RESEARCH ARTICLE OPEN ACCESS

Mining Text using Levenshtein Distance in Hierarchical Clusteing

Simranjit Kaur[1], Prof. Kiranjyoti[2]

Guru Nanak Dev Engineering College
Ludhiana, India

----------**********----------

Abstract:-

Intelligent text mining is subject that has caught up the attention of most Business house and Data researchers. In 2013, 5 Exabyte of data is produced on daily basis this data without further analysis and summarization is wasted. Hence researchers has developed many algorithm and systems to record, analyze, filter and summarize the produced data so that important business can be taken effectively, efficiently and in within no time. But small spelling or grammar error found in a textual data can register them as noise and thus losing important piece of information. Hence correcting those mistakes before realization is of paramount significance. But since the number of textual information is humongous, there is a lack of time critical algorithms. Hence this paper presents an algorithm for time effective corrective measure.

Keyword: - **Levenshtein Distance, Hierarchical Clustering, Edit Distance and Text Mining**

----------**********----------

I. INTRODUCTION:-

Mining data for business is a non-trivial process of identifying well founded, potentially useful and understandable patterns in data that is essential for the development and growth of any business. Mining helps corporate house to build business model based on recorded data to predict upcoming marketing trends.

Text mining is the ability to process unstructured text typically in very large documents and interpret their meaning and automatically identify and extract out concepts as well as the relationship between those concepts to directly answer question of interest. There are two popular types of text mining.

Keyword search: typically an individual types in the related keywords and search engine finds all the documents that have the specified keyword in them. All these documents contain the keywords related words then it is up to the individual to read all the document to find the relevant bits. Keyword search is useful and most people use it every day. Keyword search is useful when

1. Speculative browsing for information
2. Finding general information

Limitations of keyword search is that it fails when the set of result from the search is very large. These result set may spawn up to millions of documents. Most people don't have time to read all documents. To reduce the search result down, noisy and irrelevant hits have to be filtered out but manual curation may be time prohibitive and narrowing the question may mean you miss a vital result because same things can be expressed as many different ways and thus person have to type in all and when something can be expressed as many different ways it would be soon become impractical and unwieldy.

Therefore intelligent mining is preferred. Important features that any intelligent mining algorithm should provide:

1) Sentence Identification: Identify that a piece of text is a sentence or not. This alone can be very valuable especially when finding association or connections between things. While using only the standard keyword search there will be chances that the association between the search and the results may be very small. It introduces a lot of noise in the results and an individual has to re-through every document for knowing which one contain anything relevant. If a word occur together in a sentence then there's a much greater likely hood that there is an actual association between the two things,
2) Identify Groups: Secondly it should identifies noun groups, verb groups which portrays the relation between the groups.
3) Morphology: And it should group words into meaningful units. Morphology allows search for different forms of words. The ability to identify different morphological words as same gives much higher recall.

Intelligent mining is really is about helping to make smarter faster decisions and doing it with increased accuracy in a systematic and reproducible way that scale. Intelligent mining is backbone in modern Recommender systems [4].

In this fast moving competitive world there is not time to waste. Time is a useful resource, technology has of course made the lives easier by giving useful information on the click of a button and there are lots of search engines that gives all the necessary information available in the data store. So this paper proposes an algorithm that will make the mining easier, crispier and accurate making sure no important information is lost due to spelling mistake. With spelling mistake arises a lot of chances of vital information being lost due to available algorithm being inefficient to detect incorrect words (words with spelling mistakes) So this algorithm that has been proposed in this paper will eliminate this issue. And will provide an optimal solution for the search.

II. LITERATURE REVIEW:-

Kaur et al. (2011) [2], in their paper gave an overview of data mining. According to them data mining is a process for digging hidden patterns and trends in data that is not immediately apparent from summarizing the data and data mining is applicable to those data where general query does not wield any effectual results. There are several data mining techniques:

a) Neural Network – Neural Network is a structure program that mimics the working of human brain by assigning weightages. There are three stages in a neural network namely, input, hidden and output layer. This neural network are heavily used to find hidden data trend that is hard to find otherwise.
b) Visualization – Expanding the data set in some form of graph, charts and dendrogram because the stake holder can take necessary decisions.
c) Genetic algorithm – The genetic algorithm is an optimization algorithm that is based upon natural selection inspired by nature.
d) OLAP also known as online analytical processing is a database technique that gives n-dimensional view to the data. Hence enabling the collaborator with sufficient knowledge for taking decisions.

Agarwal et al. (2013)[3] reviewed text mining in their paper. They pointed out that text mining is a technique to automatically weed out useful data from the noise. They described a particular technique known as Text Classification Technique and reviewed some common techniques of data mining namely:

a) Artificial Intelligence
b) Decision Trees – Decisions are represented as Tree shape structures
c) Genetic Algorithm

d) Fuzzy C Means – Division of data elements into classes and
e) Modified Algorithm

Nithya et al. (2013)[1] in their paper compares different clustering techniques. At first they described hierarchical clustering a clustering technique where different element are grouped together to form clusters. Advantage of using this type of clustering technique is that it enable us to form specified number of clusters. Next they described about Partition methods in which database of N objects are clustered into k partitions. Thereafter they discussed density based clustering technique that is based on discovering the clusters of arbitrary shapes. At last they reviewed KNN also called K-nearest neighbor. KNN is a lazy learning algorithm rooted in Euclidean distance as a distance measure. And is further optimized by Particle swarm optimization technique.Gupta et al. (2012) [7] in their paper discussed several spell checking techniques and some of them are:

f) Similarity Keys: An Index key is associated with every word in the dictionary and these keys are searched against spelling errors
g) Rule Based Techniques – Predefined Rules that cures the any common spelling errors
h) N-Gram Based Techniques – N letter before and after the errors are checked
i) Neural Network – An unsupervised learning technique that train itself by adjusting its weight to correct spelling mistakes.
j) Probabilistic Technique – Sets the probability for the occurrence of each letter in the spelling.

III. METHODOLOGY:-

A queried result set spawned through text mining can yield hundreds of documents containing several thousand texts consequently it will be very inflexible and rocky to search for the desired text from the result set. And the techniques mentioned in the literatures not applicable to the result set having humongous texts. Thus in this paper another approach of Hierarchical clustering is taken. The working procedure that this paper has adopted to address this issue is:

1) Collect text based result set
2) Select a distance function
3) Perform Hierarchal clustering

4) Collect Texts – For this experiment huge amount of text has been scrapped from Twitter, Facebook, Wikipedia and other popular sites. Further this text is tokenized by removing non-printable characters and then it is stored in a CSV file.
5) Distance Function – Levenshtein Distance is used as a distance measure. In Levenshtein Distance we compute minimum distance between any two given word in a tabular fashion. In computing Levenshtein distance there are three operations. Namely, Insertion, Deletion and Substitution. Only one character can be either inserted, replaced or substituted at any given time. Cost for each operation is one unit. No other operation is permitted. In various paper transpose is also considered. But for the sake of experiment simplicity transposition is not taken into account. For demonstration we wish to convert the word “RACE” into “PAIR” using levenshtein distance. Initial setup, the word which need to be transformed are written horizontally and the resulting word is written vertically. And initially a table of 5x5 is constructured and then the table is populated. The conversion between empty strings to another is zero. Hence it is marked as zero in the table. And distance between converting a letter from empty string is 1 hence ‘R’ and ‘P’ is marked 1 on their first cell. Subsequently ‘A’, ‘C’ and ‘E’ comprises 2nd, 3rd and 4th letters of the word hence it is marked as 2, 3 and 4 respectively. Similarly the first column next to ‘A’, ‘I’ and ‘R’ is also marked 2, 3 and 4. And the rest of the table

is filled with blank. And this blank will be calculated next and filled by appropriate number.

		R	A	C	E
	0	1	2	3	4
P	1	-	-	-	-
A	2	-	-	-	-
I	3	-	-	-	-
R	4	-	-	-	-

Figure 1: Initial Setup

Then if column [letter] = Row [letter] then borrow the number present in the diagonal else taken the number minimum from its adjacent cells and add 1 to it. The resulting table will appear like this:

		R	A	C	E
	0	1	2	3	4
P	1	1	2	3	4
A	2	2	1	2	3
I	3	3	2	2	3
R	4	4	3	3	3

Figure 2: Final Result

The last cell of the table gives the Levenshtein distance required to convert a given text into other.

Hierarchical Clustering [5] [6] – Based on the Levenshtein distance required number of clusters are formed. And mathematically represented as:

Figure 3: Forming Clusters

$$\forall x \in C \Rightarrow x_i \in C_m \wedge m \in \{ i : i \pm \partial j, j \in Q \}$$
$$where\ C_m\ is\ Equivalence\ class$$
and Q is small rational number

The complete algorithm described above is represented mathematically below:

$$Collecting\ and\ tuning\ data$$
$$S_i \in DS_k \wedge s_j \subset S \Rightarrow s_j \cap BS = \emptyset$$

$$Initialization$$
$$M(i,0) = i$$
$$M(0,j) = j$$

$$Recurrence\ Relation$$
$$\forall\ x \in s_j$$
$$\forall\ y \in s_k where, k \neq i$$
$$x = y \Rightarrow M(i,j) = M(i-1, j-1)$$
$$x \neq y \Rightarrow M(i,j) = min \begin{Bmatrix} M(i-1,j-1) \\ M(i-1,j) \\ M(i,j-1) \end{Bmatrix} + 1$$
$$M(I,J)\ is\ the\ Levensthein\ distance$$

$$Clustering$$
$$Distance = \{ D_{i,j} : Dist(s_i \wedge s_j \forall i, j = 1 \ldots max(I,J)) \}$$
$$\forall\ max(I,J)\ iterations:$$
$$i, j = arg\,min\, D_{i,j}$$
$$where, i\ and\ j\ are\ pair\ of\ closest\ clusters$$
$$Join(C_i, C_j)$$
$$Delete(C_i, C_j)$$
$$\forall\ C_{k \neq i} \in C \Rightarrow D_{k,i+j} = min\{ D_{k,i}, D_{k,j} \}$$

Figure 4: Complete Algorithm

IV. DEMONSTRATION:-

The experimental text is retrieved from Wikipedia and some errors has been introduced such that an experiment of the proposed algorithm can be performed. The sentence is first partitioned into words and removing any punctuation or non-printable characters and assigning each word to a row. The initial experimental setup is shown in the image below:

A
computer
os
a
general
purpose
device
that
can
be
programmed
to
carry
out
a
set

Figure 5: Incorrect Sentence

Cost Estimation using Levenshtein distance

		o	s
	0	1	2
i	1	1	2
s	2	2	1

Figure 6: Cost Estimation

Hierarchical Cluster
H={{},1},{{},2},{{},3},{{},4} }before correction.
And after correction:
{{},1},{{(os,is)},2},{{},3},{{},4}.
In this manner complete Cluster set is populated and tested against result set.

V. RESULT AND CONCLUSIONS:-

Accuracy when text mining is performed in the data store is shown in the image below:

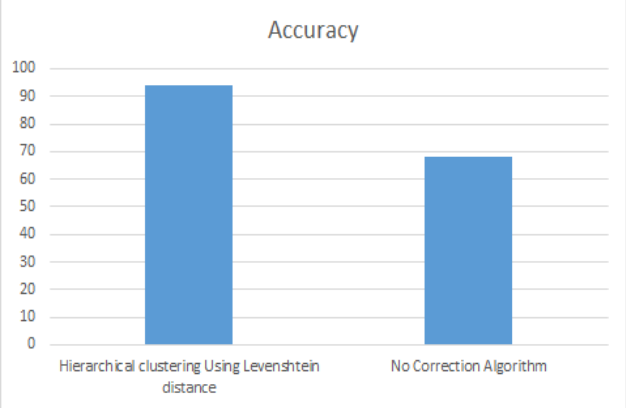

Figure 7: Accuracy

Once the clusters are thickly populated additional 0.01 second is required for a search from a database having more than 10,000 texts than the algorithm with no correction algorithm. The time taken is represented as a graph below:

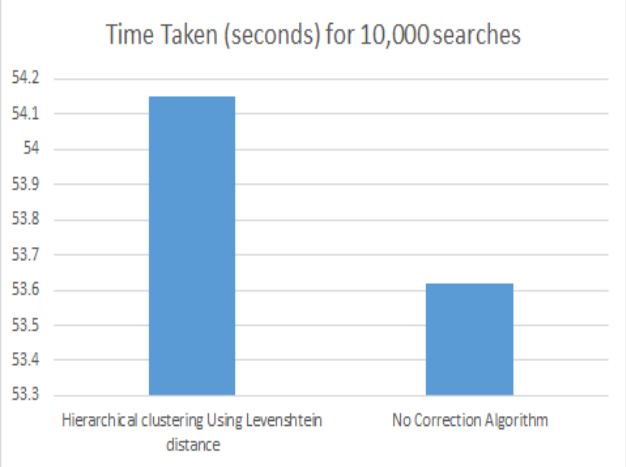

Figure 8: Time Taken

VI. FUTURE SCOPE:-

The Levensthein distance used in this algorithm only considers insertion, deletion and substitution. Leaving scope for transposition for further research and various optimization techniques such as PSO and BSO can be applied on the proposed algorithm to reduce time penalty introduced by applying Levensthein clusters.

ACKNOWLEDGMENT:-

I warmly put my sincere thanks to Prof. Kiran Jyoti, Prof. Akshay Girdhar and others who gave their precious time and support.

And special thanks to Prof. Raninder Dhillon for providing experimental data set.

REFERENCES:-

1. Diagnosis, A Survey on Clustering Techniques in Medical. N.S.Nithya, Dr.K.Duraiswamy and P.Gomathy. 2, Tamil Nadu : International Journal of Computer Science Trends and Technology, December 2013, Vol. 1. 2347-8578.

2. Data Mining: An Overview. Singh, Gurjit Kaur and Lolita. 2, Ludhiana : International Journal of Computer Science and Techonology, June 2011, Vol. 2. 0976-8491.

3. Review Paper on Punjabi Text Mining Techniques. Singla, Shruti Aggarwal and Salloni. 2, Fatehgarh Sahib : International Journal of Computer Science and Technology, June 2013, IJCST, Vol. 4. 2229-4333.

4. An Efficient Recommender System using Hierarchical Clustering Algorithm. Chalotra, Prabhat Kumar and Sherry. Ludhiana : International Journal of Computer Science Trends and Technology, August 2014, IJCST. 2347-8578.

5. Efficiently Measuring Similarities Between Objects in Different Views of Hierarchical Clustering. Mahammad Nazini, MD Roshna and Shaik. Jakeer Hussain. 2, Hayathnagar : s.n., June 2013, Vol. 4. 0976-8491.

6. Hierarchical Clustering Based Activity Tracking System in a Smart Environment. Rajkumari, Amitha. R and K. 3, Coimbatore : Intenational Journal of Comptuer Science and technology, September 2012, IJCST, Vol. 3. 2229-4333.

7. Spell Checking Techniques in NLP: A Survey. Mathur, Neha Gupta and Prathistha. s.l. : International Journal of Advanced Research in Computer Science and Software Engineering, December 2012. 2277 128X.