International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-9 Issue-1, May 2020

Categorizing Multi-Label Product Questionnaires through SVM Based Click stream

Sathya Charanya. C, V. Saravanan

Abstract: ***In this paper, Question Categorization (QC) has been studied most primarily in order to understand customers' search intention. In both of these searches, the items in the question list relate to the category label belonging to the taxonomy tree that is being examined. Despite this, search queries about the product usually vary depending on what is vague, and introduce new products over time, seasonal trends and narrow. Traditional supervised approaches to E-Commerce QC are not possible due to the high volume of traffic and high cost for manual annotation in E-Commerce search engines. Here, clickstream data is utilized to determine the effectiveness of a channel's marketplace. So, using the customer's click concept, to collect large-scale question categorization data, this paper uses unsupervised methods that means SVM algorithm is mainly used in this system. Here the data is in the multiclass and multi-label classifier is used to classify them. This paper gets on a large multi-label data set with specific and individual queries from a specific category. In this paper, a comparison of different sophisticated text classifiers is viewed. This paper calculates the micro-F1 scores of top and leaf, which are considered to be a linear SVM-ensemble.***

Keywords: Question categorization, Clickstream data, Taxonomy classification, Multi-label classifier.

I.INTRODUCTION

The first is to speculate precisely what is the goal expressed in the user's questions, and to improve user satisfaction, and improving e-commerce conversion rates, understanding the question is important to retrieve. Users'queries is often difficult to understand, as users' questions are vague and narrow. The first level of question comprehension is QC. This means here that the questions are categorized into one or multiple node-defined target types. Then this paper has to figure out which type each question one depends on. Thus by predicting, and assigning hypothesis rank signal to the search engine, in QC content searching, relevancy can be increased. This kind of ability is very important, content search in e-commerce in particular is often used as a taxonomy tree, then that is sent to specific products that are classified. Customers get a big change in terms of their choices and plan to spend. With this growing number of people, people are ready to buy the services and products available in other countries. This, in turn, makes a lot of profit for e-commerce companies. The profit and product of this e-commerce company depends on the large population.

Manuscript received on April 02, 2020.
Revised Manuscript received on May 21, 2020.
Manuscript published on May 30, 2020.
***** Correspondence Author
Sathya Charanya.C, Assistant professor, Department of computer science, Salem sowdeswari college, Self-financing courses wing, Salem-636010.
Dr.V.Saravanan, Dean - Computer Studies, Dr.SNS Rajalakshmi College of Arts and Science, (an Autonomous Institution), Coimbatore.

Paper [1] therefore used the Naive Bayes algorithm to classify the text. In paper [2], dealing with the sparse data, for that purpose, the paper adopted a graph-based solution. There are lots of ways to understand the questions. To enable all these features efficiently, an efficient product assortment is required. This paper [3], therefore, has used deep learning methods. Nowadays, people buy products through e-commerce. They compare the product's services to other vendors. Moreover, a lot of customers post their valuable feedback. Seeing those feedback makes other customers buy products. So in this paper [4], they have used the deep learning method to solve product search. In this fast-growing moment of digital marketing, sending newly updated products to the customer. And offer many offers to the client. These e-commerce companies need to understand the needs of customers. The behavior of the users is thus stored in the database. So this paper [5] is pre-processing those records and then mapping them. Generally to search for products online, text queries must match image content. Paper [6], therefore, used the multimodal method to implement the shortest model fit. The Internet has grown greatly by merging sellers and buyers. Customers generally have different options for maximum product attributes. So this paper analyzed [7], customer reviews, then the classification of the concept. Paper [8] used the unsupervised method. The question has performed actions such as relationship detection and clustering. In paper [9], Fuzzy logic used the Logic method to process and categorizes a large number of reviews.

II.PROPOSED METHODOLOGY

This section describes the QC data collection approach and the taxonomy of E-commerce products.

1. Clickstream Data

Visit or Session S_n is a timeline or chronological of the actions of a registered customer that includes clicks, purchases, search, add to cart. A customer can interact with the search engine N times in a single S_n. For each question Q_i $(1 \leq i \leq N)$, Here, a customer can redeem n_i products $[P_{i1}, P_{i2} \ldots P_{in_i}]$. This paper is collecting product pairs and questions, (Q_i, P_{ic}), Here, $(1 \leq c \leq n_i)$ they are selected by mouse clicks, and purchased or added in the shopping cart. If the time interval between two actions in a session is more than thirty minutes, they are divided into two separate sessions. After doing the above gathering process in the all session, the set and query of the product pairs with their all click frequencies are obtained. In the next step, we need to increase click information to link product type labels to each question. Then an effective and simple way is, the query type label must be assigned to the related classification category with the selected product.

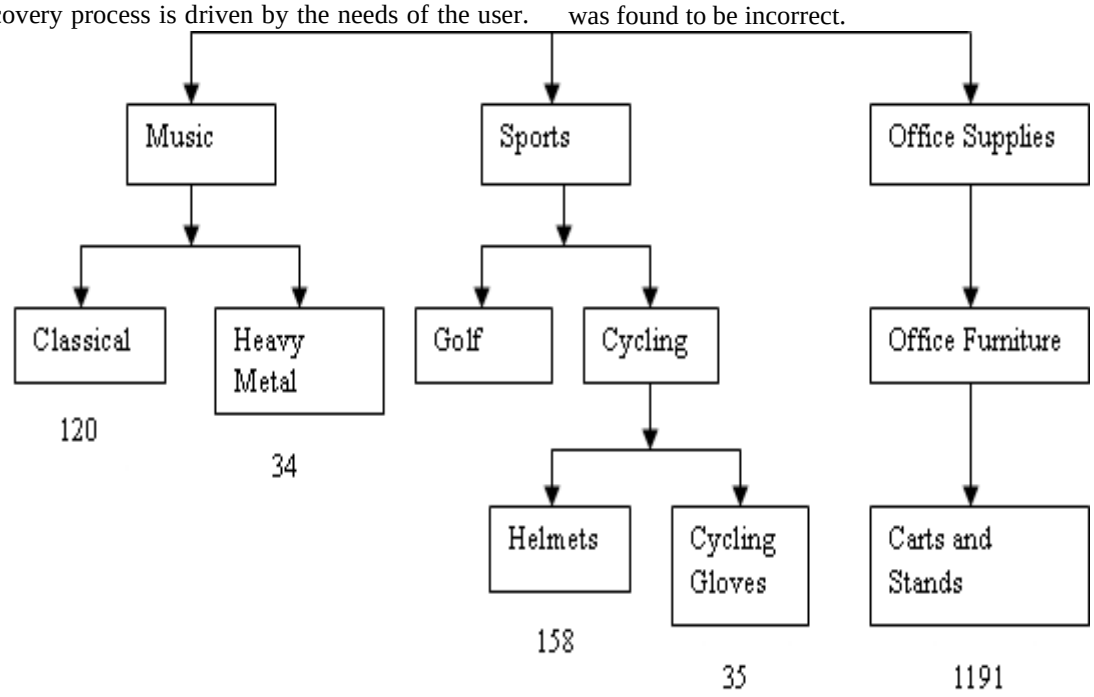
In selecting the restored product, based on the assumption that the customer's behavior is coherent, this is a reasonable approximation.

2. Product Taxonomy

Product taxonomy is a tree-based hierarchy. In it, each node is converted from a product list into another product category. Generally in the E-commerce setting, in the taxonomy tree, merchants are responsible for matching their products with the defined as leaf category. The sequence of nodes, from the root of the tree to the leaf, represents the semantic labels of an object, and this is often referred to as the most soaked in attribute-based breadcrumbs. The 1-to-1 graph is executed to convert each product to its corresponding path. As a result; previous query product pairs are converted into query label pairs. As stated earlier, each individual question can have a lot of labels.

E-commerce catalogs, collecting many products classified by the best grain taxonomic tree, and the actual product recovery process is driven by the needs of the user. Then the initial taxonomy does not reflect the format. In this context, Sports → Cycling →Bike Accessories→ Bike Fenders and Sports → Cycling → Bike Accessories→ Child Seatsare due to the small number of items classified in the taxonomy leaf, customers are unlikely to view. So combining these two nodes with their parent nodes, Sports →Cycling→ Bike Accessories. In this case, the products belonging to the removed leaf are in parent node; they get the opportunity to classify as a more general category.

Figure 1 shows a small portion of the taxonomy. Below each node is the sum of the click frequencies for all queries associated with that node. Due to the popularity of some product types the frequency of the node is equal. To re-balance the tree, if the sum of the total click frequencies of a node is less than a certain threshold, the fewer nodes are reunited with their parents. In this process, if the frequency threshold is set to 50, the number of type's decreases from 5,896 to 2,085. This hypothesis is validated by manually reviewing questions with multi-label type pairs. Then 24.8% was found to be incorrect.



1: Part of the taxonomy

3. Data Characteristics

In this paper, 403,349 questions labeled over 2,085 types were collected, the corps then derived from them were divided into test and training modules with the 3:1 split. Each question can be marked with more than one type; the mean number of question labels in the training set is 1.4. No wonder that 48 percent questions are equal to or less than two words; because, that's one of the challenges for QC. Figure 2 shows the % value of the questions with variety lengths in the training set,

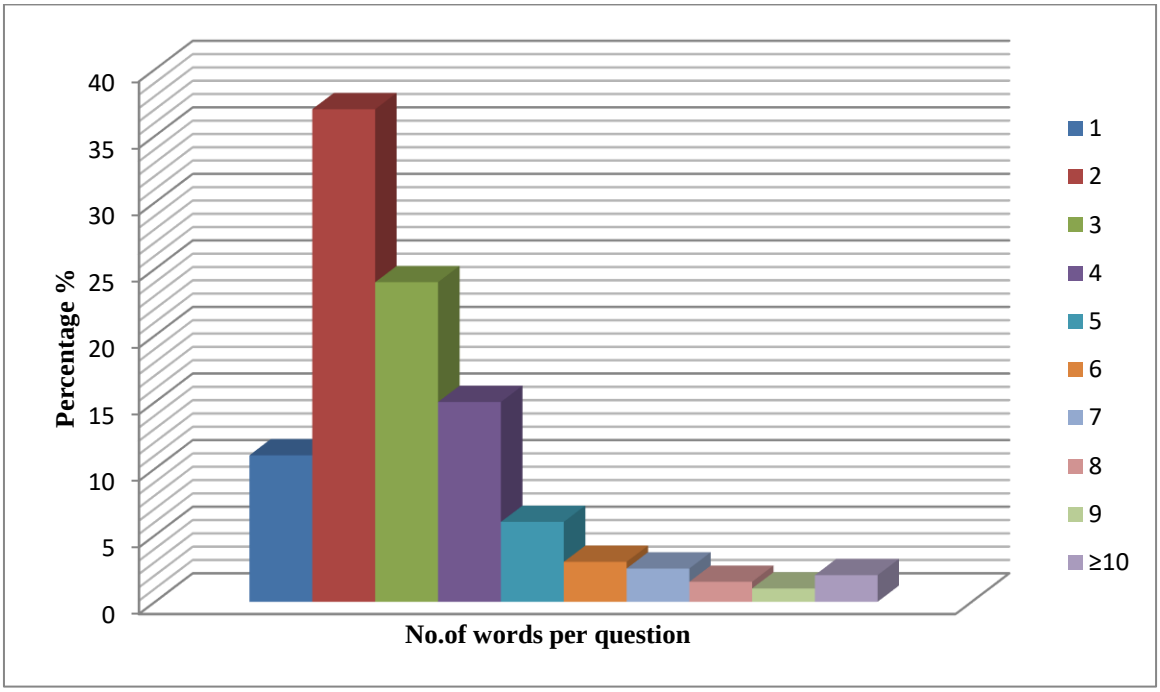


Fig.2: Percentage Value of the Questions

In the first stage of product taxonomy, there are 36 categories. The average depth of the tree is 3.37 and the maximum depth is 7. It is shown in Figure 3 that questions are treated equally in these 36 high-level categories. Less than 1 percent of questions come in very small numbers, and more than 23 percent of the questions come in large size.

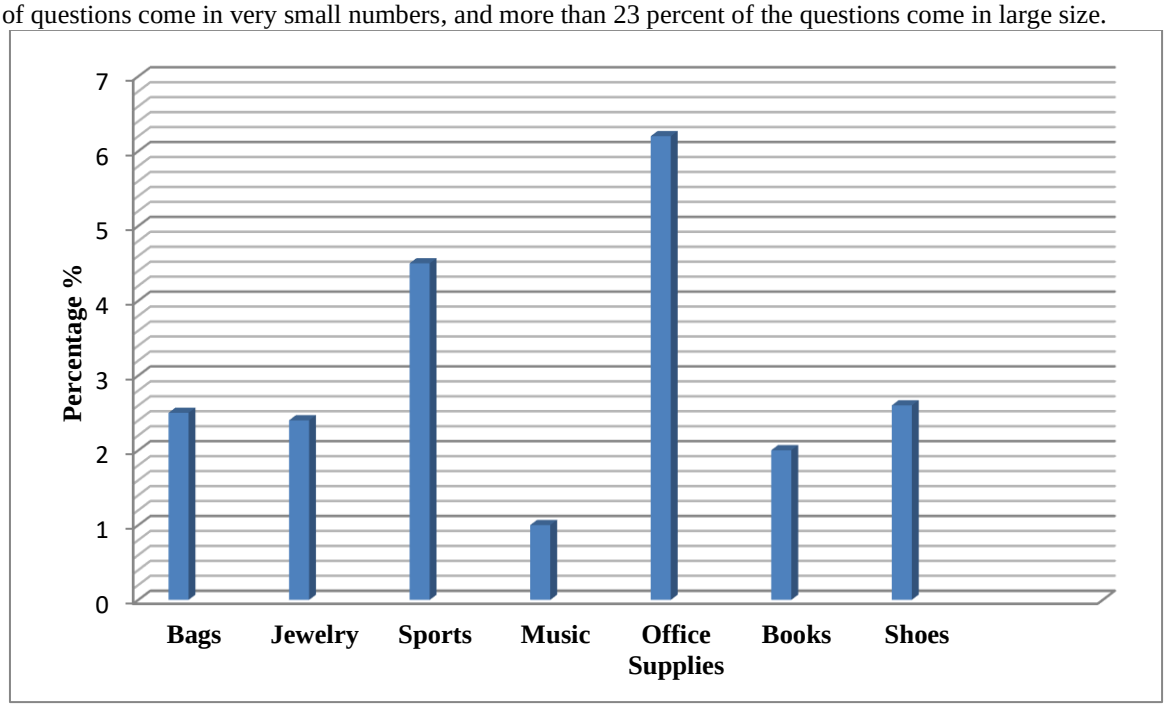


Fig.3: Percentage of questions among 36 level-1 categories in the training set

III.METHODS

With a web question, question classification can be created as a text classification task. The goal is to find the best n variety. To accomplish this task, this sheet uses some text classifiers, such as Gradient Boosting Trees (GBT), fastText, attention based CNNs, logistic regression, and SVMs.

1. GBT

In the predictive learning problem, the purpose is to find the approximate F (a) of the F*(a) function. It maps the input set a = {a_1... a_n} to the label b. At the same time, using M training data {a_i, b_i} M_i = 1, minimizes a certain loss of functionality. This GBT seeks to minimize the loss of functionality below:

$$L = E_b [L (b, F (a)) a] \quad (1)$$

F(a) can be any sort of function, like a decision tree,to find the optimal solution, it has to follow the additive strategy and number optimization paradigm.

$$F^*(a) = \sum_{n=0}^{N} f_n (a, x, p) \quad (2)$$

Where,

f_0 (a, x, p) = Initial guess,

$\sum_{n=1}^{N} f_n\,(a, x, p)$= Incremental boosts on a,

x and p = Shown as the split points of the predictors and the weight gain in the no-nodes.

2. fastText

fastText is an open source library for efficiently learning sentence classifications. Classification accuracy is said to be parallel to deep learning classifiers, which means char-CNN. In both the testing phase and training, there is a significant increase in VDCNN and char-CRNN. Each input sentence a_1, a_2... a_M specifies as M bag of m-gram features to buy some information about the local word order.Here the features are median, to create hidden and embedded variable. To calculate the probability distribution in the target classes, a SofMax function is used.

3. Attention-Based CNNs

The focusing method allows a neural network text classification model to attend to different parts of the input. Each of the entries unit, whether, word or character. From the point of view of the model, there is another emphasis on learning attention. Two-level attention-grabbing hierarchical networks, it is now sophisticated classifiers for document classification. Its two-level structures are both sentence representation and deep word representation; it has the ability to learn both. The hierarchical focus is different from the Networks, in a large-scale product title classification, the focus CNN algorithm is used here to achieve the best accuracy. Instead of using a two-stage focus framework, to learn the representation of the input sentence, the first uses attention to a context. Second, the release of the previous volume, then in the second attention seat, the environment acts as a vector. Then to predict the output class, The SofMax layer and the fully fused layer should be combined.

IV.RESULTS AND DISCUSSION

This paper, in question classification tests, it takes two different settings: One is a single label classification and the other is a multiple label classification. It sets questions with more than one label to remove ambiguity from the data. This reduced the test set into 78,053 questions and the training set into 234,157 questions, and decreases the number of labels as 1,406. The training package now includes 302,511 questions and it also includes a test set of 100,838 questions with 2,085 target types.

Many features are extracted at the level of tokenized spelling, word, and level for SVMs, XGBoost, and logistic regression. Here, the number of terms bi-gram and uni-gram is used at the word level with the word bi-gram and uni-gram, extracting the uni-gram to four-grams with their frequency counts for the character level. Feature engineering does not need to focus CNN and fastText as input source text.

1. Single-Label QC

For each text classification, the best-micro F1 score is shown in table1 using different predictive algorithms. For one classifier, only one time condition prediction is performed. The evaluation in different ways is primarily concerned with taking the best k positions out of the entire breadcrumb.

Table 1: Best micro-F1 score of multi-class single-label

Level	GBT	fastText	ACNN	LR	SVM
1	0.77	0.78	0.81	0.84	0.88
2	0.66	0.62	0.71	0.74	0.78
3	0.59	0.56	0.65	0.68	0.73
4	0.56	0.52	0.61	0.65	0.69
5	0.55	0.51	0.50	0.65	0.69
Leaf	**0.54**	**0.51**	**0.59**	**0.64**	**0.68**

In the single label QC system, the linear SVM algorithm has the highest micro-F1 score. One real reason may be related to the larger feature, this is more than three times the number of training events, but much lower. To get more accuracy in this position, the linear kernel should be sufficient, and there is no need to map data to higher dimensional space. This makes it the best classifier for query classification. It is not advisable to handle a large number of features.Then as shown in figure 3, given that the data is unbalanced in a system; the GBT algorithm implies weak performance.

Table 2: Micro-F1 for multi-class single-label SVMs classifier

Query length threshold	0	1	2	3	4
Questions	78,053	69,243	43,227	23,892	11,198
Level	Micro F1 Score				
1	0.88	0.90	0.92	0.95	0.97
2	0.78	0.80	0.83	0.86	0.88
3	0.73	0.75	0.78	0.81	0.84
4	0.69	0.72	0.74	0.77	0.79
5	0.69	0.72	0.74	0.76	0.79
Leaf	**0.68**	**0.70**	**0.72**	**0.74**	**0.77**

In the text classification process, input text is more information that can be learned from a classifier. And how the length of the query affects the performance of a text classifier can be found in a table 1, the SVM process filters out questions that are less than a certain threshold, the first row in Table 2 indicates the value of the filtration threshold, threshold equal to 1 means that all queries of equal length or less will be discarded from the data. The micro-score of each algorithm increases as the threshold rises.

2. Multi-label QC

Multiple label classification is more efficient when compared to a single label query classification system, and has been used so much in real applications. Based on the results of the single label test, this paper selects the best classification, that is, in table 3; the SVM algorithm is best shown.

Table 3: Scoreof Micro-Precision, Micro-Recall and Micro-F1

Level	Multi-label SVM only			Single label + Multi-label SVM		
	P	R	F1	P	R	F1
1	0.94	0.71	0.80	0.94	0.91	0.92
2	0.83	0.62	0.71	0.84	0.79	0.81
3	0.77	0.58	0.66	0.78	0.72	0.75
4	0.73	0.55	0.62	0.74	0.68	0.71
5	0.72	0.54	0.62	0.73	0.67	0.70
Leaf	**0.72**	**0.54**	**0.61**	**0.73**	**0.67**	**0.70**

3. Accuracy of Classification

In this paper, the SVM method used is more accurately classified than other algorithms (GBT, fastText, ACNN, LR). The information ratios for this, i.e. accuracy rates, are shown in the figure above.

Here, the classification accuracy is a very correctly classified and quality of classification.

$$\text{Accuracy} = \frac{\text{No.of correct classifications}}{\text{Total no.of classifications}}$$

To calculate the classification accuracy, it is calculated on a negative and positive basis.

$$\text{Accuracy} = \frac{TN+TP}{TN+TP+FN+FP}$$

Where,

TN = True Negatives,
TP = True Positives,
FN = False Negatives,
FP = False Positives.

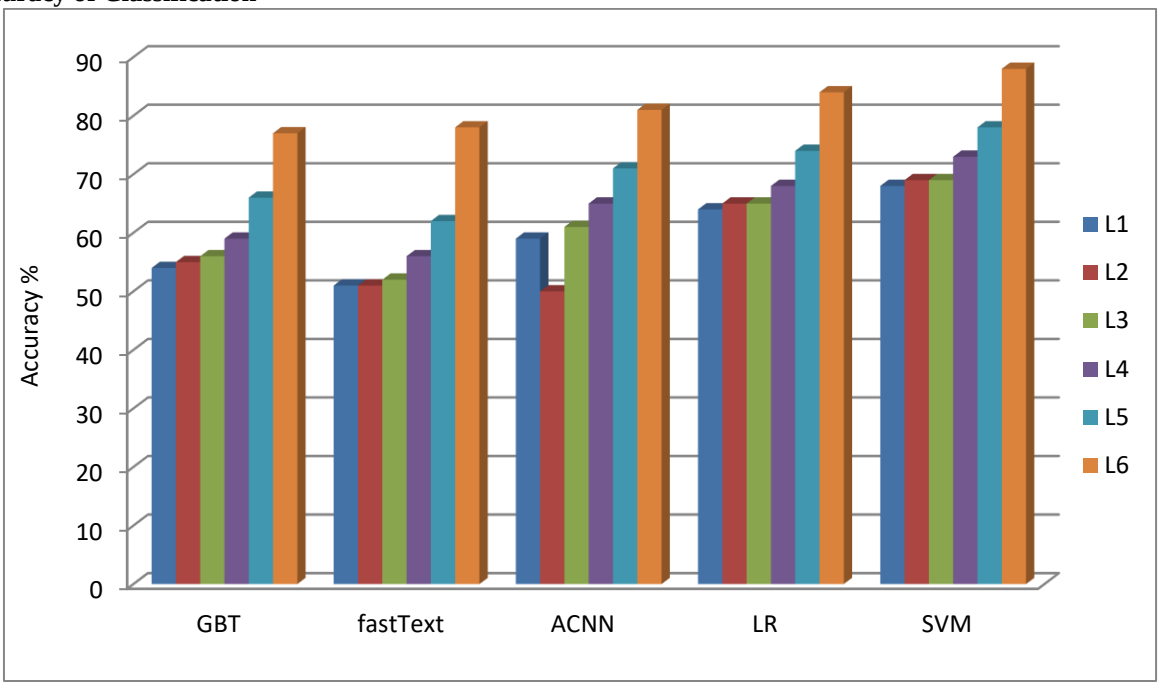


Fig.4: Accuracy Comparison

Accuracy in question classification depends on correctly categorized questions with the best values. The classification accuracy of the proposed and existing methods is classified in Figure 4. Based on customer reviews, in many assorted products, there are questions based on how many categorized and revised best estimates, depending on the accuracy of the classification system. The proposed method is more accurate than the existing methods. From this comparative accuracy conclusion, the proposed system, it has been proven that the best ratings that are adjusted and categorized will yield higher values based on questions.

V.CONCLUSION

Due to the annotation of a highly valued man with vague and short questions lack of test and training data is the two main function of query classification. In the field of e-commerce, to collecting customer feedback, this paper used an unsupervised method. For this purpose, this paper took 2,085 types and 403,349 questions. Then, the best and most sophisticated text classifiers are used as single and multiple labels. Due to the narrow nature of the questions at this work, deep learning models are affected. And it controls the neural network for learning deeper explanations. Finally in this paper, the most effective text classifier is the SVM algorithm. The next goal is to combine query expansion methods to take context information into account, and improve short queries.

REFERENCES

1. Sherry Singh, Shailja Madhwal, "Modelling Search Habits on E-commerce Websites using Supervised Learning", 2018 IEEE 8th International Advance Computing Conference (IACC), 18 April 2019.
2. Sihang Jiang, Jiaqing Liang, "Towards the Completion of a Domain-Specific Knowledge Base with Emerging Query Terms", 2019 IEEE 35th International Conference on Data Engineering (ICDE), 06 June 2019.
3. Leonidas Akritidis, Athanasios Fevgas, "Effective Products Categorization with Importance Scores and Morphological Analysis of the Titles", 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI), 17 December 2018.
4. Yin-HsiKuo, Winston H. Hsu, "Feature Learning with Rank-Based Candidate Selection for Product Search", 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 23 January 2018.

5. Prajakta Ghavare, Prashant Ahire, “Big Data Classification of Users Navigation and Behavior Using Web Server Logs”, 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 25 April 2019.
6. Amrita Saha, MeghaNawhal, “Learning Disentangled Multimodal Representations for the Fashion Domain”, 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 07 May 2018.
7. Qing Zhu, “Trust Service Discovery by Opinions Classification on Virtual Communities”, 2010 Second International Conference on Networks Security, Wireless Communications and Trusted Computing, 07 June 2010.
8. Baoqiu Wang, Yukun Zhong, “An unsupervised method for information retrieval on E-commerce websites”, Proceedings of 2nd International Conference on Information Technology and Electronic Commerce”, 14 May 2015.
9. Samaneh Nadali, Masrah Azrifah Azmi Murad, “Sentiment classification of customer reviews based on fuzzy logic”, 2010 International Symposium on Information Technology, 02 September 2010.
10. Donglin Chen, Xiaofei Li, “User-oriented intelligent service of e-catalog based on semantic web”, 2010 2nd IEEE International Conference on Information Management and Engineering, 03 June 2010.