A new approach for efficaciousPattern frames in text mining

Ramesh kumar Mojjada (Mtech CSe) email: rameshkumar.mojjada@gmail.com
Maharaja venkata gajapathi college

I.Abstract:

The discoveries on mining and to achieve the perfect and effectual patterns are still open challenge. But so many existing discoveries do not follow the offset of the prefix duplications before/after clustering. Here in this, we proposed how to get large and best patterns from huge data sets. For that we take a new approach after cleaning the documents called ***MMR*****[1]**(Marking and merging resolute)aproach. ***MMR*** is used to mark the documents individually and merging the possible patterns with respect to the markers**[2]** in a single document for further processing like search or discovery for swaped terms. Based on the final marked document the feasible patterns will be discovered based on the input term vector with efficatious way. Then the possible best and feasible combinations will be discovered from priority blocks by using ***BFT*** block framing technique based on starting terms as first two terms in search sentence. ***(index items: prefix , term , marking ,possible combinations,blocking)***

II. Introduction:

In recent years as the data is increasing the techniques for storing or preserving and to mine that particular data in effective way is the biggest challenge. So when the data is getting stored, the effective approach of mining is always benefit for data processing**[3]** further. The ancient aproach is the searching of the large data after clear clustering with respect to search input term with criteria. This criteria is totally distinct for various mining techniques. Some will be term approach, some will be categorical approach, some will be segment and some will be novel aproach.

In this work ***MMR*** (Marking and merging resolute) based on the 0 offset term the spots/markers will be placed in the available documents. The patterns(at the placed markers) will be gathered in with respect to left and right terms padding technique. If a mark is placed at p_{th} offset then the term will be discovered as follows.

$M_p(d_2)$ - Mark is placed in at p_{th} offsetin document d_2.

$\underbrace{\textbf{Pa} – [] [] []}$ t2 [] t1 [] [] [] t2 []

Pa_1 and as $t_2$is duplicated last t_2 will be eliminated from pattern. And this Pa1 (t2 [] t1 [] [] [])will be appended to the final document which will contain all the discovered patterns from all the documents. Once the final document is framed with all discovered patterns, with respect to first and second term segmentation will be generated. The internal thresholdλ**[29]**value will be auto generated to present the best patterns discovery.

III.Related work

Preprocessing of the available documents

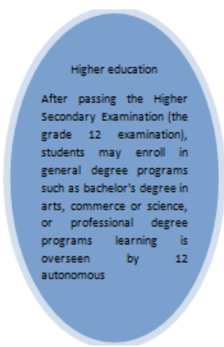
- Stop words removal
- Stemming
- Pruning
- weightage

As the above steps are common in preprocessing the next ***MMR*** and ***BFT*** approaches are unique in this experimental work. The main general data items sets are always challenge for new approaches but ***MMR*** and ***BFT*** experimentally done processing on 5 data sets which are general educational related datasets with with the following representations with representations with size .

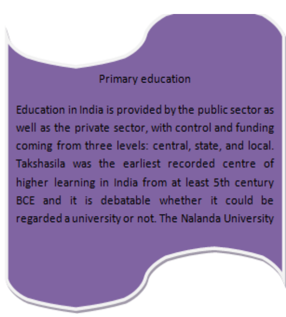
1. ***he***: higher education
2. ***pe*** : private education
3. ***pm*** : primary education
4. ***se*** : secondary education
5. ***see*** : secondary education extension

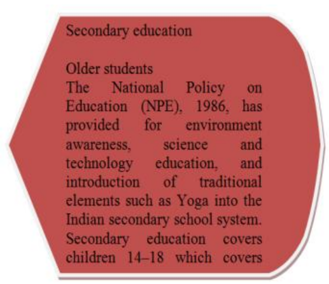
Datasets	Size in kb
he	5
pe	4
Pm	4
Se	2
See	3

The main data sets as follows and the total processing of data is on 21kb.



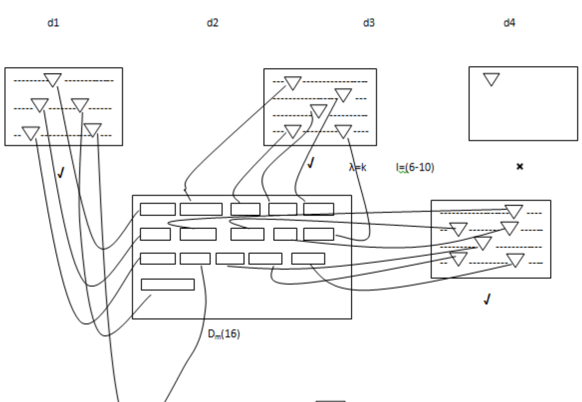
He(higher education) pe(primary education





Pe(primary education) se(secondary education)

See(secondary education)



V. Pattern model

Document	Marked weight
D1	5
D2	6
D3	6
D4	1

The pattern model is in one single model based on starting or prefix term. The consideration of +ve and –ve documents based on threshold where the markers are less with particular prefix term.The marked negative documents will be ignored for merging. Markers will be discovered in one structure so that the pre and post terms will be considered with respect to marked term. So from the marker right side ***5*** preceding terms and left side ***5*** terms will be the main sentence from that if left side ***5*** terms second term found then the pattern will be discovered from that second term . If the second term found in right side preceding ***5*** terms that term will removed and pattern will be end with pre term of that duplicated term in that pattern. This continues for all the markers in all the given documents.

Once all discovered patterns structured in one merge document based on ***t1*** and ***t2*** two blocks naming ***t1***term pattern block and ***t2***term pattern block will be framed. In these blocks the prefix ***t1*** patterns and the prefix ***t2*** patterns will be available with***t1***prefix patterns with high priority.

Considering the first block which is ***t1***term pattern block if the preceding term is ***t2*** in the patterns those patterns will be the priority patterns in the first block. The same factor will be applicable for the second block also and rest of the patterns will be outliers .So excluding the outliers remaining patterns are efficacious patterns.

V. Marking approach

***MMR*(Marking and merging resolute)**

Input : all documents Dn

Output: all marked documents vector MD_s

D_n - document vector (number (n) documents as initial documents)

$\sum_0^J T$- Total term vector

T_1 – seed / prefix term (high priority terms)

T_2 – preceding seed/ post fix

MD_s←marked documents

$\sum M\ (di, off) \leftarrow 0$ //Marking vector

n ← 0

loop start

for each di in Dn

// iterate through document from top level

If (T_1 == d(n))

M ← Mark($d_{i,n}$)

I = i+1

end if

$MD_s \epsilon$ M

n←n+1

end for

n←0

end for

In the above algorithm the output is all marked document vector.

MDs =*{ d1s , d2s , d3s , d4s, d5s}*

In this work some minimum marking range, threshold applying mechanism to eliminate the documents from the marked vector(all marked documents with marked positions)

Mark function: Basically this function will take offset and document for marking the prefix term in the relevant document. Here even preceding term also can be marked but, our to challenging technique is based on first prefix term based mark second term also will become prefix based on the offset consideration in the left and right to the first prefix term. So if preceding term starts at ***m-3***rd position(***m*** is the prefix current mark and the discovered pattern will start with second preceding term as prefix and ends with right 5, (m+position of second term) of second term. After that since second term is duplicated the second preceding term in the pattern will be deleted. This continues for all the documents and one single document will be framed with all prefixes as first term and also second term.

Merging with 5 pre post terms for marked term notation

Input: MD_s
Output: efficacious patterns document E_d
Initialization:
$T_p \leftarrow 0$ // temporary patterns
n← 0
for each marked document d_j in MD_s
m←marked value
$\boldsymbol{T_p=\{T_{m-4}\ T_{m-3}\ T_{m-2}\ T_{m-1} T_m\ T_{m+1}\ T_{m+2}\ T_{m+3}\ T_{m+4}\ T_{m+5}\}}$
P ← CHECK (TP)
Ed=P
End for
End loop
CHECK function will check for duplications with respect to terms given effective
Pattern for blocking

(*BFT*)Blocking framing technique

Initialization→ Mds (marked document stack)

With out blocking

$T_1 \leftarrow$ term1 $\boxed{\text{B1} \leftarrow \text{blocker1}\lambda=4}$

$T_2 \leftarrow$ term2 $\boxed{\text{B2} \leftarrow \text{blocker2}\lambda=3}$

ΔOL←outlear oueve

For each P in Md_s(patterns)
If (T_1 == P(0))
$B_1 \equiv P$
Else
$B_2 \equiv P$
End if
0_i = Outlier (B1)
0_s= Outlier (B2)

OL={0,u o_s} //total documents
$F_{ep}=\{B_1 \cup B_2\} - \{OL\}$ // Final efficacious patterns

VI. Example:

O*s* = **t*1* t*2* t*3* t*4* t*5* …….t*n***

Estimation: {t2, t5} - to be cleaned terms ex: ***"of" "ing"***etc..

In this case ***t1*** is the starting or prefix term

After cleaning the search or input term vector is ***{ t1 t3 t4 t6 …….tn }***

Let [29]λ = **m** (m is the threshold for search or input term length)

So the the resultant search or input term vector size would be n-1 starting with 0 means m terms will be there in the vector.

$$\underbrace{\{\, t1\ \ t3\ \ t4\ \ t6\ \ \ldots\ldots .tn \,\}}_{m}$$

so in the above term vector *tn* is in the ***m-1***th index.

Ds = *d1 d2 d3 d4 d5………dn* (total dynamically selective documents) which are to be processed with respect to t1 as the starting term which in this case as prefix term.

Dm is the single document after marking and merging with respect to first term(***t1***) and this document is the final document to locate the best pattern evolution.

VII. Future work

Once the merged document is framed taking count for all the merged patterns consider the prefix for first term and by considering more number of times with respect to second term the patterns will be collected and for each pattern invert matrix will be generated for generating prefix mining with flexibility of any distinct term.

Example if the pattern is **'one two three four five one two four'**

The generated invert matrix is as follows

1--- two 4,1 5,2
2--- five 3,1
3--- one 1,1 1,2
4---three 5,1
5--- four 2,1

Considering 'one' as prefix and possible combination of prefix mined patterns are as follows.

one three
one four
one two
one five
one three five
one four five
one two five
one three two
one four two
one five two
one three four
one two four
one five four
one four three
one two three
one five three

Considering ‘two’ as prefix and possible combination of prefix mined patterns are as follows.

two one
two three
two four
two five
two one five
two three five
two four five
two one four
two three four
two five four
two one three
two four three
two five three
two three one
two four one
two five one

	One	two	three	four	five	six
One – 1	5,1	5,2	3,3	2,4		
Two – 4	2,1	4,2	$	$		
Three -2	2,1	$	$	6,4		
Four - 5	4,1	4,2	$	3,4		
Five -6	$	$	5,3	2,4		
Six -3	6,1	$	3,3	4,4		

VIII. Conclusions

Though lot of mining techniques are available this unique approach of offset shuffling marking with blocking algorithms gave phenomenal results for generic data sets. The main aim to take generic datasets is to have unique approach to discover efficacious patterns with global data.

IX. Acknowledgements:

A good repetitive and investigative support from Mr B.Srinivas Mtech to write this paper. After lot of reviews and discussions this work has come to an end to discover efficacious patterns in text / data mining.

X. References

[1] S.T. Dumais, “Improving the Retrieval of Information from External Sources,” Behavior Research Methods, Instruments, and Computers, vol. 23, no. 2, pp. 229-236, 1991.

[2] J. Han and K.C.-C.Chang, “Data Mining for Web Intelligence,” Computer, vol. 35, no. 11, pp. 64-70, Nov. 2002.

[3] J. Han, J. Pei, and Y. Yin, “Mining Frequent Patterns without Candidate Generation,” Proc. ACM SIGMOD Int’l Conf. Manage- ment of Data (SIGMOD ’00), pp. 1-12, 2000.

[4] Y. Huang and S. Lin, “Mining Sequential Patterns Using Graph Search Techniques,” Proc. 27th Ann. Int’l Computer Software and Applications Conf., pp. 4-9, 2003.

[5] N. Jindal and B. Liu, “Identifying Comparative Sentences in Text Documents,” Proc. 29th Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ’06), pp. 244-251, 2006.[6] T. Joachims, “A Probabilistic Analysis of the Rocchio Algorithm with tfidf for Text Categorization,” Proc. 14th Int’l Conf. Machine Learning (ICML ’97), pp. 143-151, 1997.

[7] T. Joachims, “Text Categorization with Support Vector Machines: Learning with Many Relevant Features,” Proc. European Conf. Machine Learning (ICML ’98),, pp. 137-142, 1998.

[8] T. Joachims, “Transductive Inference for Text Classification Using Support Vector Machines,” Proc. 16th Int’l Conf. Machine Learning (ICML ’99), pp. 200-209, 1999.

[9] W. Lam, M.E. Ruiz, and P. Srinivasan, "Automatic Text Categor- ization and Its Application to Text Retrieval," IEEE Trans. Knowl- edge and Data Eng., vol. 11, no. 6, pp. 865-879, Nov./Dec. 1999.

[10] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text Classification Using String Kernels," J. Machine Learning Research, vol. 2, pp. 419-444, 2002.

[11] A. Maedche, Ontology Learning for the Semantic Web. Kluwer Academic, 2003.

[12] C. Manning and H. Schu ̈tze, Foundations of Statistical Natural Language Processing. MIT Press, 1999..

[13] K. Sparck Jones, S. Walker, and S.E. Robertson, "A Probabilistic Model of Information Retrieval: Development and Comparative Experiments—Part 2," Information Processing and Management, vol. 36, no. 6, pp. 809-840, 2000.

[14] R. Srikant and R. Agrawal, "Mining Generalized Association Rules," Proc. 21th Int'l Conf. Very Large Data Bases (VLDB '95), pp. 407-419, 1995.

[15] S.-T. Wu, Y. Li, and Y. Xu, "Deploying Approaches for Pattern Refinement in Text Mining," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1157-1161, 2006.

[16] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, "Automatic Pattern- Taxonomy Extraction for Web Mining," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI '04), pp. 242-248, 2004.

