Journal of Engineering Science and Technology Review 10 (3) (2017) 1- 6

Research Article

JOURNAL OF
Engineering Science and Technology Review

www.jestr.org

An Interactive Information-Retrieval Method Based on Active Learning

Lei Chen[1*], Rong Bao[1], Yi Li[2], Kailiang Zhang[1], Yuan An[1] and Nguyen Ngoc Van[3]

[1]Jiangsu Province Key Laboratory of Intelligent Industry Control Technology, Xuzhou University of Technology, Xuzhou 221000, China
[2]Xuzhou Construction Machinery Group Research Institute, Xuzhou 221000, China
[3]School of Electronics and Telecommunications, Hanoi University of Science and Technology, Hanoi 10999, Vietnam

Received 6 April 2017; Accepted 15 June 2017

Abstract

Comprehending user demands through several human–computer interactions can effectively increase information-retrieval accuracy. Mainstream active learning algorithms use uncertainty sampling strategy. However, such algorithms cannot produce satisfactory results under few interactions. To improve interactive information-retrieval efficiency and accuracy, an sampling strategy based on the error-correcting capacity of samples was proposed for active learning. This strategy evaluated the expected value of unlabeled samples by calculating their potential error-correcting capacity associated with the classifier. Based on this sampling strategy, a fast interactive information-retrieval scheme adopting reinforcement learning and low-complexity classifier was designed in this study. The effects of three sampling strategies (random sampling, uncertainty sampling, and the proposed sampling strategy based on error-correcting capacity) on information-retrieval accuracy were examined using an experiment through a text set of Reuters-21578. Experimental results demonstrated that the proposed sampling strategy achieved higher retrieval accuracy and stability than random and uncertainty samplings. The retrieval accuracy of the proposed scheme was approximately 1.6% higher than that of the sampling algorithm based on uncertainty strategy. The proposed scheme can be used for real-time information retrieval because of its low computational complexity. The production of this study can improve the accuracy and latency of interactive information-retrieval services.

Keywords: information retrieval, active learning, machine learning, human–computer interaction, quality of user experience

1. Introduction

With the acceleration of mobile network access, the netizen population with access to internet network through mobile terminals is rapidly increasing and is accompanied with computational burden transfer to cloud end [1]. Some services such as online translation, voice cloud, and information push service are based on information retrieval in large corpus. Limited by natural language understanding techniques, accurate retrieval in a large text library is often difficult, and interactive inquiries are necessary to gradually determine the retrieval target. Active machine learning strategy is required for the abovementioned interactive information retrieval. Active learning asks questions instead of receiving knowledge passively and can increase pertinence in the next interaction according to response. Active learning is conducive for discovering high-value samples in the sample space and acquiring an accurate description of user demand. Users have higher requirements on latency and accuracy of these information retrieval services. To improve quality of user experience, some information-retrieval services introduced higher requirements on retrieval accuracy and speed. For high-accuracy information retrieval, the limitations on computational expenses and the number of interactions are two great challenges for active learning in interactive information retrieval.

Active learning has been the research hotspot in the machine learning field. Lewis and Gale introduced the uncertain sampling strategy [2]. In their classification experiments, the scheme combining the strategy selecting samples which cannot be certainly classified as the problem for every inquiry and the probabilistic classifier achieved satisfying results. The aforementioned active learning strategy becomes a common machine learning strategy and is gradually combined with various machine learning algorithms in the field of nature language processing and information retrieval [3]. The common active learning strategies include Query by committee, Margin sampling, and Posterior probability [4–5]. All the aforementioned methods achieved good effect in the information-retrieval field.

However, establishing high-quality training set under few interactions and reducing the calculation complexity of text classifiers are two key problems for interactive information retrieval to guarantee a positive user experience.

2. State of the Art

Active learning algorithm is the key for interactive information retrieval. Among several machine learning approaches, support vector machine (SVM) offers a solid theoretical basis. In the field of active learning, the combination of SVM and several sampling strategies has

*E-mail address: chenlei@xzit.edu.cn

ISSN: 1791-2377 © 2017 Eastern Macedonia and Thrace Institute of Technology. All rights reserved.
doi:10.25103/jestr.103.01

achieved great progresses. Combining SVM, several active learning strategies have been developed based on uncertainty sampling analysis and sample influences on version space, such as Simple Margin, Max-min Margin, and Ratio Margin [6]. The degree of classification mode changing caused by labeling sample has attracted certain research attentions in active learning field, and became an important standard for selecting the most informative samples [7]. Nevertheless, using SVM was difficult for online services, due to its expensive space and time consumption. Although multiplexing technique can slightly shorten the training time[8], it creates a heavy computational burden when the unlabeled sample space is large. Therefore, the SVM-based active learning algorithms are difficult to meet timeliness in online retrieval services.

Reinforcement learning method was proposed later in the literature [9] and is widely used in the text classification field [10]. With the promotion of the latent semantic model, reinforcement learning methods have become more and more popular for text classification [11–12]. Among these algorithms, AdaBoost.MH with low complexity is especially applicable for applications based on online information retrieval in the mobile internet network. However, it underperforms on a small training set. If the training set is formed by traditional uncertain sampling strategy, establishing a big training set is necessary. This will increase the number of interactions and deteriorate user experience.

Interactive retrieval demands for low-complexity and high-accuracy active learning algorithm. And the key for increasing accuracy of active machine learning algorithm depends on the selection of high-informative samples [13]. Conventional active learning algorithms formed the initial training set by selecting high-representative samples through clustering analysis [14], and then labeled the most uncertain samples. However, usually achieving unsatisfying effects in interactive information retrieval is due to small initial training set and existence of outliers [15].

Given the limited initial information for retrieval, a classifier often forms a “wrong understanding” for the retrieval target in its early period. High-informative sample should correct the “wrong understanding” in few interactions and increase retrieval accuracy. Therefore, a sample value judgment strategy based on the error-correcting capacity of classification model was designed in this study. Considering the timeliness requirement of interactive information retrieval, a machine learning algorithm with low computation complexity was proposed based on the AdaBoost.MH algorithm. With low computational complexity, the algorithm formed a simplified training set through iterative interaction to increases information-retrieval accuracy.

In this study, Section 3 designs interactive retrieval steps and introduced a machine learning algorithm using sampling strategy based on error-correcting capacity. Section 4 made an interactive retrieval experiment based on text library, and compares the retrieval accuracy of the algorithms based on different sampling strategies. Section 5 presents conclusions.

3. Methodology

3.1 General framework

Interactive information retrieval has the following three characteristics: (1) No unlabeled sample set exists in the beginning of retrieval, and data on key words or behavioral habits are few. (2) During the interaction, “high-value samples” are selected to reduce the number of interaction, establish a high-quality training set, and form high-accuracy classifiers. (3) To guarantee online user experience, the latency of interactive information retrieval must not be too high, namely the formed training set must be small. These characteristics introduce certain requirements on algorithm complexity.

These problems are the research focuses of the active learning strategy. According to the characteristics of active learning method and interactive information retrieval, the basic flowchart of interactive retrieval algorithm based on traditional active learning strategy can be designed as follows:

Step 1) User submits query information;
Step 2) To calculate relevancy between query information and each document to establish the relevancy model.
Step 3) To improve relevancy model through several interactions with the user by using active learning strategy.
Step 4) To output retrieval results in the relevancy descending order according to document ranking of the relevancy model.

Usually the interactive retrieval system model based on active learning method can be expressed as $A=(C,L,S,Q,U)$, where C is classifier, $L=R_t \cup N_t$, R_t and N_t are relevant and irrelevant document set which are identified through interaction, S is classifier, Q is evaluation function that acquires simplified training set by identifying high-value samples, and U is the unlabeled sample set and evaluation object of Q.

In active learning strategy, the uncertain sampling [2], and reduction capacity of version space [6] are often viewed as an important evaluation criteria for sample values. Let C_L be the classifier formed by using $\mathbf{X_i}\Theta\mathbf{V_i}=\{x_{i1}',x_{i2}',...,x_{in}'\}$ as the training set, and $C_L(x)$ be the classifier calculated value of sample x. In binary classification, classification results are generally judged by sign of values. If $x \in U$, $|C_L(x)|$ is proportional to the certainty degree of categorization of x and samples with low certainty degree are usually chosen into the training set in active learning. Additionally, if $y(x)$ is the real category of sample x, the $\left|\{C_L', \forall C_L'(x)=y(x)\}\right|$ is size of the version size (where C_L' is set of all candidate classifiers), the samples that can reduce the version space to the greatest extent can be used as high-value samples. Reference [10] proposed using expected changing degree of the classification model as evaluation criteria for sample value and deduced that the expected change principle of the classification model is equal to the most uncertain sampling principle.

3.2 Active learning strategy based on error-correcting capacity

This study believed that a number of high certainty samples have also high values in interactive information retrieval. If these samples are recognized as different categorization by human, namely $y \cdot C_L(x) < 0$, then the high value of

$\left|C_L(x)\right|$ means the high error-correcting capacity to the classifier C_L. As x is added, the new training set $C_{L\cup\{x\}}$ will be closer to real user demands. Samples that can correct classifiers have higher values when interactions are few and the early cognition of classifier significantly deviates from the retrieval target. With consideration to timeliness requirement of interactive retrieval, a simple active learning strategy for interactive retrieval was designed according to the abovementioned principle by using the AdaBoost.MH algorithm based on term frequency analysis.

With the potential correcting capacity of unlabeled samples to judgment rules, the evaluation method of unlabeled samples value which is applicable to real-time interactive information retrieval was designed. For a specific document, this evaluation method can express sample values as:

$$\alpha \cdot po \cdot \left(\frac{\left(S_{Max} - S_d\right)}{\left(S_{Max} - S_{Min}\right)} \right) + \beta \cdot ne \cdot \left(\frac{\left(S_d - S_{Min}\right)}{\left(S_{Max} - S_{Min}\right)} \right) \tag{1}$$

where α and β are empirical coefficients, po is expected contribution (correcting capacity) of documents judged as positive sample to the classifier, ne is expected contribution (correcting capacity) of documents judged as negative sample to the categorization, S_d is the score given by classifier to the current document d (a higher score means a higher expectation for the document belonging to positive sample and a lower score translates to lower expectation), S_{Max} and S_{Min} are the highest and lowest scores of the classifier for unlabeled documents.

In Equation (1), $\frac{S_{Max} - S_d}{S_{Max} - S_{Min}}$ reflects the probability for document to be judged by the current classifier C_L as positive samples and $\frac{S_d - S_{Min}}{S_{Max} - S_{Min}}$ is the probability for document to be judged by the current classifier C_L as negative sample. Given that high certainty samples are less likely to be labeled to disagree with classifier expectation and cause interaction waste, the abovementioned two empirical coefficient require a well design.

The specific calculation methods of po and ne are the algorithm core and has to be designed according to specific classifier and document presentation method. In this study, documents were expressed by a vector space model and classifier used a weak classifier based on key term frequency with low computation complexity to meet timeliness requirements of interactive retrieval. To each unlabeled document sample, the calculation formula determining its contribution coefficients (correcting capacity) of on current classifier can be acquired as follows:

$$po = \sum_{\forall w \in W} c(w) \cdot idf(w) \tag{2}$$

and

$$ne = \sum_{\forall w \notin W} tf\text{-}idf(w,d), \tag{3}$$

where $c(w)$ is the relevancy between term w given by the classifier and the target query document. This score can be used to measure the consistence between sample and the retrieval target. And W is a set of key terms in the current document d. Let D be the document set, $d \in D$ is the current document and $Tr \subset D$ is the labeled document set. Let $|Tr|$ stand for total number of labeled document, $\#Tr(w)$ stand for number of labeled documents containing the word w and $\#(w,d)$ is the frequency of w in document d. The calculation formula of the idf function becomes $idf(w) = \log\left(|Tr| / \#Tr(w)\right)$, so the tf-idf function formula is $tf\text{-}idf(w,d) = \left(\#(w,d)\right) \cdot idf(w)$.

Based on this definition, the high value of po indicates that d contains some words which are believed by the classifier having high relevancy with retrieval target. If d is labeled as the irrelevant document, adding d into the training set can significantly correct errors of the current classifier. Next to the contribution coefficient, the probability that d is labeled as an irrelevant document is described as its relevancy ranking in the unlabeled document. The higher numerical value of ne indicates that the current document contains a number of unique words. If this document is labeled as a relevancy document, it can increase important judgment standard to the classifier and correct the classifier that is lack of key term. According to an abundant number of experiments, the empirical value of α/β in the formula (1) is 0.5.

3.3 Fast retrieval algorithm

With an active learning strategy based on error-correcting capacity, an interactive retrieval algorithm that gradually increases labeled samples from users and improves classifier performance was designed by combining the Boost algorithm and the classifier of low computation complexity. The flow chart is as follows:

Step 1) To randomly choose n documents into training set T.
Step 2) To train the classifier using Boost algorithm.
Step 2) To classify unlabeled documents by the classifier produced by Boost algorithm.
Step 4) To calculate the expected contribution of each document.
Step 5) To choose several documents with the highest contribution for user labeling and put them into training set T.
Step 6) If this Boost algorithm does not reach the iteration limit, return to Step2, otherwise, turn to Step 7.
Step 7) To rank the rest of the documents.
Step 8) To output the retrieval results.

If Step 5 is changed to “choose document with the most uncertainty classification”, this algorithm turns in to an algorithm based on the “most uncertain sampling” strategy.

4. Result Analysis and Discussion

4.1 Experimental design

Since information retrieval establishes one-class classifier, the selected learning algorithm has to take influences of absence of terms into account. Therefore, "AdaBoost.MH with real-valued predictions" algorithm was used in this experiment, and was not introduced here (see [16]).

Considering the timeliness requirement, the weak classifier based on term frequency which has low computation complexity in References [13, 16] was used and modified according to the needs of literature retrieval. The definition of the weak classifier is as follows:

$$h\left(d,l\right)=\left(\begin{array}{l} c_{0l}, if\left(w\notin d\right) \\ c_{1l}, if\left(w\in d\right) \end{array}\right., c_{jl}=\frac{1}{2}\ln\left(\frac{W_{+1}^{jl}+\varepsilon}{W_{-1}^{jl}+\varepsilon}\right), \tag{4}$$

where $\varepsilon=\frac{1}{2\left|Tr\right|}$. Let $D_j\left(w\right)$ be the set of including and excluding word w, then $D_{j=0}=\left\{d:w\notin x\right\}$ and $D_{j=1}=\left\{d:w\in x\right\}$, where d is a document, x is the key term set of document d. Let:

$$f\left(d,w,j,b\right)=\left(d\in D_j\left(w\right)\right)\wedge\left(y\left(d\right)=b\right), \tag{5}$$

where $b\in\left\{-1,+1\right\}$, $y\left(d\right)$=1 or -1, representing whether d is the target retrieval document. The value of logic expression (5) is 0 or 1. Equation (5) reflects the relationship between existence/absence of term and document classification. Meanwhile, the weighting function is set as follows:

$$B_{t+1}\left(d,l\right)=\frac{B_t e^{-\alpha_t y\left(d\right)h\left(d,l\right)}}{Z_t}, \tag{6}$$

where $l=y\left(d\right)$, t is iteration number, α_t is weight of the current iteration, and Z_t is normalized function:

$$Z_t=\sum_{d\in T_r}\sum_{l\in\left\{1,-1\right\}} B_t\exp\left(-\alpha_t y\left(d\right)h\left(d,l\right)\right). \tag{7}$$

Therefore, w_b^{jl} in the equation (4) is:

$$w_b^{jl}=\sum_{d\in D} D_t\left(d,l\right)f\left(d,w,j,b\right). \tag{8}$$

Hence, the weak classifier represented by Equation (4) means: when $j=1$, the intuitive meaning of function $h\left(d,l\right)$ without considering weight is that d is determined as the retrieval target when the d contains w and the number of documents with w which is related with retrieval target is higher than the document unrelated with the retrieval target. When $j=0$, the situation when d does not contain specific term w is considered.

One subset of Reuters-21578 was used as the sample set in this experiment. Furthermore, it contains 8 categories, namely, "bop" (105 documents), "gas" (105 documents), "gnp" (136 documents), "gold" (124 documents), "oil" (124 documents), "sugar" (162 documents), "supply" (174 documents) and "oilseed" (171 documents). In the pre-processing stage of documents, terms with an occurrence frequency smaller than three and meaningless expletives are deleted.

Given the limited initial information for real interactive retrieval, 5% of the documents were randomly selected as the training set in this experiment, which were used to collect some key term information to imitate initial input retrieval information of users. Moreover, this experiment simulated a total of 5 interactions and submitted 1% unlabelled documents for "user" labeling in each interaction, because the time spent on human computer interaction is limited in real-time services.

4.2 Experimental results

In this experiment, one category is hypothesized as the retrieval content of "user" in each interaction. Each category was retrieved 10 times to observe performances of different algorithms in interactive information retrieval. The retrieval accuracy of three sampling algorithms (random sampling algorithm, sampling algorithm based on uncertainty, and the proposed sampling algorithm based on error-correcting capacity) were compared in the experiment.

The initial training sets of each category for 10 retrievals in the experiment are all randomly produced. This enabled the observation of initial training set influences on classifier and active error-correcting capacity of the classifier in follow-up learning. Experimental results of different categories are listed from Table 1 – Table 8.

Table 1. Test result for "bop" categorization

Random	Correct	Uncertainty
0.7143	0.7794	0.7321
0.8679	0.7042	0.8511
0.56	0.7255	0.75
0.6364	0.75	0.7384
0.7333	0.7719	0.8974
0.5397	0.6	0.7794
0.9355	0.75	0.7576
0.7412	0.8	0.8679
0.8604	0.6986	0.8936
0.5941	0.7451	0.7593

Table 2. Test result for "gas" categorization

Random	Correcting	Uncertainty
0.8632	0.9540	0.8730
0.9512	0.9647	0.9322
0.8359	0.6226	0.9607
0.8119	0.9082	0.9011
0.7818	0.8989	0.9368
0.94	0.8947	0.9406
0.9462	0.9610	0.9072
0.8433	0.9432	0.9462
0.8797	0.8989	0.9681
0.8806	0.9524	0.9091

Table 3. Test result for "gnp" categorization

Random	Correcting	Uncertainty
0.8991	0.9322	0.8876
0.9259	0.9701	0.9459
0.9213	0.9125	0.9053
0.8899	0.9830	0.8667
0.8990	0.9178	0.9623
0.9126	0.9437	0.8791
0.8812	0.95	0.9070
0.9560	0.9275	0.9247
0.7589	0.9833	0.8713
0.8738	0.9138	0.8666

Table 4. Test result for "gold" categorization

Random	Correcting	Uncertainty
0.9464	0.9756	0.9775
0.7692	0.9778	0.96
0.9266	0.9670	0.9241
0.9247	1.0	0.9559
0.9316	0.9535	0.9737
0.9252	1.0	0.9571
0.9449	0.9762	0.9855
0.9286	0.9787	0.9196
0.9455	0.9324	0.9868
0.9217	0.9670	0.9737

Table 5. Test result for "oil" categorization

Random	Correcting	Uncertainty
0.9171	0.9928	0.9521
0.9545	0.9860	0.9323
0.9583	0.9430	0.9367
0.9883	0.9853	0.9227
0.8680	0.9935	0.9425
0.9693	0.9625	0.8229
0.9732	0.9612	0.9261
0.9018	0.9758	0.9589
0.8491	0.9857	0.9738
0.9217	0.9787	0.9879

Table 6. Test result for "sugar" categorization

Random	Correcting	Uncertainty
0.9621	0.9806	0.9595
0.9650	0.9630	0.9821
0.9704	0.9623	0.9515
0.9632	0.9615	0.9514
0.9632	0.9633	0.9596
0.9627	0.9661	0.96
0.9136	0.9612	0.9519
0.9627	0.9623	0.9528
0.9648	0.9630	0.9778
0.9702	0.9554	0.9583

Table 7. Test result for "supply" categorization

Random	Correcting	Uncertainty
0.9167	0.9429	0.9444
0.7158	0.9268	0.9149
0.8587	0.9219	0.9324
0.8806	0.925	0.8831
0.8293	0.9688	0.9848
0.8514	0.8906	0.9059
0.9592	1.0	0.9642
0.8704	0.9265	0.8615
0.9714	0.9744	0.8667
0.6901	0.9149	0.8315

Table 8. Test result for "oilseed" categorization

Random	Correcting	Uncertainty
0.8019	0.7778	0.8425
0.6957	0.7826	0.7788
0.6329	0.8852	0.6711
0.8182	0.8125	0.8438
0.7603	0.7910	0.8429
0.87	0.875	0.8148
0.7683	0.8780	0.9359
0.6271	0.8197	0.7955
0.7683	0.7719	0.8095
0.6642	0.8704	0.7272

To observe the overall retrieval accuracies and capacity of active offsetting initial training set limitation of three sampling strategies, the average retrieval accuracies of each category and their fluctuations were calculated. The average retrieval accuracy of each category in 10 retrievals is presented in Fig.1.

In this experiment, two active learning strategies are significantly superior to learning strategies based on random training sets. The proposed active learning based on the error-correcting capacity of classifier achieves higher retrieval accuracy compared to the two learning strategies.

To further analyze the selection capability of active learning strategy in high-value samples, the Standard deviations of different categories in 10 simulation retrievals are calculated. High Standard deviation reflects the high sensitivity to the initial training set. Low Standard deviation indicates that the algorithm can search for desired samples pertinently and is only slightly influenced by initial training set that is manifested by the small fluctuation of retrieval accuracy. Fluctuations of retrieval accuracy of three algorithms are shown in Fig.2.

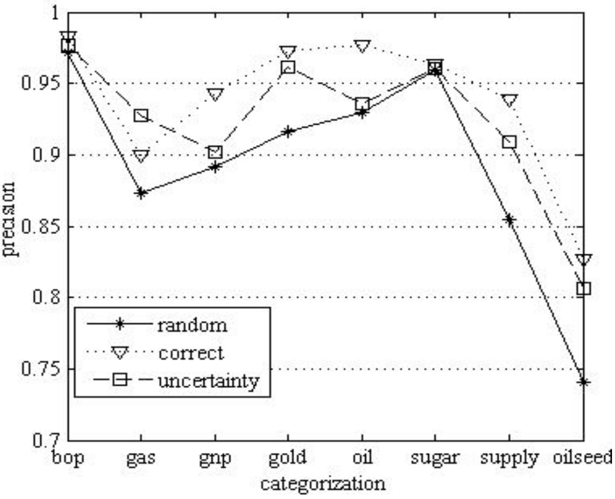

Fig. 1. Average retrieval accuracy of different categories in 10 retrievals (the y-axis is the average retrieval accuracy and the x-axis is 8 categories)

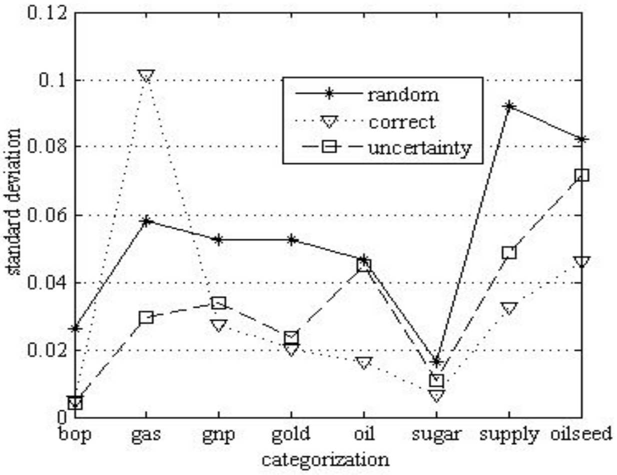

Fig. 2. Standard deviation of average retrieval accuracy of different categories (y-axis is the Standard deviation of average retrieval accuracy and x-axis is 8 categories)

The average retrieval accuracy of the proposed active learning algorithm gently fluctuates in most cases and is slightly influenced by initial training set, and is presented in Fig. 2. Therefore, the algorithm still performed with good stability under limited interaction number and training set space.

5 Conclusions

User experience in interactive information retrieval is sensitive to interaction number and service latency. To increase interactive information-retrieval accuracy and efficiency under limited interactions and limited labeled documents, and solve the outlier problem against uncertainty judgment standards, a novel sampling strategy based on expected error-correcting capacity of samples to current classifiers was proposed, and a text classification model was formed by combining the weak classifier with low computation complexity. On this basis, the active learning method based on term frequency with low computation complexity was designed based on the AdaBoost.MH algorithm to adapt with interactive information retrieval sensitive to latency. This study concludes the following:

(1) Interactive information retrieval based on reinforcement learning algorithm and weak classifier can reduce the complexity of the algorithm, thus, decreasing service delay and increasing user experience.

(2) In interactive information retrieval with limited initial training set and limited interactions, the active learning strategy based on the error-correcting capacity of samples indicates higher accuracy and stability than that based on uncertainty.

The proposed interactive retrieval algorithm considers the error-correcting capacity of samples and complexity of classifier. Furthermore, it can accelerate the accurate judgment of the retrieval target of users and provide effective online active learning techniques for real-time interactive retrieval services. However, it neglects important values of short text samples in the interaction. The effect of text length on interaction efficiency requires further exploration.

Acknowledgements

The authors are grateful for the support provided by Natural Science Foundation of Jiangsu Province (Grant No. BK20161165).

References

1. Toumi H., Brahmi Z., Benarfa Z., Gammoudi M.M., "Server load prediction using stream mining". *31st International Conference on Information Networking*, Da Nang, Vietnam: IEEE, 2017, pp. 653-661.
2. Lewis D.D., Gale W.A., "A Sequential Algorithm for Training Text Classifiers". *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval,* Dublin, Ireland: ACM, 29(2), 1994, pp.13-19.
3. Du B., Wang Z., Zhang L., Zhang L., Liu W., Shen J., Tao D., "Exploring Representativeness and Informativeness for Active Learning". *IEEE Transactions on Cybernetics*, 47(1), 2017, pp.14-26.
4. Tuia D., Ratle F., Pacifici F., Kanevski M.F., Emery W.J., "Active learning methods for remote sensing image classification". *IEEE Transactions on Geoscience & Remote Sensing*, 47(7), 2009, pp.2218-2232.
5. Liu K., Qiang X., Wang Z., "Survey on active learning algorithm". *Computer Engineering and Applications*, 48(34), 2012, pp.1-4.
6. Tong S., Koller D., "Support vector machine active learning with applications to text classification". *Journal of Machine Learning Research,* 2(1), 2002, pp.45-66.
7. Kremer J., Pedersen K.S., Igel C., "Active learning with support vector machines". *Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery*, 4(4), 2014, pp.313-326.
8. Hu R., Mac N.B., Delany S.J., "Active learning for text classification with reusability". *Expert Systems with Applications*, 45(C), 2015, pp.438-449.
9. Freund Y., Schapire R.E., "A decision-theoretic generalization of online learning and an application to boosting". *European Conference on Computational Learning Theory 1995: Computational Learning Theory*, London, UK: Springer, 1995, pp.23 -37.
10. Jiang Y., Zhou Z.H., "A Text Classification Method Based on Term Frequency Classifier Ensemble". *Journal of Computer Research and Development*, 43(10), 2006, pp.1681-1687.
11. Al-Salemi B., Ab-Aziz M.J., Noah S.A., "LDA-AdaBoost. MH: Accelerated AdaBoost. MH based on latent Dirichlet allocation for text categorization". *Journal of Information Science*, 41(1), 2015, pp.27-40.
12. Omar M., On B.W., Lee I., Choi G.S., "LDA topics: Representation and evaluation". *Journal of Information Science*, 41(5), 2015, pp.1-4.
13. Forestier G., Wemmert C., "Semi-supervised learning using multiple clusterings with limited labeled data". *Information Sciences*, 361(C), 2016, pp.48-65.
14. Zhu J., Wang H., Tsou B.K., Ma M., "Active learning with sampling by uncertainty and density for data annotations". *IEEE Transactions on Audio Speech & Language Processing*, 18(6), 2010, pp.1323-1331.
15. Zhu J., Wang H., Yao T., Tsou B.K., "Active Learning with Sampling by Uncertainty and Density for Word Sense Disambiguation and Text Classification". *Proceedings of the 22nd International Conference on Computational Linguistics,* Manchester, UK: ACL, 2008, pp.1137-1144.
16. Schapire R.E., Singer Y., "Boostexter: A boosting-based on system for text categorization". *Machine Learning*, 39(2-3), 2002, pp.135-168.