Quality & Quantity (2022) 56:1293–1326
https://doi.org/10.1007/s11135-021-01177-9

Using text mining algorithms in identifying emerging trends for recommender systems

Iman Raeesi Vanani[1] · Laya Mahmoudi[1] · Seyed Mohammad Jafar Jalali[2] · Kim-Hung Pho[3]

Accepted: 25 May 2021 / Published online: 31 May 2021

Abstract
Recommendation systems as the main e-commerce tools play an important role in business survival. Therefore, recommender systems and their challenges are a concern for scholars and professionals. Since this kind of system offers appropriate suggestions to online users using their interests and preferences, a lack of information about users and their purchase histories has negative impacts on the performance of recommender systems. This issue is known as "cold start problem" including cold-start user as well as cold start item and occurs when a new user logs in or an item is registered newly in a system. To deal with this problem, a lot of scientists have started studying and have done great researches annually. The first and most important step to optimize recommender systems is to have enough knowledge about previous studies and their proposed methods and algorithms using a review of these researches. Collecting and reading each of these articles is a difficult and time-consuming process. Accordingly, in this paper, we analyze the textual data collected from the best journal articles addressing the challenges of recommender systems to identify new and emerging fields in this area. This research can pave the way for future researchers of this field to develop more and more recommendation systems. The way to conduct this research is to first extract valid scientific articles in the domain of recommender systems challenges from the reputable scientific databases, the web of science. Then, using different text mining algorithms on keywords, titles, and abstracts of these articles, identification of emerging topics in this field is achieved.

Keywords E-commerce · Recommender systems · Cold start users · Text mining · Clustering

1 Introduction

The web world contains a great deal of information collected in various fields. The recent changes and advances in information technology have led to the emergence of diverse businesses and thus the production of large volumes of information on the websites (Raeesi

✉ Kim-Hung Pho
phokimhung@tdtu.edu.vn

Extended author information available on the last page of the article

Vanani and Jalali 2017). The more websites are created, the more difficult it is for them to survive due to the competition existing between them. It also makes it harder for internet users to choose their needs from the sheer amount of information. Therefore, companies need new marketing strategies such as one-to-one marketing and customer relationship management to survive in this competitive environment. One-to-one marketing represents an extreme form of segmentation, with a target segment of size one (Arora et al. 2008). Recommender systems offer a way to do this type of marketing. These systems automate web customization, allowing individual personalization for each customer. On the other hand, due to the increasing amount of information on the web, it is very difficult to find information that meets the user's needs. So, recommender systems as an essential part of the Web overcome the information overloading problem (Bagher Rahimpour Cami et al. 2017). For the abundance, dynamics, and unstructured nature of the data on the web, the need for recommender systems for both clients and website owners are an advantage and benefits, two groups. In other words, e-commerce businesses need such systems to attract customers, sell more, and ultimately make more profit by providing access for customers to the information they need from the huge amount of data.

1.1 Recommender systems

With the advent of e-commerce, there was a huge revolution in businesses to offer their products through websites and apps. The movement toward e-commerce makes businesses provide consumers with greater choices. The popularity of e-commerce sites has reshaped users' shopping habits and users prefer to spend more time shopping online (Zhou and Ding, 2018). These changes in businesses increase the amount of information that customers must process earlier than they are in a position to pick the items to respond to their demands. Recommender systems are the solution presented to the problem of information overload. These Systems are software tools in commercial applications or websites that suggest information (e.g. items, people, news articles) to end-users taking into account various types of knowledge and data, such as the user's preferences, actions, tasks and contextual information (Sielis et al. 2015). There are different definitions of recommender systems (RSs). These systems are a tool that assists users by presenting services or products that are most likely of their interest (Khanian and Mohd 2016). In the definition of RSs provided by (Cao and Li 2007; Liu et al. 2017), these systems are proposed to analyze past behaviors and make recommendations on current issues. In simpler terms, recommender systems (RSs) attempt to identify and make the most appropriate and closest offers to the users by discovering the way they think. In other words, recommender systems or RSs use the information describing users 'situations such as location, time, and task to generate more relevant and personalized recommendations (Adomavicius et al. 2011; Asabere 2013). But in most studies, recommender systems or RSs are known as a type of information filtering system which provides access to customized information that tracks user's interest in a specific context. RSs are divided into three main categories to drive the recommendations: collaborative, content-based, and hybrid filtering (Adomavicius and Tuzhilin 2005; Lu et al. 2015).

1.2 Collaborative filtering

Collaborative Filtering (CF) is the most successful and widely used recommendation technique (Abdi et al. 2018). The basic assumption of Collaborative Filtering or CF is that

people who had similar tastes in the past will also have similar tastes in the future (Çano and Morisio 2017). So, the recommendations, suggest to a new user, are based on the preferences or taste information obtained from the active user ratings. Accordingly, Collaborative Filtering (CF) is also called a people-to-people correlation. Generally, the main idea of CF is based on two principles:

1. The rating is given by user *u* to a new item*I*is very similar to the rating achieved by user *v* provided that users u and v have rated other goods similarly.
2. User *u* is likely to rate goods *I* and *J* comparably if these two goods are equally rated by other users.

CF methods can be classified into two categories Memory-Based and Model-Based (Gong et al. 2009).

1.3 Content-based filtering

CBF is based on the assumption that people who liked items with certain attributes in the past, will like the same kind of items in the future as well (Çano and Morisio 2017). CBF technique generally filters items according to the similarity of the contents the user is interested in (Lu et al. 2015). In this approach, the items offered to users are based on their purchase history and profile. For example, if a user saw and liked a comedy film in the past, the system will learn to recommend comedy films. CBF approaches usually provide obvious recommendations since they make use of content description (Lu et al. 2015). To make a list of recommendations, CBF tries to develop a model or profile for the user's interest based on examining the characteristics of items rated by the user previously. Once the system can create this profile, the comparison between this profile and the one for each product obtained to measure the user's interest in each item. Apparently, the higher the accuracy of the user-generated profile, the higher the quality of the results obtained.

1.4 Hybrid recommendation approach

The hybrid recommendation approach is derived from a combination of different RS methods which will improve and provide better recommendations than using a single technique. Various methods can be used to hybridize RS. This includes feature combination, feature augmentation (Aslanian et al. 2016), cascading, and switching method (Paradarami et al. 2017). Hybrid recommender systems are used either to leverage the power of multiple data sources or to improve the performance of existing RSs within a particular data modality (Li et al. 2014). A significant motivation to construct hybrid systems is to leverage the complementary strength of different kinds of RSs, such as CF and CBF to develop a system with more robustness. Hybrid recommendation systems are divided into a monolithic hybrid recommendation, parallel hybrid recommendation, and pipeline hybrid recommendation (Burke 2002).

1.5 Cold-start problem

Recommendation systems, due to their wide dimensions, have attracted many researchers in various sciences as well as some brands and big companies. Amazon and Google

News are two examples of the most famous websites that use recommender systems to suggest products and news articles in a personalized environment (Liu et al. 2010) and as a result, make huge profits from these systems annually. It is worth noting that despite numerous investigations into recommender systems, they encounter some problems. One of the most known problems in the recommender system (RS) is the cold start problem (Blerina Lika et al. 2014). The cold-start problem is one of the most challenging problems in RSs (Volkovs et al. 2017) and occurs when there is no adequate information about an item or a user to make relevant predictions (Kunaver and Požrl 2017). This problem has two variants: the new user cold-start problem and the new item cold-start problem (Le Hoang Son 2014). RSs encounters the first problem when a new user is introduced to the system and RS has no information about the user in giving recommendations. So, the most important issue which these systems are encountered with is known cold-start user problems because addressing this problem can also cover a somewhat new item issue. Cold start user happens while systems cannot present any offers to the users with no purchase histories. Recommender systems can only make suggestions for the users with information on the way they purchase, give ratings, and make comments.

Due to the aforementioned importance of recommendation systems, the development of these systems is a particular concern for many developers and large companies that makes them be motivated to study more. So, to optimize and improve the accuracy of these systems, additional information besides typical information about users and items are studied by researchers. As a result, a wide range of studies have been written and are available now in this field. Further studies are also needed to optimize the existing recommender systems and eliminate their shortcomings. Therefore, a complete overview of conducted studies to get knowledge about the previous researches and their proposed algorithms and techniques are inevitable. Traditional and manual methods to extract the desired information from the available articles are too time-consuming and it is necessary to design machine-based solutions for this purpose. Text mining techniques are applied to analyze the textual data and discover the most important issues. The purpose of this technique is to provide some understanding of how the text is processed without having a human read it.

In this paper, we focus on the identification of emerging and most important topics in the studies addressing the cold start problem using text mining algorithms. This study can be the basis for further researches in such a way that new researchers can try to enhance the accuracy of the existing systems and improve them with the background information they have gained from the analysis of previous studies. This research can also highlight the most important topics that future researchers need to address in this field by providing a list of the most important issues that have been addressed over the past years.

The remainder of the paper is arranged as follows:

In Sect. 2, the past studies which tackle the cold start problem are presented. Section 3 refers to the research methodology stated in the CRISP model. In Sect. 4, we pay to answer the research question including as follows:

1. What are the trends in the articles engaging in recommender system challenges?
2. What is the main focus of emerging trends in recommender systems?

In the final section, suggestions and conclusions are presented.

1.6 Literature review

Weng et al. (2008) propose a new information source namely item taxonomic applied for product and item classification that can be adopted by recommender systems or RSs. Using the implicit relations between users' item preferences and taxonomic preferences, this paper can make better offerings as well as mitigate the cold start problem. To overcome the cold start issue, Martinez et al. (2009) focus on the hybridization of CF algorithms with knowledge-based and as a consequence provide a certain degree of accuracy in the recommendation. Basiri et al. (2010) use also a hybrid recommendation technique results from a combination of a collaborative filtering or CF approach and ontological data to tackle the challenge of RSs called cold start problems. In a study that is done by Zhou et al. (2011), the similarity among users is measured through a metric measurement system that is used in CF approaches. As a result, this method enables RSs to offer users that have rated only a few items. Then, using a genetic algorithm, quality of prediction as well as the performance of the offers are improved. Golbandi et al. (2011) get paid to provide a model to make a profile for new users based on querying their responses to a kind of interview. This article presents subsequently a method to discover users' opinions about items existed in their profiles. Safoury and Salah (2013) introduce another information source, demographic data, to avoid the cold-start problems. Their study shows that RSs using information such as age, gender, and occupation known demographic data have better performance in recommendation generation, especially for new users. Blerina Lika et al. (2014) present a kind of recommender system that is a combination of classification, semantic and heuristic methods. Well-known classification algorithms are used to create user groups and to find the similarities among users, semantic similarity techniques are used and the result of this article shows the better performance of the recommender system for the high number of end users. In the other article, Pereira and Rasca (2015), address the problem of cold-start users on websites and to reduce this problem, a combination of CF and demographic information is offered. Gupta and Goel (2017) present a novel approach to cluster users based on their attributes. The result shows that this approach which applies the fuzzy c-means technique is more accurate than the traditional k-means clustering algorithm. Camacho and Alves-Souza (2018) analyzed articles between 2011 and 2017 on Collaborative Filtering-based recommender systems. This kind of system is embedded to solve the cold start problems using social networks. This paper refers to the influence of friends on the users and showed that to deal with cold start problem, social network information of users and their friends are considered a good source. The studies done on recommender system challenges are very broad that we provided some examples of these studies here. What distinguishes this research from earlier studies is that by analyzing the related previous articles between 2008 and 2018 as well as making a focus on the main topic and proposed methods in these articles, researches can have a more comprehensive look at the solutions offered to alleviate the problems affecting the accuracy of these systems. So, based on these research findings, it can be expected to develop more accurate recommendation systems to resolve the challenges. Due to the importance of marketing in e-commerce and the amount of increasing data called "big data", Amado et al. (2018) presented a research literature analysis on 1560 articles published from 2010 to 2015 to identify the main trends in this domain using text mining. The analysis was performed regarding to five dimensions: Big data, Marketing, Geographic location, sectors and products.

2 Materials and methods

2.1 Text data mining

Text mining (TM) is a very recent and increasingly interesting area of research that seeks to use automated techniques to investigate a high level of information from huge amounts of textual data and present it in a useful form to its potential users (Choudhary et al. 2009; Jalali et al. 2020a, b). TM or text mining is mainly used to define the procedure of extracting interesting and non-trivial data and knowledge from unstructured text (Gök et al. 2015; Jalali and Park 2018). The modeling of the different natural phenomena were studied using different data analysis techniques, such as statistical and mathematical modeling containing time series analysis, regression modeling, optimization and numerical analysis (Pan et al. 2019; Zarei et al. 2019; Mahmoudi et al. 2017; Mir et al. 2021; Kabir et al. 2021, 2020). Text mining can be considered the same as data mining, except that text mining methods deal with unstructured or semi-structured data like documents, emails, tweets, comments, and so on (Jalali et al. 2018; Jalali et al. 2019a, b, c). However, data mining is created to tackle structured data derived from databases (Gupta et al. 2009; Jalali et al. 2021a, Jalali et al. 2021b). Due to the unstructured language, the process of knowledge discovery in a set of text documents seems much more complex than the process of knowledge discovery in databases (Al-Daihani et al. 2016). Furthermore, text mining is an interrelated field with Natural Language Processing (NLP) (Salloum et al. 2018) by which it attempts to understand the meaning of the text as a whole. Due to the complexity of the knowledge discovery process from textual data, documents need to be converted to structured data that are suitable for analytical models (Jun et al. 2014). Therefore, in text mining, there is an additional level of difficulty where the unstructured text must be prepared for data mining algorithms and techniques such as clustering, classification, and visualization (Salloum et al. 2018; Talafidaryani et al. 2020). Afterward, the data will be analyzed by data mining tools. Data mining tools can analyze data, discover patterns to extract knowledge, and help select business strategies to use them in scientific and medical research (Hasani et al. 2018; Sohrabi et al. 2019; Jalali et al. 2019a, 2019b; Jalali et al. 2020a, b). The goal of knowledge discovery and data mining is to find attractive patterns or models which are hidden in the database (Han et. al. 2011; Sohrabi et al. 2017). Data mining as an interdisciplinary knowledge uses a combination of various techniques from statistics, machine learning, pattern recognition, databases, and visualization to solve the problem of extracting information from large databases (Jalali 2016) and text mining as mentioned above is used to obtain useful information from textual data by a variety of techniques. In general, TM tasks can be categorized into text classification, text clustering, text summarization, topic identification and association of rules (Kumar and Ravi 2016; Jalali et al. 2020a, b). In this article, we use text clustering methods to cluster the collected articles to discover the hot topics.

2.2 Text clustering

Text clustering is an approach for automatically finding classes, concepts, or groups of patterns from unstructured data (Souza et al. 2018). Text clustering is based on the Cluster hypothesis which proposes that relevant documents must have more similarities with one another than the non-relevant ones (Huang 2008). The clustering technique is normally used for large big data analysis and aims at converting large document collections into

smaller meaningful groups which can be more manageable in information retrieval and comprehension. Different methods and algorithms are employed in the document clustering process based on unsupervised learning that tries to group text by similarity. In the clustering process, the numbers, properties, and associations of the grouped sets are initially unknown (Salloum et al. 2018).

2.3 K-means algorithms

The K-Means algorithm is a well-known clustering partitioning method. The *k*-mean approach divides the data set into *k* clusters, where every cluster is subjected to be represented by the mean of points; called the centroid (Salloum et al. 2018). Initially, k cluster centroids are randomly chosen for k clusters, and the remaining items are allocated to the nearest centroid based on distance calculation (Garg et al. 2016). The purpose of this algorithm is to divide the dataset into some predefined clusters. The different steps of the K-means clustering algorithm are as follow:

- Define the k points randomly as the center of the clusters.
- Assign each record from the dataset to a cluster with the closest distance to the centroid. The most famous criteria for calculating record intervals is clustering Euclidean distance with the following formulation:

$$d_E(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

- Group the objects based on the distance to nearest centroid;
- Repeat stages 2 and 3 until there are no changes in the center of the clusters

2.4 Word cloud

Word cloud is one of the most commonly used methods for graphically displaying text data and is an image that contains the key words of a documents or set of documents (Uitdenbogerd, 2019). It is basically a weighted list of words with a certain spatial arrangement (sequential, circular, clustered, etc.) (Heimer et al. 2014). Word cloud are useful as a preliminary tools as well as a validation tool for comparing the results multiple sources of data (McNaught and Lam 2010). It helps to find the connection between the given text and the details needed and displays word significance in terms of importance using different font sizes and colors. We applied the word cloud visualization technique to illustrate the most important topics achieved by the most frequency of words in the collected papers.

2.5 Applying CRISP-DM

Data mining is an approach to get new information and extract hidden patterns from a huge amount of data collections (Jalali et al. 2019a, b, c). The information that is extracted by the data mining etiquette is not explicitly available in the database, whereas database application only projects the information that is available in the info bank with a restricted manipulation capacity (Agarwal 2013). With the growth of the field of data mining, several

data mining methodologies were proposed to systematize the discovery of knowledge from data, including the tool neutral and popular Cross-Industry Standard Process (CRISP-DM) for Data Mining (Clifton and Thuraisingham 2001). CRISP-DM is a method or technique that has been generally accepted as the top methodology for data mining projects (Piatetsky 2014; Raeesi Vanani and Jalali 2018).

Figure 1 shows the six-step phases of the CRISP-DM methodology for identification of emerging topics alleviating cold start problem summarized as follow:

The starting point in any data mining project is the definition of the project's goal that is determined in the first phase "Business Understanding". At this point, the project goals must be understood and then the problem of data mining is defined. The purpose of this research is to identify the most important and emerging topics using analyzing the previous studies addressing cold-start users in commercial website recommendation systems. During the "Data Understanding" phase, data is collected, described, and understood regarding the data mining project goal. In this case, the articles address the cold start problems collected from the top business journals. As a result, the publication source of this kind of documentation should be cited. For collecting the articles, we referred to each of the journals that are accessible in "Appendix 1". Then, all available articles from **2008** to **2018** are extracted and analyzed. As a result, about 300 articles were obtained. The "Data Preparation" phase includes all the steps needed to make the final dataset from the initial raw data

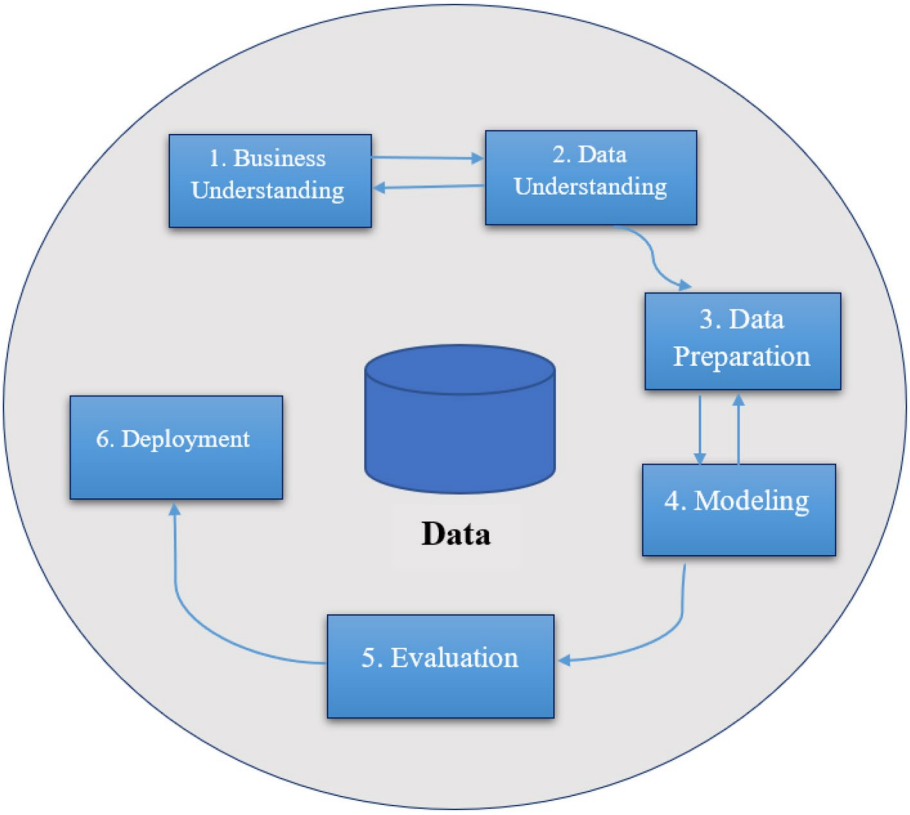

Fig. 1 CRISP-DM life cycle

for the data mining task. Accordingly, the related articles collected in the aforesaid phase are taken into consideration to make a sheet in which the title, abstract and keywords of each paper are separated by the year of publication. Then, the repetitive and meaningless words, as well as phrases with a lack of semantic information, are omitted from the data collection. In the text indexing step, texts were tokenized and converted into a list of words. To avoid considering two identical words as two tokens, all the words have been lowercased. As a final step in data cleansing, a common set of stop words were removed from the texts. Consequently, all the words obtained from the preprocessing step are weighted using the TF-IDF method, which is the best and most used weighting algorithm. The TF-IDF weighting method, which stands for "term frequency-inverse document frequency ", is used in data retrieval and text analysis to determine the importance of a word. In the process, we use the TF-IDF approach to select the most powerful weighting keywords to represent each document (Zhaowei Qu et al. 2018) and then cluster the documents. In the "Modeling" phase, different data mining techniques are applied to execute the data mining task on the preprocessed data set. In this study, the clustering techniques are used for modeling and the most accurate technique is determined based on a criterion namely Davis Bouldin Index. Thenceforth every cluster is analyzed to get knowledge about the clusters contained some articles. Within the "Evaluation" phase, the constructed model, and the steps involved in its construction are evaluated to ensure that the model is suitable for achieving the project's goal. In this study, clustering algorithms have been evaluated using the Davies-Bouldin index. After getting successful results from the model evaluation, it is deployed to implement and develop in the "Deployment" phase. In this research, the emerging topics achieved by analyzing the articles addressing RSs challenges are reported to the researchers. These topics are described in detail in the Discussion section.

2.6 Data collection

This phase involves collecting the needed data which are obtained from Web of Science listed in "Appendix 1", using the advance search on "Cold Start User" and "Recommender Systems" keywords. All articles in this area were collected within the period of 2008 and 2018 and stored separately in some sheets by the year and including information about the titles, abstracts, and keywords of the papers. The list of publications was refined to contain the records with no missing values. Non-English papers were also removed from the data set. As a result, 12 articles were eliminated and the total number of articles was reduced to 300. In Fig. 2, we demonstrated the number of extracted publications per year. In the next step, the initial dataset was processed to prepare for text mining operations.

2.7 Data preparation

Data Preparation has four steps and was applied in the following order:

- *Text Normalization* Text normalization is the process of transforming a text into a canonical form. It is useful for topic extraction where near-synonyms and spelling differences are common. For example, identical words such as "cold-start users" and "cold start users" are mapping to just "coldstartusers".
- *Text cleansing* In this step, repetitive and meaningless words and phrases that have no semantic load according to the subject of the research are removed from the data col-

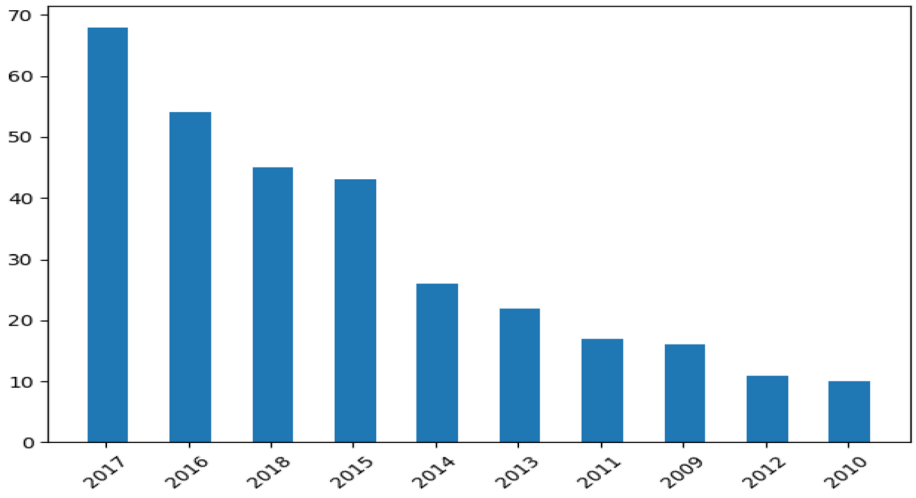

Fig. 2 Number of extracted publications per year

lection. Words such as: "academic", "paper", "research" and "become". All punctuations such as commas, question marks, and semicolons were also dropped.

- *Tokenization and removing stop words* Tokenization is a step that splits longer strings of text into smaller pieces, or tokens (Beleveslis et al. 2019). All stop words were deleted after tokenizing the words. Stop words are generally described as words that refer to the most common words in a language. Words such as "always", "who", "various" and "most" are called stop words.
- *Stemming* This step is to make sure that similar words with the same meaning used at different times or with prefixes and suffixes are considered a token. Stemming is widely used in text processing to match similar words in a text document to improve the efficiency and effectiveness of information retrieval (IR) systems (Kannan and Gurusamy 2014).

2.8 Data modelling

After the data preparation step, all the remaining words are weighted using the TF-IDF method, which is used in information retrieval and text analysis to determine the importance of a word based on the repetition of the words. Afterward, to meet the purpose of this research, it needs to identify subgroups in the data based on the similarities among the collected article's terms such that the similar contents are grouped in one cluster. To achieve these clusters with high quality and the least error, we tested different clustering algorithms including K-means, K-medoid, LDA, and DBSCAN in RapidMiner and python software within the "Modeling" phase. The best clustering algorithm has been selected based on criteria such as the Davis Boldin index. Davies Bouldin Index (DBI) is a metric for evaluating the efficiency of clusters. This index (DB) is based on the idea that for a good partition the inter-cluster separation as well as intra-cluster homogeneity and compactness should be high (Pal and Biswas 1997). DBI means to boost distance between clusters and minimize the distance between the components in a cluster. A low scatter and high distance

between clusters give the optimal number of clusters hence the minimal value of this index is desired (Sledge et al. 2010). When DBI is low, it means the clusters are not similar to each other and are separated well. It is noteworthy that the aim of this research is not to compare between the aforesaid clustering algorithms, but instead to improve the quality of clusters. Accordingly, the elected algorithm should have the best performance in clustering all the articles into subgroups with the closest similarity in their features achieved by the value of the Davis Bouldin Index. In comparison to K-medoid, LDA, and DBSCAN algorithms, K-means with the lowest Davis Bouldin index value was able to perform better not only in getting good clustering results but also in the elapsed time taken in clustering. The k-means clustering algorithm is an unsupervised learning algorithm that resolves the well-recognized clustering algorithm (Shakeel et al. 2018). The main idea in this algorithm is to define k points as the center for each cluster. The best choice for cluster centers in this algorithm is to locate the centers as far apart as possible. Thereafter, each record from the dataset is assigned to a cluster with the least distance to the center of the cluster. In the K-means algorithm, the data points are distributed evenly across all the data centers. The main objective of this algorithm is to minimize the distance between the data and the cluster centers. The aforesaid features of the k-means algorithm, as well as its implicitly and ease of implementation, make it considerably outperformed the k-medoid, LDA, and DBSCAN algorithms. The resulting clusters from the k-means algorithm in this research were created with high accuracy that makes the articles be distributed as well based on each cluster's features and content in comparison to the three above clustering algorithms.

3 Results

3.1 Most important topics in e-commerce and recommender systems articles

As discussed earlier, the advent of e-commerce has created a huge revolution in businesses and there is an exponential growth on the number of commercial websites as well as products offered on each website. Accordingly, a lot of researches have studied in the E-commerce field to develop the methods applied for trade companies. Figure 3 shows the frequency of words in this area achieved through analyzing the studies. As can be seen in Fig. 3, the Topics that are important in this area are centered around words such as user preferences, personalization, filtering systems, and recommender systems or RSs. This means that due to a large number of websites as well as items offered on each website, users become confused to select the products which meet their needs. This problem is solved by providing information filtering systems and these systems, which are also part of the recommendation systems, can provide the appropriate products concerning the behavioral, geographical and shopping history of the users. Due to the need of RSs in e-commerce, these systems have attracted a lot of attention. So, to identify more about the usages and weaknesses of these systems, we analyzed the articles collected randomly from 2008 to 2018 in the field of RSs. The most frequent and important words used in these articles were extracted and presented in Fig. 4. As illustrated in this figure, recommender system articles have been compiled in almost all fields and have attracted the attention of many researchers, which highlights the importance of RSs to study. After analyzing the RSs articles in the aforementioned areas, the problems encountered in these systems, as well as their inability to access new users, were addressed. As shown in the above figure,

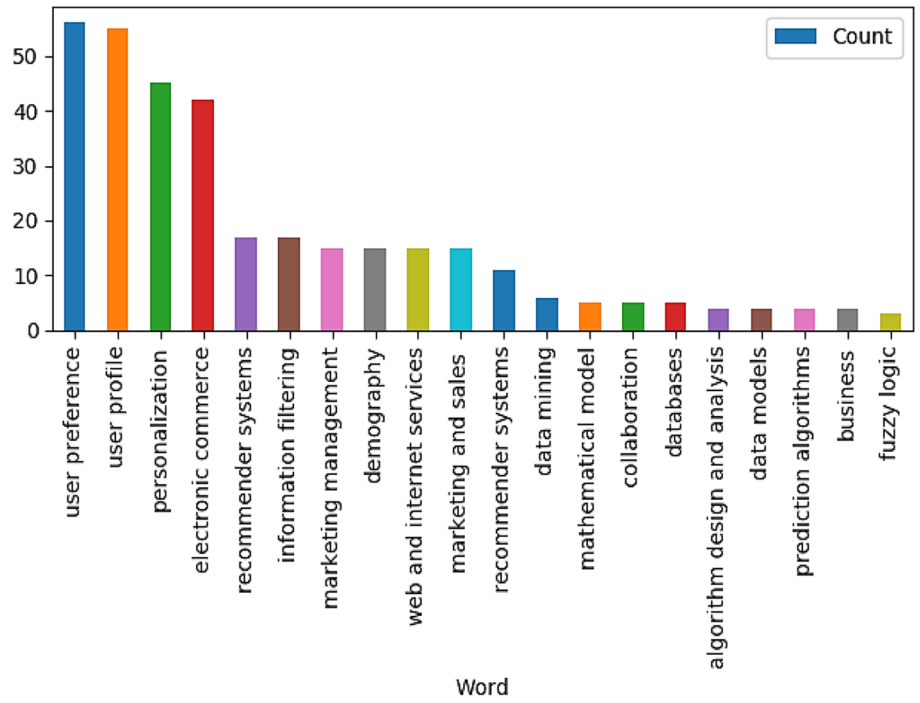

Fig. 3 The most frequent words in ecommerce articles

the problem of cold-start users (new users) has caused concern for researchers and many articles have been published to tackle this problem.

3.2 Visualizing of most important topics in cold start articles using word clouds

As shown in the figures above, 3 parts of the collected papers including titles, abstracts, and keywords are illustrated visually to more simply comprehend the main topics. Figures 5, 6, 7 represented that in addition to words such as "recommend"," system", "user" and "cold start" considered keywords and therefore they must be bold, the words including "social", "network", "inform" and "rate" are the most frequently words used in the papers addressing the cold start problem.

3.3 Most important techniques and methods alleviated cold start problem

To find the most important topics and techniques proposed in the past studies to improve RSs, all collected articles are analyzed and the most suitable clusters are created through the elected algorithm namely k-means We tried different k values from 3 to 10, and the optimal k was selected by repeated execution of the k-means algorithm with a different number of clusters. For $k=5$, the DBI was at its lowest value, meaning the clusters are separated well and there is the least similarity between the clusters. Accordingly, the articles clustered into 5 categories as follow:

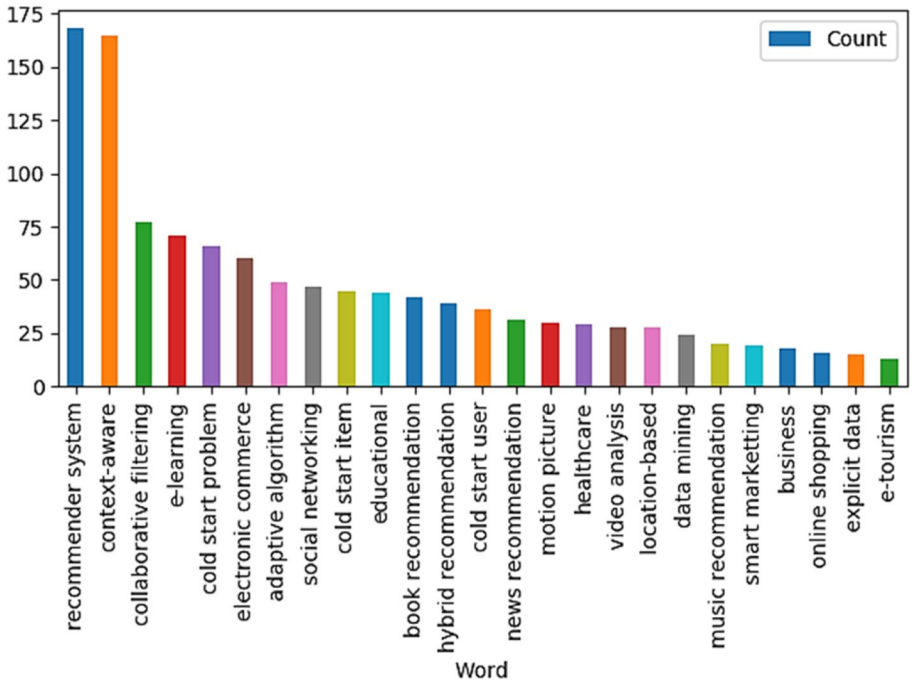

Fig. 4 Most frequent words in recommender system articles

The first cluster is including the researches who tried to find similarities between users using their additional information such as demographic info and user profile. The main idea of these studies is that people with the same demographic information have the same taste and tendencies. The proposed techniques were unable to alleviate cold start problems for the need to have information about new users. A social tagging system that has attracted the attention of many researchers and is in **the second cluster** is offered to identify new users and items using tags listed by the users. The foregoing system can predict the users' behavior by clustering the tags and find the nearest cluster to entered tag. To discover user preferences and get knowledge about their characteristic, other sources are taken into consideration to give more comprehensive information about users' opinions. **The third cluster** that is relating to social networks has been the concern of many scholars to get access to this information through the comments posted by the users about a product including film or music to optimize the recommender systems or RSs in offering cold start users. Another topic that has become a challenging issue is big data that make researchers develop machine learning applications in clustering and classification of a huge amount of data and analyze them to get knowledge about users' behavior patterns. The discussion about big data and machine learning algorithms are including in **the fourth cluster.** Strong Users or opinion leaders with high rating records can have a significant impact on predicting new users' attitudes by assigning them to the nearest opinion leaders and analyzing their characteristics to generalize the achieved features to new users. The ways and techniques to identify opinion leader are considered in **the fifth cluster.**

Fig. 5 Word cloud displaying the most frequently used words in the titles of collected articles

4 Discussion

To discover the trends of articles taking cold start problem, the findings are presented regarding the year of publication that is the main focus of this research.

4.1 Analysis of the trends of articles alleviating the cold start problem

In this section, the most important topics of the articles each year from 2008 to 2018 are presented along with a detailed description of the proposed techniques and methods to follow the trends of the articles (Table 1).

4.2 Clustering cold-start users researches of 2008

The optimal number of clusters for this year is four with the Davies Bouldin value of 0.028 presented in the Table 2.

Fig. 6 Word cloud displaying the most frequently used words in the keywords of collected articles

4.3 Clustering cold-start users researches of 2009

The optimal number of clusters for this year is five with the Davies Bouldin value of 0.016 presented in the Table 3.

4.4 Clustering cold-start users researches of 2010

The optimal number of clusters for this year is four with the Davies Bouldin value of 0.066 presented in the Table 4.

4.5 Clustering cold-start users researches of 2011

The optimal number of clusters for this year is three with the Davies Bouldin value of 0.13 presented in the Table 5.

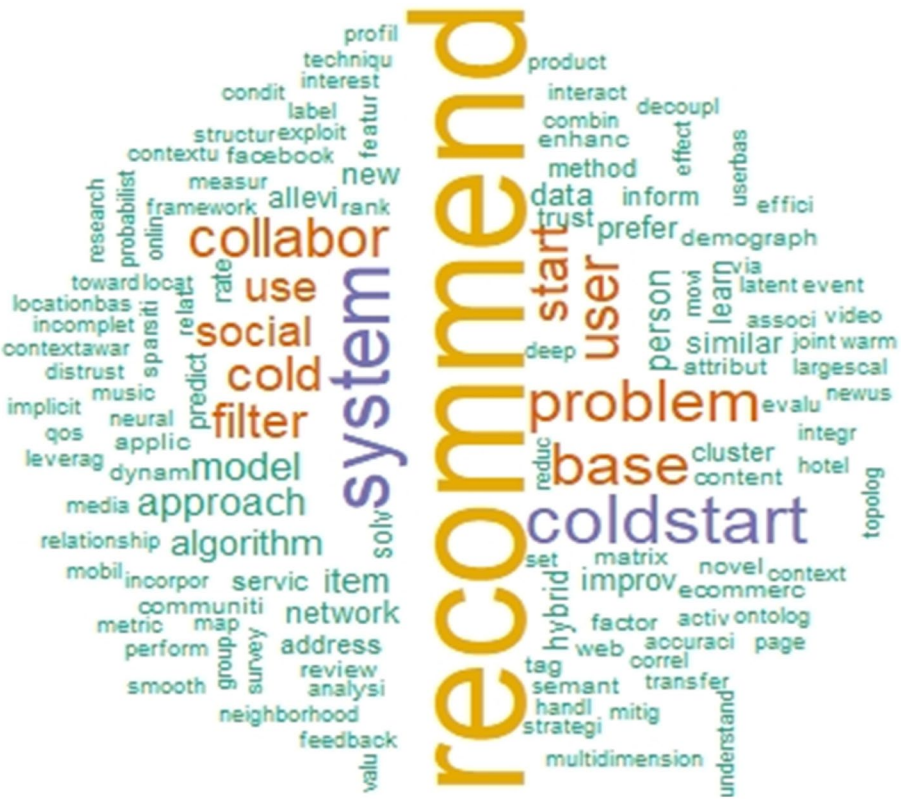

Fig. 7 Word cloud displaying the most frequently used words in the abstracts of collected articles

Table 1 Most important topics in alleviating cold start problems

Cluster	Keywords
Cluster 0	Similarity measure, semantic web, demographic information, user profile, collaborative filtering, Taxonomy, trust network, ontology, latent feature, additional information
Cluster 1	Tag, social, bookmark system, URL, fuzzy, discrete imperialist competitive, web services,
Cluster 2	Social-network, trust, similarity, friend, Facebook, online, twitter, sentiment analysis, social behavior, Genetic algorithm
Cluster 3	Neural network, deep learning, probabilistic model, big data, matrix factorization, machine learning
Cluster 4	Strong user, nearest neighbor, rating matrix, leader, social network,

4.6 Clustering cold-start users researches of 2012

The optimal number of clusters for this year is four with the Davies Bouldin value of 0.023 presented in the Table 5 (Table 6).

Table 2 Clustering cold start articles in recommendation systems for 2008

Cluster no.	Cluster name	Terms	Description
Cluster 0	Cross-level association Rules	Associ_rule Cross_level_associ Rule_clar Collabor_filter Item_relationship Algorithm_gener Gener Rule_approach Hybrid Novel_hybrid	This cluster alleviate the cold-start users problem using multi-level association rules (CLAR). One of the techniques used in recommender system is Collaborative filtering. The weakness of this technique in dealing with cold start problems is that it focuses on finding similarities between users and no longer examines product features. In addition to examining the similarities between users, the association rules find the features of products and also the relationship between the product and the user
Cluster 1	Semantic ranking	Similar_measur Pip_measur Rate Sparsity_data Hybrid_approach Heurist_similar Compar	In this cluster, by providing similarity measurements, there were valuable recommendations for users with a little ranking history. The advantage of this method is that there is no need for additional data but it can't offer any suggestions to the users with no rating records
Cluster 2	Taxonomic data classification	Taxonomi Novel_recommend Benchmark_techniqu Taxonomi_prefer	In this cluster, data are classified based on their meanings so that users can easily access the information they need. Taxonomi is a method popularly applied for product or content classification to make better quality recommendations
Cluster 3	Web 2 0	Trust-network Free-scale-network network Rate-matrix Network-recommend Web 2 0	In Web 2, users, in addition to being able to generate content themselves, could also share it with other users, thereby building trust-based relationships. A new user can connect to a network by rating comments

Table 3 Clustering cold start articles in recommendation systems for 2009

Cluster no.	Cluster name	Terms	Description
Cluster 0	Demographic attributes	Algorithm_base User_attribut Demograph Attribute_rate Brows_tree User_prefer Demograph_vector Implicit_inform	This cluster points out that in addition to implicit information, some systems use users' personal information. For example, age, gender and nationality of users can be a good source to get information from the users and offer suggestions. This type of information is called demographic information which some RSs are based and operate on this basis that users with similar demographic attributes are likely to have similar tastes and desires
Cluster 1	Knowledge-based RSs	Knowledge-base Intellig-system Fuzzi-base Fuzzi-Similar Multi-layei- perceptron Kernel	Knowledge-based systems make recommendations based on their perception of customer needs and product features. In other words, these systems make a list of recommendations using the information about customers and products. To analyze the information and get knowledge, these systems use common methods such as genetic algorithms, fuzzy, neural networks, etc
Cluster 2	Semantic web	Semant_Web Web_base Graph Multi diment Ontolog Histori Pedagog_factor	The purpose of RSs is to direct users to pages that best meet their needs and interests. Using web-based algorithms, graph partitioning, and distributed learning automation, the graphs between semantic web pages in the URL and the similarity of a website's pages with each other can be determined to obtain information to offer new pages to the users
Cluster 3	Latent features	Latent_Feautur Weight_distribut Data_sparsiti Attenu_factor Weight_gener	Two important issues in collaborative systems are the problem of cold start and data sparsity. Data sparsity arises because in most cases, users are only interested in certain items and only rank them. Therefore, in each vector of the user-items matrix, a lot of information has a null value. To solve this problem, many researchers have presented dimensional reduction techniques, which are mainly based on matrix factorization. Dimensional reduction methods involve extracting attributes using latent variables to describe the hidden data

Table 3 (continued)

Cluster no.	Cluster name	Terms	Description
Cluster 4	Trust-based RSs	Trust-metric Fine-grain Trust-user Trust-network Trust-metric Abnorm	With the advent of trust-based networks among users and according to the features of such networks, new methods and algorithms called trust-based methods are proposed to improve RSs. Trust-based RSs provide valuable information to the users based on a trust relationship. Trust is the amount of users' faith in each other formed based on the criteria such as the ability and power of people

Table 4 Clustering cold start articles in recommendation systems for 2010

Cluster no.	Cluster name	Terms	Description
Cluster 0	RSs based on clustering algorithms and classification methods	Cluster Cluster-filter associ-cluster cluster-algorithm probabilist-neural-network	Probabilistic neural network is used to identify and classify patterns. In general, they learn the patterns in the training data and then categorize new samples based on these patterns. Probabilistic neural networks are one of the most important monitoring methods used to identify and classify patterns
Cluster 1	Strong Users	Ctrong_user Nearest_neighbor Multi_attribut Rate_matrix Similar_user Activ_user	The k-nearest neighbors algorithm is the most famous method of collaborative filtering. One method to analyze a KNN algorithm is to consider it as a process that produces a graph. To solve the problem of cold start, a strong user neighbor can be assigned to cold-start users with no profiles till their profile get bigger enough to compute a separate neighbor. The term "strong users" is used to describe users who have rated a large number of items and those who have a high degree of "impact" at recommending others
Cluster 2	Computational Intelligence methods	Calcul comput_intellig Genet_algorithm Accuracy Bayesian Geographi Latent_factor	These methods include Bayesian network-based methods, clustering, artificial neural network, genetic algorithm and fuzzy set-based methods. Given that the existence of incomplete information is a major challenge in the real world. Therefore, Bayesian networks are offered to deal with these issues. Bayesian networks is a probabilistic graphical model to represent the conditional relationship between variables via a directed acyclic graph. In this type of research, the traits of age, nationality, employment, geography are directly considered as effective factors on the products that are probably of interest to the user
Cluster 3	Tag-based RSs	Tag Event Social tag Social event Histori Cluster Weigth	Social tagging systems are becoming increasingly common. These systems allow users to easily organize their resources. The increase of these systems has led to an increase in the volume of data and providing appropriate information from this vast volume of data that has been considered by researchers. These systems identify users' interests using their behavior in the past and offer appropriate suggestions to them. When new users log in, there is usually not enough information to generate suggestions for them, leading to a cold start problem. Many attempts have been made to cluster the tags. Accordingly, upon entering a new user and listing his desired tags, the k neighbors of each new tag and the resources tagged by the neighbors are extracted and weighed, and several sources with the highest weight are suggested

Table 5 Clustering cold start articles in recommendation systems for 2011

Cluster no.	Cluster name	Terms	Description
Cluster 0	Social Bookmarking	Tag bookmark tag_recommend bookmark_system social_bookmark tag_algorithm bookmark_url	Social bookmarking systems have grown significantly. These systems allow users to mark their desired URLs using keywords called tags. The tags can be used to search and categorize bookmarks. The problem with these systems is their inability to offer to new users who do not have enough history. The PUT-Tag system uses the similarity of tags, user, and content to suggest tags and it has the same performance against new and active users
Cluster 1	Hybrid algorithms	Kernel_fuzzi cmean Fuzzi_radial_basi cluster Fuzzi Network Kfcm Hybrid	The c-means fuzzy clustering algorithm, FCM, is a method often used to identify patterns. This method has advantages such as good modeling of results However, it is not able to determine the number of clusters. To address this problem, *a* kernel *fuzzy* c-means algorithm (KFCM)is proposed based on genetic algorithm. to optimize the c-means fuzzy clustering, which is a combination of improved genetic algorithm and kernel methods. The improved genetic algorithm is used to optimize the centrality of the clusters, and then the KFCM algorithm is used to classify. Thus, the efficiency of the FCM algorithm is improved
Cluster 2	Ontology	Ontologi Semant_strauctur Node Node_recommend active_node semant web	The Semantic Web seeks to make the information on the Web comprehensible to machines. Semantic web uses the methods and techniques proposed in artificial intelligence such as natural language processing, image processing, ontology, and so on. In fact, ontology represents a common understanding of a domain and its related vocabularies. Technologically, ontologies represent classified and unclassified concepts and relationships. ontologies are often conceived and visualized as a semantic network of related conceptual nodes. Accordingly, the existing shortcomings in the recommender systems can be reduced by providing suggested items based on the user profile, metadata, and extracting connections between them. In this regard, in order to extract similarities between user profiles and metadata, ontology has been used

Table 6 Clustering cold start articles in recommendation systems for 2012

Cluster no.	Cluster name	Terms	Description
Cluster 0	Genre Correlations	Genr_correl data_Sparsiti Movi Music Categori Algorithm Web	RSs need a certain amount of user preferences data to be able to identify a group of users and suggest items based on that. But if there is no data, a cold start problem arises and the systems are not able to make any suggestions. For this purpose, classification correlations are used for film and music RSs to make offerings using the genre information of films proposed by experts
Cluster 1	Trust-semantic fusion-based	Semant Fusion Trust Semant_fusion tsf approach distribut_network Knowledg	The semantic trust fusion-based recommendation approach offers a higher quality of suggestions compared to the group and semantic recommendation algorithms, and this approach is more accurate against the data sparsity problem than the mentioned approaches
Cluster 2	Psychometric measurements	Learn_system baselin_model psychometr commun User_model	Mitigating the issue of fatigue is critical for systems that support lifelong learning. This is due to the complex measurement methods required to format user cognitive models. Accordingly, a baseline cognitive modeling approach is proposed to initialize user models with psychometric measurements from the target community that alleviates the cold start problems
Cluster 3	Random Walk	Makovin Random_walk user_profil Graph Node Hybrid cf	Target users can be predicted and suggestions made to them using item content and the user's social information. The RS is widely used in social networks with a variety of methods such as collaborative filtering. A hybrid CF based on a random walk use item content and user social information to make recommendations. This method can improve the cold start problem by offering to the users with lack of rating information

4.7 Clustering cold-start users researches of 2013

The optimal number of clusters for this year is four with the Davies Bouldin value of 0.036 presented in the Table 7.

4.8 Clustering cold-start users researches of 2014

The optimal number of clusters for this year is three with the Davies Bouldin value of 0.042 presented in the Table 8.

4.9 Clustering cold-start users researches of 2015

The optimal number of clusters for this year is four with the Davies Bouldin value of 0.024 presented in the Table 9.

4.10 Clustering cold-start users researches of 2016

The optimal number of clusters for this year is four with the Davies Bouldin value of 0.35 presented in the Table 10.

4.11 Clustering cold-start users researches of 2017

The optimal number of clusters for this year is four with the Davies Bouldin value of 0.089 presented in the Table 11.

4.12 Clustering cold-start users researches of 2018

The optimal number of clusters for this year is three with the Davies Bouldin value of 0.126 presented in the Table 12.

4.13 Conclusions & future works

As mentioned earlier, e-commerce has made a major revolution in businesses to provide their consumers with more products. This kind of business caused some problems containing making users more confused in selecting items. Therefore, a recommendation system came to solve this problem. But these systems were not fully successful in offering recommendations to all users as well as identifying new items. Cold start users and cold start items are considered the most important problems that threaten these systems. The aforementioned problems have made a lot of researchers and companies to study them to develop RSs. So, in this research we analyzed the studies addressing the cold start problems to identify emerging topics and techniques in this area. As a result, additional data sources like social networks, leader opinion, and presenting hybrid methods are the most important topics that attract a lot of scholars to focus on them in their studies. Based on the results of this research and the analysis carried out in this field, it cannot be claimed that the problem of recommender systems or RSs has been completely

Table 7 Clustering cold start articles in recommendation systems for 2013

Cluster no.	Cluster name	Terms	Description
Cluster 0	Similarity matrix	Top-n Similar Cluster Similar-matrix vector-cosin method	In this cluster, the vector cosine method is proposed to obtain user similarity matrix and then cluster users into different groups. The top-N recommendations are produced by averaging of each item ratings to improve RSs in overcoming the cold-start problem
Cluster 1	Timestamp	Rate Timestamp Collabor Predict-algorithm Group-similar	Most algorithms designed so far take into account user ratings equally in all situations, ignoring the fact that user interests and requirements may change over time. In this cluster, the presented RSs try to make appropriate recommendations through the timing votes, predefined groups of items, and similarities between users to predict the ratings a user will give to an item in the close future
Cluster 2	Geographical-fuzzy clustering	Fuzzi Fuzzi-geographi demograph-cluster fuzzi-cluster motion -pictur	Fuzzy geographic clustering (FGWC) is recognized as one of the efficient algorithms in geographic and population analysis. This algorithm is a combination of fuzzy algorithm and CF that overcome the cold start problem based on the nearest neighbor.
Cluster 3	Social Network	Social-network Web-servic relationship User-prefer Collabor Histori Spars-matrix	Due to the popularity of social networks this year, as well as the increase in the number of web services, the use of the user's social information can greatly contribute to the accuracy and optimization of offering suggestions. Social networks such as Twitter, in which users share the contents they want, can help systems to identify better the users' preferences

Table 8 Clustering cold start articles in recommendation systems for 2014

Cluster no.	Cluster name	Terms	Description
Cluster 0	Prior rating	Prior-rate Data-sparsiti similar-measur Rate-confid Motiv Dimens	User rating is the basis of RSs in e-commerce. Lack of motivation to rate items after the purchase reduces the effectiveness of such systems, and due to lack of data, the cold start problem raises. Prior rating is an important source of information for RSs. So, by making the mediation environment attractive, the user can be forced to score points. Among these motivational methods, three-dimensional systems can be used to present products
Cluster 1	Facet-based model	Facet Comput-model histori Twitter context-model	Facet-based process modeling describes different aspects and preferential benefits of the whole user community. By combining the different preferences of all users with microblogging features such as content, time and social network, the benefits and advantages of this process can be tracked. Making a facet-based process can significantly reduce the user's cold start problem
Cluster 2	Discrete imperialist competitive algorithm	Tag-base-recommend system Linear-program Tag-cluster discret-imperialist competit	RSs analyze users' interests based on their past behavior and offer them the best suggestions; However, these systems are incapable to offer to the users with no or little information. Researchers proposed different algorithms to better cluster tags including the imperialist competitive algorithm in which tags are clustered using the K-medoids algorithm In this method, when a new user logs in and enters his tags, the system offers him several sources with the largest weight

Table 9 Clustering cold start articles in recommendation systems for 2015

Cluster no.	Cluster name	Terms	Description
Cluster 0	Novel hybrid methods	Collabor-filter novel-hybrid method system-inform optim Faca-dtrs Nhsm Mipegwac Arm	To alleviate cold start problems, different novel hybrid methods to optimize RSs. A novel hybrid user-based fuzzy collaborative filtering is a new hybrid method proposed to combine the advantages of different groups of methods and eliminate their disadvantages using some special methods The results of research on combined methods showed that these methods have performed better in increasing the accuracy of RSs. These algorithms include FACA-DTRS, MIPFGWC, ARM and NHSM.
Cluster 1	Implicit social trust	Sentiment-analysi Genet-algorithm Mathemat-model Trust Similar Social-network	Implicit social trust is created based on the relationship between an active user and his friends on online social networks (OSNs). The parameters of trust in a user, firstly, are obtained from the method of sentiment analysis on his comments. Secondly, they emerge from the aspects of trust that are extracted from the accounts of the user's friends. Trust can significantly solve the user's cold start problem
Cluster 2	Dynamic RSs	Uncertainty Miss-valu Hybrid-method optim Heurist-algorithm Dynam-recommend-system Adaptat-model motion-pictur	Many organizations believe that websites can attract new customers and retain old customers. To achieve this capability, web log files can be used to store customer access patterns to the website. A dynamic RS includes all registered and unregistered customers. This is a rational recommender technique that uses language patterns to make recommendation. The results of the studies indicate that the proposed system improved accuracy and quality of recommendations
Cluster 3	Matrix factorization	Matrix-factor Auxiliari-inform Latent-factor Spars-matric Collabor Big-data	Collective filtering is based on users' past performance, while content-based systems use user characteristics and items to predict and suggest. The two main areas of collaborative filtering are neighborhood methods and latent factor. Matrix factorization is one of the latent factor models that use matrix analysis to predict. The most important advantage of this method is that there is no need to obtain additional information to predict and make recommendations

Table 10 Clustering cold start articles in recommendation systems for 2016

Cluster no.	Cluster name	Terms	Description
Cluster 0	Topology	Social-network graph Topolog Link Commun Profil Social-behavior	In the approach of RSs based on graph topology, the intrinsic properties of the network structure are used to determine the similarity between nodes. The purpose of these systems is to find active and effective people in social networks through social graph analysis. By getting to know more and more active people in social networks, more quality suggestions can be given to users
Cluster 1	Big data	Machin-learn big-data mahout Deep-learn collabor Larg-scale	In this cluster, the aim is to provide an environment for the rapid development of machine learning applications and scalable algorithms in the three areas of clustering, classification و and recommendation. Mahout, which is for analyzing big data, can be used to produce more compatible and scalable machine learning programs to tackle cold start problems
Cluster 2	Novel methods	Comput-model motion-pictur matrix-factor combin demograph collabor-filter content-base	In this cluster, the researchers try to present the methods with high accuracy by combining the proposed methods. One of these combination methods played a great role in solving the cold start problem has been the combination of the collaborative filtering method with the content-based method. The content-based system is used for clustering based on demographic information and the collaborative filtering is applied to identify the nearest neighbor
Cluster 3	Social media	Motion-pictur facebook twitter Music Profil Twitt Socia- network analysi	In these articles, cold start problems are mitigated using social media. In this approach, in addition to collecting general information from users' social media accounts, the users' preferences and tastes are also extracted. The information helps the RS to get knowledge about what kind of movies and music will be most popular with the users. The results show that this approach is effective in reducing cold start problems

Table 11 Clustering cold start articles in recommendation systems for 2017

Cluster no.	Cluster name	Terms	Description
Cluster 0	Reinforcement algorithms	Contextu multi-arm bandit diversify-recommend reinforc-learn reward Punish Switch	Reinforcement learning is one of the machine learning trends inspired by behavioral psychology. These algorithms are classified as unsupervised algorithms. In these algorithms, a machine is trained to make a particular decision based on its current position (set of available variables) and permissible actions (e.g., forward, backward, etc.). The first time, this decision can be completely random and for each action or behavior that occurs, the system gives feedback. From this feedback, the machine realizes that the decision is right or not. For the dependence of the current state and behavior on previous states and behaviors, the Markov decision-making process can be one example of this group. Neural network algorithms can also be considered in this category. These methods allow machines to automatically determine the desired behavior to maximize its performance
Cluster 1	Opinion Leaders	Social-filter graph Deleg social-network psychology-test neighbor Central Node	In social filtering systems, users with similar preferences form a social community in which people are connected through trust. This system like other RSs has the cold start problem. In fact, making effective recommendations in cold start conditions is a fundamental task in a RS because giving bad advice can drastically reduce a website's revenue, market share, and also credibility. One effective way is to identify the most influential users, called opinion leaders within the trust network. To identify these people, advanced graph concepts are used
Cluster 2	Heterogenous information	Heterogeny-inform Meta-path network Latent-inform auxiliary-inform Cross-HIN	To meet the "cold start user" challenge, these articles suggest the use of heterogeneous information in users' information networks. Integrating heterogeneous information across multiple user networks can help to get more information from any aspects of users' behaviors and preferences

Table 11 (continued)

Cluster no.	Cluster name	Terms	Description
Cluster 3	Social network	Hypergraph hybrid-approach social-network intern Similar Rank Hmf	Current research on RSs has focused mainly on social networking algorithms. Because the social networking structure of recommender systems has not been fully analyzed. Therefore, many researchers tried to analyze the internal relationship of social networks. HMF, for example, is a proposed hybrid system that uses hypergraph topologies to describe and analyze the internal relationship of a social network. Numerous factors, including textual information, user attribute, item attribute, and similarity of user rankings, are achieved based on these systems

Table 12 Clustering cold start articles in recommendation systems for 2018

Cluster no.	Cluster name	Terms	Description
Cluster 0	Online Social Network	Associat-mine Dynam behavior Osn characterist cross osn Time Factor-method	Online Social networks (OSN) play an essential role in people's daily life, and large numbers of users use multiple online social networks. In this section, the articles examine the dynamic features of user behavior to track user interests in each OSN and provide a framework to create dynamic communication between OSNs. The results show that the proposed framework works well to make offerings to the new users
Cluster 1	label suggestions	Hybrid Trust Label relationship factor Valu Label-set	In this cluster, some algorithms are proposed to examine the users' preferences as well as the relationship between items on social network to increase the variety of label suggestions. Then, by combining the initial value of the label with the popularity of that, the accuracy of the label suggestion is increased
Cluster 2	Wearable sensors	Complex-activ Wearable-comput acceler Grid-cell factor-matrix mobil	Recognizing complex activity is a significant issue in mobile and wearable computing. Complex activity can be identified by user's location to make recommendations based on his data location. Detecting current activity using location data, the system will no longer have a problem with the new users

resolved and it is required to execute studies on further features of these systems in the future. So, researchers can classify similar groups based on another item called user prioritization in the product selection by tracking user clicks on items. Additionally, consideration of more social networks such as Instagram and LinkedIn besides twitter and Facebook can help achieve more accurate information about user behavior and preferences. Another subject offered to study is to analyze other information gained using users' video and photo processing in addition to the analysis of their textual data. Studies on users' medical information can also make an appropriate and accurate offer to them based on their physical condition and it will be a revolution in human life.

Appendix 1

Journals	Impact factor
Journal of Information Technology	4.953
Information Sciences	4.832
Knowledge-Based Systems	4.529
Expert Systems with Applications	3.928
Computer-Supported Collaborative Learning	3.469
Computers in Human Behavior	3.435
Neurocomputing	3.317
Decision Support Systems	3.222
Information Systems	2.777
Information Systems Research	2.763
Journal of Management Information Systems	2.356
Journal of the Association for Information Systems	2.012

References

Abdi, M.H., Okeyo, G., Mwangi, R.W.: Matrix factorization techniques for context-aware collaborative filtering recommender systems: a survey (2018)

Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng. **17**(6), 734–749 (2005)

Adomavicius, G., Mobasher, B., Ricci, F., Tuzhilin, A.: Context-aware recommender systems. AI Mag. **32**(3), 67–80 (2011)

Agarwal, S.: Data mining: data mining concepts and techniques. In *2013 International Conference on Machine Intelligence and Research Advancement* (pp. 203–207). IEEE. (2013)

Al-Daihani, S.M., Abrahams, A.: A text mining analysis of academic libraries' tweets. J. Acad. Librariansh. **42**(2), 135–143 (2016)

Amado, A., Cortez, P., Rita, P., Moro, S.: Research trends on big data in marketing: a text mining and topic modeling-based literature analysis. Eur. Res. Manag. Bus. Econ. **24**(1), 1–7 (2018)

Arora, N., Dreze, X., Ghose, A., Hess, J.D., Iyengar, R., Jing, B., Zhang, Z.J.: Putting one-to-one marketing to work: personalization, customization, and choice. Mark. Lett. **19**(3), 305–321 (2008)

Asabere, N.Y.: Towards a viewpoint of context-aware recommender systems, (CARS) and services. Int. J. Comput. Sci. Telecommun. **4**(1), 10–29 (2013)

Aslanian, E., Radmanesh, M., Jalili, M.: Hybrid recommender systems based on content feature relationships. IEEE Trans Ind Inf **3203**, 1 (2016)

Basiri, J., Shakery, A., Moshiri, B., Hayat, M. Z.: Alleviating the cold-start problem of recommender systems using a new hybrid approach. In 2010 5th International Symposium on Telecommunications (2010) (pp. 962–967). IEEE

Beleveslis, D., Tjortjis, C., Psaradelis, D., Nikoglou, D.: A hybrid method for sentiment analysis of election related tweets. In *2019 4th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)* (2019) (pp. 1–6). IEEE

Burke, R.: Hybrid recommender systems: survey and experiments. User Model. User-Adapt. Interact. **12**(4), 331–370 (2002)

Camacho, L.A.G., Alves-Souza, S.N.: Social network data to alleviate cold-start in recommender system: a systematic review. Inf. Process. Manage. **54**(4), 529–544 (2018)

Cami, B.R., Hassanpour, H., Mashayekhi, H.: A content-based movie recommender system based on temporal user preferences. In *2017 3rd Iranian Conference on Intelligent Systems and Signal Processing (ICSPIS)* (2017) (pp. 121–125). IEEE.

Çano, E., Morisio, M.: Hybrid recommender systems: a systematic literature review. Intell. Data Anal. **21**(6), 1487–1524 (2017)

Cao, Y., Li, Y.: An intelligent fuzzy-based recommendation system for consumer electronic products. Expert Syst. Appl. **33**(1), 230–240 (2007)

Choudhary, A.K., Oluikpe, P.I., Harding, J.A., Carrillo, P.M.: The needs and benefits of text mining applications on post-project reviews. Comput. Ind. **60**, 728–740 (2009)

Clifton, C., Thuraisingham, B.: Emerging standards for data mining. Comput. Stand. Interfaces **23**(3), 187–193 (2001)

Garg, N., Gupta, R.K.: Clustering techniques on text mining: a review. Int. J. Eng. Res. **5**(4), 241–243 (2016)

Gök, A., Waterworth, A., Shapira, P.: Use of web mining in studying innovation. Scientometrics **102**(1), 653–671 (2015)

Golbandi, N., Koren, Y., Lempel, R.: Adaptive bootstrapping of recommender systems using decision trees. In Proceedings of the fourth ACM international conference on Web search and data mining, WSDM '11, pages 595–604, New York, NY, USA, ACM. (2011)

Gong, S. Ye, H. Tan, H.: Combining memory-based and model-based collaborative _ltering in recommender system," in Proc. IEEE Int. Conf. Paci_c-Asia Circuits, Commun. Syst., (2009) pp. 690–693

Gupta, V., Lehal, G.S.: A survey of text mining techniques and applications. J. Emerg. Technol. Web Intell. **1**(1), 60–75 (2009)

Han, J., Pei, J., Kamber, M.: Data Mining: Concepts and Techniques. Elsevier, Amsterdam (2011)

Hasani, H., Jalali, S.M.J., Rezaei, D., Maleki, M.: A data mining framework for classification of organizational performance based on rough set theory. Asian J. Manag. Sci. Appl. **3**(2), 156–180 (2018)

Heimerl, F., Lohmann, S., Lange, S., Ertl, T.: Word cloud explorer: Text analytics based on word clouds. In 2014 47th Hawaii International Conference on System Sciences (2014) (pp. 1833–1842). IEEE

Huang, A.: Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand* (2008) (Vol. 4, pp. 9–56)

JafarJalali, S.M.: Visualizing e-government emerging and fading themes using SNA techniques. In *2016 10th International Conference on e-Commerce in Developing Countries: with focus on e-Tourism (ECDC)* (2016) (pp. 1–4). IEEE

Jalali, S.M.J., Park, H.W.: State of the art in business analytics: themes and collaborations. Qual. Quant. **52**(2), 627–633 (2018)

Jalali, S.M.J., Mahdizadeh, E., Mahmoudi, M.R., Moro, S.: Analytical assessment process of e-learning domain research between 1980 and 2014. Int. J. Manag. Educ. **12**(1), 43–55 (2018)

Jalali, S. M. J., Karimi, M., Khosravi, A., Nahavandi, S.: An efficient neuroevolution approach for heart disease detection. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)* (2019) (pp. 3771–3776). IEEE

Jalali, S.M.J., Khosravi, A., Alizadehsani, R., Salaken, S.M., Kebria, P. M., Puri, R., Nahavandi, S.: Parsimonious evolutionary-based model development for detecting artery disease. In 2019 IEEE International Conference on Industrial Technology (ICIT) (2019) (pp. 800–805). IEEE

Jalali, S.M.J., Khosravi, A., Kebria, P.M., Hedjam, R., Nahavandi, S.: Autonomous robot navigation system using the evolutionary multi-verse optimizer algorithm. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)* (2019) (pp. 1221–1226). IEEE

Jalali, S.M.J., Ahmadian, S., Khosravi, A., Mirjalili, S., Mahmoudi, M.R., Nahavandi, S.: Neuroevolution-based autonomous robot navigation: a comparative study. Cogn. Syst. Res. **62**, 35–43 (2020)

Jalali, S.M.J., Park, H.W., Vanani, I.R., Pho, K.H.: Research trends on big data domain using text mining algorithms. Digit. Scholarsh. Humanit. (2020)

Jalali, S.M.J., Ahmadian, S., Khosravi, A., Shafie-khah, M., Nahavandi, S., Catalao, J.P.: A novel evolutionary-based deep convolutional neural network model for intelligent load forecasting. IEEE Trans. Ind. Inf. (2021)

Jalali, S.M.J., Ahmadian, S., Khodayar, M., Khosravi, A., Ghasemi, V., Shafie-khah, M., Nahavandi, S. and Catalão, J.P.: Towards novel deep neuroevolution models: chaotic levy grasshopper optimization for short-term wind speed forecasting. Eng. Comput., (2021) pp.1–25

Jun, S., Park, S.S., Jang, D.S.: Document clustering method using dimension reduction and support vector clustering to overcome sparseness. Expert Syst. Appl. **41**, 3204–3212 (2014)

Kabir, H.D., Khosravi, A., Nahavandi, S., Srinivasan, D. (2020). Neural network training for uncertainty quantification over time-range. IEEE Trans. Emerg. Topics Comput. Intell.

Kabir, H.D., Khosravi, A., Kavousi-Fard, A., Nahavandi, S., Srinivasan, D.: Optimal uncertainty-guided neural network training. Appl. Soft Comput. **99**, 106878 (2021)

Kannan, S., Gurusamy, V., Preprocessing Techniques for Text Mining. Podi, RTRICS. (2014)

Khanian, M., Mohd, N.: A systematic literature review on the state of research and practice of collaborative filtering technique and implicit feedback. Artif. Intell. Rev. **45**(2), 167–201 (2016)

Kumar, B., Ravi, V.: A survey of the applications of text mining in financial domain. Knowl. Based Syst. **114**, 128–147 (2016)

Kunaver, M., Požrl, T.: Diversity in recommender systems—a survey. Knowl. Based Syst. **123**, 154–162 (2017)

Li, X., Wang, M., Liang, T.-P.: A multi-theoretical kernel-based approach to social network-based recommendation. Decis. Support Syst. **65**, 95–104 (2014)

Lika, B., Kolomvatsos, K., Hadjiefthymiades, S.: Facing the cold start problem in recommender systems. Expert Syst. Appl. **41**(4), 2065–2073 (2014)

Liu, J., Dolan, P., Pedersen, E.R.: Personalized news recommendation based on click behavior. In Proc. the 14th International Conference on Intelligent User Interfaces, Hong Kong, China, Feb. 7–10, (2010) pp.31–40

Liu, Y., Wang, W., Ghadimi, N.: Electricity load forecasting by an improved forecast engine for building level consumers. Energy **139**, 18–30 (2017)

Lu, J., Wu, D., Mao, M., Wang, W., Zhang, G.: Recommender system application developments: a survey. Decis. Support Syst. **74**, 12–32 (2015)

Mahmoudi, M.R., Behboodian, J., Maleki, M.: Large sample inference about the ratio of means in two independent populations. J. Stat. Theory Appl. **16**(3), 366–374 (2017)

Martinez, L., Perez, L. G., Barranco, M. J.: Incomplete preference relations to smooth out the cold-start in collaborative recommender systems. In *NAFIPS 2009–2009 Annual Meeting of the North American Fuzzy Information Processing Society* (2009) (pp. 1–6). IEEE

McNaught, C., Lam, P.: Using a wordle as a supplementary research tool. Qual. Rep. **15**(3), 630–643 (2010)

Mir, M., Kabir, H.D., Nasirzadeh, F., Khosravi, A.: Neural network-based interval forecasting of construction material prices. J. Build. Eng. **39**, 102288 (2021)

Pal, N.R., Biswas, J.: Cluster validation using graph theoretic concepts. Pattern Recogn. **30**(6), 847–857 (1997)

Pan, J.J., Mahmoudi, M.R., Baleanu, D., Maleki, M.: On comparing and classifying several independent linear and non-linear regression models with symmetric errors. Symmetry **11**(6), 820 (2019)

Paradarami, T.K., Bastian, N.D., Wightman, J.L.: A hybrid recommender system using artificial neural networks. Expert Syst. Appl. **83**, 300–313 (2017)

Pereira, A.L.V., Hruschka, E.R.: Simultaneous co-clustering and learning to address the cold start problem in recommender systems. Knowl. Based Syst. **82**, 11–19 (2015)

Piatetsky, G.: CRISP-DM, still the top methodology for analytics, data mining, or data science projects. KDD News. (2014)

Qu, Z., Song, X., Zheng, S., Wang, X., Song, X., & Li, Z.: Improved Bayes method based on TF-IDF feature and grade factor feature for chinese information classification. In 2018 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 677–680). IEEE. (2018)

Raeesi Vanani, I., Jalali, S.M.J.: Analytical evaluation of emerging scientific trends in business intelligence through the utilization of burst detection algorithm. Int. J. Bibliometr. Bus. Manag. **1**(1), 70–79 (2017)

Raeesi Vanani, I., Jalali, S.M.J.: A comparative analysis of emerging scientific themes in business analytics. Int. J. Bus. Inf. Syst. **29**(2), 183–205 (2018)

Salloum, S. A., Al-Emran, M., Monem, A.A., Shaalan, K.: Using text mining techniques for extracting information from research articles. In: Intelligent natural language processing: trends and applications (2018) (pp. 373–397). Springer, Cham

Shakeel, P.M., Baskar, S., Dhulipala, V.S., Jaber, M.M.: Cloud based framework for diagnosis of diabetes mellitus using K-means clustering. Health Inf. Sci. Syst. **6**(1), 15 (2018)

Sielis, G. A., Tzanavari, A., Papadopoulos, G.A. (2015). Recommender systems review of types, techniques, and applications. In: Encyclopedia of Information Science and Technology, Third Edition (pp. 7260–7270). IGI Global

Sledge, I.J., Bezdek, J.C., Havens, T.C., Keller, J.M.: Relational generalizations of cluster validity indices. IEEE Trans. Fuzzy Syst. **18**(4), 771–785 (2010)

Sohrabi, B., Raeesi Vanani, I., Baranizade Shineh, M.: Designing a predictive analytics solution for evaluating the scientific trends in information systems domain. Webology **14**(1), 32–52 (2017)

Sohrabi, B., Raeesi Vanani, I., Gooyavar, A., Naderi, N.: Predicting the readmission of heart failure patients through data analytics. J. Inf. Knowl. Manag. **18**(1), 1950012–190020 (2019)

Son, L.H.: HU-FCF: A hybrid user-based fuzzy collaborative filtering method in recommender systems. Expert Syst. Appl. **41**(15), 6861–6870 (2014)

Souza, E., Santos, D., Oliveira, G., Silva, A., Oliveira, A. L. (2018). Swarm optimization clustering methods for opinion mining. Nat. Comput., pp 1–29

Talafidaryani, M., Jalali, S.M.J., Moro, S. Digital transformation: toward new research themes and collaborations yet to be explored. Bus. Inf. Rev., 0266382120986035

Uitdenbogerd, A.L.: World cloud: a prototype data choralification of text documents. J. New Music Res. **48**(3), 253–263 (2019)

Volkovs, M., Yu, G., Poutanen, T.: Dropoutnet: Addressing cold start in recommender systems. Adv. Neural Inf. Process. Syst. (2017) (pp. 4957–4966)

Weng, L.T., Xu, Y., Li, Y., Nayak, R.: Exploiting item taxonomy for solving cold-start problem in recommendation making. In 2008 20th IEEE International Conference on Tools with Artificial Intelligence (2008) (Vol. 2, pp. 113–120). IEEE

Zarei, A.R., Shabani, A., Mahmoudi, M.R.: Comparison of the climate indices based on the relationship between yield loss of rain-fed winter wheat and changes of climate indices using GEE model. Sci. Total Environ. **661**, 711–722 (2019)

Zhou, M.: Micro behaviors: a new perspective in e-commerce recommender systems, Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, (2018) pp. 727–735

Zhou, K., Yang, S. H., Zha, H. Functional matrix factorizations for cold-start recommendation. In: Proceedings of the 34th international ACM SIGIR. (2011)

Authors and Affiliations

Iman Raeesi Vanani[1] · Laya Mahmoudi[1] · Seyed Mohammad Jafar Jalali[2] · Kim-Hung Pho[3]

Iman Raeesi Vanani
Imanraeesi@atu.ac.ir

Laya Mahmoudi
Laya.mahmoudii@gmail.com

Seyed Mohammad Jafar Jalali
sjalali@deakin.edu.au

[1] Faculty of Management and Accounting, Allameh Tabataba'i University, Tehran, Iran

[2] Institute for Intelligent Systems Research and Innovation (IISRI), Deakin University, Geelong, Australia

[3] Fractional Calculus, Optimization and Algebra Research Group, Faculty of Mathematics and Statistics, Ton Duc Thang University, Ho Chi Minh City, Vietnam