

# Patient-Case Similarity

<sup>1</sup>Vishakha Sharma, <sup>2</sup>Tannya Mittal, <sup>3</sup>Shrey Shukla, <sup>4</sup>Tushar Babbar,  
<sup>5</sup>Ms. Priyanka Bhardwaj

Department Of Computer Science and Engineering

MIET, Meerut

{vishakha.sharma.cs.2016, tannya.mittal.cs.2016, shrey.shukla.cs.2016, tushar.babbar.cs.2016,  
priyanka.bhardwaj}@miet.ac.in

---

**Abstract:** The aim of patient-case similarity is to recognize identical patients on the basis of their medical reports. Recognition of such cases of patients can be used in improving the results of drug proposal for new patients, clinical decision support, prediction of clinical outcome, research on these cases. We have applied machine learning algorithm to accomplish the specified aim with higher accuracy.

**Keywords:** patient-case, word2vec, vectorizing.

---

## 1. INTRODUCTION

In this era of high number of patients and increasing risk of emergency medical condition we are highly in the need of system where emergency situations can be tackled in a better and effective way. When in case the doctor is unavailable the patient may not risk his/her life. Thus, we put forth this system through which we introduce this model. The goal of this approach is to recognize patients who are identical to index patient and derive insights from the records of identical patients to provide personalised predictions. It is to summarize & review and published studies describing computer-based approaches for predicting patients' future health status based on health data & patient similarity, recognize gaps, and provide a starting point for related future research.

The model can be used as backend application for various hospital management systems in order to maintain the patients' data which can be utilized in future. When emergency arises and the doctor in the worst situation is not available to attend the patient at very moment. This application will help to tackle emergency situation of a patient in an efficient way by taking reference from the patients' records from the database.

## 2. LITERATURE REVIEW

The major work in the area of patient-case similarity are:

[<sup>1</sup>] In this model the patient similarity method is applied for recognition of patient suffering from hepatocellular carcinoma who have gone Transarterial Chemoembolization. TACE is preferred only if some conditions are satisfied example no vascular invasion, etc.

[<sup>2</sup>] This model is used to summarize and review system oriented approaches for identifying patient health condition based on their data as similarity status providing a starting point for further research.

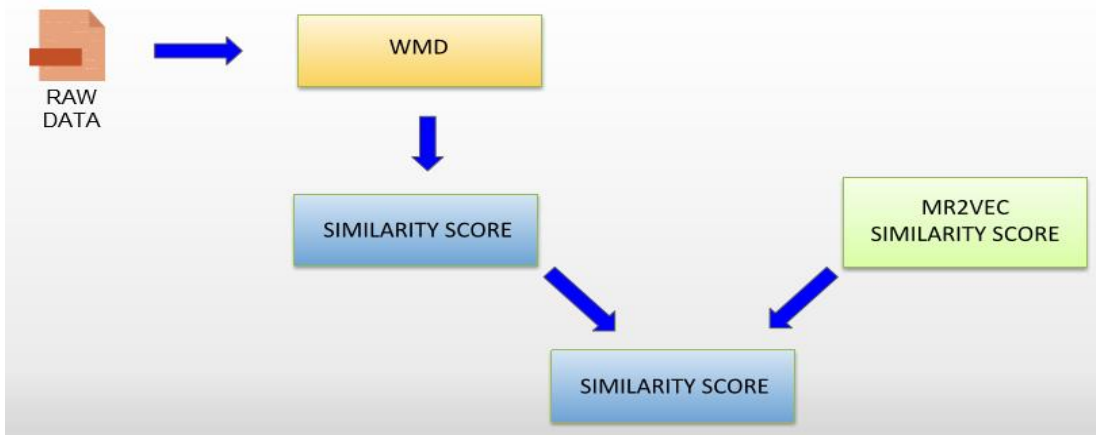
[<sup>3</sup>] This model explains about machine learning algorithms & statistical strategies that have been executed on big data mining. So as to make these algorithms more effective & avoid the problem of over fitting.

Word2vec is a neural network that evaluates the text data. This model helps in understanding that word2vec is not only one algorithm instead includes two more learning models named as CONTINUOUS BAGS OF WORDS (CBOW) & Skip-gram.

[4] This model helps in learning the data mining techniques which is the process of extracting meaningful patterns & models from data sets. Data mining entirely depends on the quality of text, the presence of missing values, partial data & outlier data etc. Therefore, it is of prime importance that data must be processed before mining.

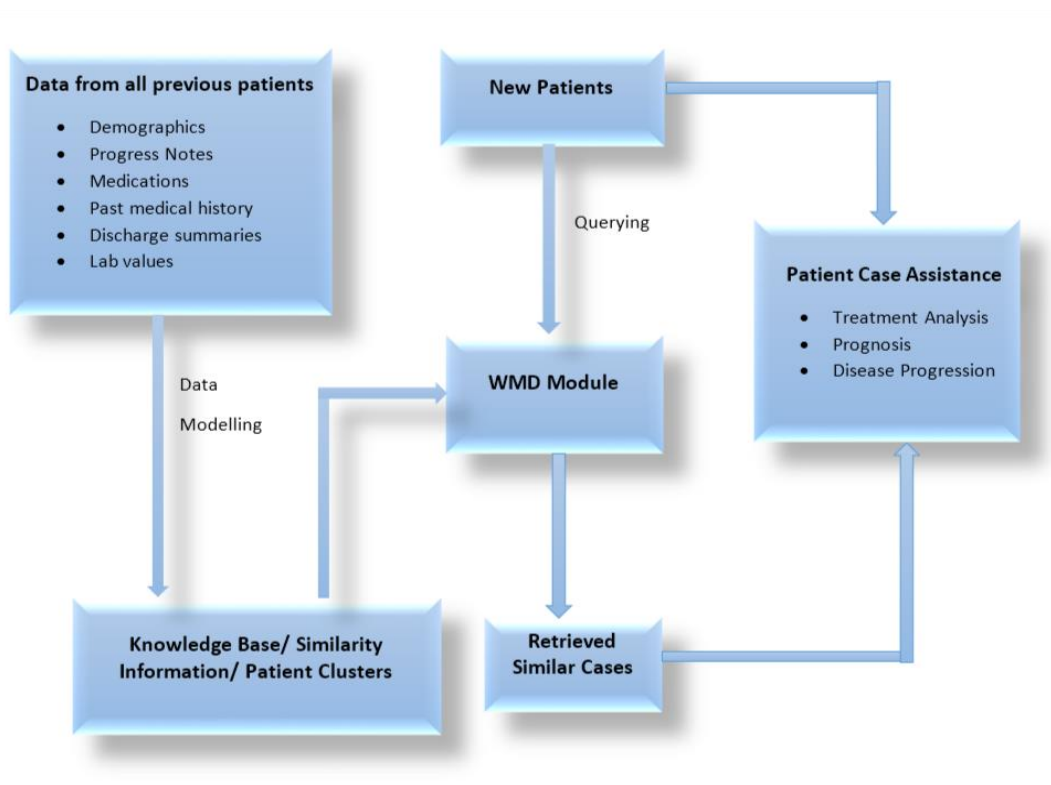
### 3. WORK FLOW OF DESIGN SYSTEM

The goal of this approach is to recognise patients who are identical to an index patient & derive insights from the records of identical patients to provide personalized predictions. It is to summarize & review published studies describing computer-based approaches for predicting patients' future health status based on health data & patient similarity, recognise gaps, & provide a starting point for related future research.



The algorithm used is word2vec which produces distributed depiction of words, and by that we mean word types; i.e. any given word in a vocabulary, such as sit or settle or sat has its own word vector & those vectors are adequately represented in a lookup table or dictionary.

#### 3.1 Module Description



**Fig 2: Data Flow in Patient Case Similarity System**

#### **Data from all previous patients:**

This module comprises of the data taken from patient their past histories old prescription and progress reports. All the medications recommended by the doctors to patients relating to their diseases. It consist of all the information regarding when patient has been admitted and when he has discharged from the hospital. The positive and negative results of the all the tests made are recorded (for eg: MRI ,ECG etc). Here, the main focus revolves around conducting the reviews by performing the automated searches. Hence a database is generated to store all the information regarding patient hospital.

#### **Knowledge base /similarity information /patient cluster:**

This module is used to store complex structured and unstructured information of patient's history. Here the patient clusters are formed where there is an aggregation of cases of diseases and problems or another health related conditions such as cancer, heart stroke, brain hammarage closely grouped in time & space .This module is used to recognize the group of patients with combinations of comorbid conditions .This is basically used to develop targeted management care that is merely needed in future outcomes.

#### **New patient:**

This module deals with the queries which patients posses. In this module, patient brings the queries regarding his health condition in the form of either soft copy or hard copy which is analysed by the hospitals or the administration.

#### **Wmd module:**

WMD stands for "WORDS MOVER'S DISTANCE". It defines that distances & between embedded word vectors are to some level systematically meaningful. It utilizes the above mentioned property of embedding & treats text document as a weighted point cloud. It overcomes basic distance measurement limitation. It is used to compare two pre-processed files where one file is of patients' condition and the other file is stored with the hospital's staff.

#### **Retrieve similar cases:**

This module comprises of the similar clusters which are obtained by the processing of files from the above module. This module helps us in predicting the health status of particular patients.

#### **Patient case assistance:**

This module is used to provide the services on behalf of patients report and meetings. Hence it predicts the:

- Time duration required for the treatment.
- Defines the types of disease & diagnosis required.
- Doctor's recommendation.
- Facilities offered by varied hospitals.

#### **ALGORITHM**

```
F1=open(file1) //description of file 1
F2=open(file2) //description of file 2
sentence 1=f1.read() //reading of file 1
sentence 2=f2.read() //reading of file 2
sentence1.lower().split() //converting sentence 1 and sentence 2 in lower case and splitting
model=Keyed Vectors .Load_word2vec(...) //loading of algorithm functions and vector files
model.init_sims(...) // normalising of vector in word2vec class
distance=model.wmdistance(sen 1,sen 2) //comparing of sentences extracted from files in the form of WMD
print() //printing of similarity index percentage .
```

### 3.2 Description of Dataset

#### MIMIC DATASET:

MIMIC is a relational database carrying tables of data relating to patients. A table is a data storage structure which resemble to a spreadsheet: each column consists consistent information (e.g., patient identifiers), and each row consists an instantiation of that information (e.g. a row could contain the integer 340 in the patient identifier column which would imply that the row's patient identifier is 340).

#### WORD2VEC:

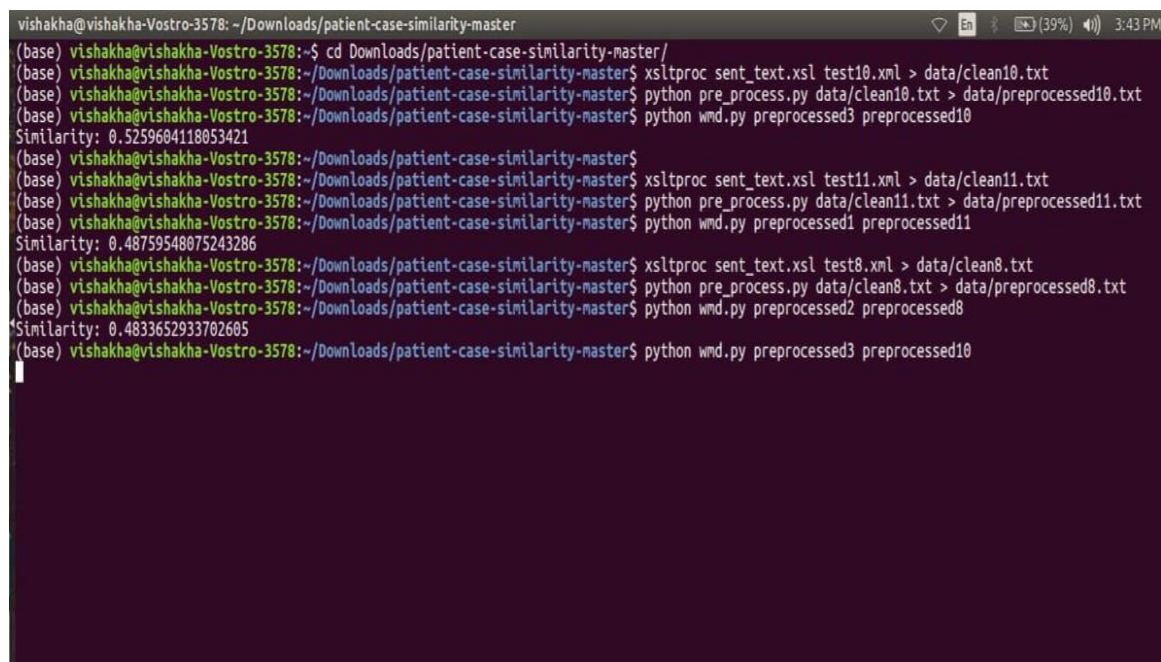
Word2vec is a technique to learn word embeddings with the help of shallow neural web. It processes data by vectorizing words and comprises of input i.e. text corpus and output i.e. set of vectors called as feature vectors which represent words in that corpus. Word2vec is not actually a deep neural web, it changes text into numerical format that deep neural web is able to understand.

Word2vec has a wide range of applications. Its applications extends beyond parsing the sentences. It is applied to those texts in which patterns are recognized easily i.e. genes, codes, social media demographics, symbolic series etc.

Word2vec algorithm is used to combine the vectors of identical words in vector space i.e. It recognizes similarities mathematically. It creates vectors which are numerical depiction of word features. Features like, context of individual words. It does so without any human intercession.

Provided enough data context & usages Word2vec can easily make accurate assumption of word meanings based on past occurrences. Those assumptions are used to maintain word's combination with other words or cluster documents and classify by topic.

## 4. RESULT ANALYSIS



```
vishakha@vishakha-Vostro-3578: ~/Downloads/patient-case-similarity-master
(base) vishakha@vishakha-Vostro-3578:~$ cd Downloads/patient-case-similarity-master/
(base) vishakha@vishakha-Vostro-3578:~/Downloads/patient-case-similarity-master$ xsltproc sent_text.xml test10.xml > data/clean10.txt
(base) vishakha@vishakha-Vostro-3578:~/Downloads/patient-case-similarity-master$ python pre_process.py data/clean10.txt > data/preprocessed10.txt
(base) vishakha@vishakha-Vostro-3578:~/Downloads/patient-case-similarity-master$ python wmd.py preprocessed3 preprocessed10
Similarity: 0.5259604118053421
(base) vishakha@vishakha-Vostro-3578:~/Downloads/patient-case-similarity-master$
(base) vishakha@vishakha-Vostro-3578:~/Downloads/patient-case-similarity-master$ xsltproc sent_text.xml test11.xml > data/clean11.txt
(base) vishakha@vishakha-Vostro-3578:~/Downloads/patient-case-similarity-master$ python pre_process.py data/clean11.txt > data/preprocessed11.txt
(base) vishakha@vishakha-Vostro-3578:~/Downloads/patient-case-similarity-master$ python wmd.py preprocessed1 preprocessed11
Similarity: 0.48759548075243286
(base) vishakha@vishakha-Vostro-3578:~/Downloads/patient-case-similarity-master$ xsltproc sent_text.xml test8.xml > data/clean8.txt
(base) vishakha@vishakha-Vostro-3578:~/Downloads/patient-case-similarity-master$ python pre_process.py data/clean8.txt > data/preprocessed8.txt
(base) vishakha@vishakha-Vostro-3578:~/Downloads/patient-case-similarity-master$ python wmd.py preprocessed2 preprocessed8
Similarity: 0.4833652933702605
(base) vishakha@vishakha-Vostro-3578:~/Downloads/patient-case-similarity-master$ python wmd.py preprocessed3 preprocessed10
```

Patient case similarity uses similarity index to mention the output. On bases of the percentage mentioned it is judged how well the two files are matching on the basis of which correction recommendations are given by the doctors and hospital staff.

Initial command cleans the xml file in which the vectored data is present followed by the command that pre-processes the cleaned text file and wmd command compare the pre-processed text file with the one present in the database of hospital staff and yields out how identical patient's file is with the files existing in database.

Hence on computing the two pre processed file we get the similarity in the form of percentage as 52.59%, 48.75%, 48.33% this is how the patient's case resembles with each other, thus it helps in predicting the types of ailment.

## 5. CONCLUSION

Many patient similarity algorithms have been used & found significant by enhancing clinical efficiency, similar patients' identification, predicting patients' trajectory, providing clinical decision support & avert unanticipated drug reactions.

This analytics create a cheaper portable alternative to affirmation based clinical guidelines & unplanned controlled trials specifically if trial data are unavailable for condition or patient characteristics specific to a query individual. Thus similarity algorithms synthesized with systems machine tools generates workable insights in precision medicine. Patient-case similarity is an indispensable step that allows assortment of patients into clinically meaningful subgroups.

Cloud-based-image-integrated similarity search in big data ;Value System Similarity: Investigation of patient therapist days ; Comparing accuracy of cosine-based similarity & correlation based similarity algorithm in TOURIST RECOMMENDATION SYSTEM ; Distance similarity can be used as CRB technique for early detection of breast cancer ; Suicidal risk detection using a similarity based classifier ; Content based image retrieval in radiology : current status & future directions.

## REFERENCES

- [1] Long Ma and Yanqing Zhang, "Using Word2Vec to Process Big Text Data",IEEE International Conference on October 2015,DOI:10.1109/BigData.2015.7364114(Big Data).
- [2] Anis Sharafoddini, Joel A Dubin and Joon Lee, "Patient Similarity in Prediction Models Based on Health Data", JMIR Med Inform 2017 Jan-Mar, 5(1): e7.
- [3] LWC Chan, T Chan, LF Cheng, WS Mak, "Machine Learning of Patient Similarity",Bioinformatics and Biomedicines Workshops, 2010 IEEE International Conference. .
- [4] Suad A .Alasadi, Wesam S. Bhaya, "Review of data preprocessing techniques in data mining", in Journal of Engineering and Applied Sciences on September 2017.