RESEARCH

Open Access

Improved content recommendation algorithm integrating semantic information

Ran Huang[1*]

*Correspondence:
hr_sdyu@163.com

[1] Shandong Youth University of Political Science, Jinan, Shandong, China

Abstract

Content-based recommendation technology is widely used in the field of e-commerce and education because of its intuitive and easy to explain advantages. However, due to the congenital defect of insufficient semantic analysis of TF-IDF vector space model, the traditional content-based recommendation technology has the problem of insufficient semantic analysis in item modeling, fails to consider the role of semantic information in knowledge expression and similarity calculation, and is not accurate enough in calculating item content similarity. The items with semantic relevance in content can not be well mined. The research goal of this paper is to improve the semantic analysis ability of the traditional content-based recommendation algorithm by integrating semantic information with TF-IDF vector space model for item modeling and similarity calculation and proposed an improved content recommendation algorithm integrating semantic information. In order to prove the effectiveness of the proposed method, several groups of experiments are carried out. The experiments results showed that the overall performance of the proposed algorithm in this paper is the best and relatively stable. This verified the validity of our method.

Keywords: Semantic information, TF-IDF, Content recommendation, Word vector

Introduction

With the continuous development of information technology, accelerating the construction of digital economy, digital society and digital government, and driving the transformation of production mode, lifestyle and governance mode as a whole with digital transformation have become the main theme of current social development. The application of recommendation algorithm is an important way to activate the potential of data elements and explore the value of data. The recommendation system generates a list of recommended items that users are most likely to be interested in through in-depth analysis and mining of user's interactive behavior data such as browsing, collecting, purchasing, and item attribute data. Content-based recommendation technology is widely used in e-commerce, education, entertainment and other fields because of its intuitive and easy-to-explain advantages.

Relevant research

The recommendation system attempts to mine users' interests and recommend possible items that users may be interested in according to users' historical behavior and real-time application scenarios. The recommendation technology has been widely used in information service platforms [1, 2] such as e-commerce [3], social networks and news websites [4]. In 2000, Mooney designed a book recommendation system named Libra by using the content-based recommendation method. In the system, the content information of books came from Amazon e-commerce website. The book model was constructed and the recommendation was generated with them. In 2004, Middleton designed an online academic research paper recommendation system named Quickstep. The system also adopted the technology of content-based recommendation, constructed the user interest model through the classification information of the papers viewed by the user, and calculated the similarity between the user interests characteristics and the paper characteristics to generate the paper recommendation. In 2009, Khribi et al. developed a recommendation system that established a user preference model based on learners' browsing history and generated resource recommendations combined with learning resource content. In 2017, Soulef Benhamdi et al. designed a personalized learning material recommendation system named NPR-EL by using collaborative filtering and content-based recommendation technology, which integrated recommendation technology into the learning environment to provide learners with personalized learning materials. Sunandanad et al. developed a movie recommendation system using enhanced content-based filtering algorithm based on user demographic data [5]. Bagul et al. designed a novel content-based recommendation approach based on LDA topic modeling for literature recommendation [6]. Tai et al. designed content-based recommendation using machine learning [7].

Through the study of relevant literature, we found that content-based recommendation technology is widely used in the field of education [8] because it is intuitive and easy to explain. The research on course recommendation in this paper is also based on the content-based recommendation method [9]. However, due to the congenital defect of insufficient semantic analysis of TF-IDF vector space model, the traditional content-based recommendation technology has the problem of insufficient semantic analysis in item modeling. It does not consider the role of semantic information in knowledge expression, and the calculation of items content similarity is not accurate enough. The items with semantic relevance in content can not be well mined. The research goal of this paper is to improve the traditional content-based recommendation algorithm through semantic information representation and proposed an improved content recommendation algorithm integrating semantic information.

Improved content-based recommendation algorithm integrating semantic information

Content-based recommendation algorithm model

The content-based recommendation algorithm attempts to recommend other items similar in content to the items that specific users are paying attention to or liked in the past. The key technology is building feature models for item and user [10]. The recommendation process is to match the item characteristics with the user interest characteristics, find the items most similar to the user interests and generate a recommendation list. The structure of content-based recommendation is shown in Fig. 1.

Content analyzer: Used for object feature modeling. The content-based recommendation algorithm mainly uses the descriptive information of items, which can be structured or unstructured data. Unstructured data is preprocessed to extract the structured information and convert the original information into a specific format. This step is usually completed by TF-IDF [11] dimensional space model.

Information learner: Used to collect user interaction information for user modeling. When the system explicitly requires users to evaluate the goods or services provided, such feedback information obtained is called explicit feedback. The system monitors and analyzes users' online behaviors to mine users' interest information. This kind of feedback is called implicit feedback, which usually does not require the participation of active users. The recommendation model designed in this paper models users' interests and generates recommendation lists according to the implicit feedback information of users' online course browsing behaviors.

Filtering component: Match the user interest information and item description information according to some similarity measurement algorithm, such as cosine similarity, Euclidean distance, Pearson correlation coefficient, etc. The matching results are mainly divided into two categories. One is binary correlation, such as like and dislike and the other is continuous correlation, which generates a recommendation list in order of degree of possible interest.

Semantic information representation

Word embedding technology can reflect the semantic information of words to a certain extent. The semantic distance between words can be calculated by word vectors.

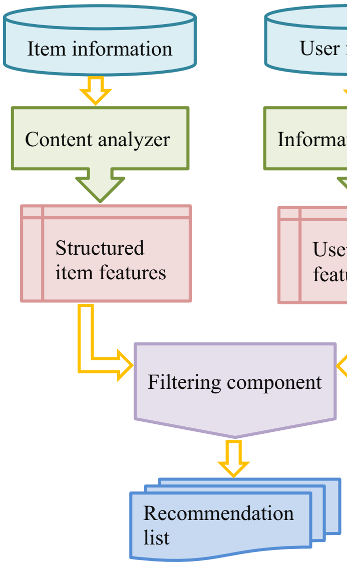

Fig. 1 Content-based recommendation algorithm model

At present, the commonly used word vectors are mainly based on word2vec and fasttext models.

Word2vec model uses deep learning method [12] to obtain the spatial distribution representation of words. It is a language model that learns low dimensional word vectors rich in semantic information from massive text corpus in an unsupervised way [13, 14]. It is the result of the application of neural network in the field of natural language processing. Word2vec word vector model maps words to low dimensional space, so that semantically similar words are close in the space. The semantic similarity between words is expressed by calculating the spatial distance between word vectors. Therefore, word2vec word vector has good semantic characteristics and can be used for natural language processing tasks such as text classification [15–18], emotion calculation, dictionary construction and so on.

Word2vec includes two training models that are continuous word bag model CBOW and Skip-gram model. CBOW model predicts the current word w_t based on the 2n known words in the context. The Skip-gram model predicts 2n words in the context according to the word w_t.

Taking Skip-gram model as an example, it is actually to maximize the conditional probability of the context with the window length of 2n around word w_t,as in (1). In order to simplify the calculation, we convert (1) into (2). That is find the minimum conditional probability of (2).

$$\mathrm{E} = P(w_{t-n}, \ldots, w_{t+n}|w_t) \tag{1}$$

$$\mathrm{E} = -\log P(w_{t-n}, \ldots, w_{t+n}|w_t) \tag{2}$$

And by the same token, CBOW model is actually to maximize the conditional probability of word w_t, as in (3). That is find the minimum conditional probability of (4).

$$\mathrm{E} = P(w_t|w_{t-n}, \ldots, w_{t+n}) \tag{3}$$

$$\mathrm{E} = -\log P(w_t|w_{t-n}, \ldots, w_{t+n}) \tag{4}$$

Fasttext is an optimization of word2vec [19], adding n-gram information on the basis of word2vec. Because Skip-gram model ignores the internal structure of word, fasttext split word into sub-word and calculates word vector through n-gram vector. Therefore, it can learn the expression of rare words and solve the problem of oov to some extent.

Improved content recommendation algorithm integrating semantic information

The content-based recommendation algorithm is based on a single TF-IDF vector space model in the process of item feature extraction, and the item similarity calculation is based on a single cosine similarity calculation method. TF-IDF vector space model is based on term-frequency statistics of corpus. This method assumes that feature items exist independently. Each dimension of space vector represents a feature item, and the weight of feature items in all dimensions together constitutes the feature vector representation of the item. Using this method, we can intuitively understand all the characteristics of the item, which has the advantages of easy interpretation and easy operation.

The disadvantage is that we fail to consider the impact of the semantic correlation between characteristic items on the calculation of item content similarity, and the resulting recommendation lists need to be improved in the recommendation accuracy. Word embedding technology has the advantage of representing the semantic information of words [20, 21]. The feature words with similar semantics have the closer distance in spatial distribution. Then the semantic correlation between feature words can be obtained by similarity calculation or distance measurement. In order to improve the semantic analysis defects and recommendation accuracy of content-based recommendation algorithm, we combined word embedding technology with TF-IDF vector space model for item modeling and similarity calculation, and proposed a new recommendation algorithm. The algorithm model is shown in Fig. 2.

Item modeling based on tf-idf and word embedding technology

The item modeling method based on TF-IDF and word embedding technology comprehensively utilizes the term-frequency information and semantic information between feature words, so as to more comprehensively represent the content information of the item from different perspectives.

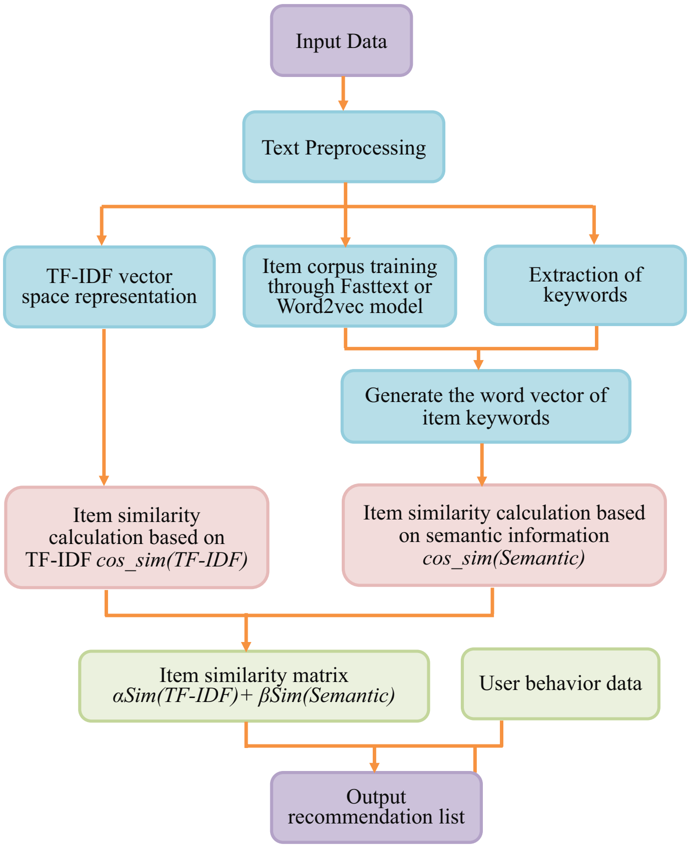
Input Data
Text Preprocessing
TF-IDF vector space representation
Item corpus training through Fasttext or Word2vec model
Extraction of keywords
Generate the word vector of item keywords
Item similarity calculation based on TF-IDF $cos\_sim(TF\text{-}IDF)$
Item similarity calculation based on semantic information $cos\_sim(Semantic)$
Item similarity matrix $\alpha Sim(TF\text{-}IDF)+\beta Sim(Semantic)$
User behavior data
Output recommendation list

Fig. 2 Improved content-based recommendation algorithm integrating semantic information

1. Item representation based on tf-idf weighting algorithm
   TF-IDF weighting algorithm is a common method to represent text in vector space [22], and it is also an item representation method commonly used in content-based recommendation algorithm. TF-IDF weighting algorithm makes statistics based on term frequency and comprehensively weighs the importance of feature words to an independent text and the whole corpus to give different weight values to each word [23]. It assumes that if a word appears many times in an article, represented by TF weight value, and appears simultaneously less in the whole corpus, represented by IDF weight value, it is likely to be related to the topic of the article. TF-IDF can also be understood as TF*IDF. Assuming that a corpus contains m texts, each text is represented by n feature items, and each feature item is represented by wi, then the TF-IDF calculation is as (5).

   $$TF - IDF(w_i) = \frac{c_{w_i}}{\sum_{i=1}^{n} c_{w_i}} * log\frac{m}{m(w_i)} \tag{5}$$

   C_{w_i} indicates the number of occurrences of the word w_i in the text. $\sum_{i=1}^{n} C_{w_i}$ indicates the total number of occurrences of all words in the text. $\frac{C_{w_i}}{\sum_{i=1}^{n} C_{w_i}}$ indicates the normalization of word weight value in independent text. *m(w*$_i$*)* represents the number of texts containing the feature item w_i.
2. Extraction of item text keywords
   The text data related to the item usually includes hundreds or even thousands of words, and each word reflects the item to different degrees. In order to better represent the semantic information of the item text data and improve the accuracy of the recommendation algorithm, it is necessary to extract the keywords from the relevant corpus of each item and mine the core vocabularies of the item [24]. On the one hand, TF-IDF weighting algorithm can be applied to generate text vector space representation to transform the processing of text into mathematical operations. On the other hand, it can extract item keywords according to the weight values of different feature words generated by TF-IDF weighting algorithm. The higher the TF-IDF weight value, the more feature words can reflect the key information of the item. Therefore, TF-IDF weighting algorithm can effectively mine item keywords. According to the word vector of the extracted item keywords, the item representation based on semantic information is generated.
3. Item representation based on tf-idf and word embedding technology
   Integrating the item keywords word vector into the item representation comprehensively utilizes both the term frequency in the corpus and the semantic information of item keywords, providing a more accurate reference for the calculation of item content similarity, so as to mine the correlation between different items, which is also an important innovation on the content-based recommendation algorithm in this paper. Generate the word vector of the extracted item keywords through text corpus training by word2vec or fasttext model and then sum the item keyword word vectors to obtain the semantic representation of the item based on word vector [25, 26]. The importance of each keyword to the item is different, which is reflected in the different TF-IDF weight values of keywords. Therefore, the importance of different key-

words should be considered when representing the item through word vector. The keyword word vector is summed according to TF-IDF weight to represent each item. The calculation formula of item representation based on TF-IDF word weight and keyword word vector is as (6). v_k indicates the word2vec or fasttext word vector of item keywords. tfidf(k,C) indicates the TF-IDF weight value of keyword k in corpus of item C.

$$vec(C) = \sum\nolimits_{k \in C} v_k \cdot tfidf(k, C) \tag{6}$$

Item similarity calculation based on tf-idf and semantic information

1. Item similarity calculation based on tf-idf

   The content-based recommendation algorithm models the item with the TF-IDF weighting algorithm, calculates the similarity between the items through the cosine similarity method [27, 28], and generates the potential item recommendation list according to the item similarity matrix. Cosine similarity measures the difference between items by calculating the cosine value of the angle between two vectors in the vector space. It mainly focuses on the difference in the direction of the vector rather than the absolute distance of the space. The calculation of cosine similarity is as (7).

$$Sim\left(t_j, t_k\right) = \cos\theta = \frac{\overrightarrow{v_{t_j}} \cdot \overrightarrow{v_{t_k}}}{\|\overrightarrow{v_{t_j}}\|\|\overrightarrow{v_{t_k}}\|} \tag{7}$$

2. Item similarity calculation based on tf-idf and semantic information

   The improved recommendation algorithm uses both TF-IDF and word embedding technology for item representation. Therefore, in the recommendation engine filtering component, the item similarity based on TF-IDF and the item semantic similarity based on word vectors should be calculated respectively. The total similarity between items is obtained by weighted summation. The optimal weight coefficient combination is adjusted and determined through the specific experimental process.

Experiment and results

In order to verify the effectiveness of the proposed method, we conducted experiments based on MOOC datasets. Considering the characteristics of content-based recommendation algorithm, multiple text data such as course title, course overview and course teaching objectives are used to conduct comparative experiments to observe the differences of data tendency of the same recommendation algorithm, so as to facilitate the horizontal comparison between the performances of the same recommendation algorithm on different data sets and the optimal performances of different recommendation algorithms, taking course title,course overview and course teaching objectives as data sets respectively, recorded as TF-IDF (course-title),TF-IDF (course-overview) and TF-IDF (teaching-objective). At the same time, we introduced multiple groups of comparative experiments to verify the effectiveness of semantic information for content recommendation.

Datasets

The data set used in the experiment was collected by the scrapy crawler framework from the icourses website. The courses involved include computer, economic management, psychology, foreign language, literature and history, arts, engineering, science, life science, philosophy, law, ideological and political education, pedagogy, innovation and entrepreneurship and many other categories. The crawled data about course mainly includes course title, course overview and course page links in Chinese text format. Before the experiment, adding up to 3000 online open courses.

Measurement

In order to prove the validity of the proposed model, this paper selected the prediction accuracy [29] as the index to evaluate the performance of the algorithm. The accuracy of prediction reflects the ability of recommendation system in recommendation accuracy, and also affects the reliability of recommendation system. Four possible recommendation results are shown in Table 1.

Calculation of prediction accuracy is as:

$$precision = \frac{tp}{tp + fp} \tag{8}$$

Experimental procedure

The experimental steps are as follows:

Step 1: Data preprocessing

The data preprocessing process is mainly to segment Chinese words and remove stop words. In the experiment, Jieba Chinese word segmentation tool is used to segment the text data of course. In order to reduce the dimension of course features and improve the accuracy of recommendation, the stop words are removed from the course text after word segmentation. This paper selected a comprehensive stop word list which integrated the stop word list of Harbin Institute of technology, Baidu and Sichuan University, including 2792 stop words.

Step 2: keyword extraction

After the preprocessing of Chinese text word segmentation and removal of stop words, we extracted keywords from the text data of course overview by TF-IDF weighting algorithm. Due to the limitations of the algorithm, there will be deviations in the keyword extraction results, so the number of keywords extracted will affect

Table 1 Classification of recommended results

	Recommended	Not recommended
Similar items	Number of positives (*tp*)	Number of false negatives (*fn*)
Dissimilar items	Number of false positives (*fp*)	Number of negatives (*tn*)

the course representation and subsequent course content similarity calculation to a certain extent. In the experiment, we chose to extract different numbers of keywords and observe the changes of the final experimental results, so as to mine the semantic information of the course as comprehensively and accurately as possible. After many experimental tests, we finally set the range at 10.

Step 3: word vector training
In the experiment, word2vec and fasttext model are used to train word vector. Word2vec includes two training models, namely continuous word bag model CBOW and Skip-gram model. At the same time, each model has two sets of frameworks, which are based on Hierarchical softmax and Negative sampling. Among them, negative sampling does not use complex Huffman tree. It can improve the training quality of word vector while improving the training speed. Word2vec and fasttext word vector dimension are both set at 300. The relevant training parameters are shown in Table 2.

Step 4: item modeling
Taking the course named how teachers make research as an example, the course corpus is calculated through the TF-IDF weighting algorithm to obtain the word frequency vector space representation based on statistics and the vector dimension is 18,809. In the meantime the course corpus is calculated through the word2vec to obtain the semantic vector space representation and the vector dimension is 300, greatly reducing the dimension.

Step 5: similarity calculation
Cosine similarity based on TF-IDF and semantic information are calculated respectively and the total similarity between courses are obtained by weighted summation, recorded as $\alpha\mathrm{Sim(TF\text{-}IDF)} + \beta\mathrm{Sim(semantic)}$. The optimal weight coefficient of α and β is set at about 0.7 and 0.3, at which the experimental resutls is the best. Then we sequence the courses in reverse order according to the cosine similarity, take the top n courses most similar to each course, query corresponding course ID and course title, generate and save the course similarity matrix. Because this calculation process can be completed offline, the system can directly recommend courses for online learners according to the similarity matrix.

To further illustrate the innovation of this paper, Tables 3, 4 and 5 respectively showed the cosine similarity results between some course items based on TF-IDF weight value, word2vec word vector and the algorithm in this paper. From the data, we can see that

Table 2 Word2vec training parameters

Training Corpus	Chinese Wikipedia
Corpus size	1.3G
Vector dimension	300
Word segmentation tool	jieba
Training tools	Word2Vec of Gensim
Training model	Skip-Gram with Negative Sampling
Training parameters	The dynamic window size is 5. The minimum word frequency is 10. The number of iterations is 5

Table 3 Cosine similarity between some course items based on TF-IDF

Item ID	1	2	3	4	5	6
1	1.0	0.088	0.134	0.115	0.036	0.156
2	0.088	1.0	0.106	0.085	0.085	0.121
3	0.134	0.106	1.0	0.079	0.034	0.172
4	0.115	0.085	0.079	1.0	0.028	0.124
5	0.036	0.073	0.034	0.028	1.0	0.037
6	0.156	0.121	0.172	0.124	0.037	1.0

Table 4 Cosine similarity between some course items based on word vector

Item ID	1	2	3	4	5	6
1	1.0	0.486	0.437	0.474	0.482	0.581
2	0.486	1.0	0.434	0.530	0.599	0.625
3	0.437	0.434	1.0	0.342	0.395	0.386
4	0.474	0.530	0.342	1.0	0.595	0.559
5	0.484	0.599	0.395	0.595	1.0	0.655
6	0.581	0.625	0.386	0.559	0.655	1.0

Table 5 Cosine similarity between some course items based on algorithm in this paper

Item ID	1	2	3	4	5	6
1	1.0	0.199	0.219	0.215	0.161	0.275
2	0.199	1.0	0.197	0.209	0.219	0.262
3	0.219	0.197	1.0	0.153	0.135	0.232
4	0.215	0.209	0.153	1.0	0.187	0.246
5	0.161	0.219	0.135	0.187	1.0	0.210
6	0.275	0.262	0.232	0.246	0.210	1.0

the word vector representation is a useful complement to the semantic aspect of calculating the similarity between items.

Results

The comparison of recommendation accuracy of classic content-based recommendation algorithm on different data sets are shown in Fig. 3.The performance comparison of different recommendation algorithms are shown in Table 6 and Fig. 4. We can see from the formula of prediction accuracy that recommendation accuracy is related to the length of the recommendation list and the number of courses similar to the course browsed by the learner in the recommendation list. In the experiment, we set the length range of the list to 10, which means recommending at least 1 similar course and at most 10 similar courses. Then, the statistics analysis about the average recommendation accuracy of each recommendation model and the change with the length of the recommendation list was performed.

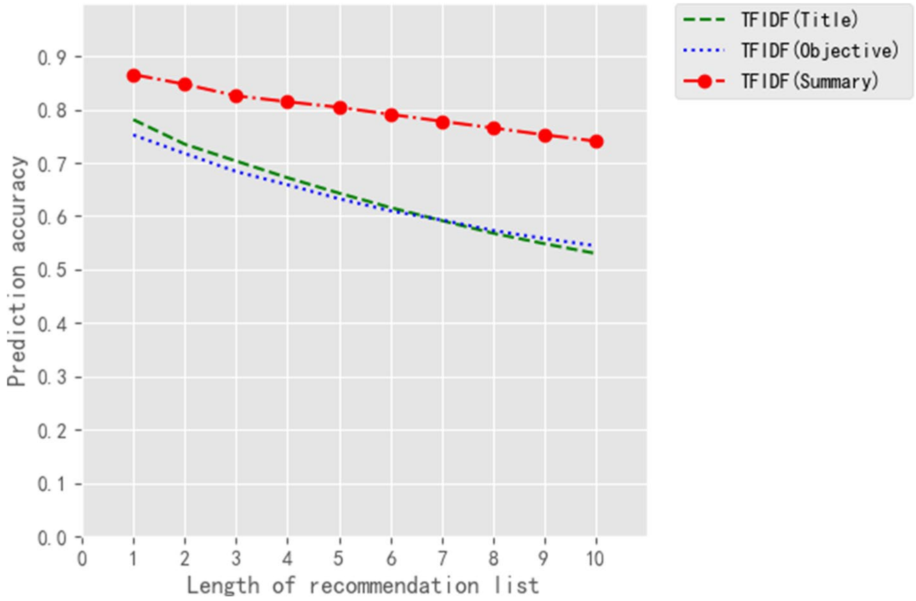

Fig. 3 Performance comparison of content-based recommendation algorithm on different data sets

Table 6 Comparison of prediction accuracy of different methods

The length of recommendation list	TF-IDF	word2vec	Lstm	Fasttext	TF-IDF + word2vec	TF-IDF + Textfast
1	0.866	0.827	0.751	0.829	0.893	0.894
2	0.847	0.798	0.732	0.801	0.879	0.883
3	0.826	0.775	0.714	0.778	0.870	0.875
4	0.815	0.754	0.701	0.757	0.854	0.861
5	0.804	0.733	0.692	0.739	0.840	0.845
6	0.791	0.716	0.681	0.720	0.829	0.832
7	0.778	0.706	0.673	0.708	0.816	0.820
8	0.766	0.695	0.663	0.697	0.803	0.811
9	0.753	0.689	0.655	0.690	0.792	0.797
10	0.741	0.677	0.647	0.679	0.780	0.789

Discussion and conclusion

Data tendency comparison

1. Data tendency of content-based recommendation algorithm

   Three classic content-based recommendation algorithms experiments were carried out, taking course title, course overview and course teaching objectives as data sets respectively. The experimental results are shown in Fig. 3. We observed that under the same length of recommendation list, taking course overview as course content for classic content-based course recommendation, the recommendation accuracy is significantly higher than the experimental results obtained by using course title and teaching objectives as course content respectively. This is because the course overview contains richer and more detailed vocabularies related to the course content,

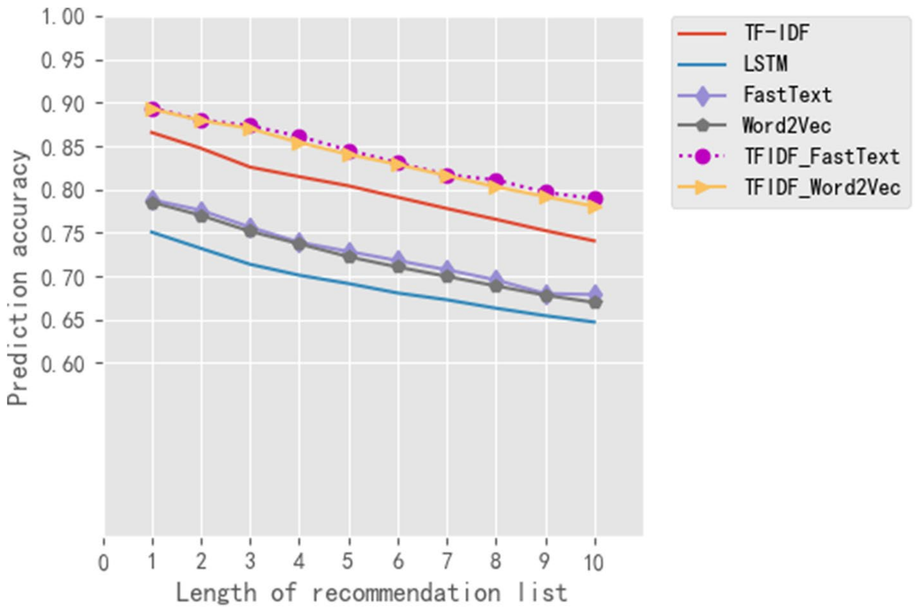

Fig. 4 Performance comparison of different recommendation algorithms

which can better summarize the main contents of the course. Compared with the course overview, the course title and teaching objectives contain less vocabularies, especially the course title. Then, after the course is transformed into vector space through the method of TF-IDF, the probability of co-occurrence words between different courses will be reduced, thus reducing the similarity and recommendation accuracy between courses. The experimental results also proved this.

The analysis showed that when using classic content-based recommendation algorithm, the richer and more comprehensive the text vocabularies, the higher the accuracy of the recommendation results, which can be understood as the requirement for the amount of vocabularies to some extent.

2. Data tendency of recommendation with item modeling through word embedding technology
   In order to fully mine semantic information and obtain the best experimental results, we also carried out comparative experiments on the recommendation algorithm with item modeling only throuth word embedding, taking course title, course overview as data sets respectively. Taking word2vec as an example, the recommendation accuracy of content recommendation algorithm with item modeling throuth word2vec word vector is shown in Fig. 3. We observed that the accuracy of recommendation based on word2vec by course title is slightly better than that based on course overview. This result is contrary to the result of classic content-based recommendation algorithm. This may be because the course title is the most refined expression of the semantics of the course content, and the course overview text can not reflect the semantics of the course content more prominently because it contains redundant words.

The analysis shows that when using word vector for course modeling and recommendation, the stronger the generality of the text vocabularies introducing the course, the higher the accuracy of the recommendation, which can be understood as the requirement for the quality of vocabularies to some extent.

Performance comparison of different recommendation algorithms

In the experiment, we have implemented the classic content-based recommendation algorithm and improved content recommendation algorithm integrating semantic information through Python programming language. At the same time, we have carried out comparative experiments on the dataset with the FastText, Word2Vec and LSTM recommendation models. The experimental results are shown in Table 6 and Fig. 4.

The data shows that the recommendation accuracy of the improved content recommendation algorithm integrating semantic information is the highest. Compared with the classical content-based recommendation algorithm, the recommendation accuracy of content recommendation algorithm with item modeling through word embedding technology rises by 2.7%. Compared with the recommendation algorithm with item modeling based only on the word2vec or fasttext word embedding, it rises by 7.6%. Compared with LSTM recommendation algorithm, it rises by 7.6%. It can be seen that it is effective to improve the classic content-based recommendation algorithm by integrating semantic information into item modeling through word embedding technology. On the other hand, we also see that the recommendation accuracy of the recommendation algorithm that only uses word2vec or fasttext word embedding technology to model items is not higher than that of the classic content-based recommendation algorithm model. This is because the content representation of items through TF-IDF is still a very effective method, but the item modeling and similarity calculation by integrating semantic information take into account word frequency information and semantic information, make the content representation and calculation evaluation of the items more comprehensive and scientific, the scope of the items to be recommended more inclusive, and the recommendation accuracy is higher. The experimental results also prove the rationality of the improvement.

Case analysis of course recommendation

This section mainly analyze the course instance sequence generated by the content-based recommendation algorithm and improved content recommendation algorithm integrating semantic information. Taking the courses Appreciation of hundreds of ancient poems and Pediatrics as examples, two different recommendation algorithms generated different course recommendation lists, as shown in Tables 7 and 8.

By analyzing the course recommendation lists in Tables 7 and 8 from the perspective of artificial semantic understanding, it can be seen that the course recommendation list generated by improved content recommendation algorithm integrating semantic information is significantly better than content-based recommendation algorithm in course content relevance and course ranking. As shown in Table 7, taking the course appreciation of hundreds of ancient poems as an example, the recommended courses generated by improved content recommendation algorithm

Table 7 Recommendation lists of Appreciation of hundreds of ancient poems

Recommendation lists	Content-based recommendation algorithm	Improved content recommendation algorithm integrating semantic information
1	Modern reading of ancient poetry	Modern reading of ancient poetry
2	Theoretical and practical research of socialism with Chinese characteristics	Tang poetry, Song Ci and traditional culture
3	Tang poetry, Song Ci and traditional culture	Reading and recitation of Chinese Classical Poetry
4	Chinese culture English speaking	Research on verve Poetry
5	The world of wisdom and practice	Urban writing of classical literature
6	Chinese traditional culture	Chinese poetry art
7	Outline of Chinese modern history	Sixteen lectures on Chinese language and culture
8	Research on Chinese curriculum standards and teaching materials	Research on Chinese curriculum standards and teaching materials

Table 8 Recommendation lists of Pediatrics

Recommendation lists	Content-based recommendation algorithm	Improved content recommendation algorithm integrating semantic information
1	Pediatrics of traditional Chinese Medicine	Pediatrics of traditional Chinese Medicine
2	Preventive medicine	Neonatology
3	Medicopsychology	Pediatric nursing
4	Engineering mechanics	Surgical skill teaching
5	College Physics	Medicopsychology
6	Neonatology	Medical imaging
7	History of Chinese Philosophy	Chinese internal medicine
8	Green rehabilitation	Materials science research methods

integrating semantic information are almost all related to ancient poetry or literature, and the ranking of recommended courses is relatively more reasonable. The recommended courses based on content-based recommendation algorithm include courses such as Theoretical and practical research of socialism with Chinese characteristics and Outline of Chinese modern history, which deviate greatly from the contents of the courses visited, and the less relevant course, namely theoretical and practical research of socialism with Chinese characteristics, ranks second in the recommended list. Similarly, taking the course Pediatrics as an example, the course recommendation list generated by the improved method proposed in this paper, from pediatrics to medicine, is related to the course content, while the recommended courses of the content-based recommendation method include courses that deviate greatly from the target courses, such as Engineering Mechanics, College Physics, history of Chinese philosophy, etc. It is also unreasonable in course ranking. For example, it is obviously somewhat abrupt and illogical for the two recommended courses Engineering Mechanics and College Physics to be arranged before the course neonatal science.

Conclusion

Considering that the classic content-based recommendation algorithm has insufficient semantic analysis in terms of item representation and similarity calculation, this paper proposed an improved content recommendation algorithm integrating semantic information. The embedded representation containing semantic information is generated through the word2vec model or the fasttext model, and is combined with TF_IDF to generate the item representation that integrates statistical information and semantic information, and then calculates the item similarity based on the fusion information to generate the item recommendation list, so as to improve the accuracy and inclusiveness of the recommendation algorithm. Finally, a number of comparative experiments on the data set verified the effectiveness of the improved recommendation algorithm.

Acknowledgements
Not applicable.

Author contributions
RH designed and implemented the proposed model, carried out the experiments and wrote the manuscript. The author read and approved the final manuscript.

Funding
Not applicable. This research received no specific grant from any funding agency.

Availability of data and materials
The datasets analysed during the current study are not publicly available but are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
No competing of interests.

Received: 7 May 2022 Accepted: 17 May 2023
Published online: 28 May 2023

References

1. Chang JW, Chiou CY, Liao JY. Music recommender using deep embedding-based features and behavior-based reinforcement learning. Multimedia Tools Appl. 2019;1–28.
2. Wu S, Sun F, Zhang W, et al. Graph neural networks in recommender systems: a survey[J]. 2020.
3. Felfernig A, Friedrich G, Dietmar J. An integrated environment for the development of knowledge-based recommender applications. Int J Electron Commer. 2006;11(2):11–34.
4. Girsang AS. Recommendation System Journalist For Getting Top News Based On Twitter Data. Paper presented at the 2019 International Conference Of Science and Information Technology in Smart Administration, Balikpapan, Indonesia. 2019; 16-17.
5. Sunandana G, Reshma M, Pratyusha Y, et al. Movie recommendation system using enhanced content-based filtering algorithm based on user demographic data[C]//2021 6th International Conference on Communication and Electronics Systems (ICCES). Coimbatore: IEEE Press, 2021;1–5.
6. Bagul DV, Barve S. A novel content-based recommendation approach based on LDA topic modeling for literature recommendation[C]//2021 6th International Conference on Inventive Computation Technologies (ICICT). Coimbatore: IEEE Press, 2021; 954–961.
7. Tai Y, Sun Z, Yao Z. Content—Based Recommendation Using Machine Learning[C]//2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP). Gold Coast: IEEE Press, 2021;1–4.
8. Dascalu MI, Bodea CN, Mihailescu MN. Educational recommender systems and their application in lifelong learning. Behav Inf Technol. 2016;35(4):290–7.
9. Rahimpour CB, Hamid H, Hoda M. User trends modeling for a content-based recommender system. Expert Syst Appl. 2017;87:209–19.
10. Ran H, Ran L. Research on content-based MOOC recommender model. Paper presented at 2018 5th International Conference on Systems and Informatics (ICSAI). 2018;10–12.

11. Rani U, Bidhan K. Comparative assessment of extractive summarization: TextRank, TF-IDF and LDA. J Sci Res. 2021;65(01):304–11.
12. Li S, Gong B. Word embedding and text classification based on deep learning methods. Paper presented at CSCNS2020.2020;22–23.
13. Mikolov T. Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst. 2013;26:3111–9.
14. Mikolov T. Efficient estimation of word representations in vector space. Comput Sci. 2013;25(05):213–9.
15. Wang J, Luo L, Wang DQ. Research on Chinese short text classification based on Word2Vec. Comput Syst Appl. 2018;27(5):209–15.
16. Sifeng J. Correlation analysis and text classification of chemical accident cases based on word embedding. Process Saf Environ Prot. 2022;158:698–710.
17. Lei C, Jun L. Research on text feature selection method based on word vector. J Chin Comput Syst. 2018;39(5):991–4.
18. Yuankun C, Yan J, Guang C. Research on website topic classification based on word2vec. Comput Digital Eng. 2019;47(01):169–73.
19. Joulin A, et al. Bag of tricks for efficient text classification. CoRR, 2016, abs/1607.01759.
20. Tang M, Zhu L, Zou XC. A document vector representation based on word2vec. Comput Sci. 2016;43(6):214–7.
21. Xiao L, Hui X, Lijie L. Research on sentence semantic similarity calculation based on word2vec. Comput Sci. 2017;44(9):256–60.
22. Qaiser S, Ali R. Text mining: use of TF-IDF to examine the relevance of words to documents. Int J Comput Appl. 2018;181(01):25–9.
23. Xitao L, Lei G. Research on Chinese word segmentation and part of speech tagging. Comput Technol Dev. 2015;25(2):175–80.
24. Ping N, Degen H. Research on Chinese keyword automatic extraction based on TF-IDF and rules. J Chin Comput Syst. 2016;37(4):711–5.
25. Lei C, Jun L. Text feature selection method based on word vector. J Chin Comput Syst. 2018;39(5):991–4.
26. Jie C, Cai C, Yi L. Document classification method based on word2vec. Comput Syst Appl. 2017;11:159–64.
27. Liu WC. A review of text similarity approaches. Inf Sci. 2019;3:158–68.
28. Erjing C, Enbo J. Review of text similarity calculation methods. New Technol Libr Inf Service. 2017;1(6):1–11.
29. Kaminskas M, Bridge D. Diversity, serendipity, novelty, and coverage. A survey and empirical analysis of beyond-accuracy objectives in recommender systems. ACM Trans Interact Intel Syst. 2016;7(1):1–42.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.