Comparative analysis of similarity measures on Marathi text Document for plagiarism detection

[1]Ramesh R. Naik, [2]Maheshkumar B. Landge, [3]C. Namrata Mahender

[1,2]Research scholar, [3]Assistant professor, Department of CS and IT, Dr.B.A.M.University, Aurangabad (MS), INDIA.

[1]ramesh.naik31@yahoo.com, [2]maheshkumar.landge@gmail.com, [3]nam.mah@gmail.com

Abstract - This paper provides a comparative analysis of Marathi text documents based on similarity measures. Various similarity measures have been applied on of textual contents for many different languages especially English, French etc. In this paper selective similarity measures are applied for similarity analysis considering their parameters which were found helpful according to the structure of Marathi language for text plagiarism detection. The first task was development of corpus as no standard corpus is available for Marathi language, second collection and normalization [conversion in same format] of data as numerous fonts are available. The collected data is converted to Unicode format standard UTF-8. Comparative analysis shows that sequence matcher is best similarity measure as compare to cosine similarity and Jaccard similarity.

Keywords —Similarity measures, cosine similarity, sequence matcher, Jaccard similarity, plagiarism detection.

I. INTRODUCTION

The plagiarism is a wide spread and growing problem in the academic process. Simi-larity has been a subject of great interest in human history since a long time ago. Plagiarism in a text document by observing similarities between it and other documents is called plagiarism detection[1].The measurement of similarity between different things is the important function of any information retrieval, data mining and plagiarism detection. There are a number of ways to compute similarity among various things. The majority of the current systems are only pattern discovery techniques based on basic similarity measures. The similarity measure is applied widely in many text applications which include classification and clustering [2]. In this paper we have used three similarity measures, first is sequence matcher this calculates the similarity ratio between two files. Secondly, cosine similarity is a measure taking the cosine of the angle between two vectors and third is Jaccard similarity deals with the similarity between the finite sets of sample which is regarded as the size of the intersection which is then divided by the size of the union of the sample sets [3].

II. LITERATURE SURVEY

Similarity measure is an important factor for plagiarism detection. This measure tells the degree of closeness or separation between a pair of objects. This literature survey covers different similarity measures we have studied in detail. The below description shows in brief various similarity measures studied.

A. *Minkowski Distance*

This is generic form of metric distance calculation for multidimensional data. Thennorm Minkowski distance measure can be defined as the distance Dij between two parts i and j as, [4].

B. *Manhattan distance*

The Manhattan is Minkowski distance at norm value of 1. It is the measure of Ab-solute difference between any two points. It is described as, [4]

$$Dij = \sum_{l=1}^{d} | xil - xjl |$$

C. *Euclidean distance*

The Minkowski distance at norm value of 2 is described as Euclidean distance. It is the most commonly used measure to determine distance between two points. It is described as [4].

$$Dij = (\sum_{l=1}^{d} | xil - xjl |^{1/2})^{2}$$

D. *Chebyshev distance*

Minkowski distance is termed as Chebyshev distance. It

$$Dij = (\sum_{l=1}^{d} | xil - xjl |^{1/2})^{n}$$

represents the greatest dis-tance between two vectors along any coordinate dimension. It is shown in given for-mula [4].

$$dj = \max_{1 \le l \le d} | xil - xjl |$$

E. *Dice's Coefficient*

Dice coefficient similarity measure is defined as twice the number of terms com-mon to compared entities/strings (nt) divided by the total number of terms in both tested strings [5].

$$s = \frac{2nt}{n_x + n_y}$$

F. *Hamming distance*

Hamming distance measure is used for binary attributes. It is defined as the number of bits which differ between two binary strings [6].

G. *Levenshtein Distance*

It is also referred to as edit distance and is a generalized form of hamming distance. The distance between two strings is given simply by the minimum edit operations needed to convert one string into the other. The edit operations are insertion, deletion, or substitution of a single character [7].

Cosine similarity is a measure of similarity between two vectors

of an inner product space that measures the cosine of the angle between them [5].

Jaccard similarity is computed as the number of shared terms over the number of all unique terms in both strings [8].

Matching Coefficient is a very simple vector based approach which simply counts the number of similar terms, on which both vectors are non-zero.

Overlap coefficient is similar to the Dice's coefficient, but considers two strings a full match if one is a subset of the other.

III. APPLIED MEASURES FOR COMPARATIVE STUDY

As per now as no foolproof similarity measure have been tested on Marathi text, to start off with we selected following three measures as per our requirement.

1. Sequence matcher is used for comparing pairs of sequences of any type [9].

2. Jaccard similarity is computed as the number of shared terms over the number of all unique terms in both strings. 3.cosine similarity is used to measure similarity between two vectors of an inner product space that measures the cosine of the angle between them [10].

To start with this section first provides the brief information on the database collected, then the measures applied and finally the comparative analysis is discussed in section 4.

A. Database description

We have created the Marathi text database, which contains 20 files for research article Marathi text files. From which 15 files for training database purpose and five text files for testing database. The Marathi text database manuscript is having minimum size of research article is 4 pages and maximum size is 7 pages available in the database. As many different fonts are used by Marathi writers eg. Shivaji font, kruti dev font, akrutidev Priya font, lekhani, Mangal, Aparajita font, Priya font. There is lot variation in the writing style of the fonts available thus before preparing database we normalize the data in one form that is Unicode format standard utf-8. This 1.29MB of information is available in the database in the format of word file for the further processing.

B. Sequence Matcher

In the result of table1 we have taken 15 Marathi text documents for training. In that we compared 15 documents with all 15 documents. In result of table2 we have taken 15 Marathi text documents for training and five Marathi text documents for testing. In that we compared 5 documents with all 15 documents. We have calculated similarity ratio between two files using sequence matcher similarity measure.

Table1: sequence matcher results training files

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	100.0	1.0920	0.4593	1.2125	0.4470	0.8173	2.4002	1.2722	1.9455	1.1806	1.1179	0.5711	0.7889	0.9914	1.2459
2	0.9678	100.0	0.9261	18.73	1.1948	1.5897	1.4169	0.8563	0.8005	1.4047	1.5300	0.9921	0.4778	0.9001	1.7299
3	0.3320	1.0675	100.0	1.7484	0.8180	0.9959	0.5277	1.7805	1.0926	0.8109	0.9126	0.3172	0.4304	1.4172	1.2814
4	0.7019	16.148	2.0206	100.0	1.4117	0.8635	0.9503	1.9295	1.4970	1.2350	1.8430	1.4414	0.4889	0.6397	1.0752
5	0.4806	1.3706	1.1964	1.1550	100.0	1.1325	0.5413	1.3950	1.0856	1.5842	1.9221	1.0206	1.1631	1.1575	0.4222
6	1.0275	1.0581	0.7590	0.7286	2.144	100.0	0.9034	1.2436	2.2815	1.9039	1.1699	1.0775	0.7357	1.5654	1.0785
7	1.7012	1.3431	0.9136	1.4454	1.0158	0.9327	100.0	1.2100	2.3195	1.1179	0.8442	0.6325	1.2195	1.4007	0.6344
8	1.6810	0.9668	1.4399	1.8815	0.7393	2.1656	2.0254	100.0	3.4184	1.4544	1.4779	1.4642	1.4048	2.0826	1.8089
9	1.7259	1.3471	0.6806	1.3413	1.0278	1.0263	1.0513	2.3929	100.0	1.4251	1.8241	1.3239	0.9630	2.6219	1.6319
10	1.3089	1.6934	0.4445	1.4037	1.2673	1.3439	0.7210	0.7026	0.6669	100.0	1.1449	0.8843	1.1359	1.2367	1.5858
11	1.0340	2.0021	1.4984	1.4225	1.9819	1.3553	0.7898	2.4343	2.3453	1.0282	100.0	0.9741	0.9640	2.1050	1.6414
12	1.5760	1.8867	0.8150	2.7956	1.0206	1.5227	0.7219	1.9254	1.7921	1.1644	1.7200	100.0	0.4402	1.5508	1.6206
13	1.0895	1.3693	0.9685	1.6001	2.0252	2.4372	1.6892	0.8966	2.1799	1.8356	1.7414	0.8097	100.0	1.9229	1.2581
14	1.4272	1.4804	0.8954	0.9241	1.5480	1.1908	1.5079	2.0979	1.8368	2.2200	1.6681	1.1169	1.0765	100.0	0.8955
15	0.5888	1.5888	1.2474	1.2552	1.2564	0.6940	0.4719	1.7869	0.9506	1.7682	1.2395	0.8845	0.4638	1.4788	100.0

Table2. Sequence matcher result training and testing files.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1.0518	1.7611	1.8098	1.4843	1.4559	1.9657	1.2814	2.2685	3.5791	2.3499	1.3978	1.1183	1.0639	1.4745	1.0920
2	1.4913	1.9123	1.1449	1.5342	0.7996	1.0973	1.0453	1.0646	2.2042	3.5993	1.5546	1.6295	0.4482	1.7252	2.1788

3	1.3340	1.6893	1.3837	1.4324	1.8885	1.1477	1.1638	1.92639	0.5649	1.6835	2.0283	0.5068	0.9251	1.8510	1.6928
4	0.4811	0.9423	0.8409	2.0754	1.714	1.7196	2.0644	0.9196	2.4055	2.8019	1.7452	1.8615	0.8014	1.3929	1.5889
5	0.9708	1.7134	0.3384	1.3510	0.5432	1.0984	1.9302	2.0710	1.3090	2.7158	1.0870	1.5645	1.5239	3.6214	1.8614

C. Jaccard similarity

In the result of table3 we have taken 15 Marathi text documents for training. In that we compared 15 documents with all 15 documents. In result of table4 we have taken 15Marathi text documents for training and five Marathi text documents for testing. In that we compared 5 documents with all 15 documents. We can compute similarity between two documents by using given formula.

$$J_{\delta(A,B)=1-J(A,B)=\frac{|A\cup B|-|A\cap B|}{|A\cup B|}}$$

Table3.Jaccard similarity result training files.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1.0	0.6416	0.6608	0.75	0.60377	0.5255	0.8804	0.7938	0.7849	0.5963	0.8152	0.5655	0.75	0.7956	0.7857
2	0.6416	1.0	0.7258	0.7982	0.5196	0.7709	0.6198	0.6416	0.6153	0.4924	0.6379	0.7258	0.6283	0.6101	0.7217
3	0.6608	0.7258	1.0	0.7610	0.5080	0.7519	0.6521	0.6608	0.6194	0.5158	0.6428	0.7478	0.6481	0.6	0.7454
4	0.75	0.7982	0.7610	1.0	0.5614	0.6564	0.7075	0.75	0.6730	0.5423	0.6990	0.7155	0.7244	0.6666	0.7941
5	0.6037	0.5196	0.5080	0.5614	1.0	0.4436	0.625	0.6190	0.6701	0.8888	0.7157	0.4841	0.6702	0.6632	0.6442
6	0.5255	0.7709	0.7519	0.6564	0.4436	1.0	0.5522	0.5597	0.5343	0.4513	0.5658	0.7656	0.5806	0.5419	0.6279
7	0.8804	0.6198	0.6521	0.7075	0.625	0.5522	1.0	0.7653	0.8333	0.6168	0.8651	0.6239	0.7977	0.8444	0.8315
8	0.7938	0.6416	0.6608	0.75	0.6190	0.5597	0.7653	1.0	0.7659	0.6111	0.7956	0.6050	0.8295	0.7578	0.8041
9	0.7849	0.6153	0.6194	0.6730	0.6701	0.5343	0.8333	0.7659	1.0	0.6274	0.9156	0.6194	0.8	0.9156	0.8152
10	0.5963	0.4924	0.5158	0.5423	0.8888	0.4513	0.6168	0.6111	0.6274	1.0	0.67	0.4692	0.6428	0.6213	0.6355
11	0.8152	0.6379	0.6428	0.6990	0.7157	0.5658	0.8651	0.7956	0.9156	0.67	1.0	0.6140	0.8554	0.9277	0.8461
12	0.5655	0.7258	0.7478	0.7155	0.4841	0.7656	0.6239	0.6050	0.6194	0.4692	0.6140	1.0	0.6330	0.6140	0.6991
13	0.75	0.6283	0.6481	0.7244	0.6702	0.5806	0.7977	0.8295	0.8	0.6428	0.8554	0.6330	1.0	0.8117	0.8202
14	0.7956	0.6101	0.6	0.6666	0.6632	0.5419	0.8444	0.7578	0.9156	0.6213	0.9277	0.6140	0.8117	1.0	0.7872
15	0.7857	0.7217	0.7454	0.7941	0.6442	0.6279	0.8315	0.8041	0.8152	0.6355	0.8461	0.6991	0.8202	0.7872	1.0

Table4. Jaccard similarity result training and testing files.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.7956	0.6666	0.6880	0.7326	0.6632	0.5905	0.7849	0.8351	0.8275	0.6213	0.8604	0.6428	0.8554	0.7977	0.8461
2	0.8222	0.6120	0.6454	0.6862	0.6494	0.5426	0.8314	0.7826	0.8352	0.6237	0.8915	0.5877	0.8414	0.8915	0.7741
3	0.5419	0.7578	0.7380	0.6535	0.4666	0.8091	0.5703	0.5781	0.5645	0.4532	0.5725	0.7804	0.5882	0.56	0.6504
4	0.6875	0.6517	0.6574	0.7171	0.6288	0.5634	0.75	0.7608	0.7701	0.5882	0.7816	0.6728	0.7951	0.7415	0.8314
5	0.8043	0.6153	0.6339	0.6893	0.6701	0.5581	0.875	0.7849	0.9036	0.6435	0.9390	0.6339	0.8433	0.9156	0.8152

D. Cosine similarity

Cosine similarity is a popular vector based similarity measure in text mining and Information retrieval. In this approach compared strings are transformed into vector space so that the Euclidean cosine rule can be used to calculate similarity. This approach is often paired with other approaches to limit the dimensionality of the vector space [11].

In the result of table5 we have taken 15 Marathi text documents for training. In that we compared 15 documents with all 15 documents. In result of table6

We have taken 15Marathi text documents for training and five Marathi text documents for testing. In that we compared 5 documents with all 15 documents. We can compute similarity between two documents by using given formula.

$$similarity = \cos\Theta = \frac{A.B}{\| A \| \| B \|} = \frac{\sum_{t=1}^{n} At * Bt}{\sqrt{\sum_{t=1}^{n} A_t^2} * \sqrt{\sum_{t=1}^{n} B_t^2}}$$

Table5. Cosine similarity result training files.

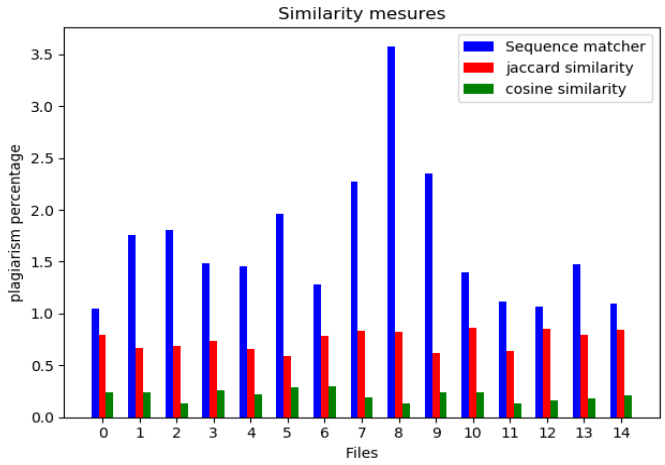
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1.0	0.2558	0.2080	0.3098	0.2735	0.3229	0.5403	0.35947	0.2305	0.3007	0.3725	0.2394	0.2360	0.2931	0.3395
2	0.2558	1.0	0.2951	0.7807	0.2163	0.3352	0.3072	0.2192	0.2035	0.2697	0.2677	0.3824	0.2093	0.1965	0.3326
3	0.2080	0.2951	1.0	0.3190	0.1402	0.2501	0.2292	0.1864	0.1471	0.1634	0.1636	0.2610	0.1448	0.1159	0.2392
4	0.3098	0.7807	0.3190	1.0	0.2212	0.4088	0.3656	0.2539	0.2192	0.3093	0.2847	0.3728	0.2218	0.2070	0.3578
5	0.2735	0.2163	0.1402	0.2212	1.0	0.2313	0.3179	0.2316	0.3333	0.1811	0.2357	0.1296	0.2093	0.2229	0.2348
6	0.3229	0.3352	0.2501	0.4088	0.2313	1.0	0.3998	0.2951	0.2221	0.3344	0.2632	0.3096	0.2720	0.2146	0.3116
7	0.5403	0.3072	0.2292	0.3656	0.3179	0.3998	1.0	0.4003	0.2779	0.3697	0.3797	0.2947	0.2954	0.3237	0.3612
8	0.3594	0.2192	0.1864	0.2539	0.2316	0.2951	0.4003	1.0	0.2779	0.3494	0.2680	0.2331	0.2863	0.2825	0.2472
9	0.2305	0.2035	0.1471	0.2192	0.3333	0.2221	0.2779	0.2779	1.0	0.2887	0.2063	0.2049	0.2290	0.3043	0.1903
10	0.3007	0.2697	0.1634	0.3093	0.1811	0.3344	0.3697	0.3494	0.2887	1.0	0.2745	0.2147	0.2763	0.2306	0.2958
11	0.3725	0.2677	0.1636	0.2847	0.2357	0.2632	0.3797	0.2680	0.2063	0.2745	1.0	0.1675	0.2072	0.2732	0.2955
12	0.2394	0.3824	0.2610	0.3728	0.1296	0.3096	0.2947	0.2331	0.2049	0.2147	0.1675	1.0	0.2025	0.1231	0.2531
13	0.2360	0.2093	0.1448	0.2218	0.2093	0.2720	0.2954	0.2863	0.2290	0.2763	0.2072	0.2025	1.0	0.1709	0.1716
14	0.2931	0.1965	0.1159	0.2070	0.2229	0.2146	0.3237	0.2825	0.3043	0.2306	0.2732	0.1231	0.1709	1.0	0.2322
15	0.3395	0.3326	0.2392	0.3578	0.2348	0.3116	0.3612	0.2472	0.1903	0.2958	0.2955	0.2531	0.1716	0.2322	1.0

Table6. Cosine similarity result training and testing files.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.2412	0.2440	0.1323	0.2643	0.2244	0.2928	0.2986	0.1918	0.1316	0.2362	0.2427	0.1358	0.1652	0.1825	0.2082
2	0.3208	0.2449	0.1823	0.3157	0.2195	0.3415	0.3995	0.3539	0.2316	0.3346	0.2897	0.2352	0.2842	0.2165	0.2631
3	0.1185	0.3184	0.1558	0.2875	0.1367	0.1947	0.1472	0.1004	0.1006	0.1247	0.1337	0.1928	0.1262	0.1049	0.1534
4	0.2665	0.3449	0.3070	0.3863	0.1840	0.3582	0.3171	0.2332	0.1762	0.2198	0.2213	0.2948	0.2023	0.1539	0.2737
5	0.2979	0.2141	0.1359	0.2233	0.2354	0.2232	0.3568	0.3099	0.1804	0.2408	0.2435	0.1444	0.1908	0.3344	0.2407

IV. ANALYSIS

Graph1. Comparative analysis of three different similarity measures.



Calculating similarity between words is a basic part of text similarity which is then used as a primary stage for sentence, paragraph and document similarities. Till now above similarity measures sequence matcher, Jaccard similarity, cosine similarity has not used to detect similarity in Marathi text. We have applied these similarity measures on Marathi text to detect plagiarism. Above graph1 shows the comparative analysis of three different similarity measures, in that we have taken Marathi text documents for training and testing. In that we compared one testing document with all 15 documents. We have computed similarity between two documents and by using sequence matcher the highest similarity is 3.5791 and lowest is 1.0518, in Jaccard similarity the highest similarity is 0.8554 and lowest is 0.5905 and in cosine similarity the highest similarity is 0.2986 and lowest is 0.1323. From above comparative analysis we found Sequence matcher is best similarity measure as compare to cosine similarity and Jaccard similarity.

V. CONCLUSION

Text document Similarity is a process where two text documents are compared to find the Similarity between them. This paper covers the brief introduction of similarity measures and literature survey covers different similarity measures. We have taken Marathi text documents for training and testing. After that we have calculated similarity of documents by using three different similarity measures and in result we have been done comparative analysis for plagiarism detection from that we found Sequence matcher is best similarity measure as compare to cosine similarity and Jaccard similarity.

ACKNOWLEDGMENTS

We are thankful to the Computational and Psycho-linguistic Research Lab, Department of Computer Science & Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS) for providing the facility for carrying out the research.

REFERENCES

[1] Paul Clough.: Plagiarism in natural and programming languages an overview of current tools and technologies. Technical report, University of Sheffeld, Sheffeld, UK, Jun 2000.

[2] Anna Huang. : Similarity Measures for Text Document Clustering. NZCSRSC 2008, Christchurch, New Zealand, 2008.

[3] J. Han and M. Kamber. : Data Mining: Concepts and Techniques.2nd ed. San Francisco, CA, USA: Morgan Kaufmann; Boston, MA, USA: Elsevier, 2006.

[4] Peter Grabusts. : The choice of matrics for clustering algorithms. 8th International Scientific and Practical conference, vol-2, 2011.

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

[6] Ankita Vimal, Satyanarayana R Valluri, Kamalakar Karlapalem. : An Experiment with Distance Measures for Clustering. Technical Report: IIIT/TR/2008/132.

[7] L. Gravano, P. G. Ipeirotis, H. V. Jagadish, N. Koudas, S. Muthukrishnan, and D. Srivastava. : Approximate string joins in a database (almost) for free. In VLDB, pages 491–500, 2001.

[8] Jaccard, P. 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, and 1901.

[9] https://docs.python.org/2/library/difflib.html#difflib.SequenceMatcher, last accessed 2018/01/25

[10] Wael H. Gomaa, Aly A. Fahmy: A Survey of Text Similarity Approaches, International Journal of Computer Applications (0975 – 8887) Volume 68– No.13, April 2013

[11] P.-N. Tan, M. Steinbach, and V. Kumar. : Introduction to Data Mining. Boston, MA, USA: Addision-Wesley, 2006.

