A SURVEY ON SIMILARITY MEASURES IN TEXT MINING

M.K.Vijaymeena[1] and K.Kavitha[2]

[1]M.E. Scholar, Department of Computer Science & Engineering, Nandha Engineering College, Erode-638052, Tamil Nadu, India
[2] Assistant Professor, Department of Computer Science & Engineering, Nandha Engineering College, Erode-638052, Tamil Nadu, India

ABSTRACT

The Volume of text resources have been increasing in digital libraries and internet. Organizing these text documents has become a practical need. For organizing great number of objects into small or minimum number of coherent groups automatically, Clustering technique is used. These documents are widely used for information retrieval and Natural Language processing tasks. Different Clustering algorithms require a metric for quantifying how dissimilar two given documents are. This difference is often measured by similarity measure such as Euclidean distance, Cosine similarity etc. The similarity measure process in text mining can be used to identify the suitable clustering algorithm for a specific problem. This survey discusses the existing works on text similarity by partitioning them into three significant approaches; String-based, Knowledge based and Corpus-based similarities.

Keywords

Similarity measure, Corpus, Co-occurrence, Semantics, Lexical analysis, Synsets.

1. INTRODUCTION

Document clustering is a process that involves a computational burden in measuring the similarity between document pairs. Similarity measure is the function which assigns a real number between 0 and 1 to the documents. A zero value means that the documents are dissimilar completely whereas one indicates that the documents are identical practically. Vector-based models have been used for computing the similarity in document, traditionally. The Different features presented in the documents are represented by vector- based models.

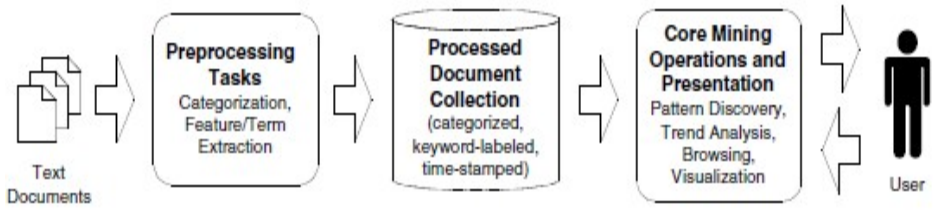

Figure 1: High-level text mining functional architecture.

Text similarity measures play an increasingly vital role in text related research and applications in several tasks such as text classification, information retrieval, topic tracking, document clustering, questions generation, question answering, short answer scoring, machine translation, essay scoring, text summarization, topic detection and others.

DOI : 10.5121/mlaij.2016.3103

Finding similarity between words is a necessary part of text similarity. It is the primary stage for sentence and document similarities. Figure 1 represents High-level text mining functional architecture. Words can be possibly similar in two ways lexically and semantically. Words are lexically similar if they have a similar character sequence. If the words have same theme, then they are semantically similar and if they have dissimilar theme but used in same context, then they are not related to each other. String-Based algorithms are used to measure Lexical similarity; Corpus-Based and Knowledge-Based algorithms are based on Semantic similarity.

String-Based measures operate on sequence of strings and composition of characters. For measuring the similarity and dissimilarity between the text strings, String metric is used. It is used for string matching or comparison but it is approximate. The semantic similarity determines the similarity between words based on the information gained from large corpora is called as Corpus-Based similarity. Knowledge-Based similarity is a semantic similarity measure which is used to determine the degree of similarity between words based on information derived from semantic networks. Each type will be described briefly.

This paper is arranged as follows: Section 2 presents String-Based algorithms by divided them into character-based and term-based measures. Sections 3 and 4 introduce Corpus-Based and knowledge-Based algorithms respectively. Concept of advanced text mining system is introduced in section 5 and section 6 concludes the survey.

2. STRING-BASED SIMILARITY

String similarity measures are operated on string sequences and composition of characters. A metric that is used for measuring the distance between the text strings is called String metric and used for string matching and comparison. This survey represents various string similarity measures and they were implemented in SimMetrics package [1]. Various algorithms are introduced in which seven are character based while the other is term-based distance measures and they are discussed in the following section.

2.1. Character-based Similarity Measures

The similarity between the strings is based on the contiguous chain of characters length which are present in both strings. This is known as Longest Common Substring algorithm (LCS) measures.

- The distance between two strings is measured by counting process. The minimum number of operations required transforming one string into another string and the operations such as insertion, deletion, or substitution of a single character and transposition of two characters which are adjacent is defined by Damerau-Levenshtein [2, 3]. The Mathematical computation of Levenshtein distance between two strings a,b where

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j) + 1 \\ \mathrm{lev}_{a,b}(i,j-1) + 1 \\ \mathrm{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

  where $1_{(ai \neq bj)}$ is the indicator function and it is equal to 0 if $a_i = b_j$,equal to 1 otherwise, and $\mathrm{lev}_{a,b}(i,j)$ is the distance between the first i characters in a and the first j characters in b.
- Jaro distance measure is based on the number or order of the characters between two strings which are common; It is mainly used for record linkage [4, 5]. Jaro–Winkler is an extension

of Jaro distance; a prefix scale gives more ratings related to strings and it matches from the beginning for a set prefix length [6]. The Jaro distance d_j of two given strings S_1 and S_2 is

$$d_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases}$$

Where m is the number that denotes the matching characters and t denotes half the number of transpositions.

- The algorithm which follows dynamic programming is Needleman-Wunsch algorithm [7]. It was one of the applications of dynamic programming for biological sequence comparison. It also performs a global alignment in order to find the best alignment over the entire of two sequences and suitable only when the two sequences are of similar length.
- Smith-Waterman [8] is an example of dynamic programming. For finding the best alignment over the conserved domain of sequences, local alignment is performed. It is very useful for dissimilar sequences with similar regions. A matrix H is built as follows:

$$H(i,0) = 0,\ 0 \le i \le m$$
$$H(0,j) = 0,\ 0 \le j \le n$$
$$H(i,j) = \max \left\{ \begin{array}{cc} 0 & \\ H(i-1,j-1) + s(a_i,b_j) & \text{Match/Mismatch} \\ \max_{k\ge 1}\{H(i-k,j) + W_k\} & \text{Deletion} \\ \max_{l\ge 1}\{H(i,j-l) + W_l\} & \text{Insertion} \end{array} \right\},\ 1 \le i \le m, 1 \le j \le n$$

Where a,b = Strings (Alphabet Σ), m is equal to length(a), n is equal to length(b), S(a,b) is a similarity function. The maximum Similarity-Score between a suffix of a[1...i] and a suffix of b[1...j] is denoted by H(i,j),W_i is the gap-scoring scheme.

- N-gram [9] is a sub-sequence of n items in a given text sequence. N-gram similarity algorithms are used to compare the n-grams from each character in two strings. Distance is computed to divide the number of similar n-grams by maximal number of n-grams. For statistical natural language processing, N-gram models are used widely. Phonemes and its sequences are modelled using a n-gram distribution in speech recognition. For parsing, words are modelled such that each n-gram is composed of n words and used for language identification.

2.2. Term-based Similarity Measures

- The Block Distance [10] is known as Manhattan distance. It is used to compute the distance that would be travelled to get from one data point to another only if a grid-like path is followed. The Block distance between the items is the sum of the differences of their components correspondingly. In *n*-dimensional real vector space with fixed Cartesian coordinate system, the taxicab distance (d_1) between vectors p,q is the sum of the lengths of the line segment projections between the points onto the coordinate axes.

$$d_1(\mathbf{p},\mathbf{q}) = \|\mathbf{p}-\mathbf{q}\|_1 = \sum_{i=1}^{n} |p_i - q_i|,$$

where (p,q) are vectors.

- The Similarity measure which is primarily used to measure the cosine angle between the documents is known as Cosine Similarity.

$$S_{Cos}(\mathbf{d}_1, \mathbf{d}_2) = \frac{\mathbf{d}_1 \cdot \mathbf{d}_2}{(\mathbf{d}_1 \cdot \mathbf{d}_1)^{1/2}(\mathbf{d}_2 \cdot \mathbf{d}_2)^{1/2}}.$$

- Dice's coefficient [11] is defined as two times the number of terms which are common in the compared strings and divided by the total number of terms present in both strings.

$$S_{Dic}(\mathbf{d}_1, \mathbf{d}_2) = \frac{2\mathbf{d}_1 \cdot \mathbf{d}_2}{\mathbf{d}_1 \cdot \mathbf{d}_1 + \mathbf{d}_2 \cdot \mathbf{d}_2}.$$

- Euclidean distance is measured by calculating the square root of the sum of squared differences between the two vectors elements.

$$d_{Euc}(\mathbf{d}_1, \mathbf{d}_2) = [(\mathbf{d}_1 - \mathbf{d}_2)\cdot(\mathbf{d}_1 - \mathbf{d}_2)]^{1/2}$$

- Jaccard similarity [12] is computed in such a way that the number of shared terms divided by the number of all unique terms present in both strings.

$$S_{EJ}(\mathbf{d}_1, \mathbf{d}_2) = \frac{\mathbf{d}_1 \cdot \mathbf{d}_2}{\mathbf{d}_1 \cdot \mathbf{d}_1 + \mathbf{d}_2 \cdot \mathbf{d}_2 - \mathbf{d}_1 \cdot \mathbf{d}_2}$$

- A simple vector based approach is named as Matching Coefficient. It simply counts the number of similar terms and both vectors should be non zero. The Simple Matching Coefficient (SMC) is a statistics mainly used to compare the similarity of the sample sets and its diversity. Consider two objects named as A and B and each have n binary attributes.SMC is defined as:

$$SMC = \frac{\text{Number of Matching Attributes}}{\text{Number of Attributes}} = \frac{M_{00} + M_{11}}{M_{00} + M_{01} + M_{10} + M_{11}}$$

  Where:
  The total number of attributes is represented as M_{11} and *A*, *B* both are equal to 1.
  The total number of attributes is represented as M_{01} and where attribute *A* is 0 and attribute *B* is 1.
  The total number of attributes is represented as M_{10} where attribute *A* is 1 and attribute *B* is 0.
  The total number of attributes is represented as M_{00} where *A* and *B* values are equal to 0.

- Overlap coefficient considers two strings a full match if one is a subset of another and it is similar to Dice Coefficient. The overlap coefficient also called as Szymkiewicz-Simpson coefficient is a similarity measure that is related to the Jaccard index. It measures the overlap between two sets. The measure is calculated by dividing the size of the intersection by the smaller of the size of the two sets:

$$\text{overlap}(X, Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)}$$

  If set *X* is a subset of *Y* or Vice-versa, it shows that the overlap coefficient value is 1.

3. CORPUS-BASED SIMILARITY

Corpus-Based similarity is a semantic similarity measure which determines the similarity between the words based on the information gained from corpora. Figure 2 shows the Corpus-Based similarity measures. A Corpus is a large collection of texts and it is used for language research.
Hyperspace Analogue to Language (HAL) creates a space semantically and the strength of association between the word are represented by the row and the words are represented by the column [13,14]. The user of the algorithm could drop out the low entropy columns in the matrix. In the beginning of the window, a focus word is placed if the text is analyzed. The neighbouring words are recorded while co-occurring. Word-ordering information is recorded by treating the co-occurrence based on whether the neighbouring word appeared before and after the focus word.
Latent Semantic Analysis (LSA) is a important technique of Corpus-Based similarity. It assumes that the words with close meaning will mostly occur in the similar group. A matrix which contains word counts per paragraph was constructed from a text dataset. To reduce the number of columns, singular value decomposition (SVD) is used and it preserves the similarity structure among the rows.

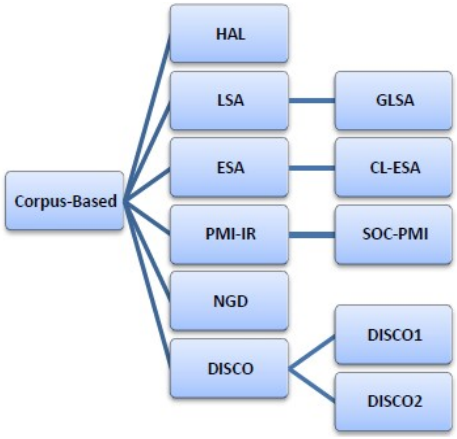

Figure 2: Corpus-Based Similarity Measures

Generalized Latent Semantic Analysis (GLSA) is a framework for computing semantically motivated term. The LSA approach is extended by focusing on term vectors instead of the representation of dual document-term. A measure used for computing the semantic relatedness of arbitrary texts is called as Explicit Semantic Analysis (ESA). The cosine measures between the vectors are expressed based on the semantic relatedness between two texts.

4. KNOWLEDGE-BASED SIMILARITY

Knowledge-Based Similarity is a semantic similarity measures and it based on identifying the degree of similarity between words and it uses various information derived from semantic networks [15]. The most popular semantic network is known as WordNet which is a big lexical database of English language. Nouns, verbs, adverbs and adjectives are grouped into sets of cognitive synonyms known as synsets. By means of conceptual-semantic and lexical relations, Synsets are interlinked. Knowledge-based similarity measures can be divided into two groups such as measures of semantic similarity and semantic relatedness. Measures of semantic similarity have been defined between words and concepts. The prominence on word-to-word similarity metrics is due to the resources availability and specifically encode relations between words and concepts (e.g. WordNet), and the various test beds that allow for their evaluation (e.g. TOEFL or SAT analogy/synonymy tests). Consider that the input text segments are T1 and T2. The similarity between the text segments could be determined using the below scoring function:

$$sim(T_1, T_2) = \frac{1}{2}\left(\frac{\sum_{w\in\{T_1\}} (maxSim(w,T_2)*idf(w))}{\sum_{w\in\{T_1\}} idf(w)} + \frac{\sum_{w\in\{T_2\}} (maxSim(w,T_1)*idf(w))}{\sum_{w\in\{T_2\}} idf(w)}\right)$$

This similarity score value should be between 0 and 1. The score of 1 indicates identical text segments and score of 0 indicates that there is no semantic overlap between the two segments. The following word similarity metrics: Lesk, Leacock & Chodorow, Wu & Palmer, Lin, Resnik, and Jiang & Conrath are word-to-word similarity metrics which selects any given pair of words with highest concept-to-concept similarity.

- The Leacock & Chodorow (1998) similarity is determined as follows:

$$Sim_{lch} = -\log \frac{length}{2 * D}$$

Here, length is the shortest path length between two concepts which uses node-counting and D is the maximum taxonomy depth.

- The Lesk similarity between two concepts is a function of the overlap between the definitions correspondingly and it is provided by a dictionary. Based on an algorithm proposed by Lesk (1986), it is the solution for word sense disambiguation process. The application of the Lesk similarity measure is used in semantic networks and in conjunction with any dictionary but it should provide definitions for words.
- The Wu and Palmer (1994) similarity metric is used to measure the depth of the given concepts in the Word Net taxonomy, the least common subsumer (LCS) depth and combines these figures into a similarity score:

$$Sim_{wup} = \frac{2 * depth(LCS)}{depth(concept_1) + depth(concept_2)}$$

- The measure introduced by Resnik (1995) is used to return the information content (IC) of the least common subsumer of two concepts:

$$Sim_{res} = IC(LCS)$$

where IC is defined as:

$$IC(c) = -\log P(c)$$

and P(c) = Probability of encountering an instance of concept c in a big corpus. Lin (1998) metric was builded on Resnik's measure which adds a normalization factor consisting of the information content of the two input concepts as follows:

$$Sim_{lin} = \frac{2 * IC(LCS)}{IC(concept_1) + IC(concept_2)}$$

5. ADVANCED TEXT MINING SYSTEM

The general model provided by various data mining applications is followed by the Text mining system. The main areas are pre-processing tasks, core mining tasks, components of presentation layer and browsing functionality, and refinement techniques. Figure 3 shows the System architecture for an advanced text mining system.

- Pre-processing is a technique that prepares data for knowledge discovery operations and removes the noise present in the data. Generally, Pre-processing tasks convert the information from each original source into a canonical format.
- Core Mining Operations are considered as the heart of a text mining system. It includes trend analysis, pattern discovery and incremental knowledge discovery algorithms.
- Presentation Layer Components consists of GUI and pattern browsing functionality. Visualization tools and user-facing query editors also fall under this architectural category. Presentation layer components include graphical tools for creating and modifying concept clusters and it creates annotated profiles for various specific concepts.
- Refinement Techniques are methods used to filter redundant information and it is used for clustering closely related data. It represents a full, comprehensive suite of pruning, ordering, generalization, and clustering approaches.

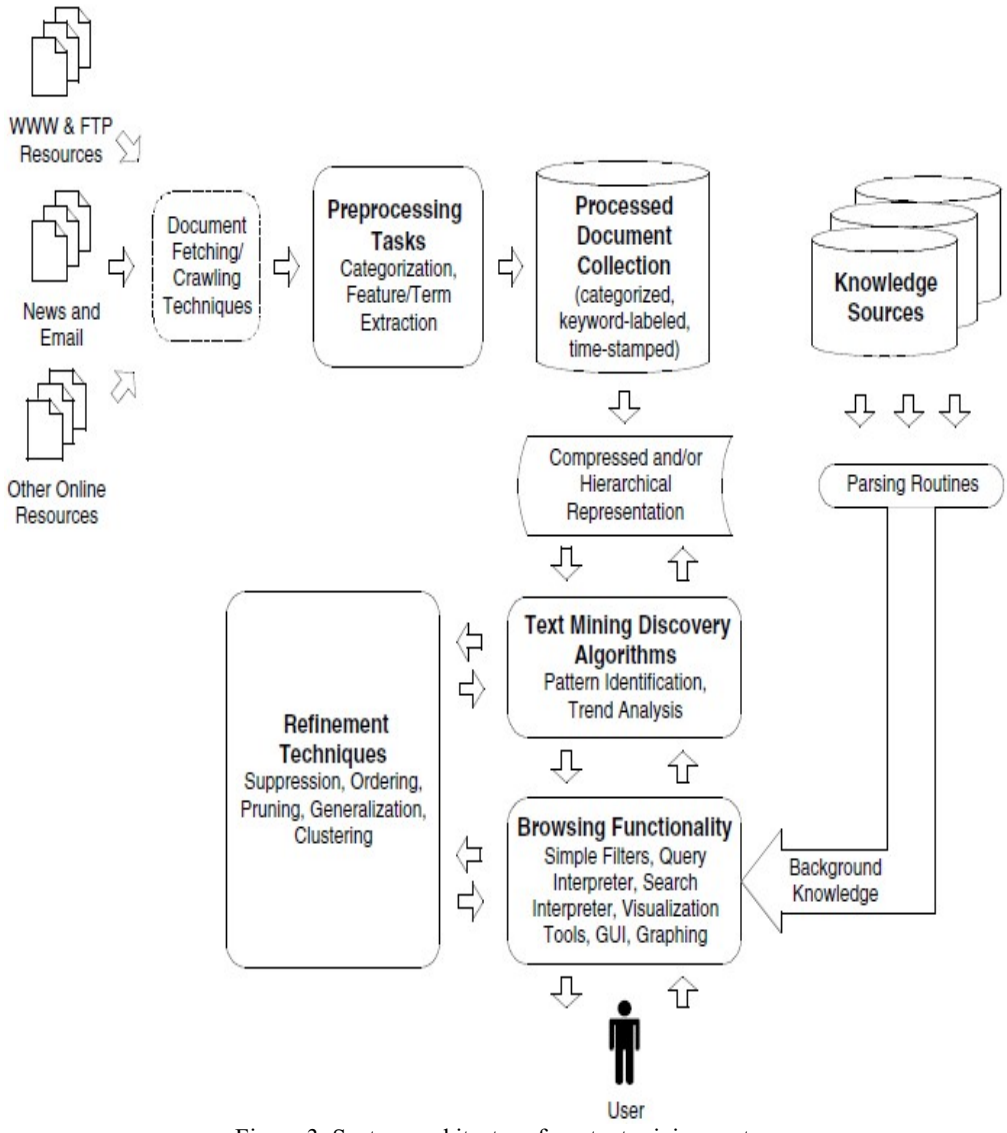

Figure 3: System architecture for a text mining system.

6. APPLICATIONS OF TEXT MINING

Similarity measures in Text mining are increasing function of common features and decreasing function of dissimilar features. The Similarity measures will be applied in various fields and for different purposes in Text mining domain. Unstructured text mining is an area which is gaining significant importance in adoptions for business applications. Text mining is being applied to answer various business questions and to optimize day-to-day operational efficiencies and used for improving long-term strategic decisions. Following are the practical real-world instances where text mining has been successfully applied.

6.1. Text Mining in the Automotive Industry

It is essential that auto companies must explore all the ways for reducing costs. One of the very important levers in the cost equation is Optimizing warranty cost for automobile manufacturers. One of the underutilized aspects of optimizing warranty cost is input from service technician's comments. Automobile companies can take following actions which will optimize dealer inventory for spare parts, reduce warranty-related cost erosion and help suppliers deliver quality components:

- Auto manufacturers can share the results of technician's comments with specific product suppliers. It is required to undertake joint initiatives to reduce the number of defective components.
- Automobile companies can build an early warning system and this could happen by considering the frequency of occurrence of keywords. They are specific in a watch list like "brake failure," "short circuit,","low mileage" etc.,. And this may cause legal liability in few cases.
- If the component was manufactured internally, then the specific manufacturing process that was responsible for the defective component can be re-engineered and it eliminates reoccurrence.
- By calculating the select spare parts or auto component's frequency of occurrence, it can be used as an input to forecast the regional needs for spare parts. This is known as auto spare part inventory optimization.
- Most dealer management systems have a preliminary taxonomy for classifying defects. This taxonomy changes depending upon how well it was defined originally. If most defects get classified under a miscellaneous section, then keyword and theme frequency analysis can guide in the creation of new defect classifications and used for text mining.

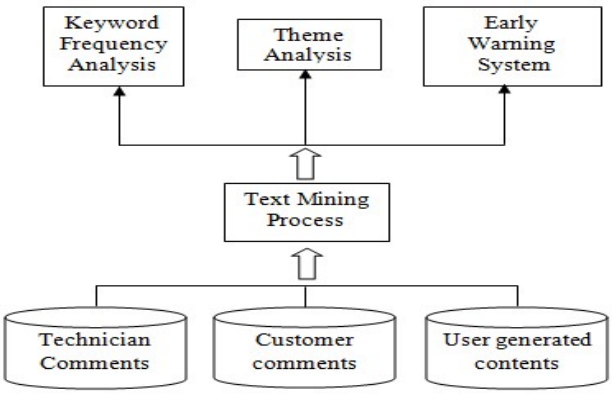

Figure 4: Text mining in Auto Industry.

6.2. Text Mining in the Healthcare Industry

The healthcare industry is a huge spender on technology with the proliferation of hospital management systems and to log patient statistics, low-cost devices are used. Since there is an increase in the breadth and depth of patient data suddenly, it is required to mine the comments in doctor's diagnosis transcripts. Outputs of the mining process can yield information which benefits the healthcare industry in following numerous ways.

- The top ten diseases are isolated by frequencies of keyword per region. The mix of tablets/medicines to stock is leveraged to optimize on the limited outlets store. It should be noted that the frequency of disease related keywords will change.
- Based on doctor's comments, an early warning system can be built within text mining outputs in order to detect sudden changes. For example, if the keyword frequency of lungs or breathing exceeds 60 appearances, that too specifically in the last month for a given region, it can be a hint to excessive environmental conditions that results in respiratory problems. To remedy the situation, a proactive intervention can be activated.

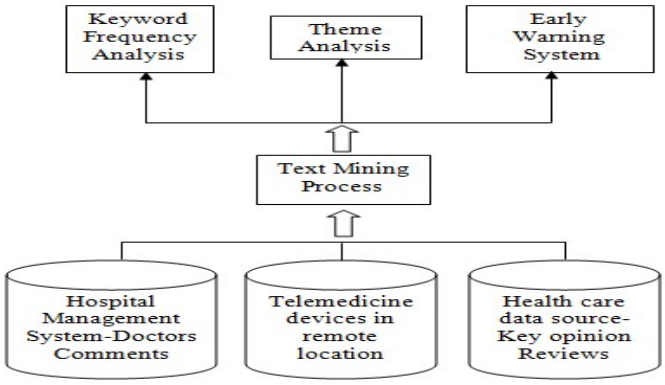

Figure 5: Text mining in Healthcare Industry.

7. CONCLUSION

Text Mining is a significant research area which is gaining an increasing popularity in the recent years. Measuring the similarity between the text documents is an important operation of text mining. In this survey, three text similarity approaches such as String-based, Corpus-based and Knowledge-based similarities are discussed. String-Based measures are operated on character composition and string sequences. Corpus-Based similarity is a semantic similarity measure which determines the similarity between words based on the information gained from large corpus. A semantic similarity measures known as Knowledge-Based similarity is based on the degree of similarity between the words and concepts. Some of these algorithms were combined together in many researches and they are hybrid similarity measures. Useful similarity packages such as SimMetrics, WordNet Similarity and NLTK were mentioned. The System architecture for an advanced text mining system and the functionalities of various components are described. Different real time applications of text mining used in environments such as Automobile industry and Healthcare industry have been discussed.

REFERENCES

[1] Chapman, S. (2006). SimMetrics : a java & c# .net library of similarity metrics, http://sourceforge.net/projects/simmetrics/.

[2] Hall , P. A. V. & Dowling, G. R. (1980) Approximate string matching, Comput. Surveys, 12:381-402.

[3] Peterson, J. L. (1980). Computer programs for detecting and correcting spelling errors, Comm. Assoc. Comput. Mach., 23:676-687.

[4] Jaro, M. A. (1989). Advances in record linkage methodology as applied to the 1985 census of Tampa Florida, Journal of the American Statistical Society, vol. 84, 406, pp 414-420.

[5] Jaro, M. A. (1995). Probabilistic linkage of large public health data file, Statistics in Medicine 14 (5-7), 491-8.

[6] Winkler W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage, Proceedings of the Section on Survey Research Methods, American Statistical Association, 354–359.

[7] Needleman, B. S. & Wunsch, D. C.(1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins", Journal of Molecular Biology 48(3): 443–53.

[8] Smith, F. T. & Waterman, S. M. (1981). Identification of Common Molecular Subsequences, Journal of Molecular Biology 147: 195–197.

[9] Alberto, B. , Paolo, R., Eneko A. & Gorka L. (2010). Plagiarism Detection across Distant Language Pairs, In Proceedings of the 23rd International Conference on Computational Linguistics, pages 37–45.

[10] Eugene F. K. (1987). Taxicab Geometry , Dover. ISBN 0-486-25202-7.

[11] Dice, L. (1945). Measures of the amount of ecologic association between species. Ecology, 26(3).

[12] Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles 37, 547-579.

[13] Lund, K., Burgess, C. & Atchley, R. A. (1995). Semantic and associative priming in a high-dimensional semantic space. Cognitive Science Proceedings (LEA), 660-665.

[14] Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instruments & Computers, 28(2),203-208.

[15] Mihalcea, R., Corley, C. & Strapparava, C. (2006). Corpus based and knowledge-based measures of text semantic similarity. In Proceedings of the American Association for Artificial Intelligence.(Boston, MA).

AUTHORS

M.K.VIJAYMEENA received the B.E degree in computer science and engineering from Jansons Institute of Technology in 2014. She is currently pursuing the M.E. degree in Computer science and Engineering at Nandha engineering college, Erode, India. Her current research interests include data mining, text mining and information retrieval.

Mrs.K.KAVITHA received the M.Tech. degree in Information Technology from SNS College of Technology, Coimbatore in 2011.She is currently working as an Assistant Professor in Nandha engineering college, Erode, India. Her current research interests include data mining and Network Security.