Text Documents Classification Using Word Intersections

Mohamed Ali AlShaari

***Abstract*—There are many methods that can be used to classify documents, some of these methods depend on discipline and others depend upon human orientation. Nonetheless all of them have certain degree and type of difficulty. This work was designed to introduce a simple idea that can be used to classify any text document. The hypothesis was based on the fact that every science, discipline or any field of knowledge has its own terminologies. Therefore, the algorithm (software system) was developed employing some operations in set theory to extract the terminologies within a certain discipline. These terminologies were used to classify text documents, whether they are related to specific discipline. The algorithm written to carry out all necessary operations was implemented using Matlab. The system developed was tested, and the results obtained were accurate. It is anticipated that this system can be used to facilitate e-translation of documents to produce more meaningful translation.**

***Index Terms*—Classification, intersection, set, terminology.**

I. Introduction

Highlight a Document classification or document categorization is a problem treated by several disciplines like computer science, library science, and information science [1]. The aim of document classification is to determine the field that a document is related to, i.e. Assign a document to certain classes or categories [2]. As the library science is concerning manual documents classification, the information and computer science develops algorithms to carry out documents classification.

There are many kinds of documents such as text documents, images, etc. The project presented in this paper deals with the classification of text documents, which could help in the classification of other kinds of documents.

Documents might be classified according to subject or attributes such as author, type, publishing date, etc.

The subject classification is considered. There are many methods used to make subject classification of documents. The most significant methods are the content based approach method and the request based approach method.

A. Content Based Classification

In this classification the document is assigned to a certain class according to its content, i.e. if at least 20% of the document covers the subject, the document assigned to it. This may be accomplished using software designed to count words in the document related to a specific subject. The redundancy of words is used to classify the document. [3]

Manuscript received May 7, 2013; revised July 23, 2013.

Mohamed AlShaari is with the Department of computer science, Benghazi University, Benghazi, Libya (e-mail: Mohamed.shaari@gmail.com).

B. Request Oriented Classification (Or -Indexing)

This type of classification is based on orientation ability of users to manipulate the categories of documents. The classifier chooses which descriptors should the entity relate? This classification depends on the standards of library indexing, which requires user suggestions. The request oriented classification consider a user-based approach if practical data about using or users are applied.

C. Automatic Document Classification

"To improve the effectiveness and efficiency of document categorization at the library setting, more in-depth studies of using automatic document classification methods are required."[4]

There are two methods to classify documents automatically in supervised and unsupervised document classification. In the supervised method some human comments represented in the mechanisms offer evidence to correct classification for documents, but in unsupervised document classification or document clustering the classification finalized without any human feedback; however the classification is completed without any exterior information [5].

In addition, there is a semi-supervised document classification, cooperates supervised and unsupervised mechanisms in which some of the fragments of the documents are categorized by human interventions.

II. The Idea

Every discipline has its own specialized terminologies, and each document related to certain discipline must contain a number of its terminologies. Therefore, if we know the specified terminologies of a discipline, then we can know whether or not a document is related to a discipline according to its contained terminologies, hence documents could be classified.

III. The Proposed Classification Method

A. Part One

1) No less than seven documents were selected, which contained fundamental texts related to a certain discipline, e.g. overview, review, or any introductory papers
2) Common words between those documents were found, e.g. every word found in all documents
3) Only the necessary terminologies were kept, i.e. All unnecessary words such as propositions, articles, ordinary verbs such as and, in a, the were cancelled,
4) Most iterative terminologies T were determined in the

DOI: 10.7763/IJET.2014.V6.678

discipline S.

B. Part Two

1) Several text documents were selected, some related to discipline S, and others related to another discipline Y
2) The extent of terminologies T or some of it found in selected documents were recorded, each counted, and the ratio of T terminologies in each document was calculated.
3) The highest number or ratio means that the document is related to the discipline X, and the smallest number or ratio means that the document is not related to the discipline

Therefore, we can classify any document according to its contained terminology.

IV. The Algorithm

- Let $D_1 ... D_n$ be documents related to scientific field S, so:
- D_1={ $word_1$, $word_2$, … }
- D_2={ $word_1$, $word_2$, … }
- … …
- D_n={ $word_1$, $word_2$, … }
- I = $\cap_{i=1..n}$ D_i % I={ $word_1$, $word_2$, … }; common words between all documents

- Let N be a set of neutral words % like When, Give, Is, a, an, …
- I=I-N % remove neutral words from set I
- Let T be a Document related to S, X related to discipline that related to S, and Y related to another science
- K= T ∩ I
- L = X ∩ I
- M = Y ∩ I
- ∴ |K| ≅ |I| % Close to S
- ∴ |L| < |I|
- ∴ |M| << |I| % Much less than, perhaps very much less than (<<<)

V. Results

Two data sets were used to test the algorithm, one was related to Genetic Algorithm (GA), the other was related to Linear Programing (LP).
First test:

1) Seven documents were used with subject like Introductions to Genetic Algorithms (GA), the program calculates the intersection between the words in all documents, and after removing the neutral words the program produced the following 25 terms as a result(see Table I):
" Algorithms', 'Crossover', 'Fitness', 'GA', 'Genetic', 'Mutation', 'Selection', 'crossover', 'evolution', 'fitness', 'function', 'mutation', 'number', 'point', 'population', 'probability', 'random', 'randomly', 'search', 'selection', 'solutions', 'string', 'time', 'work' "
2) Four files were selected to test, the first one was related the Genetic Algorithms, the second was related to Artificial Intelligence, the third was related to Computer Science, and the last one was related to Management Science.
3) The program calculates the relations between 25 terms and test files and shows the results which illustrate the belonging of each test file document to Genetic Algorithms Field.

TABLE I: Repeated Words in Reference Files

Subject of seven reference files	Genetic Algorithms
Repeated key words in reference files	" Algorithms', 'Crossover', 'Fitness', 'GA', 'Genetic', 'Mutation', 'Selection', 'crossover ', 'evolution', 'fitness', 'function', 'mutation', 'number', 'point', 'population', 'probability ', 'random', 'randomly', 'search', 'selection', 'solutions', 'string', 'time', 'work' "
Number of repeated words	25

TABLE II: Extent of Relation of GA File to Reference File

Subject of first test file	Genetic Algorithm (GA)
Repeated words between test file and reference files	'Algorithms' 'Crossover' 'Fitness' 'Genetic' 'Mutation' 'Selection' 'crossover' 'evolution' 'fitness' 'function' 'mutation' 'number' 'point' 'population' 'probability' 'random' 'search' 'selection'
Number of words	18
Extent of relation of first test file to GA	72%

TABLE III: Extent of Relation of AI File to Reference File

Subject of second test file	Artificial Intelligence (AI)
Repeated words between test file and reference files	'Algorithms' 'Genetic' 'Mutation' 'crossover' ' evolution' 'fitness' 'function' 'mutation' 'number' 'point' 'population' 'random ' 'search' 'solutions' 'time' 'work'
Number of words	16
Extent of relation of second test file to GA	64%

TABLE IV: Extent of Relation of CS File to Reference File

Subject of third test file	Computer Science (CS)
Repeated words between test file and reference files	'Selection' 'function' 'number' 'random' 'randomly' 'search' 'selection' 'solutions' 'string' 'time' 'work'
Number of words	11
Extent of relation of third test file to GA	44%

TABLE V: EXTENT OF RELATION OF MS FILE TO REFERENCE FILE

Subject of selected file	Management Science (MS)
Repeated words between test file and reference files	'Algorithms' 'Genetic' 'Selection' 'point' 'solutions' 'time' 'work'
Number of words	7
Extent of relation of last test file to GA	28%

The GA test file scored 72% (see Table II) which means that this file closely related to reference files. The AI, CS and MS test files scored lower percentages than GA (64%, 44% and 28%, respectively, see Tables III, IV and V), which indicates that as the test file discipline is less related to reference files discipline the percentage decreases.

Second test:

1) Seven documents were used with subject Introductions to LP, the program calculates the intersection between the words in all documents, and after removing the neutral words the program produced 73 terms as a result of disciplines relationship extent (see Table VI): 'Linear' 'Programming' 'additional' 'amount' 'analysis' 'based' 'cases' 'constraint' 'constraints' 'cost' 'costs' ,…
2) Four files were selected to test the system, the first one was related Linear Programming, the second was related to Management Science , the third was related to Operations Research, and the last one was related to History.
3) The results illustrate the extent of relationship of each test file document to LP Field.

TABLE VI: REPEATED WORDS IN REFERENCE FILES

Subject of seven reference files	Linear programming
Repeated key words in reference files	'Linear' 'Programming' 'additional' 'amount' 'analysis' 'based' 'cases' 'constraint' 'constraints' 'cost' 'costs' 'current' 'equal' 'example' 'feasible' 'first' 'form' 'found' …
Number of repeated words	73

TABLE VII: EXTENT OF RELATION OF LP FILE TO REFERENCE FILE

Subject of first test file	Linear programming
Repeated words between test file and reference files	'Linear' 'Programming' 'additional' 'amount' 'analysis' 'based' 'cases' 'constraint' 'constraints' 'cost' 'costs' 'current' 'equal' 'example' 'feasible' 'first' 'form' 'found' …
Number of words	71
Extent of relation of first test file to LP	97.3%

TABLE VIII: EXTENT OF RELATION OF MS FILE TO REFERENCE FILE

Subject of second test file	Management Science
Repeated words between test file and reference files	'amount' 'analysis' 'based' 'cost' 'costs' 'example' 'first' 'form' 'found' 'function' 'general' 'important' 'large' 'least' 'method' …
Number of words	42
Extent of relation of second test file to LP	57.5%

TABLE IX: EXTENT OF RELATION OF OR FILE TO REFERENCE FILE

Subject of third test file	Operations Research (OR)
Repeated words between test file and reference files	'Linear' 'Programming' 'additional' 'amount' 'analysis' 'based' 'cases' 'constraint' 'constraints' 'cost' 'costs' 'current' 'equal' 'example' 'feasible' 'first' 'form'…
Number of words	72
Extent of relation of third test file to LP	98.6 %

TABLE X: EXTENT OF RELATION OF HISTORY FILE TO REFERENCE FILE

Subject of selected file	History
Repeated words between test file and reference files	'amount' 'analysis' 'based' 'current' 'equal' 'example' 'first' 'found' 'function' 'important' 'large' 'least' 'linear' 'method' …
Number of words	30
Extent of relation of last test file to LP	41 %

The test file (OR file) scored 98.6% (see Table IX) because LP as a topic is related to OR discipline, but the test file related to MS scored 57.5% (see Table IIIV) because MS test file was a general test file, but the History test file scored 41%(see Table X), because History is remotely related to LP discipline .

VI. CONCLUSION

I anticipate that the idea proposed and successfully tested in this paper would ease text classification. The simple fact that every discipline has its own terminologies gives us the ability to categorize any document, whether it is related to one discipline or another. Furthermore, the paper presented the algorithm that demonstrates how to apply the idea. The algorithm was implemented using Matlap software language.

The software was tested using text documents related many disciplines, in every test a certain discipline was used as the main subject to retrieve its terminologies and a text document belonging to the main subject or other discipline was used as a test document to find the extent of the

relationship between the test document and the main subject discipline.

The idea, the algorithm and the software developed is anticipated to help researchers and developers in many fields that employ information retrieval, natural language processing especially text mining, and they can be used in many other scientific fields.

The unnecessary words, i.e. words commonly used in every text document need more research to be determined accurately in order to be removed from the test and reference files without affecting the necessary terms.

In addition, the software could be developed using intelligent programming language, and tested with other natural languages to be used in a broader sense.

REFERENCES

[1] K. Golub, "Automated subject classification of textual web documents," *Journal of Documentation*, vol. 62, Mar 2006.

[1] Y. Li and A. Jain, "Classification of text documents," *The Computer Journal*, vol. 41, no. 8, pp. 537-546, 1998.

[2] M. Rafi, S. Hassan, and M. Shaikh, "Content-based text categorization using wikitology," *IJCSI*, vol. 9, no 2, July 2012.

[3] A. Khan, B. Baharudin, and K. Khan, "A review of machine learning algorithms for text-documents classification," *Journal of Advances in Information Technology*, vol. 1, no. 1, Feb 2010

[4] H. Cui, "Unsupervised learning of soft patterns for generating definitions from online news," in *Proc. the 13th International Conference on WWW*, New York, May 17 - 22, 2004, pp. 90 – 99.

Mohamed Ali AlShaari is a full time lecturer in the department of computer science at the Benghazi University, Libya. He obtained his Bsc in computer science from Garyounis University, Libya and Master of artificial intelligence in faculty of information system in Benghazi University. His main research lines are natural language processing, genetic algorithms, bioinformatics.