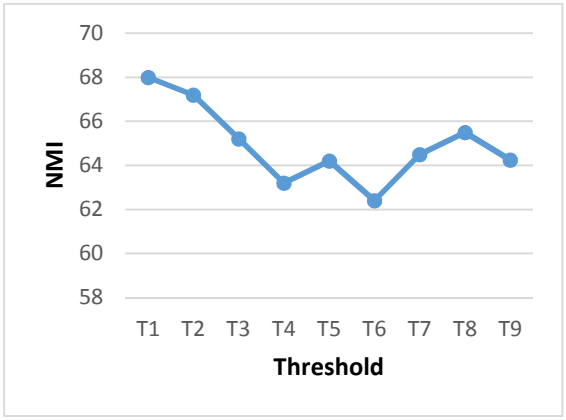

Figure 10. Threshold for ontology Figure.

The results demonstrated in the figure above are also the average of 10 iterations for each threshold, which indicated that T1 yielded the best results.

V. WEIGHING WORD VECTOR FEATURES

After creating the word vector, each of the documents in the word vector dataset must be weighed. The norm TF-IDF algorithm is used to weigh the news text features. Each document is prepared, saved in the form of news code and word vector, and the proposed cluster is applied on each of the word vectors enriched by ontology according to figure 2, section 3, and the clustering operation is completed. Figure 11 illustrates the results of implementing the ontology.

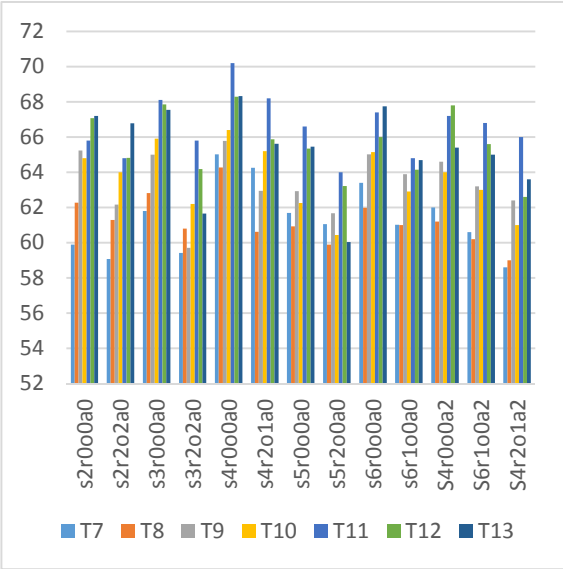

Figure 11. Investigation of the effect of combining.

different semantic relationships on clustering It can be observed from figure 11 that the best result is obtained from the s4r0o0a0 combination with the threshold of T11.

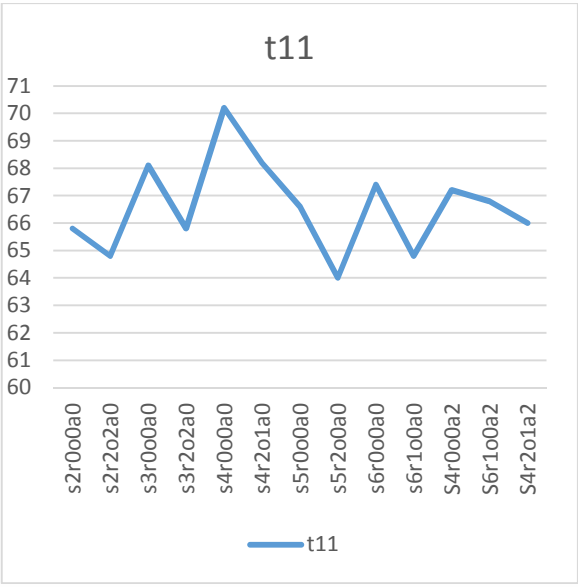

Figure 12. Investigation of the effect of combining

different semantic relationships on clustering at the 11T threshold.

[DOI: 10.52547/itrc.14.3.37]

It can be observed from Figures 11 and 12 that the combination of the semantic relationships only including synonyms is more productive to the combinations of semantic relationships including the four relationships of synonyms, antonyms, hypernyms, and hyponyms according to the NMI criterion. The reason for this can be justified using an example:

Consider two texts in the field of animals. The first text discusses whales and the other discusses sheep. If hypernyms are used to enrich the word vector of both documents, features such as mammals will be added to both documents' word vectors which would create ambiguity in distinguishing the documents and will reduce clustering accuracy.

As another example, consider two texts, one regarding elephants and the other regarding sheep. If hyponyms are used to enrich the word vector of both documents, features such as quadruped will be added to both documents' word vectors which would create ambiguity in distinguishing the documents and will reduce clustering accuracy.

VI. COMPARING THE PROPOSED METHOD TO OTHER METHODS

The first section compares the proposed method to the base method in terms of using or not using ontology.

TABLE III. EVALUATION OF NMI STANDARD PERFORMANCE OF BOTH STANDARD AND PROPOSED METHODS (USING ONTOLOGY)

Threshold	Ontology(NMI)	Standard(NMI)
۷T	٦٥,٠١	٦٤,٦
۸T	٦٤,٢٦	٦٤
۹T	٦٥,٧٧	٦٤,٢
١٠T	٦٦,٤	٦٥
١١T	٧٠,٢	٦٦,٢
١٢T	٦٨,٢٨	٦٦
١٣T	٦٨,٣٣	٦٣,٥

In this section, the results of the proposed method will be reported and compared to the other two methods mentioned in the first section. The compared methods are as follows:

The first rival method is introduced by Hourali and Nozari using two-stage clustering [22]

The second rival method is introduced by Boras and Sokas [23] using the K-means algorithm and the well-known TF-IDF criterion. This method will hence forth be referred to as KmeansTFIDF.

The dataset used by Nozari is applied to the proposed clustering method in this comparison, and the following results are obtained.

TABLE IV. COMPARISON OF THE PROPOSED METHOD WITH TWO OTHER COMPETING METHODS.

Number of Members	NMI	Methods
١١٠	٠,٧۶	Our proposed method
۵٠	٠,۶٨٧٨	Nozari & Hourali(1396)
-	٠,۶۶٩٣	KmeansTFIDF

As Table 4 demonstrates, the proposed method had a higher NMI criterion of 8.5.

A. *Commparing with other recommender system*

Figure 14 demonstrates the results of comparing the proposed recommender system and the system proposed by Savalanpout. The criterion of accuracy has been used to evaluate the recommender system.

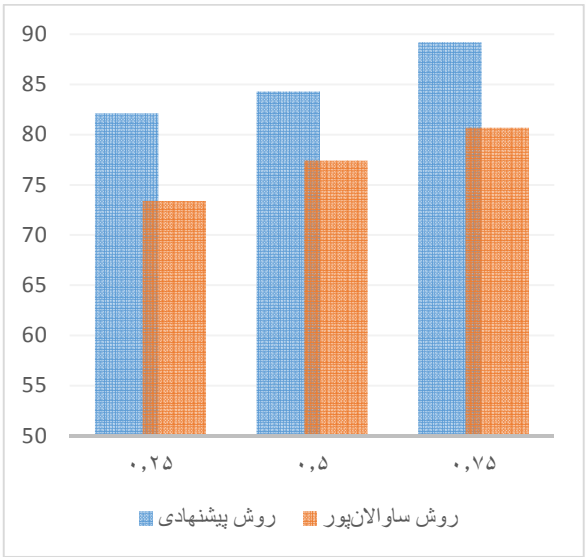

Figure 13. Comparison of proposed system, and Savalanpour system.

The expansion of news publications on the internet highlights the need to refine it to find the news that the users find interesting. News websites such as Google News, Yahoo News, Daily Me, and Tabnak news collect news from various sources and provide a general view of the news across the world. The problem that these internet services face is the great volume of news reports that drown the users. The main challenge is to help users find articles they find interesting and to reduce the users' decision-making in reading news and provide a small portion of user-related news items that the users' time and energy.

On the other hand, these systems figure out the information and services that the user has not discovered yet but might be interested in since they are capable of analyzing users' past behavior.

The news themselves are used to calculate similarities on content-based recommender systems. Using a set of news reports that have been recently published and the user's history, content-based systems try to find the contents that comply with user history.

[DOI: 10.52547/itrc.14.3.37]

When it comes to news, users' taste is more inclined to the latent content compared to the news objects themselves. Thus, news recommendation is different from the recommendation of products such as music, movies, or books. News reports have specific requirements. These requirements include a set of dynamic objects, non-complete characteristics of the users, and the difference between independent news portals. The set of dynamic objects refers to the rate of objects entering or exiting the system. Thus, the following challenges will be faced when recommending news objects:

The freshness and popularity of news reports change significantly over time, as a result of which news users rarely read old news again while in the case of music and movies, a user might use one movie or music several times.

Obtaining the taste of each news user is impossible since users' taste develops over time. A large number of online news articles requires a high-performance speed and scalability in news systems.

News systems deal with great changes, news sets with higher rates, and additions and deletions compared to music and movie sets.

Scatter, popularity imbalance, the dynamic set of objects, content criterion

VII. Conclusion

Since many studies have been conducted on recommender systems in the Persian language but a few studies focused on news recommendation because of the challenges mentioned in chapters one and above, the present study concentrated on content-based systems to improve news recommendation by analyzing the content of news items. Further focus on analyzing the text of news items can help use the algorithms in other text processing fields for this purpose and take steps towards the improvement of news recommender systems. It is recommended to combine these algorithms with the existing algorithms in recommender systems to achieve further improvement in this regard. In future, we will use deep learning methods in recognizing neighboring users and checking its effectiveness in making suggestions.

VIII. References

[1] Feng, J.; Fengs, X.; Zhang, N.; Peng, J. An improved collaborative filtering method based on similarity. PLoS ONE , 2018, 13, 1–18.

[2] Su, Z.; Lin, Z.; Ai, J.; Li, H. Rating Prediction in Recommender Systems based on User Behavior Probability and Complex Network Modeling. IEEE Access 2021, 9, 30739–30749.

[3] Sardianos, C.; Papadatos, G.B.; Varlamis, I. Optimizing parallel collaborative filtering approaches for improving recommendation systems performance. Information 2019, 10, 155.

[4] Ortega, F.; Mayor, J.; López-Fernández, D.; Lara-Cabrera, R. CF4J 2.0: Adapting Collaborative Filtering for Java to new challenges of collaborative filtering-based recommender systems. Knowledge Based Syst. 2020, 215, 106629.

[5] Zhang, F.; Qi, S.; Liu, Q.; Mao, M.; Zeng, A. Alleviating the data sparsity problem of recommender systems by clustering nodes in bipartite networks. Expert Syst. Appl. 2020, 149, 113346.

[6] Alhijawi, B.; Al-Naymat, G.; Obeid, N.; Awajan, A. Novel predictive model to improve the accuracy of collaborative filtering recommender systems. Inf. Syst. 2021, 96, 101670.

[7] Wang, D.; Yih, Y.; Ventresca, M. Improving neighbor-based collaborative filtering by using a hybrid similarity measurement. Expert Syst. Appl. 2020, 160, 113651.

[8] Liu, H.; Hu, Z.; Mian, A.; Tian, H.; Zhu, X. A new user similarity model to improve the accuracy of collaborative filtering. Knowl.-Based Syst. 2014, 56, 156–166.

[9] Patra, B.K.; Launonen, R.; Ollikainen, V.; Nandi, S. A new similarity measure using Bhattacharyya coefficient for collaborative filtering in sparse data. Knowl.-Based Syst. 2015, 82, 163–177.

[10] Polatidis, N.; Georgiadis, C.K. A multi-level collaborative filtering method that improves recommendations. Expert Syst. Appl. 2016, 48, 100–110.

[11] Zhang, F.; Zhou, W.; Sun, L.; Lin, X.; Liu, H.; He, Z. Improvement of Pearson similarity coefficient based on item frequency. Int. Conf. Wavelet Anal. Pattern Recognit. 2017, 1, 248–253.

[12] Sun, S.B.; Zhang, Z.H.; Dong, X.L.; Zhang, H.R.; Li, T.J.; Zhang, L.; Min, F. Integrating triangle and jaccard similarities for recommendation. PLoS ONE 2017, 12, e183570.

[13] Wu, C.; Wu, J.; Luo, C.; Wu, Q.; Liu, C.; Wu, Y.; Yang, F. Recommendation algorithm based on user score probability and project type. Eurasip J. Wirel. Commun. Netw., 2019, 80.

[14] Widiyaningtyas, T.; Hidayah, I.; Adji, T.B. User profile correlation-based similarity (UPCSim) algorithm in movie recommendation system. J. Big Data 2021, 8, 52.

[15] Lestari, S.; Adji, T.B.; Permanasari, A.E. WP-Rank: Rank Aggregation based Collaborative Filtering Method in Recommender System. Int. J. Eng. Technol. 2018, 7, 193–197.

[16] Tran, C.; Kim, J.Y.; Shin, W.Y.; Kim, S.W. Clustering-Based Collaborative Filtering Using an Incentivized/Penalized User Model. IEEE Access 2019, 7, 62115–62125.

[17] Vellaichamy, V.; Kalimuthu, V. Hybrid collaborative movie recommender system using clustering and bat optimization. Int. J. Intell. Eng. Syst. 2017, 10, 38–47.

[18] Widiyaningtyas, T., Indriana H., and Teguh B. A. "User profile correlation-based similarity (UPCSim) algorithm in movie recommendation system." Journal of Big Data 8.1 (2021): 1-21.

[19] Dhanani, J., Rupa, M., and Dipti, R. "Legal document recommendation system: A cluster based pairwise similarity computation." Journal of Intelligent & Fuzzy Systems Preprint (2021): 1-13.

[20] Raina, V., and Srinath K., "Natural language processing." Building an Effective Data Science Practice. Apress, Berkeley, CA, 2022. 63-73.

[21] Dudhabaware, R.S., Mangala S.M., "Review on natural language processing tasks for text documents." 2014 IEEE International Conference on Computational Intelligence and Computing Research. IEEE, 2014.

[22] Nozari, A., Hourali, M. , Tarhani. T., Improving the recognition of the subject of Persian news using two-stage clustering, Malek-Ashtar University of Technology, Tehran, 1396.

[23] Bouras, C. and Tsogkas, V. A clustering technique for news articles using word Net Knowledge-Based Systems 36 (2012) 115–128.

[24] Widmann, M. "Cohen's Kappa: What It Is, When to Use It, and How to Avoid Its Pitfalls." Retrieved June 12 (2020).

[DOI: 10.52547/itrc.14.3.37]

Maryam Hourali received her Ph.D. degree from Tarbiat Modares University in 2012. Her main research area includes Natural Language Processing, IT Engineering and Fuzzy Systems.

Mansoureh Hourali received her B.Sc. degree from Iran University of Science and Technology (IUST). And Ph.D. degree in Industrial Engineering from Payam Noor University of Tehran. She received the. Her main research area includes Technology Management, E-Government, E-Commerce, Future Research and Decision Making.

[DOI: 10.52547/itrc.14.3.37]