# Unstructured Document Categorization: A Study

Debnath Bhattacharyya[1], Poulami Das[1], Debashis Ganguly[1],
Kheyali Mitra[1], Purnendu Das[1], Samir Kumar Bandyopadhyay[2], Tai-hoon Kim[3]

[1] Computer Science and Engineering Department, Heritage Institute of Technology
Kolkata-700107, India
{debnathb,dasp88,DebashisGanguly,kheyalimitra}@gmail.com,
purnendu_das@yahoo.com

[2] Department of Computer Science and Engineering, University of Calcutta
Kolkata-700009, India
skb1@vsnl.com

[3] Hannam University, Daejeon – 306791, Korea
taihoonn@empal.com

**Abstract.** The main purpose of communication is to transfer information from one corner to another of the world. The information is basically stored in forms of documents or files created on the basis of requirements. So, the randomness of creation and storage makes them unstructured in nature. As a consequence, data retrieval and modification become hard nut to crack. The data, that is required frequently, should maintain certain pattern. Otherwise, problems like retrieving erroneous data or anomalies in modification or time consumption in retrieving process may hike. As every problem has its own solution, these unstructured documents have also given the solution named unstructured document categorization. That means, the collected unstructured documents will be categorized based on some given constraints. This paper is a review which deals with different techniques like text and data mining, genetic algorithm, lexical chaining, binarization method to reach the fulfillment of desired unstructured document categorization appeared in the literature.

**Keywords:** Unstructured Documents, Categorization, Text and Data mining, Genetic Algorithm, Lexical Chaining, Binarization.

## 1 Introduction

The advent of e-commerce and corporate intranets has led to the growth of organizational repositories containing large and unstructured document collection. It is relatively less cumbersome to define categories broadly classifying the information contained in the collection. So, efficient storage and transmission of documents as well as archiving and information retrieval for document databases have become important research issues.

Structured documents must maintain the structure where in addition to pure textual information, the meaning of different sections likes author, title, abstract, heading of section or subsection, paragraph, etc are also stored within the same document.

Document summarization has been a well-known field of computational linguists among all. Automatically classifying text documents is of great practical importance given the massive volume of online text available through the World Wide Web. The existing statistical text learning algorithms can be trained to approximately classify documents. Various methods and applications were developed and used for the purpose of automatic document classification. The discipline of Text Mining merges methods of information retrieval, computational linguistics and data mining to achieve means for an automatic analysis of large and unstructured corpora.

In this paper, we have tried to present a detail survey on unstructured document categorization and we hope that this work will definitely provide a concrete overview on the past, present and future aspects in this field.

## 2 Overview

Document categorizations are not the new mechanism to categorize the unstructured documents. It has been coming from long pats days. The history of document categorization is discussed below.

### A. Past

Information science is the studying the collection, classification, manipulation, storage, retrieval and dissemination of information. Institutionally, information science emerged in the 19th Century along with many other social science disciplines. As a science, however, it finds its institutional roots in the history of science, beginning with publication of the first issues of Philosophical Transactions, generally considered the first scientific journal, in 1665 by the Royal Society.

The discipline of European Documentation, which marks the earliest theoretical foundations of modern information science, emerged in the late part of the 19th Century together with several more scientific indexes. Paul Otlet, a founder of modern information science proceeded to build a structured document collection that involved standardized paper sheets and cards filed in custom-designed cabinets according to an ever-expanding ontology and a commercial information retrieval service.

With the 1950's came increasing awareness of the potential of automatic devices for literature searching and information storage and retrieval. By the 1960s and 70s, there was a move from batch processing to online modes.

### B. Present

In 1992 the US Department of Defense, along with the National Institute of Standards and Technology (NIST), cosponsored the Text Retrieval Conference (TREC) to look into the information retrieval community by supplying the infrastructure that was needed for evaluation of text retrieval methodologies on a very large text collection.

Labour-intensive manual text-mining approaches first surfaced in the mid-1980s, but technological advances have enabled the field to advance swiftly during the past decade.

The concept of neural networks started in the late-1800s as an effort to describe how the human mind performed.

Computer simulations of evolution started as early as in 1954 with the work of Nils Aall Barricelli, who was using the computer at the Institute for Advanced Study. Starting in 1957, the Australian quantitative geneticist Alex Fraser published a series of papers on simulation of artificial selection of organisms with multiple loci controlling a measurable trait. From these beginnings, computer simulation of evolution by biologists became more common in the early 1960s, and the methods were described in books by Fraser and Burnell (1970) and Crosby (1973).

### C. Future

Many have done, several are eager to be invented. For example, the concept of Neural Network has included neuroscience as a future researching field. A particularly important part of the investigation in recent years has been the exploration of the role of neuro-modulators such as dopamine, acetylcholine, and serotonin on behavior and learning.

Biophysical models, such as BCM theory, have been important in understanding mechanisms for synaptic plasticity. Research is ongoing in understanding the computational algorithms used in the brain, with some recent biological evidence for radial basis networks and neural back propagation as mechanisms for processing data.

## 3 Approach

### A. Text and Data mining Technique

This technique has a great impact in the field of document categorization. Generally this is used in the following approaches.

- GENERATION OF TAXONOMY FOR LARGE DOCUMENT COLLECTION:

Text and data mining are two closely related approaches for information retrieval. The main aim here is to create automatic taxonomies by using the text and data mining concept and comparing documents on the basis of certain characteristic features they contain.

For automatic generation of Taxonomy of collected documents Lexical Affinity (LA) followed by Linguistic Features (LF) extraction and hierarchical clustering algorithm (HCA) are used. These methods are (LA, LF) generating the output for clustering. Now HCA is an algorithm which follows bottom-up approach. Starting from the individual documents this algorithm first generates the lower (bottom) cluster. Then it goes higher by grouping the lower level clusters according to necessity. As a result, the automatic taxonomy is generated for the given set of documents.

The HTML output from this technique (LF based Taxonomy) is given in the Fig. 1. The output is generated based on names, terms etc. The result is quite satisfactory and stable at about 5000 documents.

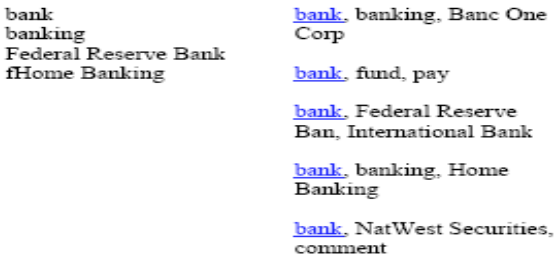


**Fig.1.** Node within LF based Taxonomy

- SIMPLE AND FAST TERM SELECTION PROCEDURE FOR TEXT CLUSTERING:

One algorithm that is proposed for enhancing the text clustering process by reducing the dimensionality of feature space and thus reducing time for processing and enhancing clustering procedure is Vector–Space Model [3]. The algorithm has used the document as a document to term matrix for calculation procedure.

Thus it gives the limitation to the higher ranked terms of a document. And this is purely concern with saving the time and labor in computation. The result is also quite satisfactory as it does clustering with pure efficiency and the number terms involved in computation is very less which yields computation speed.

### B. Lexical Chain Technique

Document summarization has been working as a Computational linguist for past few decades. Lexical Chain and concept of natural language processing (NLP) are used to

generate the concept of document summarization [4]. This summarizer can process the simple" flat" documents as well as complex ones.

Now for generating document summarizer; certain step-by-step approach is followed.
(i) Analyzing the structure of documents,
(ii) Classification of the documents and make sets according to predefined categories,
(iii) Using of natural language technique for summarization of documents,
(iv) The summaries are amalgamated with the structure of the actual document to give the exact summary.

Through the structured analysis and concept of lexical chain method, structured as well as unstructured documents can be summarized in a perfect manner. Ultimately the output in XML is given in the Fig.2 and Fig.3.

**BCL Corpus**

This document describes the creation, maintenance and modification of the BCL Corpus created at BCL Technologies. BCL Technologies develops software solutions necessary for document management and web publishing. It specializes in developing software that analyzes, manipulates and uses information that is stored in different file formats. As part of the customer support BCL Technologies responds to individual queries from customers who are using BCL products and who have questions regarding the products we sell.

The BCL corpus is a written corpus comprised of email messages we receive from our customers. These email messages contain questions, comments and general inquiries regarding our document-conversion products. These email messages were collected between June 2000 and May 2001. We modified the raw email programmatically by deleting the attachments, html and other tags, header files, and senders' information. In addition, we manually deleted salutations, greetings, and any information that was not directly related to customer support. There are around 34,640 lines and 170,000 words in the BCL Corpus. We constantly update our corpus with new email from our customers.

We further pruned down our corpus to create subsets of testing corpora in order to test various modules of the Spoken Language User Interface Toolkit (SLUITK) system. For example, from the BCL corpus, we created a sample test corpus of 1000 mono-clausal inquiry-format sentences to test the end-to-end frame generation module of our system. Similarly, we created a sample test corpus of 50 generic sentences from our corpus to do a preliminary testing of the whole system.

**Fig.2.** Given Text

<Head>Support **BCL Corpus** <\Head> <ContentWeight>
<1412><\ContentWeight><ImageWeight>0<\ImageWeight><LinkWeight>0<LinkWeight>

**Fig. 3.** After extraction from the above text

### C. Genetic Algorithm (GA) Technique

This is one of the techniques that are followed in the world of structured documentation. However, very few researchers use this technique. This is specially used in information retrieval (IR) method [5].

In this approach, it was expected that the combination of GA along with the well-known matching functions would give the satisfactory result than using single matching function. The following figure expresses the general method for information retrieval.

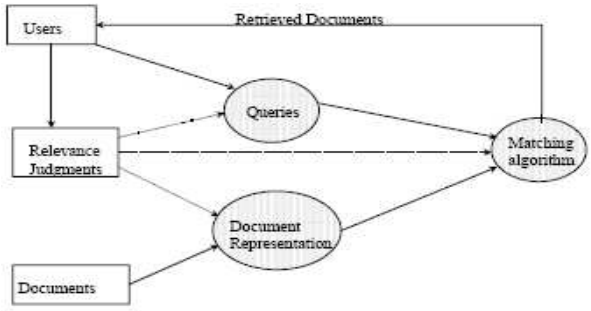


**Fig.4.** Information Retrieval System

GA is used to change the descriptions of documents or queries and also the matching functions those are used for IR from any document. In this case, documents and queries are spaced in multi-dimensional vector space for convenience of work. Now, while retrieving the documents those are accepted who are close to the vector-associated with query. For implementing the algorithm (shown in the Fig.5), some processes are followed.

- GENERATE MATCHING FUNCTION VARIANTS*:* Randomly chosen values are assigned to each and every matching function. The ranges start from 0 to 1.0.And here 50 different populations is selected for examination.
- MATCHING FUNCTION VARIANTS FITNESS EVALUATION: Overall matching value corresponding to respective documents of a population is computed and then the calculated documents are organized in non-ascending order.
- *GENETIC MODIFICATION:* It includes four different stages.
- SELECTION AND REPRODUCTION: The selected generations are exposed for reproduction in the coming generation.
- CROSSOVER: The information exchange is happening through two points those are arbitrarily selected (two-point cross over [5]).
- MUTATION: This process is done by the method given by Gaussian noise.
- PROCESS TERMINATION*:* Now the total process will be terminated when it is found that after applying genetic modification to individual generation, no such modified out put is found through few consecutive generations.

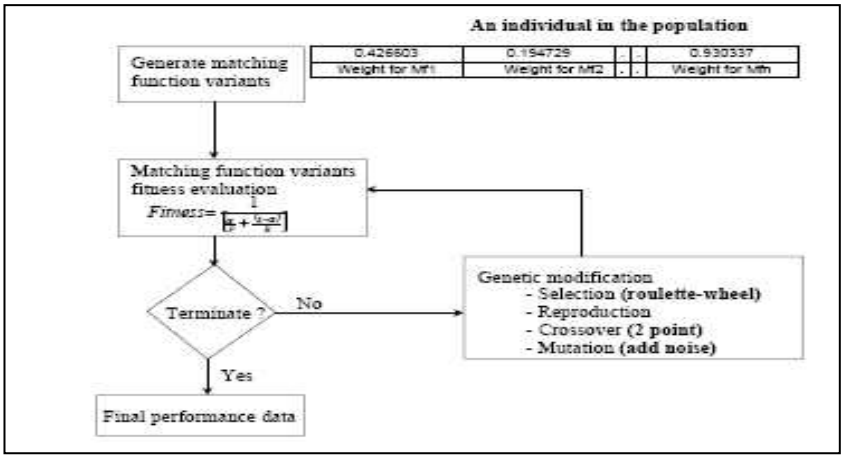


**Figure 5:** Genetic Procedure

### D. Artificial Neural Network technique

This is a well-known approach for classification of documents. Automatic document classifications are done using the concept of Self Organizing Maps (SOM) and Learning Vector Quantization (LVQ) algorithm [6].

Self Organizing Maps (SOM) do automatic classifications .SOM is based on Artificial Neural Network concept. Through this concept automatic clustering can be done. Learning Vector Quantization (LVQ) is for maximizing the correctness of classification of data.

The two algorithms SOM and LVQ are used individually for checking the validity. For each case the algorithms are generated then trained and tested .Two simulations are used for both the processes (SOM, LVQ). Each simulation is arranged in different pattern. In one case 70% are training set and 30% are for testing. For the other one 30% are for training, and 70% are used for testing. The result is tabulated in the following Table 1.

Table 1

| | SOM | (Unsupervised) | | LVQ | (Supervised) | |
|---|---|---|---|---|---|---|
| | 100% | 70%-30% | 30%-70% | 100% | 70%-30% | 30%-70% |
| Learn | 72.69 | 70.84 | 72.98 | 74.74 | 74.43 | 76.71 |
| Test | NA | 67.08 | 56.72 | NA | 67.39 | 64.45 |

**Table 1:** The experimental Result

### E. Binarization

Paleography is the study of ancient handwritten manuscripts. It generally deals with timing and localizing of ancient scripts. Most of the analysis of the manuscripts is based on the character shape. That is why thresholding with acuteness is required to use in this field. To get rid of the corrupted condition of the documents, multi-stage accurate binarization scheme is applied [7]. The multistage threshold [8] technique is adapted in Paleographical analysis. The steps that are followed are:

- GLOBAL THRESHOLDING: This stage reduces the volume of search space of foreground elements.
- DISCARDING IRRELEVANT OBJECTS: After passing through the 1st stage, the document will be free from small blobs and letters. The line extraction method is there to extract the text lines from the documents.
- LOCAL COMPONENT PROCESSING: Now the foreground pixels are accumulated and sets are generated. The neighbors of a certain data have the effect on it.
- POST PROCESSING: At last the holes that are generated by corrupted parts of some words i.e. the faded parts are processed accordingly.

The total procedure is explained by the given figures (Figure 6, Figure7) of Hebrew text. The 1st one is the input and 2nd one is the processed text that is the desired output.

**Fig. 6.** Input Text

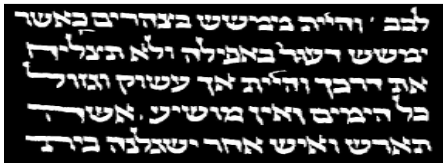

**Fig.7.** Modified output from given text

## 4 Application

The main reason for document categorization is to access the desired document in a sophisticated manner so that in future the data or the document itself can be modified and retrieved with out loosing any information.

In the business world document categorizations are used to achieve the above-mentioned goal so that the data will be securely stored in data repository. This is also used in industrial field for betterment of information storage.

The text and data mining concept used in taxonomy and automatic text clustering methods are giving quick access to data from repository. As they are giving their desired output; those concept can be used in large number of document for being categorized in faster process. That is why those can be used in banking system, organizations or in industry. The binarization process can easily be implemented for extracting data from corrupted (specially the ancient manuscript) documents.

In future, the concept of automatic taxonomy of documents can be amalgamated with the concept of paleographical search technique of information retrieval process. The technique used for taxonomy generation along with data retrieval process of binarization can lead to the better result than before. The data extractions from corrupted documents are quite similar to image processing. So amalgamating the concept of image processing

and the data extraction method can emerge a new and more concrete concept of categorization of documents in future.

## 5 Evaluation

When we follow any kind of approach to obtain categorization, we always keep focusing on not losing any information after modifying that. So, before we follow a certain technique, we have to evaluate that to match our desired goal. The evolutions are described below.

### A. TEXT AND DATA MINING TECHNIQUE:

For generation of taxonomy of documents, the process that have discussed, gives satisfactory and stable result at about 5000 documents. But as we increased the size of the sample, the saturation point changes some how. For LA based input data when it is tested, the stability is gained above 4000 documents .It is required to increase the size of the sample by 50% of data for the higher dimensional Input data.

In case of text clustering process, the result is also quite satisfactory as it does clustering with pure efficiency and the number terms involved in computation is very less which yields computation speed.

### B. LEXICAL CHAIN MECHANISM:

The commercial summarizer, which follows this technique, is still under process. So there is a vast field to improve the desired out come from this approach.

### C. GENETIC ALGORITMIC APPROACH:

The information retrieval process using this algorithm gives expected results in certain fields like Simulated and Canfield for document retrieval .Yet other fields are there where this algorithm must be checked and also with new different kinds of matching functions (except those which are used).

### D. ARTIFICIAL NEURAL NETWORK:

Automatic document classification based on ANN, gives result, where it is found that the supervised learning is giving better than the unsupervised one. This is limited to small cluster size. More improvement is required in this field.

### E. BINARIZATION:

Extraction of information from corrupted documents using this method is quite satisfactory. But in future the following must be achieved for using this process in paleographical work;

(i) writer's authentication checking mechanism
(ii) Writing style identification
(iii) Dating, timing and localization of manuscript

## 6 Discussion & Conclusion

Document categorization is enhancing the level of data storage as well as data access and modification process. We have already discussed methods related to this field shown the drawbacks of each and every approach. Many techniques and algorithms for categorization have been devised. Nothing is sufficient by its own. We should keep track on the limitations and the problems of the given approach. Like genetic algorithm, it must be tested in many other fields except the mentioned ones. Not only that the matching functions must enhance its flexibility towards the problems. For neural network, the concept of neuroscience must be combined in future so that it will provide a new era in document categorization. Text data and images can be combined together and may yield an efficient result to achieve the goal. Medical field gives such type of environment where we can use document categorization efficiently. The interest in automating the collection, organization and analysis of biological data is creating the platform for this.

The way of categorizing the unstructured documents advances quickly. A good categorizer or classifier efficiently categorizes large sets of documents in a reasonable time frame and with an acceptable accuracy, and that provides classification rules that are human readable for possible fine-tuning. If the training of the classifier is also quick, this could become in some application domains a good asset for the classifier. Data will be secured only when it is well structured and some security or protection is available. So document categorization can be merged with several security processes. These are also developing themselves for future revolution.

## References

1. Adrian Müller, Jochen Dörre, Peter Gerstl, Roland Seiffert: The TaxGen Framework: Automating the Generation of a Taxonomy for a Large Document Collection
2. http://www.gc.ssr.upm.es/inves/neural/ann1/concepts/Suunsupm.htm
3. Luiz Gonzaga (1), Marco Grivet (2), Ana TerezaVasconcelos (1) Laboratório Nacional de Computação Científica - LNCC (1) Pontifícia Universidade Católica do Rio de Janeiro - PUCRIO (2) lgonzaga@lncc.br, mgrivet@cetuc.puc-rio.br, atrv@lncc.br: A Simple and Fast Term Selection Procedure for Text Clustering
4. Hassan Alam, Aman Kumar, Mikako Nakamura, Fuad Rahman1, Yuliya Tarnikova and Che WilcoxBCL Technologies Inc.fuad@bcltechnologies.com: Structured and Unstructured Document Summarization: Design of a Commercial Summarizer using Lexical Chains
5. Praveen Pathak Michael Gordon Weiguo Fan: Effective Information Retrieval using Genetic Algorithms based MatchingFunctions Adaptation
6. Dina Goren-Bar, Tsvi Kuflik, Dror Lev Information Systems Engineering Department, Ben Gurion University of the Negev, Beer-Sheva: Supervised Learning for Automatic Classification of Documents using Self-Organizing Maps
7. Itay Bar Yosef1, Klara Kedem1, Its'hak Dinstein2, Malachi Beit-Arie3 and Edna Engel3: Classi_cation of Hebrew Calligraphic Handwriting Styles: PreliminaryResults.
8. Kamran Etemad, David Doermann, and Rama Chellappa: Multiscale Segmentation of Unstructured Document Pages Using Soft Decision Integration.
9. Sourav Sengupta, Bernard J. Jansen: Designing a Value Based Niche Search Engine Using Evolutionary Strategies
10. Jane Cleland-Huang, Raffaella Settimi, Xuchang Zou, Peter Solc: The Detection and Classification of Non-Functional Requirements with Application to Early Aspects
11. Bing Liu, Philip S. Yu, Xiaoli Li: Partially Supervised Classification of Text Documents.
12. Tomek Strzalkowski, Wang Jin and Fang Lin: Integration of Document Detection and Information Extraction
13. http://bioinformatics.oxfordjournals.org/cgi/content/abstract/ as visited on 24/4/569