Intelligent Automation & Soft Computing
DOI:10.32604/iasc.2021.016434
Article

Tech Science Press

Improved Algorithm Based on Decision Tree for Semantic Information Retrieval

Zhe Wang[1,2], Yingying Zhao[1], Hai Dong[3], Yulong Xu[1,*] and Yali Lv[1]

[1]School of Information Technology, Henan University of Chinese Medicine, Zhengzhou, 450046, China
[2]School of Information Science and Technology, Donghua University, Shanghai, 201620, China
[3]School of Computing Technologies, RMIT University, Melbourne, VIC 3001, Australia
[*]Corresponding Author: Yulong Xu. Email: wzhe_wz@126.com

Received: 02 January 2021; Accepted: 30 April 2021

Abstract: The quick retrieval of target information from a massive amount of information has become a core research area in the field of information retrieval. Semantic information retrieval provides effective methods based on semantic comprehension, whose traditional models focus on multiple rounds of detection to differentiate information. Since a large amount of information must be excluded, retrieval efficiency is low. One of the most common methods used in classification, the decision tree algorithm, first selects attributes with higher information entropy to construct a decision tree. However, the tree only matches words on the grammatical level and does not consider the semantic of the information and lacks understanding of the information; meanwhile, it increases the amount of calculation and the complexity of the algorithm on synonymous fields, and the classification quality is not high. We investigate the retrieval method, unstructured processing with different semantic data, extracting the attribute features of semantic information, creating a multi-layered structure for the attribute features, calculating the window function according to the theory of multi-level analytic fusion, and fusing different levels of data. Then, we calculate the expected entropy of semantic information, undertake the boundary treatment of the attributes, calculate the information gain and information gain ratio of the attributes, and set the largest gain ratio of semantic data as the nodes of the decision tree. Our results reveal the algorithm's superior effectiveness in semantic information retrieval. Experimental results verify that the algorithm improves the expressing ability of knowledge in the information retrieval system and improves the time efficiency of semantic information retrieval.

Keywords: Semantic; information retrieval; decision tree

1 Introduction

With the advent of the internet, search engines are widely used [1], and the quick retrieval of target information from a massive amount of information is a core research area. Efficiency is an important measure of information retrieval methods [2,3]. With the development of natural language processing and artificial intelligence, semantic retrieval research has developed rapidly in recent years. The semantic web adds semantic information to networked documents, enabling computers to understand terms, concepts, and their logical relationships, allowing the maximum sharing and reuse of information. Therefore, semantic information retrieval methods can understand human language expressions so as to simplify communication and interactions between humans and computers.

Retrieval based on a semantic level can enable the analysis of semantics embedded in query words, enriching knowledge bases, improving the ability of computers to understand natural languages, and mining knowledge not clearly expressed by keywords. These features make the identification of retrieved content more accurate and improve the retrieval performance of search systems. Therefore, semantic information retrieval has become a topic of great interest in the field of computer data processing [4,5].

Semantic information retrieval extends the meanings of search words. It can consider the relationships between classes and attributes, establish semantic index items, and enhance the logical reasoning ability of a retrieval system [6]. It is not limited to the use of retrieval words as a starting point. Its results are semantic entities that contain the attributes of search words and relationships between them, instead of mechanically matching these strings, so as to increase the retrieval result space and precision. A semantic retrieval model enhances human-computer interaction. One way to improve its efficiency is to expand the query words input by users. Continuous updating of a query-extended word set can ensure that retrieval intentions are better understood, and provide a better retrieval experience. The process of semantic information retrieval is shown in Fig. 1.

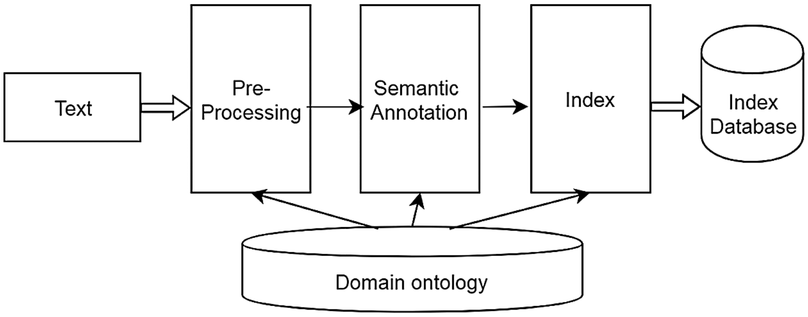

Figure 1: Process of semantic information retrieval

Semantic information and processing are attracting the attention of researchers. Its main retrieval methods include the statistical information retrieval model, Boolean and extended Boolean models, Bayesian model, and vector space model. Other models are based on ontology, and include the k-means algorithm [7], fuzzy c-means algorithm [8], Markov algorithm [9], semantic similarity algorithm [10], and notably, the cloud-computing-based clustering algorithm [11,12], which can be used to mine complex data, and whose broad development prospects have made it a focus of many experts.

Semantic information includes much complex, disordered, and highly variable information [13,14]. The traditional algorithm has been used for semantic information mining, which converges slowly, has high computational complexity, and may significantly reduce the efficiency of data processing [15]. An improved information retrieval algorithm based on a decision tree was proposed to avoid such limitations [16,17], using multi-level analytic fusion theory to obtain a window function for different information, and using this to fuse various levels of information. The results can be utilized to build a decision tree for semantic information retrieval.

2 Traditional Semantic Information Retrieval Principles

A traditional retrieval model (especially the Boolean model) is based on the literal matching of keywords or subject words; it ignores semantic information contained in keywords and lacks the ability to conduct semantic matching. This often leads to low recall and precision, and a poor user experience [18–20]. With the development of semantic web technologies, semantic retrieval research has developed rapidly [21], enabling users to input natural language and retrieve more keywords related to search keywords, instead of manually listing search information that only matches search terms [22,23].

A semantic retrieval model integrates all kinds of knowledge and information objects, intelligent and non-intelligent theories, and methods and technologies, including retrieval based on the knowledge structure, knowledge content, and expert heuristics; intelligent browsing retrieval based on knowledge navigation; and distributed multidimensional retrieval [24]. Common models include classification retrieval, multidimensional cognitive retrieval, and distributed retrieval [25]. The classification retrieval model uses the most essential relationship between things to organize resource objects; it has semantic inheritance, reveals the hierarchical and reference relationships of resource objects, and fully expresses the multidimensional combinational demands of users. The multidimensional cognitive retrieval model is based on the neural network, which simulates the structure of the human brain, organizes information resources into a semantic network structure, and constantly improves retrieval results through a learning mechanism and dynamic feedback technology. The distributed retrieval model uses a variety of technologies to evaluate the relevance of information resources to users' needs.

The semantic retrieval system, in addition to providing keywords to achieve subject retrieval, combines natural language processing and a knowledge representation language to represent various structured, semi-structured, and unstructured information, and to provide multi-channel and multi-functional retrieval [26]. Natural language [27] is the language that people use every day, and natural language processing technology can effectively improve retrieval efficiency. Its task is to establish a computer model that can imitate the human brain to understand, analyze, and answer natural language questions. From a practical point of view, the computer needs the ability to recognize a basic human-computer conversation and other language processing functions. For the Chinese language, there is a need for technology for Chinese word segmentation, phrase segmentation, and synonym processing.

Semantic retrieval is a search technology based on knowledge [28], using machine learning and artificial intelligence to simulate or expand people's thinking and improve the relevance of information content. It has obvious advantages: break through the limitation of single text matching, have an intelligent understanding of the purpose of user query, and deal with more complex information retrieval needs. Through various analysis, processing, and intelligent technologies, semantic retrieval can actively learn users' knowledge, provide personalized services, and improve its efficiency.

2.1 Common Algorithms of Information Retrieval Model

2.1.1 Boolean Models

The earliest information retrieval model, the Boolean model [29], is a simple model based on set theory and Boolean algebra. It is a strict matching model based on feature items, whose purpose is to find documents returned as "true" by a query word. The matching rules of a text query follow the rules of Boolean operations. Users can submit a query according to the Boolean logical relationship of search items in a document, and the search engine determines the results according to a pre-established inverted file structure. The standard Boolean logic model uses a binary decision criterion by which searched documents are either query-related or not. The results are usually not sorted by relevance.

In the Boolean model, a document is represented by a collection of key terms, which all come from a dictionary. While matching a query with a document, the model depends on whether its terms meet the query criteria. The model has a simple form and is easy to implement, but its exact matching will return too many or too few documents.

2.1.2 Vector Space Models

Vector space models [30] represent texts in the information base and a user's query as points (vectors) in a vector space. A vector value is the weight calculated by term frequency-inverse document frequency (TF-IDF). The similarity between documents is measured by that between vectors. The most commonly used similarity measure, the cosine distance formula, calculates the angle between two vectors as a measure of correlation. In the same space, if the angle is smaller, then the cosine is larger and the two vectors are more similar. Hence we can easily obtain the similarity between vectors by using the cosine theorem. The vector space model is the basis of text retrieval systems and web search engines.

In the vector space model, if the information retrieval system involves n keyword terms, then an n-dimensional vector space is established, where each dimension represents a keyword term. We must first establish the vector of texts and the user query. Each coordinate of a document vector is represented by the weight of the corresponding keyword, which indicates its importance to the user. Then the similarity between the query and text vectors is calculated. Based on the matching results, relevant feedback can be obtained to optimize the user's query.

2.1.3 Probabilistic Models

Probabilistic models [31] are based on the principle of probability ranking, and they consider the internal relationship between keywords and documents. Based on the Bayesian principle, a probabilistic model uses the probability dependence between keywords and between keywords and documents for information retrieval. It calculates the probability that a document is relevant, and sorts the documents based on that. If documents are sorted by decreasing probability, then those most likely to be obtained are ranked highest. This model aims to identify the uncertainty of relevance judgments and the fuzziness of query information representation in information retrieval.

2.2 Computing the Relevance of Semantic Information

The relevance calculation of semantic information mainly includes two categories: semantic similarity calculation based on distance and semantic similarity calculation based on attribute features [32–34].

2.2.1 Semantic Similarity Calculation Based on Distance

The semantic distance is an important factor to calculate the semantic similarity between concepts. Its main idea is: denote the value range of semantic distance as [0,∞), when the semantic distance between concepts is smaller, the semantic similarity between concepts is larger, otherwise the semantic similarity is smaller. The detailed formulas are shown in Eq. (1).

$$\mathrm{Sim}(\mathrm{x},\mathrm{y}) = \frac{\alpha}{dis(x,y)+\alpha} \tag{1}$$

In Eq. (1), Sim(x,y) is used to describe the semantic similarity based on distance, dis(x,y) is used to describe semantic distance, and α is the variable adjustment parameter.

2.2.2 Semantic Similarity Calculation Based on Attribute Features

Its core idea is that the attribute features of an instance object can reflect itself. It can determine whether there is a similar relationship between two instance objects according to the same or similar attribute features. The detailed formulas are shown in Eq. (2).

$$\mathrm{Sim(x,y)} = \frac{f_{com}(x,y)}{f_{total}(x,y)} = \frac{f(x \cap y)}{f(x \cap y) + \alpha(x-y) + \beta f(y-x)} \tag{2}$$

In Eq. (2), Sim(x,y) is used to describe the semantic similarity based on attribute features, f($x \cap y$) is the common attributes of x and y, f(x-y) is the attributes containing x but not y; however, f(y-x) is the opposite. α and β are the variable adjustment parameters which are used to describe the importance of x or y. For example, when the importance of the attribute features of x is higher than y, then a > β, or else a < β; only when the attributes of x and y are almost the same, a = β, can we get Sim(x,y) = 1.

2.3 Establishment of Markov Model for Semantic Information

The hierarchy of semantic information in a Markov model [35–37] can be described by N, N = (T, $\mathrm{B_T}$, S_T^b, Q_T^b, W), where T is the set of semantic state parameters of the semantic information retrieval system, $\mathrm{B_T}$ is the mapping from the semantic state parameter dataset to the semantic information dataset, Q is the mapping from the semantic state parameter dataset to T, S is the probability of a retrieval decision, and W is the retrieval time parameter. The probability of a retrieval decision is calculated as Eq. (3).

$$S_T^b = F\{s_{u+1} + \eta s_{u+2} + \eta^2 s_{u+3} + \ldots + \eta^{s-1} s_{u+p} | t_u = t, t_{u+p} = t^s\} \tag{3}$$

By the use of unstructured processing with different semantic data, the state parameters of retrieval decisions can be obtained as Eq. (4).

$$W_i(t) = F_i\left(\sum_{i=1}^{p} \eta(i) S_i | t_0 = t\right) \tag{4}$$

$$W^+(t) = \sum_b \omega(t,b) \sum_t q_t^b (S_t^b + \eta \mathrm{W}(t^b)) \tag{5}$$

In Eq. (4), η describes the decision to retrieve the i-th item of semantic information, and η(i) = $\eta^{\mathrm{y1+y2+\ldots+yp}}$.

Therefore, the state parameters of the optimal semantic information decision can be obtained. We can use Eq. (5) to calculate the optimal state parameters of different retrieval decisions.

Based on the above formulas, we can design the Markov model for semantic information retrieval,

$$R^+(t,b) = \sum_t Q_t^{b_2}\left(S_t^b + \eta \max\left(\sum_t q_t^b (S_t^b + \eta \mathrm{W}(t^b))\right)\right) \tag{6}$$

Using the above methods, different semantic information parameters can be initialized that can provide accurate data for decision-making in semantic retrieval.

3 Semantic Information Retrieval Method Based on Decision Tree Algorithm

The decision tree algorithm is one of the most common methods employed in classification, and is also used in semantic information retrieval. By recursively dividing the feature space of data, sample data are divided into clusters, and the classification rules are displayed in the form of a tree (as shown in Fig. 2) to discover and represent the knowledge contained in the data [38].

The traditional semantic information retrieval method is used to mine complex semantic data; however, it suffers from slow convergence and reduced efficiency of data processing due to the complexity and the massive volume of data. Aimed at these disadvantages, an improved semantic information retrieval algorithm is proposed based on the decision tree.

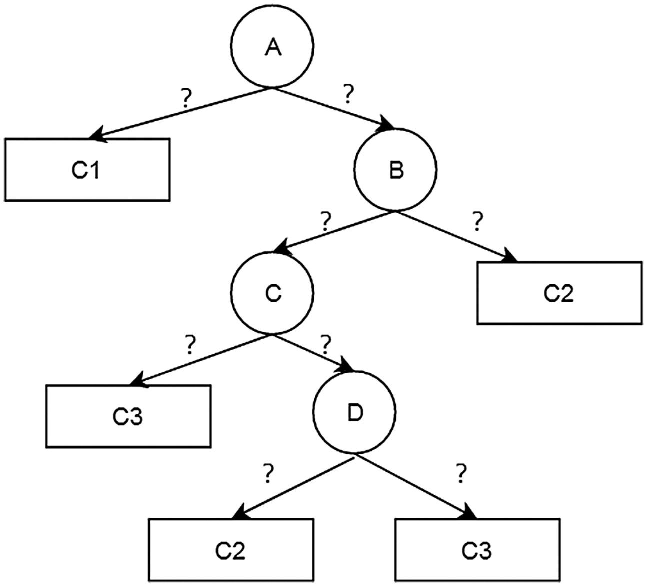

Figure 2: Decision tree model

3.1 Fusing Semantic Information

According to the multi-level analysis and fusion algorithm of data features, semantic information can be fused to achieve its retrieval [39–41].

There is much semantic information involved in semantic information retrieval, with quite different features. Semantic information types can be obtained. The detailed formulas are shown in Eq. (7).

$$\begin{aligned}\mathrm{PN(k,l)} &= \left|\frac{\partial^2 K(k,l)}{\partial k^2}\right| + \left|\frac{\partial^2 K(k,l)}{\partial l^2}\right| = |2K(k,l) - K(k-step,l) - K(k+step,l)| \\ &+ |2K(k,l) - K(k,l-step) - K(k,l+step)| UON(k,l) = \sum_k \sum_l NP(k,l)\end{aligned} \tag{7}$$

Using the window function in information retrieval, different information can be fused, the detailed formulas are shown in Eq. (8).

$$NP(k,l) = \sum_r \sum_s Y(r,s)[PN(k+r,l+s)]^2 \tag{8}$$

We can create a multi-layered structure for the data features of different semantic information. The detailed formulas are shown in Eq. (9).

$$\begin{aligned}\varphi_C &= \frac{PN^c(k,l)}{PN^c(k,l) + PN^D(k,l)} \\ \varphi_D &= \frac{PN^d(k,l)}{PN^c(k,l) + PN^D(k,l)}\end{aligned} \tag{9}$$

In the semantic information retrieval system, we can represent sub-detection systems by C and D, and fuse semantic data as Eq. (10).

$$\begin{array}{llll} K_H(\mathrm{k,l}) & = & K_C(\mathrm{k,l}) & NP_C(\mathrm{k,l}) - NP_D(\mathrm{k,l}) > \upsilon \\ K_H(\mathrm{k,l}) & = & K_C(\mathrm{k,l}) \times \psi\mathrm{C} + K_D(\mathrm{k,l}) \times \psi\mathrm{D} & |PN_C(\mathrm{k,l}) - PN_D(\mathrm{k,l})| \leq \upsilon \\ K_H(\mathrm{k,l}) & = & K_D(\mathrm{k,l}) & PN_C(\mathrm{k,l}) - PN_D(\mathrm{k,l}) < \upsilon \end{array} \tag{10}$$

Thus we can extract the features of different semantic information and carry out semantic fusion to increase the efficiency of semantic information retrieval.

3.2 Establishing the Decision Tree for Semantic Information Retrieval

We denote the semantic dataset as Z, $Z = \{(z_k, a_k)|k = 1,2,|\ldots, \mathrm{total}\}$, $z_k = (z_{k1}, z_{k2}, \ldots, z_{kf})$, where z_k is the semantic set of all dynamic information, with attributes $\{C_1, C_2, \ldots, C_f\}$.

We calculate the expected entropy for the semantic information data as Eq. (11).

$$\mathrm{K}(q_1, q_2, \ldots, q_p) = -\sum_{l=1}^{p} \frac{q_l}{total} log_2\left(\frac{q_l}{total}\right) \tag{11}$$

We denote the attributes of the semantic information dataset as C, C_h ($h = 1,2, \ldots, f$), which includes the attribute values of s. We undertake the boundary treatment for the attributes of all the semantic information data. The detailed formulas are shown in Eq. (12).

$$\mathrm{G}(C_h) = \sum_{u=1}^{s} \frac{q_{1u} + \ldots + q_{pu}}{total} K(q_{1u}, q_{2u}, \ldots, q_{pu}) \tag{12}$$

In Eq. (13), we calculate the information gain and the information gain ratio of the attributes, and get semantic information optimization parameters for the decision tree. The semantic data with the maximum gain ratio are regarded as the node of the decision tree, and are used to construct the semantic information retrieval decision tree.

$$\mathrm{E}(C_h) = \mathrm{K}(q_1, q_2, \ldots, q_p) - \mathrm{G}(C_h) \tag{13}$$

$$\mathrm{u}(C_h) = -\sum_{u=1}^{s} \frac{q_u}{total} \times nd\left(\frac{q_u}{total}\right)$$

$$C_k = MIN + \frac{MAX - MIN}{Q} \times k \ (k = 1, 2, \ldots, Q)$$

4 Experimental Results

The experimental environment was Visual C++. The number of semantic information samples in the database was 566, and there were 10 types of semantic information. Tab. 1 shows the semantic information sample data and attributes.

The traditional and improved algorithms were applied to mine the semantic information, with results as shown in Fig. 3, which show that when the semantic information sample has low complexity, the algorithms take almost the same time for data mining.

We then took all the semantic information sample data as the experimental object, and performed the same experiment, with results as shown in Fig. 4.

We can observe that when the semantic information is complex, the improved algorithm spends about 15 ms on data mining, whereas the traditional algorithm spends about 24 ms. This demonstrates the advantages of the improved algorithm when the sample complexity is high.

Table 1: Sample data information

Serial number	Attribute	Count
1	Port	65
2	World	68
3	Industry	99
4	Economy	20
5	Politics	39
6	Culture	41
7	Government	61
8	Security	59
9	Agriculture	35
10	Traffic	79

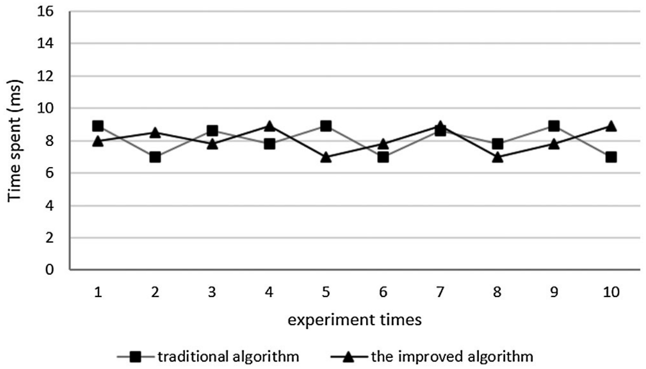

Figure 3: Retrieval results on semantic information sample with low complexity

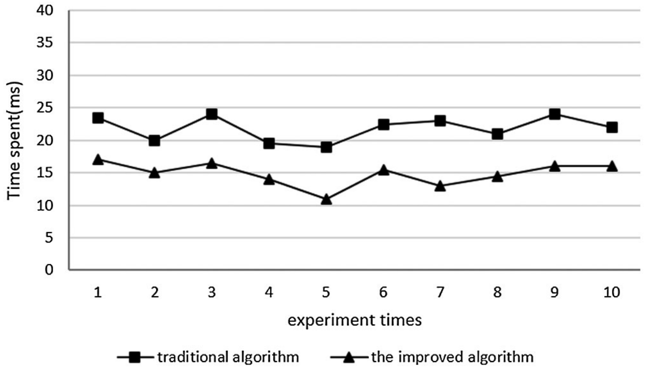

Figure 4: Retrieval results on semantic information sample with high complexity

Our experimental results show that the improved algorithm improves the speed of data mining for complex semantic information compared to the traditional algorithm.

5 Conclusion

Traditional semantic information retrieval is prone to slow convergence when the quantity of data is large and the data are complex. We conducted a study of a semantic information retrieval method based on a decision tree algorithm and the theory of multi-level analytic fusion. We obtained the state parameters of the optimal semantic information decision through the unstructured processing of semantic data, calculated the window function, and fused its levels of data to obtain information fusion results. We created a multi-layered structure for the data features of diverse kinds of semantic information and extracted the attribute features of semantic information. We calculated the ratio between information gain and the information gain of the dynamic data attributes of the semantic information and set the largest gain ratio of semantic data as the nodes of the decision tree. We thereupon developed a semantic data optimization parameter decision tree, based on which we built a decision tree to achieve semantic information retrieval. Our experimental results showed that the algorithm thus improved can add to the efficiency of semantic information retrieval, and help to avoid the disadvantages commonly associated with traditional algorithms. Due to the influence of experimental conditions, the decision tree based on the semantic information retrieval system selects part of the data as the experimental object. In the future, we need to make further evaluation for the system on large-scale datasets.

Funding Statement: The author(s) received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] D. Carmel, E. YomTov, A. Darlow and D. Pelleg, "What makes a query difficult?," in *Proc. 29th, ACM SIGIR*, New York, USA, pp. 390–397, 2006.

[2] B. T. Kumar and S. M. Pavithra, "Evaluating the searching capabilities of search engines and metasearch engines: A comparative study, Annals of Library and Information Studies," *National Institute of Science Communication and Information Resources (India)*, vol. 57, no. 2, pp. 87–97, 2010.

[3] M. Caramia, G. Felici and A. Pezzoli, "Improving search results with data mining in a thematic search engine," *Computers and Operations Research*, vol. 31, no. 14, pp. 2387–2404, 2004.

[4] Z. M. Xing, L. N. Wang, W. B. Xing, Y. J. Ren, J. Y. Xia *et al.,* "Application of ontology in the web information retrieval," *Journal on Big Data*, vol. 1, no. 2, pp. 79–88, 2019.

[5] D. Bollegala, Y. Matsuo and M. Ishizuka, "A web search engine-based approach to measure semantic similarity between words," *IEEE Transactions on Knowledge & Data Engineering*, vol. 23, no. 7, pp. 977–990, 2011.

[6] A. Alhroob, W. Alzyadat, A. T. Imam and G. M. Jaradat, "The genetic algorithm and binary search technique in the program path coverage for improving software testing using big data," *Intelligent Automation & Soft Computing*, vol. 26, no. 4, pp. 725–733, 2020.

[7] A. A. Ahmed and B. Akay, "A survey and systematic categorization of parallel k-means and fuzzy-c-means algorithms," *Computer Systems Science and Engineering*, vol. 34, no. 5, pp. 259–281, 2019.

[8] J. M. Liang, "Image retrieval based on genetic FCM algorithm and support vector machines," *Computer Engineering & Applications*, vol. 45, no. 20, pp. 165–168, 2009.

[9] R. H. David, L. T. Miller, M. Richard and S. Schwartz, "Using hidden markov models for information retrieval," in *Proc. TREC-7*, Gaithersburg, USA, pp. 273–284, 1998.

[10] N. Cao, S. Li, K. Shen, S. Bin, G. Sun *et al.,* "Semantics analytics of origin-destination flows from crowd sensed big data," *Computers, Materials & Continua*, vol. 61, no. 1, pp. 227–241, 2019.

[11] L. Yao, J. L. Gu and Y. F. Gao, "Optimized ciphertext retrieval for cloud computing based on dynamic clustering," in *Proc. 3rd ACM Workshop on Information Hiding and Multimedia Security*, New York, USA, pp. 353–359, 2016.

[12] H. Zhang, G. Chen and X. Li, "Resource management in cloud computing with optimal pricing policies," *Computer Systems Science and Engineering*, vol. 34, no. 4, pp. 249–254, 2019.

[13] H. Dong, F. K. Hussain and E. Chang, "Ontology-learning-based focused crawling for online service advertising information discovery and classification," in *Proc. ICSOC 2012*, Shanghai, China, pp. 591–598, 2012.

[14] S. G. Small and L. Medsker, "Review of information extraction technologies and applications," *Neural Computing & Applications*, vol. 25, no. 4, pp. 533–548, 2014.

[15] Z. W. Tian, S. B. Tian, T. Wang, Z. Gong and Z. Q. Jiang, "Design and implementation of open source online evaluation system based on cloud platform," *Journal on Big Data*, vol. 2, no. 3, pp. 117–123, 2020.

[16] J. Guan, J. Li and Z. Jiang, "The design and implementation of a multidimensional and hierarchical web anomaly detection system," *Intelligent Automation & Soft Computing*, vol. 25, no. 1, pp. 131–141, 2019.

[17] D. Liu, T. R. Li and D. C. Liang, "Incorporating logistic regression to decision-theoretic rough set for classifications," *International Journal of Approximate Reasoning*, vol. 55, no. 2, pp. 197–210, 2014.

[18] M. S. Kasture, A. J. Shivagaje, C. G. Shelake and A. J. Nalavade, "An application of naive bayes classifier to explore big data using XLSTAT," *International Research Journal of Agricultural Economics and Statistics*, vol. 9, no. 2, pp. 285–289, 2018.

[19] N. Poornima and B. Saleena, "Multi-modal features and correlation incorporated naive bayes classifier for a semantic-enriched lecture video retrieval system," *Imaging Science Journal*, vol. 66, no. 5, pp. 263–277, 2018.

[20] S. Wu and S. I. Mcclean, "Performance prediction of data fusion for information retrieval, Information Processing and Management," *Elsevier*, vol. 42, no. 4, pp. 899–915, 2006.

[21] J. He, E. Meij and M. D. Rijke, "Result diversification based on query-specific cluster ranking," *Journal of the Association for Information Science and Technology, Wiley-Blackwell*, vol. 62, no. 3, pp. 550–571, 2011.

[22] H. Dong, F. K. Hussain and E. Chang, "A transport service ontology-based focused crawler," in *Proc. SKG 2008*, Beijing, China, pp. 49–56, 2008.

[23] L. X. Jiang, Z. H. Cai, H. Zhang and D. H. Wang, "Naive bayes text classifiers: A locally weighted learning approach," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 25, no. 2, pp. 273–286, 2013.

[24] R. Alagarsamy, S. A. Sahaaya and A. Mary, "Intelligent rule-based approach for effective information retrieval and dynamic storage in local repositories," *Journal of Supercomputing: An International Journal of High-Performance Computer Design, Analysis*, vol. 76, no. 8, pp. 3984–3998, 2020.

[25] A. Islam, D. Ikpen and I. Kiringa, "Applications of corpus-based semantic similarity and word segmentation to database schema matching," *VLDB Journal*, vol. 17, no. 5, pp. 1293–1320, 2008.

[26] R. Wang, B. C. Li and W. Q. Du, "Entity disambiguation based on context word vector and topic models," *Chinese Journal of information technology*, vol. 33, no. 11, pp. 46–56, 2019.

[27] Y. Zhang and Q. Zhang, "An algorithm based on HNC theory for semantic calculation between words," *Computer Engineering and Applications*, vol. 41, no. 34, pp. 1–3, 2005.

[28] S. K. Karen and V. Douglas, "Complexity reduction in lattice-based information retrieval," *Information Retrieval*, vol. 8, no. 2, pp. 285–299, 2005.

[29] P. Wang and L. Wang, "Design of intelligent English translation algorithms based on a fuzzy semantic network," *Intelligent Automation & Soft Computing*, vol. 26, no. 3, pp. 519–529, 2020.

[30] J. Zhen, "Detection of number of wideband signals based on support vector machine," *Computers, Materials & Continua*, vol. 63, no. 1, pp. 445–455, 2020.

[31] D. Huang, E. Frank and H. Ian, "Learning a concept-based document similarity measure," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 8, pp. 1593–1608, 2012.

[32] S. Sagar, M. Khemaja, M. Lefrancois and I. Rebai, "Dynamic reconfiguration of smart sensors: A semantic web based approach," *IEEE Sensors Journal*, vol. 20, no. 3, pp. 1619–1629, 2019.

[33] P. Barnaghi, W. Wang and J. Kurian, "Semantic association analysis in ontology-based information retrieval," *Handbook of Research on Digital Libraries Design Development and Impact*, vol. 33, no. 1, pp. 131–141, 2009.

[34] D. Lewandowski, "Evaluating the retrieval effectiveness of web search engines using a representative query sample," *Journal of the Association for Information Science and Technology*, vol. 66, no. 9, pp. 1763–1775, 2015.

[35] I. E. Edem, S. A. Oke and K. A. Adebiyi, “A novel grey-fuzzy–markov and pattern recognition model for industrial accident forecasting,” *Journal of Industrial Engineering International*, vol. 14, no. 3, pp. 455–489, 2018.

[36] Q. S. Lian, S. Song, S. Z. Chen and B. S. Shi, “A phase retrieval algorithm based on higher-order markov random fields and nonlinear compressed sensing,” *Acta Electronica Sinica*, vol. 45, no. 9, pp. 2210–2217, 2017.

[37] J. Y. Xiao, L. M. Zou and C. Q. Li, “Optimization of hidden markov model by a genetic algorithm for web information extraction,” in *Proc. IJCIS 2007*, Chengdu, China, pp. 301–306, 2007.

[38] J. Mantas and J. Abellan, “Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data,” *Expert Systems with Application*, vol. 41, no. 10, pp. 4625–4637, 2014.

[39] S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. Grossman, O. Frieder *et al.,* “Fusion of effective retrieval strategies in the same information retrieval system,” *Journal of the Association for Information Science and Technology*, vol. 55, no. 10, pp. 859–868, 2004.

[40] B. Abdullah, H. Daowd and S. Mallappa, “Semantic analysis techniques using twitter datasets on big data: Comparative analysis study,” *Computer Systems Science and Engineering*, vol. 35, no. 6, pp. 495–512, 2020.

[41] S. Wu, C. Huang and L. Li, “Fusion-based methods for result diversification in web search,” *Information Fusion*, vol. 41, no. 10, pp. 16–26, 2018.