Multimedia Tools and Applications (2022) 81:32995–33015
https://doi.org/10.1007/s11042-022-12506-x

Design and Development of Density-Based Effective Document Clustering Method Using Ontology

Giridhar Urkude[1] · Manju Pandey[1]

Received: 14 January 2021 / Revised: 6 April 2021 / Accepted: 25 January 2022 /
Published online: 16 April 2022

Abstract
Text document clustering is used to separate a collection of documents into several clusters by allowing the documents in a cluster to be substantially similar. The documents in one cluster are distinct from documents in other clusters. The high-dimensional sparse document term matrix reduces the clustering process efficiency. This study proposes a new way of clustering documents using domain ontology and WordNet ontology. The main objective of this work is to increase cluster output quality. This work aims to investigate and examine the method of selecting feature dimensions to minimize the features of the document name matrix. The sports documents are clustered using conventional K-Means with the dimension reduction features selection process and density-based clustering. A novel approach named ontology-based document clustering is proposed for grouping the text documents. Three critical steps were used in order to develop this technique. The initial step for an ontology-based clustering approach starts with data pre-processing, and the characteristics of the DR method are reduced with the Info-Gain collection. The documents are clustered using two clustering methods: K-Means and Density-Based clustering with DR Feature Selection Process. These methods validate the findings of ontology-based clustering, and this study compared them using the measurement metrics. The second step of this study examines the sports field ontology development and describes the principles and relationship of the terms using sports-related documents. The semantic web rational process is used to test the ontology for validation purposes. An algorithm for the synonym retrieval of the sports domain ontology terms has been proposed and implemented. The retrieved terms from the documents and sport ontology concepts are mapped to the retrieved synonym set words from the WorldNet ontology. The suggested technique is based on synonyms of mapped concepts. The proposed ontology approach employs the reduced feature set in order to clustering the

✉ Giridhar Urkude
giridharurkude@gmail.com

Manju Pandey
manjutiwa@gmail.com

[1] Department of Computer Applications, National Institute of Technology Raipur, Raipur, India

text documents. The results are compared with two traditional approaches on two datasets. The proposed ontology-based clustering approach is found to be effective in clustering the documents with high precision, recall, and accuracy. In addition, this study also compared the different RDF serialization formats for sports ontology.

Keywords Ontology · Conventional k-means · Density-based clustering · Precision

1 Introduction

The clustering of the text documents is used to cluster the relevant documents. For example, the clustering process plays a significant role in search engines. The term-based document retrieval produces a mixture of relevant and irrelevant documents, as the term is syntactically represented. The use of ontology in clustering documents returns more important documents than traditional term methods. The documents are returned depending on the search word when the user searches the document in the search engine. The collected documents are information that is both relevant and irrelevant. The search result depends on the syntactic representation of the document search terms, and thus the sense of the word is not be taken into account. For example, the users searching the net types in search engines, the documents fetched, for instance, include fisherman net types, net documents, and UGC network documents relevant to research. The clustering of documents plays a significant role in order to maximize the hunt for correct documents. The presented study focuses on clustering documents based on the meaning of the term search to overcome the traditional clustering method problem.

This research work primarily focuses on the ability to increase the efficiency of a clustering text document using domain ontology and includes the use and pre-processing of the BBCSports dataset and BBCSports Simulated dataset. The info-gain feature selection Reduction technique eliminates the feature characteristics, and conventional K-Means clustering is compared with DBC (Density-based clustering) DR (Dimension Reduction) technique. The Ontology of the Sports Domain has been established, and its concepts and synonyms are extended to K-Means. WordNet ontology is used in order to find the synonyms of the definitions. The suggested clustering of the document uses the ontology of the sports domain along with WordNet ontology. The performance of the ontology-based clustering algorithm outperforms the other two approaches.

A strong cluster has a high intra-cluster resemblance and a low inter-cluster resemblance. Thus, the main goal of clustering is to reduce the inter-similarity between clusters and maximize their intra-similarity. The consistency of the clustering method depends on both the representation of the text and the measure of similarity. The grouping of documents comprises the following steps,

1. Collection of Documents,
2. Pre-processing of Data
    - 2.1 Tokenization
    - 2.2 Stemming
    - 2.3 Stop-Word Removal,
3. Selection of the right similarity measure,
4. Construction of document term matrix,
5. Selection of the proper clustering method,

6. Implementation of the clustering method in an efficient way,
7. Validation of the clustering quality.

The clustering cycle of documents starts with the selection of the clustering documents. The documents are checked beforehand by the removal of stumps and stop-words. The method then converts the resources into a mathematical model; consequently, records are translated into indexed terms in the mathematical model using a space model vector for each word. This model consists of a matrix of documents where the weight or number of events of documents are allocated to each word. For instance,

D is a set of Documents, and t is a set of terms.

$$D = \{d_1, d_2, d_3, d_4, \ldots\ldots\ldots d_n\}$$

$$t = \{t_1, t_2, t_2, t_4, \ldots\ldots\ldots\ldots t_m\}$$

Each term is represented by a vector

$$D_i = \{w_{1i}, w_{2i}, w_{3i}, \ldots\ldots\ldots\ldots\ldots\ldots w_{mi}\}$$

Where $i \epsilon \{1,2,3,\ldots..m\}$,

Where w_{mi} is the weight of a term m in document D_i.

Here, w_{ij} is the product of 'frequency' and the 'reverse document frequency'. The term frequency is used to calculate the number of occurrences in a database of a given word. Many documents vary in length, so a word in lengthy documents can occur more frequently than in a shorter one. Therefore, the term frequency is normalized by dividing the entire number of conditions in the document. The reverse document frequency (rdf), on the other hand, expresses how important a term is. All terms are equally relevant when measuring term frequency (tf).

$$tf = \frac{\text{Number of occurrences of the particular term in the document}}{\text{Total number of terms in the entire document}} \tag{1.1}$$

$$rdf\ (t) = \frac{\text{Log e (total number of documents)}}{\text{Number of documents with term t in it}} \tag{1.2}$$

$w_{i,j} = tf_{i,j} * idf_{i,j}$

tf-rdf representation of document d is

$$D_{tf-rdf} = [tf_1 \log (n/df_1), tf_2 \log(n/df_2), \ldots\ldots\ldots\ldots, tf_d \log(n/df_d)] \tag{1.3}$$

To consider documents of different lengths, each document vector value is normalized to a unit vector $||D_{tf\text{-}rdf}|| = 1$.

1.1 Challenges in Document Clustering

Document clustering has been an important research subject for decades due to its numerous applications; for example, clustering of records, optimization of the search, data extraction, and information recovery are needed in many other sectors. However, certain challenges in

clustering documents must be targeted to improve clustering performance. Some of the problems with clustering documents are,

- DTM (Document term matrix) is a high-dimensional matrix, which consists of the documents obtained and all the special words used to collect documents. This DTM is a very sparse matrix in large dimensions. When the dataset is limited, the clustering algorithms handle this DTM. Nevertheless, when the dataset becomes enormous, it is difficult for the grouping algorithm to manage DTM, and thus it makes the clustering process more complex and affects efficiency.
- DR (Dimension Reduction) method selection – the appropriate selection of DR method may be used to reduce the complexity of the DTM. Some methods of feature selection are incredibly costly, and therefore, selecting the most relevant features is very difficult for the function selection process.
- Selection of cluster numbers – the precision of clustering algorithms depends on the number of clusters chosen if the classmark is not already defined. It is complicated to determine the cluster number prior to the clustering process.

The rest of the paper is organized as follows, section 2 presents a literature review of text document clustering and provides an extensive survey of WordNet ontology and ontology approach, section 3 describes the objective of the research work and problem statement, section 4 explains the methodology for the development of sports domain ontology. Section 5 presents the result discussion. Section 6 describes the strength of this research work and also provides suggestions for future work.

2 Related Work

A literature review has been undertaken in the ontology domain, document clustering, WordNet ontology, and ontology-based document clustering. The significance of ontology in the seminal search and clustering of documents is all defined by these principles. Clustering algorithms are researched to understand how search optimization is useful.

The study [7] has summarized many ways to reduce the dimensionality of high microarray data. Two approaches, namely, feature extraction and feature selection procedures, can be used to minimize dimensionality. This paper studied different types of feature selection methods and various types of microarray data extraction techniques.

Balabantaray et al. [2] compared the K-Means Cluster and K-Medoids Clustering algorithms. K-Means clustering was carried out using the WEKA method and Euclidean and Manhattan distance. K-Medoids was implemented through Java programming on the text word matrix. The experimental evidence demonstrated that K-Means give better outcomes than K-medoids.

Filters, wrapper-based, embedded, and modern hybrid feature selection approaches have been investigated in the study [10]. For several technology fields, such as image processing, text mining, bioinformatics, and industrial applications, the study summarized the results of specific feature selection methods. This research concludes that the hybrid selection methods for all domains, including text mining, industrial applications, image processing, and bioinformatics, would yield better results.

The study in [4] introduced the general concept of semantic data mining and investigated why ontology has the potential to help semantic data mining and how formal semantics in ontologies can be incorporated into the data mining process and provided the state of the art of ontology-based approaches for the use of ontology from data pre-processing to mining results. The paper summarizes many semantic data mining activities, including Association rule Mining, ontology-based clustering, and ontology-based knowledge extraction.

The Corpus-based enrichment approach is used for the clustering of short texts described in the study [23]. It provides a collection of conjugate meanings to explain the subject matter and the word form. It proposes a virtual iterative method for short text. The extension of short text data is achieved by introducing new terms that do not appear in a short text document with a virtual term frequency. The posterior probability of new terms, provided all the terms in that document, is used to measure the virtual frequency.

The length feature weight (LFW) scheme allows the term weighting technique to cluster text documents. It is based on unique cluster terms to boost the algorithms presented in [1]. The proposed method assigns a favourable term weight based on the information obtained from the collection of documents. It defines words/terms that are unique to each cluster and enhances their weights based on the estimated factors at the document level. The approach is applied to the benchmarked datasets.

The study in [5] proposed an approach to automatically merge the recovered documents into the list of relevant documents. Various methods have been used for clustering documents. In this method, a novel technique is used for document clustering called ABK-Means clustering, which clusters documents based on the similarity between the terms of the text. The NLP (Natural Language Processing) speech tag element extracts the information in the document and extracts the information using the chunk category. This document clustering has used descriptors and descriptor extraction to improve document clustering performance.

The study [6] proposed a new clustering algorithm, which used the weight attribute for clustering documents. The probabilistic distribution of the attribute in the document called "benefit ratio" is used for attribute weighting. The clustering result was strengthened with the aid of the weighting attribute in work.

The work in [14] suggested a new method that integrated domain awareness into the data mining process. This method has been established by the school ontology using protégé. This method clusters documents using the K-Means cluster algorithm. Clusters are reconstructed by separating and combining operations using the ontology of the school. The main objective of this approach was to boost clustering results by filling the gap between knowledge and semantics.

The study in [18] reviewed 23 papers and established four major areas of semantic clusterings such as latent semantic indexing, graph-based, ontological-based, and lexical-based. This work concluded that the semantic approach is better than the conventional keyword approach in terms of cluster accuracy and consistency. WordNet ontology has been used as context information for the work, and WordNet ontology lexical chains are incorporated to classify features using the synonym relationship. This work also proposed the combination of lexical chain definition and ontology to solve many of the problems of current clustering algorithms.

The work described in [21] implemented lexical strings to extract semantically related words from the text. This work aims to incorporate lexical chains into document clustering to reduce the high dimensional space of the applications, and it was achieved by the extraction process of the disconnected core elements. Disabling polysemous and interchangeable nouns

increases the consistency of clusters. The study used the wu & palmer test and called the glosses of words for the word sense uncertainty (WSD). WSD is a method that substitutes the most suitable language determined by the meaning of a document for the original words in a document.

The experiment in [3] illustrated the defined and evolving meaning systems known as artistic collectives. Collocations were used as a semantic framework to construct a social context in the study of socio-semantic networks. The study limitation is the use of statistical models focused on local micro-patterns of social relations and conceptual connections.

The model shown in work [20] retains the meaning of the concept by considering the sequence of terms. In the phrase text matrix, phrase weight was assigned to the number of phrases in each cell. The work in [11] enhanced mean converge score, and high precision shows the efficiency of this technique. Abstract summarization of multi-documents using semantic similarity measures and attribute weights. The limitation found in this work is the use of a semantic graph for abstract multi-document descriptions.

A multi-class classification model separates artificial language constructs such as tables, formulas, and pseudo-codes from natural languages to efficiently cluster documents [8]. It detects unnatural language components divided into four categories: table, code, mathematical formula, and miscellaneous. Features accessible from the plain text are explored to create a statistical model based on a newly annotated corpus containing a collection of documents, such as PPT, PDF, HTML, etc.

Documents having little text and less commonality are clustered in the study [9]. Most of the time, text similarity measure results into zero values, leading to the sparse matrix, and thus, term correlation was used to express semantic relativity known as statistical semantics. A matrix was used for document clustering based on selected terms showing a correlation between them in the presented work.

A new methodology for ontology development in the field of the lesson plan was proposed in the study [17]. The approach consists of different phases, which include study, development, implementation, assessment, and maintenance of requirements. Moreover, this approach gave framework developers a systematic guideline for establishing ontology in other domains.

Semantic clustering has been used in [19] based on the semantics of the documents in order to group the relevant documents. A search engine was used to obtain information; subsequently, these results were pre-processed, and feature extraction was performed. The characteristics have been developed using ontological and semantic principles. Using the Floyd-Warshall algorithm [15], a dissension matrix was developed for the documents. The hierarchical clustering algorithm was used to cluster the functional vectors in the similarity matrix. Current approaches to web search results were obtained in terms of high precision performance. Hierarchical agglomerative clusters might be used for massive data sets as it is computationally costly. The use of the seminal clusters to evaluate related text documents was also suggested in the study [22]. The study also includes the extraction and pre-processing of the document corpus. The frequency-inverse document frequency algorithm is employed to classify frequent terms and create a document matrix. A field ontology is developed from a text corpus to give the related words a vocabulary. A fuzzy equivalence relationship was used to evaluate the degree of membership in the text corpus. A single value decomposition is used to transform the document matrix into a concept space. The K-Means algorithm was used in work to cluster the concept space. The use of pre-processing in the domain ontology enhances cluster outcomes. The drawback of the method is that the method's success depends entirely on the accuracy and validity of the ontology used. In work [16], customer reviews were graded, and the text corpus

was stripped from the pages and pre-processed by crawling customer reviews. In the text corpus, a definition mapping was created from domain ontology. Euclidean distance metrics were used for the measurement of the similarity of the phrases in the vector model word pocket. The modified K-Means algorithm was used to group the bag of terms. Experimental findings have demonstrated that the precise use of ontology in the pre-processing stage of the produced clusters increases.

The presented new approach concentrates on the problems of conventional clustering methods. Firstly, sports ontology is established using sports field concepts and relations. The new synonym recuperation algorithm is introduced in the study to extract the definition of sport ontology with WordNet ontology. Sport ontology definitions are related to the terms derived from the datasets. The algorithm retrieves the synonyms and returns the most important synonym along with the definition of ontology as a consequence. The proposed ontology-based clustering approach uses this reduced feature set.

3 Problem Statement

The features of a high dimensional thin document term matrix that affects the efficiency of clustering can be difficult to reduce with traditional clustering methods. Feature selection techniques are used to restrict functionality; however, their tests are unreliable. Many predefined criteria are used, and the study observed that most practical selection methods are very expensive and often highly computational. Therefore, the ontological clustering approach was developed to reduce the feature set and improve clustering quality efficiency.

The main aim of this work is to enhance the consistency of ontology clusters. The density-based clustering on domain ontology is used to cluster documents semantically to provide better retrieval than standard K-Means. The goals are as follows:

- Analyze the text document clustering approaches and their challenges.
- To reduce the features set for BBCSports dataset and BBCSports Simulated Dataset, analyze the feature selection methods of dimension reduction technique.
- Design and develop sports domain ontology with important games such as Athletics, Cricket, Football, Rugby, and Tennis.
- Study and evaluate cluster methods based on ontology.
- Cluster approaches based on ontology are tested and evaluated.
- Identify the most relevant terms of WordNet ontology synsets.
- Develop a framework for sports ontology text document clustering.
- Validate the ontology-based methodology for BBCSports and BB Sports Simulated dataset with other two conventional approaches using evaluation metrics.

This new method concentrates on the problems of conventional clustering methods. Firstly, sports ontology is established using sports field concepts and relations. For the extraction of the definition of sport ontology with WordNet ontology, the new synonym recuperation algorithm is introduced. Sport ontology definitions are related to the terms derived from the data sets. The algorithm retrieves the synonyms and returns the most important synonym along with the definition of ontology as a consequence. The proposed ontology-based clustering approach uses this reduced feature set.

4 Methodology

The primary goal of text document clustering is to organize documents with high similarity inside the cluster and low similarity to other clusters. The resemblance between intra-clusters is solely inside the clusters, and the texts are inextricably related to one another. Inter-cluster similarity, on the other hand, clusters terms are related to other clusters. In order to cluster the relevant documents using an ontology approach, one must follow the knowledge management process, which includes obtaining, organizing, and reusing semantic knowledge to develop a system with a common understanding [12].

The developed ontology-based clustering considers the terminology and synonyms for the clustering process. This study employed WordNet ontology and the created sport domain ontology to improve clustering efficiency. The WordNet Ontology is utilized to retrieve synonyms of the extracted terms using the synonym recovery algorithm. Due to the semantic terms used for the clustering process, the size of the document matrix has been decreased significantly. As a result, the text document using ontology clustering groups the related documents more effectively, which clusters related sports documents based on terms semantics. The overall methodology is depicted in Fig. 1, which shows all crucial components.

In order to implement this methodology, BBCSports and BBCSports Simulated datasets are employed. This study offers an overview of the phases of this technique and the methodological principles of the ontological clustering approach for sports.

The following five steps are part of the suggested methodology;

1. Term Extraction and Preprocessing
2. Non-Semantic methods of clustering (K-Means and DBC with DR),
3. Creation of sports field ontology,
4. Clustering ontology.
5. Assessment of the clusters.

4.1 Preprocessing

The pre-processing technique collects terms from the documents that significantly impact the outcome of any grouping process. After applying the preprocessing method, the mathematical model produces the relevant terms available in a text document. This model is the very first process on the input set of documents and, at the same time, relatively straightforward to study. The parsing technique includes tokenization, stop word deletion, and stemming processes.

The tokenization process converts the content of a document into a collection of words. Tokenization is primarily concerned with learning the words in a sentence. The collection of words is used for further investigation and also used to describe specific terms in documents. The stemming process reduces the derived words to their root or base word. It is used to minimize the number of distinct phrases by stemming them to their roots, such as “jumps” into “jump.”

This experiment used the snowball stemming method to convert words into their base word. Many words, such as “and”, “a”, “about” appear frequently but do not have any significance because the goal of these words is to assimilate words into a phrase. The interpretation of every content (term) of the document is complicated due to the enormous volume of texts. These are considered as noise and must be reduced in order to improve the efficiency of the machine.

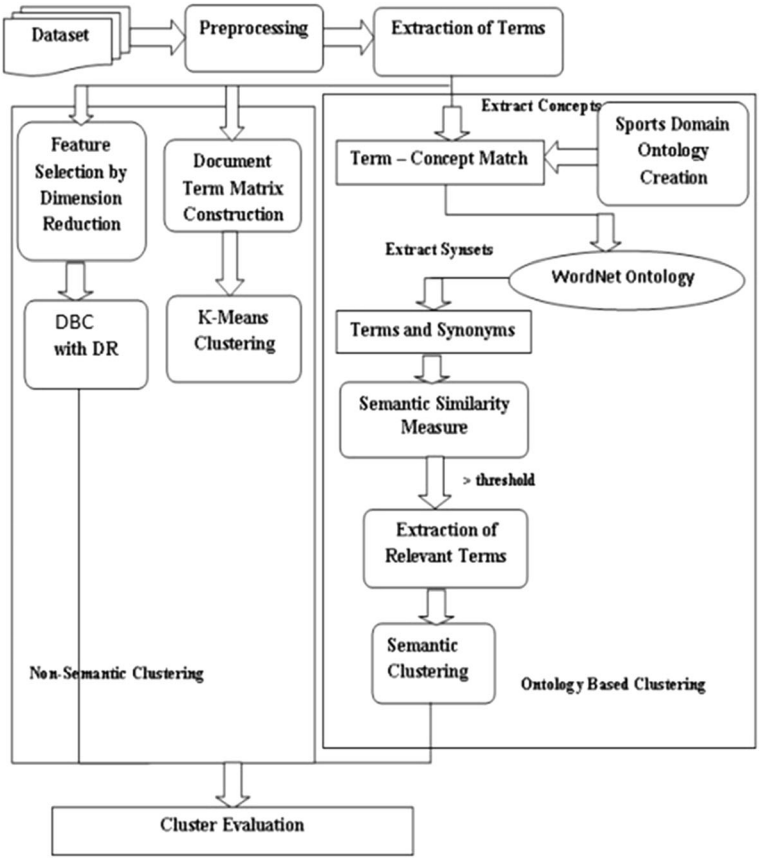

Fig. 1 Methodology of Proposed System

The advantages of pre-processing include,

1. Reducing the difficulty of text document clustering by increasing the scale of the word matrix.
2. Improve the efficiency of the clustering process.

Pre-processing is used to lessen the complexity of the text document clustering process, which is accomplished by dimension reduction of the document matrix in order to match the relevant terms, which boosts the efficiency of the clustering process. It is also used to determine which word is more relevant to which group of documents. The texts are prepared through a series of steps known as pre-processing techniques. The pre-processing steps are described as,

- Tokenization
- Stop word removal
- Stemming

Tokenization The majority of dataset documents contain numbers, dates, and common terms that are not related to any document and should be omitted from the word list. The pre-

processing starts with the tokenization process. A text stream is broken into different tokens, which include a collection of characters, images, phrases, or other terms of meaning. The tokens list provides an interface for additional pre-processing steps. The token list consists of numbers, dates, times, and relevant words and abbreviations after tokenization. The abbreviations must be converted into their complete form to get the insightful clusters.

Stop-word Removal Stop-words are words like "also", "or", "the", "which", etc that appear frequently but have no significance in the clustering process because their purpose is to join terms in a sentence. Understanding the substance of the documents is challenging due to the widespread circulation of stop-word within the document. The stop-words are not needed in the clustering process, and therefore, they are eliminated to improve clustering performance quality.

Stemming The derived words are reduced to their root or base word in the stemming process; for example, "jumps", "leaping", "jumped", "jumping" must be reduced to a famous "jump" core. There is a probability of two faults in stemming, one is over, and another is under stemming. Two words with distinct stemming are derived from the same stem and are known as false positives. Under stemming, two words have the same token, yet they are not in the process. The stemming algorithm is used to eliminate the normal morphological and inflexional terminations of words.

4.2 Non-Semantic Clustering

The non-semantic clustering approach mainly relies on K-Means and DBC clustering with the DR feature selection methods. In these methods, the document word matrix is constructed using the Euclidean distance as a metric of similarity, but the sense of the terms has not been concerned.

K-Means Clustering The K-Means classification technique is a robust unattended research algorithm designed by MacQueen [13] to handle the well-known clustering problem. The K-Means algorithm aims to partition a set of documents into 'K' clusters, where 'K' is a predefined constant dependent on their features. The centroids represent the average mean value of each cluster. As the key concept, 'K' center lines are provided for each cluster. The cluster's core is so well established that it is strongly related to similarity functions, where the Euclidean distance is determined from all documents in the cluster. In this study, the BBCSports datasets and BBCSport Simulated data sets are clustered with K-Means.

Different BBCSports datasets are composed of five games, and Euclidean distance is used for the syntactical similarity of the word.

DBC with Feature Selection DR Density connectivity functions govern density-based clustering techniques. Density-based approaches try to frame the data profile by splitting the space into dense areas. High-density areas are separated from the low-density regions, defined by the threshold density. This approach aims to expand the defined cluster to a certain density. These techniques are employed if irregular clusters of shape occur, and the data are noisy. The popular density-based approaches to clustering static data are DBSCAN, OPTICS, and DENCLUE.

4.3 Semantic clustering

Semantic features are added in order to incorporate the semantic dimension in the clustering process. The semantic value should be translated into concepts (i.e. corresponds to definition marks in the ontology of reference instead of basic modalities); therefore, the comparisons between values that represent classes can be made using a feature of semantic resemblance. Comparing objects in a semantic instance, the concept of distance/similarity measurements between the values of a semantic pair is a critical approach to the cluster. The similarity is quantified by how concepts are alike in some information (e.g. ontology or a corpus) based on semantical proof. The knowledge used to approximate the similarity between words enables these functions to be categorized in various classes. In some approaches, taxonomies and ontologies are generally regarded as graphical, where semantic relations (properties) are used to connect concepts. The similarity typically depends on the minimum number of connections between concepts (minimum path). Similarity may also rely on other features like the breadth of the taxonomy concepts. This taxonomy measure has the key advantage of depending only on ontology to determine the similarity. However, the degree of completeness, homogeneity and coverage of ontology affect them.

4.4 Evaluation Metrics

The study described in this section evaluates clustering consistency using evaluation steps such as recall, accuracy, and precision. The uncertainty matrix is used to calculate the metrics of accuracy, recall, and precision. The formulas for accuracy, recall, and precision are observed using TP (True Positive), FP (False Positive), FN (False Negative), and TN (True Negative) instances of the datasets with the different clustering methods. These formulas are defined as:

$$\text{Precision} = \frac{\text{Number of relevant documents Retrieved (TP)}}{\text{Number of documents Retrieved (TP + FP)}} \tag{4.1}$$

$$\text{Recall} = \frac{\text{Number of relevant documents Retrieved (TP)}}{\text{Number of relevant documents (TP + FN)}} \tag{4.2}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})} \tag{4.3}$$

4.5 Sports Ontology

The generated sport domain ontology consists of five sports classes: Athletics, Cricket, Football, Rugby, and Tennis. With the aid of domain expert knowledge (BBCSports documents), the sports domain ontology is established. This sports ontology includes definitions, sub-concepts and their relations, and constraints. The classes for sports ontology with its subclasses and relations are described in Fig. 2. The circle represents the ontology classes, and the arrow represents a relation between the two classes.

The experiment used the protégé tool in order to develop the sports ontology (shown in Fig. 3). An ontology is the description of a domain with a finite number of classes and relations. The classes and relations in the ontology can be taken as per the requirement of the application, and it should be scalable. For the document clustering process, the study used a limited number of classes and relations. It can be further extended for all the other sports classes.

4.6 Synonyms Retrieval and Similarity Measure Calculation

The presented algorithm accepts the word as a source and returns the count and terms along with its most important synonyms, the extracted words, and the sports domain ontology concepts (flow graph is shown in Fig. 4).

The threshold value is set as '0.8' for the similarity index since it offered us the best performance value for precision and recall. The threshold value equal to 'one' gives a similar term. The downloaded WordNet database is loaded, and an instance of that object is created to use WordNet ontology. Sports field ontology concepts are used to validate the extracted

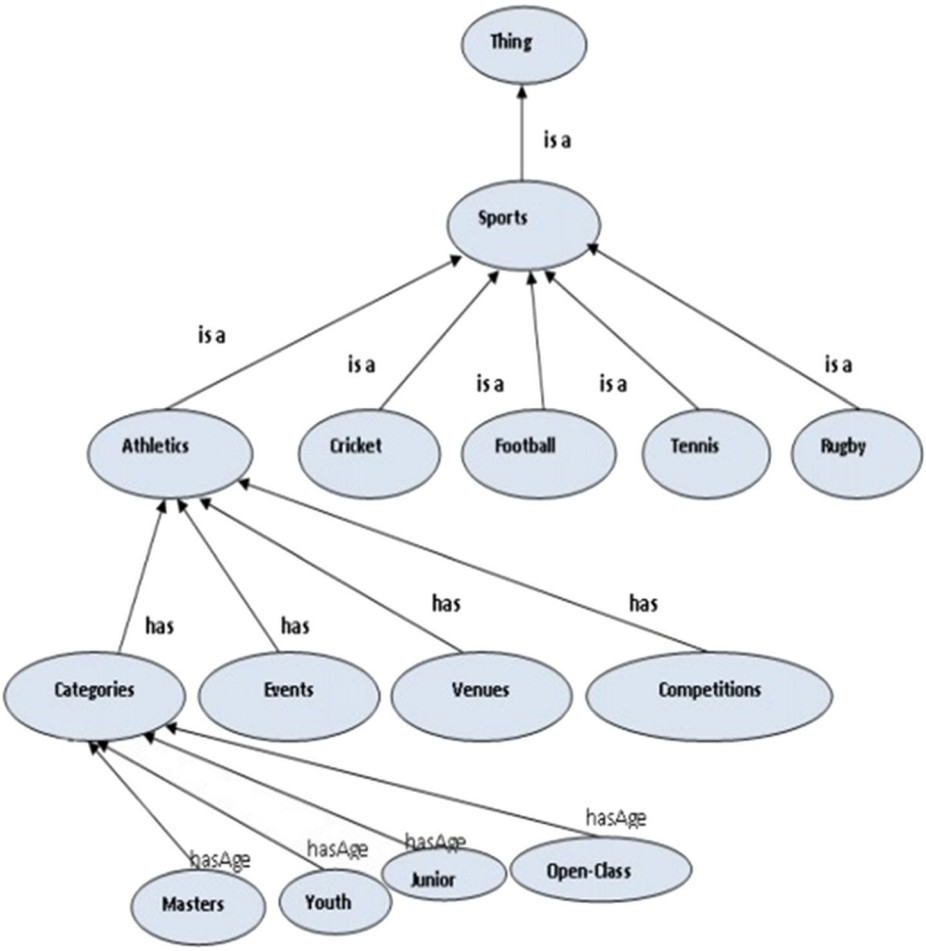

Fig. 2 Main Classes of Sports-Ontology

dataset terms. If the terms are the same, it is chosen for further processing. The pseudo-code for the synonym recovery algorithm is shown in Algorithm 1.

(To retrieve the synonym for a given word and check for relevancy. Relevant synonym is added with the terms. The terms and relevant terms are applied to k means clustering)

Input: Extracted terms, Ontology Concepts and Subconcepts

Output: Synonyms

1. Set count=0, threshold=0.8
2. Load WordNetDatabase and create object.
3. For all terms and concepts
4. Check if term = concept
5. Increment count
6. Get Synsets of term
7. If length of Synsets >0
    - 7.1 Get WordForms
    - 7.2 For each pair of WordForms
    - 7.3 Calculate similarity measure S
    - 7.4 If S > threshold
    - 7.5 Add synonym with terms
    - 7.6 Else ignore
    - 7.7 Repeat step 7.2 until WordForms exists.
8. Add terms.
9. Repeat step 3 until terms and concepts exists.
10. Count is returned.
11. Terms and synonyms are returned.

The Semantic Similarity Measure is defined over a set of terms and calculated on the basis of the relationship between the meanings of the words. This semantic similarity measure differs from a similarity measure based on the syntactic representation of the terms.

The WordNet database synonym contains the WordForms definition and array, and the length of that particular synonym array is determined and a similarity measure taken for each pair of WordForms. Similarity measure S returns with terms and increments the count value

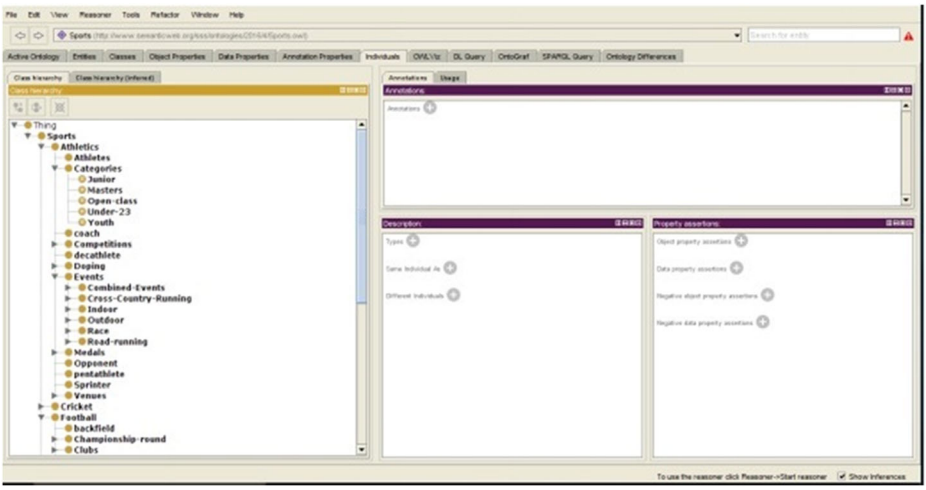

Fig. 3 Class Hierarchy

when it reaches the predetermined threshold value. Terms with similarity measure values less than the cutoff are ignored. The preceding procedure is performed for all WordForms pairings.

The synonym recovery procedure is depicted in Fig. 5, which explains synonyms with a similarity value larger than the threshold value are returned by the synonym recovery algorithm. The synsets consist of WordForms (synonym terms), and the similarity calculation is determined and tested for each pair of terms by the threshold value. All the similar semantic terms were used for the clustering process. The K-Means clustering approach performed the clustering for all the similar semantic terms.

The ontology-based document clustering uses the semantics of the terms to cluster the relevant documents. The semantic has been obtained from the synonym set, which was retrieved from the WordNet ontology. The WordNet ontology contains similar types of words for a term (called synonym set), which act as a dictionary. Every WordForm in the synonym set was compared to sports ontology concepts and pre-processed extracted terms. The ontology-based clustering approach uses similar semantic terms for K-Means clustering to classify the documents into the same semantic cluster, whereas the simple K-Means clustering approach uses only syntactical similarity.

5 Results and Discussion

The BBCSports Dataset has been used in this experiment, which is open source and available for download from the BBC website. The BBCSports dataset contains 737 text documents divided into five categories: Athletics, Cricket, Football, Rugby, and Tennis. The different clustering methods, along with the newly developed ontology-based clustering method, have been implemented on the BBCSports dataset. The result evaluation and discussion of these methods are discussed in this section.

5.1 Comparison of Different Clustering Methods

When the K-Means method was applied to the dataset, it was found that 123 documents out of 737 were grouped in the incorrect cluster, and 614 documents were grouped in the correct

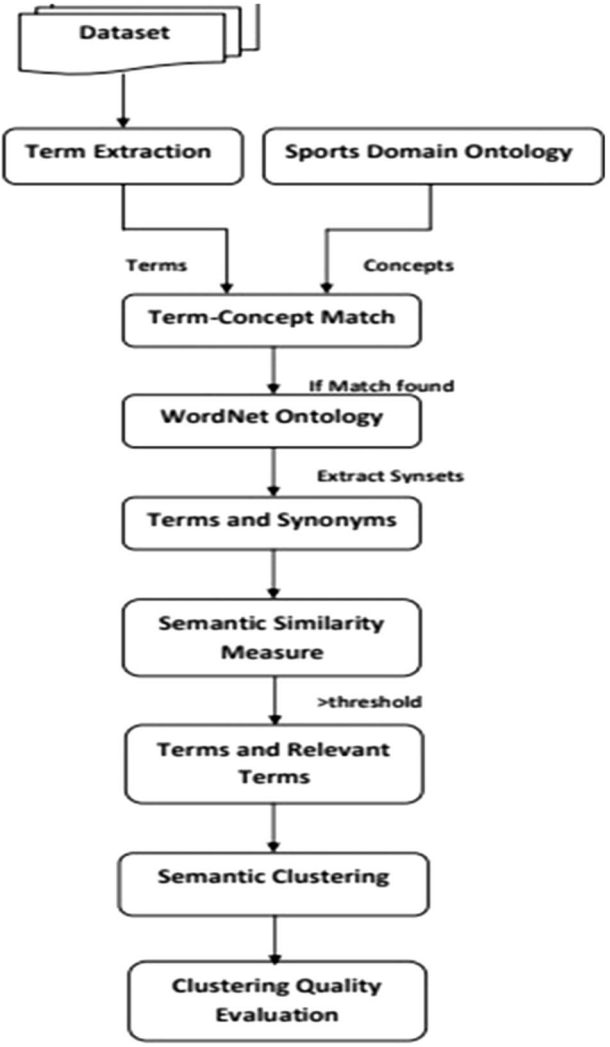

Fig. 4 Methodology of Ontology-based Clustering

cluster. The number of incorrect grouped documents was found to be 16.7%, the average recall value was 85%, with an average accuracy was 83.3%, and the precision value was 83% for the K-Means method.

In the Density-Based clustering method, the dimension reduction technique was implemented to reduce the dimensions of the document matrix. As per this method, 69 out of 737 incorrect BBCSports documents were clustered. The percentage of erroneously categorized documents was found to be 9.3%, which performs better than the K-Means clustering method. The average recall was found to be 89.2%, which is 4.2% higher than the K-Means method, the average accuracy was found to be 90.7%, which is 7.4% greater than the accuracy of the K-Means method, and the precision was found to be 93.4% that is 10.4% greater than the K-Means approach. According to these results, the DBC approach with the DR outperforms the K-Means method for clustering the BBCSports documents.

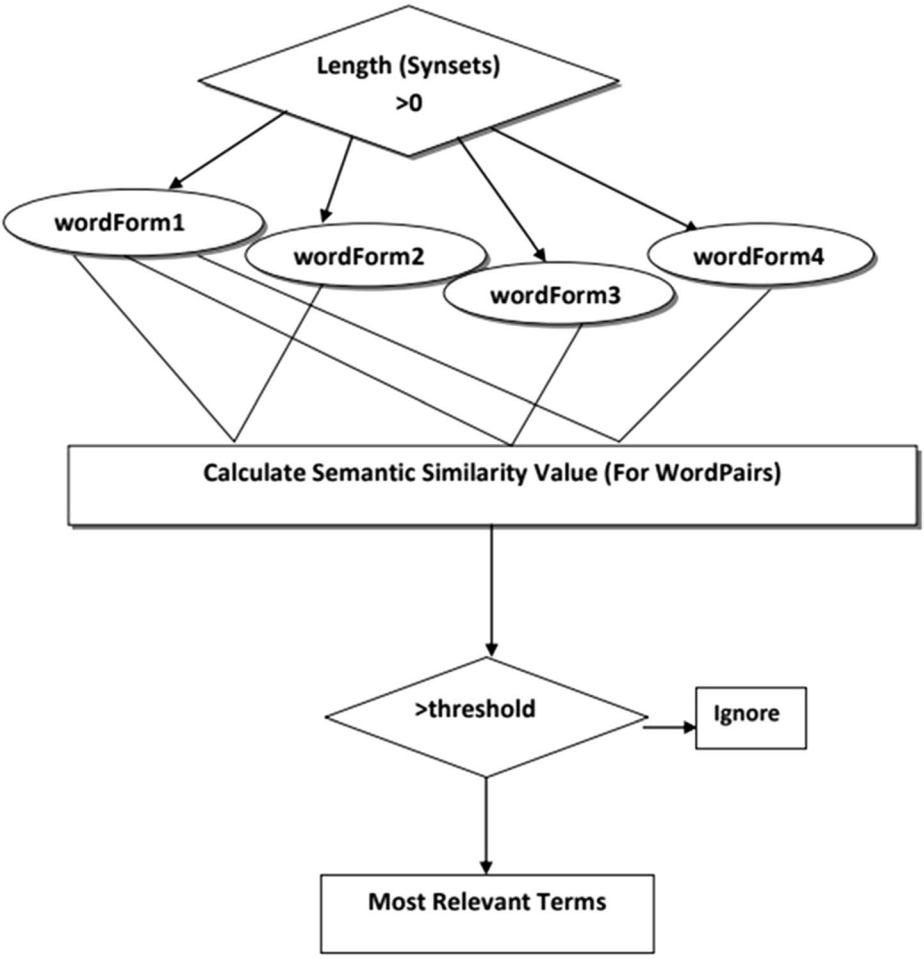

Fig. 5 Procedure for synonym recovery using WordForms

In an ontology-based approach, the K-Means Clustering Algorithm is subject to the words mapped with the principles of sports field ontology and synonyms. In this method, 43 out of 737 documents were wrongly clustered and 5.8 percent inaccurate.

Finally, the newly developed ontology-based approach is implemented with the K-Means clustering scheme. In this method, the sports ontology classes and WordNet synonyms are mapped with the extracted terms. When this approach was applied to the dataset, it was found that 43 out of 737 incorrect documents were clustered, which was 5.8% of the total document. With the ontology-based semantic similarity method, the recall was found to be 93.7%, which

Table 1 Different clustering methods comparison

Metrics	K-Means Clustering (in %)	DBC with DR Clustring (in %)	Ontology-based Clustring (in %)
Precision	83.0	93.4	**96.6**
Recall	85.0	89.2	**93.7**
Accuracy	83.3	90.7	**94.2**

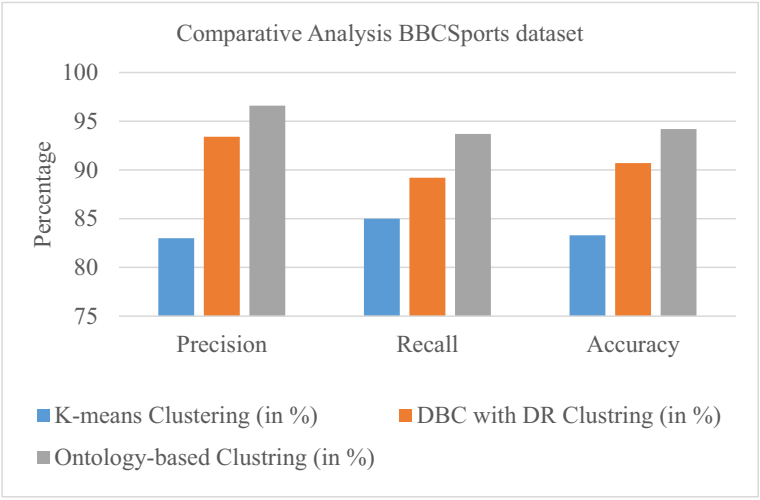

Fig. 6 Comparison of different approaches for BBCSports datasets

is 4.5% higher than the DBC approach, the accuracy was 94.2%, which is 3.5% higher than the DBC approach, and the precision was 96.6%, which is 3.2% higher than the DBC approach. The K-Means, density-based clustering, and ontology-based technique were compared in terms of precision, recall, and precision, shown in Table 1. The DBC approach performs better as compared to simple K-Mean clustering. At the same time, the ontology clustering approach performs better in every parameter (accuracy, precision, and recall) compared to the K-Means and DBC methods. As a result, it is concluded that the ontology-based clustering process (using the semantic of the words) surpasses the other two approaches.

The above-mentioned methods are experimented and tested on the BBCSports dataset. A significantly improved performance has been obtained in the ontology-based document clustering technique on the basis of recall, precision, and accuracy compared to K-Means and DBC with DR approaches. The bar chart in Fig. 6 clearly shows that density-based document clustering using the ontology approach outperforms the other two methods in all the parameters.

The Simulated BBCSports dataset includes 1080 text papers and five natural classes: Athletics, Cricket, Football, Rugby, and Tennis.

In the K-Means clustering, the accuracy was 76%, i.e., there was a total of 819 correctly clustered documents, and 261 records were wrongly clustered. Therefore, the wrong cluster documents were 24% in this clustering process. In the DBC clustering algorithm with the feature selection DR method, the total accuracy was 89%. Overall, 960 documents were accurately clustered, and 120 out of 1080 documents were wrongly grouped. In the ontology

Table 2 Comparison of different approaches for BBCSports Simulated dataset

Metrics	K-Means Clustering (in %)	DBC with DR (in %)	Ontology-based Clustring (in %)
Precision	75	87.6	**93**
Recall	87.3	92.4	**93**
Accuracy	76	89	**92**

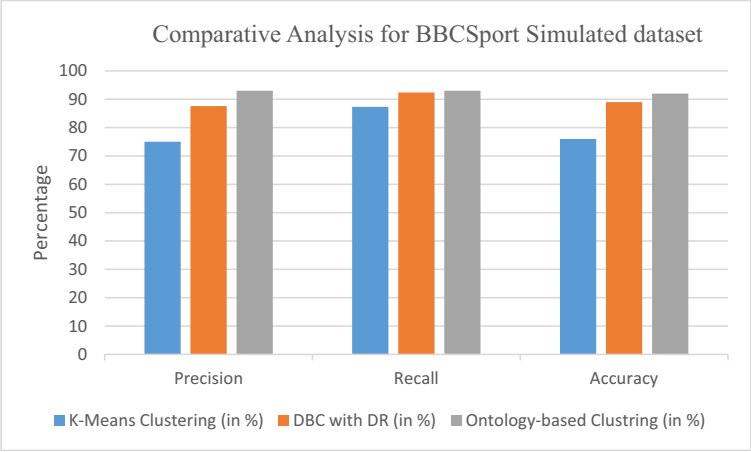

Fig. 7 BBCSports Simulated three clustering approaches

method for BBCSports Simulated data set, 87 out of 1080 documents were incorrectly grouped, which was 8%, and the total accuracy was 92% using this method.

These results prove the importance of semantics in order to cluster the text documents; here, the involvement of ontology (semantic graphs) improves the overall clustering process. The presented small feature set was used to build a document term matrix and for clustering. The results show that the InfoGain feature selection method improves clustering efficiency even though the dataset is huge.

The clustering results using the different methods, including the developed ontology method over the BBCSport Simulated dataset, are shown in Table 2. It was found that the precision of ontology-based clustering was 18% higher than the K-Means clustering procedure and 5.4 percent higher than the DBC with DR method. The recall using the ontology-based method was 5.7% higher than the K-Means clustering procedure and 0.6% higher than the DBC with DR method. Similarly, the ontology-based method's accuracy was obtained 16%

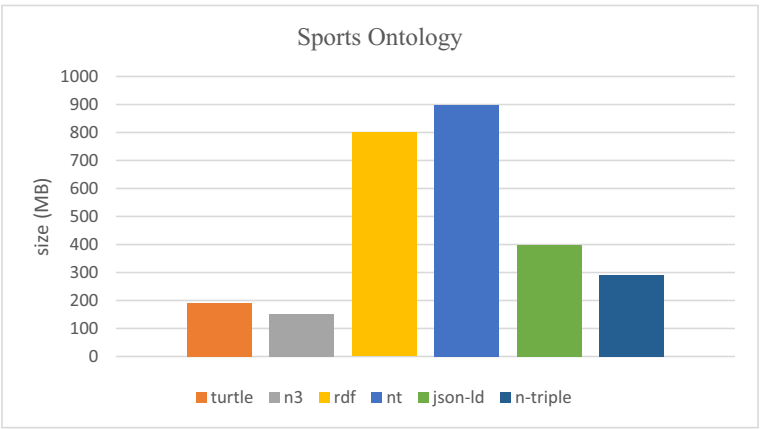

Fig. 8 Comparison of different data formats in sports ontology

higher than the K-Means clustering method and 3% higher than the DBC with DR method. The experiment results in the graphical format are shown in Fig. 7 for the BBCSports Simulated dataset, which includes various parameters (accuracy, precision, and recall).

5.2 Comparison of different data formats

The data formats are generally used to represent the ontology in order to provide descriptions of the domain. The new sports ontology has been developed and discussed in this section. The newly created sports ontology was stored in different data formats to offer compatibility with other ontologies, and a comparison was made with various data format sizes. This experiment included major RDF data formats, namely RDF/XML, NT, TURTLE, JSON-LD, N-Triple, and N3, for data size comparison of the sports ontology.

The sports ontology description consists of five major classes, namely Athletics, Cricket, Football, Rugby, and Tennis. There were various subclasses and properties involved in the ontology in order to relate terms of the document. The sports ontology classes and properties are encoded with different data formats and compared in order to achieve the smallest size. In Fig. 8, six different data formats for sports ontology has been compared; the X-axis represent the data format while Y-axis represents the data size.

The N3 data format size was approximately five times smaller than the NT and RDF data format for the Sports ontology. N-Triple format, on the other hand, was twice as massive as the N3 format. The size of the JSON-LD data format was approximately three times larger than the N3 data format size. So N3, N-triple, and turtle were found to be the best formats. The data description using JSON-LD and N-triple are medium size as compared to the NT data format. As a result, it is concluded that the N3 data format takes less space on the disk among all the data formats and is suitable for ontology representation for limited space devices.

6 Conclusion and Future Scope

Clustering using ontology is an important area that calls for a lot of work in developing new methods for successful clustering. The K-Means and DBC clustering approaches have been compared with the ontology-based document clustering. The traditional methods for document clustering, such as K-Means and Density-based clustering, use the syntactical approach, which considers only the similarity of the collection of characters (lexemes). Whereas, The newly developed ontology-based method considers the semantics of the words for document clustering. Due to this, it improves the performance of ontology-based clustering as compared to the other two approaches.

In order to implement ontology-based document clustering, ontology development for a domain is mandatory. An ontology has been created for the sports domain that contains the five sports classes, namely Athletics, Cricket, Football, Rugby, and Tennis. This ontology also includes the properties which provide the relationship between the classes. The synonyms are retrieved from the WordNet ontology using the sports ontology classes, which are then used in the clustering process. The WordNet ontology is a lexical database for cognitive synonyms, also known as synsets, and this ontology is available on the web. The WordForms are elements of a synset array that are compared in pairs with the sports ontology concepts and based on the threshold value, and the synonyms have been chosen. The ontology-based document clustering uses the K-Means clustering algorithm to cluster the semantic terms. The result obtained

from this method is significantly high. Finally, it can be concluded that the developed approach increases the efficiency of document clustering.

The comparative study of various data formats of different sizes for sports ontology is also discussed. The ontology is represented using the various data formats such as RDF/XML, NT, TURTLE, JSON-LD, N-Triple, and N3 data formats. However, every data format takes a different amount of space in order to store the ontology. The evolution of the data format brings new sets of formats along with the compatibility issue. In order to encounter this issue, various ontology formats are developed for current and future uses. The data sizes for each data format are compared, and it is found that among all the data formats, Notation-3 (N3) takes less space to save the ontology.

From the results, it is concluded that the proposed ontology-based clustering approach has surpassed both the traditional method in order to clustering process. The ontology-based clustering method for the BBCSports dataset precision was 96.6%, the recall was 93.7%, and the accuracy was 94.2%. For the ontology-based clustering method on the BBCSports simulated dataset, the precision, recall, and accuracy were 93%, 93%, and 92%. This study applies the ontology-based clustering approach only to text-based documents. Therefore, there is a scope for the experiment with other types of documents using the ontology-based clustering approach.

93, recall of 93, and accuracy of 92%, which is 18%,5.7%, and 16% higher when compared to the K-Means clustering approach and 5.4%,0.6% and 3% higher when compared to DBC with Feature Selection DR clustering approach for the BBCSports Simulated Dataset. Compare the size of different data formats of sports ontology. Therefore, the proposed ontology-based clustering approach increases clustering efficiency for large and small data sets. The future scope of the work is the proposed system clusters the text documents; in the future, it can be extended to clustering all types of documents.

References

1. Abualigah LM, Khader ATA, Hanandeh ES (2018) A Novel Weighting Scheme Applied to Improve the Text Document Clustering Techniques
2. Balabantaray R, Sarma C, Jha M (2015) Document Clustering using K-Means and K-Medoids. ArXiv abs/1502.0
3. Basov N, Nooy W, Nenko A (2018) Emergent Meaning Structures: A Socio-Semantic Network Analysis of Artistic Collectives
4. Dou D, Wang H, Liu H (2015) Semantic Data Mining: A Survey of Ontology-based Approaches
5. Gangavane HN, Nikose MC, Chavan PC (2015) A novel approach for document clustering to criminal identification by using ABK-means algorithm. In: 2015 International Conference on Computer, Communication and Control (IC4). pp 1–6
6. Gupta M, Garg K (2016) Attribute Weighted K-means For Document Clustering. Int Res J Eng Technol 03:1583–1589
7. Hira ZM, Gillies DF (2015) A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. Adv Bioinforma 2015:198363. https://doi.org/10.1155/2015/198363
8. Jang M, Choi J, Allan J (2017) Improving Document Clustering by Removing Unnatural Language
9. Jia C, Carson MB, Wang X, Yu J (2018) Concept Decompositions for Short Text Clustering by Identifying Word Communities. Pattern Recogn 76:691–703. https://doi.org/10.1016/j.patcog.2017.09.045
10. Jovic A, Brkic K, Bogunovic N (2015) A review of feature selection methods with applications. 2015 38th Int Conv Inf Commun Technol Electron Microelectron 1200–1205
11. Khan A, Salim N, Jaya Kumar Y (2015) A framework for multi-document abstractive summarization based on semantic role labelling. Appl Soft Comput 30:737–747. https://doi.org/10.1016/j.asoc.2015.01.070

12. Li P (2016) Semantic Reasoning on the Edge of Internet of Things
13. MacQueen JB (1967) Some Methods for Classification and Analysis of MultiVariate Observations. In: Le CLM, Neyman J (eds) Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, pp 281–297
14. Narzary J, Basumatary M (2016) A Methodology for Incorporation of Domain Ontology in Knowledge Discovery Process for Interpretation and Improvement of Mining Results
15. Oladele TO, Adegun AA, Ogundokun RO, Okeyinka AE, Ayeni L (2019) Application of Floyd-Warshall's algorithm in air freight service in Nigeria. Int J Eng Res Technol 12:2529–2535
16. Razia Sulthana A, Subburaj R (2016) An Improvised Ontology based K-Means Clustering Approach for Classification of Customer Reviews. Indian J Sci Technol 9:1–6. https://doi.org/10.17485/ijst/2016/v9i15/87328
17. Saad A, Shaharin S (2016) The Methodology for Ontology Development in Lesson Plan Domain. Int J Adv Comput Sci Appl 7
18. Saiyad NY, Prajapati HB, Dabhi VK (2016) A survey of document clustering using semantic approach. In: 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT). IEEE, pp 2555–2562
19. Soliman SS, El-Sayed MF, Hassan YF (2015) Semantic Clustering of Search Engine Results. Sci World J 2015:931258. https://doi.org/10.1155/2015/931258
20. Thijs B, Glänzel W, Meyer M (2015) Using noun phrases extraction for the improvement of hybrid clustering with text- and citation-based components. The example of "information System Research". CEUR Workshop Proc 1384:28–33
21. Wei T, Lu Y, Chang H, Zhou Q, Bao X (2015) A Semantic Approach for Text Clustering Using WordNet and Lexical Chains. Expert Syst Appl 42:2264–2275. https://doi.org/10.1016/j.eswa.2014.10.023
22. Yue L, Zuo W, Peng T, Wang Y, Han X (2015) A fuzzy document clustering approach based on domain-specified ontology. Data Knowl Eng 100:148–166. https://doi.org/10.1016/j.datak.2015.04.008
23. Zheng CT, Liu C, Wong HS (2018) Corpus-Based Topic Diffusion for Short Text Clustering. Neurocomput 275:2444–2458. https://doi.org/10.1016/j.neucom.2017.11.019