Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 55 (2015) 843 – 851

Procedia Computer Science

Information Technology and Quantitative Management (ITQM 2015)

Personalized Financial News Recommendation Algorithm Based on Ontology

Rui Ren[a,b,c], Lingling Zhang[a,b,c]*, Limeng Cui[b,c,d], Bo Deng[d,e], Yong Shi[b,c]

[a]*School of Management, University of Chinese Academy of Sciences, Beijing100190,China*
[b]*Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing100190, China*
[c]*Research Centre on Fictitious Economy and Data Science, Chinese Academy of Sciences, Beijing100190, China*
[d]*School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing100190,China*
[e]*Institute of Computing Technology, Chinese Academy of Sciences, Beijing100190, China*

Abstract

To deal with the challenge of information overload, in this paper, we propose a financial news recommendation algorithm which help users find the articles that are interesting to read. To settle the ambiguity problem, a new presented OF-IDF method is employed to represent the unstructured text data in the form of key concepts, synonyms and synsets which are all stored in the domain ontology. For users, the recommendation algorithm build the profiles based on their behaviors to detect the genuine interests and predict current interests automatically and in real time by applying the thinking of relevance feedback. Finally, the experiment conducted on a financial news dataset demonstrates that the proposed algorithm significantly outperforms the performance of a traditional recommender.

Peer-review under responsibility of the Organizing Committee of ITQM 2015

Keyword: news recommendation algorithm; ontology; relevence feedback; OF-IDF

1. Introduction

With the rapid development of World Wide Web, people tend to read news on line instead of reading via physical newspaper subscription. Meanwhile, finance plays a more and more crucial role in the world, people are more curious about financial events and knowledge nowadays than ever before. The critical problem people are confronted with is the information overload which means the volumes of financial news are overwhelming to the users. Accordingly, in this paper, we propose a financial news recommendation approach which aims to deliver the interesting news articles to users not only to help individuals save time and energy but also help

* Corresponding author. Tel.:+86-10-82680676; fax: +86-10-82680698.
E-mail address: rr2010jolin@163.com, zhangll@ucas.ac.cn.

Peer-review under responsibility of the Organizing Committee of ITQM 2015
doi:10.1016/j.procs.2015.07.151

enterprises improve user loyalty.

There are three different ways of news recommendation so far: a content-based recommendation, a collaborative filtering recommendation, and a hybrid recommendation. Content-based methods plays a central part in recommender systems, as it is able to recommend information that has not been rated before and accommodates the individual differences between users [1,2]. So in this paper, we focus on the content-based recommendation. One of the disadvantage of it is the lack of semantics [3]. To deal with the issue, we use an ontology to store lexicalized financial domain concepts, relations and synsets. An accurate user profile is a critical part of content-based recommendation approach. We employ the thinking of relevance feedback to discover users' genuine interests automatically rather than create and update profiles manually.

The contribution of this paper is three-fold. Firstly, we construct a specific financial domain ontology which try to store all the paramount information about the certain amount of financial news. Secondly, this paper proposes a new method called OF-IDF to represent the unstructured text data. Thirdly, we construct user profile by applying the thinking of relevance feedback to detect the users' genuine interests automatically.

The rest of paper is organized as follows. Section 2 reviews some relevant literatures about the news recommendation and ontology. Section 3 introduces the way to construct financial domain ontology. Section 4 proposes the method OF-IDF to represent the text data. Section 5 elaborates on the personalized financial news recommendation algorithm. Section 6 presents the experimental results and its evaluation. Section 7 concludes the paper and points out our future work.

2. Related Work

Three different technologies are commonly used in news recommender systems: a collaborative filtering recommendation, a content-based recommendation and a hybrid recommendation.

Collaborative filtering recommendation focuses on user similarity or item similarity. It has been applied to personalized news reading applications, such as GroupLens which is the first system for collaborative filtering of netnews, to help people find articles they may like in the huge stream of available articles[4] and the first version of Google News recommender [5]. Although it can detect the potential interest of the users, the approach does not consider the content of items and is arduous to solve the problem of scalability, sparsity and cold-start. Content-based recommendation concentrates on user profiles which are built by analyzing the content of items that the user accessed and favored in the past. It excels in tackling text data, can recommend unpopular products and list the characteristics of the products so as to explain the reason. The hybrid method is a mix of the two previously mentioned recommendation systems and tries to combine the best of both worlds [6], however, it cannot resolve the difficulty from the root.

In this paper, we limit the discussion to content-based recommendation in that it doesn't need to collect the opinions of the peer users, can fully understand the content of the news and generate recommendations with good explanation.

Term frequency-inverse document frequency (TF-IDF) is a well-known and the most common content-based recommendation method which is put forward by Salton et al in 1975 [7] and often used in information retrieval and text mining. It is a statistical measure used to evaluate how important a word is to a document in a document collection or corpus. The importance increases proportionally to the number of times that a word appears in the document but is offset by the frequency of the word in the corpus. Combined with the vector space model and similarity calculation, TF-IDF can be applied to recommend news items to a specific user [8]. TF-IDF is term-based which need calculate the TF-IDF weights of all the terms of a text, as a result, it is not suitable to process big data. Furthermore, the term-based method may cause linguistic ambiguity.

In order to settle the ambiguity problem, in this paper, we try to employ ontology to optimize TF-IDF method. Ontology not only can present key concepts and relationship but process the content of the information as well [9]. The term "ontology" comes from the field of philosophy that is concerned with the study of being

or existence. In computer and information science, ontology is a technical term denoting an artifact that is designed for a purpose, which is to enable the modeling of knowledge about some domain, real or imagined[10]. According to Thomas R. Gruber (1993), an ontology is an explicit specification of a conceptualization [11]. Borst (1997) believed ontologies are formal descriptions of shared knowledge in a domain [12]. Rudi Studer et.al (1998) defined an ontology as a formal, explicit specification of a shared conceptualization [13], which demonstrates that four factors are essential to an ontology: conceptualization, explicit, formal and shared. The reason for ontologies being so popular is in large part due to what they promise: a shared and common understanding of some domain that can be communicated across people and computers [13]. In natural-language applications, ontologies can be used for natural-language processing [14], or automatically or semi-automatically extract knowledge from texts. Wordnet is one of the largest lexical ontologies in English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations [15].

3. Financial Domain Ontology

In this paper, based on 1823 financial news which comes from the news database of 2010 proposed by Hogenboom F, Vandic D, Frasincar F, et al [3], we construct a specific financial domain ontology which try to store all the paramount information about the financial news. The ontology is a shared and common understanding of some domain that can be communicated across people and computers [10]. The domain ontology includes the key concepts that capture the semantic context of the articles as well as the relations between them such as hasCompetitor and isCEOof. For example, if a user is interested in the news of Google, it is likely that he is fond of the information of Yahoo. Moreover, for every concept in the ontology, a synonym property is defined to eliminate the ambiguity with lexical representations of a concept. A WordNet sense property is also defined to retrieve synsets of concepts. The ontology is stored in OWL format which is well supported in multiple software environments. Recommenders that focus on the ontology might produce faster and more accurate recommendations than the term-based recommenders, since they don't need to consider all words. And unlike words, concepts are not ambiguous. Besides, the relationship between the concepts may detect the potential interest of the users to meet their demand from multiple dimensions.

4. The Structured Representation of Financial News

The financial news article to be recommended is represented as $d_j (1 \le j \le N)$, N is the total number of the articles. The concept in the ontology is represented as k_i （$1 \le i \le M$）, M is the total amount of the concepts in the given article to be recommended. $w_{i,j}$ is the weight of *ith* keyword in the *jth* article. $f_{i,j}$ is the number of k_i in the article d_j. Consequently, the ontology frequency $OF_{i,j}$ of k_i in the article d_j is calculated as follows, m is the amount of the concepts in the article d_j in the total

$$OF_{i,j} = \frac{f_{i,j}}{\max_i f_{i,j}} \quad （1 \le i \le m ）$$

Ontology frequency measures how frequently a concept occurs in a article, while inverse article frequency measures how important a concept is by weighing down the frequent ones while scaling up the rare ones. The inverse article frequency of concept k_i is IDF, n_i shows how many times k_i occurs in the articles

$$IDF_i = \log \frac{N}{n_i}$$

OF-IDF of the k_i is:

$$w_{i,j} = OF_{i,j} \times IDF_i$$

Thus, the article d_j is represented as the vector containing all OF-IDF values for the different words from an unread news document

$$content(d_j) = (w_{1j,} w_{2j,} \cdots, w_{mj})$$

5. Personalized Financial News Recommendation Algorithm

$$\begin{array}{c} \\ \mathrm{I} \\ \mathrm{II} \\ \mathrm{III} \\ \vdots \\ M \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & \cdots & N \end{array} \\ \begin{bmatrix} & a & & b & & & c \\ & & d & e & & & \\ f & g & & & h & & k \\ & & & & & & \\ & p & & & q & & r \end{bmatrix} \end{array}$$

Fig. 1. User-Concept Matrix

$$\begin{array}{c} \\ \mathrm{I} \\ \mathrm{II} \\ \mathrm{III} \\ \vdots \\ M \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & \cdots & N \end{array} \\ \begin{bmatrix} & a+\alpha & & b+\alpha & & & c+\alpha \\ & & d+\beta & e+\beta & & & \\ f+\gamma & g+\gamma & & & h+\gamma & & k+\gamma \\ & & & & & & \\ & p & & & q & & r \end{bmatrix} \end{array}$$

Fig. 2. Modified User-Concept Matrix

Assume User *i* have read *M* articles, and the articles have already been represented as the ontology weight vector by the method in Sector4, *N* is the total number of concepts appeared in the *M* articles, a,b,c,⋯,r stands for the OF-IDF weight. Consequently, the user profile is intuitively represented as User-concept matrix shown in Fig.1.

Assume User *i* clicked Article Ⅰ, User *i* not only read Article Ⅱ but also liked it, User *i* clicked ArticleⅢ but didn't like it, under this condition, we use the thinking of relevance feedback to modify the interest profile. The modified User-concept matrix is demonstrated in Fig.2, α, β, γ are the modified parameters, obviously, α, β are positive while γ is negative. The modified user profile is calculated as the equation below which use the thinking of relevance feedback, where S^{α} shows that User *i* clicked m articles, S^{β} shows that User *i* read n articles and also liked them, S^{γ} shows that User *i* clicked *l* articles but didn't like them, S^{α}, S^{β}, S^{γ} are all vectors. Thus, the user profile of is represented as the vector $profile_i$ containing all OF-IDF values for the different words from the user profile's news document.

$$profile_i = \sum_m (\alpha + S^{\alpha}) + \sum_n (\beta + S^{\beta}) + \sum_l (\gamma + S^{\gamma})$$

The final step is to compute the similarity between the news article $content(d_j)$ and the user profile $profile_i$, and the similarity is described as $u(d_j, p_i)$, where $\|\cdot\|$ is the norm of the vector.

$$u(d_j, p_i) = sim(content(d_j), proflie_i) = \frac{content(d_j) \square proflie_i}{\|content(d_j)\| \square \|proflie_i\|}$$

All of the results are sorted in a descending order, and those news document which have a similarity value higher than the cut-off value are recommended to the user.

The procedures to implement the algorithm is manifested in the flow chart.

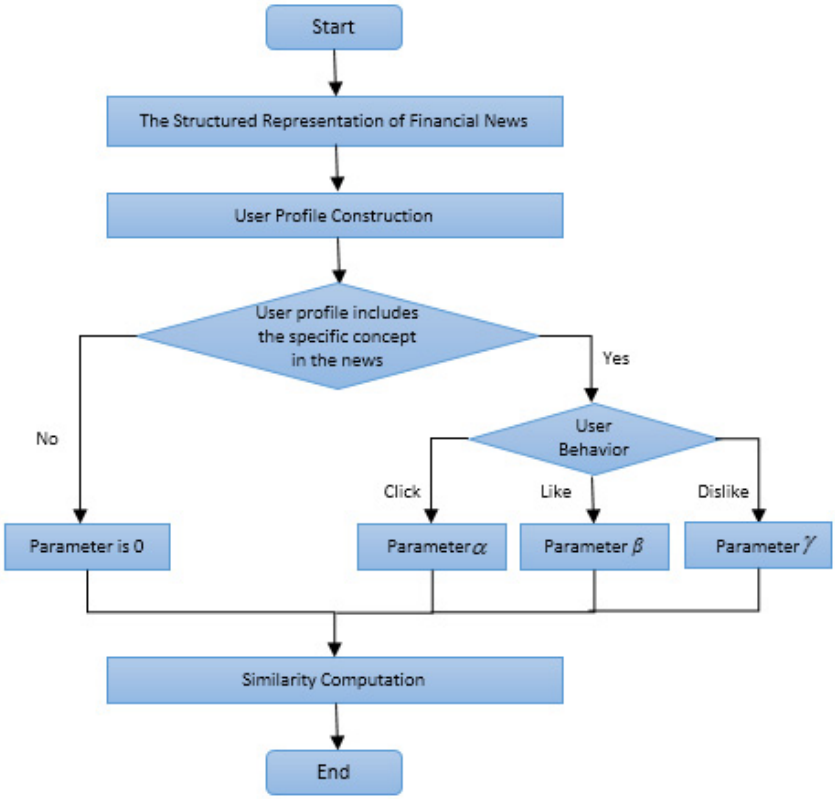

Fig. 3. Flow Chart of the Algorithm

6. Experiment

6.1. Experimental Design

In this section, we conduct experimental studies of the algorithm we proposed. The news data comes from the news database of 2010 proposed by Hogenboom F, Vandic D, Frasincar F, et al [3], including 1823 financial news which covered international financial news in 2010 from different aspects.

We built a temporary website to collect users' browsing history. For each article, the user can choose to indicate whether it is interesting or not. We've selected 33 registered users' browsing data which contains 3600 records.

After removing stop words, stemming and calculating OF-IDF, every news item was vectorized. The processing of the data set is based on supervised learning. Four-fifths data of each user's browsing history is treated as training set, while one-fifth data is regarded as test set. The training set is used to create a user profile and recommend interesting news items to the users by implementing the algorithm we've proposed above and the baseline method which is the classic TF-IDF weighting and the cosine similarity. The test set is used to evaluate the algorithms by comparing the similarities between the recommendation results and the test set which contains the true users' behaviors.

6.2. Evaluating Indicators

Precision and recall are common indexes in the domain of information retrieval. And F1 Score considers both the precision and the recall.

Precision: the number of web pages recommended correctly divide by the total pages recommended.

Recall: the number of web pages recommended correctly divide by the number of pages that users visited.

F1 Score: the harmonic mean of precision and recall.

Runtime: the total time cost of running the algorithm.

6.3. Experimental Implementation

We applied 5-fold cross validation method to deal with every user's browsing data, and randomly divided it into 5 parts, 4 parts of which were treated as training data with the rest part as test data. And then we executed recommendation algorithm under the same parameter setting. In addition, precision and recall was calculated according to the recommendation results of certain algorithm. Finally, we computed the average of precision and recall in five different training set.

In this experiment, we compared our method with the baseline method. According to the foregoing experimental results, the algorithm we proposed outperforms the baseline method at precision, recall and F1 Score. It indicates that the financial ontology can characterize user's reading habit more precisely. Furthermore, our proposed method can capture user's genuine interest better. We summarized the recommendation results of 33 users and made comparison of two algorithms in the following.

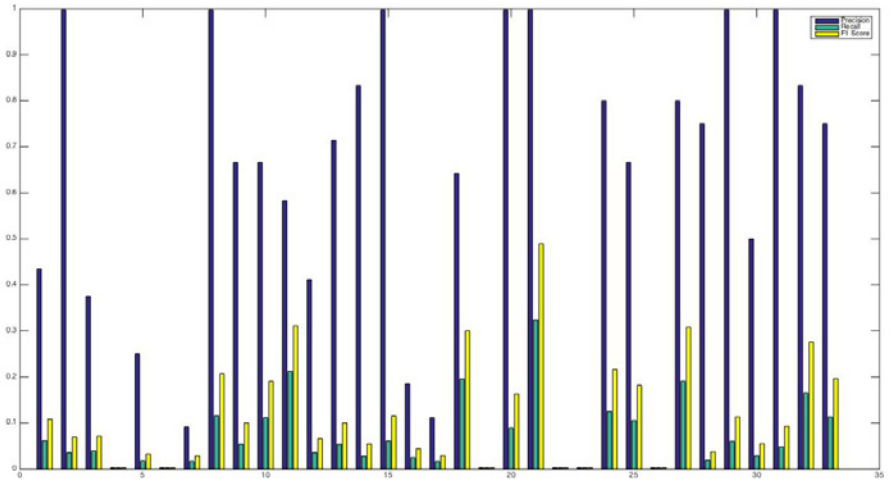

Fig. 4. Precision, Recall and F1 Score of the baseline method

Fig.4. shows the precision, recall and F1 Score of all users using the baseline method. From the figure we can see that for some users, the precision is pretty high. But for some users, the precision, recall and F1 Score are all very low. Subsequently, we can conclude that the baseline method cannot infer user's interest and generate new concept that user may be interested.

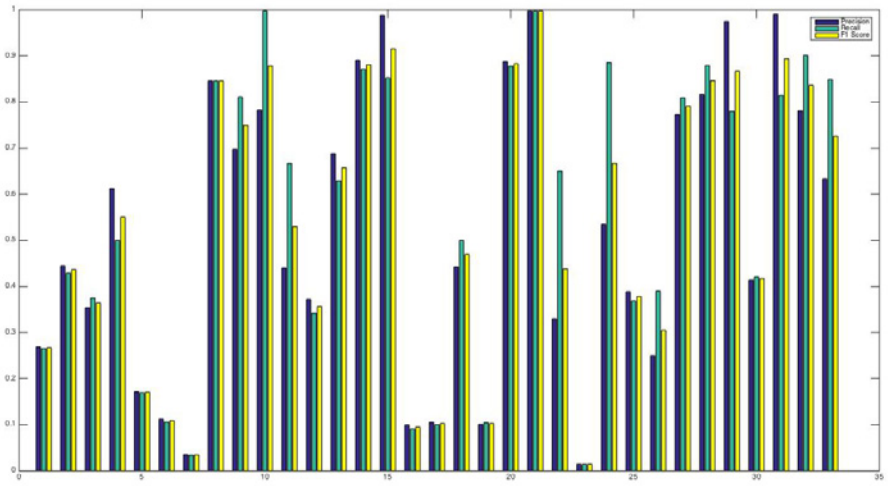

Fig. 5. Precision, Recall and F1 Score of proposed algorithm

Fig.5. demonstrates the precision, recall and F1 Score of all users using our method. Notice that the precision under some users' data has been declined, we may deduce that although the ontology gives user a broad view of different portal of news, it may cause the loss of precision.

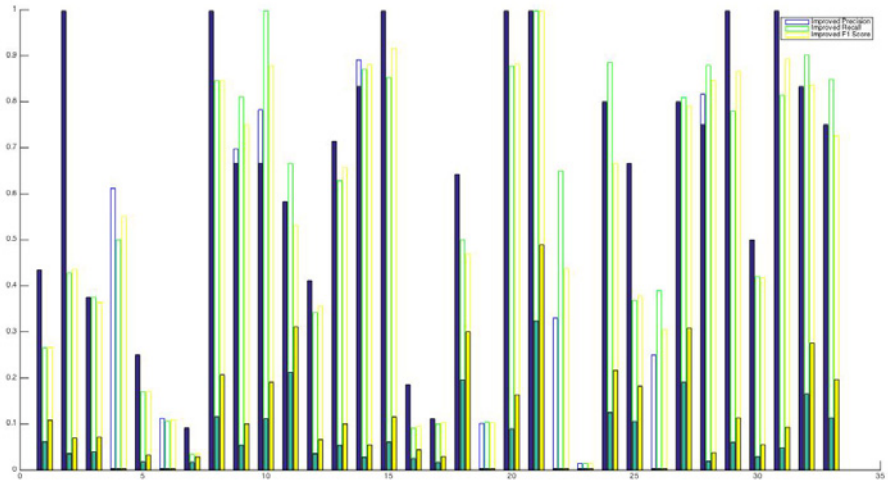

Fig.6. Improvements of proposed algorithm

Fig.6. manifests the improvements between our method and the baseline method. The transparent bars are the improvement of our method. Obviously, we can see that the precision, recall and F1 Score are improved profoundly.

Table. 1. Runtime of two methods

Number of news	**500**		**1000**		**2000**	
Method	Baseline	Our method	Baseline	Our method	Baseline	Our method
Runtime(s)	0.002	0.428	5.433	0.937	20.677	1.897

From Table.1, we can indicates that the runtime of baseline method is increased rapidly as the volume of news data. The size of word dictionary will grow exponentially when the number of news is increasing. And the computing cost will be intolerable. However, in our method, the dictionary depends on the concept in the ontology database. Therefore, our method achieves better when it comes to large dataset.

7. Conclusion

In this paper, we propose a financial news recommendation approach which aims to deliver the interesting news articles to the users, not only to help individuals save time and energy but also help enterprises improve user loyalty. To settle the ambiguity problem, we first constructed a specific financial domain ontology which stores key concepts and relationship between them. In the second place, a new presented method for called OF-IDF is used to represent the unstructured text data. In addition, a recommendation algorithm is proposed to model a user's genuine interests and predict current interests automatically by applying the thinking of relevance feedback. Finally, we conducted an experiment to compare the new approach with the baseline method which applies the classic TF-IDF weighting and the cosine similarity. The experimental results demonstrated that the algorithm we proposed markedly improved the quality of news recommendation.

In future work, we would like to combine our algorithm with other semantic approaches proposed in [19, 20]. Besides, position data can be used to extend our model to improve the performance of news recommendation. Eventually, it would be better to compare our algorithm with some other traditional methods.

Acknowledgements

The authors are very grateful to National Natural Science Foundation of China (Grant No. 71471161, 71071151, 71331005, 71110107026) for research support.

References

[1] Carreira, R., Crato, J. M., Gonçalves, D., Jorge, J. A. Evaluating adaptive user profiles for news classification, Proceedings of the 9th international conference on Intelligent user interfaces, 2004

[2] Good, N., Schafer, J. B., Konstan, J. A., Borchers, A.,Sarwar, B., Herlocker, J., Riedl, J. Combining collaborative filtering with personal agents for better recommendations, Proceedings of the 16th national conference on Artificial intelligence and the 11th Innovative applications of artificial intelligence conference innovative applications of artificial intelligence, 1999.

[3] Hogenboom F, Vandic D, Frasincar F, et al. A query language and ranking algorithm for news items in the Hermes news processing framework. Science of Computer Programming, 2014, 94: 32-52.

[4] Resnick P, Iacovou N, Suchak M, et al. GroupLens: an open architecture for collaborative filtering of netnews. Proceedings of the 1994 ACM conference on Computer supported cooperative work. ACM, 1994: 175-186.

[5] Liu J, Dolan P, Pedersen E R. Personalized news recommendation based on click behaviour. Proceedings of the 15th international conference on Intelligent user interfaces. ACM, 2010: 31-40.

[6] Adomavicius, G., Tuzhilin, A. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Transactions on Knowledge and Data Engineering 2005,17(6):734–749.

[7] Salton G, Wong A, Yang C S. A vector space model for automatic indexing. Communications of the ACM, 1975, 18(11): 613-620.

[8] Goossen F, IJntema W, Frasincar F, et al. News personalization using the CF-IDF semantic recommender. Proceedings of the International Conference on Web Intelligence, Mining and Semantics. ACM, 2011: 10.

[9] McGuinness D L, Van Harmelen F. OWL web ontology language overview. W3C recommendation, 2004, 10(10): 2004.

[10] Tom Gruber. The Encyclopedia of Database Systems, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag, 2009.

http://tomgruber.org/writing/ontology-definition-2007.htm
[11] Gruber T R. A translation approach to portable ontology specifications. Knowledge acquisition, 1993, 5(2): 199-220.
[12] Borst W N. Construction of engineering ontologies for knowledge sharing and reuse. Universiteit Twente, 1997.
[13] Studer R, Benjamins V R, Fense ID. Knowledge Engineering, Principles and Methods. Data and Knowledge Engineering,1998,25(12):161~197
[14]]Bateman J A. On the relationship between ontology construction and natural language: a socio-semiotic view. International Journal of Human-Computer Studies, 1995, 43(5): 929-944.
[15] http://wordnet.princeton.edu/
[16] Lavrenko V, Schmill M, Lawrie D, et al. Language models for financial news recommendation. Proceedings of the ninth international conference on Information and knowledge management. ACM, 2000: 389-396.
[17] Li Q, Wang J, Chen Y P, et al. User comments for news recommendation in forum-based social media. Information Sciences, 2010, 180(24): 4929-4939.
[18] Das A S, Datar M, Garg A, et al. Google news personalization: scalable online collaborative filtering. Proceedings of the 16th international conference on World Wide Web. ACM, 2007: 271-280.
[19] Business Intelligence Applications and the Web: Models, Systems and Technologies. Business Science Reference, 2012.
[20] IJntema W, Goossen F, Frasincar F, et al. Ontology-based news recommendation. Proceedings of the 2010 EDBT/ICDT Workshops. ACM, 2010: 16.
[21] Cui L, Shi Y. A Method based on One-class SVM for News Recommendation. Procedia Computer Science, 2014, 31: 281-290.
[22] Cui L, Meng F, Shi Y, et al. A Hierarchy Method Based on LDA and SVM for News Classification. Data Mining Workshop (ICDMW), 2014 IEEE International Conference on. IEEE, 2014: 60-64.