INTERNATIONAL RESEARCH JOURNAL ON ADVANCED SCIENCE HUB
e-ISSN : 2582 - 4376
Open Access
RSP SCIENCE HUB
(The Hub of Research Ideas)
Available online at www.rspsciencehub.com

Special Issue of First International Conference on Social Work, Science & Technology (ICSST 2021)

Enriching and Clustering Short Text Using KNN

Ms. Shalika[1], Mr. Veepin Kumar[2]
[1]Assistant Professor, Department of Computer Applications, KIET Group of Institutions, Ghaziabad, India.
[2]Assistant Professor, Department of Information Technology, KIET Group of Institutions, Ghaziabad, India.
shalika.mca@kiet.edu[1], veepin.kumar@kiet.edu[2]

Abstract

Semantic Hashing technique wraps the meaning of short texts into compressed binary codes. So, to find out that whether two short texts are alike or not in their meaning, their binary codes need to be matched. A deep neural network is used for encoding. Bag-of-words representation of texts is used to train the neural network. Unfortunately, the fundamental semantics are not sufficiently captured by the above mentioned form of representation for short texts such as titles, tweets, or queries. We propose adding additional semantic signals to better group short texts using their meaning. More specifically, we procure the co-occurring terms and concepts of every term in the short text via a knowledge database to further enhance the short text. Additionally, we use a k-Nearest Neighbor based approach id for hashing. Multiple experiments provide evidence that by increasing the number of semantic signals, our neural network is better capable to capture the meaning of short texts, which enables various uses like retrieving information, classifying data, and processing of short texts.

keywords: Short Text, k-Nearest Neighbor, Semantic Enrichment, Hashing.

1. Introduction

Nowadays, it has become very difficult to search anything on the web that has multiple meanings. To overcome this challenge, a method needs to be defined that not only gives us a broad view about how many variations about a particular product are there but also about how diverse the world has become and how technology has flourished in this modern era. Clustering, a very common but useful word is being used in a variety of forms and for different types of applications. The main aim of any clustering technique is to minimize similarity among members of the different cluster and maximize among members belonging to same clusters. The short text is a word or a group of words that form a phrase or a sentence(s) such as a news title, tweet or a search query etc. They came to attention when the use of natural language processing applications increased and search engines began gaining popularity. Understanding short texts have become an important research topic in the past few years since it has used in many natural language processing applications. A pressing need for better enrichment and clustering of short texts is in search optimization. As short texts are different from written language and do not contain normal statistical markers like written language normal natural language processing methods such as topic modeling cannot be used to process them. Many methods exist to understand short texts by enriching them. For example, using syntax and semantically enriched data derived from various resources like Wikipedia, WordNet, the Open Directory Project (ODP), etc. allow mapping the short text to Wikipedia titles and then enriching them with corresponding articles. The short texts can be enriched with search engine results. Available methods for semantic hashing are Latent Semantic Analysis which analyses connections among a group of files and the words

they already contain by generating concepts related to the files and words. LSA functions on the assumptions that words alike in their meaning will appear in the same pieces of sentences. The present system uses the concept of Deep Neural Network for clustering and enhancing of short texts. The system converts the short text that is searched into equivalent binary codes with the help of auto-encoders. The auto-encoders perform the conversion by semantic enrichment and hashing.

All of these existing methods have shortcomings. Search based method is likely to show irrelevant results for very commonly used words causing the enriched short text to be filled with noise. On the other hand, external resource using methods are restricted by the limit of their resources. For example, WordNet falls short in the department of proper nouns [1-5].

2. Preliminaries

2.1 k-Nearest Neighbor

The k-nearest neighbor algorithm store available cases and classifies new cases on the basis of some similarity measure. It is a non-parametric method. It is mainly used for regression and classification. The input is the k closest examples in both cases and output depends on the case.

- In k-NN regression object's property value is the output.
- In k-NN classification object's class membership is the output.

Since k-NN is a method which is based on instances or shallow learning all functions are approximated raw-fully and computation is negotiated until classifying data is done. The neighbors are taken from the set of objects whose property value or class is known. It can be considered as the training set.

Algorithm

Training examples consist of vectors in a multidimensional space, each having a class label. During the training phase vectors and class label of samples are stored. In classification phase, k is a constant defined by the user, and an unlabeled vector is labeled by selecting a label which is frequent among the k samples nearest to the unlabeled vector.

2.2 Clustering

The main purpose of clustering is to determine how data can be grouped together into an intrinsic cluster from a given set of raw data. It only and only depends on the user and his/her criteria that makes out a good cluster and is independent of the final aim of clustering. The main focus should be on increasing similarity within the same cluster and minimizing similarity among different clusters so as to obtain the highest accuracy.[6-10].

2.3 Methods of Clustering

Clustering is broadly divided into two approaches-

1. Hierarchical Clustering keeps on linking pairs of clusters until each data object is included in the hierarchy. Further two main methods are used -

a) Agglomerative Methods start with each data point as a single cluster and after each step, combine other instances to form further clusters until there is only one big cluster.

b) Divisive Methods begin with one large cluster and proceed to split into smaller and smaller clusters of items that are the most different.

2. Non-hierarchical Clustering is based on segmentation of data, which is further based on some measure of dissimilarity between segments and the much-needed degree of segmentation. It begins with a few seed clusters and then each object is tested based on which it is either placed within the predefined cluster or a new cluster is created. There are three kinds of methods used-

a) Partitioning Methods begin with a number of arbitrary data points. The most popular algorithm is the k-means algorithm. It begins with initialization of a cluster center for each of the data point. Then each data point is assigned to the group which has its center point closest to the data point (determined by Euclidean distance). After assigning all the data points, new cluster centers are calculated. Finally, this process is repeated again and again until there is no further change in cluster centers.

b) Density-based Methods are best used when clusters have irregular shapes. It works by putting data points together in high-density areas as members belonging to the same cluster and treating low-density areas as boundaries between any two clusters. Boundaries are expanded until a point is reached where the required density is not met.

c) Probability-based Methods is based on probability distribution. It is also known as Expectation Maximization (EM) algorithm. First of all posterior probability of a given class is computed and then multiplied with the value of the

data point, further all these values are combined into a suitable function, which is finally maximized to get the final result.

2.4 Fundamental Steps in Clustering of Data

For clustering of data, the very first step is to extract data from the knowledge/database. It deals with problem identification followed by collation of data and finally, it's pre-processing. The very next step is to choose a suitable algorithm. Hit and Trial method is most commonly is used. However, parametric algorithms are relatively much more suitable as they involve choosing the optimal parameters to get the best solution. Data is further processed in order to make it compatible with the chosen algorithm, involving data normalization, data transformation, and data integration. Now a software package is found that automatically develops a model from the training dataset.

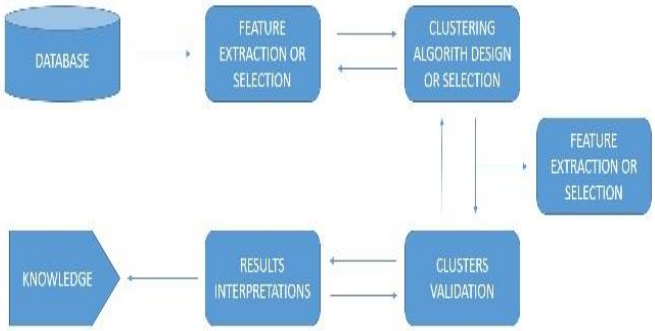

Fig. 1: Fundamental Steps of Clustering

Validating the clusters is also an important task. There is a need to test whether the given clusters formed are significantly different from each other. For this, there are certain tests that are undertaken. If there are only a couple of clusters, computing the mean and standard deviation of each of the clusters and then dividing the difference of means by the pooled standard deviation gives the 't' value. This value is compared with the value in the table if the computed value is greater than the critical value; the two clusters are significantly different. What if there are more than two clusters, then a chi-square test is carried out where the mean of each cluster is subtracted from the overall mean of the dataset. This difference is then squared and divided by the overall mean. Sum of all these values from all clusters is the computed chi-square. The last option is to split the dataset into test and training cases randomly. The centroids of clusters from the training cases can be used to cluster the test cases. The centroids of clusters formed by the test data are then computed and compared with the training data which finally validate the clustering done. Finally, the results are analyzed and interpreted, from which knowledge is gained as a result of processing the information in the form of data.[11-15].

3. Enrichment Using K-NN

The proposed model uses "k nearest neighbor" based approach is for enrichment of the short text. The proposed system semantically enriches short texts with the aid of both concepts and co-occurring terms, such as knowledge derived from external sources that are theorized from probase. The model provides a method to merge neural network and knowledge information for analyzing short texts so that it helps machine understands short texts much better.

K Nearest Neighbors

a) Use the local neighborhood to obtain a prediction.

b) k memorized examples more similar to the one that is being classified are retrieved.

c) A distance function is needed to compare the examples similarity. There are number of distance functions namely Euclidean Distance, Manhattan Distance, Hamming Distance and Minkowski Distance.

d) By changing the distance, we change how examples are classified.

3.1 Algorithm Used

The main approach used for grouping is k nearest neighbor (K-NN). It stores all possible cases and classifies new cases based on a distance function. The algorithm proceeds as follows,

- A case is classified based on the majority of votes from its neighbors and is finally assigned to the group most common among its k nearest neighbors.
- Distance is measured by a distance function. If k=1, the case is simply assigned to the group of its nearest neighbor.
- Basically, three distance measures are used namely Euclidean Distance, Manhattan Distance and Minkowski Distance but Euclidean is the most commonly used.
- All these three functions are valid only for continuous variables. For categorical data, hamming distance is better suited.

- Choose an optimal value for K by firstly inspecting the data. Larger the K value, more precise is the result but without any guarantees.
- Cross-validation is another technique to determine K value by comparing independent data sets for validation.
- The ideal value of K lies between 3 -10.
- However, if both numerical, as well as categorical data are used in a mixture within a single dataset, then standardization of the numerical variables between 0 and 1 must be carried out.

4. Model Design

4.1 Problem Formulation

In the existing system, Deep Neural Network (DNN) has been used for making decisions and the algorithm used in the proposed system is K Nearest Neighbor (K-NN). K-NN is a supervised learning approach whereas DNN is an unsupervised learning approach. DNN needs high-performance hardware for training the data and execution which is not the case when implementing k-NN algorithm requires no training of dataset whereas DNN requires very long time to train and it is extremely computationally expensive. Also, DNN is not feasible for everyone since the most complex models generally take weeks to train using hundreds of machines equipped with expensive GPUs. DNN requires a lot of unlabeled training data to make concise conclusions. Unlike DNN, k-NN can provide enough transparency for its decisions. In case of DNN, what is learned is not very easily comprehended. It is much easier to understand what's going on during training using other classifiers like decision trees, logistic regression etc. DNN can only give good results when the training data set is very large otherwise it is highly unlikely that DNN will perform better than other approaches.[16-20]

4.2 Model Architecture

To begin with, the user, first of all, enters a query on the search engine. The engine then provides the user with a list of URLs, which are either uploaded as a probabilistic base or already saved in the database. These URLs are ranked using top K search algorithm. Admin, on the other hand, has the responsibility of firstly extracting the information from the internet and then processing it so that it can further proceed for grouping. A web crawler can be used to browse the World Wide Web systematically for the purpose of indexing the websites and allowing a user to interact with the most popular website before the opposite. But here we have used k-NN (k nearest neighbor) algorithm, a method which doesn't use any parameters for classifying and regression. Output in k-NN is always a member of a class. The object of a class is classified by majority votes given by its neighbors i.e. if k=1, then simply the object is assigned to the single nearest neighbor class. Neighbors are chosen from a given group of objects for which either the class or the object property value is known beforehand. It is quite sensitive towards local and raw data.[21-25].

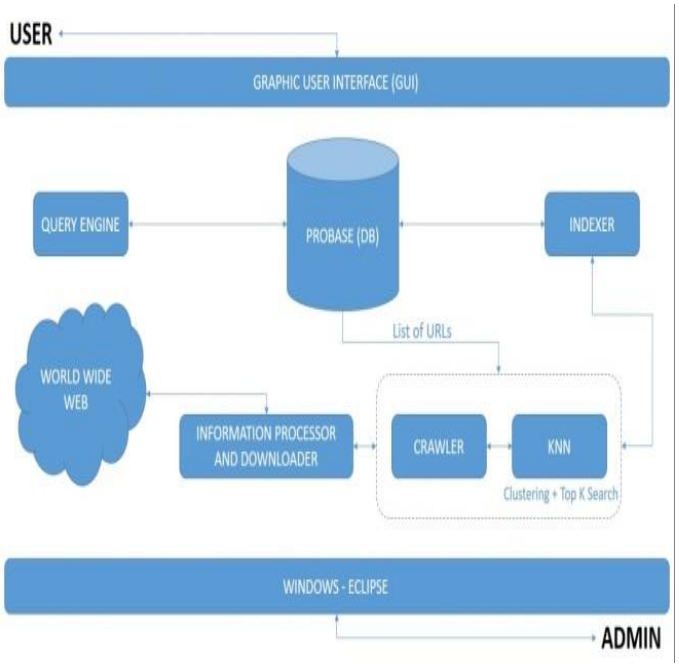

Fig 2: Architecture

Conclusion

In this research paper, we propose a technique for enriching and clustering the short texts using k-nearest neighbor approach. We first enrich the short texts by using external sources. A web crawler crawls the website on the internet and stores the enriched short texts into a database. The experiments performed on the proposed model show that the model is able to provide meaningful results for proper nouns. The existing systems did not provide very accurate results in case of a proper noun and proper nouns are an important part of the language. The comparison of experimental results with existing systems showed that the proposed model was a significant improvement in space and time complexity.

Future Work

In the future, K-NN is not a very cheap prediction algorithm; hence a different algorithm can be used to reduce the cost of prediction. It has a lazy approach, the entire data set might be used in decision making thus a different machine learning algorithm with more self-learning approach and using lesser data could be used. While using K-NN, in a very low dimensional space we can use an RP-Tree or KD-Tree to improve its performance. The value of k can (and should) be optimized using a hold-out training data set, which in turn will optimize the results. K-NN needs to be carefully tuned. The choice of K and the metric (distance) to be used are critical. MYSQL is being used at local host to store our data. Although there are many advantages of self-hosting as compared to paid hosting we can use paid hosting.

References

[1]. G. E. Hinton, "Training products of experts by minimizing contrastive divergence," Neural Computation, vol. 14, no. 8, pp. 1771–1800, 2002.

[2]. D. Meyer, F. Leisch, and K. Hornik, "The support vector machine under test," Neurocomputing, vol. 55, no. 1-2, pp. 169–186, 2003.

[3]. "Latent Dirichlet Allocation", David M. Blei, Andrew Y. Ng and Michael I. Jordan, Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.

[4]. "Feature generation for text categorization using world knowledge," in IJCAI, 2005, pp. 1048–1053.

[5]. "Feature Generation for Text Categorization Using World Knowledge", Evgeniy Gabrilovich and Shaul Markovitch, 2005

[6]. M. Sahami and T. D. Heilman, "A web-based kernel function for measuring the similarity of short text snippets," in WWW, 2006, pp.377–386.

[7]. D. Shen, R. Pan, J.-T. Sun, J. J. Pan, K. Wu, J. Yin, and Q. Yang, "Query enrichment for web-query classification," ACM Trans. Inf. Syst., vol. 24, no. 3, pp. 320–352, 2006.

[8]. G. E. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, pp. 504–507, 2006.

[9]. P. V. Gehler, A. Holub, and M. Welling, "The rate adapting Poisson model for information retrieval and object recognition," in ICML, 2006, pp. 337–344.

[10]. W. tau Yih and C. Meek, "Improving similarity measures for short segments of text," in AAAI, 2007, pp. 1489–1494.

[11]. S. Banerjee, K. Ramanathan, and A. Gupta, "Clustering short texts using Wikipedia," in SIGIR, 2007, pp. 787–788.

[12]. B. Stein, "Principles of hash-based text retrieval," in Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2007, pp. 527–534.

[13]. "Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis", Evgeniy Gabrilovich and Shaul Markovitch, in IJCAI, 2007, pp. 1606–1611.

[14]. C. D. Manning, P. Raghavan, and H. Sch¨utze, Introduction to information retrieval. Cambridge University Press, 2008.

[15]. X. H. Phan, M. L. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections," in WWW, 2008, pp. 91–100.

[16]. O. Egozi, E. Gabrilovich, and S. Markovitch, "Concept-based feature generation and selection for information retrieval," in AAAI, 2008, pp.1132–1137.

[17]. X. Hu, N. Sun, C. Zhang, and T.-S. Chua, "Exploiting internal and external semantics for the clustering of short texts using world knowledge," in CIKM, 2009, pp. 919–928.

[18]. Y. Bengio, "Learning deep architectures for AI," Foundations and Trends in Machine Learning, vol. 2, no. 1, pp. 1–127, 2009.

[19]. R. Salakhutdinov and G. E. Hinton, "Semantic hashing," Int. J. Approx. Reasoning, vol. 50, no. 7, pp. 969–978, 2009.

[20]. Y. Song, H. Wang, Z. Wang, H. Li, and W. Chen, "Short text conceptualization using a probabilistic knowledge base," in IJCAI, 2011, pp. 2330–2336.

[21]. W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: a probabilistic taxonomy for text understanding," in SIGMOD Conference, 2012, pp. 481–492.

[22]. P. Li, H. Wang, K. Q. Zhu, Z. Wang, and X. Wu, "Computing term similarity by large probabilistic ISA knowledge," in CIKM, 2013,

pp. 1401–1410.
[23]. “Context-Dependent Conceptualization”, Dongwoo Kim, Haixum Wang and Alice Oh, 2013
[24]. “Understanding Short Texts Through Semantic Enrichment and Hashing”, Zheng Yu, Haixun Wang, Xuemin Lin and Min Wang, 2015
[25]."Short Text Understanding Through Lexical-Semantic Analysis" Wen Hua, Zhongyuan Wang, Haixun Wang, Kai Zheng, Xiaofang Zhou, 2015.