AUTOMATED CATEGORIZATION SCHEME FOR DIGITAL LIBRARIES IN DISTANCE LEARNING: A Pattern Recognition Approach

Dr. Serkan GUNAL
Faculty of Engineering and Architecture
Department of Computer Engineering
Anadolu University, Eskisehir, TURKIYE

ABSTRACT

Digital libraries play a crucial role in distance learning. Nowadays, they are one of the fundamental information sources for the students enrolled in this learning system. These libraries contain huge amount of instructional data (text, audio and video) offered by the distance learning program. Organization of the digital libraries is therefore very important for easy and fast access to the desired information. Improper categorization of data may mislead the students searching the library. Since manual categorization of huge amount of data might be challenging, an automatic and reliable method is needed. In this sense, this paper proposes an automated categorization scheme for digital libraries in distance learning. The categorization scheme is designed and developed by a pattern recognition approach. Effectiveness of the proposed scheme is evaluated on widely used *Reuters* database. The results of the experimental study verify that the proposed scheme is a good candidate for categorization of digital libraries in distance learning programs.

Keywords: **Pattern recognition, text categorization, digital library, distance learning.**

INTRODUCTION

The recent advances in the field of communication and internet technology have created a new aspect in education called as distance learning. In this learning system, the students do not have to be physically inside a classroom and geographic distances have of no importance. One of the primary information sources for distance learning programs is digital library. Yunus Emre learning portal of Anadolu University [Anadolu University, 2008] and OpenCourseWare system of Massachusetts Institute of Technology [MIT, 2008] are just two good examples of the digital libraries in distance learning. Students enrolled in these programs can quickly retrieve instructional material (text, audio and video) from the digital libraries via web access [Lau, 2000]. In contrast to printed materials, this is probably the most efficient way of accessing data. Depending on the type of distance learning program and the number of lectures offered by the program, the amount of data inside the digital library might be quite high. Therefore, proper organization and categorization of the digital library is vital for easy and fast access to desired information. Improper categorization may mislead the students browsing the library. Since vast amount of data is of concern, manual categorization might be very time-consuming and unreliable. Consequently, automated categorization approaches are needed.

Automated categorization of data is actually a pattern recognition job. A pattern recognition system simply categorizes or classifies objects or events into an appropriate category or class [Duda et al., 2001; Theodoridis and Koutroumbas, 2003; Webb, 2002]. The objects or events are simply named as patterns. The patterns go through three major stages in the recognition process. These stages are feature extraction, feature selection and classification.

The patterns to be recognized correspond to the instructional materials available in the digital library of distance learning program. As mentioned before, instructional material can be in different forms such as text, audio and video. However, one can represent all these forms with text data. Text material is already in this form. Audio and video materials can be associated with a text-based description explaining the contents of those materials. In this way, all instructional data in any form can be represented with text. Thus, the patterns to be recognized are limited to text data, and the pattern recognition problem becomes actually a text recognition or text categorization problem for this study.

In the literature, plenty of work has been done related to text categorization or classification in different fields. Spam e-mail detection process is handled as a 2-category (legitimate and spam e-mail) text classification problem in [Gunal et al., 2006]. Web pages are classified with a text categorization approach in [Selamat and Omatu, 2004]. The study [Shang et al., 2007] performs text categorization on a digital news database. A survey discussing the text categorization approaches from machine learning perspective is also presented in [Sebastiani, 2002].

In this sense, this paper proposes a framework of automated data categorization for digital libraries in distance learning. As required by a pattern recognition system, a feature extraction mechanism for text based data is developed together with a feature selection strategy. Then, the selected features are fed into a pattern classifier to finalize the recognition process. In the experimental study, classical Reuter's database is used to simulate a digital library containing considerable amount of data on different topics. Experimental results indicate that the categorization performance obtained by the proposed scheme is pretty good, which makes it a good candidate for categorization of digital libraries in distance learning programs. Rest of the paper is organized as follows: Firstly, pattern recognition and its fundamental stages are briefly explained. Next, the proposed scheme of automated data categorization for digital libraries is introduced. Then, the experimental study is described and results of the study are given. Finally, the conclusion of the paper is presented.

PATTERN RECOGNITION

Pattern recognition is a multi-disciplinary subject covering the fields of statistics, engineering, artificial intelligence, computer science, psychology, physiology, etc [Duda et al., 2001; Theodoridis and Koutroumbas, 2003; Webb, 2002]. The aim of pattern recognition is to classify objects or events into an appropriate class or category. The objects or events are simply called as patterns. Speech signal, human face, retina, fingerprint, text message are just some examples of the patterns. Human beings are able to recognize certain patterns with their five senses: sight, hearing, taste, smell and touch.

However, computer-based automated pattern recognition systems are required when the human senses fail to recognize patterns, or if there is a need for automating and speeding up the recognition process. Recognition of text, image, speech, speaker, biomedical data are some well known applications of the pattern recognition. The importance of pattern recognition, which has been an active research area over 50 years, is constantly increasing together with emerging application fields.

The three fundamental stages of the pattern recognition process are feature extraction, feature selection and classification as shown in Figure: 1.

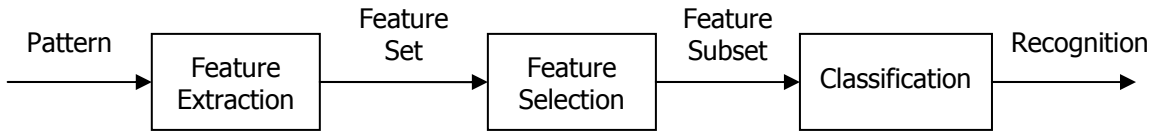

Figure: 1
Pattern recognition process

Feature Extraction

Feature is the measurable or observable data corresponding to pattern. Feature extraction eliminates redundant data and retrieves characteristic information about the pattern. Elimination of redundant information is of vital importance for the processing time of recognition process. If a pattern is represented by more than one feature, a *feature set* is of concern. A feature set having d features is represented with d - dimensional *feature vector*. The d - dimensional space $¡^{d}$ is called as *feature space* [Kuncheva, 2004]. Features belonging to patterns could be *quantitative* or *qualitative*. For instance, maximum speed information of a vehicle is quantitative feature while model information is qualitative. Statistical pattern recognition mostly deals with the quantitative features where syntactic pattern recognition uses the qualitative ones [Duda et al., 2001; Theodoridis and Koutroumbas, 2003; Webb, 2002].

Feature Selection

Feature extraction removes irrelevant data and retrieves characteristic information about pattern, as mentioned before. In this way, particular amount of dimension reduction is achieved and a feature set is obtained.

In the case of feature selection, a more discriminative subset of the regarding feature set is attained using some selection techniques. This process aims not only to increase dimension reduction rate but also to prevent the effect of *curse of dimensionality* [Jain and Zongker, 1997; Theodoridis and Koutroumbas, 2003].

Classification

For recognizing an unknown pattern, classification is carried out following the feature extraction and feature selection stages. In classification, a dataset consisting of a number of features, whose classes are previously known, goes through a training process.

A decision rule or mechanism is then employed at the end of the training. This approach, which uses a training set and some priori knowledge, is called as *supervised classification*.

In contrary, the classification approach which does not use any training set and priori knowledge is *unsupervised classification* [Duda et al., 2001; Theodoridis and Koutroumbas, 2003; Webb, 2002]. One should note that the quantitative features and supervised classification approach are used in this paper.

AUTOMATED CATEGORIZATION

Since automated categorization process is handled as a pattern recognition problem, where patterns correspond to texts, the stages of feature extraction, feature selection and classification should be all carried out.

The features in text recognition problem usually correspond to keyword frequencies, that is, the number of appearance of keywords in a given text. The keywords are special and potentially discriminative words for different subjects. Hence, a multi-dimensional feature vectors consisting of the values of keyword frequencies is obtained for the texts available in the digital library at the end of feature extraction stage. Here, it should be noted that all the letters in the texts are converted to small case before processing for case insensitivity.

Not all the words in a text can actually be discriminative. Since there are many common words (i.e., conjunctions, adverbs, pronouns, etc.) present in the texts from different categories, those words would have no contribution to discrimination and classification. For this reason, they can not be selected as the keywords and they should be eliminated. It is therefore a feature selection operation, the second stage of the pattern recognition process.

One way of selecting the keywords is to define them logically for specific categories; that is, keywords with high probability of appearance in a specific subject are selected. For instance, one can choose the keywords given in Table: 1 for a health related category.

Table: 1
Sample keywords for health related category

No	Keyword
1	hospital
2	doctor
3	nurse
4	pill
5	patient

This approach may work well for small number of categories. However, it would fail as the number of categories increases or if there is no prior information related to categories within the digital library. Therefore, a more methodical way is needed for defining the keywords. For this purpose,

- **Frequencies of all the words are computed among all documents for each category in the library. During this computation, irrelevant words are ignored and only the roots of words are considered by removing plural suffix for generalization (i.e., the word "computer" is considered when "computers" is encountered).**

- **Once the word frequencies are obtained for each category, the words with the highest frequencies are obtained.**
- **Among each category, the words with the *N* highest frequencies are selected as the keywords for the regarding category. The number *N* is defined empirically.**
- **The keywords from each category are then grouped together. After grouping, if there are still common keywords from different categories, they are ignored for preventing possible confusion in classification.**
- **Thus, *M x N* (or smaller if common words are found) final keyword set is obtained for all categories within the library where M indicates the total number of categories.**

Thus, frequencies of the selected keywords will be the elements of the multi-dimensional feature vector for a given text. The last attribute of the feature vector is defined as the size of text in bytes. This last information is used to measure the frequencies of the keywords relative to the text size.

Final feature vector structure for a given text is as shown in Figure: 2. After obtaining the feature vectors, they can be fed into the classifier.

Keyword#1 Freq.	Keyword#2 Freq.	...	Keyword#(*M x N*) Freq.	Text Size

Figure: 2
Structure of the feature vector for a given text

In classification, which is the final stage in the pattern recognition process, feature vectors, which are obtained among all the categories within the digital library, constitute a training set. This training set is then employed in a selected pattern classifier to obtain a decision criterion so that new instructional materials, whose categories are not defined yet, can be classified.

For this study, the selected classifier is Support Vector Machine (SVM). SVM is a kernel-based classifier and provides considerable recognition performance even for confusing and overlapping classes. It is therefore pretty suitable for automated categorization job.

Theory of SVM is not given here for preserving the focus of paper. The readers who are interested in this classifier are referred to [Schölkopf and Smola, 2001] for further details.

EXPERIMENTAL STUDY

In the experimental study, Reuters dataset [Hettich and Bay, 1999] is used for simulating a digital library. This dataset is a collection of news bulletin on a wide range of categories. For evaluating the proposed categorization scheme, 6 different categories, where each category has 500 texts, are selected from this dataset.

The selected categories are summarized in Table: 2. They are actually overlapping and confusable categories which makes the correct categorization difficult. Among 500 texts, 400 texts are used in training while remaining 100 texts are reserved for testing.

Table: 2
Selected categories of Reuters dataset for experimental work

No	Category	Description
1	acq	Acquisitions
2	crude	Crude Oil
3	earn	Earnings and Earnings Forecasts
4	grain	Grain
5	money-fx	Money/Foreign Exchange
6	trade	Trade

First, irrelevant words are defined logically as described in previous section. While defining these words, certain conjunctions, complements, adverbs, pronouns, letters and numbers are considered based on English grammatical rules.

All the irrelevant words defined for this study are listed in Table: 3.

This list can be updated or expanded with respect to the subjects of categories. Then, all categories are analyzed for feature selection as described in the previous section.

Thus, the keywords frequencies are obtained for each category and the keywords with the highest 25 frequencies are retrieved from every single category as listed in Table: 4.

Table: 3
List of irrelevant words

No	Word	No	Word	No	Word	No	Word
1	a	21	for	41	off	61	we
2	about	22	from	42	on	62	were
3	after	23	has	43	or	63	when
4	all	24	have	44	other	64	whether
5	also	25	he	45	out	65	which
6	an	26	her	46	said	66	while
7	and	27	him	47	she	67	will
8	are	28	his	48	so	68	with
9	as	29	i	49	than	69	would
10	at	30	if	50	that	70	you
11	be	31	in	51	the	71	0
12	been	32	into	52	them	72	1
13	before	33	is	53	then	73	2
14	but	34	it	54	therefore	74	3
15	by	35	its	55	they	75	4
16	can	36	may	56	this	76	5
17	did	37	me	57	to	77	6
18	does	38	might	58	until	78	7
19	done	39	not	59	want	79	8
20	each	40	of	60	was	80	9

Table: 4
List of keywords the highest 25 frequencies for each category

No	acq	crude	earn	grain	money-fx	trade
1	mln	oil	mln	tonnes	bank	trade
2	pct	mln	profit	wheat	dollar	dlr
3	company	price	share	grain	market	export
4	share	pct	oper	corn	mln	deficit
5	corp	crude	record	agriculture	pct	pct
6	merger	opec	sale	export	exchange	mln
7	usair	barrel	company	department	stg	surplus
8	stake	production	corp	crop	yen	import
9	bank	energy	pay	price	money	washington
10	board	price	earning	program	billion	state
11	acquisition	petroleum	stock	farmer	currency	foreign
12	shareholder	company	dividend	maize	rate	government
13	agreement	exploration	gain	production	central	tariff
14	sell	industry	tax	trade	japan	market
15	cash	market	income	market	dealer	good
16	york	reserve	operation	official	foreign	economic
17	american	demand	extraordinary	government	paris	world
18	twa	minister	split	acres	trade	official
19	management	tax	payable	rice	london	minister
20	sale	saudi	york	winter	intervention	reagan
21	acquire	government	calif	barley	japanese	bill
22	securities	foreign	gain	bill	system	tokyo
23	exchange	gulf	revenue	season	bill	meeting
24	price	ecuador	shareholder	acreage	major	industry
25	unit	corp	credit	certificate	england	house

Among this list, the keywords common to at least two categories are detected (Table: 5) and eliminated for improving discrimination. Hence, final keyword list with 90 members are obtained and feature vectors of the texts in the dataset are extracted as visualized in Figure: 2.

Table: 5
List of keywords common to at least two categories in Table: 4

No	Word	No	Word	No	Word	No	Word
1	bank	7	gain	13	official	19	shareholder
2	bill	8	government	14	pct	20	tax
3	company	9	industry	15	price	21	trade
4	corp	10	market	16	production	22	york
5	exchange	11	minister	17	sale		
6	foreign	12	mln	18	share		

Following the feature extraction, feature vectors are fed to SVM classifier and recognition accuracy for each category is computed.

The correct recognition rates for 100 texts reserved for testing are given in Table: 6. Average recognition rate among all categories is almost 90% which is a good accuracy considering that the selected categories are overlapping and confusing for discrimination.

Table: 6
Recognition accuracies for each category

Category	Recognition Rate (%)
acq	81,00
crude	98,00
earn	86,00
grain	95,00
money-fx	89,00
trade	88,00
Average	90,00

CONCLUSIONS

Automated categorization of digital libraries, which has become vital information source in distance learning programs due to rapid developments in the field of communication and internet technology, is addressed in this paper.

A pattern recognition approach consisting of feature extraction, feature selection and classification stages is used to implement the proposed scheme. Extensive testing of the proposed automated categorization scheme on a widely used database provides promising results for even overlapping and confusable categories.

Therefore, this scheme can be adapted and used in digital libraries of distance learning programs successfully.

BIODATA and CONTACT ADDRESS of AUTHOR

The author received B.S., M.S. and Ph.D. degrees in Electrical and Electronics Engineering from Eskisehir Osmangazi University, Eskisehir, Turkey in 1999, 2003 and 2008, respectively. He worked as a Research Assistant in the same university between 1999 and 2001. He then spent several years as an R&D engineer in technology companies. Currently, he is with the Department of Computer Engineering, Anadolu University, Eskisehir, Turkey. His main research areas include pattern recognition and digital signal processing.

Dr. Serkan GUNAL
Faculty of Engineering and Architecture
Department of Computer Engineering
Anadolu University, Eskisehir, TURKIYE
Phone : +90 (222) 3213550 – 6567
Fax : +90 (222) 3239501
Email : serkangunal@anadolu.edu.tr

REFERENCES

Anadolu University, Yunus Emre Learning Portal. (2008). Website: http://yunusemre.anadolu.edu.tr

Duda, R. O., Hart, P.E., and Stork, D.G. (2001). *Pattern Classification*, John Wiley & Sons Inc., USA.

Gunal, S., Ergin, S., Gulmezoglu, M. B., Gerek, O. N. (2006). "On feature extraction for spam e-mail detection", *Lecture Notes in Computer Science*, vol.4105, pp.635–642.

Hettich, S. and Bay, S. D. (1999). The UCI KDD Archive [http://kdd.ics.uci.edu]. Irvine, CA: University of California, Department of Information and Computer Science.

Jain, A., Zongker, D. (1997). "Feature selection: evaluation, application, and small sample performance", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.19, no.2, pp.153–158.

Kuncheva, L. I. (2004). *Combining Pattern Classifiers*, John Wiley & Sons Inc., New Jersey.

Lau, L. K. (2000). *Distance Learning Technologies: Issues, Trends and Opportunities*, Idea Group Pub.

MIT (Massachusetts Institute of Technology), OpenCourseWare system. (2008). Website: http://ocw.mit.edu

Sebastiani, F. (2002). "Machine learning in automated text categorization", *ACM Computing Surveys*, vol.34, no.1, March 2002, pp.1–47.

Selamat, A. and Omatu, S. (2004). "Web page feature selection and classification using neural networks", *Information Sciences*, vol.158, pp.69–88.

Schölkopf, B., Smola, A.J. (2001) *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press.

Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y. and Wang, Z. (2007). "A novel feature selection algorithm for text categorization", *Expert Systems with Applications*, vol.33, pp.1–5.

Theodoridis, S. and Koutroumbas, K. (2003) *Pattern Recognition*, Academic Press, USA.

Webb, A. (2002). *Statistical Pattern Recognition*, John Wiley & Sons Ltd., England.