Building a System based on Intelligent Agencies for Assisting in Classification and Circulation of Books and Periodicals Automatically

M. E. ElAlami
Computer Science Department
Faculty of Specific Education
Mansoura University

Hosnia M. M. Ahmed
Computer Science Department
Faculty of Specific Education
Mansoura University

Mahy E. Elemam
Computer Science Department
Faculty of Specific Education
Mansoura University

ABSTRACT
This paper presents an intelligent system to assist librarians in classification and circulation of books and periodicals. The system is based on a database containing the categories of scientific departments in the Faculty of Specific Education. It is also contains knowledge base that includes most of the words related to these categories. The Distance Vector Classifier Engine is used for text pre-processing classification. K-Nearest Neighbor algorithm and Term Frequency – Inverses Document Frequency are used for classification books and periodicals. The proposed system has achieved accuracy reached to 94.3%.

Keywords
Artificial Intelligent, Intelligent Systems, Text mining, Classification of Books

1. INTRODUCTION
The library is a structured summation, accessible by a determined society for reference or borrowing [1]. Classification has imposed important meanings in our life. In general, the definition of classification means grouping together like things according to common qualities [2]. Classification in libraries is the arrangement, organization and division of books and educational materials into groups according to a specific scientific content ,such as 'Languages', 'Education', 'History', and 'Business', 'Travel ' … et al [3]. There are many bibliographic classification systems, like the Dewey Decimal Classification (DDC), Library of Congress Classification (LC), and so on [4]. It was noted that the contents of the library are in urgent need to be organized in a technical way for readers to obtain the required books easily [5]. Intelligent information agents are computer programs that exist in some environments to retrieve information independently and adaptable [6]. There are many systems used in classification of books such as Naïve Bayes (NB), and Neural Networks (NN), K nearest neighbor (KNN), and Support Vector Machines (SVM) [7].

2. RELATED WORK
There are many studies have been applied in classification of books and periodicals. Chen, Shufeng has proposed a method to improve the representative model. It was depended on CURE (Clustering Using REpresentatives) and QKNN (Quick k-nearest neighbor) algorithms. Empirical results showed that this method could effectively decrease the number of samples and accelerate the search for the k nearest neighbor samples [8].

S.N. Bharath Bhushan, et .al has proposed an active likeness be made to calculate the approximation of two sets of text documents. Also, a similar pattern compression standard for text documents was suggested. It was relied on choosing a representation model and choosing a similarity scale and choosing learning algorithms. Empirical results showed that the f -measure score acquired from suggested likeness metric is better than the existing state of the art technique [9].

Yang, Lianbao, et .al had proposed a new smart classification sample to tackle the imbalanced error text data. It was based on SMOTE (Synthetic Minority Over-sampling TEchnique) algorithm. It was used to generate the fault data randomly for making the data balanced. Result showed that the sample has significantly improved the error classification [10].

Jorge Gonzalez-Lopez, et. Al strategies to distribute Ml-KNN (Multi-Label k-Nearest Neighbor) over Spark had been offered and estimated. It was based on three strategies: Brute force, tree-based index, and locally sensitive fragmentation. The effect of these strategies in Ml-KNN had been deliberated into separate in view of multiple measures. The results indicated that the tree-based indexing strategy is superior to other methods, with a speed of 266x for the largest dataset [11].

Latif, Syukriyant, et .al had aimed to classify summary content of English language magazines. It was built using NB (Naive Bayes) algorithms. The feature chosen operations have been used term weighting to give weight to each word. Results displayed that the better results of the classification during the practice data test were acquired by 75% [12].

June-Jei Kuo two layered book classification system had been proposed. The first layer had based on SVM (Support Vector Machine) and NB (Naive Bayes). Further, the decision tree had been used in the second layer. It was based on summary and table of contents for automatic classification books. The strategy has demonstrated its ability to perform better than traditional classification using a single classifier [13].

Guo Pengfei, et.al the robot model has been proposed in the classification of books based on a multi-agent. The intelligent robot could perceive the barcode of books automatically. The result showed that would improve the efficiency of the library management. It also reduced the hard work of the librarian [14].

3. THE PROPOSED SYSTEM

The main challenges that meet researchers in the classification of books and scientific journals are the lack of dataset. There is no public or standard dataset published over the internet for this type of problem, so the dataset was collected manually. 7000 words were collected in all specializations in the following scientific departments (Computer Science - Home Economics - Educational Media - Art Education, and finally Music Education). These words were collected from August 2017 to March 2018. Some of these words were collected from libraries and search engines such as (Yahoo, Google, Bing, etc.).

The following table shows some examples of categories of scientific specialties.

Table 1: Examples of Categories of Scientific Specialties

Scientific Specialties	Categories
Computer Science	- Operating systems - Programming languages - Computer Security - Databases
Home Economics	- Nutrition and food science -Home management -Textile and Clothing
Music Education	- Composition and music theory - Arab Music - Solfege and Rhythm exercises
Art Education	- Wood work - Sculpture - Printing textile
Education Media	- School Journalism - Educational radio and television - Educational theatre

The following table shows some examples of words for each category.

Table 2: Examples of Words for Each Category

Categories	Words
Computer Science	Decision Tree – Solutions – Algorithm – adversary – Developer – Access – Browser – Chrome – Cloud – BASIC – Assembler – ASCII – Branch – Char - C++ .
Home Economics	Anorectic – Bariatric – Primatology – cook – Roughage – appliance – caregivers – Corruption – decision – Camouflage – creation – cloth – Color – impact.
Music Education	Music – Arabic – Drum – Harmonic – Rhythm – Scales – agonic - Arrhythmic – duration – dynamic – Guitar – accelerator – Band – Mandolin.
Art Education	Bead – Cabinet – Bolster – Cluck – Crotch - Crook – Grain – Aperture – Bitmap – Compression – Draw – Color- plasticize – Sculpture.
Education Media	Conversation – Journal – Newsroom – Write – documentaries – Drama – Educator – Reader – Radio – Theatre – Television.

The framework for the proposed system is shown in figure 1

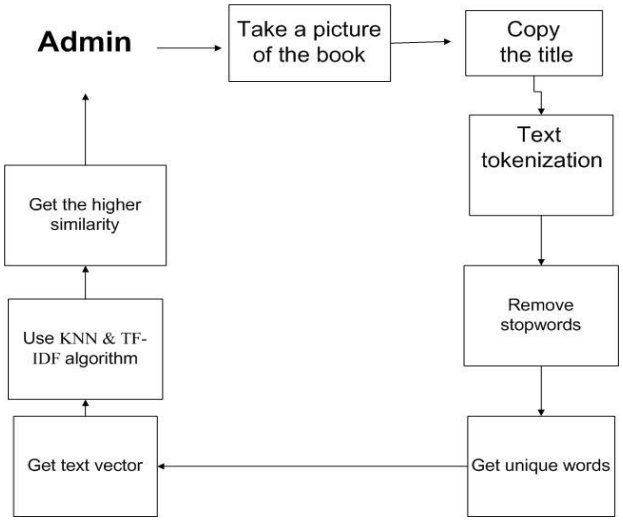

Fig1: The Framework for the Proposed System

In the following part the proposed system steps will be illustrated:

- The librarian (Admin) takes a picture of the book to be classified through the program.
- A copy of the title for book OCR (Optical Character Recognition) is taken from the proposed system.
- Feature are extracted through the following steps

1. Tokenization: A title of book is treated as a string, and is broken down into a list of symbols.
2. Removing stop words: Stop words like "the", "a", "and"… and so on, are much happening. Therefore, non-important words should be removed.
3. Stemming word: Applying the stemming algorithm that converts different word form into similar canonical form. This step is the process of combining icons with their own root shape, e.g. connection to connect, computing to calculate etc.

- The system deletes repeated words and gets unique words only from the text.
    - After Removing stop words and getting unique words from the text, the title vector will be applied by checking each word in the book and periodical Title, with words in each category; if the word in the text existed in the mentioned category or not be existed finally the book and Periodical vector will be gotten.
    - By using K-NN the distance between two points in the plane with coordinates p=(x, y) and q= (a, b) can be calculated as shown in equation (1). After calculating this, the similarity between these vectors will be gotten [15] .

$$\boldsymbol{d(p,q) = d(q,p) = \sqrt{(x-a)^2 + (y-b)^2}} \quad (1)$$

- The higher similarity is gotten and this means that category with the higher similarity included the classified book and Periodical.

The process of KNN algorithm to classify sample X is as follow [16] :

- Suppose there are j training categories $C_1, C_2, \ldots, C_j$ and the sum of the training samples is N after feature reduction, they become m-dimension feature vector .
- Make sample X to be the same feature vector of the form $(X_1, X_2, \ldots, X_m)$, as all training samples .
- Calculate the similarities between all training samples and X. Taking the i^{th} sample di $(d_{i1}, d_{i2}, \ldots, d_{im})$ as an example, the similarity SIM(X, d_i) is as following :

$$SIM(X, di) = \frac{\sum x.di}{\sqrt{(\sum_{i=1}^{m} xi)^2 \ (\sum_{i=1}^{m} di)^2}} \quad (2)$$

- Choose k samples which are larger from N similarities of SIM(X, di), (i=1, 2,…, N), and treat them as a KNN collection of X. Then, calculate the probability of X belong to each category respectively with the following formula.

$$\boldsymbol{p(x, cj) = \sum_d SIM(x, di).y(di, cj)} \quad (3)$$

Where

$y(d_i, C_j)$ is a category attribute function, which satisfied:

$$y(di, cj) = \begin{cases} 1, di, \in c\ j \\ 0, di, \notin cj \end{cases} \quad (4)$$

- Judge sample X to be the category which has the largest $P(X, C_j)$.

4. APPLICATIONS AND RESULTS

The proposed system is designed using PHP and JavaScript languages. KNN algorithm and TF-IDF method are used to get high similarity between the proposed categories. The proposed system can be accessed through the World Wide Web at http://mahy.onlinewebshop.net/moha/login. Also it can be used through Personal Computer (PC). Running the proposed system on PC requires the availability of the XAMPP software to convert the PC to server.

The proposed system graphical user interface is shown in figure 2. It is divided in to six categories namely; Users, category, keyword, books, periodicals and logout. All classified books and periodical are displayed as well as the category to which they belong. The books and periodical data can be modified or deleted.

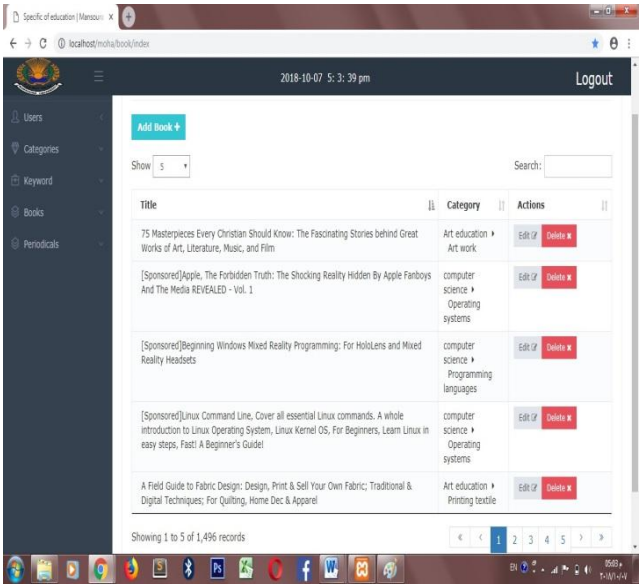

Fig.2: The Proposed System Graphical User Interface

Add new book to be classified in the proposed system is shown in figure 3. The title of the book is added. By clicking on a classify command, the field to which the book is categorized appears. All information about the book is entered.

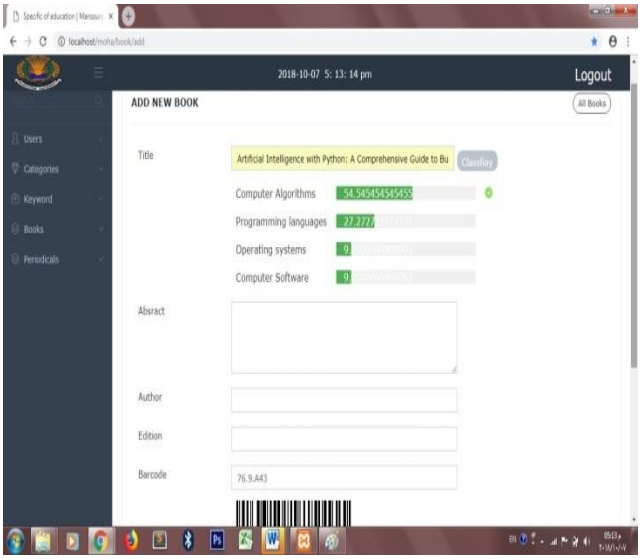

Fig.3: Add New Book to be classified in the Proposed System

Edit information about book in the proposed system is shown in figure 4. By clicking on the Edit button, this window appears, in which information about the book, such as its title, author name and more, is modified. It also enables us to reclassify again.

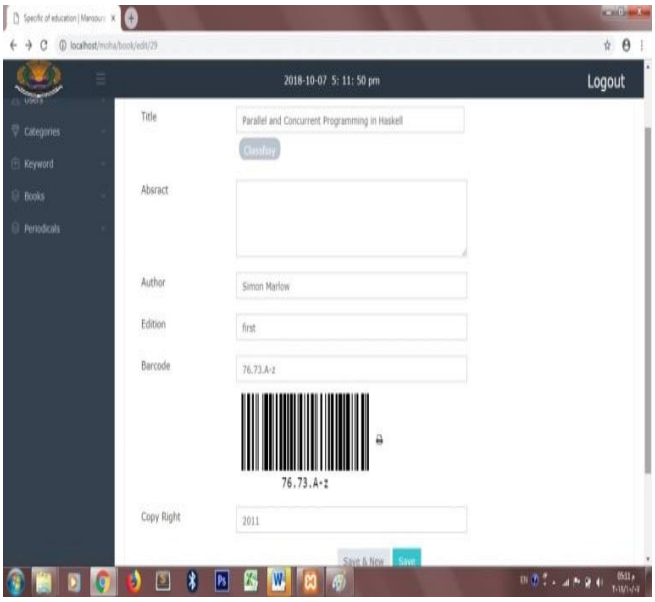

Fig .4: Edit Information about Book in the Proposed System

Parent category of Scientific Specialties which is divided in to five categories is shown in figure 5. This screen contains the mains as well as their label code. Within each major category are the sub-categories.

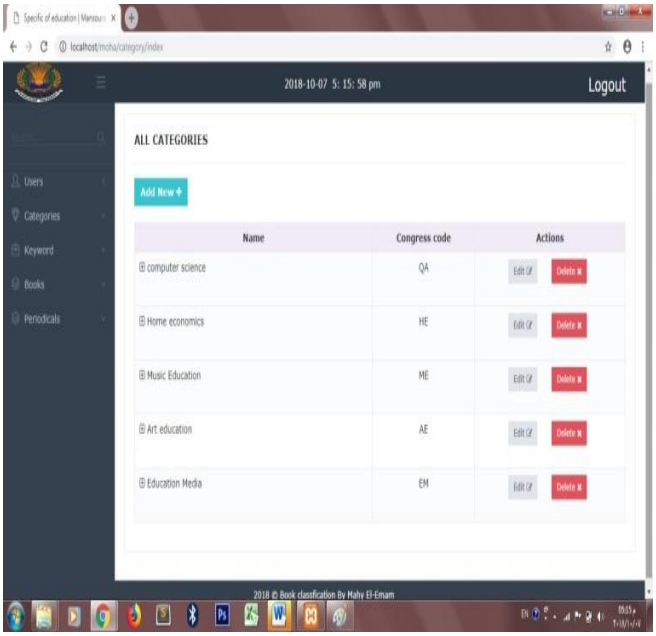

Fig. 5: Parent Category of Scientific Specialties

In the proposed system, there are about one hundred four hundred and ninety-five books, and also about fifty periodicals have been tested. In this section, Comparison between the basic classification and the proposed system classification of books and periodicals is shown in table 3

.Table 3: Comparison between the Basic Classification and the Proposed System Classification

Category	Basic Classification	The proposed system Classification
Computer Science		
programming books	147	121
Software Engineering	114	103
Security & Cryptography	101	100
Databases	87	85
Simulation	101	100
Internet	62	61
Operating systems	64	63
Computer Algorithms	115	113
Home Economics		
Nutrition and food science	33	32
Home management	34	33
textiles and clothing	38	37
Music Education		
Theories and Composition of Music	32	31
Arabic Music	37	34
Solvege and motor rhythm	36	34
The Performance	37	36
Art Education		

Metal work	34	33
Wood work	55	52
Sculpture	48	47
Ceramics	48	45
Painting	57	54
Weaving and textile	68	60
Printing textile	48	45
Art work	45	41
Education Media		
School Journalism	46	43
Educational radio and television	48	42
Educational theatre	51	50
Periodicals		
Periodicals	55	50
Sum	**1641**	**1545**

The performance measures used for the evaluation of classification books are precision, recall and F-measure, as shown in equations (5-7) respectively [17-18].

$$\text{Precision} = \frac{TP}{TP+FP} \quad \ldots\ldots\ldots\ldots \quad (5)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (6)$$

$$\text{F} - \text{measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad \ldots\ldots \quad (7)$$

Recall, precision and f-measure for classification books using the proposed system in Computer Science are shown in table 4 and figure 6.

Table 4: Recall, Precision and F-measure for Classification Books in Computer Science

Category	Recall	Precision	f- measure
programming books	0.83	0.88	0.86
Software Engineering	0.91	0.87	0.89
Security & Cryptography	0.99	0.86	0.92
Databases	0.98	0.85	0.91
Simulation	0.99	0.86	0.92
Internet	0.98	0.79	0.85
Operating systems	0.98	0.8	0.88
Computer Algorithms	0.98	0.88	0.93

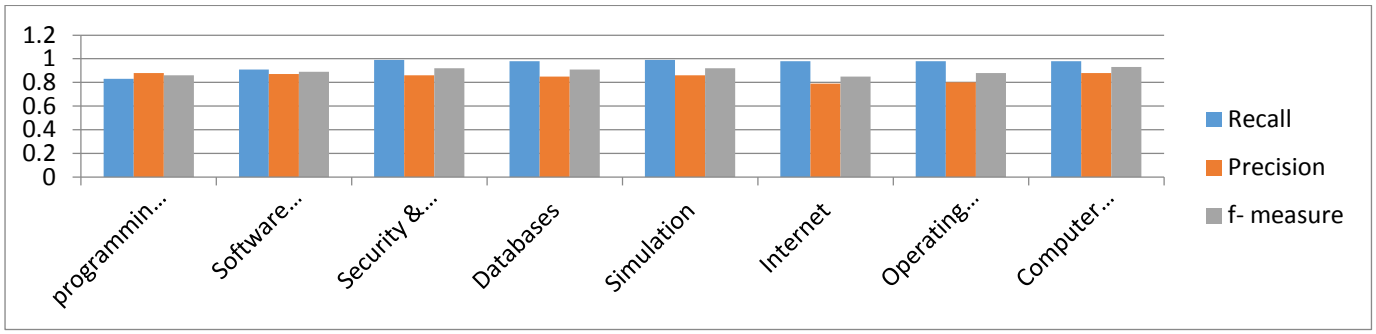

Fig. 6: Recall, Precision and F-measure for Classification Books in Computer Science

Recall, precision and f-measure for classification books using the proposed system in Home Economics are shown in table 5 and figure 7.

Table 5: Recall, Precision and F-measure for Classification Books in Home Economics

Category	Recall	Precision	f- measure
Nutrition and food science	0.97	0.67	0.8
Home management	0.97	0.68	0.8
textiles and clothing	0.97	0.7	0.82

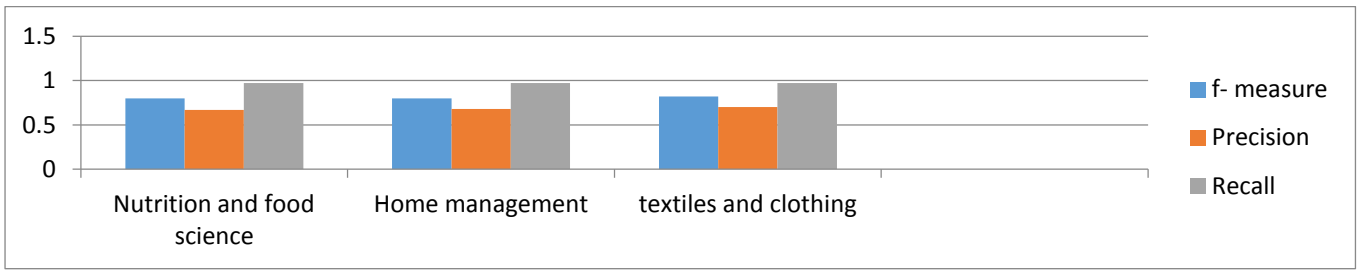

Fig. 7: Recall, Precision and F-measure for Classification Books in Home Economics

Recall, precision and f-measure for classification books using the proposed system in Music Education are shown in table 6 and figure 8.

Table 6: Recall, Precision, and F-measure for Classification Books in Music Education

Category	Recall	Precision	f- measure
Theories and Composition of Music	0.97	0.66	0.79
Arabic Music	0.92	0.69	0.79
Solvege and motor rhythm	0.95	0.69	0.8
The Performance	0.97	0.7	0.8

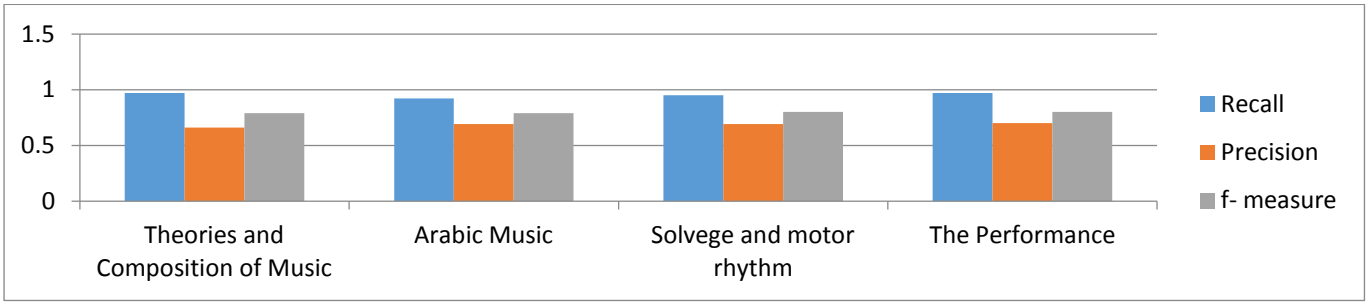

Fig. 8: Recall, Precision, and F-measure for Classification Books in Music Education

Recall, precision and f-measure for classification books using the proposed system in Art Education are shown in table 7 and figure 9.

Table 7: Recall, Precision and F-measure for Classification Books in Art Education

Category	Recall	Precision	f- measure
Metal work	0.97	0.68	0.8
Wood work	0.95	0.77	0.85
Sculpture	0.98	0.75	0.85
Ceramics	0.94	0.74	0.83
Painting	0.95	0.78	0.86
Weaving and textile	0.89	0.79	0.84
Printing textile	0.94	0.74	0.83
Art work	0.92	0.72	0.81

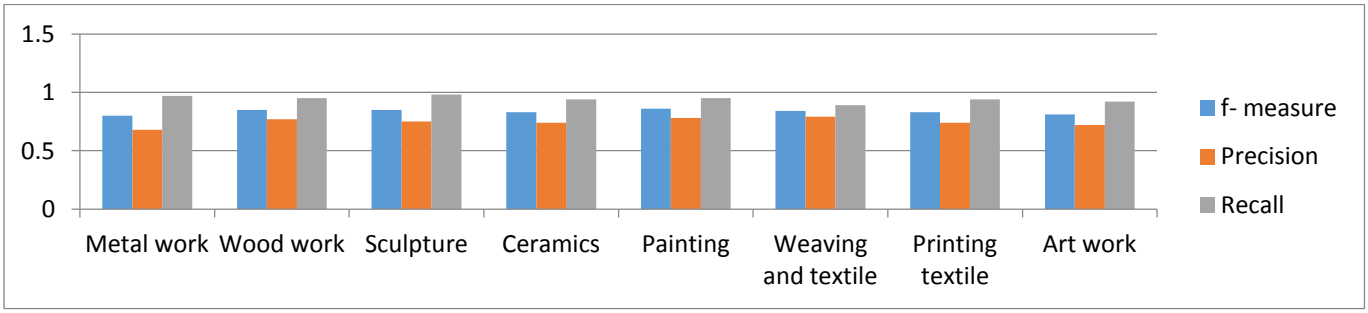

Fig. 9: Recall, Precision and F-measure for Classification Books in Art Education

Recall, precision and f-measure for classification books using the proposed system in Education Media are shown in table 8 and figure 10.

Table 8: Recall, Precision and F-measure for Classification Books in Education Media

Category	Recall	Precision	f- measure
School Journalism	0.94	0.73	0.83
Educational radio and television	0.88	0.73	0.8
Educational theatre	0.98	0.76	0.86

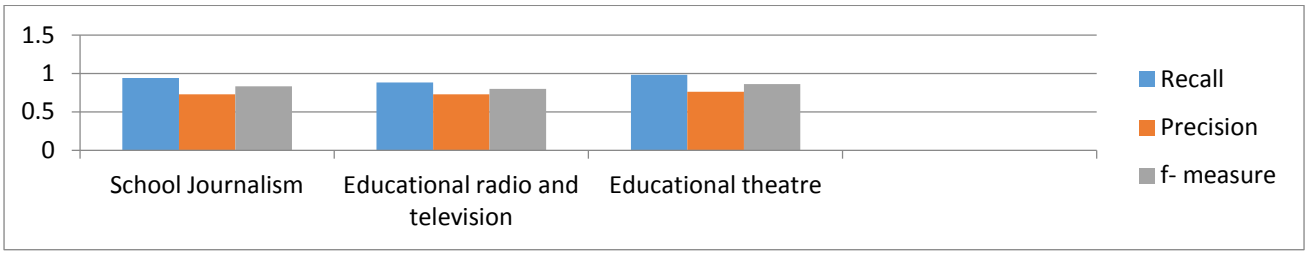

Fig. 10: Recall, Precision and F-measure for Classification Books in Education Media

Recall, precision and f-measure for classification Periodicals using the proposed system are shown in table 9 and figure 11.

Table 9: Recall, Precision and F-measure for Classification Periodicals

Category	Recall	Precision	f- measure
Periodicals	0.95	0.77	0.85

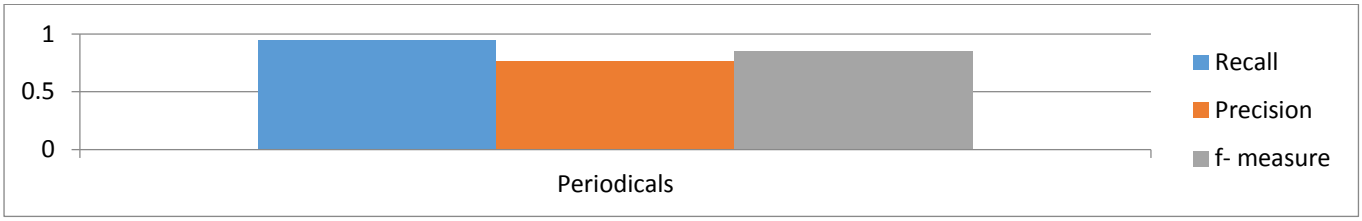

Fig. 11: Recall, Precision and F-measure for Classification Periodicals

The accuracy of the proposed system

$$= \frac{\text{Number of corrected classified books}}{\text{Total number of books}} * 100$$

$$\text{The classification accuracy} = \frac{1545}{1641} * 100 = 94.3\%$$

This means that the proposed system can classify books with satisfying results.

5. CONCLUSIONS AND FUTURE WORK

This paper used a novel way for classification of books and Periodicals. It uses KNN, and TF-IDF method for building the proposed system. The results show that the proposed system achieves high percentage reached to 94.3 for books classification. Results could be amended in future work by applying the proposed system on various languages like Arabic.

6. REFERENCES

[1] Dewa Gede Hendra Divayana, I Made Sugiarta ,et.al (2015),"Digital Library of Expert System Based at Indonesia Technology University", International Journal of Advanced Research in Artificial Intelligence, Vol. 4, No.3.

[2] Cabonero, David A. and Dolendo, Russell B., (2013). "Cataloging and Classification Skills of Library and Information Science Practitioners in their Workplaces: A Case Analysis" Library Philosophy and Practice (e-journal). PP.958- 960.

[3] Guo Pengfei,Du Liangxian,Qi Junxia (2012) , "ALibrary Book Intelligence Classification System based on Multi-agent " ,PP .2187 – 2193.

[4] Fadaie Araghi, Gholamreza (2004). "A New Scheme for Library Classification. Cataloging & Classification Quarterly", Vol. 38, PP. 75-99.

[5] SUMAN (S) and KARMARKAR (Debanshu). (2002) "The role of library classification in organizing the web". In Workshop on information resource management. PP . 1-3.

[6] Mohammed Abbas Kadhim . et .al (2016), " A Multi-intelligent Agent System for Automatic Construction of Rule-based Expert System " ,Vol. 9,PP. 62-68.

[7] Sholom, W., Nitin, I., Tong, Z., Fred , D. (2005),"Text mining: predictive methods for analyzing unstructured information" , New York , Springer.

[8] Shufeng Chen (2017), "K-Nearest Neighbor Algorithm Optimization in Text Categorization" , IOP Conf. Series: Earth and Environmental ScienceVol.108.

[9] S.N. Bharath Bhushan *, Ajit Danti (2018), "Classification of compressed and uncompressed text documents" , Future Generation Computer Systems , Vol. 88 ,PP. 614–623.

[10] Lianbao Yang ,et.al (2018), "Intelligent classification model for railway signal equipment fault based on SMOTE and ensemble learning" , IOP Conf. Series: Materials Science and Engineering ,Vol. 383

[11] Jorge Gonzalez-Lopez ,et .al (2018), "Distributed nearest neighbor classification for large-scale multi-label data on spark" , Future Generation Computer Systems ,Vol. 87 ,PP. 66–82.

[12] Syukriyanto Latif ,et.al (2018)," Content Abstract Classification Using Naive Bayes" , Journal of Physics: Conf. Series 979.

[13] June-Jei Kuo (2014) ," An Automatic Library Data Classification System Using Layer Structure and Voting Strategy" , Springer International Publishing Switzerland , LNCS 8839, PP. 279–287,

[14] Guo Pengfei,et.al (2012) ," The Application of Semantics Web in Digital Library Knowledge Management" , International Conference on Applied Physics and Industrial Engineering ,vol. 24 ,PP. 2180 – 2186.

[15] Xindong Wu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh (2008) , "Top 10 algorithms in data mining", Knowl Inf Syst , Vol. 14, PP. 1–37

[16] Sadegh Bafandeh Imandoust And Mohammad Bolandraftar ,(2013), "A k-Nearest Neighbor Based Algorithm for Multi-label Classification ",Journal of Engineering Research and Applications , Vol. 3, Issue 5, PP.605-610.

[17] Boris Neubert, Sören Pirk,et .al (2010) , "Precision and Recall as Appearance Space Quality Measure for Simplified Aggregate Details", Eurographics Symposium on Rendering.

[18] David M.W. Powers,(2014)," What the F-measure doesn'measure",https://www.researchgate.net/publication/273761
233_What_the_Fmeasure_doesn%27t_measure.,availabl e at date .10/5/2018 , time.12.50 PM.