An Efficient method of Web Document Clustering with Semantic Representation of Documents

Raja Varma Pamba[1], Dr Elizabeth Sherly[*2]

[1]*Assistant Professor,*
Dept of CSE,LBSITW,Trivandrum
[2]*Professor & Head,*
IIITMK,Technopark,Trivandrum

***Abstract*—The huge influx of information being retrieved before the user at the time of search poses the biggest challenge of overcoming the irrelevant information from the search space .This paper focusses on its novel approach in handling this issue of irrelevancy in two aspects. Firstly by representing documents in a much more effective manner taking into account the concepts or the meaning of the terms in the document. Second by using Frequent Pattern Growth principle of association analysis to help find frequent item sets is also discussed. The evaluation reveals an enhancement in Accuracy for the proposed approach compared to traditional term frequency approach**

***Keywords*— Universal Networking Language, ontologies,**

I. Introduction

A beeline of research is in the area of application of machine learning and information technology on web platforms. The data from the web are rather more of complex nature to handle as they are generated asynchronously. The web data is more of dynamic ,unstructured with complex attributes. To deal with web data of complex attributes we need a finer representation of these data .Bag of words representation commonly used cannot in these cases work intelligently due to synonym problem and polysemy problem. For instance having a word "semantics" in one and "meaning" in another in the normal syntactic search parlance is considered as two different words with no semantic relation, adding upto two different frequency words. In the vice versa if a word "bear" appears as a noun and as a verb the existing system fails in identifying and counts the frequency to two .In this proposed system first measure taken to confront this issue of synonymity is by using concept vectors.i.e a single concept vector mapping "semantic" and "meaning" to a common word could help resolve the issue.

The paper is organized on the following lines: Section 2 covers the related work. Section 3 describes the methodology used. In section 4, we describe proposed approach adopted. The results are covered in section 5 and finally conclusion is presented in section 6..

II. Related works

Chaudhary and P.Bhattacharyya [1] described a new method of feature vector representation using UNL[2].In UNL every document is explicitly represented as a semantic graph with universal words as concepts and semantic relation between them as links. When this new breed of data is well represented for a semantic search space, then various extended nature inspired clustering algorithms are used for better results. To help better clustering Association analysis especially FP Growth principle is used for discovering interesting relationship hidden in large data sets. FP-growth is a divide and conquer strategy that mines a complete set of frequent item sets without candidate generation.

3. Universal Networking Language

UNL is an inter lingua in the form of semantic network to represent and exchange information.In UNL a sentence can be considered as a hyper graph where each node is the concept and the links or arcs represent the relation between the concepts.UNL consists of Universal words(UWs),Relations and Attributes and Knowledge Base. UNL Ontology is lattice structure where UWs are inter connected through relations including hierarchical relations such as icl(a-kind-of) and iof(an-instance-of).Universal Words are used as components of the document vectors as in [1].Each UWs is represented in an unambiguous manner because of which multiple words in the document get automatically differentiated producing correct frequency count. For Example,in the sentence {"A bear can bear very cold temperatures"} the word bear has two different notions, bear(icl>animal) and bear(icl>tolerate).In a normal statistical representation the frequent counts to 2 rather in UNL, due to its unambiguous manner these two words find different places in document vectors. With the property inheritance characteristic of the UW system, possible relations between lower UWs are deduced from their upper UWs and thus reduces the number of binary relation descriptions of the UNL Ontology .For example Ram went to China from India by aeroplane to attend a conference, are shown in Fig. 1All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

```
{unl}
$agt(go(icl>move>do,plt>place,plf>place,
agt>thing).@entry.@past,Ram(icl>person))
plt(go(icl>move>do,plt>place,plf>place,a
gt>thing).@entry.@past,China(iof>asian_c
ountry>thing))
plf(go(icl>move>do,plt>place,plf>place,a
gt>thing).@entry.@past,India(iof>asian_c
ountry>thing))
```

```
met(go(icl>move>do,plt>place,plf>place,a
gt>thing).@en
try.@past,aeroplane(icl>heavier-than
air_craft>thingequ>airplane))
obj:01(attend(icl>go_to>do,agt>person,ob
j>place).@entry,conference(icl>meeting>t
hing).@indef)
pur(go(icl>move>do,plt>place,plf>place,a
gt>thing).@ent ry.@past,:01)$
  {/unl}
```

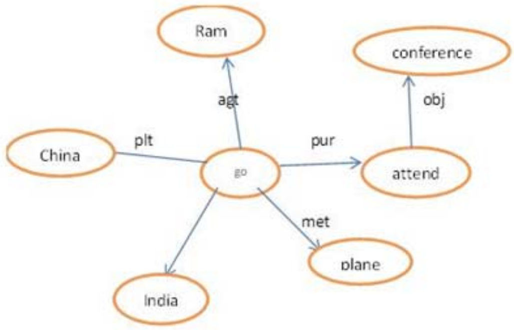

Fig1:Semantic Graph

3.1 UNL Document Vector Generation Algorithm

The basic assumption taken in this contexts[1] is that the more the number of links to and from a Universal Word, the more is the importance of the word in the document. Thereby calculating the weight of the concepts in the semantic graph is determined by the number of links to the node. The generation of document vectors and UWIDF from the UNL ontology is discussed in [1] as follows:

1. parse the UNL document to construct the UNL graph
2. For each UW in the UNL graph count the links to/from other UWS from ii
3. Adjust the count depending upon the relation connecting given UWs like transferable relation, equal weight relations, partial transferable relations and non-transferable relations.
4. if transferable relation in case of {agt,obj,aoj,cag,pur,ptn,rsn},weight w_c of the child node w_$_{c}$=weight of the parent node w_$_{p+1.}$ If equal weight relations in cases of relations like {and,or,cnt,scn,pof,pos,coo,seq} then $_{w\_c= w\_p.}$
5. if in partial transferable relation cases like {ben,ins,met,opl,plc,plf,plt,to, via} then w_$_{c}$=w_$_{default}$+1
6. if non-transferable relation w_{link} =1
7. Construct the concept vectors by merging the counts of each UWs for a given web document got from step 3
8. Multiply this count with its IDF where IDF(t)=log(N/N$_t$),N is the total number of documents ad N$_t$ is the number of times the term has occurred.
9. Generate the final concept vectors by assigning counts of UWs to their corresponding positions in the vector.

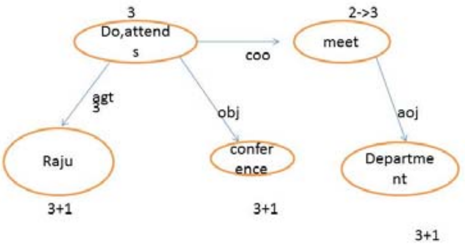

Fig2: Raju attends conference after the department meeting

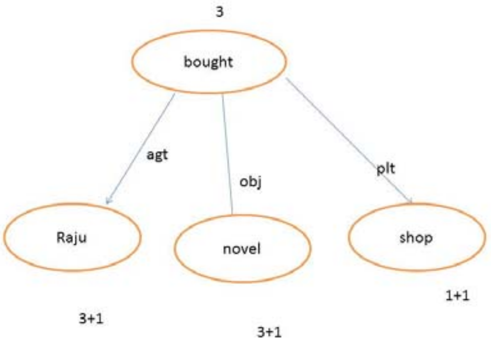

Fig3: Raju bought the novel from the shop

Considering Fig2 and Fig3 the weights for each UWs for the same can be computed as follows based no step 3 of Algorithm 1:

X$_1$=<4,3,4,3,4>
X2=<4,3,4,2,4>

The Universal words or concepts considered are {Raju, read, novel, watch, television ,bought, shop} and the final document vectors corresponding to the UWs computed as per step 4 of Algorithm 1 are given below:

V$_1$ =<4,3,4,3,4,0,0>
V$_2$ =<4,0,4,0,0,3,2> accordingly.

3.2 FP Growth Algorithm

3.2.1 Frequent Pattern Tree Construction

Let I={a$_1$,a$_2$,....,a$_m$} be a set of items and a transaction database DB={T$_1$,T$_2$,....,T$_n$} is a transaction which contains a set of items in I. This is assumed equivalent to term document matrix. In this Transaction are equivalent to Documents D$_k$=D$_1$,D$_2$,.....,D$_n$ and items equivalent to keywords or terms as extracted by UNL.

3.2.2 FP Tree Algorithm[2]

Input: A Term Document Matrix (Transaction database, D_B) and a minimum support threshold taken as 3 in this proposed approach.

Output: FP Tree, frequent pattern tree

Method:

- Two main scanning. First scanning of the Tern Document Matrix or Transaction database. Collect the set of frequent items(F) and their support counts. Sort F according to descending support count as F_{list},the list of frequent items.
- Create the root of an FP-tree, T, and label it as "null". For each document considered as transaction I_{Tran} in D do the following, Select and sort the frequent items in I_{Tran} according to the order of F_{list}. Let the sorted frequent list in I_{Tran} be [p | P], where p is the first particle as per the nomenclature used in Fuzzy Particle Swarm Clustering and P be the remaining list of particles in the population or the search space. Call inserttree([p|P],T), which is performed as follows. If T has a child N such that N.item-name=p.item-name then increment N by 1;else create a new node N,with a value =1,its parent link linked to T,and its node link linked to the nodes with same item-name.If P is nonempty, call inserttree(P,N)recursively.

3.2.3 FP Growth Algorithm:Mining frequent patterns with FP tree by pattern fragment growth[2]

Input: A database D, represented by FP tree constructed according to Algorithm 1 and minimum support threshold, xi

Output: Set of Frequent Patterns

Method: call FP-growth(FP-tree, null)

- if Tree T contains a single prefix path then {
- let P be the single prefix path part of tree
- let Q be the multipath part with the top branching node replaced by a null root
- for each combination of the nodes in the path P do
- generate pattern β U α with support=minimum support of nodes in β
- let fpset(P) be the set of patterns so generated else
- let Q be tree
- for each item a_i in Q do
- generate pattern $\beta=a_i$ Ǔ α with support $=a_i$.support

  construct β conditional pattern base and then conditional FP Treeβ,

  if tree β#0,then call FPgrowth($Tree_\beta$,β)

    - let fpset(Q) be the set of patterns so generated .
- return(freq pattern set(P) ∪ freq pattern set(Q) ∪ (freq pattern set(P) × freq pattern set(Q)

4. PROPOSED METHODOLOGY

- Generation of Bag of concepts
- Generation of Document vectors using UNL Ontology
- Create UW document matrix which is same as transaction database for FP tree construction
- FP Growth principle to find the frequent item sets which is considered as number of clusters
- Evalution of Accuracy where accuracy=No of documents correctly clustered/Total number of documents.

5. EXPERIMENTAL RESULTS

Method	Accuracy
TF-IDF+FP growth	.85
UNL-IDF+FP growth	.96

Fig 4:Accuracy Computation.

The results shows a promising way ahead for using this methodology for clustering.

6. CONCLUSION

In our proposed approach we have tried to find an efficient methodology for clustering of documents based upon semantic representation of documents and enhancing the clustering by using FP Growth algorithm. The accuracy predicted shows that our model have attempted to develop a system more meaningful and concept oriented. We have proposed a model taking into account the constraints of the traditional approaches of clustering. This methodology proposed assures for a semantic representation of documents along with computation of clusters with the help of FP growth algorithm. Also, there is a lot of room for improvement. For example, clustering algorithm is time consuming, which also produces increase of computational complexity of the method. We need to speed up clustering algorithm in future.

REFERENCES

[1] Choudhary B and P.Bhattacharyya,Constructing Better Document vectors using Universal Networking Language in the ELeventh Internation World Wide Web Conference,2002.

[2] Jiawei Han,Jian Pei,Yiwen Yin, Mining Frequent Patterns without Candidate Generation:A Frequent-Pattern Tree Approach, Data Mining and Knowledge Discovery, 8,53-87, 2004

[3] Uchida,H and M.Zhu,The Universal networking Language,Specification Version 3.0,United Nations University:Tokyo,Japan,1998.

[4] Walaa K.Gad and Mohammed S Kamel, Enhancing Text Clustering Performance Using Semantic Similarity, ICEIS,LNBIP 24, PP.325-335, 2009.