

Music Information Retrieval System based on Vector Space Model

Kyawt Yin Khaing, Hnin Min Oo
University of Computer Studies, Mandalay
kyawtyinkhaing15791@gmail.com

ABSTRACT

Nowadays, Information Technology is rapidly improving to retrieve and integrate information from multiple data source especially on the Internet. To facilitate the information retrieval system to be more efficient and effective, the system implements information retrieval which the documents related with the music information. Information Retrieval (IR) addresses the problem of retrieving specific information from a collection of documents. The Information Retrieval System has the task of retrieving text documents that answer a user query. In this system, users run queries for retrieving data about composer name, artist name, song title and album title. The system imposes music information retrieval system by applying vector space model and cosine similarity measures. By using this system, the user can instantly retrieve and get the information of music information.

Keywords: information retrieval, vector space model, cosine similarity, music information

1. INTRODUCTION

Information Retrieval finds information that matches their information needs. The user who needs information issues a query to the retrieval system through the query operations module. In IR system, the information is not structured; it is contained in free form in text. As information technology becomes indispensable for daily life, a huge amount of information is proliferated in the world. The speed and amount of the proliferation has been further accelerated by the advent of Internet and the available information is almost flooding [10]. From such a huge amount of various and noisy information, new tools are needed to

discover useful information or knowledge that demands of individual user. Aims at discovering valuable knowledge for users are information retrieval (IR). The heart of an IR system is its retrieval model. The model is used to capture the meaning of documents and queries, and determine from that the relevance of documents with respect to queries. The quality of a retrieval system can principally only be determined through the degree of satisfaction of its users. This paper uses vector space model (VSM) for music information retrieval and cosine similarity measure.

The rest of this paper is organized as follows. Section 2, related work of IR is presented. Section 3 proposes a background theory of IR system. This system retrieves musical information such as (artists, album, song title and composer). In section 4, VSM and Cosine Similarity are explained. In section 5, detail of the system design is presented. Section 6 presents the Implementation of the system. The conclusion is given in section 7.

2. RELATED WORK

A.B. Manwar, HemantS.Mahalle, K.D. Chinchkhede and Dr.VinayChavan [1] implemented vector space model, considering term-frequency, inverse document frequency measures, achieves utmost relevancy in retrieving documents in information retrieval.

Nicole Kelly [8] used IR using VSM. The VSM of IR provides to search enormous compilations of documents and return documents in a database as relevant or not by modeling information as a vector and evaluating the cosine between those vectors.

In IR documents are usually given as extensional vectorial representation, in which the dimensions (features) of the vector representing a document are the terms occurring in the document. The approach to term representation that the IR community has almost universally adopted is

known as the bag-of-words approach presented by Nicola Poletini [9].

There have been multiple methods of IR used to solve the dilemma, but this system uses the vector space model. In the VSM, the dimensions of the vector representing a document are the terms occurring in the document. For VSM, the weight is positive and non-binary. Weights are used for calculating degree of similarity between each document and the query. The ranked document set in the decreasing order of degree of similarity is precise than the result of Boolean model.

3. INFORMATION RETRIEVAL SYSTEM

Information Retrieval (IR) System selects and returns to the user desired documents from a large set of documents in accordance with criteria specified by the user functions document. This system also presents the task of representing, storing, organizing, and offering access to information items. An IR model governs how a document and query are represented and how the relevance of a document to a user query is defined.

IR is finding material of an unstructured nature that satisfies information need from within large collections. There are three main IR models: Boolean model, Vector space model, Language model or probabilistic model. The most commonly used models in IR systems and on the Web are these above models [3].

Vector Space Model is an algebraic model for representing text documents as vector of identified. It is used information filtering, information retrieval, indexing and relevancy ranking. VSM is used in this system. Firstly, when the user input enters, this system is split into individual words. These individual words are comparing with the synonym table in the database. After comparing with the synonym table, the resulted output words are used by VSM. And the system shows the rank list of music information by calculating cosine similarity measure. Then, this system also gives the user download and listen song.

3.1. Tokenization

The first part of the process prepares text for further analysis. Tokenization is the task of segmenting a document into words. Input keywords are tokenized into individual words. In this system, tokenization is the removal of space between each of the words. The example of input words is tokenized into individual words is as follows:

For example:

Input: Chit Thu Nge Chin Myar Swar

Output: Chit/Thu/Nge/Chin/Myar/Swar

3.2. Converting to Target Word Using Synonym Table

The dictionary defines synonym as a word that means exactly the same as another word, or a word that can be interchanged with another word. The task of administrator is to define equivalent names for an input word in the database because there are many spelling for a word in Myanglish. The synonym table converts the specific word to correctly retrieve information for this system. In this system, the administrator creates a synonyms table in the database for singer name, composer name, song title and album title. Thus, the synonym table defines for a specific word (target word) has about four various spelling or different words. The example of synonym table for singer name is as follows:

Table 1: The example of synonym table for singer name

No	Input Word	Target Word
1	linn	lin
2	lyn	lin
3	lynn	lin
4	lin	lin

4. VECTOR SPACE MODEL

Vector space model is one of the techniques to get the necessary and extract information. It is used as a way to compare document to each other. Each document is vector [3]. Each vector has as many dimensions as there are terms in the documents

space. Vector represents documents and models both textual contents and structure information. Similarities are measured on the basis of feature vectors.

VSM creates a space between the documents that are represented by vector. Each dimension corresponds to a separate term. If a term occurs in the document, its value in the vector is non-zero. Several different ways of computing these values, also known as (term) weights, have been developed. One of the best known schemes is tf-idf weighting. Vector operations can be used to compare documents with queries.

$$tf_{ij} = \frac{f_{ij}}{\max\{f_{1j}, f_{2j}, \dots, f_{vj}\}} \quad (1)$$

Where,

tf= term frequency
t=number of term
d=number of document

The inverse document frequency (idf_i) of term t_i is :

$$idf_i = \log \frac{N}{df_i} \quad (2)$$

Where,

idf = inverse document frequency
N = Number of document
df = document frequency

The final TF-IDF term weight formula is:

$$w_{ij} = tf_{ij} \times idf_i \quad (3)$$

Where,

w_{ij} = term weight
tf_{ij} = term frequency
idf_i = inverse document frequency

4.1. Document Relevance Ranking with Cosine Similarity Measures

It is often difficult to make a binary decision on whether a document is relevant to a given query. Unlike the Boolean model, the vector space model does not make such a decision. Instead, the documents are ranked according to their degrees of relevance to the query. One way to compute the degree of relevance is to calculate the similarity of

the query q to each document d_i, the document collection D.

There are many similarity measures. The most well known one is the cosine similarity, which is the cosine of the angle between the query vector q and the document vector d_j. Cosine similarity is also widely used in text/document clustering. Ranking of the documents is done using their similarity values. The top ranked documents are regarded as more relevant to the query. Another way to assess the degree of relevance is to directly compute a relevance score for each document to the query.

Cosine similarity in VSM is determined by using associative coefficients based on the inner product of the document vector and query vector, where word overlap indicates similarity. The inner product is usually normalized. The cosine coefficient measures the angle between the document vector and the query vector. Cosine similarity is one of the most popular similarity measure applied to text documents, such as in numerous information retrieval applications.

$$\text{cosine}(\vec{Q}, \vec{d}) = \frac{\sum_{i=1}^{|v|} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{|v|} (w_{ij})^2} \times \sqrt{\sum_{i=1}^{|v|} (w_{iq})^2}} \quad (4)$$

Where,

\vec{Q} = vector of querying document
 \vec{d} = vector of document
w_{ij} = weight of term document
w_{iq} = weight of querying document

4.2. Working Flow of Vector Space Model

In this section, the calculation steps of the vector space model applied in the system are mentioned with detail calculation steps. It can be supposed that an IR system for the query "music information". The database collection consists of three documents (D = 3) with the following content:

D₁: "Chit Thu Nge Chin Myar Swar"
D₂: "A Chit Sone Thu Nge Chin "
D₃: " Lwan Yet Myar Swar"

Table 2: Retrieval Result Table

Term vector model based on $w_{ij} = tf_{ij} \times idf_i$												
Query, Q: Thu Nge Chin												
D ₁ : "Chit Thu Nge Chin Myar Swar"												
D ₂ : "A Chit Sone Thu Nge Chin"												
D ₃ : "Lwan Yet Myar Swar"												
D=3; idf= Log(D/df)												
Terms	Q	Counts, tf_i			df _i	D/d f _i	idf _i	weights, $w_i = tf_i \times idf_i$				
		D1	D2	D3				Q	D1	D2	D3	
Chit	0	1	1	0	2	3/2 =1.5	0.176	0	0.176	0.176	0	0
Thu	1	1	1	0	2	3/2 =1.5	0.176	0.176	0.176	0.176	0.176	0
Nge	1	1	1	0	2	3/2 =1.5	0.176	0.176	0.176	0.176	0.176	0
Chin	1	1	1	0	2	3/2 =1.5	0.176	0.176	0.176	0.176	0.176	0
Myar	0	1	0	1	2	3/2 =1.5	0.176	0	0.176	0	0	0.176
Swar	0	1	0	1	2	3/2 =1.5	0.176	0	0.176	0	0	0.176
A	0	0	1	0	1	3/1 =3	0.477	0	0	0.477	0	0
Sone	0	0	1	0	1	3/1 =3	0.477	0	0	0.477	0	0
Lwan	0	0	0	1	1	3/1 =3	0.477	0	0	0	0	0.477
Yet	0	0	0	1	1	3/1 =3	0.477	0	0	0	0	0.477

First of all, total documents (D) to retrieve in this example are three documents. Depending on the terms included in these documents and the terms in the user input query (music information), the system calculates the weight (w_i), term frequency or number of times a term occur in a document (tf_i), document frequency or number of document containing a term (df_i) and inverse document frequency (idf_i). Depending on these results, the system will calculate the similar values of vector space model. Similarity analysis of vector space model is described in section 4.3.

4.3. Similarities Analysis of Vector Space Model

First for each document and query, VSM computes all vector lengths (zero terms ignored).

The normalization of the weight of document (D_1, D_2 and D_3) is as follow:

$$|D_i| = \sqrt{\sum_t w_{i,j}^2}$$

$$|D_1| = \sqrt{0.176^2 + 0.176^2 + 0.176^2 + 0.176^2 + 0.176^2 + 0.176^2} = 0.43$$

$$|D_2| = 0.76$$

$$|D_3| = 0.72$$

The calculation of the normalization of Query Document (Q) is as follow:

$$|Q| = \sqrt{\sum_i w_{Q,j}^2}$$

$$|Q| = \sqrt{0.176^2 + 0.176^2 + 0.176^2} = 0.305$$

The vector space model calculates the dot product as the next step after above calculation.

$$Q \cdot D_i = \sum_t w_{Q,j} w_{i,j}$$

$$Q \cdot D_1 = (0.176 \cdot 0.176) + (0.176 \cdot 0.176) + (0.176 \cdot 0.176) = 0.093$$

$$Q \cdot D_2 = 0.093$$

$$Q \cdot D_3 = 0$$

In the next step, the calculation steps of cosine similarity are as follows:

$$\text{cosine } \theta_{D_i} = \text{sim}(Q, D_i)$$

$$\text{Sim}(Q, D_i) = \frac{\sum w_{Q,j} w_{i,j}}{\sqrt{\sum_j w_{Q,j}^2} \sqrt{\sum_i w_{i,j}^2}}$$

$$\text{cosine } \theta_{D_1} = \frac{Q \cdot D_1}{|Q| \cdot |D_1|} = \frac{0.093}{0.305 \cdot 0.43} = 0.71$$

$$\text{cosine } \theta_{D_2} = \frac{Q \cdot D_2}{|Q| \cdot |D_2|} = \frac{0.093}{0.305 \cdot 0.76} = 0.40$$

$$\text{cosine } \theta_{D_3} = \frac{Q \cdot D_3}{|Q| * |D_3|} = 0$$

Above mentioned answers are the results of similarities of the documents and query of songs (Q, D₁, D₂ and D₃) by calculating cosine similarity.

4.4. Ranking Documents

When the similarities of the documents are calculated, they are arranged according to the rank order. The ranking function sorts the documents according to the degree of similarity to the query. In this system, it sorts and ranks the documents in descending order according to the similarity values and shows these results to the users.

Rank 1: Doc 1 = 0.71
 Rank 2: Doc 2 = 0.40
 Rank 3: Doc 3 = 0

5. PROPOSED SYSTEM DESIGN

This system consists of two parts: administrator and user. Figure 1 shows the proposed system design for the administrator. In this part; the administrator must enter name and password. If the name and password is correct, the administrator can either upload or delete music information in the song database.

The proposed system design for the user is presented in figure 2. From the user point of view, the user can search music information by album title, singer name, composer name and song title. Vector space model is used for searching music information.

In this system, firstly the user input is split into individual word. For example, let the user input is "Chit Thu Nge Chin Myar Swar" is split into individual words (Chit/Thu/Nge/Chin/Myar/Swar). After tokenization, each individual word is compare with the word in synonym table in the database. For example the spelling or word "lin" may be "lin" or "lynn" or "linn" or "lyn". The synonym table converts the specific word (target word) to correctly retrieve information for this system. And then, the resulted output word compared with the word in synonym table is used by VSM. And the system shows the rank list of music information by using cosine similarity

measure. Finally, this system gives the user can listen and download the MP3 song format.

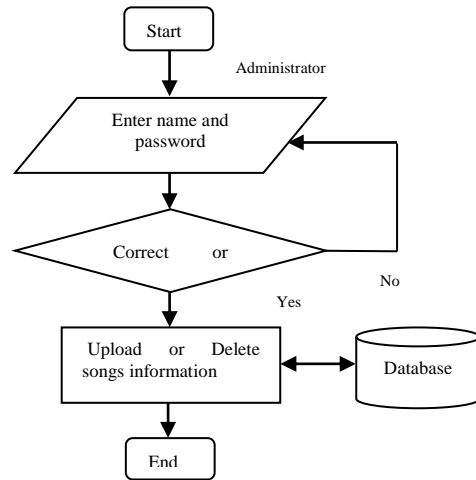


Figure 1: The Proposed System Design for the administrator

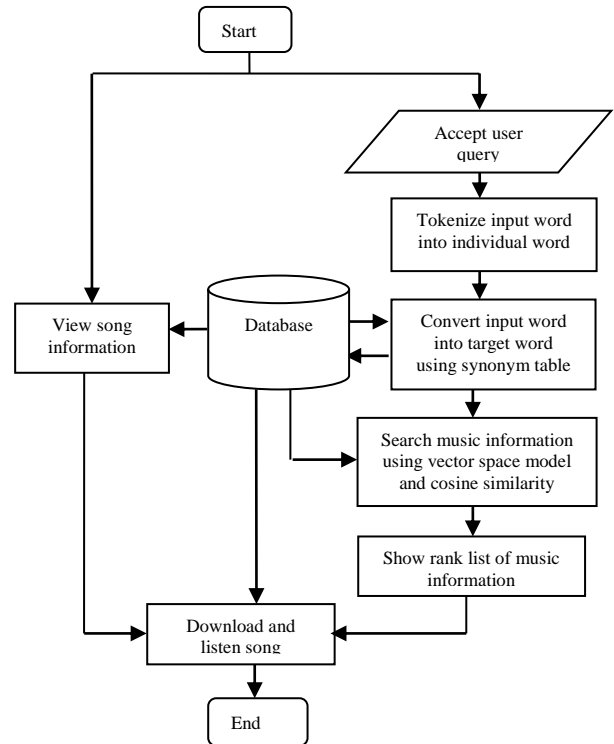


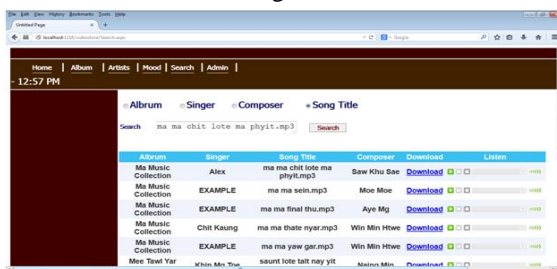
Figure 2: The Proposed System Design for the user

6. IMPLEMENTATION OF THE SYSTEM

This system emphasis on the music information which helps people to search information about music. This system is capable of extracting music information from Database by using Vector Space Model. Vector Space Model and cosine similarity are used to calculate the similarity of the music information by the user input. This system has two parts: user view and administrator view.

From the side of user, the user can search music information from Home Page, Album Page, Artist Page and Mood Page. In Home Page, the user can view the singer name by selecting A to Z of user appropriate choice. Album Page consists of the Album Name which is sorted years by descending order. By selecting Male, Female or Group, user can find singer name in the Artists Page. In this system, Mood Page consists of the song type (Classic, Pop, Rock, Country, Hip-Hop and R&B). By choosing one of the above song types, user can view the song information of the corresponding song type. And then, user can download and listened song of the users' search.

On the other hand, user can search music information by song title, album title, singer name and composer name by using VSM. And the system shows the rank list of songs information by using cosine similarity. Then, user can listen and download the song information of their searching. Figure 3 proposes that the user searches the song title *“ma ma chit lote ma phyt”* when the system shows the rank list of song information as follows:



Album	Singer	Song Title	Composer	Download	Listen
Ma Music Collection	Alex	ma ma chit lote ma phyt.mp3	Saw Khu See	Download	Listen
Ma Music Collection	EXAMPLE	ma ma sein.mp3	Moe Moe	Download	Listen
Ma Music Collection	EXAMPLE	ma ma final thu.mp3	Aye Mg	Download	Listen
Ma Music Collection	Chit Kaung	ma ma thate nyar.mp3	Win Min Hwa	Download	Listen
Ma Music Collection	EXAMPLE	ma ma yaw gar.mp3	Win Min Hwa	Download	Listen
More Than Year	Mhoo Min Thoo	saunt lote lalt day yit	Mahon Min	Download	Listen

Figure 3: Searching “song title” in Search Page of the System

When the user searches the music information by singer name, the system shows the singer’s name. Secondly, if the user clicks the “View Details”, the system describes the detail of the music

information of the singer. Then, the user can listen and download the MP3 song format.

In the admin view, the administrator must enter Name and Password. If the Name and Password is correct, the administrator can either Upload or Delete music information.

7. CONCLUSION

This system is more flexible and convenient way for user to access. This system will present more relevant music information to the user ordered by similarity. The main components of a search system are the task of collecting music information and IR system which has the task of retrieving text documents that retrieve a user query.

This system uses the VSM and cosine similarity. VSM and cosine similarity are used for retrieving songs information and their similarity songs by the user input. Furthermore, the system will satisfy the user by typing different spelling (e.g. linn or lynn or lyn or lin take a word lin) by using synonym table in the database. Finally, the system gives the user that can listen and download the MP3 song format.

8. REFERENCES

- [1] A.B.Manwar, HemantS.Mahalle, K.D. Chinchkhede & Dr. VinayChavan, “A Vector Space Model for Information Retrieval: A Matlab Approach”, Indian Journal of Computer Science and Engineering (IJCSE), Apr-May 2012.
- [2] Bing Liu, “Web Data Mining: Exploring Hyperlinks, Content and Usage Data”, Second Edition, Department of Computer Science, University of Illinois, Chicago, ISBN 9783642194597.
- [3] Gerard Salton & Michael J.McGill, “Introduction to Modern Information Retrieval”, MC Graw-Hill, 1983.
- [4] Nicole Kelly, “Information Retrieval Using Vector Spaces”, May 10, 2012.
- [5] S. Karen “Vector Space Model in IR”, San Francisco, California, 2007.

