Information Retrieval System using Vector Space Model

Kyu Kyu Swe, Thein Htay Zaw
Computer University, Mandalay
kkyu1312@gmail.com

Abstract

The heart of an information retrieval 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. As large sets of documents are now increasingly common, there is a growing need for fast and efficient information retrieval algorithms. The algorithms are embedded in the vector space model. The simple vector space model is based on literal matching of terms in the documents and the queries. This paper implements digital library information retrieval system. In this paper, when user's query is input to the system, system computes terms-document matrix of weight for information retrieval. Then the similarity is computed between query and all documents and the retrieved documents are ranked using similarity measure methods. Finally the system analyzes with precision and recall of information retrieval results.

1. Introduction

Information Retrieval System is used to handle records and data, and retrieve complete user information need. A Ranking algorithm operates according to basic premises regarding the notion of document relevance. There are three classic models in Boolean model, Vector model, and Probabilities model. In this paper, vector model is used to develop ranking algorithm. Retrieval technique is proposed that rank the result based on phase-based similarity measure between the documents and the query. In the vector space model, we represent documents as vectors. It matches the search words to the words in the index and assigns a weight to each word based on the number and position of the word in each document. And then it uses the assigned weight to rank the documents based on their similarity. Display the ranked retrieval list according their order. The system searches its database for documents that are related to the user's query, and presents those that are most relevant.

The system presents information retrieval system for library information based on vector space model. The rest of the paper is described as follows: section 2 presents the related works of the system. Section 3 is included with theory of information retrieval. Section 4 is fulfilled with the design and section 5 is implementation of the system. The conclusion of this system will be combined at the last section 6.

2. Related works

T. Mandl proposes that neural network based information retrieval system. This paper describes Neural networks are well suited for Information Retrieval (IR) from large text or multimedia databases. Their capacity for tolerant and intuitive processing offers new perspectives in IR where the vague nature of human relevance judgments has confronted theory and systems with considerable problems. Latent Semantic Indexing (LSI) to information retrieval using a neural back propagation network. The transformation between two representation schemes is enabled through preprocessing by LSI which is based on Singular Value Decomposition (SVD) [6].

Waban, J. Lofstrom, S. Perttu and K. Valtonen [7] combine link analysis with discrete PCA (a semantic component method) to develop an auxiliary score for information retrieval that is used in post-filtering documents retrieved via regular tf-idf methods. For this, they use a topic-specific version of link analysis based on topics developed automatically via discrete PCA methods. To evaluate the resultant topic and link based scoring, a demonstration has been built using the Wikipedia, the public domain encyclopedia on the web.

3. Information retrieval system

Information retrieval system also supports the user in the query formulation, e.g. through visual interfaces. Similarity of the structured representations is used to model relevance of information for users. As a result a selection of relevant information items or a ranked result can be presented to the user. Since information retrieval

system deals usually with large information collections and/or large user communities, the efficiency of an information retrieval system is crucial. This imposes fundamental constraints on the retrieval model. Retrieval models that would capture relevance very well, but are computationally prohibitively expensive are not suitable for an information retrieval system.

To make searching more efficient, a retrieval system stores documents in an abstract representation that can efficiently isolate those documents likely to relate to the users' needs. The response to a query is constructed using the indices and the operations provided by the underlying conceptual model. The user has to know something about this model in order to structure queries correctly and efficaciously.

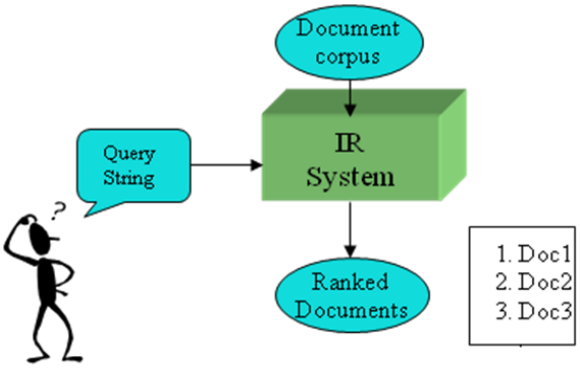

Figure 1. IR system

An information retrieval system's conceptual model defines how documents in it database are compared. In a system using an inverted file of keywords, two documents may be considered "similar" if they if they have a certain number of keywords in common. A system that uses a term frequency vector model would consider two documents to be similar when their term vectors lie close together in the vector space [4].

3.1. Vector space model

The vector space model is implemented by creating the term-document matrix and a vector of query. Let the list of relevant terms be numerated from 1 to m and documents be numerated from 1 to n. The term-document matrix is an m × n matrix A = [a_{ij}], where a_{ij} represents the weight of term i in document j [5]. On the other side, there is a query or customer's request. In the vector space model, queries are presented as m-dimensional vectors. The simple vector space model is based on literal matching of terms in the documents and the queries but literal matching of terms does not necessarily retrieve all relevant documents.

When document vectors reflect the frequencies with which terms appear, documents are considered similar if their term vectors lie close together in vector space. Before determining distance, the dimensions of the space should be normalized in a way that reflects the differing importance of different words. Importance is generally measured simply on the basis of word frequency, rare words being more salient than common ones.

Each term is weighted by the number of documents in which it appears, so that a single appearance of the counts far less than a single appearance of, say, Jezebel. The components of the term vector are not simply term frequencies, but term frequencies divided by the number of documents in which that term appears. This is called term-frequency time inverse document frequency weighting, or tf- idf. Documents are represented by using the vector- space model. This model uses weight each term based on its inverse document frequency (idf) in the document collection [2]. The weight of a term t_i in a document is given by

$$w_i = tf_i \times idf_t \quad (1)$$

where tf_i, i= 1,....,n is the term frequency of the term t_i in the document. The term frequency can also use to the normalization of term frequency that is defined as

$$tf_i = 0.5 + 0.5\,(tf_i) / \max(tf_i) \quad (2)$$

"idf" is inverse document frequency which is defined by

$$idf = \log(N / df_i) \quad (3)$$

where df_i is the number of documents in a collection of N documents in which term t_i occurs [3].

3.2. Similarity measure

A key factor in the success of information retrieval system is the similarity measure between query and documents. In order to be able to group similar data objects a proximity metric has to be used to find which documents (or clusters) are similar. The calculation of the dissimilarity between two objects is achieved through some distance function, sometimes also referred to a dissimilarity function. Given two feature vectors **x** and **y** representing two objects it is required to find the degree of similarity or dissimilarity between them. There is a large number of similarity metrics. A very common class of distance functions is known as Euclidean distance described as [1]:

$$\| x - y \| = \sqrt{\sum_{i=1}^{n} | x_i - y_i |^2} \quad (4)$$

A more common similarity measure that is used specifically in information retrieval is the cosine correlation measure, defined as:

$$\cos(x, y) = \frac{x.y}{\| x \| \| y \|} \quad (5)$$

where ||.|| indicates the length of the vector [3].

$$\text{sim}(D_i, D_j) = \frac{\sum_{t=i}^{N} w_{it} * w_{jt}}{\sqrt{\sum_{t=i}^{N} (w_{it})^2 * \sum_{t=j}^{N} (w_{jt})^2}} \quad (6)$$

3.3. Precision and recall

Precision and recall are two widely used measures for evaluating the quality of results in domains such as Information Retrieval and Statistical classification.

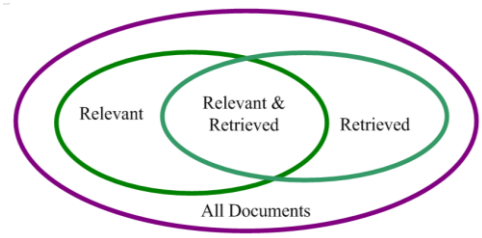

Figure 2. Precision and recall for IR system

$$precision = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|} \quad (7)$$

$$\text{Recall} = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|} \quad (8)$$

Precision can be seen as a measure of exactness or fidelity, whereas Recall is a measure of completeness. In an Information Retrieval scenario, Precision is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search, and Recall is defined as the number of relevant documents retrieved by a search divided total number of existing relevant documents (which should have been retrieved) [1].

4. System design

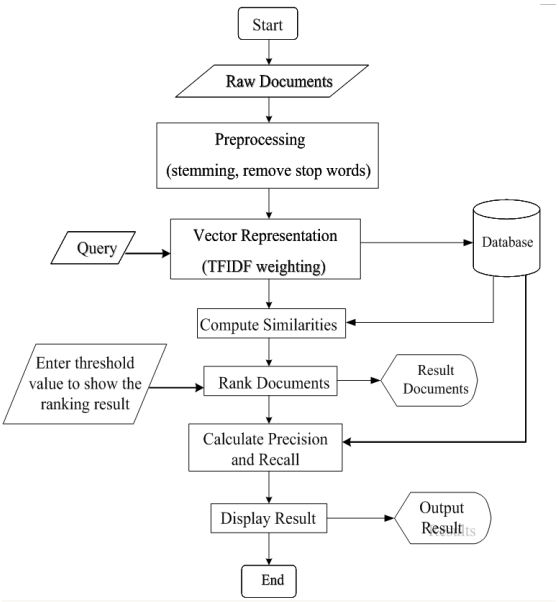

Figure 3. System flow for IR system

In this system, raw documents are abstract files. These files are put by the administrator. Preprocessing stage determines stemming (most useable words) and remove stop words (a, an, the, etc). Vector Representation calculates the weight of abstract files without stemming and removes stop words. Then, the vector representation put these abstract files into the database. Moreover, vector representation clusters and determine the weight of input abstract files from the query. The system compares the input abstract files and the existing abstract files in the compute similarities stage. Then, rank documents produce the similar abstract files. Moreover, this stage represents the most similarities results. System produces the results based on rank results to calculate precision and recall.

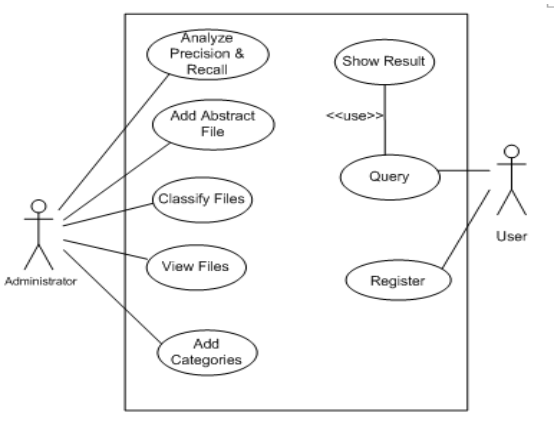

Figure 4. Use case diagram of IR system

This use case diagram is used for two users, administrator and user. Administrator can do the process of add new categories or classes, view and

update abstract files, classify files according to their classes, add new abstract files and analyze the output results by precision and recall. User can query as he/she wants to know the information, and then the system shows the results. User can view the abstract files.

5. System implementation

At the start of system, the user must input the query. And then, the system searches relevant document from database. The relevant document can be viewed depending on Ranking Algorithm. This Information Retrieval System has four processes. They are:

1. preprocessing the document with text mining
2. assigned a weight using vector space model
3. measuring document similarity
4. rank the documents based on their similarity.

If user wants to know the information, he/she must enter the query that wanted to know and clicks "search" button, the system displays the values of global weights of each term or keyword and the similarity values of each abstract files. The result of information retrieval is ranked according to similarity values. Query can be typed between 20 keywords and the system takes the query as a document. Figure 5 shows the weight of the abstract files from query. It is also shows the rank results by comparing data from the database.

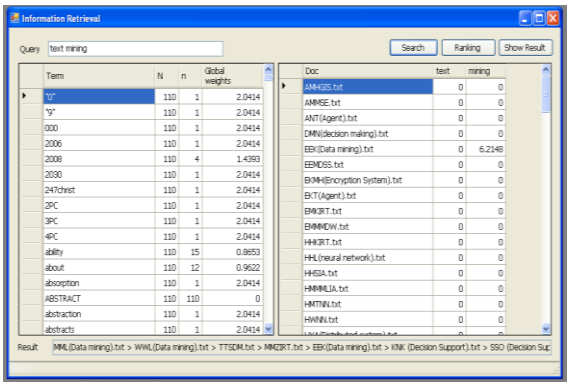

Figure 5. Query Result

If the user wants to know the most similarity results, user must enter the highest rank level. In Figure 6, click " Ranking " button, user can enter the threshold value of similarity to show the ranking result, and then the system displays the result.

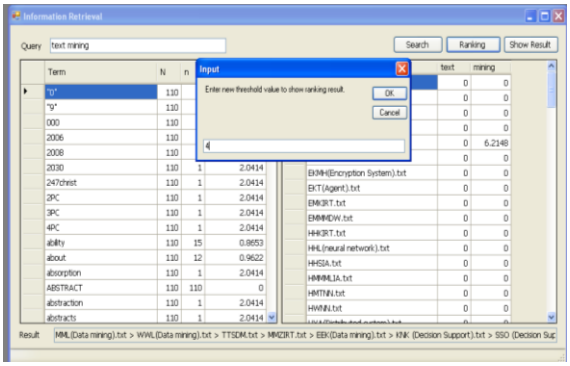

Figure 6. Entry of new threshold value from user

Figure 7 shows the threshold value of similarity results from the rank results. User can view the detail of the abstract file when click "Show file" button.

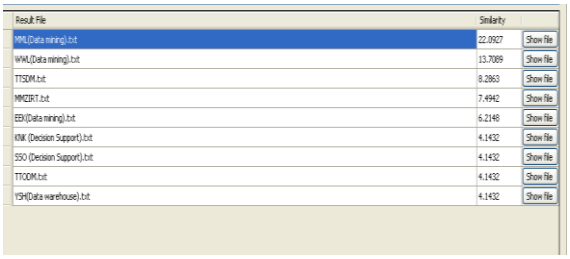

Figure 7. Result of user defined threshold value

In Define Document and Class, administrator must define the document with manual class. The administrator can define the abstract file with one or more class name by selecting the check box.

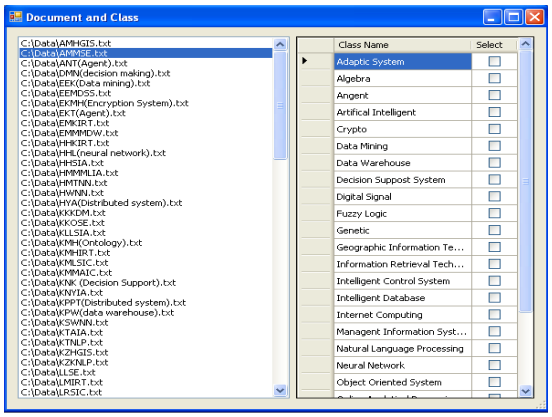

Figure 8. Define manual class menu form

The accuracy of the information retrieval system is analyzed by using precision and recall. When the value of precision and recall is near 1, the result is the best.

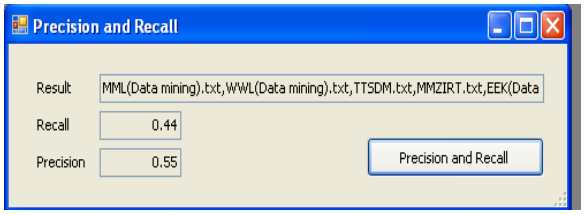

Figure 9. Result of precision and recall

6. Conclusion

All information retrieval can be performed manually, but automation has many benefits: larger document collections can be processed more quickly and consistently. The benefits of the system are as follow:

- using vector space model for IR can give more relevant user information.
- user can select the appropriate ones that are closed related with their information need conceptually.
- the vector model's term weighting scheme improves retrieval performance.
- its partial matching strategy allows retrieval of documents that approximate the query conditions.

In this system, the documents can be retrieved literally and not the document with polysemy and synonymy. One limitation of the system is just can read only text file. This system can be extended to compare the performance by using other uninformed search algorithms such as genetic algorithm, clustering algorithm, etc. This system can be extended to support latent semantic indexing and concept indexing, and both methods can retrieve all relevant documents and solve the problem of polysemy and synonymy.

References

[1] G. Salton and C. Buckley, “Term-Weighting Approaches in Automatic Text Retrieval”, Department of Computer Science, Cornell University Ithaca.

[2] IR Models: “The Vector Space Model”, Lecture 7.

[3] J. Han and M. Kamber, "Data mining Concepts and Techniques", Mining Text and Web Data, Department of Compute Science, University of Illinois at Urbana-Champaign.

[4] K. Abere, “Information Retrieval and Data Mining”, EPFL-IC, Laboratoire de system d'informations répartis.

[5] N. Polettini,,“The Vector Space Model in Information Retrieval - Term Weighting Problem”, Department of Information and Communication Technology, Italy University of Trento.

[6] T. Mandl, "Efficient Preprocessing for Information Retrieval with Neural Networks", Zimmermann, 7th European Congress on Intelligent Techniques and Soft Computing. Aachen, Germany.

[7] W. Buntine, J. Lofstrom, S. Perttu and K.Valtonen "Topic-Specific Link Analysis using Independent Components for Information Retrieval" Helsinki Institute for Information Technology (HIIT) Finland.