CONCEPT BASED CATEGORIZATION OF DOCUMENTS FOR SEARCH ENGINES

Soumen Swarnakar[1], Sangita Karmakar[2]

[1]*Assistant Professor, Department of Information Technology, Netaji Subhash Engineering College, W.B., India*
[2]*B.Tech Student, Department of Information Technology, Netaji Subhash Engineering College, W.B., India*

Abstract

Now days, information retrieval is a challenging work for search engines. In this paper we will discuss about text categorization. Text documents categorization is the process to classify documents according to some predefined knowledge. Documents with same concept are grouped together, and documents with different concept are formed other group according to their similarity of context of the documents. This grouping technique is called document categorization. So the related documents will be in same category and non related documents in other category. In this paper we have concentrated on the document context, according to their context, categorization process is done. So we are trying to propose Link base document categorization according to the document context of a particular concept. In this way we can retrieve the proper information about the document and also find about the document's main concept and about what sub concept according to the percentage of weights of domains of a document. According to percentages of different concepts of different domain and indexing of documents, the categorization can be improved for information retrieval process of a search engine.

Keywords: *Context address, Mixture category, pure category, concept dictionary, Domain.*

----------***----------

1. INTRODUCTION

Text document concept analysis is a part of information extraction. Document categorization is happened according to the concept dictionary or on which way documents are coming in sequence. Automatic text categorization has many practical applications, including indexing for document retrieval, automatically extracting metadata, word sense disambiguation by detecting the topics a document covers, and organizing and maintaining large catalogues of Web resources and It is also used in automatic document organization topic extraction and information retrieval or filtering information. A fuzzy based approach for multilabel text categorization and similar document retrieval has been suggested by Rubiya P U et al. (2015). Ontology based document clustering has been proposed by Soumen Swarnakar (2012) whereas a new approach to concept base document clustering has also been proposed by Soumen Swarnakar et al. (2015). Correlated concept based dynamic document clustering algorithms for newsgroups and scientific literature suggested by Jayaraj Jayabharathy and Selvadurai Kanmani (2014). A new term weighting Scheme for clustering dynamic data streams is also proposed by Joel W. Reed et al. (2006). Traditional search engines output a list of results that are ranked according to their relevance to the query. Our proposed approach will help to categorize the documents automatically according to concept which will improve search engine performance.

2. PROPOSED WORK

In this paper focus has been given on concept of documents. At first all the text documents are converted in to lower case. Next, preprocessing is done by removing articles, prepositions, conjunctions and finally stemming operations on different words are applied, i.e., if word in document is *injured*, then after stemming the word would be *injure*. Next, according to concept dictionary and synonyms dictionary, occurrence matrix has been computed. From occurrence matrix, percentages of different concepts of different domains for documents have been calculated and according to percentage of concepts, the concept of the document will be decided. After that, indexing of documents is done according to concept based category.

2.1 TERMINOLOGIES USED

2.1.1 Concept address

Concept address is specific location into the domain according to the specific keywords coming from the documents. Concept address for any particular document gives the specific address for a particular domain in which it reside.

Here concept address has been described below for any document numbers i = 0, 1, 2, 3..... n, co^i signifies concept of document i where W1, W2, W3 are the occurrence of concept 1, occurrence of concept 2 and occurrence of concept3 respectively.

co^i	W1	co^i	W2	co^i	W3

In this way we can implement the concept address of the any new document.

2.1.2 Mixture Category

In this paper, it is shown that if a document has more than one concept, based on domain, then this document is categorized as mixture concept. This type of mixture concept based document will hold base address or starting address with some specific category and linked to some other sub category.

2.1.3 Pure Category

Pure category is nothing but a class of documents holding the pure type of concept, based on a specific domain within it. In this category document will be present in a particular class and no link will be from this category.

2.1.4 Concept Dictionary

Concept dictionary is nothing but a dictionary where many types of concept and related objects are stored. These objects are called the keywords of the specific concept. These keywords reside in the domain of any concept.
For example, if we consider Medical is the concept, so domain is also the medical, and objects of the domain are specified Keywords like doctor, nurse, which are related to the medical concept. In this paper we are considering the concept dictionary is static type, means at run time the dictionary is not changed. Concept dictionary is shown below in Table1.

Table 1: Concept Dictionary

CONCEPT	KEYWORDS
MEDICAL	Doctors, nurse, hospital, medical, nursing home, patient, medicine, x-ray, disease, treatment, ambulance, cardiologist, neurologist.
TRANSPORT	Car, motor cycle, bus, taxi, cycle, school van, transportation system, truck, ambulance, driver.
EDUCATION	Teacher, class, college, school, professor, intuition, student, book, copy.

2.1.5 Domain

Domain is the classification of the concept dictionary, by which we can decide in which category the document is actually resides in, means according to the document's concept a particular document is categorized in to a specific domain or a mixer domain.
For an example, we have 3 domains like MEDICAL, EDUCATION and TRASPORT. So the minimum no. of domain is 3. But, different combination of categorization is possible which are actually holds the mixture concept based document.

2.1.6 Percentage of Weight

Percentage of weight of the any document is defined as total occurrence of objects of a specific domain over total objects of different domain present in the document .So it is basically define the count of the no. of the terms is occurring in the document. The formula of weight is defined below,
W= ((total occurrence of objects of a specific domain) /(total objects of different domain present in the document))*100;
Example:
Suppose for document 1,

1^1	6	2^1	0	3^1	2
Medical		Transport		Education	

From this above table we get
1^1 =means that $document_1$ is medical concept. .
6=means that keyword of the medical (domain) concept is occurred 6 times in the document.
So, the weight of the $document_1$ is for the medical (domain).
W1=6/ (6+2) =6/8=0.75=75% (medical domain)
W2=0/ 0=0% (transport domain)
W3=2/ (6+2) =2/8=0.25=25% (education domain)
So, from the above result, we conclude that document 1 is medical oriented education concept.

2.2 CONCEPT BASED DOCUMENT CATEGORIZATION ALGORITHM

Step 1: Sorting percentage of weights

For each document i = 0, 1, 2, 3.....n find the weight of each domain.
Suppose document i (doc $_i$) has different percentage of objects of different domain.
Supposes doc_i has percentage of weight for 3 concepts are p_j(medical), p_k(transport), p_l(education) respectively.
Now, sort percentage and arrange in w1, w2, w3 respectively in descending order of percentage.

Step 2: Categorization of Document

if(p_j!=0 && p_k!=0 && p_l!=0)
{
if(p_j>p_k>p_l)
{
then, w1=p_j; w2=p_k;w3=p_l;
Now, the category of doc_i will be p_j centric p_k sub centric and p_l oriented. (Mixer Category).
}
else if(p_j>p_l>p_k)
{
then, w1=p_j; w2=p_l;w3=p_k;
Now, the category of doc_i will be p_j centric p_l sub centric and p_k oriented. (Mixer Category).
}
else if(p_k>p_j>p_l)
{
then, w1=p_k; w2=p_j;w3=p_l;
Now, the category of doc_i will be p_k centric p_j sub centric and p_l oriented. (Mixer Category).
}
else if(p_k>p_l>p_j)
{
then, w1=p_k; w2=p_l;w3=p_j;

Now, the category of doc_i will be p_k centric p_l sub centric and p_j oriented. (Mixer Category).
}
else if(p_l>p_j>p_k)
{
then, w1=p_l; w2=p_j;w3=p_k;
Now, the category of doc_i will be p_l centric p_j sub centric and p_k oriented. (Mixer Category).
}
else
{
then, w1=p_l; w2=p_k;w3=p_j;
Now, the category of doc_i will be p_l centric p_k sub centric and p_j oriented. (Mixer Category).
}
}
else if(p_j!=0 && p_k!=0)
{
if(p_j>p_k)
{
then, w1=p_j; w2=p_k;
Now, the category of doc_i will be p_j centric p_k oriented. (Mixer Category).
}
else
{
then, w1=p_k; w2=p_j;
Now, the category of doc_i will be p_k centric p_j oriented. (Mixer Category).
}
}
else if(p_j!=0 && p_l!=0)
{
if(p_j>p_l)
{
then, w1=p_j; w2=p_l;
Now, the category of doc_i will be p_j centric p_l oriented. (Mixer Category).
}
else
{
then, w1=p_l; w2=p_j;
Now, the category of doc_i will be p_l centric p_j oriented. (Mixer Category).
}
}
else if(p_k!=0 && p_l!=0)
{
if(p_k>p_l)
{
then, w1=p_k; w2=p_l;
Now, the category of doc_i will be p_k centric p_l oriented. (Mixer Category).
}
else
{
then, w1=p_l; w2=p_k;
Now, the category of doc_i will be p_l centric p_k oriented. (Mixer Category).
}
}
else
{
then w1=p_j or p_k or p_l;
Now, the category of doc_i will be p_j centric or p_k centric or p_l centric. (Pure Category).
}

2.3 EXPERIMENTAL RESULTS

Doc_1: Rita is a doctor. Her appointment is as the cardiologist in the hospital. She is good doctor .She is educated from very good medical college. Recently she is also joined into the medical institution as a professor. She teaches the students of that intuition. Now she is become the greatest cardiologist in India.

Doc_2: Swati is a school student of class five. She goes to the school every day by school van. She is very good student. Her teachers are proud of her grades. She is always comes first in the class and she wants to become a doctor in her life. She wants to become a neurologist.

Doc_3 : Bus is used for public transport whereas truck is used for heavy transportation system. Taxi is also used for public transport. But car is used for private transport. Ambulance is used for medical purpose for any emergency condition.

Doc_4 : Susmit is a neurologist. He is a one of the best doctor in west Bengal. He is educated from the best medical college. Now he has joined in the medical college as a professor. In this hospital he has joined as a neurologist and he has been awarded as the best doctor in the hospital. As a doctor he tries to do the best treatment of his patients.

For document 1 (Doc_1)

1^1	8	2^1	0	3^1	5
Medical		**Transport**		**Education**	

P_j=8/(8+5)=0.615*100=61.5%(medical concept).
P_k=0/0=0% (transport concept).
P_l=5/(8+5)=0.38*100=38%(education concept)
} Doc1

For document 2 (Doc_2)

1^2	2	2^2	1	3^2	8
Medical		**Transport**		**Education**	

P_j=0.18*100= 18% (medical concept).
P_k=9%(transport concept).
P_l=73% (education concept)
} Doc2

For document 3 (Doc_3)

1^3	1	2^3	9	3^3	0
Medical		**Transport**		**Education**	

P_j=10% (medical concept)
P_k=90% (transport concept)
P_l=0% (education concept)
} Doc3

For document 4 (Doc_4)

1^4	11	2^4	0	3^4	4
Medical		**Transport**		**Education**	

P_j=73% (medical concept)
P_k=0% (transport concept)
P_l=27% (education concept)
} Doc4

For the categorization part of documents, each **document concept class matrix** has been divided into 4 parts. First

part is document name, 2nd part is main concept, based on highest percentage, 3rd part is next sub concept or NULL (if Pure class) based on next highest percentage of concept, followed by 4th part with next sub category or NULL. The structure of **concept class Matrix** is shown below in table 2.

Table2: Document Concept Class Matrix

Document No.	Main concept	Sub concept/NULL	Next Category/NULL

The figure 1 shows link based document categorization based on concept analysis, where the category of Doc_1 is medical centric education oriented document whereas category of Doc_2 is education concept centric, medical concept sub oriented transport concept oriented document. The category of Doc_3 is transport concept centric medical concept oriented document whereas the category of Doc_4 is same as Doc_1.
So, we can categorize the above documents in 3 categories: (Doc_1, Doc4), (Doc_2), (Doc_3).

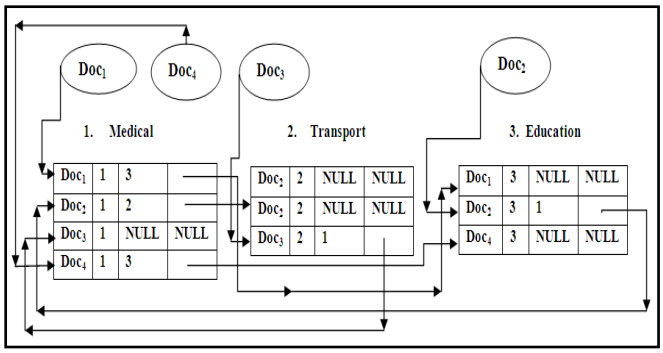

Fig 1: Link based document categorization based on concept Analysis

Now, after categorization, if we think about search engine, then to improve searching purpose, according to domain concepts, the documents which are of same category, they are ranked or arranged according to highest percentage of 1st domain weight, if 1st domain weight percentages are same, they are ranked according to 2nd domain and so on. As 1st domain weight percentage of Doc_4 is higher than 1st domain percentage of Doc_1, so for indexing purpose Doc_4 will come before Doc_1. Doc_2 and Doc_3 are of different concepts documents, so indexing has been done differently. This concept is shown in figure 2 below:

According to concept ranking:

Document	Index
Doc_4	1
Doc_1	2

Document	Index
Doc_2	1

Document	Index
Doc_3	1

Fig 2: Indexing of documents according to concept based category

3. CONCLUSION

Concept extraction of a particular document is the main focus of this paper. After that, the documents are linked for different domains of concept. In real life, sometime it happens that when a document belongs to different concept of different domains, it is very much essential to categorize them according to their real concept. In this paper we have shown how to improve the methodology of concept based document categorization based on link as well as we have shown the methodology for indexing documents for search engine.

REFERENCES

[1] Rubiya P U and Cinita Mary Mathew, "A Fuzzy Based Approach for Multilabel Text Categorization and Similar Document Retrieval," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, Issue 9, 2015.

[2] Soumen Swarnakar, Shubhalakshmi Ray and Tithi Mitra Chowdhury, **"**Comparative Study on Context-Based Document Clustering**,"** International Journal of Innovations in Engineering and Technology (IJIET), Vol. 4 Issue 1, PP 219-227, June 2014.

[3] Soumen Swarnakar , Roshni Roy , Shriti Singh, Ritika Rohini and Paulami Gorai, "A new Approach to concept based document clustering and comparative study with Hierarchical Clustering," International Journal of Computer Engineering and Applications, Volume IX, Issue IV, April 15.

[4] Jayaraj Jayabharathy and Selvadurai Kanmani, "Correlated concept based dynamic document clustering algorithms for newsgroups and scientific literature," Decision Analytics Journal, February 2014.

[5] Joel W. Reed et al., "TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams," The Fifth International Conference on Machine Learning and Applications, ICMLA 2006, Orlando, Florida, USA, 14-16 December 2006.

BIOGRAPHIES

Prof. Soumen Swarnakar, Assistant Professor, Department of Information Technology, Netaji Subhash Engineering College, Techno City, Garia, Kolkata-700152, India.
Email id: soumen_swarnakar@yahoo.co.in

Sangita Karmakar, B.Tech Student, Department of Information Technology, Netaji Subhash Engineering College, Technocity, Garia, Kolkata-700152, India.
Email id: sangitakarmakar1995@gmail.com