Open Access Library Journal
2018, Volume 5, e5005
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Research on Artificial Intelligence Frontier Recognition Based on LDA

Ting Xie, Ping Qin, Juehu Yan

Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing, China
Email: chilli6279@163.com

How to cite this paper: Xie, T., Qin, P. and Yan, J.H. (2018) Research on Artificial Intelligence Frontier Recognition Based on LDA. *Open Access Library Journal*, **5**: e5005.
https://doi.org/10.4236/oalib.1105005

Received: October 26, 2018
Accepted: December 2, 2018
Published: December 5, 2018

Abstract

Research frontier is the focus of scientific frontier and guides the direction of scientific development. It is of great significance for the state, institutions and researchers to grasp the research frontier in a timely and accurate manner. Based on LDA model, this paper uses Python language to carry out standardized processing, stop words removal, stem extraction and word shape restoration on foreign artificial intelligence data from 2013 to 2017. The processed data are imported into LDA model to output topic—vocabulary matrix and document—topic matrix. The topic is described on the basis of the topic—vocabulary matrix, and the research frontier is calculated in the light of the document topic matrix and the constructed frontier identification index, the research frontier of artificial intelligence abroad is obtained, which includes three categories: computer vision research, application of artificial intelligence in various fields and data mining and clustering research.

Subject Areas

Information Science, Library, Intelligence and Philology

Keywords

Artificial Intelligence, LDA, Formatting, Research Frontier, Python

1. Introduction

Artificial intelligence has entered the rapid development stage, and becomes the important strategic development direction of global science and technology [1]. Researchers from various disciplines, whether in computer science, philosophy or economics, are all working hard on artificial intelligence in response to the national strategy of seizing the highest point of artificial intelligence. Intelligence research should contribute to the strategy of artificial intelligence. Grasping the frontier of AI research is the basis of effective research, and the research frontier

identification and tracking are the strength and superiority of information science research. It is of great significance for the state, institutions and researchers to grasp the research frontier in a timely and accurate manner. At home and abroad, the research on the research frontier, based on citation and word frequency analysis, has been very rich, and the comparative study of various methods has also emerged one after another. However, there are still three problems in the two methods: the lag of citation analysis, the lack of semantic information support and the inability to effectively integrate data sources. Therefore, this paper intends to use LDA model based on topic model to explore the research frontier of artificial intelligence from the semantic level.

2. Data Selection and Processing

2.1. Data Sources

Scopus is the largest A & I database in the world today. Daily updates of Scopus help researchers keep abreast of the frontiers of research [2]. Therefore, this paper selects the Scopus database, the year is limited to 2013 to 2017, the subject category is limited to "artificial intelligence (1702)", the type of literature for journals and conferences. Conferences were limited by reference to the Category A of the International Academic Conferences on Artificial Intelligence recommended by CCF of the Chinese Computer Society. In the end, 9000 journals and 6058 conferences were obtained.

2.2. Data Processing

Data processing is mainly for the above data standardization processing, stop words removing, stemming and lemmatization, the entire data processing of this study is based on Python [3].

1) Standardization Processing, The smallest processing unit that a machine needs to understand is the word (participle). The Word_tokenize () method is a generic and powerful method for identifying all types of corpuses.

2) Stop Words Removing, Stop word removing is one of the most commonly used preprocessing steps. In general, conjunctions, articles and pronouns are listed as stoppages. Some specific parts of English, such as conjunctions such as for, or, or the word "the" have no meaning to the subject model. These words are called stop words and need to be removed from our list of words.

3) Stemming, Stemming is a process of pruning and cutting leaves.

4) Lemmatization, Lemmatization is a structured approach that covers all the grammatical and variational forms of the root. The morphological reduction operation uses context and morphology to determine the change form of the relevant words, and uses different standardization rules to obtain the relevant roots according to the part of speech.

3. LDA Topic Model Construction

3.1. Model Principle

In 2003, as Blei *et al.* proposed the latent Dirichlet Allocation model in [4],

which is a three-layer Bayesian probability generation model, which uses iterative estimation to calculate the subject vocabulary of the document. The main idea is to assume that each document is a mixture of multiple topics, and each topic is a probability distribution over multiple vocabularies. There are D-documents and W-vocabulary in the corpus. Assuming that the documents have a K-topic, the process of generating the LDA model theme is:

1) For each document $d \epsilon D$, according to the Dirichlet distribution $\theta_d \sim \mathrm{Dir}(\alpha)$, the subject distribution parameters of the document d are obtained;

2) For each topic $z \epsilon K$, according to the Dirichlet distribution $\varphi_z \sim \mathrm{Dir}(\beta)$, the multi-distribution of the vocabulary on the subject z is obtained;

3) For the i-th vocabulary in document D, the subject is obtained according to the polynomial distribution $Z_{d,i} \sim \mathrm{Mult}(\theta_d)$.

In the LDA model, parameter settings, parameters, and most empirical studies are based on the rule of thumb, *i.e.*, setting α = 50/K, β = 0.01; the number of topics K is constrained by the subject consistency score, that is, when the consistency score is the highest, K takes the most Excellent value; two Dirichlet distributions θ_d, φ_z, which cannot be directly obtained, are estimated by Gibbs sampling algorithm in actual research [5] [6] [7] [8].

3.2. Consistency Score

In order to determine the number of topics in the text set, this paper uses the evaluation index consistency score in the statistical language model to determine the optimal number of topics [9]. The formula for consistency score is described as follows:

$$\mathrm{coherence}(V) = \sum_{(v_i,,)\epsilon V} \mathrm{score}(v_i, v_j, \epsilon) \tag{1}$$

$$\mathrm{score}(v_i, v_j, \epsilon) = \log p\left[(v_i, v_j) + \epsilon \Big/ p(v_i)\, p(v_j)\right] \tag{2}$$

In formulas (1) and (2), the probability of each word appearing in the text set, and N is the number of all words appearing in the text set. The consistency score is calculated by the co-occurrence frequency of the words in the sliding window, which increases with the increase of sentence similarity, so the higher the consistency score is, the better (**Table 1** and **Figure 1**).

4. Construction of Frontier Indicators Based on LDA Model

4.1. Topic Intensity

According to the conceptual connotation and related theories of research frontier, the view of Small and Griffith in [10] and the point of Persson in [11], those study hold that the core feature of research frontier is "Topic intensity", which represents the degree of research in this field. The higher the topic attention, the more researchers pay attention to a certain field and the higher the research output.

Theme intensity is the proportion of topics in the document, and is a quantitative

Table 1. The consistency score of different the number of Topic.

The number of Topic	15	20	25	30	35	40	45
consistency score	0.448717	0.461249	0.445073	0.450096	0.452616	0.442233	0.474482
The number of Topic	50	55	60	65	70	75	80
consistency score	0.448378	0.441599	0.349508	0.352140	0.342051	0.356929	0.328431

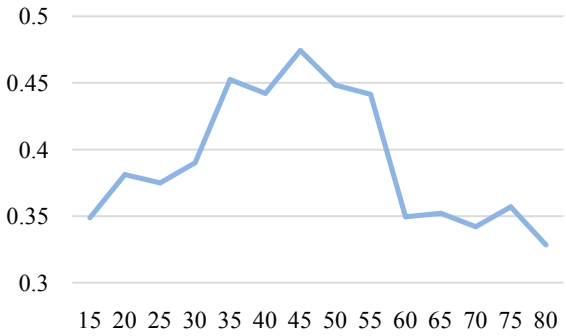

Figure 1. Consistency score.

index to measure whether the research topic is a frontier topic [12]. Theme intensity can be expressed as the ratio of the sum of the weights of the subject in all scientific literature to the total amount of literature:

$$\theta_j = \sum_d \theta_j^{(d)} \Big/ M \tag{3}$$

4.2. Theme Novelty

In order to quantitatively analyze the novelty of research frontiers, the average publication time of research frontiers is an important identification index. The novelty of topic is represented by the average time of issuing the theme, and it is a quantitative index to measure the novelty of the theme [13]. The closer the topic is, the higher the novelty of the topic is. The research frontiers are new scientific discoveries and problems in the field, so their novelty is relatively high. n the process of calculating topic novelty, a supporting document for a topic is composed of the probability distribution of the document. The document probability value can be understood as the strength of the document. For example, in topic *K*, the greater the probability value of document d, the higher the document's support for the topic.

The novelty of a topic represents the timeliness of a research topic, which can be reflected by the average age of the theme. The closer the subject's age of publication is, the higher the novelty of its theme. The novelty of the topic is an important index to discover new research frontiers.

5. Frontier Recognition Based on LDA Model

5.1. Topic Extraction Based on LDA Model

Through the estimation of the consistency score, the number of topics $k = 45$, *a*

= 50/k, and b = 200/w, the data were imported into the constructed LDA model to be tested, and the topic—vocabulary matrix and document—topic matrix based on probability were obtained.

1) Topic-Vocabulary Matrix

The topic vocabulary distribution results extracted from topics are shown in Table 2. For the sake of space, only the top ten words with the highest probability in each topic are listed in the table.

2) Topic—vocabulary matrix

The document—topic probability distribution generated by LDA model topic recognition can visually see the relevant information of each topic, including

Table 2. Topic—Vocabulary Probability Distribution Table (TOP12).

Topic 0	Topic 1	Topic 2	Topic 3
data (0.044)	detect (0.042)	infer (0.038)	object (0.037)
market (0.020)	system (0.029)	rule (0.037)	video (0.032)
big (0.018)	fault (0.014)	probabl (0.022)	actio (0.015)
energi (0.016)	test (0.013)	distribut (0.022)	motion (0.013)
price (0.014)	argumen (0.010)	weight (0.021)	method (0.011)
intell (0.013)	the (0.010)	probabilist (0.019)	contro (0.011)
learn (0.010)	schedu (0.009)	bayesian (0.011)	trajector (0.011)
distan (0.010)	analys (0.009)	approxim (0.011)	3D (0.010)
trade (0.009)	anomal (0.008)	model (0.009)	system (0.009)
system (0.009)		sampl (0.009)	detect (0.009)
Big data analysis based on intelligent learning	**Fault Location Analysis Based on Clustering**	**Bayesian Probability Model Inference**	**Machine Vision 3D Inspection System**
Topic 4	Topic 5	Topic 6	Topic 7
tree (0.053)	support (0.085)	facial (0.021)	imag (0.030)
algorithm (0.046)	vector (0.079)	comput (0.021)	descriptor (0.014)
search (0.022)	machin (0.042)	memori (0.021)	extract (0.014)
forest (0.017)	system (0.016)	express (0.014)	comput (0.012)
optim (0.017)	twin (0.016)	human (0.011)	the (0.012)
random (0.010)	svm (0.014)	face (0.010)	vision (0.012)
intell (0.009)	quadrat (0.012)	system (0.010)	detect (0.011)
genet (0.008)	classif (0.011)	interact (0.009)	shape (0.011)
proble (0.008)	the (0.007)	analysi (0.009)	process (0.010)
the (0.008)	propos (0.007)	imag (0.008)	local (0.009)
Artificial intelligence-related algorithm estimation	**Classification Based on Dual Support Vector**	**Face Recognition System and Analysis**	**Computer vision and image processing**
Topic 8	Topic 9	Topic 10	Topic 11
prefer (0.038)	risk (0.029)	featur (0.042)	sourc (0.032)
match (0.022)	2016 (0.024)	data (0.031)	transfer (0.032)
vote (0.014)	cancer (0.017)	cluster (0.031)	music (0.016)
intellig (0.013)	breast (0.017)	classif (0.030)	target (0.016)
artifici (0.009)	divers (0.015)	method (0.027)	spatial (0.015)
comput (0.009)	ordin (0.014)	select (0.026)	data (0.013)
choic (0.008)	electr (0.010)	learn (0.018)	learn (0.013)
incomplet (0.008)	closur (0.009)	algorithm (0.015)	risk (0.010)
problem (0.008)	patholog (0.009)	sampl (0.012)	system (0.009)
system (0.007)	intellig (0.008)	propos (0.012)	discoveri (0.008)
Research on artificial Intelligence matching problem	**Application of Artificial Intelligence in Cancer**	**Clustering Selection Algorithm Based on Feature Vector**	**Method for manually generating music**

title, keywords and abstract. The document—topic probability distribution table is a 15,058 × 4 list, with 15,058 rows representing one document and 5 columns including topic number, topic strength, vocabulary, text information. Due to the large content, it is presented in the appendix. The relevant calculations of frontier identification are all calculated according to the probability distribution table of document—topic distribution. Table 3 presents the relevant information of the literature information with the highest contribution rate in each topic, which can provide the relevant information of each topic in a very detailed manner for easy understanding and analysis.

Table 3. Probability Distribution Table of Topics in Documents (Topic Intensity Top1) - (Top11 Categories).

T_Num	Per_Contrib	Keywords	Text
0	0.96749997	data, market, big, energi, price, intellig, learn, distanc, trade, system	Granularities and inconsistencies in big data analysis
1	0.85110002	detect, system, 2015, fault, test, argument, the, schedul, analysi, anomali	A clustering-based strategy to identify coincidental correctness in fault localization
2	0.81900000	infer, rule, probabl, distribut, weight, probabilist, bayesian, approxim, model, sampl	Annealed importance sampling for structure learning in Bayesian networks
3	0.74830001	object, video, action, motion, method, control, trajectori, 3D, system, detect	Design and assessment of a machine vision system for automatic vehicle wheel alignment
4	0.87309998	tree, algorithm, search, forest, optim, random, intellig, genet, problem, the	Generalized rapid action value estimation
5	0.79299998	support, vector, machin, system, twin, svm, quadrat, classif, the, propos	Twin support vector hypersphere (TSVH) classifier for pattern recognition
6	0.83329999	facial, comput, memori, express, human, face, system, interact, analysi, imag	Dynamic adaptation of pedestrians: To a model guided by the perception
7	0.85350000	imag, descriptor, extract, comput, the, vision, detect, shape, process, local	Analysis of the best production condition of cleaner froth in bauxite flotation process based on froth texture coarseness measurement
8	0.88370001	prefer, match, vote, intellig, artifici, comput, choic, incomplet, problem, system	Elicitation and approximately stable matching with partial preferences
9	0.63230001	risk, 2016, cancer, breast, divers, ordin, electr, closur, patholog, intellig	Essential closures and supports of multivariate copulas
10	0.93599999	featur, data, cluster, classif, method, select, learn, algorithm, sampl, propos	Feature selection for high-dimensional imbalanced data
11	0.62050002	sourc, transfer, music, target, spatial, data, learn, risk, system, discoveri	Chaotic music generation system using music conductor gesture

The 45 research topics in the field of artificial intelligence in the past five years reflect the current research situation in this field, and the research contents in this field can be summarized and analyzed, mainly involving the following aspects:

The first category: research on the basic theories and methods of artificial intelligence, focusing on the research on the basic theories and related methods of artificial intelligence, the focus is on the big data analysis based on intelligent learning (Topic 0), Bayesian probabilistic model inference (Topic 2), constraint algorithm problem Topic 17), factorization algorithm of matrix (Topic 42) and data presentation based on implicit distribution model (Topic 43).

The second category: research on the design and stability of intelligent systems, mainly including robot intelligent systems (Topic 12), user information recommendation systems (Topic 13) and model design based on intelligent learning (Topic 16).

The third category: applied research for decision support, mainly on decision support, risk and uncertainty, matching value fuzzy set measure (Topic 21), machine intelligent prediction model (Topic 25) and expert intelligent method research of cloud network (Topic 27).

The fourth category: cognitive and neuroscience-inspired artificial intelligence research, mainly for neural networks, brains and other related research, cognitive systems and decision models (Topic 15), neural network learning algorithms (Topic 19), deep learning and Convolutional neural networks (Topic 30), neural network spatial learning models (Topic 33).

The fifth category: computer vision research, research on recognition, face recognition and image segmentation, machine vision 3D detection system (Topic 3), face recognition system and analysis (Topic 6), computer vision and image processing (Topics 7), related research on artificial intelligence matching problem (Topic 8), image fusion method evaluation (Topic 20), neural network based vision system research (Topic 24) and artificial intelligence based visual recognition (Topic 34).

The sixth category: data mining and clustering research. There are fault location analysis based on clustering (Topic 1), classification based on dual support vectors (Topic 5), clustering selection algorithm based on feature vectors (Topic 10), machine learning algorithm model (Topic 26), big data text classification (Topic 27), big data text classification (Topic 28), feature learning and classification method (Topic 36), machine learning algorithm based on sensitivity analysis (Topic 39) and data analysis and classification (Topic 41).

The seventh category: research on intelligent optimization algorithms, including artificial intelligence related algorithm optimization (Topic 4) and particle swarm optimization algorithm (Topic 18) for optimization and algorithm optimization.

The eighth category: natural language processing research, which focuses on information retrieval and natural language processing, includes query based on

semantic knowledge (Topic 14), intelligent analysis algorithm and intelligent search (Topic 31), natural language emotion processing (Topic 32) and intelligent search algorithm (Topic 44).

The ninth category: the application of artificial intelligence in various fields, such as medical, education, transportation, music, security, etc., the application of artificial intelligence in cancer (Topic 9), the artificial generation method of music (Topic 11), the game's intelligent algorithm game theory and Nash equilibrium (Topic 22), intelligent transportation design (Topic 23), the application of intelligent learning models in education (Topic 29), genetic prediction of clinical medicine (Topic 37), social network online communication Intelligent (Topic 38) and artificial intelligence applications in the field of network security (Topic 40).

5.2. Frontier Recognition

After LDA topic extraction, two important probability distributions are obtained: document—topic probability distribution and topic—word probability distribution. According to the document—topic probability distribution, topic strength and topic novelty can be calculated.

1) Frontier Recognition Based on Topic Intensity

According to the formula for quantitatively calculating the topic strength of the index, the document—topic probability distribution matrix obtained above is calculated to obtain a topic strength list, including the topic number, topic name and topic intensity value. Draw the topic intensity of all topics into a line chart, as shown in Figure 2. The line is the average value of the topic intensity, which can be used to visually see the topic number above the average.

Figure 2 shows a line chart of topic intensity, the horizontal axis represents the topic, the vertical axis represents the intensity of the topic, and the horizontal lines for the topic intensity average. Among them, topic 42 (the application of artificial intelligence in the field of network security) has the highest topic novelty, reaching 0.42066372, and topics 6, 10, 11, 12, 15, 16, 17, 18, 25, 30, 31, and 36 have higher topic novelty values. The topic of 35 topics in Figure 2 is more novel than the average.

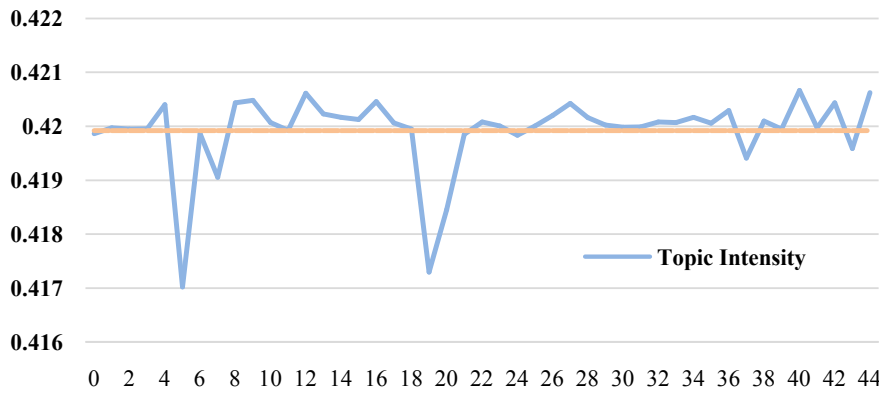

Figure 2. Example of a figure caption (figure caption).

There are 35 topics whose topic intensity value exceeds the topic average value, accounting for 77.78% of all topics. Among them, the top three topics are classified as category 9, category 6, category 2 and category 8, with 21% of the topics being category 9. The application of artificial intelligence in various fields, such as medical treatment, education, transportation, music, safety, etc. 19% of the topics come from the sixth category of data mining and clustering research, which focuses on classification, mining, clustering and machine learning. 12% of the topics come from the research on the design and stability of the second type of intelligent system, the research on artificial intelligence related systems and their stability, and 12% from the research on the eighth type of natural language processing, and the research on information retrieval and natural language processing.

2) Frontier Identification Based on Topic Novelty

According to the formula for quantitatively calculating the topic novelty of the index, the document—topic probability distribution matrix obtained above is calculated to obtain a list of topic novelty, including topic number, topic name and topic novelty. The topic intensity of all topics is plotted as a bar chart. As shown in Figure 3, the straight line is the average value of topic novelty, and the topic number higher than the average value can be visually seen.

Figure 3 represents the topic novelty value histogram, the column height represents the topic novelty, the horizontal axis is the topic number, the horizontal line is the average topic intensity value, which is 2015.22748. Among them, topic 5 (classification based on dual support vectors) has the highest topic novelty, reaching 2015.38204, and topics 0, 1, 3, 5, 6, 7, 11, 18, 19, 20, 21, 24, 25, 30, 37, and 43 have higher topic novelty values.

There are 16 topics with topic novelty values exceeding the average of topic novelty, accounting for 35.56% of all topics. Among them, the topic classification with the highest proportion is the fifth category, with 31.25% of the topics being the fifth category, computer vision research, research on recognition, face

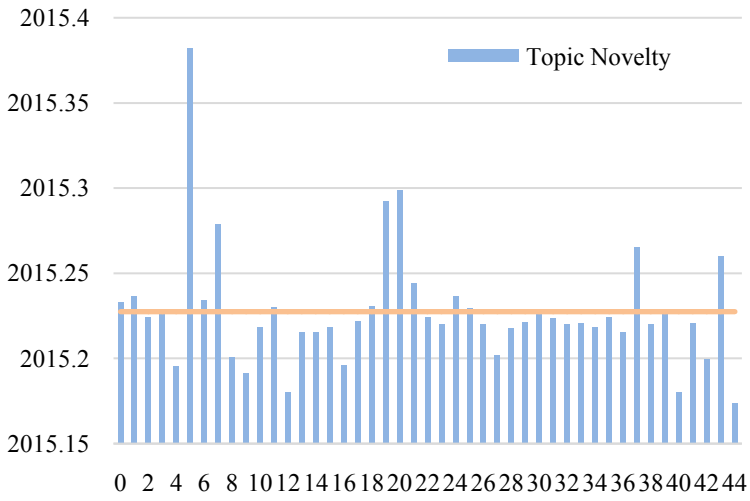

Figure 3. Distribution diagram of topic novelty value.

recognition and image segmentation, etc. In the second category (research on intelligent system design and stability, research on artificial intelligence related systems and their stability), the seventh category (research on intelligent optimization algorithms, research on optimization, algorithm optimization, etc.) and the eighth category (research on natural language processing, research on information retrieval, natural language processing, etc.), the value of topic novelty does not exceed the average value of topic novelty.

3) Frontier Identification

The frontiers calculated by topic intensity and theme novelty are shown in Table 4.

Table 4. Frontier Distribution Table.

	Topic Name	Vocabulary	Topic Novelty	Topic Strength
0	Big data analysis based on intelligent learning	data, market, big, energi, price, intellig, learn, distanc, trade, system	2015.23283	0.41986308
1	Fault Location Analysis Based on Clustering	detect, system, fault, test, argument, the, schedul, analysi, anomali	2015.2367	0.41996971
3	Machine Vision 3D Inspection System	object, video, action, motion, method, control, trajectori, 3D, system, detect	2015.22767	0.41995171
5	Classification Based on Dual Support Vector	support, vector, machin, system, twin, svm, quadrat, classif, the, propos	2015.38204	0.41701993
6	Face Recognition System and Analysis	facial, comput, memori, express, human, face, system, interact, analysi, imag	2015.23432	0.41985814
7	Computer vision and image processing	imag, descriptor, extract, comput, the, vision, detect, shape, process, local	2015.27885	0.41905205
11	Method for manually generating music	sourc, transfer, music, target, spatial, data, learn, risk, system, discoveri	2015.23034	0.41992369
18	Particle swarm optimization algorithm	optim, swarm, algorithm, particl, intellig, pso, financi, hybrid, learn, evolut	2015.2308	0.41994495
19	Neural network learning algorithms	network, neural, learn, algorithm, the, cnn, machin, comput, signal, markov	2015.29242	0.4172966
20	Image fusion method evaluation	method, featur, fusion, estim, propos, imag, model, detect, comput, local	2015.29904	0.41845379
21	Matching value fuzzy set measure	fuzzi, measur, set, function, relat, oper, intellig, gener, valu	2015.24441	0.41985219
24	Neural network based vision system research	network, system, track, neural, control, vision, visual, comput, sensor	2015.23655	0.41983066

Continued

25	Machine intelligent prediction model	predict, forecast, time, seri, network, model, traffic, data, perform, machin	2015.2296	0.42000201
30	Deep learning and Convolutional neural networks	deep, convolut, network, logic, knowledg, reason, model, neural, intellig, represent	2015.22756	0.41998241
37	Genetic prediction of clinical medicine	system, attribut, medic, network, data, predict, the, gene, model, clinic	2015.26544	0.41940783
43	Data presentation based on implicit distribution model	model, person, implicit, curv, system, narr, differ, data, feedback, advertis	2015.26018	0.41958783

Research frontiers include: Big data analysis based on intelligent learning, Fault Location Analysis Based on Clustering, Machine Vision 3D Inspection System, Classification Based on Dual Support Vector, Face Recognition System and Analysis, Computer vision and image processing, Method for manually generating music, particle swarm optimization algorithm, neural network learning algorithms, image fusion method evaluation, matching value fuzzy set measure, neural network based vision system research, machine intelligent prediction model, deep learning and Convolutional neural networks, genetic prediction of clinical medicine, data presentation based on implicit distribution model.

6. Conclusions

Based on LDA model, this paper studies the literature in the field of artificial intelligence abroad from 2013 to 2017: on the one hand, data collection and processing are completed; on the other hand, based on LDA model, the research frontier of artificial intelligence in foreign countries from 2013 to 2017 is concluded as follows:

1) Computer vision research is on recognition, face recognition and image segmentation, including 3D detection system of machine vision, face recognition system and analysis, computer vision and image processing, research on artificial intelligence matching, evaluation of image fusion methods, research on vision system based on neural network and vision recognition based on artificial intelligence.

2) The application of artificial intelligence in various fields, such as medical treatment, education, transportation, music, safety, etc., includes the application of artificial intelligence in cancer, artificial generation of music, intelligent algorithm game theory and Nash equilibrium for games, intelligent transportation design, application of intelligent learning model in education, gene prediction of clinical medicine, online communication intelligence of social networks, and application of artificial intelligence in network security. So far, artificial intelligence has produced many successful application areas, especially in the commercial field. The popularity of other fields is an inevitable trend in the context of the

increasing marketization of artificial intelligence.

3) Data mining and clustering research, for classification, mining, clustering and machine learning, there are cluster-based fault location analysis, dual support vector based classification, feature vector based cluster selection algorithm, machine learning algorithm models, big data text classification, big data text classification, feature learning and classification methods, machine learning algorithms based on sensitivity analysis, and data analysis and classification. At present, support vector machine algorithms have been successfully applied in many fields such as pattern recognition, image processing and bioinformatics.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Yu, H.Q., Cao, J.J. and Wang, Y.F. (2018) Frontier Analysis of International Artificial Intelligence Research from the Perspective of Information Science. *Intelligence magazine*, **37**, 21-26.

[2] Meho, L.I. and Yang, K. (2007) Impact of Data Sources on Citation Counts and Rankings of Lis Faculty: Web of Science versus Scopus and Google Scholar. *Journal of the American Society for Information Science & Technology*, **58**, 2105-2125. https://doi.org/10.1002/asi.20677

[3] Loper, E. and Steven (2002) Nltk: The Natural Language Toolkit. *ETMTNLP*'02 *Proceedings of the ACL*-02 *Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, Philadelphia, Pennsylvania, 7 July 2002, 63-70. https://doi.org/10.3115/1118108.1118117

[4] Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003) Latent Dirichlet Allocation. *J Machine Learning Research Archive*, **3**, 993-1022.

[5] Hu, J.M. and Chen, G. (2014) Mining and Evolution of Centent Topics Based on Dynamic LDA. *Library and Information Work*, **58**, 138-142. https://doi.org/10.13266/j.issn.0252-3116.2014.02.023

[6] Guan, P., Wang, Y.F. and Fu, Z. (2016) Effect Analysis of Scientific Literature Topic Extraction Based on LDA Topic Model with Different Corpus. *Library and Information Service*, **2**, 112-121. https://doi.org/10.13266/j.issn.0252-3116.2016.02.018

[7] Wang, P., Gao, C. and Chen, X.M. (2015) Research on LDA Model Based on Text Clustering. *Information Science*, **1**, 63-68.

[8] Ruan, G.C. and Xia, L. (2017) Retrieval Results Clustering Application Research Based on LDA. *Journal of Intelligence*, **36**, 179-184.

[9] Stevens, K., Kegelmeyer, P., Andrzejewski, D., *et al.* (2012) Exploring Topic Coherence over Many Models and Many Topics. *Conference on Empirical Methods in Natural Language Processing*.

[10] Small, H. and Griffith, B.C. (1974) The Structure of Scientific Literatures. Identifying and Graphing Specialties. *Science Studies*, **4**, 4-17. https://doi.org/10.1177/030631277400400102

[11] Persson, O. (1994) The Intellectural Base and Research Fronts of JASIS 1986-1990. *Journal of the American Society for Information Science*, **45**, 31-38. https://doi.org/10.1002/(SICI)1097-4571(199401)45:1<31::AID-ASI4>3.0.CO;2-G

[12] Zhang, S.S. (2016) A Comparative Study of Measurement Methods and Indicators of Scientific Frontier Features. Ph.D. Thesis, Dalian University of Technology, Dalian.

[13] Feng, J. and Zhang, Y.Q. (2017) Research on Scientific Frontier Identification and Analysis Methods Based on LDA and Ontology. *Information Theory and Practice*, **40**, 49-54. https://doi.org/10.16353/j.cnki.1000-7490.2017.08.009