INFORMATION RETRIEVAL BASED ON CLUSTER ANALYSIS APPROACH

Orabe Almanaseer

Department of Information Technology, The University of Jordan, Aqaba, Jordan

ABSTRACT

The huge volume of text documents available on the internet has made it difficult to find valuable information for specific users. In fact, the need for efficient applications to extract interested knowledge from textual documents is vitally important. This paper addresses the problem of responding to user queries by fetching the most relevant documents from a clustered set of documents. For this purpose, a cluster-based information retrieval framework was proposed in this paper, in order to design and develop a system for analysing and extracting useful patterns from text documents. In this approach, a pre-processing step is first performed to find frequent and high-utility patterns in the data set. Then a Vector Space Model (VSM) is performed to represent the dataset. The system was implemented through two main phases. In phase 1, the clustering analysis process is designed and implemented to group documents into several clusters, while in phase 2, an information retrieval process was implemented to rank clusters according to the user queries in order to retrieve the relevant documents from specific clusters deemed relevant to the query. Then the results are evaluated according to evaluation criteria. Recall and Precision (P@5, P@10) of the retrieved results. P@5 was 0.660 and P@10 was 0.655.

KEYWORDS

Cluster Analysis, Documents analysis, Information Retrieval, Text Mining.

1. INTRODUCTION

The data mining is an interdisciplinary field that deals with the extraction of information from a large set of data and transformation into an easily interpretable structure for further use [1]. However, text mining is an emerging technology for analysing large collections of unstructured documents, which offers powerful possibilities for extracting interesting patterns from huge amount of unstructured data available online [2-6].

Moreover, text mining involves the application of techniques from areas such as Information Retrieval and Clustering Analysis [7]. Information retrieval (IR) has been changed dramatically in the last few years as the storage spaces and World Wide Web are considerably getting expanded. Web search is one of the information retrieval scenarios but it is not the only one, there are many other scenarios like searching emails, searching the contents of a laptop computer, finding stuff in some companies' knowledge base and so on. Clustering analysis, on the other hand, is a newly developed computer-oriented data analysis technique. It is a common method used in the psychological, social, and physical sciences to identify subgroups or profiles of individuals within the larger population who share similar patterns on a set of variables [8]. For the purpose of this paper, we will focus on the areas of information retrieval based on clustering analysis process.

2. BACKGROUND

2.1. Information Extraction

Information Extraction (IE) is analysing unstructured text in order to extract information about pre-specified types of events, entities or relationships [9]. In other words, it is the activity of automatically extracting structured information from unstructured sources [5]. IE systems can be used in natural language processing applications for finding and understanding limited relevant parts of texts and gathering information from many pieces of text in order to produce a structured representation of relevant information. The main goals of information extraction systems are organizing information so that it is useful to people, and putting information in a semantically precise form that allows further inferences to be made by computer algorithms [9].

2.2. Information Retrieval (IR)

Information Retrieval (IR) is the topic most commonly associated with online documents, and the main task of information retrieval is to retrieve relevant documents in response to a query. Figure 1, illustrates the objectives of information retrieval of documents, where (a) a general description is given of the query, (b) the document collection is searched, and (c) a subset of relevant documents is returned [4].

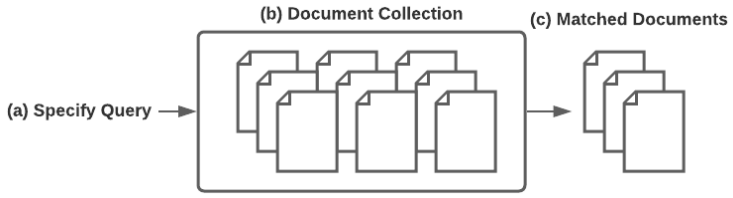

Figure 1. Information Retrieval Objectives [4]

According to Djenouri et al. [1], in Information Retrieval problem consider the set of m objects $\Lambda = \{\Lambda_1, \Lambda_2 ... \Lambda_m\}$ and the set of n terms $T = \{T_1, T_2, ..., T_n\}$. Each object Λ_i is a subset of terms in T ($\Lambda_i \subset T$, $\forall_i \in [1 ...m]$). Given the set of queries $Q = \{Q_1, Q_2, ..., Q_i\}$, where each query Q_i is composed by the set of terms, that is, $Qi \subset T$, the IR problem aims at finding, for each query $Q_i \in Q$, the most relevant subset of objects Λ', such that $\Lambda' \subset \Lambda$.

2.3. Clustering Analysis

Perhaps the most common theme in analysing data is cluster analysis. Clustering is a fully automatic process through which a collection of documents is classified into groups. The documents within each group are more closely related to one another than documents assigned to different groups [5]. Figure 2 illustrates the overall task of cluster analysis.

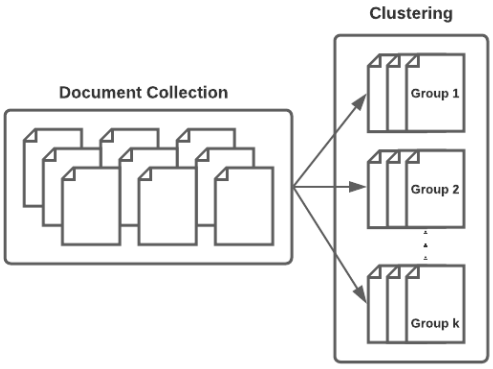

Figure 2. Clustering a Document Collection [5]

According to Tkach [10], the clustering tool identifies a list of terms or words which are common in the documents within one group, and it can also be done with respect to combinations of the properties of documents, such as their length, cost, date, etc. The most common clustering techniques are K-means clustering, fuzzy C-means clustering, mountain clustering and subtractive clustering [11]. For the purpose of this paper, we will use the K-means clustering analysis.

2.4. Cluster-Based IR

The traditional IR solutions need to scan the whole objects for every user query. This process is highly time-consuming, particularly for a large number of objects and queries. To deal with this problem, cluster-based retrieval solutions have been largely studied in the last decade [1, 12–14].
In cluster-Based retrieval, consider a set of k clusters $G = \{G_1, G_2...G_k\}$, where each G_i is represented by the set of objects $\{\Lambda^i_1, \Lambda^i_2, \ldots, \Lambda^i_{|Gi|}\}$ and consider a set of queries $Q = \{Q_1, Q_2, ..., Q_i\}$. Cluster-based retrieval aims at retrieving one or more clusters in G in response to every query in Q. The task is to match the query against clusters of objects instead of individual objects and rank clusters based on their similarity to the query. Solutions to cluster-based retrieval are aimed at reducing the time performance of the information retrieval process. Instead of processing the whole object databases, only the relevant clusters to the user query are explored [1]. In general, the existing cluster-based approaches are much faster than traditional approaches when applied to large collections.

3. PROPOSED FRAMEWORK

The purpose of this project lies on cluster-based retrieval in order to design a framework for to textual analysis to automatically analyse a collection of documents that are written in English language. The main task is to group documents into several clusters such that similar objects are grouped in the same cluster, and then the information retrieval is only performed on the clusters deemed relevant to a given user query (see Figure 3).

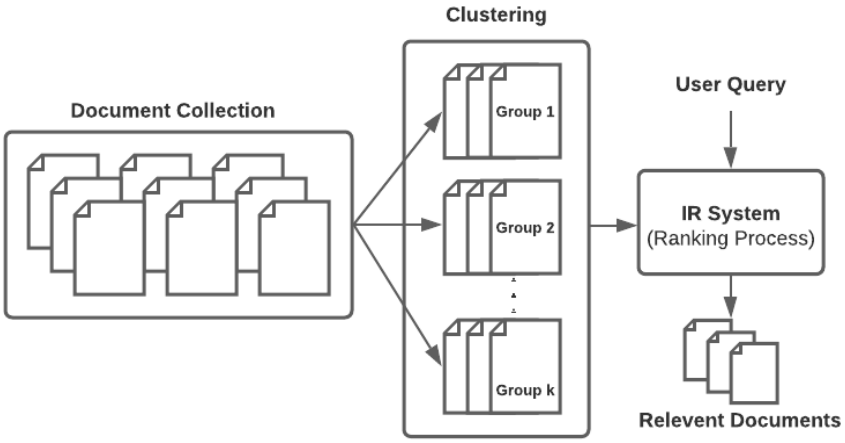

Figure 3. The System Framework

4. METHODOLOGY

In this paper, a proposal framework based on Cluster-Based Retrieval is adopted in order to design and develop a system for clustering and retrieving documents from documents collection (see Figure 3). The project is implemented through two stages. In stage 1, the clustering analysis is implemented to group documents into several clusters, while in stage 2; our task is to perform information retrieval process on the clusters deemed relevant to a specific query. Then the results are evaluated according to evaluation criteria.

5. DOCUMENT COLLECTION

For the purpose of this paper, thousands of documents were collected in a digital format. The selected data set contains of (180286) documents that are stored in (HTML) format and written in English language. The data sets differ in terms of document size, number of categories and average category size.

6. IMPLEMENTATION

In this project, the overall implementation process was based on the proposed framework (see Figure 3). The process consists of many different courses of actions including, document pre-processing, weight calculation, document representation and K-Means algorithm. In order to accomplish the overall implementation process, different tools and algorithms have been used in each step. The open source natural language processing library (genism) - which is implemented in Python – has been used to implement the pre-processing steps and (Cython) is used to build the document vectors.

6.1. Document Pre-Processing

The document pre-processing consists of two main steps including: tokenization and normalization. However, all HTML tags were removed from documents before starting processing them. The main use of tokenization is identifying the meaningful keywords called tokens. Tokenization splits sentences into individual tokens, typically words. In the normalization process, we should remove some unnecessary tokens in order to get proper result, these tokens

include stop words, special characters, unifying data, URLs, pronouns, adverbs, preposition and so on. This will reduce the storage size used for representing these tokens in the main memory.

The conversion into tokens and the pre-processing is mainly done using the open source libraries spaCy and regular expressions. spaCy is written in Python and Cython. Besides spaCy there is a wide range of libraries for NLP. Depending on the application, some of them are more suitable than others (cf. [15]). For this paper spaCy was mainly chosen because of its high performance and claimed accuracy for syntactic analysis in [15].

6.2. Weight Calculation

This step involves calculating the weight of each word using term frequency-inverse document frequency (tf-idf).

$$Wi = TF * IDF \tag{1}$$

whereas (TF) is the count of token (t) in character sequence cs, then the term frequency (TF) is defined by,

$$TF\ (t,cs) = sqrt(count(t,cs)) \tag{2}$$

And the df(t,docs) is the number of documents in which the token (t) appears, then the inverse term frequency (IDF) of (t) is defined by,

$$IDF(t,docs) = \log (D/df(t,docs)) \tag{3}$$

whereas, D is the number of the documents in the dataset. Therefore, the weight is calculated for each token as follows.

$$Wi = TF * IDF \tag{4}$$

6.3. Document Representation

For efficient retrieval, we employ the idea from Vector Space Model (VSM) which is an algebraic model for representing text documents and it is essential data structure used in information retrieval. In this project, a document vectors model is built as an object containing all terms in the documents collection, whereas each term in our index is an object holding the document frequency (df) of this term, and pointing to a posting list that contains the (ids) of the documents where the term occurs and the term weight (Wi) in each document. (see Fig. 4).

	Tw1	Tw2	Tw3	Tw4	Tw5
$d1$	0.5	0.25	0.25	0.125	0.0
$d2$	0.15	0.234	0.44	0.67	0.56
$d3$	0.55	0.56	0.23	0.25	0.125
$d4$	0.45	0.26	0.5	0.55	0.0

Figure 4. The Document Vector Model

6.4. Clustering Based Algorithms

Among various clustering based algorithm, we have selected K-means algorithm. Implementation of K-means algorithm was carried out via WEKA tool. The weighted matrix is the input for the K-means algorithm, so an (.arff) file is formed (which is the common file extension for WEKA) which consists of the matrix formed earlier. Two types of attributes are used for this experiment, one is the numerical attribute which represents weights of each term in the matrix and other is the nominal attribute which represents list of documents to be clustered.

6.5. Query processing

Now we can easily retrieve the relevant documents from the weighted matrix for any one-term-query by retrieve its posting list. However, to handle queries contain more than one term, we have to find the union between the posting lists for all terms in the query. As mentioned previously, the main task of this project is to perform information retrieval process on the clusters deemed relevant to a specific query. Therefore, we have to give a rank for each clusters of documents according to the user query, and then ranking documents on relevant clusters against the same query. However, an algorithm (see Equation. *) is implemented to assign a rank to the retrieved documents based on tf-idf weighting along with the vector space model for scoring. The following equations have been used to score documents (d) based on given query, whereas document (d) is belong to relevant cluster G.

$$Score(q,d) = \sum_{t \in q}^{n} tf - idf_{t,d} \qquad (5)$$

The major Information Retrieval system process is based on the previous equation and is fully implemented according to the following algorithm [13].

COSINESCORE(*q*)

1. float Score[*N*] = 0
2. Initialize Length[*N*]
3. **for each** query term *t*
4. **do** calculate $w_{t,q}$ and fetch posting list for *t*
5. **for each** pair(*d*, $tf_{t,d}$)in postings list
6. **do** Scores[*d*] +=$wf_{t,d} \times w_{t,q}$
7. Read the array Length[*d*]
8. **for each** *d*
9. **do** Score[*d*] = Score[*d*]/Length[*d*]
10. **return** Top *K* components of Score[]

7. TESTING AND RESULTS

Beginning of the testing was carried out by taking hundred documents from the following domains:

1. Political News (70132 documents)
3. Economic Article (45247 documents)
4. Sport (64907 documents)

The result was obtained after applying the K-means algorithm once on the weight calculated using term frequency and the other using tf-idf method. For testing purposes documents clusters are ranked against 5 different queries. The result is illustrated in (Table 1). Now, it becomes easier to retrieve the most relevant document according to a user query. The relevant document for each query is retrieved form the highly ranked cluster.

Table 1. Clusters Ranking Against Queries

Category / Query	Political News	Economic Article	Sport Article
Query1	0.94	0.03	0.01
Query2	0.91	0.04	0.05
Query3	0.00	0.10	0.90
Query4	0.06	0.1	0.93
Query5	0.00	0.99	0.01

8. CLUSTER-BASED IR EVALUATION

The classic IR notations of precision and recall are adapted to evaluate the performance of the system result. The Precision (P) is the fraction of retrieved documents that are relevant (see Equation 2), and the Recall (R) is the fraction of relevant documents that are retrieved (see Equation 3).

$$\text{Precision} = \frac{(\#(\text{relevant items retrieved}))}{(\#(\text{retrieved items})} = \text{P}(relevant \mid retrieved) \quad (6)$$

$$\text{Recall} = \frac{(\#(\text{relevant items retrieved}))}{(\#(\text{relevant items})} = \text{P}(retrieved \mid relevant) \quad (7)$$

However, the performance measures (Precision and Recall) may be misleading when examined alone. Therefore, another measure called precision at n (P@n) is considered for evaluating our results, whereas P@5 & P@10 results returned by the system is calculated (see Table 2).

Table 2: The Cluster-Based IR Evaluation

Recall	Precision
0.00	0.831363636363636
010	0.693203168137631
0.20	0.590004733641134
0.30	0.505625749816590
0.40	0.458497377440170
0.50	0.365829471838338
0.60	0.313554213795717
0.70	0.143064322503194
0.80	0.111995554035567
0.90	0.031617647058824
1	0.00
P@5	0.660
p@10	0.655

9. CONCLUSION

A cluster-based information retrieval approach for information retrieval was proposed in this paper, in order to design and develop a system for mining and extracting useful patterns from document collection that are written in English language. In this approach, a pre-processing step was first performed to find frequent and high-utility patterns in the data set. Then a Vector Space Model (VSM) was performed to represent the dataset.

The system was implemented through two main phases. In phase 1, the clustering analysis process is designed and implemented to group documents into several clusters, while in phase 2, an information retrieval process was implemented to rank clusters according to the user queries in order to retrieve the relevant documents from specific clusters deemed relevant to the query. Then the results are evaluated according to evaluation criteria. Recall and Precision (P@5, P@10) of the retrieved results. P@5 was 0.660 and P@10 was 0.655.

REFERENCES

[1] Y. Djenouri, A. Belhadi, D. Djenouri and J. Lin, "Cluster-based information retrieval using pattern mining", *Applied Intelligence*, vol. 51, no. 4, pp. 1888-1903, 2020. Available: https://doi.org/10.1007/s10489-020-01922-x. [Accessed 17 September 2021].

[2] M. Hearst, "Issues, Techniques, and the Relationship to information Access", *Presentation Notes for UM/MS Workshop on Data Mining*, 1997. [Online]. Available: https://people.ischool.berkeley.edu/~hearst/talks/dm-talk/. [Accessed: 17- Sep- 2021].

[3] P. Zorn, M. Emanoil, L. Marshall and M. Panek, "Finding needles in the haystack : Mining meets the Web", *Online*, vol. 23, pp. 16-28, 1999. [Accessed 17 September 2021].

[4] S. Weiss, N. Indurkhya, T. Zhang and F. Damerau, *Text Mining: Predictive Methods for Analyzing Unstructured Information*. New York: Springer-Verlag, 2005.

[5] R. Feldman and J. Sanger, *The text mining Handbook*. London: Cambridge University Press, 2007.

[6] H. Karanikas and B. Theodoulidis, "Knowledge discovery in text and text mining software", Centre for Research in Information Management (CRIM), Department of Computation, UMIST, Manchester, UK, 2002.

[7] P. Losiewicz, D. Oard and R. Kostoff, "Textual data mining to support science and technology management", *Journal of Intelligent Information Systems*, vol. 15, no. 2, pp. 99-119, 2000. Available: 10.1023/a:1008777222412 [Accessed 17 September 2021].

[8] B. Everitt, S. Landau, M. Leese and D. Stahl, *Cluster Analysis*, 5th ed. Chichester: Wiley, 2011.

[9] C. Manning, P. Raghavan and H. Schütze, *Introduction to information retrieval*, 1st ed. Cambridge: Cambridge University Press, 2008.

[10] D. Tkach, *Text Mining Technology: Turning Information into Knowledge*. A White Paper from IBM Software Solutions, 1998.

[11] H. Guldemir and A. Sengur, "Comparison of clustering algorithms for analog modulation classification", *Expert Systems with Applications*, vol. 30, no. 4, pp. 642-649, 2006. Available: https://doi.org/10.1016/j.eswa.2005.07.014. [Accessed 17 September 2021].

[12] F. Raiber and O. Kurland, "Ranking document clusters using markov random fields", *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 333–342, 2013. Available: 10.1145/2484028.2484042 [Accessed 17 September 2021].

[13] K. Naini, I. Altingovde and W. Siberski, "Scalable and Efficient Web Search Result Diversification", *ACM Transactions on the Web (TWEB)*, vol. 10, no. 3, pp. 1-30, 2016. Available: https://dl.acm.org/doi/10.1145/2907948. [Accessed 17 September 2021].

[14] A. Bhopale and A. Tiwari, "Swarm optimized cluster based framework for information retrieval", *Expert Systems with Applications*, vol. 154, no. 2, p. 113441, 2020. Available: https://doi.org/10.1016/j.eswa.2020.113441. [Accessed 17 September 2021].

[15] F. Al Omran and C. Treude, "Choosing an NLP Library for Analyzing Software Documentation: A Systematic Literature Review and a Series of Experiments", *2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)*, pp. 187-197, 2017. Available: 10.1109/msr.2017.42 [Accessed 17 September 2021].

AUTHORS

Orabe Almanaseer (Jordan, 1980), Msc Business Information Technology, Manchester Business School (MBS), University of Manchester, Manchester, United Kingdom, 2008. He is a lecturer with the Faculty of information Technology and Systems