Noble Approach for Documents Clustering Using Semantics Relations and K-means Algorithm

Wai Wai Lwin

University of Computer Studies, Yangon

ABSTRACT

Fast and high-quality document clustering algorithms play an important role in effectively navigating, summarizing, and organizing information. Recent studies have shown that partitional clustering algorithms are more suitable for clustering large datasets. However, the K-means algorithm, the most commonly used partitional clustering algorithm, can only generate a local optimal solution. In this research work, we present the k-means clustering algorithm with semantic wordnet to perform a globalized search in the entire solution space. We also used the semantic relations of dataset using wordnet. The proposed method can support the efficient clustering approach for document clustering.

Keyword : *document clustering, Semantic relations, k-means, wordnet.*

1. INTRODUCTION

Near about 90 % web data is unstructured and needed to be structure as it greatly reduces the efficiency in using web information. Web text feature extraction and clustering are the main challenging tasks in web data mining, which requires an efficient clustering technique [1]. Data mining is the process of extracting the implicit, previously unknown and potentially useful information from data. That information can be used to decrease search time and cuts costs. Digitized text documents are increasing exponentially. As such, clustering becomes imperative for ever increasing digitized data [2]. Vector Space Model is a widely used method for document representation in information retrieval. In this model, each document is represented by a feature vector. The unique terms occurring in the whole document collection are identified as the attributes (or features) of feature vector. There are some common used term weighting methods such as binary method, tf (term frequency) method, and tf-idf (term frequency-inverse document frequency) method etc. in the vector space model. The "bag of words" feature representation is not able to reflect the semantic content of a document because of the synonym problem and polysemy problem [3].

In this paper, we present a semantic relation approach that using a wordnet and K-means algorithm to cluster the Business related news of News Websites and Financial Websites. We generate a domain ontology for representing the term with enriched semantic relation and synonyms and then clustering of the documents using K-means clustering technique to achieve the semantic concepts of terms, to filter web documents precisely and a better performance for inter and intra cluster similarity.

2. SEMANTIC RELATIONSHIPS

Semantic relationships are the associations that there exist between the meanings of words (semantic relationships at word level), between the meanings of phrases, or between the meanings of sentences (semantic relationships at phrase or sentence level). Following is a description of such relationships.

2.1 Semantic Relationships at Word Level

At word level, we will study semantic relationships like the following: synonymy, antonymy, homonymy, polysemy and metonymy.

2.1.1 Synonymy

Synonymy is the semantic relationship that exists between two (or more) words that have the same (or nearly the same) meaning and belong to the same part of speech, but are spelled differently. In other words, we can say that synonymy is the semantic equivalence between lexical items. The (pairs of) words that have this kind of semantic relationship are called synonyms, or are said to be synonymous.

E.g.

big = large hide = conceal small = little

couch = sofa to begin = to start kind = courteous

beginning = start to cease = to stop fast = quickly = rapidly

Pairs of words that are synonymous are believed to share all (or almost all) their semantic features or properties. However, no two words have exactly the same meaning in all the contexts in which they can occur. For example, the verbs employ and use are synonymous in the expression We used/employed effective strategies to solve the problem; however, only use can be used in the following sentence: We used a jimmy bar to open the door. If we used employ, the sentence would sound awkward *We employed a jimmy bar to open the door. In short, we can say that there are no absolute synonyms, i.e., pairs of words that have the same meaning (or share the same semantic features) in all the situational and syntactic contexts in which they can appear.

2.1.2 Antonymy

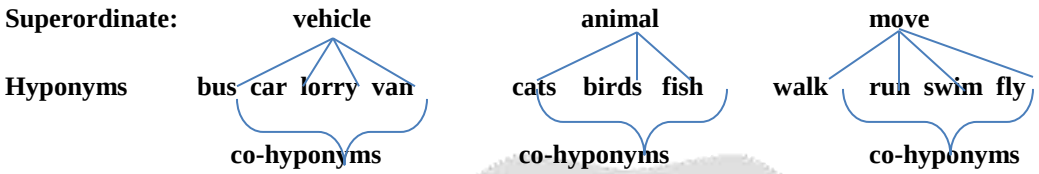
Antonymy is the semantic relationship that exists between two (or more) words that have opposite meanings. The pairs of words which have opposite meanings are called antonyms.
Homonymy Homonymy is the relationship that exits between two (or more) words which belong to the same grammatical category, have the same spelling, may or may not have the same pronunciation, but have different meanings and origins (i.e., they are etymologically and semantically unrelated). E.g., to lie (= to rest, be, remain, be situated in a certain position) and to lie (= not to tell the truth); to bear (= to give birth to) and to bear (= to tolerate); bank (= the ground near a river) and bank (= financial institution); lead [li…d] (= the first place or position, an example behavior for others to copy) and lead [led] (= heavy metal); bass [beIs] (= musical instrument) and bass [bœs] (= edible fish). The pairs of words that exhibit this kind of relationship are called homonyms. Homonyms usually have different entries in dictionaries, often indicated by superscripted little numbers; e.g., lie1, lie2. In isolated spoken sentences, homophonic homonyms can also give rise to lexical ambiguity. For example, in the following sentences it is almost impossible to know the intended meanings of bank and bear. Notice the following sentences. John went to the [bœNk] (the financial institution or the ground by the river?) Mary can't [bE´r] (have or tolerate?) children.

2.1.3 Hyponymy

Hyponymy ([ÆhaI"pÓAn´mi] or [hI"pÓAn´mi]) or inclusion is the semantic relationship that exists between two (or more) words in such a way that the meaning of one word includes (or contains) the meaning of other words(s). We say that the term whose meaning is included in the meaning of the other term(s) is the general term; linguists usually refer to it as a superordinate or hypernym. The term whose meaning includes the meaning of the other term is the specific term; linguists usually refer to it as a hyponym. If the meaning of a superordinate term is included in the

meaning of several other more specific words, the set of specific terms which are hyponyms of the same superordinate term and are called cohyponyms (cf. Crystal, 1991).

Examples:



2.1.4 Polysemy

Polysemy ([p´"lIs´mi]) is the semantic relationship that exists between a word and its multiple conceptually and historically related meanings (cf. Crystal, 1991; Fromkin & Rodman, 1998; Richards et al., 1992).

E.g., foot = 1. part of body; 2. lower part of something

plain = 1. clear; 2. unadorned; 3. obvious.

nice = 1. pleasant; 2. kind; 3. friendly; etc.

The different meanings of a word are not interchangeable; in fact, they are context-specific. [4]

3. WORDNET

Name used for lexical databases derived from the original Princeton WordNet; they group words into synonym sets and interlink them using lexical and conceptual-semantic relations. Used for computational linguistics and natural language processing.Lexical semantic database with concepts represented by synonyms in a language, so-called synsets, with semantic relations between these concepts.Acommonsense knowledge base implementation based on a semantic net structure.

WordNet is a semantic lexicon for the English language. It groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. The purpose is twofold: to produce a combination of dictionary and thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications. WordNet was developed by the Cognitive Science Laboratory (http:// www.cogsci.princeton.edu/) at Princeton University under the direction of Professor George A. Miller (Principal Investigator). WordNet is considered to be the most important resource available to researchers in computational linguistics, text analysis, and many related areas. Its design is inspired by current psycholinguistic and computational theories of human lexical memory.

It groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. WordNet was created and is being maintained at the Cognitive Science Laboratory of Princeton University under the direction of psychology professor George A. Miller. Development began in 1985.

4. K-MEANS CLUSTERING ALGORITHM

Given a dataset of n data points x1, x2, …, xn such that each data point is in Rd , the problem of finding the minimum.

$$\frac{1}{n}\sum_{i=1}^{n}\left[min_j \; d^2(x_i, m_j)\right] \qquad (1)$$

Where d(xi, mj) denotes the Euclidean distance between xi and mj. The points {mj} (j=1, 2, …,k) are known as cluster centroids. The problem in Eq.(1) is to find k cluster centroids, such that the average squared Euclidean distance (mean squared error, MSE) between a data point and its nearest cluster centroid is minimized.

The k-means algorithm provides an easy method to implement approximate solution to Eq.(1). The reasons for the popularity of k-means are ease and simplicity of implementation, scalability, speed of convergence and adaptability to sparse data.

The k-means algorithm can be thought of as a gradient descent procedure, which begins at starting cluster centroids, and iteratively updates these centroids to decrease the objective function in Eq.(1). The k-means always converge to a local minimum. The particular local minimum found depends on the starting cluster centroids. The problem of finding the global minimum is NP-complete. The k-means algorithm updates cluster centroids till local minimum is found. Fig.1 shows the generalized pseudocodes of k-means algorithm; and traditional k-means algorithm is presented in fig. 2 respectively.

Before the k-means algorithm converges, distance and centroid calculations are done while loops are executed a number of times, say l, where the positive integer l is known as the number of k-means iterations. The precise value of l varies depending on the initial starting cluster centroids even on the same dataset. So the computational time complexity of the algorithm is O(nkl), where n is the total number of objects in the dataset, k is the required number of clusters we identified and l is the number of iterations, $k \leq n$, $l \leq n$ [6].

Step 1: Accept the number of clusters to group data into and the dataset to cluster as input values

Step 2: Initialize the first K clusters

- Take first k instances or
- Take Random sampling of k elements

Step 3: Calculate the arithmetic means of each cluster formed in the dataset.

Step 4: K-means assigns each record in the dataset to only one of the initial clusters

- Each record is assigned to the nearest cluster using a measure of distance (e.g Euclidean distance).

Step 5: K-means re-assigns each record in the dataset to the most similar cluster and re-calculates the arithmetic mean of all the clusters in the dataset.

Figure 1: Generalised Pseudocode of Traditional k-means

```
1   MSE = largenumber;
2   Select initial cluster centroids {mj}j  K = 1;
3   Do
4   OldMSE = MSE;
5   MSE1 = 0;
6   For j = 1 to k
7   mj = 0; nj = 0;
8   endfor
9   For i = 1 to n
10  For j = 1 to k
11  Compute squared Euclidean  distance d2(xi, mj);
12  endfor
13  Find the closest centroid mj to xi;
        14  mj = mj + xi; nj = nj+1;
15  MSE1=MSE1+ d2(xi, mj);
16   endfor
17  For j = 1 to k
18  nj = max(nj, 1); mj = mj/nj;
19  endfor
20  MSE=MSE1;
    while (MSE<OldMSE)
```

Figure 2: Traditional k-means algorithm

5. CONCLUSION

In this paper, we proposed the noble approach for document clustering using semantic relations and k-means clustering algorithm.. By using Semantic representation and k-means clustering algorithm will achieve the high performance of intra cluster similarity. Moreover, it is found that the inter cluster similarity achieves the appropriate results also. The proposed system is ongoing work and the future work of the system will upgrade the Wordnet and make effort to change the efficient algorithm for clustering web document data.

REFERENCES

[1] J. Ghorpade-Aher, R. Bagdiya, *"A Review on Clustering Web Data using PSO"*, International Journal of Computer Applications (0975 – 8887) Volume 108 – No. 6, December 2014

[2] R. Bhagel, and R. Dhir, *"A Frequent Concept Based Document Clustering Algorithm", IJCA*, vol 4, no.5, 2010

[3]W. K. Gad and M. S. Kamel, *"Enhancing Text Clustering Performance Using Semantic Similarity",ICEIS, LNBIP 24*, pp. 325–335, 2009

[4] Prof. Argenis A. Zapata, Universidad de Los Andes, Facultad de Humanidades y Educación, Escuela de Idiomas Modernos, "*Inglés IV (B-2008)"*

[5] Fahim A. M., Salem A. M., Torkey F. A. and Ramadan M. A., "An efficient enhanced k-means clustering algorithm," Journal of Zhejiang University Science A., pp. 1626–1633, 2006

[6] Oyelade, O. J., Oladipupo, O. O., Obagbuwa, I. C., "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance," IJCSIS, pp. 292–295, Vol. 7, No. 1, 2010