Advances in Computational Sciences and Technology
ISSN 0973-6107 Volume 10, Number 5 (2017) pp. 923-932

http://www.ripublication.com

Computational Model for Document Density Based Classification

Swatantra Kumar Sahu
M.G.C.G.V.V. Satna, Madhya Pradesh, India.

Bharat Mishra
M.G.C.G.V.V. Satna, Madhya Pradesh, India.

R. S. Thakur
MANIT Bhopal, Madhya Pradesh, India.

Abstract

This paper presents a new approach document density (DD) based classification. The proposed approach is based on some parts of speech count and total words count (WC) ratio in the documents. In this paper we have used some parts of speech count and total words count ratio for finding high level, middle level and low level document density, and generated clusters of high level, middle level and low level document density. The experimental results are evaluated using the numerical computing MATLAB7.14. The Experimental results show the proposed approach is a optimistic solution for document classification.

Keywords: Document Density, Parts of speech Count, Words Count, and Document Classification.

1. INTRODUCTION

Document density (DD) based classification used some parts of speech count and words count of documents for high level DD > 0.6, middle level DD < = 0.6 and DD > = 0.4 and low level DD < 0.4. We use a document sets are D = {D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9} and start words count with different level of document density high level DD > 0.6 are {D_1,D_3,D_8}, middle level DD < = 0.6 and DD > = 0.4 {D_4, D_7, D_9} and low level DD < 0.4 {D_2, D_5, D_6}.

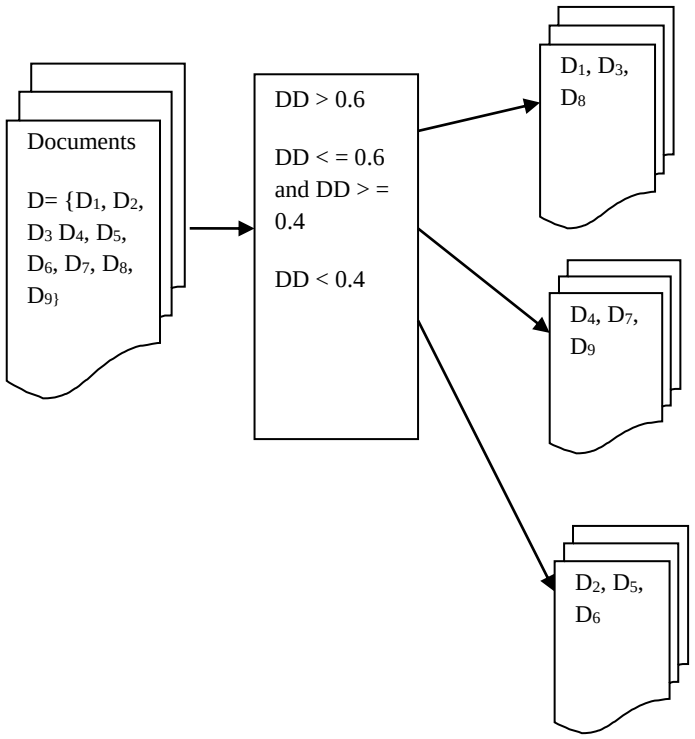
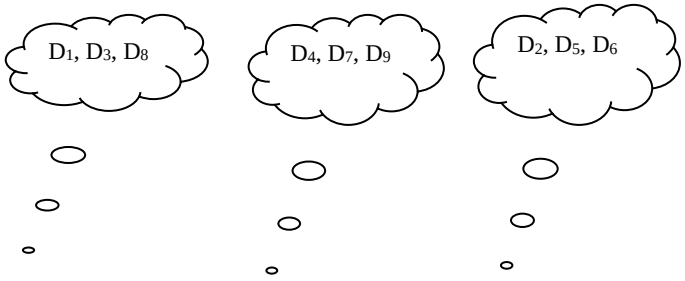


Figure 1: Documents density at Different levels

The figure1 shows documents density classification with conditional statements and figure2 shows different level document cloud. The document D_1,D_3,D_8 comes under the same cloud i.e. High level. Similarly the document D_4,D_7,D_9 comes together under middle level and document D_2,D_5,D_6 comes under low level as shown by figure 2.



(a) High level (b) middle level (c) low level

Figure 2: Documents Cloud at Different levels

In 1956 Rosenblatt, M. gave Remarks on some nonparametric estimates of a density function [2].

In 1962 Parzen, E. gave On Estimation of a Probability Density Function and Mode[7].

In 1972 Jones, K. gave A statistical interpretation of term specificity and its application in retrieval [5].

In 1974 Finkel, R. and Bentley, J. gave Quad trees — a data structure for retrieval on composite keys[4]

In 2008 Manning, C. D., Raghavan, P., and Schtze, H. gave Introduction to Information Retrieval [3].

In 2009 Michael A. gave "Idea Density –A Potentially Informative Characteristic of Retrieved Documents"[1], Palen, L., Vieweg, S., Liu, S., and Hughes, A. gave Crisis in a networked world: features of computer-mediated communication [6].

In 2011 Wing, B. P. and Baldridge, J. gave Simple supervised document geolocation with geodesic grids[8],Bosch, H., Thom, D., Wörner, M., Koch, S., Püttmann, E., Jäckle, D., and Ertl, T. gave ScatterBlogs: Geo-spatial document analysis[10].

In 2012 Roller, S., Speriosu, M., Rallapalli, S., Wing, B., and Baldridge, J. gave Supervised text-based geolocation using language models on an adaptive grid[9].

2. DOCUMENT DENSITY BASED CLASSIFICATION

This section discussed different formulae F_V, F_P, F_{ADJ}, F_{ADV}, F_{CON} that are shown in Table1.Table2 describe different parts of speech and there descriptions.

Table 1: Function using Different parts of speech

DD → / FORMULAE	FORMULAE DESCRIPTION
F_V	$\frac{Verbs\ Count\ of\ Document}{Total\ Words\ Count\ of\ Document}$
F_P	$\frac{Preposition\ Count\ of\ Document}{Total\ Words\ Count\ of\ Document}$
F_{ADJ}	$\frac{Adjectives\ Count\ of\ Document}{Total\ Words\ Count\ of\ Document}$
F_{ADV}	$\frac{Adverbs\ Count\ of\ Document}{Total\ Words\ Count\ of\ Document}$
F_{CON}	$\frac{Conjunctions\ Count\ of\ Document}{Total\ Words\ Count\ of\ Document}$

Table 2. Documents Density with Different parts of speech

Parts of speech name	**Descriptions**
Verbs	Be, have, do, say, get, make, go, know, take, see, come, Think, look, want, give, use, find, tell, ask, work, seem, feel, try, leave, call
Prepositions	of, in, to, for, with, on, at, from, by, about, as, into, like, through, after, over, between, out, against, during, without, before, under, around, among.
Adjectives	Good, new, first ,last, long, great, little, own, other, old, right, big, high, different, small, large, Next, early, young, important, few, public, bad, same, able.
Adverbs	Up, so, out, just, now, how, then, more, also, here, well, only, very, even, back, there, down, still In ,as, too, when, never, really, most.
Conjunctions	And, that, but, or, as, If, when, than, because, while, where, after, so, though, since, until, whether Before, although, nor, like, once, unless, now, except

3. PROPOSED METHODOLOGY FOR DOCUMENT CLASSIFICATION

In the Classification of document the different steps are used. The steps are as follows:

3.1 Data Collection: In this phase collect relevant documents like e-mail, news, web pages etc. from various heterogeneous sources. These text documents are stored in a variety of formats depending on the nature of the data. The datasets can also download from UCI KDD Archive. This is an online repository of large datasets and has wide variety of data types.

3.2 Review of existing Classification Methods: Initial step is to complete review of literature in the field of data mining. Next step is a detailed study of existing Algorithms for Classification. In this area lot of work done by various researchers. After studying their work, it would be attempted to find the disadvantages of existing Classification algorithms.

3.3 Design of Classification Process: In this phase a new algorithm developed for Classification Process. Classification Process means transform documents into a suitable determined classes for the Classification task. In Classification Process we performed Different tasks.

3.4 Classification Results: In this Experiment the calculation of document density (DD) is done which lead to document classification process. Document density (DD) based Document Classification is an efficient and accurate compare to other Classification method.

4. ALGORITHM FOR DOCUMENT CLASSIFICATION

The algorithm 4.1 describes document density for parts of speech.

Algorithm 4.1: This Algorithm obtains density of documents for parts of speech

Step 1: Input document sets D = {D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9}, document density formulae F={ F_V, F_P, F_{ADJ}, F_{ADV}, F_{CON} }

Step 2: Read documents for calculating correspondent document density.

Step 3: Produce and compare document words one by one from set of parts speech POS = {verbs, prepositions, adjectives, adverbs, conjunctions}.

Step 4: Count verbs, prepositions, adjectives, adverbs, conjunctions from document sets D = {D_1, D_2, D_3, D_4, D_5, D_6, D_7,D_8, D_9}.

Step 5: Count Words from document sets D = {D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9}.

Step 6: Select correspondent document density formulae F={ F_V, F_P, $F_{ADJ,}$ F_{ADV}, F_{CON}}

Step 7: Calculate Document density for correspondent parts of speech.

Step 8: Generates the clusters cloud of high level, middle level and low level document density for correspondent parts of speech.

Table 3: Documents Density with Different parts of speech

→ **DD** ↓ **LEVEL**	**DD_1 Verbs**	**DD_2 Pre positions**	**DD_3 Adjec tives**	**DD_4 Ad verbs**	**DD_5 Conjun ctions**
L_H	{ D_2, D_6, D_8 }	{ D_1, D_2, D_5,D_9}	{ D_4, D_6, D_8 ,D_9}	{ D_4}	{ D_5}
L_M	{ D_1, D_9 }	{ D_7, D_8 }	{ D_1, D_2, D_8 }	{ D_3, D_5, D_6 }	{ D_2, D_7, D_8 }
L_L	{ D_4,D_5,D_7 }	{ D_3, D_4, D_6 }	{ D_3, D_5}	{ D_1, D_2, D_7 D_8,D_9}	{ D_1, D_3, D_4 ,D_6, D_9}

5. EXPERIMENTAL RESULTS

In this section we calculate the document density (DD) for document classification. Density of documents for some parts of speech like verbs, prepositions, adjectives, adverbs, conjunctions. Verbs, prepositions, adjectives, adverbs, conjunctions counted and find the density of the documents D = {D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9} which is describe in Table 3.

The words count is set with different level of document density i.e. high level DD > 0.6 , middle level DD < = 0.6 and DD > = 0.4 and low level DD < 0.4 as shown in Table 4.This Experiment shows document density based approach is more accurate and efficient as shown in table 5.

Table 4: Documents sets with Different level

→ DD ↓ DOCUMENT	**DD_1 Verbs**	**DD_2 Pre positions**	**DD_3 Adjec tives**	**DD_4 Ad verbs**	**DD_5 Conjun ctions**
D_1	0.478	0.718	0.578	0.355	0.318
D_2	0.645	0.625	0.465	0.315	0.425
D_3	0.213	0.333	0.213	0.425	0.133
D_4	0.123	0.323	0.623	0.637	0.223
D_5	0.345	0.745	0.345	0.517	0.645
D_6	0.665	0.395	0.665	0.418	0.365
D_7	0.265	0.465	0.465	0.325	0.465
D_8	0.667	0.457	0.667	0.213	0.457
D_9	0.597	0.747	0.697	0.123	0.347

The figure3 describes the documents Density with Different parts of speech for documents {D_1, D_2, D_3}.

The figure4 describes the documents Density with Different parts of speech for documents {D_4, D_5, D_6}

The figure5 describes the documents Density with Different parts of speech for document {D_7,D_8,D_9}

The figure6 show the accuracy of proposed documents Density based method for document {D_1,D_2,D_3}.

The figure7 show the accuracy of proposed documents Density based method in percentages for document {D_4,D_5,D_6}.

The figure8 show the accuracy of proposed documents Density based method in percentages for document {D_7,D_8,D_9}

Table 5: Accuracy (in percentage) of proposed Document density based method

→ DD DOCUMENT ↓	**DD_1 Verbs**	**DD_2 Pre positions**	**DD_3 Adjec tives**	**DD_4 Ad verbs**	**DD_5 Conjun ctions**
D_1	79	91	77	74	76
D_2	88	93	86	79	86
D_3	90	89	91	84	85
D_4	87	88	76	73	75
D_5	78	84	82	88	78
D_6	73	91	75	81	84
D_7	83	93	85	76	78
D_8	76	87	89	87	86
D_9	86	87	89	88	82

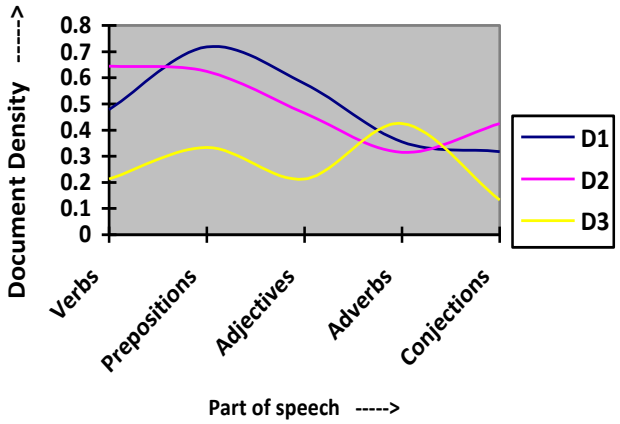


Figure 3: Documents {D_1,D_2,D_3} Density with Different parts of speech

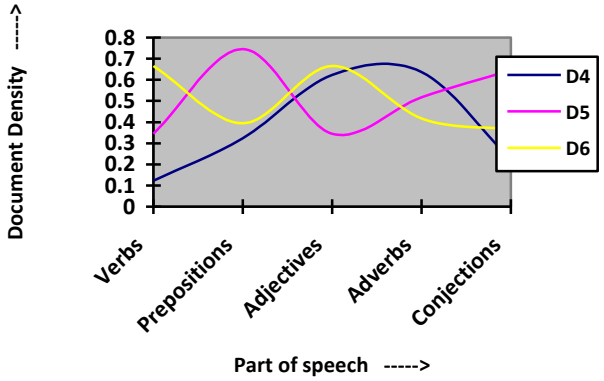


Figure 4: Documents {D_4,D_5,D_6} Density with Different parts of speech

Figure 5: Documents {D_7,D_8,D_9} Density with Different parts of speech

Figure 6: Accuracy result with Documents {D_1,D_2,D_3}

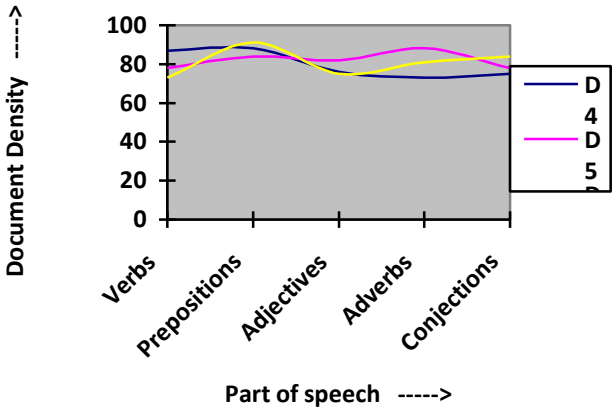


Figure 7: Accuracy result with Documents {D_4,D_5,D_6}

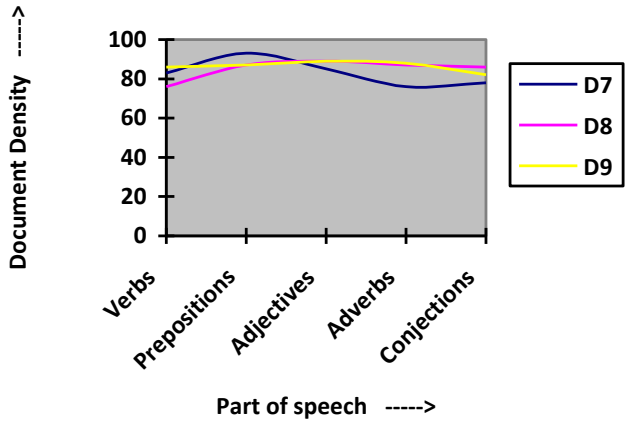


Figure 8: Accuracy result with Documents {D_7,D_8,D_9}

6. CONCLUSION

This paper discussed and analyzed the density based classification algorithm for document classification. The document density based classification algorithm is efficient for document cluster and generate more accurate classes for document data. In this analysis we can easily understand the various conditions which are responsible for the classification of various types of document datasets. This analysis also shows that this method works efficiently, for large text data.

REFERENCES

[1] Michael A. Covington "Idea Density –A Potentially Informative Characteristic of Retrieved Documents" Institute for Artificial Intelligence 2009.

[2] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. TheAnnals of Mathematical Statistics, 27(3):832–837.

[3] Manning, C. D., Raghavan, P., and Schtze, H. (2008). Introduction to Information Retrieval.Cambridge University Press, New York, NY, USA.

[4] Finkel, R. and Bentley, J. (1974). Quad trees — a data structure for retrieval on composite keys. Acta informatica, 4(1):1–9.

[5] Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of documentation, 28(1):11–21.

[6] Palen, L., Vieweg, S., Liu, S., and Hughes, A. (2009). Crisis in a networked world: features of computer-mediated communication in the april 16, 2007, virginia tech event. Social Science Computer Review.

[7] Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. The Annals of Mathematical Statistics, 33(3):1065–1076.

[8] Wing, B. P. and Baldridge, J. (2011). Simple supervised document geolocation with geodesic grids. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:Human Language Technologies - Volume 1, HLT '11, pages 955–964, Stroudsburg, PA, USA.Association for Computational Linguistics

[9] Roller, S., Speriosu, M., Rallapalli, S., Wing, B., and Baldridge, J. (2012). Supervised text-based geolocation using language models on an adaptive grid. In EMNLP-CoNLL, pages 1500–1510. ACL.

[10] Bosch, H., Thom, D., Wörner, M., Koch, S., Püttmann, E., Jäckle, D., and Ertl, T. (2011).ScatterBlogs: Geo-spatial document analysis. In Visual Analytics Science and Technology, 2011.VAST 2011. IEEE Conference on.