Thematic Context Derivator Algorithm for Enhanced Context Vector Machine: eCVM

Vaibhav Khatavkar, Makarand Velankar, Parag Kulkarni

Abstract: ***Natural Language Processing uses word embeddings to map words into vectors. Context vector is one of the techniques to map words into vectors. The context vector gives importance of terms in the document corpus. The derivation of context vector is done using various methods such as neural networks, latent semantic analysis, knowledge base methods etc. This paper proposes a novel system to devise an enhanced context vector machine called eCVM. eCVM is able to determine the context phrases and its importance. eCVM uses latent semantic analysis, existing context vector machine, dependency parsing, named entities, topics from latent dirichlet allocation and various forms of words like nouns, adjectives and verbs for building the context. eCVM uses context vector and Pagerank algorithm to find the importance of the term in document and is tested on BBC news dataset. Results of eCVM are compared with compared with the state of the art for context detrivation. The proposed system shows improved performance over existing systems for standard evaluation parameters.***

Keywords: ***Context Vector Machine, Natural Language Processing, PageRank, Named Entities.***

I. INTRODUCTION

Typically word embeddings involve various language modelling and feature learning techniques in Natural Language Processing (NLP) to map words into vectors. The mapping of words include neural networks, dimensionality reduction of term matrix, probabilistic models, knowledge base methods and context vectors. The general tasks in NLP like POS tagging, semantic relatedness has proved to have good performance if word embeddings are used. But tasks like named entity recognition do not benefit from word embeddings [1]. Many attempts are made to derive context of word using Latent Semantic Analysis (LSA), vector space model etc. There is lack of use of named entities in deriving context. When a context is defined using word embeddings there might be loss in information about a named entity. So a modified context vector machine (eCVM) is designed which will find context by combining using both word embeddings and named entities.

Revised Manuscript Received on December 30, 2019.

* Correspondence Author

Vaibhav Khatavkar*, Research Scholar, Department of Computer Engineering, College of Engineering, Pune, Shivajinagar, Pune, Maharashtra, India. E-mail: vkk.comp@coep.ac.in.

Makarand Velankar, Assistant Professor, Department of IT, MKSSS's Cummins College of Engineering, Kothrud, Pune, Maharashtra , India.

Parag Kulkarni, Adjunct Professor, Department of Computer Engineering, College of Engineering, Pune, Shivajinagar, Pune, Maharashtra, India.

The context derived in context vector machine [5] and LSA [2] give context as various terms which define a single context using single document or paragraph. We combine them along with named entities to improve the performance of the system.

The aim of the proposed work is as follows :

1. To design a Context Vector Machine (CVM) such that it make use of word embeddings and named entities to generate desired global context.
2. To compare the results of Context Vector Machine with existing techniques.

The arrangement of paper is as follows : Section I is an introduction and motivation of work. Section II explains the literature review of the work followed by proposed method in section III. Section IV gives experimentation and results of proposed work. Section V concludes the paper.

II. LITERATURE REVIEW

Various attempts have been made in deriving the context of a document. Some of them are Vector Space Model, word embeddings using Latent Semantic Analysis . Well established method for voluminous document corpus is Latent Semantic Analysis. These approaches are explored by different researchers and are covered in the following subsections.

Latent Semantic Analysis- The context of a document can be a theme of a document. Typically, Latent Semantic Analysis (LSA) is used for theme generation. Khatavkar V and Kulkarni P in [2], devised an algorithm to get the context of a document that gave context terms from the document and their weights. Khatavkar V and Kulkarni P in [3], compared SVM with LSA and without LSA. LSA proved to be effective with SVM when the corpus is large. The LSA is performed using Singular Value Decomposition (SVD). The matrix M can be derived using the equation :

$M = U*S*V$ ……(1)

where U = Reduced rank term matrix ; S = Singular Value Diagonal Matrix ; V = Document Matrix.

Named entity analysis- In text documents, Named Entities play a vital role in the identification of context. Many attempts are made to extract named entities from the document and then either classify them or derive context from them [4,5]. Shu L et. al. in [6] resolved entities and modeled the topics from the document.

Forms of words- Many times Noun and Verb phrases play an important role in the same. Forms of words like Nouns, Adjectives, Verbs are used in the identification of document context [7].

Context Vector Machine: Researchers in [2, 5] derived using context using Term Frequency Inverse Document Frequency (TF-IDF). The work presented in this paper is an extension of the work mentioned in [8]. In this work a system is developed, a modified Context Vector Machine (eCVM), which applies dependency parser to resolve the word dependencies in the paragraph. The dependency parser is an English Language Parser which parses the paragraph and resolves the dependencies present in the sentences. Stanford's Dependency Parser serves the purpose [9]. The context frequency form term t document d from document corpus D and context vectors are formulated as shown in equation 1,2 and 3 :

$$cf(t,d) = \frac{Number\ of\ context\ term\ t\ in\ document\ d}{Total\ number\ of\ context\ terms\ in\ document\ d} \quad .....\ (2)$$

$$icf(t,D) = \frac{\log|D|}{\{d' \varepsilon D \mid t\ \varepsilon d'\}} \quad .....\ (3)$$

$$CV(t,d,D) = cf(t,d) * icf(t,D) \quad ......(4)$$

Topic Analysis- Blei D M et. al. in [10] researched that LDA is a very powerful tool to find out the structure that is not known in a large set of text. The reason for the development of LDA was to solve the issues of the previous model known as PLSA proposed by Hofmann in [11]. PLSA was a probabilistic version of Latent semantic analysis done by Deerwester et. al. in [12]. Latent Dirichlet Allocation (LDA) is performed on the documents in order to get the topics from the documents.

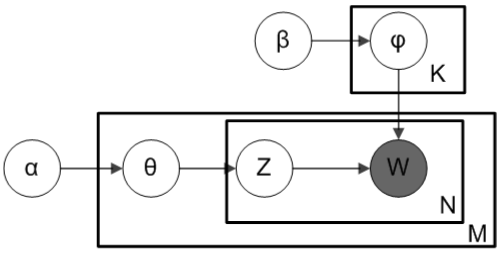

Figure 1. Plate Notation for LDA [10]

The plate notation of LDA with its standard notations is shown as in Figure 1. . The parameter Z will be the topics for document with words W where α, β, θ and φ are prior topic dirichlet, post topic dirichlet, topic and word distributions respectively.

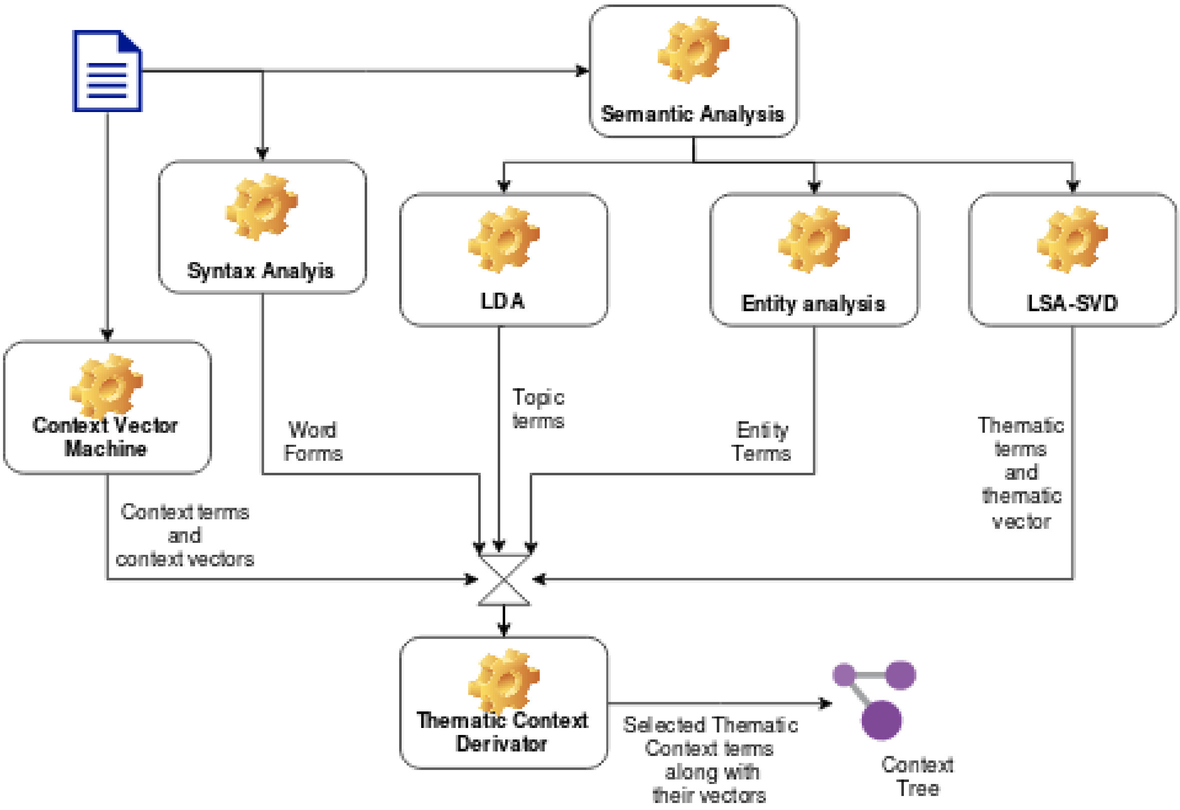
Semantic Analysis
Syntax Analyis
LDA
Entity analysis
LSA-SVD
Context Vector Machine
Word Forms
Topic terms
Entity Terms
Thematic terms and thematic vector
Context terms and context vectors
Thematic Context Derivator
Selected Thematic Context terms along with their vectors
Context Tree

Figure2. Proposed System

III. PROPOSED SYSTEM

Figure 2 depicts the proposed system. The overall working of the system is stated as below :

1. Take input document,
2. Build vector v_0 using LDA,
3. Build vector v_1 for all named entities from document,
4. Build vector v_f such that it contains terms from v_0, v_1 and all nouns, verbs and adjectives related to terms in v_0 and v_1.

5. Build vector v_2 using LSA SVD,
6. Build vector v_3 by building context vector using CVM,
7. Build v_4 for thematic context derivation by integrating vectors v_f, $v_2$and v_3.
8. Perform thematic context derivation algorithm in order to get context phrases.

The system will derive the context of the document by using thematic context derivator algorithm. Thematic context derivation algorithm is given below :

1. Identify subjects, verbs and objects from v_f, v_2 and v_3 and save the identity in 'id' for document i
2. Select term t from v_f, $v_2$and v_3
3. If t is in 'id' :
4. Add t in v_4 along with 'id'
5. Perform step 2 tillv_f, $v_2$and v_3 are covered.
6. For all terms in v_4 add t as nodes in graph G and id as relationship.
7. Change i to i+1 (consider next document from the corpus) go to step 1 if document i+1 is in corpus.
8. Calculate pagerank for all nodes in G to get the modified context vector and importance of terms in v_4.
9. Return context vector in v_4 along with importance of each term.

The thematic context derivator algorithm gives the context based on pagerank. It gives us modified context vector of the term in the document.

The page rank of context term t in document d is derived when all the context terms are in graph G [13]. The terms $t\ and\ j$ are connected by an edge e in G are connected such that the weight w of e, $w = max\ (wup(t,j))$ and ,

$$PR(t,d) = \sum_{j \in B_t} \frac{PR(j)}{L(j)} \;....\;(5)$$

where,

$t\ and\ j\ are\ context\ terms;$

$B_t\ is\ set\ containing\ all\ nodes\ linking\ to\ term\ t;$

$L(j)\ is\ the\ number\ of\ nodes\ from\ node\ t;$

The Wu-Palmer similarity is used to find the relation score between two terms, [14], given as :

$$wup(t_1,t_2) = \frac{(2*depth(lcs(t_1,t_2)))}{(len(t_1,t_2) + 2*depth(lcs(t_1,t_2)))} \;.\;(6)$$

where, $depth(lcs(t_1,t_2))$ is lcs the depth of the lowest node in the thesaurus hierarchy of terms t_1 and t_2.

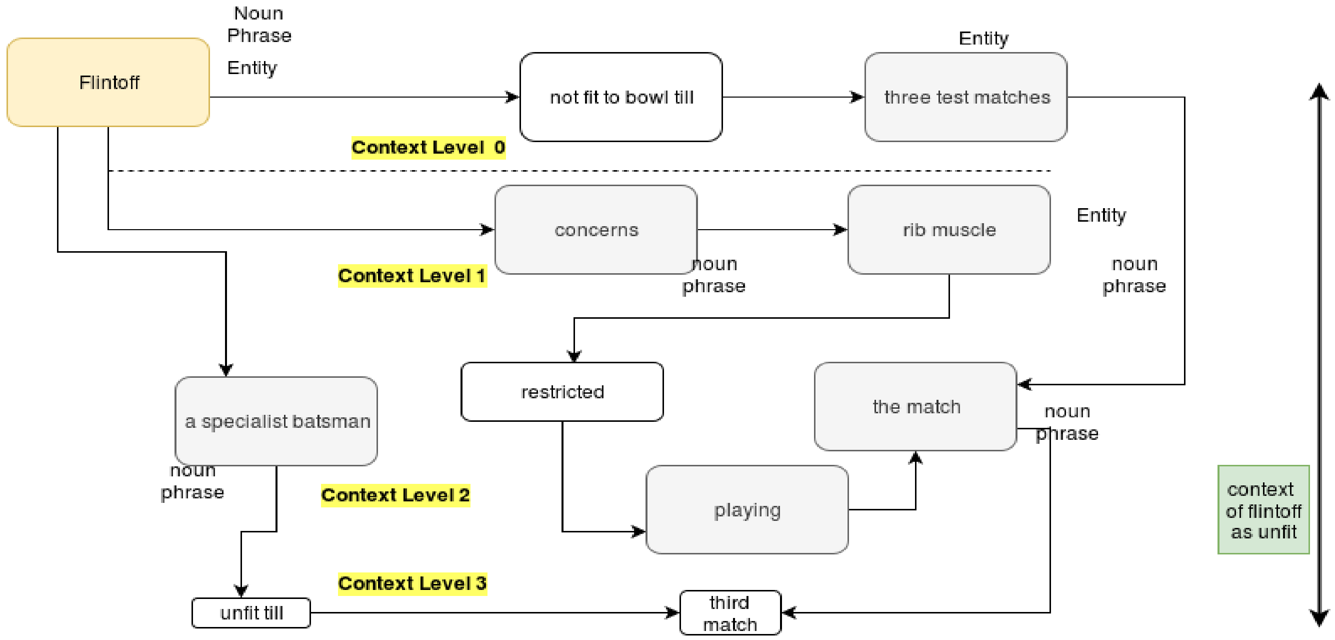

Figure 3. Sample Context for document 1

The context vectors and page ranks of the terms are derived using Equations 3 and 14. The page rank will give the importance of the term from the set of terms in the graph G. In a graph G, nodes representing terms may have the same page rank. Also, the Context Vectors may come out to be the same. So it is necessary to modify the context vector. The modified context vector $mcv(t,d)$ for term $t$in document d is given in Equation 7.

$$mcv(t,d) = cv(t,d) * PR(t,d) \ldots (7)$$

The importance (i_t) of the context term t can be given as :

$$i_t = \frac{mcv(t,d)}{dim(v_4)} \;....(8)$$

IV. EXPERIMENTATION AND RESULTS

The system is tested on BBC news dataset [15]. The dataset consists of five categories of news articles namely, business, entertainment, politics, sports and tech. The total news articles in the dataset are 2225. The sample text of 043.txt and 0.42.txt documents from sports category is taken for explanation.

Sample text from document 1 : "Flintoff fit to bowl at Wanderers Fourth Test, Wanderers: Captain Michael Vaughan

said: "He's had a bowl and came out fine so he is fully fit to play as an all-rounder."

Sample text from document 2 : "England's main concern continues to be the fitness of Andrew Flintoff, who is recovering from a torn side muscle and may not be able to bowl in the Test."

The various context vector terms for document 1 with document 2 in consideration are :

Topic according to LDA = v_0 = Flintoff, fit , bowl, play

Named Entities = v_1 = Andrew Flintoff, Captain Michael Vaughan, test match, England

LSA with SVD = v_2 = fit, bowl, play, rib, muscle

Terms from CVM = v_3 = fit, play, vaughan, bowl, concern, batsman, special

eCVM = v_4 = flintoff, fit, bowl, play, all-rounder, wanderers , Captain Michael Vaughan, test match, rib muscles, restricted.

For v_4 a context graph as depicted in figure is built and modified context vector is calculated. The

PageRank, Context Vector , Modified Context Vector and importance for sample context terms is shown in Table 1.

The context of the document 1 given from eCVM is "flintoff fitness" , "test match", "Flintoff performance" , " injury of Flintoff " , "previous matches ", "Marcus Trescothick", "England Team spirit", "Steve Fitness", "South Africa Team", "Captains", "Players " and "England team". One context derivation is shown in Figure 3. The figure shows named entities, subject , object and phrases used to derive the context. The context level is defined at each sentence of the document. The links between the nodes are generated using thematic context derivation algorithm.

Table 1. PageRank, Context Vector , Modified Context Vector and importance for sample context terms

PR(t,d)	cv(t,d)	mcv(t,d)	t ε G	i_t
0.358	1.153	0.413	play	0.0612
0.358	2.656	0.952	fit	0.0079
0.529	6.01	3.179	bowl	0.0183

For deriving F-measure, recall and precision the system is tested on BBC news dataset.

Table 2 gives F-measure, recall and precision of various systems when compared to LDA. The CVM is more effective than Named entity vector, Forms of word vector and LSA with SVD. But when used combined with others the performance is increased in F-measure and recall. The precision is the same for CVM and eCVM.

Table 2. Context vectors terms compared to LDA term vectors

Techniqu e	F-measure	Recall	Preci sion
Named Entity Vector	0.5	0.4	0.68
Forms of word vector	0.48	0.37	0.68
LSA with SVD	0.52	0.42	0.75
CVM	0.61	0.48	0.83
eCVM	0.63	0.5	0.83

Table 3 shows F-measure, recall and precision of various methods used to derive context of document when compared with eCVM. The precision of the system when compared with eCVM is same for Named Entity, LSA with SVD and LDA but the F-measure and recall are varying.

Table 3. Context vector terms compared to eCVM

Techniqu e	F-measure	Recall	Precision
Named Entity Vector	0.59	0.46	0.83
Forms of word vector	0.56	0.45	0.75
LSA with SVD	0.58	0.45	0.83
CVM	0.62	0.47	0.86
LDA	0.63	0.5	0.83

Table 4. F-measure of proposed system with existing systems

Context Derivation Technique	F-measure
Existing method in [16]	60.5%
CVM in [2]	62.5%
Proposed method using eCVM	63.5%

The F-measure of the eCVM can be compared to that of existing system in [2,16]. The comparison of the proposed system with other methods of context derivation is done in Table 4. eCVM performs better than [16] and uses named entities along with other NLP tasks. Also it performs better than CVM presented in [2].

V. CONCLUSION

The system proposed in this paper devised a context vector machine, eCVM, which considers named entities along with word embeddings. eCVM gives modified context vector for the term in the document corpus along with its Pagerank and importance. eCVM is compared with existing methods of context derivation. The eCVM shows increase in its F-measure by 4%, 7%, 3%, and 1% when compared to context derived from Named Entities, forms of words, LSA with SVD and CVM respectively. eCVM gives context phrases for each document along with their importance based on the context with respect to other documents as well. The context phrases can be used to cluster documents which are similar in context. eCVM performs better than existing context derivation methods. Further, eCVM can use attention networks and transfer learning methods for word embeddings.

REFERENCES

1. Jiwei Li and Dan Jurafsky 2015 Do Multi-Sense Embeddings Improve Natural Language Understanding? Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics: 1722–1732.
2. Khatavkar V and Kulkarni P 2019 Trends in Document Analysis. In Data Management Analytics and Innovation 810 : 249-262 Springer Singapore
3. Khatavkar Vand Kulkarni P 2019 Comparison of Support Vector Machines With and Without Latent Semantic Analysis for Document Classification. In Data Management Analytics and Innovation : 263-274 Springer Singapore
4. Nagrale D Khatavkar V and Kulkarni K 2019 Document Theme Extraction Using Named-Entity Recognition. In Computing Communication and Signal Processing : 499-509 Springer Singapore
5. Benedetti F Beneventano D Bergamaschi S and Simonini G 2019 Computing inter-document similarity with context semantic analysis. Information Systems 80 : 136-147
6. Shu L Long B and Meng W 2009 A latent topic model for complete entity resolution. In Proceedings of the 2009 IEEE International Conference on Data Engineering 880-891 IEEE Computer Society
7. Khatavkar V and Kulkarni P 2017 December Use of noun phrases in identification of a website. In 2017 International Conference on Big Data IoT and Data Science (BID) : 103-109 IEEE
8. Khatavkar V and Kulkarni P 2016 Context vector machine for information retrieval. In International Conference on Communication and Signal Processing 2016 ICCASP 2016 Atlantis Press 137 375-379
9. Chen D and Manning C 2014 A fast and accurate dependency parser using neural networks. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) : 740-750
10. Blei D M Ng A Y and Jordan M I 2003 Latent dirichlet allocation. Journal of machine learning research 3 : 993-1022
11. Hofmann T 1999 Probabilistic latent semantic analysis. In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence 289-296 Morgan Kaufmann Publishers Inc
12. Deerwester S Dumais S T Furnas G W Landauer T K and Harshman R 1990 Indexing by latent semantic analysis. Journal of the American society for information science 41-6 : 391-407
13. Wang J Liu J and Wang C 2007 May Keyword extraction based on pagerank. In Pacific-Asia Conference on Knowledge Discovery and Data Mining 857-864 Springer Berlin Heidelberg
14. Meng L Huang R and Gu J 2013 A review of semantic similarity measures in wordnet. International Journal of Hybrid Information Technology 6(1) : 1-12
15. BBC news dataset available at : http://mlg.ucd.ie/datasets/bbc.html
16. Das M and Cui R 2019 Comparison of Quality Indicators in User-generated Content Using Social Media and Scholarly Text. arXiv

AUTHORS PROFILE

Vaibhav Khatavkar completed Master of Engineering in Computer Science-Information Technology, from Vishwakarma Institute of Technology, Pune in May 2010. He completed his Bachelor of Engineering in Computer Engineering from Pune University in May 2007. He is actively involved in various Linux awareness programs, technical talks and workshops on topics like machine learning, artificial intelligence, Python programming, latex etc . He previously worked with Ascellent technologies from June 2007 to July 2008 as an exposure to the industry. He is presently working as Assistant Professor at Computer Engineering Department and pursuing his PhD from College of Engineering, Savitribai Phule Pune University. His research interests include machine learning, computer networks and operating systems. In International Conference on Communication and Signal Processing 2016, he was awarded the best paper award for the paper titled 'Context Vector Machine for information retrieval'.

Makarand Velankar completed basic graduation in computer engineering from Walchand college of Engineering, Sangli in 1990. After working in the industry for 11 years as development engineer he joined MKSSS's Cummins College of Engineering, Pune in 2001 and is presently working as Assistant Professor in Information Technology Department. HE completed a masters in Computer Engineering in 2003 from PICT, Pune University. He is pursuing a PhD from PICT, Pune University in the domain of Computational musicology. His research interests include music analysis, machine learning, artificial intelligence, Information retrieval, algorithms, soft computing etc. He has presented research papers in different journals and prestigious conferences such as Coling 2012, Acoustics 2013 etc. and delivered invited talks at ICIASP 2013, FRSM 2017 etc. During springer conference ICIC 2017 held at Pune, he received a best paper award for the paper titled 'Unified Algorithm for Melodic Music Similarity and Retrieval in Query by Humming'. He is also actively involved in developing entrepreneurship culture and mentoring startups.

Dr. Parag Kulkarni is one of the world's leading authorities on Business Strategy, Knowledge Innovation, Machine Learning, Systemic Learning and Building Innovative Knowledge Corporations in the knowledge economy. He is consultant on Innovation and Strategies for start-ups and SMEs and contributed to make many startups successful.

He holds a PhD from IIT Kharagpur, Management education from IIM Kolkata. UGSM Monarch Business School – Switzerland conferred higher doctorate - DSc on him for his contribution towards innovation and knowledge management. Recipient of Oriental foundations scholarship, he is also a Fellow of IETE and The IET. He has been visiting professor/researcher at technical and B-schools of repute including TIU - Japan, IIM, Masaryk University – Brno, COEP Pune. Parag headed various organizations and research labs and contributed to the success of more one dozen organizations through his strategic and Business acumen and innovative product building. He is core contributor to more than dozen commercially successful products. He headed Research Labs at Siemens, IDeaS and many other organizations. One of the well-known Product Innovation and Business Innovation strategist, he is an advisor to many industries and academic institutes. His core work include innovation and knowledge strategies for startups, knowledge strategies for organizations, building innovative products, breaking away from competitive landscape. His work is on systemic innovation and learning is published with many reputed journals.