Knowledge Retrieval System Development by using Vector Space Model

Sasithorn Lertariyatham*, Pongpisit Wuttidittachotti*, Somchai Prakancharoen* ,
and Sakda Arj-ong Vallipakorn**
* Faculty of Information Technology, King Mongkut's University of Technology North
Bangkok 10800
** Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand
10400
E-Mail: mali3fja@gmail.com

Abstract

This research was conducted to develop a system for data retrieval stored in the form of questions and answers by the Long text Matching for facilitate to user in Vanich International Co., Ltd. Make Index from word that consist from Long text Matching process by the Inverted Indexing. Sort and search output by Vector Space Model(VSM). The experimental results of 100 questions process by 50 queries using Cosine formula. The result is 87.50 percent of the precision, 89.70 percent of the recall, and 85.50 percent of accuracy by F-measurement.

Key words: Long text Matching, Information Retrieval, Vector Space Model, Lucene

1. Introduction

Vanich International Co., Ltd., Thailand has been used Google Drive for storage and retrieval of knowledge system. When data has large size, Google Drive will inconvenient for retrieval knowledge . Therefore ,development of knowledge retrieval system and show output in the form of questions and answers for increase quality of knowledge retrieval. The system is develop VB.Net language and cut words by Long text Matching, then make Index from word that consist from Long text Matching process by the Inverted Indexing. User can retrieval any words in question. System will show question list , example question that has words retrieval by user and link to show answers. System admin must add keywords, Question and Answer to system before user used system

2. Theory and Related Research

2.1 Theory

Information Retrieval is process that manages the representation, storage and access of document or data.

Information Retrieval System is tool that link between user and Information Collection. The purpose is retrieval from large data with high quality. Nowadays, Retrieval system is importance factor of IT SYSTEMS. Example Search Engine and knowledge retrieval system of national library.

Question-Answering System is one type of Information Retrieval was a request (Query) in form of natural language, and got correct and fast answer results. The process of information retrieval system selected and retrieved only information that user needs. The main concept is focus on the comparative importance of the documents, and term from users (Keywords Matching). The components of this information retrieval system are divided into two steps. Detail is in Figure 2-1) [1].

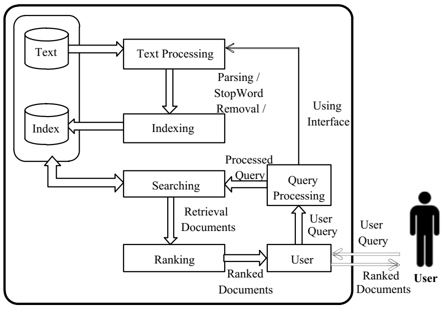


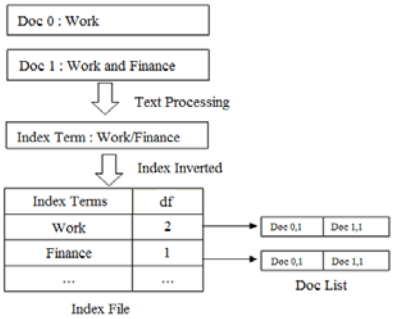
Figure 1 Architecture of retrieval system[5]

2.2 Inverted Indexing

Inverted File Index structure is favor method for general retrieval system because it has high quality in velocity and save space. Inverted File Index structure is consist by 2 elements [2].

- Words : All words in document.
- Document list : The document that include word.

Example Structure of Inverted File Index in Figure2
Document1 has "work".
Document 2 has "word and finance"



Index Terms	df
Work	2
Finance	1
...	...

Figure 2 Create index by Inverted File Index [3]

2.3 Vector Space Model

We use words from document rearrange into a vector format, each vector represented words in each documents. They compared the similarities of each document by measuring angle between vectors axis using cosine formula or dot product to measure angle of differences. From results, lower value showed similarities of result [4].

The cosine value is between 0 and 1. If the cosine is 0, it means no similarity. But if cosine value is 1, it means that document was very similar, as shown in Figure 3.

X is the horizontal main axis.

Y is the main vertical axis.

Doc 1 is words in a document 1 that has been sorted into a vector format.

Doc 2 is words in a document 2 that has been sorted into a vector format.

“q” is query terms of user retrieval.

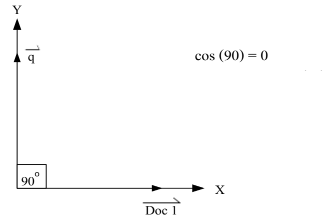


Figure 3: Document is not resembled with Query.

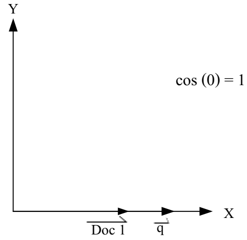


Figure 4: Document is most resembled with Query.

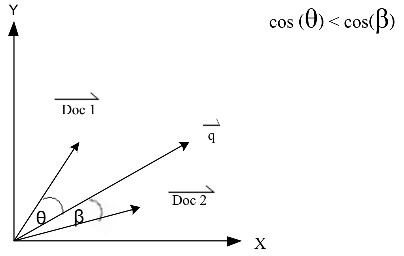


Figure 5: Query was resembled to the Doc 1 less than the Doc2.

The retrieval system will showed results from high to low similarity scores between correlation of queries and documents of cosine or dot-product of vector of Vector Space Model.

TF is frequency of retrieval term.

$$\text{TF (t in d)} = \sqrt{frequency} \qquad (1)$$

IDF(t) is the frequency of retrieval term show in documents

$$\text{IDF (t)} = 1 + \log\left(\frac{numDocs}{docFreq+1}\right) \qquad (2)$$

by numDocs = number of document
docFreq = number of document that include t word

Find weight form formula below.

$$\text{weight} = \text{TF} * \text{IDF} \qquad (3)$$

So, the order of document from 1 to N will assign as Doc and write in the matrix system as below.

$$\text{Doc} = [\text{Doc}_1 \ \ \text{Doc}_2 \ \ldots \ \text{Doc}_j \ \ldots \ \text{Doc}_n]$$

Table 1: Sort Weight of words in each document.

	Term $_1$	**Term $_2$**	**Term $_3$**	**…**	**Term $_n$**
d_1	$w_{1,1}$	$w_{1,2}$	$w_{1,3}$	…	$w_{1,n}$
d_2	$w_{2,1}$	$w_{2,2}$	$w_{2,3}$	…	$w_{2,n}$
…	…	…	…	…	…
d_N	$w_{N,1}$	$w_{N,2}$	$w_{N,3}$	…	$w_{N,n}$

(1)

Table 1 showed results of weight of word in each documents and change in to vector format as below;

Vector of document 1.

$$\{ w_{1,1} , w_{1,2} , w_{1,3}, \ldots , w_{1,2n}\}$$

Vector of document 2. (2)

$$\{ w_{2,1} , w_{2,2} , w_{2,3}, \ldots , w_{2,2n}\}$$

Vector of document N. (3)

$$\{ w_{N,1} , w_{N,2} , w_{n,3}, \ldots , w_{n,n}\} \qquad (4)$$

Docj can wrote in to

$$\begin{aligned} \text{Doc}_j &= [w_{1,j}\ w_{2,j} \ldots\ w_{i,j}\ \ldots\ w_{M,j}] \\ w_{i,j} &= tf_{i,j} \text{ x } idf_i \qquad (5) \\ tf_{i,j} &= f_{i,j} \ / \ (\text{Max}_k f_{k,j}) \\ ifd_i &= \log\ (N \ / \ n_i) \end{aligned}$$

Similarity calculation is similarity score

$$\text{Sim}(q, d_j) = \frac{\sum_{i=1}^{M} w_{i,q} \times w_{i,j}}{\sqrt{\sum_{i=1}^{M} w_{i,q}^2} \times \sqrt{\sum_{i=1}^{M} w_{i,j}^2}} = \cos\theta \qquad (6)$$

In case of q show same direction as d_j will got cosine = 1, that is maximum of matched. If q has angle of 60 degree d_j will got cosine = 0.5. If q has angle of 90 degree d_j will got cosine =0 , that is unmatched.

For sorting document based on similarity from 1 to N and the results will use to compare by d_k important ranking than sequences of d_j e.g. Sim (q, d_k) >Sim (q, d_j).

3. Testing

3.1 System Design

System was developed under web application in VB.Net language using POI library from Apache Software Foundation. Cut word process used ThaiAnalyzer library for Thai language and used word database from Lexitron Dictionary from National Electronics and Computer Technology Center (NECTEC). Use Lucene library from Apache Software Foundation for create index and retrieval information to cut word by Long text Matching ,then find TF and Weight to save in to database.

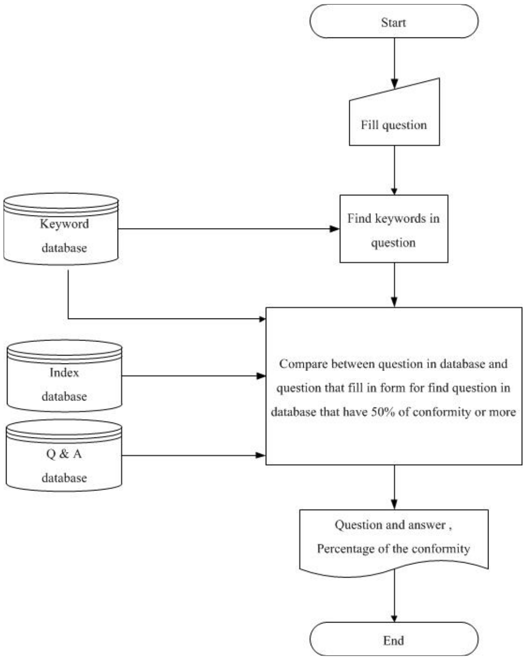
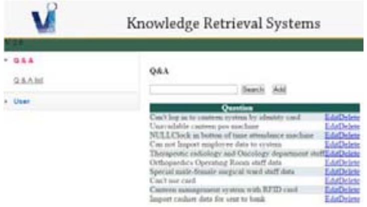

Start
Fill question
Keyword database
Find keywords in question
Index database
Q & A database
Compare between question in database and question that fill in form for find question in database that have 50% of conformity or more
Question and answer , Percentage of the conformity
End

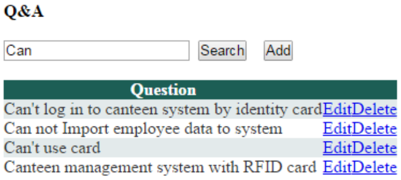
Figure 6: Workflow processes of the Knowledge Retrieval System



Q&A

Question	
Can't log in to canteen system by identity card	EditDelete
Unavailable canteen pos machine	EditDelete
NULLClock in bottom of time attendance machine	EditDelete
Can not Import employee data to system	EditDelete
Therapeutic radiology and Oncology department staff	EditDelete
Orthopaedics Operating Room staff data	EditDelete
Special male-female surgical ward staff data	EditDelete
Can't use card	EditDelete
Canteen management system with RFID card	EditDelete
Import cashier data for sent to bank	EditDelete

Figure 7: Screen shows all question in system (Admin).

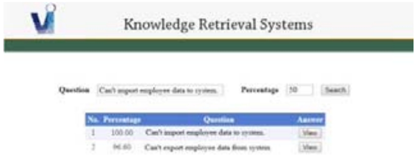


Q&A

Can

Question	
Can't log in to canteen system by identity card	EditDelete
Can not Import employee data to system	EditDelete
Can't use card	EditDelete
Canteen management system with RFID card	EditDelete

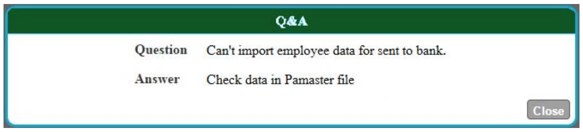
Figure 8: Screen shows result of question search (Admin).



Question: Can't import employee data to system. Percentage: 50

No.	Percentage	Question	Answer
1	100.00	Can't import employee data to system.	View
2	96.60	Can't export employee data from system	View

Figure 9: Screen shows result of question search (User).



Question: Can't import employee data for sent to bank.

Answer: Check data in Pamaster file

Figure 10: Pop up shows question and answer

5. Conclusion

Knowledge retrieval system based on VSM is tested by 100 questions process by 50 queries using Cosine formula. The result is 87.50 percent of the precision, 89.70 percent of the recall, and 85.50 percent of accuracy by F-measurement.

6. References

[1] Qiu, J., Yao, Y., Wang, Y. and Wang, X. (2006). Research of E-Government Knowledge Navigation System Based on XTM. *Proceedings of the 2006 IEEE/WIC/ ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IATW'06)*.

[2] Haruechaiyasak C.(1994). Developing IR System via Lucene. *National Electronics and Computer Technology Center (NECTEC).*

[3] Wuttikriengkraipol, A. (2011). Optimization to Information electronics file retrieval system, by Fuzzy logic. Special problem of Science , Master of Science in Information Technology, Faculty of Information Technology, *King Mongkut's University of Technology North Bangkok*.

[4] Welukanon, T. and Prakancharoen, S. (2011). *A Case Study of Administrative Court The 23rd National Graduate research Conference*, pp.190-196.

[5] Panyalerkchai, S. and Nuchitprasitchai, S. *Information Retrieval System Using***N-Gram** *Technique. King Mongkut'sUniversity of Technology North Bangkok. 2010*