Case Selection Strategy Based on K-Means Clustering

Heba Ayeldeen, Osman Hegazy and Aboul Ella Hassanien

Abstract Knowledge acquisition is considered as an extraordinary issue concerning organizations and decision makers nowadays. Learning from previous failures and successes saves plenty of time in understanding the problems and visualizing data. Case-based Reasoning (CBR) as a process is one of the most used methods to solve the problem of knowledge capture and data understanding. In this paper we proposed an approach for clustering theses documents based on CBR combined with lexical similarity and k-means algorithm for cluster-dependent keyword weighting. The cluster dependent keyword weighting help in partitioning and categorizing the theses documents into more meaningful categories. The proposed approach yield to 91.95 % increase of using CBR in comparison to human assessments.

Keywords Knowledge management · Semantic similarity · Case-based reasoning · K-means

1 Introduction

Saving a lot of time in finding the optimum solution is considered as a win-win situation. Organizations nowadays, focus on reducing time; effort and resources as well in every single cycle process they do [1]. Case-based Reasoning is con-

H. Ayeldeen (✉) · A.E. Hassanien
Scientific Research Group in Egypt (SRGE), Cairo, Egypt
e-mail: heba.ayeldeen@gmail.com

A.E. Hassanien
e-mail: aboitcairo@gmail.com

H. Ayeldeen · O. Hegazy · A.E. Hassanien
Faculty of Computers and Information, Cairo University, Cairo, Egypt
e-mail: osmanhegazy@gmail.com

J.K. Mandal et al. (eds.), *Information Systems Design and Intelligent Applications*,
Advances in Intelligent Systems and Computing 339,
DOI 10.1007/978-81-322-2250-7_39

sidered as a full whole integrated system that aid in decision making and planning as well [2].

Case-based Reasoning (CBR) system is a full computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. Simply, CBR refers to extracting or "mining" knowledge from large amounts of preexisting data [3–6]. Spending a lot of time in search as well as finding the symmetry of different sets/article/documents and even the relations between objects is considered as the domain problem for organizations. For instance in universities, graduate students take a lot of time in searching, sorting and finding related work for research. On the other hand, staff members spend time in categorizing and classifying related articles based on certain research trend(s) (i.e. year/topic/point of research/results) [5, 7].

Case-based reasoning mainly focuses on overcoming the withdrawals within organization. As a concept, CBR deals with learning from previous experiences to solve new problems. The main advantages of CBR systems are [7–9]:

- *Problem definition and understanding*.
  In situations where insufficient or imprecise data and concepts exists, a case-based reasoner can still be developed using only a small set of cases from the problem domain. As an important step in CBR is the problem representation, where cases are briefly explained and indexed with specific attributes/properties.
- *Reducing the knowledge acquisition*.
  After the case or problem is well represented, the waste of time and the need to extract a solution from scratch would be eliminated. The Knowledge acquisition tasks of CBR consist primarily of the collection of relevant existing experiences/cases/problems and their representation and storage within the data warehouse.
- *Avoiding repeating mistakes made in the past*.
  A system like CBR system where failures are recorded as well as successes, and perhaps the reason for those failures, information about what caused failures in the past can be used to predict potential failures in the future.
- *Making predictions*.
  When information is stored whether success or failure, the case-based reasoner would be able to easily predict the success of the solution suggested for a current problem.
- *Avoiding repeating steps*.
  In problem domains that require significant processes to create a solution from scratch, an earlier solution or maybe modification can be easily found by reusing a previous solution for solving other problems.

Finally it can be concluded that Case-based reasoning (CBR) is the process of solving new problems based on the solutions of similar past problems.

The rest of this paper is organized as follows. Section 2 discusses main fundamentals of CBR used in solving/selecting solutions. Section 3 briefly describes the importance of using lexical similarity to find relatedness of texts and documents.

While Sect. 4 shows a case study and steps involved in selecting the case by lexical similarity with k-means clustering. Section 5 shows the interpreting results of the case study. The last section, presents experimental results and conclusion.

2 Case-Based Reasoning: Overview

Case-Based Reasoning is able to utilize the specific knowledge of previously experienced problems situations or cases.

2.1 CBR Life-Cycle Processes

In order to ease time taken to find the optimum solutions or even alternative solutions CBR is the answer. CBR is considered as a group of methodologies combined together to predict and make the process of knowledge acquisition easier.

Below are the main processes involved within the problem solving life cycle in a CBR system [7, 9, 10]:

1. Retrieve
   As previously shown in Fig. 1, when we have a new case or problem for instance to solve, similar previously experienced cases can be easily retrieved from the data warehouse of the CBR system.
2. Reuse
   The reusing of the cases would be an option by copying or even integrating the solutions from the cases retrieved.
3. Revise
   Revising or adapting the solution(s) retrieved in an attempt to solve the new ones.
4. Retain
   Retaining the new solution once it has been validated, hence the process of knowledge acquisition and understanding the problem is valid now (Fig. 2).

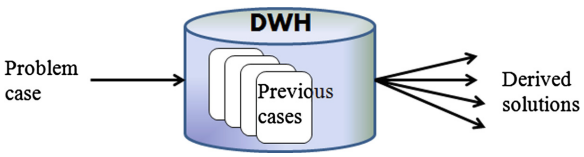

Fig. 1 Case-based reasoning system

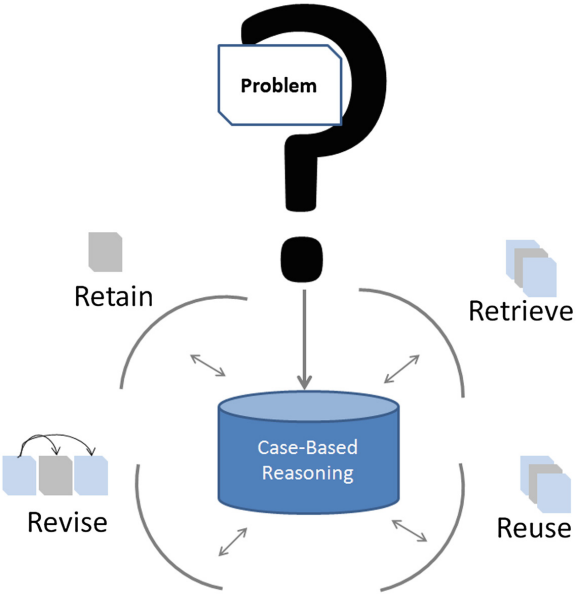

Fig. 2 Case-based reasoning life cycle process

Understanding the problem domain to easily capture the knowledge needed is concerning organizations nowadays. Text representation and data visualizations can be easily accessed by using CBR and make use of every single process within life-cycle of CBR system.

3 Lexical Similarity for Text and Document Mining

It is not easy to find the similarity between objects and get the correlation between them. Semantic and lexical similarity as well as text clustering are important means and methods of mining in texts. Text clustering is an unsupervised classification of documents and objects, which divides a text collection into several subsets called clusters, the text of each cluster has greater similarity than the one in different cluster in mean of categorizations of objects [11].

Making predictions and planning concern all decision makers. Properly classifying and clustering texts based on certain criteria and trend improves the understanding and relatedness of data to easily extract the knowledge based on good classification [11, 12]. The processes of retrieving data; reusing; validating and retaining solutions are considered as the best way to make use of the information rather than rebuilding solutions from scratch. Analyzing the relationships between documents based on concepts and terms is one of the semantic analysis methods [11].

The classification and prediction models are two data analysis techniques that are used to describe classes and predict future data classes. Classification is the process of finding a model/function of technique that describes and differentiates data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown [13, 14].

On the other hand, cluster analysis is a method that organizes a large set of unordered text documents into a small number of meaningful and coherent clusters/categories by which similar records are grouped together [5]. A clustering is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. An organization can take the hierarchy of classes that group similar events. Text document clustering groups similar documents that to form a coherent cluster, while documents that are different have separated apart into different clusters [4].

3.1 K-Mean Clustering Algorithm

K-means clustering is an algorithm to group and categorize objects based on specific features into k number of groups. Clustering is achieved by minimizing the sum of the squares of distances between data and the corresponding centroid of the cluster. The main idea is to assign k-centers for each cluster; however, a better way to select k is to place them as far away from each other as possible and associate each data point in a given data set with the nearest centroid. At this point, k new centroid must be recalculated as bar centers of the clusters resulting from the previous step. Given these new k new centers, a new binding between the same data set points and the nearest new centroid must be performed [4, 6]. The k centers change their location iteratively until no more changes occur. Finally, the k-means algorithm aims to minimize an objective function, in this case a squared error function. The objective function is defined as [15]:

$$J = \sum_{j=1}^{x} \sum_{i=1}^{k} ||x_i^j - C_i||^2 \tag{1}$$

where $||x_i^j - C_i||^2$ is a distance measure between a data point and the cluster center. This is an indicator of the distance between the n data points and their respective cluster centers. Algorithm 1 shows the steps for segmenting images into regions using the k-means clustering algorithm.

4 Case Study

4.1 Data Collection

Data was collected from the digital library of Faculty of Medicine, Cairo University. Faculty of Medicine in Cairo University is classified into 35 departments. The data collected was theses documents including the title of the theses and the abstract with keywords. The theses documents are categorized into Master and doctorate theses. Documents are tracked within the last 10 years separated and categorized based on the departments within the Medicine school. About 4,878 theses data was collected and about 15,808 keyword in the theses data.

Algorithm 1 K-means clustering algorithm

1. Compute the intensity distribution /*the histogram of intensities*/.
2. Initialize the centroids with k random intensities /*the number of clusters to be found*/.
3. Initialize $\mu_i i^k = 1$
4. FOR: Each cluster C_j
5. REPEAT:
6. Cluster the points based on distance of their intensities from the centroid intensities.

$$C^{(i)} := \arg\min_i ||x^{(i)} - \mu^i||^2 \tag{2}$$

7. Compute the new centroid for each of the clusters

$$\mu^i = \frac{\sum_{i=1}^{m} 1c_{(i)=j} x^i}{\sum_{i=1}^{m} 1c_{(i)=j}} \tag{3}$$

   where i iterates over the all intensities, j iterates over all the centroids, and mi is the centroid intensity.
8. UNTIL: cluster labels of the image does not change anymore.
9. ENDFOR.

4.2 Problem Definition

The aim of the work is to find out the departments that can work together easily to increase the research field in the different faculties of Cairo University. To do so, we focused on the theses mining concept for instance for Faculty of Medicine, Cairo University.

4.3 Case Selection by Lexical Similarity and K-Means Algorithm

After refining the keywords within the data sets and removing the stop words, the steps below were used: With the huge amount of data collected, knowledge is then extracted stating that certain departments have potential impact in working together for the purpose of increasing the level of scientific research and helping students in the information retrieval though the theses mining. After calculating the score wait for the tile and the abstract as well, we have (Table 1):

Algorithm 2 Lexical similarity following K-means

- Iteration 1
  After looping on the keywords, use lexical similarity and start getting all departments in which keyword exists in theses data
- Iteration 2
  Based on combinations from the first iteration, calculate the score weight of keywords over all theses document title
- Iteration 3
  Calculate the score weight of keywords over all theses document abstract
- Iteration 4
  Use k-means algorithm (section 3A) to calculate the distance between all documents
- Iteration 5
  Compare and get highest score for better combinations

To affirm that the combinations are the optimum ones and the best choice that departments can work together in the scientific research iteration was done. By using cluster analysis, making clusters for the results of iteration 2 and comparing it with the new clusters from the last iteration based on the abstract on the theses (Table 2).

Table 1 Sample of the data: the best combination of departments based on the theses titles

Department 1	Department 2	Weight[a]
Ophthalmology	Orthopedic surgery	8
General surgery	Medical parasitology	8
Otorhinolaryngology	General surgery	1
General surgery	Orthopedic surgery	1
Obstetrics and gynecology	General surgery	2

[a] The weight represents the repetition number of keyword in the theses title

Table 2 Sample of data: the best combination of departments based on the theses abstracts

Department 1	Department 2	Weight[b]
Ophthalmology	Orthopedic surgery	4
General surgery	Orthopedic surgery	5
Otorhinolaryngology	General surgery	4
General surgery	Medical parasitology	4
Obstetrics and gynecology	General surgery	6

[b] The weight represents the repetition number of keyword in the theses abstracts

4.4 Experimental Results

After applying the k-means algorithm, we find out that there is a correlation between the combinations of the abstract with respect to the title. The accuracy percentage of the data affirming that the combination of departments are the optimum based on the objective function defined earlier 93.21, 6.79 % of the data do not affirm that the combinations in the title and keywords match the combination of the abstract (Fig. 3).

Although the documents are classified based on the departments we focused on applying a cluster analysis and then applying the Euclidean distance to get more accurate combination of departments that can potentially work together (Table 3).

Other than a single word, we applied also the use of a phrase (terms) for better results. For example like "Primary care" and "Diabetes mellitus". So the bag of word can be treated either as a single or a term as a whole. In the case study, we assigned the weights as the number of occurrences of the word in the title of the

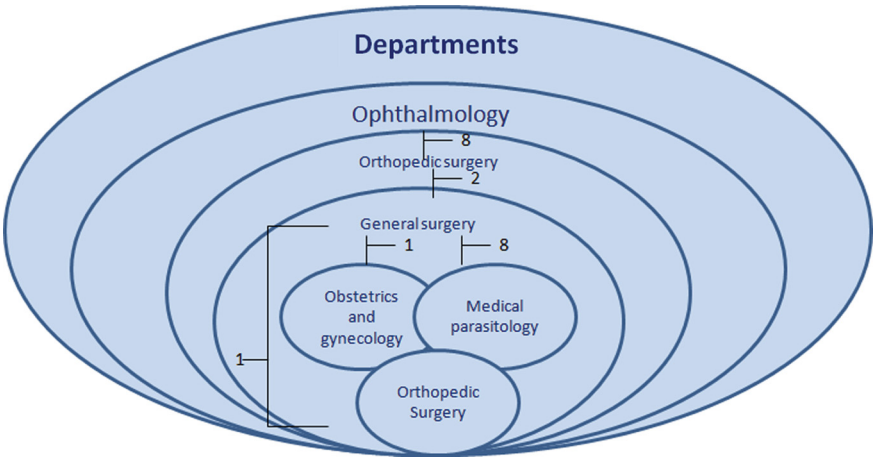

Fig. 3 Using K-means to get the nearest departments based on titles

Table 3 Bags of words representation of the theses documents (titles)

Word	Occurrences
Ultrasound	87
Biopsy	16
Pacemaker	3
⋮	⋮
Catheters	1
Stents	3

document normalized by the document title length to get the word frequency in each document.

In our case let's set the threshold change/stopping condition to 0.001 where there is no big change in the values of each documents in the cluster. We continue the steps by calculating the new cluster centers based on K-means algorithm and updated the values.

After several iterations, we see that T1, T3, T5 and T7 are belonging to same cluster 1 which after knowledge extraction can be categorized to easily work together. On the other hand, T2, T4, T6 and T8 are classified to cluster 2 on basis of high membership values in both clusters. Finally it can be concluded that using data warehouse and CBR techniques and methods are much better for human assessment for biomedical data and that applying lexical similarity and K-means clustering algorithm results in better results with 91.95 % than not using CBR.

5 Conclusion

Recently, the use of Case-Based Reasoning, semantic similarity measures as well as data mining methods leads to the improvement of many applications. Based on the experimental evaluations it is indicated that the proposed approach yields results that correlate more closely with human assessments than other.

In this paper, we showed mathematically how texts can be clustered and classified by using CBR methods and the lexical similarity k-means algorithm.

Other algorithms can be considered as well for future work, like applying the genetic programming; neural networks and comparing the results simultaneously within the cycle of CBR.

References

1. Pressman, R.S.: Software Engineering—A Practitioner's Approach, 5th edn. McGraw-Hill International Edition, New York (chap. 5) (2001)
2. Kolodner, J.L.: An introduction to case-based reasoning. Artif. Intell. Rev. **6**, 3–34 (1992)
3. Singh, S. K.: Database Systems: Concepts, Designs and Applications. Pearson Education, India (2006)

4. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. Elsevier Inc., Amsterdam (2006)
5. Rainer, R.K., Snyder C.A. et al.: Decision Support systems, pp. 333–341 (1992)
6. Ranjan, J.: Managing student data: a data mining-based framework for business schools. Int. J. Inf. Oper. Manage. Edu. **4**(1), 83–98 (2011)
7. Pal, S.K., Shiu, S.C.K.: Foundations of Soft Case-Based Reasoning. Wiley, Hoboken (2004)
8. Park, M.-K., Lee, I., Shon, K.-M.: Using case based reasoning for problem solving in a complex production process. Expert Syst. Appl. **15**, 69–75 (1998)
9. Grupe, F.H., Urwiler, R., Ramarapu, N.K., Owrang, M.: The application of case-based reasoning to the software development process. Inf. Softw. Technol. **40**, 493–499 (1998)
10. Rezvana, M.T., Zeinal Hamadania, A., Shalbafzadehb, A.: Case-based reasoning for classification in the mixed data sets employing the compound distance methods. Eng. Appl. Artif. Intell. **26**(9), 2001–2009 (2013)
11. Al-Mubaid, H., Nguyen, H.A.: Measuring semantic similarity between biomedical concepts within multiple ontologies. IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. **39**, 389–398 (2009)
12. Metzler, D., Dumais, S., Meek, C.: Similarity measures for short segments of text. In: Proceeding ECIR'07. Proceedings of the 29th European Conference on IR Research, pp. 16–27 (2007)
13. Nelson, S.J., Johnston, W.D., Humphreys, B.L.: Relationships in medical subject headings. Relationships in the Organization of Knowledge. K.A. Publishers, New Delhi (2001)
14. Ayeldeen, H., Hassanien, A.E., Fahmy, A.A.: Evaluation of semantic similarity across MeSH ontology: a Cairo University thesis mining case study. In: 12th Mexican International Conference on Artificial Intelligence, pp. 139–144. Mexico City, Mexico (2013)
15. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. **31**, 264–323 (1999)