International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 03 | Mar 2019 www.irjet.net p-ISSN: 2395-0072

Deduplication detection for similarity in document analysis via vector analysis

Mr. P. Sathiyanarayanan [1], Ms. P. Banushree [2], Ms. S. Subashree[3]

[1]Assistant Professor of CSE, [2,3] UG Scholar

Department of Computer Science and Engineering

Manakula Vinayagar Institue of Technology

Puducherrys

Abstract - ***Similarity paraphrase analysis is a machine learning approach in which the system investigate and group the human's opinions, feelings, etc in the form of text or speech about some topic. Nowadays, the textual form of data has great impact among the users. The textual information may be in structured, unstructured or semi-structured form. In accord to improve their products, brands etc., the opinion of the users are rated which leads to the data storage in a huge amount. The analysis of large amount of data is known as big data. This paper intends to survey about the current challenges in the similarity analysis and its scope in the field of real time applications.***

Keywords – Deduplicate , paraphrase, Bigdata, analytics, data duplication.

1. INTRODUCTION

Word information is limited when compared with article information. The information carried by a sentence is between that of a word and an article. Semantics in word level can be easily matched but hard to be recalled as users just use different word to express the same meaning. Semantics in sentence level carries a single topic with its context. Semantics in article level is complex with multiple topics and complicated structures. As a result, the information retrieval among these three levels is one obstacle that impedes the development of natural language understanding.

1.1 DATA MINING

Data mining is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large data sets("big data") involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations-interestingness-metrics, complexity considera-tions, post processing of discovered structures, visualization, and online updating. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics.

For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting is part of the data mining step, but do belong to the overall KDD process as additional steps.

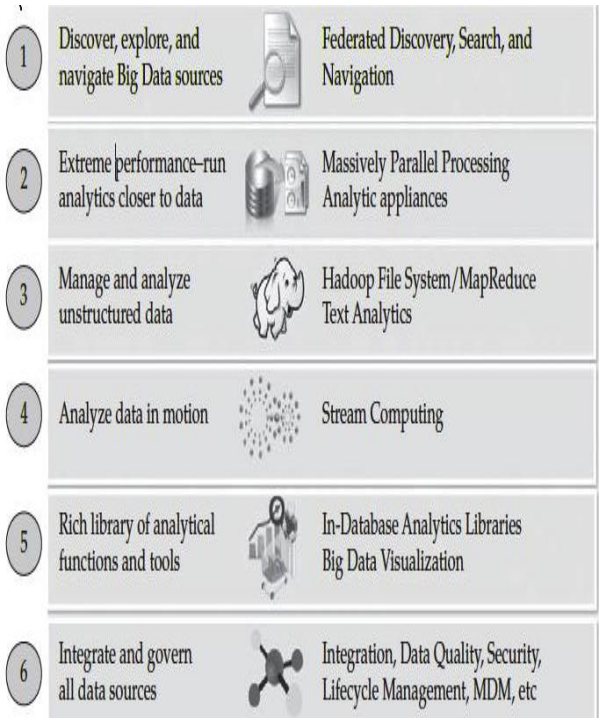
The related terms data dredging, data fishing, and data snooping refer to the use of data mining methods to sample parts of a larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These methods can, however, be used in creating new hypotheses to test against the larger data populations.

Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. This paper presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective.

This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. We analyze the challenging issues in the data-driven model and also in the Big Data revolution.

1.2 BIG DATA

Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools. The challenges include capture, curation, storage, search, sharing, analysis, and visualization. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions. Put another way, big data is the realization of greater business intelligence by storing, processing, and analyzing data that was previously ignored due to the limitations of traditional data management technologies.



1.3 Some concepts

- No sql (not only SQL), Database that "move beyond" relational data models (ie., no tables, limited or no use of SQL).
- Focus on retrieval of data and appending new data (not necessarile tables).
- Focus on key value data stores that can be used to locate data objects.
- Focus on supporting storage of large quantities of unstructured data.
- SQL is not used for storage or retrieval of data.
- No ACID (atomicity, consistency, isolation, durability).

1.4 HADOOP

Hadoop is a distributed file system and data processing engine that is designed to handle extremely high volumes of data in any structure. Hadoop has two components,

- The Hadoop distributed file system (HDFS), which supports data in structured relational form, in unstructured form, and in any form in between
- The Map reduce programming paradigm for managing applications on multiple distributed server.
- The focus is on supporting redundancy, distributed architectures, and parallel processing

1.4.1Some Hadoop Related Names to Know

- **Apache Avro**: designed for communication between Hadoop nodes through data serialization
- **Cassandra and Hbase**: a non-relational database designed for use with Hadoop
- **Hive:** a query language similar to SQL (HiveQL) but ompatible with Hadoop
- **Mahout**: an AI tool designed for machine learning; that is, to assist with filtering data for analysis and exploration
- **Pig Latin**: A data-flow language and execution framework for parallel computation
- **ZooKeeper**: Keeps all the parts coordinated and working together

What to do with the data

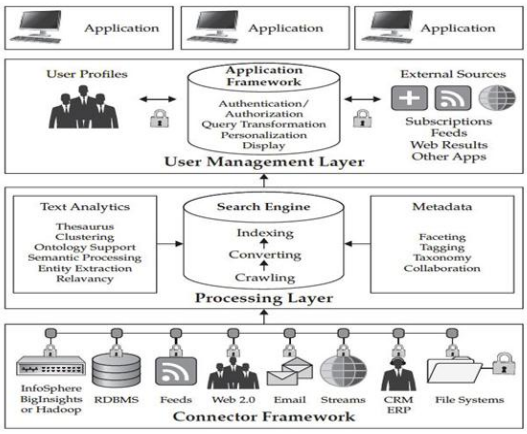


Figure 2 processing layer

The Knowledge Discovery in Databases (KDD) process is commonly defined with the stages:

(1) Selection

(2) Pre-processing

(3) Transformation

(4) Data Mining

(5) Interpretation/Evaluation.

It exists, however, in many variations on this theme, such as the Cross Industry Standard Process for Data Mining (CRISP-DM) which defines six phases:

(1) Business Understanding

(2) Data Understanding

(3) Data Preparation

(4) Modeling

(5) Evaluation

(6) Deployment

or a simplified process such as

(1) pre-processing,

(2) data mining, and

(3) results validation.

2. EXISTING SYSTEM

In the current system, word vector and topic model can help retrieve information semantically. To overcome the above problems, this paper proposes a new vector computation model for text named s2v. Words, sentences, and paragraphs are represented in a unified way in the model.Sentence vectors and paragraph vectors are trained along with word vectors. Based on the unified representation, word and sentence (with different length) retrieval are experimentally studied. The results show that information with similar meaning can be retrieved even if the information is expressed with different words.

3. PROPOSED WORK

The similarity paraphrase analysis is done by extracting the abstract content for comparing the document. Word information is limited when compared with article information. The information carried by a sentence is between that of a word and an article. Semantics in word level can be easily matched but hard to be recalled as users just use different word to express the same meaning. Semantics in sentence level carries a single topic with its context. Semantics in article level is complex with multiple topics and complicated structures. As a result, the information retrieval among these three levels is one obstacle that impedes the development of natural language understanding.

Then separation of words are combined in the form of image by using word cloud net. The frequency of words have been showed in the form of bar graph. By this result, we could determine whether the document is duplication occurs or not.

3.1 COMPLEXITY INVOLED IN THE PROPOSAL

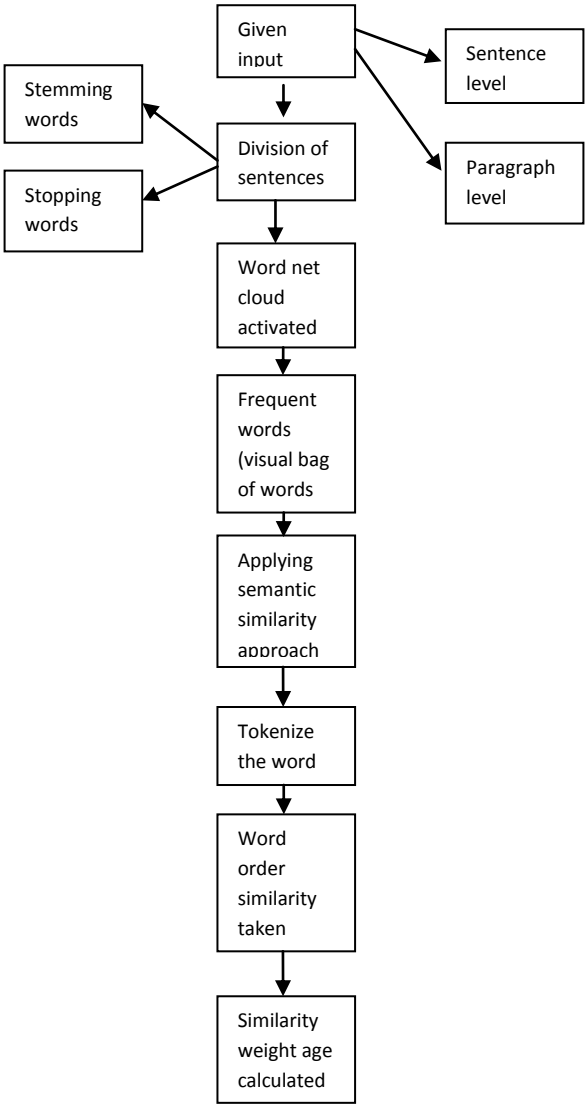
1) Antonyms share high similarity when clustered through word vectors.

2) Vectors for name entities cannot be fully trained, as name entities may appear limited times in specific corpus.

3) Words, sentences, and paragraphs, sharing the same meaning but with no overlapping words, are hard to be recognized.

BLOCK DIAGRAM:

In this block diagram it can identify the given input in the sentences are in pargraph level And then it divides the sentences so that we can identify stemming and stopping words.



3.2 Sentence level

sentences are essentially made up of words, it may be reasonable to argue that simply taking the sum or the average of the constituent word vectors should give a decent sentence representation. This is akin to a bag-of-words representation, and hence suffers from the same limitations, i.e.

- It ignores the order of words in the sentence.
- It ignores the sentence semantics completely.

Other word vector based approaches are also similarly constrained. For instance, a weighted average technique again loses word order within the sentence. To remedy this issue, Socher et al. combined the words in the order given by the parse tree of the sentence. While this technique may be suitable for complete sentences, it does not work for phrases or paragraphs.

3.3 Paragraph level

Paragraph Vectors has been recently proposed as an unsupervised method for learning distributed representations for pieces of texts. In their work, the authors showed that the method can learn an embedding of movie review texts which can be leveraged for sentiment analysis. That proof of concept, while encouraging, was rather narrow. Here we consider tasks other than sentiment analysis, provide a more thorough comparison of Paragraph Vectors to other document modelling algorithms such as Latent Dirichlet Allocation, and evaluate performance of the method as we vary the dimensionality of the learned representation.

We benchmarked the models on two document similarity data sets, one from Wikipedia, one from arXiv. We observe that the Paragraph Vector method performs significantly better than other methods, and propose a simple improvement to enhance embedding quality. Somewhat surprisingly, we also show that much like word embeddings, vector operations on Paragraph Vectors can perform useful semantic results.

3.4 Stemming words

In linguistic morphology and information retrieval, stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root. Algorithms for stemming have been studied in computer science since the 1960s. Many search engines treat words with the same stem as synonyms as a kind of query expansion, a process called conflation.

3.5 Stopping words

A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query.

We would not want these words taking up space in our database, or taking up valuable processing time. For this, we can remove them easily, by storing a list of words that you consider to be stop words. NLTK(Natural Language Toolkit) in python has a list of stopwords stored in 16 different languages. You can find them in the nltk_data directory.

4. CONCLUSION AND FUTURE WORK

In our project New vector computation model was used. Words, sentences, and paragraphs are represented in a unified way in the model. Sentence vectors and paragraph vectors are trained along with word vectors. It shows that information with similar meaning can be retrieved even if the information is expressed with different words. Data Deduplication technology usually identifies the redundant data quickly, which can be used in corporate or in banking sector. The textual information may be in structured or semi-structured form. Whenever user uploads a file in cloud ,System checks the file whether it is existing or not by using vector analysis.

REFERENCES

[1]. (2015) Ben, W. “Every Day Big Data Statistics – 2.5 quintillion bytes of data created daily", Available-http://www.vcloudnews.com/every-day-big data-statistics-2-5-quintillion-bytes-of-data-created-daily/

[2] (2015) Jaspreet, S. "Understanding Data Deduplication", Available:http://www.druva.com/blog/understanding-data-de-duplication/

[3] Y. Zhang and D. Feng and H. Jiang and W. Xia and M.Fu and F. Huang and Y. Zhou. “a fast asymmetri extremum cont-ent defined chunking algorithm for data deduplication in backup storage systems”, IEEE Transactions on Computers, pp. issue: 99, 1-1, 2016.

[4] A. Venish,K. Siva Sankar. “Study of Chunking Algorithm in Data Deduplication,” Proceedings of the International Conference on Soft Computing Systems, Springer India vol. 2, pp 13-20, 2015.