

Document Clustering with Semantic Analysis using RapidMiner

Somya Chauhan¹ and G.Hannah Grace²

¹VIT Vellore, India

somyachauhan08@gmail.com

²VIT Chennai, India

hannahgrace.g@vit.ac.in

Abstract

Document Clustering is the process of forming clusters from the whole document and is used in multiple fields like information retrieval, text mining. Earlier, when we were having large document, it was tedious task to parse the whole document and determine all the context of it. In this paper, we investigate how the different wordNet operators helps in improving the performance of the document clustering system. In this paper, we also applied the EM clustering algorithm and the comparison of EM and K-Means clustering algorithms.

Key Words: Document Clustering, WordNet, Semantic Analysis, EM-Expectation Maximization

1 Introduction

Document Clustering is a collection of textual and numeric data. Document clustering plays an important role in providing intuitive navigation and browsing mechanisms by organizing large amounts of information into a small number of meaningful clusters[1]. Document clustering can be applied to a document database so that similar documents are related in the same cluster. During the retrieval process, documents belonging to the same cluster as the

retrieved documents can also be returned to the user[2]. Traditionally, single or compound words occurring in the document set are used as the features. Because of the synonym problem and polysemous problem, generally such a bag of words cannot reflect the semantic content of a document[5]. In this paper, we investigate using semantic analysis to identify the semantic content of documents. In our system we applied various semantic related operators. WordNet is a thesaurus for English, provided by rapidminer on the document set, which helps in improving the quality of the cluster.

2 Tools used

2.1 WordNet:

Unlike Traditional dictionaries are arranged alphabetically, WordNet is arranged semantically[10]. It is a lexical database for English language[16]. It groups English words into set of synonyms called synsets, and records a number of relations among these synonym sets or their members[9]. It is a freely available software package of Princeton that makes it possible to measure the semantic similarity or relatedness between a pair of concepts (or word senses)[12]. WordNet groups nouns, verbs, adjectives and adverbs into sets of synonyms called synsets[8]. Synsets are organized into senses, giving the synonyms of each word, also hyponym/hypernym (i.e. Is-A), and meronym/holonym (i.e. Part-Of) relationships[11].

2.2 Rapid Miner:

It provides an integrated environment for machine learning, data mining, text mining, predictive analytics and other analytic methods. It is used for research, education, training, rapid prototyping and application development and supports all steps of the data mining process including data preparation, results visualization, validation and optimization[6]. It provides several operators which supports fast processing and also less overhead of implementing the core logic. These operators can be applied directly to the document.

3 Process for Semantic Analysis

The document set is processed from file. For Semantic Analysis of the document, the document is passed to the following operators in sequence- Tokenize, Transform Cases, Filter Stopwords(English), FilterTokens(By Length), WordNet Dictionary. The processed document contains the terms. A term is any sequence of characters separated from other terms which is also known as vector space model by some delimiter.

3.1 Tokenize:

This operator splits the text of a document into a sequence of tokens. Tokens are the smallest unit of a word.

3.2 Transform Cases:

This operator transforms all characters in a document to either lower case or upper case.

3.3 Filter Stopwords:

This operator filters English stopwords from a document by removing every token which equals a stopword from the built-in stopword list.

3.4 Filter Tokens(by Length):

This operator filters tokens based on their length (i.e. the number of characters they contain).

3.5 Wordnet dictionary:

Directory containing the Wordnet dictionary[11].

4 Probabilistic Approach:

Probabilistic model determines a soft clustering, in which each document has a membership probability of the cluster, as opposed to

a hard segmentation of the documents[14]. A probabilistic model describes a set of possible probability distributions for a set of observed data, and the goal is to use the observed data to learn the distribution[4].

4.1 Expectation Maximization Algorithm:

The EM (Expectation Maximization) extends the combination of K-Means and probabilistic approach. The expectation maximization algorithm enables parameter estimation in probabilistic models with incomplete data[3]. In K-Means given a fixed number of k clusters, assign observations to those clusters so that the means across clusters are as different from each other as possible. The EM algorithm extends the K-Means approach to clustering in two different ways: Instead of assigning examples to clusters to maximize the differences in means for continuous variables, the EM clustering algorithm compute probabilities of cluster memberships based on one or more probability distributions. The aim of this clustering algorithm is to maximize the overall probability or likelihood of the data, given the (final) clusters[15]. Also unlike the classic implementation of k-means clustering, the general EM algorithm can be applied to both continuous and categorical variables which provides wider range of input type support[20].

The K-means will assign observations to clusters to maximize the distances between clusters but the EM algorithm does not compute actual assignments of observations to clusters, but classification probabilities i.e., each observation belongs to each cluster with a certain probability[7].

There are two main applications of the EM algorithm. The first occurs when data indeed has missing values, due to problems with or limitations of the observation process. Second occurs when optimizing the likelihood function is analytically intractable but when the likelihood function can be simplified by assuming the existence of and values for additional but missing (or hidden) parameters[21].

The EM algorithm input are the data set (x), the total number of clusters (M), and the maximum number of iterations. The EM algorithm looks to find the MLE of the marginal likelihood by iteratively applying these steps:

K-Means Resultset	EM Resultset
Cluster 0: 50 items Cluster 1: 86 items Total number of items: 136	Cluster 0: 72 items Cluster 1: 64 items Total number of items: 136 cluster probabilities: Cluster 0: 0.5294293232557149 Cluster 1: 0.4705706767442852

Table 1: Comparison of Resultset of EM and K-Means applied on the same document

Input: Cluster number k , a database.

Output: A set of k -clusters with weight that maximize log-likelihood function.

1. E step: For each record x , compute the membership probability of x in each cluster $h = 1 \dots k$.
2. M step: Update mixture model parameter (probability weight).
3. Stopping criteria: If it is satisfied stop, else set $j = j + 1$ and go to (1).

The iterative EM algorithm uses a random variable and, is a general method to find the optimal parameters of the hidden distribution function from the given data, when the data are incomplete or has missing values[22].

5 Process Flow Diagram:

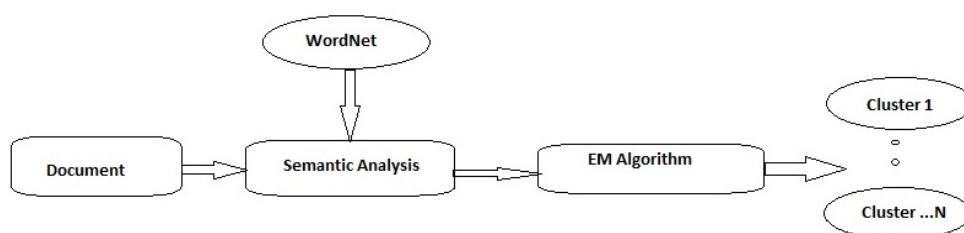


Figure 1: Process Flow Diagram

ExampleSet (6 examples, 4 special attributes, 6 regular attributes) Filter (6 / 6 examples): all

Row No.	label	metadata_file	metadata_d...	metadata_p...	Annual Clim...	Hourly Preci...	Monthly Su...	PK	! 0...	PK
1	Bigdatawitho...	Bigdataset1...	Feb 6, 2017 1...	C:\Users\So...	0	0	0	0		1
2	Bigdatawitho...	Colleted.docx	Feb 9, 2017 8...	C:\Users\So...	0	0	0	1		0
3	Bigdatawitho...	Daily.txt	Jul 14, 2016 ...	C:\Users\So...	0	0	0	0		0
4	Bigdatawitho...	Hourly precipi...	Jul 14, 2016 ...	C:\Users\So...	0	1	0	0		0
5	Bigdatawitho...	Monthly.txt	Jul 14, 2016 ...	C:\Users\So...	0	0	1	0		0
6	Bigdatawitho...	New Text Doc...	Jul 14, 2016 ...	C:\Users\So...	1	0	0	0		0

Figure 2: Resultset Before Applying Semantic Analysis

ExampleSet (6 examples, 5 special attributes, 1362 regular attributes) Filter (6 / 6 examples): all

Row No.	label	sentiment	metadata_file	metadata_d...	metadata_p...	aaiaæf	access	accumulation	accum
1	WeatherSet ...	0.167	Bigdataset1...	Feb 6, 2017 1...	C:\Users\So...	0.034	0	0	0
2	WeatherSet ...	0.167	Colleted.docx	Feb 9, 2017 8...	C:\Users\So...	0	0	0	0
3	WeatherSet ...	0.028	Daily.txt	Jul 14, 2016 ...	C:\Users\So...	0	0	0	0
4	WeatherSet ...	0.026	Hourly precipi...	Jul 14, 2016 ...	C:\Users\So...	0	0	0.194	0
5	WeatherSet ...	-0.043	Monthly.txt	Jul 14, 2016 ...	C:\Users\So...	0	0.160	0	0.160
6	WeatherSet ...	0.021	New Text Doc...	Jul 14, 2016 ...	C:\Users\So...	0	0	0	0

Figure 3: Resultset After Applying Semantic Analysis

The later image has more attributes which means the document is categorized more meaningfully. Thus, Semantic Analysis helps in reducing the document size meaningfully. The document is further applied over EM operator, the clusters obtained were more accurate.

6 Cluster Quality Result

Q1(Performace % Before SA)	Q2(Performance % After SA)
0.670	0.870

Table 2: Comparison of Quality

The clustering of documents, usually on the basis of the terms that have in common[14]. After the clustering processes, clusters should be widely spaced, i.e., the separation between clusters should be maximized, while the members of each cluster should be as close to each other as possible, i.e., the compactness in clusters should be minimized[18]. Clustering analysis attempts to discover distribution patterns of data objects[17]. The Result shows the improve in quality when the same document sets are applied with and without the Semantic Analysis over EM clustering algorithm.

7 Discussion

The Document Clustering with Semantic Analysis using Rapid-Miner provides more accurate clusters. Performing syntactic analysis to find the important word in a context. All the words or compound words in a sentence are considered to be independent and of the same importance. Combining the semantic analysis and syntactic analysis to analyze a further improvement.

7.1 Conclusion

We have evaluated the cluster quality to validate our result over the same set of documents and applied EM algorithm. We have taken the documents without semantic analysis and using semantic analysis. From the result in Table (1) it is clear that using semantic analysis improves the clustering solution. As a future work, we can combine the semantic and syntactic information in the form of the senses of the key words, which can further enhance the quality of clusters.

References

- [1] Neepa Shah, Sunita Mahajan, *Semantic based Document Clustering: A Detailed Review* IJAIS Foundation of Computer Science FCS, Vol 4- No. 5, Oct 2012.
- [2] Yong Wang¹ and Julia Hodges, *Document Clustering with Semantic Analysis* Proceedings of the 39th Hawaii International Conference on System Sciences 2006.
- [3] Sean Borman, *The Expectation Maximization Algorithm*, July 18, 2004.
- [4] Charu C. Aggarwal, ChengXiang Zhai. *A SURVEY OF TEXT CLUSTERING ALGORITHM*.
- [5] Textbook by Cliff Goddard, *Semantic Analysis: A Practical Introduction*.
- [6] docs.rapidminer.com/downloads/RapidMiner-v6-user-manual.pdf
- [7] Dr. David Banks, Dr. Frederick R. McMorris, Dr. Phipps Arabie, Prof. Dr. Wolfgang Gaul, *Classification Clustering and Data Mining Applications* - Springer
- [8] <https://wordnet.princeton.edu/wordnet/publications>.

- [9] GA Miller, *Introduction to WordNet - Description* - Princeton University
- [10] S. Banerjee and T. Pedersen, *An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet*, In Proc. of the Fourth International Conference of Computational Linguistics and Intelligent Text Processing (CICLING-02), Mexico City, 2002.
- [11] C. Leacock and M. Chodorow, *Combining Local Context and WordNet Similarity for Word sense Identification*, 1998, In C. Fellbaum, editor, WordNet; An Electronic Lexical Database, pp.265-283, MIT Press.
- [12] T. Pedersen, S. Patwardhan, and J. Michelizzi, *WordNet::Similarity - Measuring the Relatedness of Concepts*, In Proc. of the Proceedings of the 39th Hawaii International Conference on System Sciences - 2006 9 Nineteenth National Conference on Artificial Intelligence, San Jose, 2004.
- [13] Kaufman, L. and Rousseeuw, P., *Finding Groups in Data*, Wiley, New York, NY, 1990.
- [14] P. Willett, *Recent Trends in Hierarchic Document Clustering: A Critical Review*, Information Processing and Management, 24(5), pp.577-597, 1988.
- [15] <https://engineering.purdue.edu/kak/Tutorials/ExpectationMaximization.pdf>
- [16] jwordnet.sourceforge.net/handbook.html
- [17] <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0090109>
- [18] Jain AK, Murty MN, Flynn PJ. *Data clustering: a review*. ACM Comput. Surveys. 1999;31:264323.
- [19] Jain AK, Topchy A, Law MHC, Buhmann JM. *Landscape of clustering algorithms*. Proc. 17th Int. Conf. Pattern Recognit. (ICPR '04); 2004. pp. 260263.
- [20] Madhulatha TS. *An overview on clustering methods*. IOSR J Eng. 2012;2(4):719725.
- [21] Bilmes JA, *A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden Markov models*. Berkeley: CA: International Computer Science Institute; 1998.
- [22] <https://en.wikipedia.org/wiki/Expectation>

