Computer Program for Modeling the Patterns of Parts of Speech

M.M.S.Rauthan, Pritam Singh Negi and H.S.Dhami*

Dept. of Computer Science,
H.N.B.Garhwal University,
Srinagar, (Garhwal) Uttarakhand

ABSTRACT
The present work is an attempt in the direction of writing of computer programs for defining texts in the form of vector algebra and their basis so that pattern of occurrence of parts of speech could be modeled in the form of Markov Chain.

INTRODUCTION

Enabling computers to understand language remains one of the hardest problems in present world. One of the biggest obstacles in making full use of the power of computers is that they currently understand very little of the meaning of human language. Recent progress in search engine technology is only scratching the surface of human language, and yet the impact on society and the economy is already immense.

Stephanie Chua (2008) has explored the effects of different POS on text categorization effectiveness. Turney and Pantel (2010) have presented a survey of Vector Space Models and their relation with the distributional hypothesis as an approach to representing some aspects of natural language semantics. VSMs perform well on tasks that involve measuring the similarity of meaning between words, phrases, and documents. Most search engines use VSMs to measure the similarity between a query and a document. Andrews and Vigliocoo (2010) have described a model, which they refer as Hidden Markov Topics model, for learning semantic representations from the distributional statistics of language.

Pandey,Rauthan and Dhami , in their earlier work (2009) have attempted in the direction of modelling the patterns of parts of speech in texts and have been able to build a theoretical base for studying the stylistic pattern of different writers by adopting vector space approach and making use of Markov chain . The present work in an attempt in the direction of making software to translate the mathematical formulae into a form that CPU can understand.We have taken the corpora available at the cite

http://www.geocites.com/theloepa/gand_eng.html and have selected the book entitled "Godmen of India" written by Sean Richards.

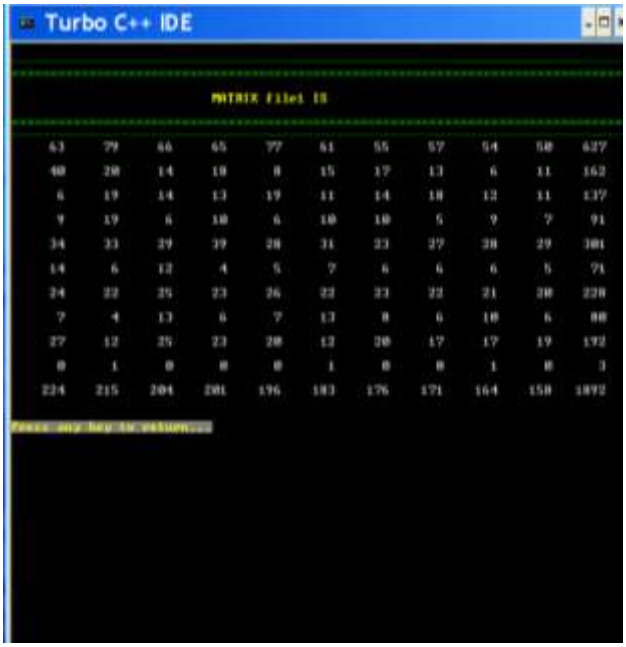
*Prof.H.S.Dhami, Head, Dept. of Mathematics, Kumaun University, Nainital

Logic, flow and description of various important algorithms/ functions used in the program and related snapshots for Modeling of Parts of Speech Patterns and Exhibition of Relation between Stylistic Patterns of Different Writers are being given in ensuing pages:

(Main Module)

Algorithm for Entering the Matrix value creating with the help of Parts of Speech present in the Corpus

1. Read input file name
2. Create an array a[11]
3. Count<- 1;
4. Do while count<-11
5. for i<-0 to 10
6. enter integer value store in a[i]
7. count<- count+1
   (end of for)
8. write content of a[ ] into file
   (end of while)



(Original matrix according to the parts of speech present in the corpus)

Probability Module:
This module takes an input as a 2D matrix of 11x11order and produce output as a probability matrix of 10x10 orders. Original matrix is saved in a file. Probability module calculates conditional probability, that a particular part of speech occupies a particular position when other parts of speech may also exhibit that position.

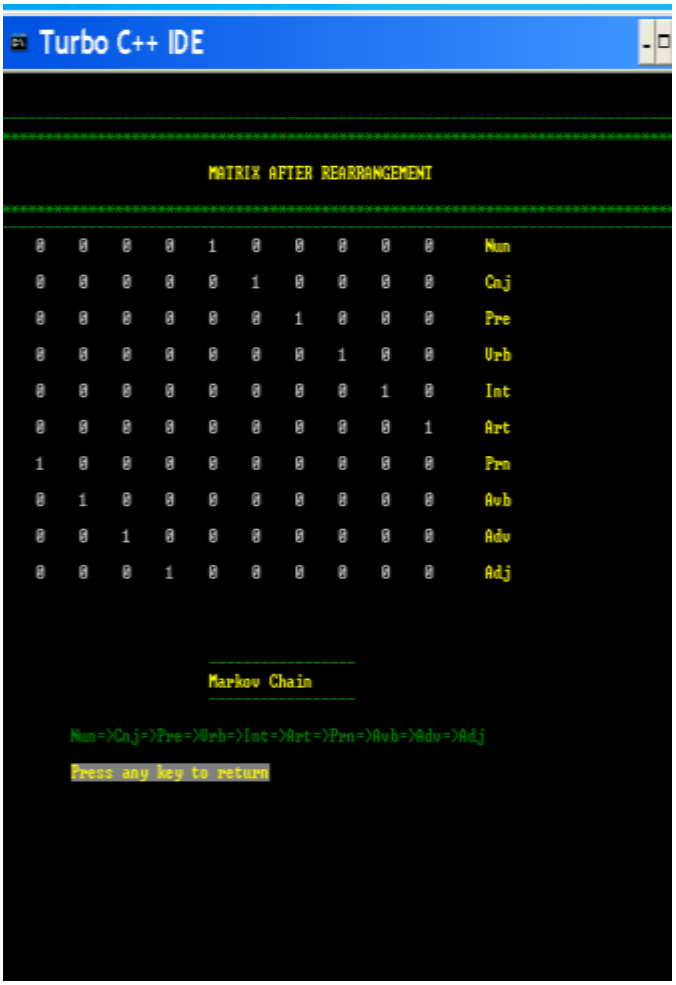
Algorithm for probability:
This function calculates conditional probability of a matrix.

1. Create float array prob[10][10]
2. Read the file where original matrix is present
3. Read content of file and copy them into matrix[11][11]
4. for i<-0 to 9
5. for j<-0 to 9
6. prob[i][j]<-matrix[i][j]/matrix[10][j];
   (end of for)
   (end of for)
   (end)

Standard Basis :

The style of a writer for a particular text can be defined in the form of vector algebra, and their basis and patterns of parts of speeches for particular writer can be defined by Makov Chain. This module calculates Standard Basis for vector space which represents the position of ten parts of speech in a selected corpus. The output of this module includes

(Standard Basis Module)

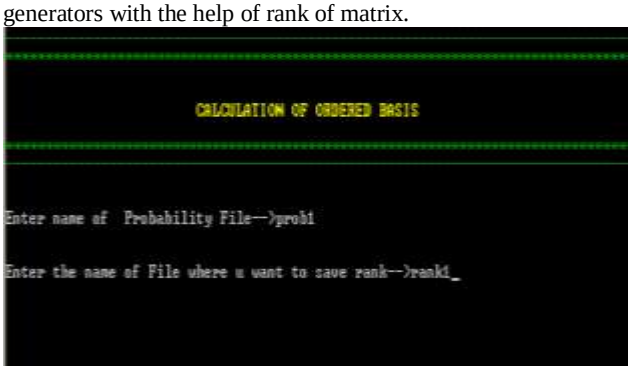


(Markov Chain)

Ordered Basis :

By existence theorem, it can be said that the supposed finite dimensional vector space shall have at least one basis and by another theorem it can infer that for the vector space of 10 dimensions, any linearly independent set with 10

elements of the vector space shall form a basis of the vector space. The basis have been generated for every vector space by deriving a linear independent set from the set of generators with the help of rank of matrix.



(Ordered Basis Module)

Algorithm for Solution()
This function calculates determinant and transpose of inverse of matrix of 10x10 order and save inverse in a file.

1. Enter the name of file where matrix is saved
2. Read contents of matrix and copy them in d1[10][10] (transpose d1[10][10] and save it into d[10][10])
3. For i<-0 to 9
4. For j<-0 to 9
5. D[i][j]<-d1[j][i]
(end of for)
(end of for)
(calculation of determinant)
6. Deter<-0
7. For i<-0 to 9
8. Deter<-Deter + $(-1)^I$ *d[0][i]*det9(d,I,10)
(end of for)
(calculation of Cofactor)
9. For i<-0 to 9
10. For j<-0 to 9
11. Po<-i+j
12. Cofact[i][j]<- $(-1)^{po}$ * cofactor(d,i,j,10)
(end of for)
(end of for)

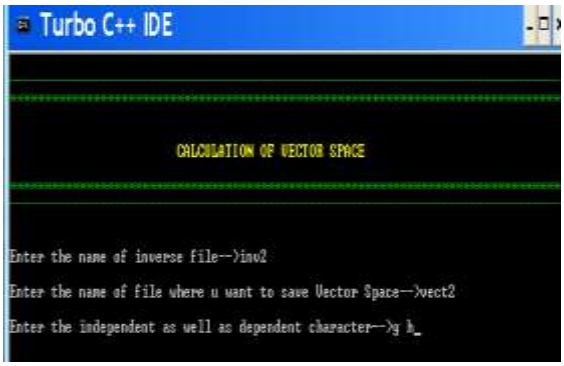
(calculation of inverse)
13. For i<-0 to 9
14. For j<-0 to 9
15. T[i][j]<-cofact[j][i]
16. T[i][j]<-t[i][j]/deter
(end of for)
(end of for)
17. Save inverse in a file
(end)

Vector Space:

Within sentences, whether simple or non simple, there are various kinds of part. For example, all the clause of a complex or compound sentences are constituents of the sentence as a whole; in a simple in a sentence all the words forms are constituents and group of words may constitute phrases , which are also constituents, of the sentence. This notion of constituency, coupled with a somewhat more general version of the traditional concepts of the phrases, is at the very heart of the formalization of grammatical structure in Chomskyan generative grammar. The 10 tuples taken are basically the linguistic constituents of all types of sentences. Each element of a vector space is formed by a linear combination of basis elements. This module gives output as a vector space having 10 tuples.

Algorithm of Vector space()

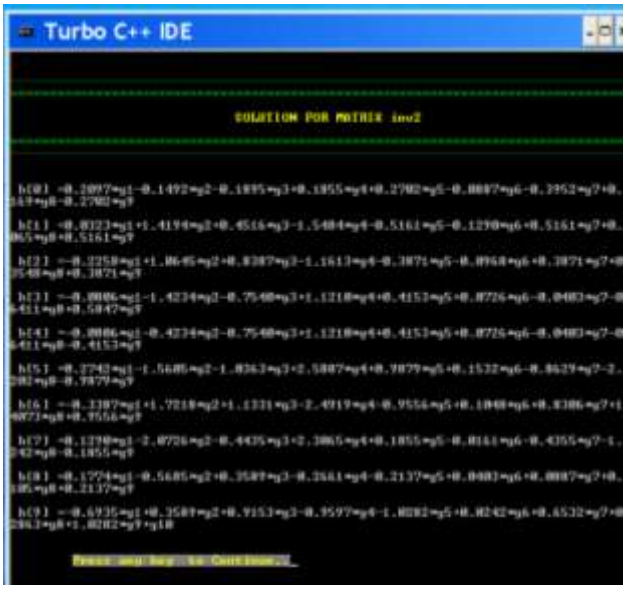
1. Enter name of file containing inverse matrix
2. Read inverse from file & copy it into inv1[10][10] matrix
3. Create a character array r[10][3]



(Vector Space Module)

4. Create a character array solu[10][160]
5. For i<-0 to 9
6. r[i] <- i+1
(end of for)
7. For i<-0 to 9
8. For k<-0 to 9
9. If inv1[i][k]!=1.000 && inv1[i][k]!=0.0000 then
10. Change element inv1[i][k] into string and concatenate it to solu[i]

```
                & now concate '*' and r[k] into
                solu[i]
11.         Else if inv1[][k]=1.0000 then
12.         Concate solu[i] and r[k]
13.         Else if inv1[i][k]=0.0000 then
14.         Continue the for loop
15.     If k!=10 then
16.     Concate solu[i] and "+"
        (end of if-else block)
        (end of for)
    (end of for)
    (Vector representation)
17. print "(a1,a2,a4,a5,a6,a7,a8,a9,a10)="
18. for i<-0 to 9
19.     print "("
20.     for j<-0 to 9
21.     print inv[i][j]*aj
22.     if j<9 then
23.         print "+"
        (end of if)
        (end of for)
24.     print ")e"i
    (end of for)
    (end of function)
```



(Inverse Module Showing Solution Of Matrix)

REFERENCES

1. Mark Andrews, Gabriella Vigliocco(2010) Topics in Cognitive Science 2 (2010) 101–113,Copyright 2009 Cognitive Science Society, Inc. ISSN: 1756-8757 print / 1756-8765
2. Peter D. Turney and Patrick Pantel (2010) From Frequency to Meaning: Vector Space Models of Semantics, Journal of Artificial Intelligence Research 37 (2010) 141-188.
3. Rakesh Pande, H.S.Dhami and M.M.S. Rauthan (2009) Modelling of part of speech patterns and exhibition of stylistic patterns of different writers, Shekhar (New Series) International Journal of Mathematics, Vol.I, Issue I, pp.139-153.
4. Stephanie Chua (2008) The Role of Parts-of-Speech in Feature Selection, Proceedings of the International Multi-Conference of Engineers and Computer Scientists 2008 Vol I, IMECS 2008, 19-21 March, 2008, Hong Kong.